Journal of Intelligent Systems and Internet of Things
Vol. 17, No. 02, PP. 50-63, 2025
DOI: https://doi.org/10.54216/JISIoT.170205

Efficient Spam Email Detection Model based on Dynamic Embedding with Deep Learning Classification

Salam Al-augby[1,*], Zahraa Ch. Oleiwi[2], Hasanen Alyasiri[1], Fahad Ghalib Abdulkadhim[1]
[1] Faculty of Computer Science and Mathematics, University of Kufa, Najaf, Iraq
[2] Faculty of Computer Science and Information Technology, University of Al-Qadisiyah, Dewaniyah, Iraq
Emails: salam.alaugby@uokufa.edu.iq ; zahraa.chaffat@qu.edu.iq; hasanen.alyasiri@uokufa.edu.iq; fahadg.abdulkadhim@uokufa.edu.iq

Abstract

One of the major concerns when transitioning emails is the potential influx of unsolicited and unwanted spam emails. These unwanted emails can clog inboxes, causing recipients to overlook important messages and opportunities. To ensure security and avoid the destructive and dangerous effect of these spam emails, machine learning and deep learning methods have been conducted to design spam detection models. In this work, a combination of embedding models and multi-layer artificial neural networks as deep learning classification models is utilized in order to introduce an approach to spam detection. The proposed classifier leverages the Bidirectional Encoder Representations from Transformers (BERT) model for word embedding, applied to the Enron-Spam dataset, offering a noteworthy technique for considerable spam detection. Experimental results demonstrate that the proposed spam detection model achieved a 99% recall rate for detecting spam emails. Notably, this model is a step forward in generality and improving the efficiency of spam detection. It presents a good attempt at presenting a solution for detecting spam emails and fake text within communication environments.

Received: January 09, 2025 Revised: March 19, 2025 Accepted: May 23, 2025

Keywords: Sam email; BERT model; Embedding models; Deep learning

1. Introduction

Email presence has changed the way of communication because it facilitates data exchange easily in addition to spread the concepts and written messages across the world[1]. The email can be considered a way of transferring files that contain text, images, web links, as well as an electronic communication between senders and recipients either individuals or groups [2] .

Spam emails can be defined as an unwanted message that can be sent in massive amounts, which pose significant challenges. Email is widely adopted by many organizations due to its ability to distribute information to the wide public. However, spammers exploit this by employing phishing attacks to trick users into clicking on malicious links. Email spoofing, where emails appear to be from known contacts, is another common threat. To address these issues, companies have developed advanced spam detection tools and techniques. For example, Google's Gmail boasts a 99.9% success rate in filtering out spam. The concept is applied through machine/deep learning techniques, automatically detecting and eliminating spam emails according to the trained models [3].

Classification technique was applied in machine learning to classify data using different strategies[4, 5]. Such strategies include Naïve Bayes (NB) and Random Forest (RF) techniques, Decision Trees, Support Vector Machine (SVM), and AdaBoost[6-8]. For example, supervised learning refers to cases where data entries are marked as “ham” or “spam.” SVM uses hyperplanes to divide data points in multi-dimensional space with examples including spam and ham emails[7]. Random forests, however, can provide classification,

regression, and other tasks based on an ensemble learning background[9]. AdaBoost is a supervised learning algorithm which belongs to ensemble models and its main function is to combine diverse weak learners into one strong learner[10]. Classification is made in the statistical model of logistic regression (LR) by making classifications based on the probabilities of events [11].

One of the first steps in the text email classification involves the process of word embedding which means transforming textual data into numbers through the use of methods like BERT[12]. Word embedding is an important part of Natural Language Processing (NLP), it economizes on features, helps retrieve semantic meanings of the words, gets contextual information, and helps in learning through the use of ready-made embeddings [13]. Therefore, this method makes it possible for proper performance even when there is a small amount of task-related data. As an example, word embedding which does not need any supervision is an important procedure for text categorization with sentiment orientation advice among other applications. One of the most useful approaches is generating word embeddings by BERT[14]. By definition, BERT is usually taken as a generic pre-trained language model but it can be adapted to various NLP tasks [15]. More precisely, to build up and develop a text categorization system, the use of embedding models, together with classification models based on machine or deep learning is an absolute necessity.

However, even with the advancements in spam detection technologies, it is often a challenge for the more traditional methods to properly handle the newly emerging spam practices. Most existing approaches use either simple keyword-based filtering or very shallow learning algorithms which are not resilient to changes in the spam.

Studies are scarce on the application of BERT technology with deep-learning network layers hence there is a research gap. The majority of the studies done so far have spent most of their time using BERT for feature extraction or word embedding and have not used a deep-learning classifier with BERT. The goal of this study is to construct a new, more efficient design for the detection of spam using BERT by leveraging its possibilities.

For this purpose, taken as embedding models BERT model combinations and as classification models of deep learning instead of static SoftMax used multi-layer artificial neural networks including dynamic embedding with deep learning classifier.

In order to address these challenges, this paper proposes a spam email classification model which is more advanced than the existing models. The model utilizes dynamic embedding techniques and deep learning classifiers to dynamically represent email content and implement advanced deep learning architectures. The goal is to significantly improve accuracy in detecting spam emails and ensure the ability to deal with new spamming techniques. This research aims to make a meaningful contribution to the cybersecurity field by thoroughly evaluating our proposed approach against established benchmarks and providing a scalable solution to the widespread problem of spam emails.

The paper is structured as follows: Section 2 provides an overview of the related work. Section outlines the materials and methods utilized in this study. Section 4 details the proposed models. Section 5 examines the results and analysis of the proposed techniques. Lastly, Section 6 offers the conclusions.

2. Related Work

Email has revolutionized the way we communicate, allowing us to effortlessly exchange data, concepts, and written messages across the world. Numerous researchers have dedicated extensive efforts for developing effective spam detection mechanisms. This literature review covers the latest research focused on utilizing deep learning for the purpose of spam detection.

A study in [16] was conducted to develop a method for detecting spammers in the Facebook social network. The approach relied on analyzing various features at both the content and user levels. Two different learning algorithms, Naive Bayes and decision tree induction, were employed to identify spammers. In an effort to enhance spammer detection, an integrated approach that combined the strengths of both algorithms was proposed. The performance of each algorithm was evaluated based on its accuracy in detecting spammers and non-spammers. The integrated approach outperformed other methods by achieving an overall accuracy of 94.1% and accurately identifying non-spammers with 99% accuracy. However, the algorithm's accuracy in detecting spammers was lower at 68.5%. It is evident that the proposed integrated algorithm was able to identify an account as spammer or non-spammer with 88.1% accuracy. The algorithms accuracy for finding non spammers was higher i.e., 99%. But at the same time, the accuracy of spam detection by spammers is

lower, i.e., 68.5%. In addition, this work achieves an overall recall of 81.13% and an overall F1-score of 88.27.

The authors of [17] describe the implementation of a method for encoding emails using the orders, requests, and general and classifying emails using imperative sentences. This method utilises the Word2Vec model for generating vectors and employs two deep learning techniques: Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). The experiment was performed using a dataset of 1,000 emails, comprising personal Gmail accounts and Enron emails. The results of the experiment indicated enhanced accuracy for the RNNs as compared to the CNNs. More specifically, the RNNs reached an accuracy of 94.9%, with a 90% recall, while the CNNs registered 86% accuracy and 93% recall. These results were achieved under the 90:10 condition of separating training and testing datasets. The experiment showed that accuracy dropped when the size of the training dataset was reduced. This indicates that the accuracy of the model, which the experiment specifically focused on, is highly sensitive to the size of the training dataset.

The Universal Spam Detection Model (USDM) was unveiled in [18] as an advanced tool for classifying spam using BERT. This model utilized a combination of diverse datasets as input and integrated individually trained models for multiple datasets. It also featured the addition of dropout layers above and below the batch normalization layers. The USDM exhibited significant performance improvements, delivering an impressive F1 score, as well as demonstrating acceptable precision and recall values. Ultimately, the model achieved an outstanding 97% accuracy, with an impressive F1 score of 0.97. One of this research pros is increasing the dataset size which is crucial for improving model accuracy, especially in deep learning. Consolidating multiple datasets can contribute to achieving better results by considering the unique nature, complexity, and specific attributes of each dataset, though it may also lead to a reduction in accuracy and performance metrics. The study did not explore potential solutions to overcome the inconsistency among the merged data. The data merging process might have achieved a better outcome if the imbalance had been tackled.

This dissertation in [19] centered around the multiple methods of text classification through the use of embeddings like Word to Vector (Word2Vec), Global Vectors (GloVe), and BERT. It achieved classifier methods with the highest accuracy and the lowest false positive rates through the synergistic use of machine/deep learning. The achievement of the classification model showed that BERT with embeddings and machine learning in combination surpasses all other methods. Deep learning self-feature extraction and machine learning combination models for classification were evaluated across 4 datasets, with the BERT+SVM hybrid model yielding the highest accuracy and the lowest false positive rate. Despite the results of this work are promising, the recall metric for classifying spam emails was consistently around 90%, which is within the acceptable range of 90%. This is considered an issue due to the fact that the significance of such metric is crucial for determining spam efficiently.

In [20], a significant approach for pointing out spam emails has been introduced. The method leverages a pre-trained BERT model, a powerful transformer-based language representation, in conjunction with machine learning algorithms. By processing email texts through BERT, the study extracted valuable features to represent the content. Subsequently, four different machine learning classifier algorithms were employed to categorize the text features as either legitimate (ham) or spam. The performance of the suggested model was evaluated using two distinct public datasets, revealing that the logistic regression algorithm exhibited the best classification accuracy across both datasets. These findings underscore the efficacy of the developed model in effectively discerning spam emails from legitimate ones. The assessment of the research results would be simplified by utilizing a confusion matrix, classification report, and a well-defined method for splitting the data into test and training sets.

The work in [21] examined the use of machine learning and deep learning algorithms to classify and eliminate spam emails from large collections. To ensure comprehensive coverage, four diverse datasets were used: Track Dataset-2007, Enron dataset, PU dataset, and Lingspam dataset. Both XGBoost algorithm and word embedding combined with Long Short-Term Memory (LSTM) were employed, yielding impressive results in terms of accuracy across the utilized datasets. Furthermore, word frequency analysis was complemented with word clouds of spam words within each dataset. This involved extensive feature engineering and data cleaning to extract and prepare comprehensive features, which were then transformed into numerical vectors.

Moreover, the dataset was carefully split into training, testing, and cross-validation sets for thorough model validation.

To summarize, the above studies employ a variety of machine learning models, for instance Logistic Regression (LR), XGBoost, Support Vector Machines (SVM), and Random Forest (RF), as well as deep learning approaches such as word embedding and LSTM. This combination allowed for a thorough evaluation of spam detection methods. In this work, BERT model, a prominent word embedding technique, is suggested for enhancing outcomes of this research and may generate more favorable results. Nonetheless, it is important to recognize that the results generated by the machine learning models utilized should not be underestimated. All aforementioned studies of literature analysis are summarized in Table 1.

Table 1: Summary of the related works of spam detection

Ref.	Dataset	Study objective	Used method	Evaluation result	Key Findings
[16]	Facebook dataset	Detecting spammers in the Facebook	Naïve Bayes, Decision Tree, Integrated Approach	**Accuracy:**94.1% (overall),68.5% (spam) **Recall:**81.18% **F1-score:** 88.27%	Integrated algorithm effective in identifying non-spammers (99% accuracy) but **68.5%.** for spam.
[17]	**SMS spam dataset**	Email categorization	Word2Vec, CNN, RNN	**Accuracy:** 94.9% (RNN), 86% (CNN) **Recall:** 90% (RNN), 93% (CNN)	-The RNN model achieved higher accuracy than the CNN model. -The performance of the models is significantly affected by the size of the training set.
[18]	Ling-spam dataset, Spam text messages dataset, Enron dataset, and Spam assassin	Universal Spam Detection Model for spam classification	BERT	**Accuracy:**97% **F1-score:**97%	Combining multiple datasets enhanced performance; however, data inconsistency posed a challenge.
[19]	Dataset 1: SMS Spam Collection Dataset, Dataset 2: Enron Dataset, Dataset 3: Spam Assassin Dataset, Dataset 4: Ling Dataset	Text classification	Word2Vec, GloVe, BERT, SVM	**Recall:** below 90% for spam detection	The BERT-based model achieved high accuracy in spam detection; however, there were concerns regarding its recall.
[20]	Dataset 1: Enron-Spam dataset Dataset 2: spam or not spam dataset	Classification of Spam email	Pre-trained BERT, Logistic Regression	**Precision:**97.85% for dataset 1 and 95.9% for dataset 2 **Recall:** 97.85% for dataset 1 and 96% for dataset 2 **F1-score:** 97.85% for dataset 1 and 95.9% for dataset 2	BERT effectively extracted features, and logistic regression demonstrated the best performance.

[21]	**Enron-Spam dataset**	Classifying and eliminating spam emails **from large collections of emails**	Different machine and deep learning methods: LR, RF, XGBoost, SVM, and LSTM	**Recall:** 92% for RF **Precision:**91% for RF	This approach effectively assessed spam detection methods. Although the BERT model is recommended for improvement.

3. Material and methods

3.1 Data Description

The proposed model was utilized with the Enron-Spam dataset, which is a significant collection of data compiled by V. Metsis, I. Androutsopoulos, and G. Paliouras. This dataset is detailed in their publication "Spam Filtering with Naive Bayes - Which Naive Bayes?"[22]. Nevertheless, the main dataset is organized in a way that keeps every email in an own text file, which are then distributed across multiple directories [2]. This dataset, which can be accessed by the public on the Kaggle platform, consists of a substantial collection of 33,716 emails. It is a useful tool for assessing how well our model performs. The Enron-Spam dataset stands out due to its well-balanced distribution of spam and ham emails, establishing an environment that closely resembles email exchange in the real world[23]. The data has been distributed as in Table 2

Table 2: The distribution of Enron-Spam dataset

Class label	Original dataset	train	test
Ham	16545	12,409	4136
Spam	17171	12878	4293
total	33716	25287	8,429

3.2 BERT Model

The BERT, short for Bidirectional Encoder Representations from Transformers, along with other Transformer encoder architectures, has achieved remarkable success across a wide range of tasks in NLP. These architectures are capable of generating vector-space representations of natural language that are well-suited for utilization in deep learning models [24, 25]. The BERT family of models leverages the Transformer encoder architecture to process each token of input text within the comprehensive context of all tokens both preceding and following it, hence the name: Bidirectional Encoder Representations from Transformers. Typically, BERT models are first pre-trained on a substantial corpus of text and subsequently fine-tuned for specific tasks [26].

Its architecture is equipped with transformers that efficiently capture long-range dependencies, as well as self-attention mechanisms and feedforward neural networks. BERT's attention mechanism operates using Query (Q), Key (K), and Value (V). After performing a linear transformation, these constituents are utilized in the scaling dot product to dynamically generate weights for various connections. In the context of self-attention, Q\ is equal to K. The dimension of Q and K is represented by d_k. Scaling the dot product is important to prevent it from growing too quickly; if left unaddressed, this problem may lead may cause the gradient of the softmax function to become too small as in equation (1):

$$\text{Attention}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \text{softmax}\left(\frac{\mathrm{Q}\mathrm{K}^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{V} \quad (1)$$

The multi-head attention mechanism involves running N self-attention processes concurrently. Each of thes e self-attention processes is referred to as a "head." To increase the flexibility of self-attention, each $head_i$, where $I \in N$ is an element of N, does not operate on the original Q, K, and V. Instead, it is assigned a unique

set of random parameter matrices for Query (W_i^Q), Key (W_i^K), and Value (W_i^V). This allows each $head_i$ to i ndependently learn its own attention map as in equation (2) [27]:

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (2)$$

The BERT model operates without the need for labeled data (unsupervised), and its multi-head attention mechanism allows it to focus on various aspects of the input data independently, enabling a more comprehensive understanding of the context [28].

The feedforward network (FFN) is a fundamental component of neural networks and is composed of two lin ear transformations with a Rectified Linear Unit (ReLU) activation function in equation (3) applied in betw een as:

$$f(x) = max(0, x) \quad (3)$$

This architecture enables the model to effectively capture and process features located at various positions w ithin the input sequence. Notably, these feedforward networks are incorporated within both the encoder and decoder components subsequent to the attention mechanism, demonstrating their importance in sequence pr ocessing tasks. The FFN function is defined in equation (4),

$$FFN(x, W_1, W_2, b_1, b_2) = max(0, xW_1 + b_1)\, W_2 + b_2 \quad (4)$$

Where W_1, W_2, b_1, b_2 are learnable parameters.

Sometimes the GELU (Gaussian Error Linear Unit) activation in equations (5) and (6) is also used instead o f ReLU[29]:

$$x\Phi(x) \quad (5)$$

Where:

$$\Phi(x) = P(X \leq x), X \sim N(0,1) \quad (6)$$

In addition, BERT utilizes bidirectional encoding to thoroughly analyze the context of each word by considering all the words in the sentence. It also possesses pre-training and fine-tuning capabilities, multi-layer stacking to capture complex contextual information, Word Piece tokenization for embedding layers, and Masked Language Modeling (MLM) to comprehend the context of words in relation to their surroundings [30, 31].

4. Proposed Model

The proposed model started with attaining the Enron email dataset, which is often regarded as Sen ’s, among others, precious e-mail corpus for anti-spam research. In order to understand the ordering and understanding of the structures of the dataset, its types and the spam and non-spam ratios, the comprehensive survey of the dataset was performed. This critical evaluation is essential so as to ascertain that the data set meets the requirements for the advanced model that is being put forth.

Data preprocessing is the most important and one of the first steps in every aspect of NLP, since a lot of processes depend on the data quality. In this research, generally available software Natural Language Toolkit (NLTK) package was used for the stop words elimination from the corpus. This step involves use of the list of stop words internally provided by NLTK consisting of 40 words inclusive of a, an, the, of, whilst allowing for some words that may change the meaning of the emails and that change the meaning of the emails are carefully avoided as this step aims to preserve the literal meaning of the text as much as possible.

One of the main goals of the preprocessing phase is to enhance the quality of the data by trying to eliminate undesired information in the dataset, by concentrating on useful words. This step is an added advantage to researchers and practitioners who deal with NLP tasks because it enables the training of models more intelligently and efficiently, thus doing away with some constraints when making predictions. One of the main steps of data preparation to improve the computational efficiency and remove the redundant information to stop word removal, but it is important to take into consideration that this process may implicitly cause

translation problems due to the loss of some intricacies of the textual context. This balance is of utmost importance during the preprocessing phase.

NLP tasks cannot be carried out effectively without engaging in data preprocessing for as mentioned in [32]. The text data is altered in a series of procedures to rid it of any imperfections and present it in a manner suitable for the task at hand. One such approach relates to the reduction of the sequence length of the inputs to be fed to the model. This is the procedure which removes unnecessary words such as the, be or and which is performed frequently, removal of stop words. Train less resources and less time understand that as the size of the dataset is shrunk, it helps in using even more benefits. But most importantly, removing stop words is a great means of advantage but at the same time many times will totally remove the proper meaning of the sentences which must not lose. However, it's important to note that removing stop words can potentially lead to a loss of the correct meaning in sentences[33]. The proposed methodology is illustrated in Algorithm 1.

Algorithm 1: dynamic embedding with a deep learning

Input: raw dataset (spam and ham emails)

Output: confusion matrix

1. **foreach** email ∈ dataset **do**
2. remove stop words
3. join word for word with the label and save it in a csv file
4. encode labels (spam (1) and ham (0))
5. **end**
6. Use Keras to preprocess the csv file
7. Split the csv file into train_X: train features, train_y: train labels, test_X: test features, test_y: test labels
8. Encode the preprocessed text using BERT.
9. Normalize the encoded text.
10. Apply dropout to prevent overfitting.
11. Apply 32 neurons with leaky ReLU activation.
12. Normalize the output.
13. Apply dropout.
14. Apply 64 neurons with leaky ReLU activation.
15. Normalize the output.
16. Apply dropout.
17. Apply 128 neurons with leaky ReLU activation.
18. Normalize the output.
19. Apply dropout.
20. Use sigmoid as activation function
21. Run built model on test_X: test features, test_y: test labels
22. Return confusion matrix

A. Model Architecture

The proposed dynamic embedding with a deep learning classifier is designed with a multi-layered architecture is based on BERT architecture, as shown in Figure 1. The architecture consists of several key components:

- Input Layer: This layer takes the preprocessed email text as input.
- Preprocessing layer, which is BERT's preprocessor directly imported from Keras, is utilized to preprocess plain text inputs into the input format expected by BERT, resulting in a sequence length of 128. This layer provides multiple methods for converting one or more batches of text segments (encoded as UTF-8 plain text) into the inputs required for the Transformer encoder model for ensuring compatibility with the BERT model, which expects inputs in a specific format.
- BERT Layer: The BERT layer utilizes transformer encoders to create a deep neural network for natural language processing. It is designed to understand the context of words in a sentence by considering the surrounding words, resulting in more accurate language understanding and representation. The BERT layer (transformer encoders) does not consider any embedding or encoding separately, and borrows

embedding from transformer neural network rather than applying some conventional embedding and encoding techniques.

- Multi-Layer Architecture: The defined four blocks interleaved with the batch normalization layers, dropout layers and dense layers with different number of neurons which has been activated with Relu. The effectiveness of dropout layers is clearly observed since overfitting is reduced and on the side of batch normalization, such helps in enhancing the performance of the model because it standardizes the outputs from the previous layers. This process of normalization prevents drastic changes in the learning process and enables reduction of the time taken to train the model.
- Output Layer: The last network layer corresponding to a classifier part of a transfer learning model is a sigmoid activation function, which is used to distinguish spam from non-spam in a text email. It is this layer that performs the prediction tasks in a classification problem as it helps in determining the output from the learned representations of the former layers enunciated.

B. Training Procedure

The proposed model refers to using supervised learning prominently. In this case the dataset was divided into training and test sets as shown in Table (1) at the proportions of 75 - 25, standard sizing rule. There are some crucial steps that should be adopted during the training phase such as:

- Loss Function: This phase needs to use a function that is appropriate for multi-class spare targets; therefore; this model employed the binary cross-entropy loss function, which can measure the error of the target class and the probability distribution produced by the model that gives a clear view of the innovative model.
- Optimizer: The Adam Optimizer was used for optimizing the training process. The reason behind adapting Adam Optimizer is because of its reliability in tending to change the learning rates efficiently while training particularly if the gradients are sparse. This model is utilized in our model due to its ability to efficiently train of the deep learning model.

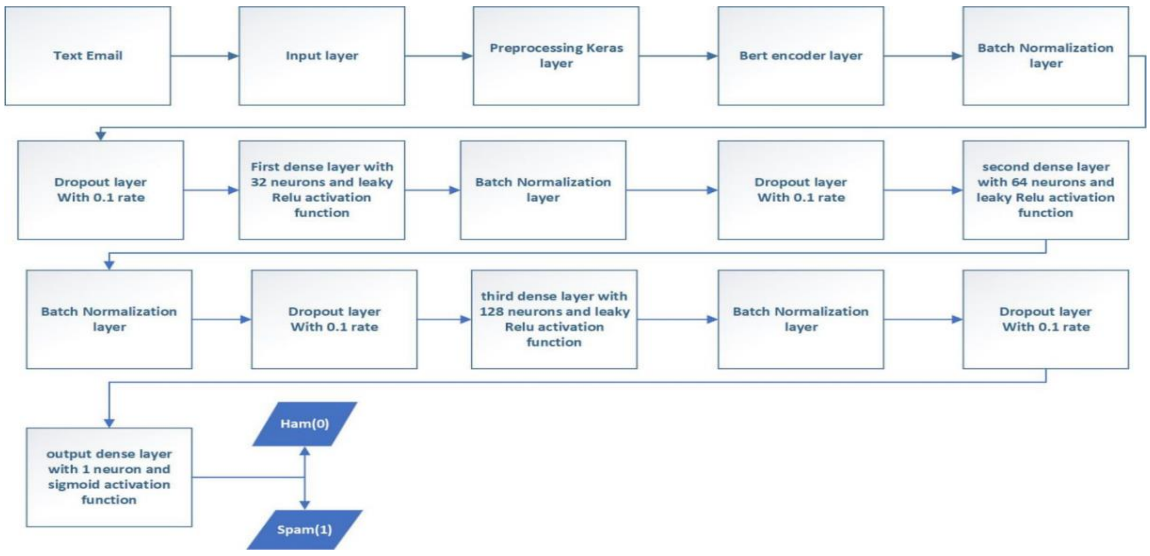

Figure 1: The general architecture of proposed dynamic embedding with a deep learning classifier model.

- Evaluation Metrics: A variation of evaluation metrics such as accuracy, precision, recall and F1 score was employed for the purpose of evaluating the comprehensive performance of our model. Using these metrics will assist in measuring the performance of the model due to its capability of accurate email classification. The performance of this model on different classes was achieved by using a confusion matrix that gives us a perspective about the performance in the positive class and in the negative class.

Results and Discussion

The Enron-Spam data set was split into training and tested parts when conducting the proposed model. Furthermore, for the purpose of assessing the performance number of metrics was used: Accuracy (ACC), recall (true positive ratio (TPR)), precision (positive prediction value (PPV)), and f1-score. The terms used for calculating these metrics are true positive samples (TP), true negative samples (TN), false positive samples (FP), and false negative samples (FN) [34]. Table 3 depicts the performance of the proposed spam detection model. The proposed model employing dynamic embedding with the deep learning classifier.

Table 3: The performance results of the proposed spam detection model using Enron-Spam dataset.

Class label	Precision%	Recall%	F1_score%	Accuracy %	TP	TN	FN	FP
Ham(0)	99	81	89	90	3354	4259	782	34
Spam(1)	84	99	91					
Macro average	92	90	90					
Weighted average	92	90	90					
Class label	Precision%	Recall%	F1_score%	Accuracy %	TP	TN	FN	FP
Ham(0)	99	81	89	90	3354	4259	782	34
Spam(1)	84	99	91					
Macro average	92	90	90					
Weighted average	92	90	90					

One can indicate that the used classifier is categorized the majority of the emails for the overall accuracy that is about 90%. This model is dependable option for email filtering system that's come from the model's ability in distinguishing between spam and non-spam emails. One of the main problems in analyzing dataset is the overfitting therefore, it's essential to contextualize accuracy within the distribution of email classes, especially in datasets where one class may significantly outnumber the other. The obtained results showed a comprehensive understanding of the model's performance and ability in real-world scenarios.

As an interpretation of the given results the ham emails are well classified because of the high precision (99%) that is crucial in spam detection because misclassifying legitimate emails as spam can cause significant disruptions for users. The precision spam ratio of this model is (84%) in spite of this ratio the model accurately identifies a good part of spam emails, there is still room for improvement to reduce the number of false positives.

The recall ratio is 99% of spam classification means the capability of this model to successfully identify actual spam emails and hence improving the process of lowering the risk of sneaking through the filter. The recall

high ratio leads to increase the probability of exposed the fishing attacks and unwanted advertisements. The main aim of classifying process is to identify and classify spam emails. Recall ratio is the crucial metric for evaluating the performance in this area is the recall specifically for spam. It's worth noting that the recall rate for detecting spam is an impressive 99%, indicating a strong potential for successfully accomplishing the primary objective of this research.

The F1-score serves as a trade-off measure of precision and recall, especially valuable in scenarios where class distribution is imbalanced. With an F1-score of 90% for both ham and spam classes, the model effectively balances the trade-off between precision and recall, demonstrating its robust performance across both classes.

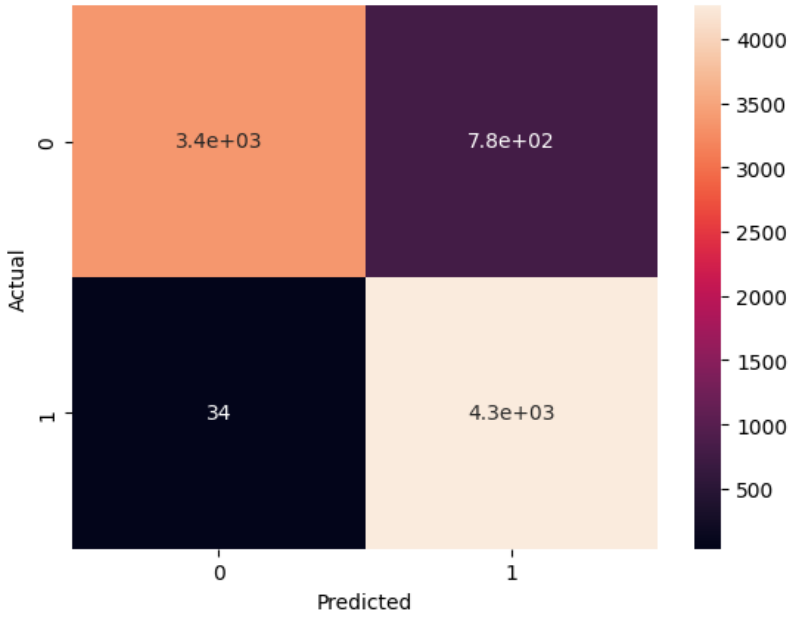

Figure 2. Confusion matrix of the proposed spam detection model using the testing dataset from the Enron-Spam dataset

Figure 2 shows the confusion matrix that employed in analyzing the suggested spam detection model, The confusion matrix was used for assessing and visualizing the efficiency the proposed model. Based on this matrix one can noticed that model shows accurately predicted 3354 out of 4136 ham samples and 4259 out of 4293 spam samples that means only 34 spam samples out of 4293 instances are misclassified which demonstrating lower computational complexity.

One of the main steps in processing and classifying the dataset is the feature extraction step thus if this step is improved the performance of model will be improved. The suggested model improved the performance of this model by increasing the number of hidden layers and structure of deep learning model. The imbalanced datasets are one of main issues of dataset processing which leads to diminishing returns and performance degradation. In our experiments, it is observed that while the recall for spam reached an impressive 100%, the recall for ham plummeted to less than 20%.

This phenomenon can be attributed to overfitting, where the model becomes excessively complex and begins to memorize the training data instead of generalizing from it. As the model's depth increases, it may start to prioritize the classification of spam emails, effectively learning to identify subtle features specific to spam while neglecting the characteristics of ham emails. This imbalance reflects a critical trade-off in model training: optimizing for one class can inadvertently compromise the performance on the other, particularly in datasets with inherent class imbalances. Consequently, while achieving perfect recall for spam may seem advantageous, the substantial drop in ham recall indicates that the model fails to maintain a balanced perspective necessary for effective spam detection. Employing regularization, careful architecture design,

and balanced training data was confirmed by achieving perfect recall for effective spam detection to preserve the model's robustness across all classes, which led finally to boosting its utility in real-world applications.

Table 4 shows a comprehensive evaluation of the proposed model for spam detection, performance comparison with other reviewed studies. This comparison included a set of factors that have a crucial effect in achieving outcomes. These factors are the used dataset, the employed methodology, the dataset size, in addition to the evaluation metrics used, such as precision, recall, and F1-score.

Table 4: Comparison of the proposed spam detection model with previous related works

Ref.	dataset	model	Number of samples	precision	recall	F1-score
[17]	SMS spam dataset	BERT-deep neural network	5574	100%	87%	93%
[21]	Enron-Spam dataset	Logistic Regression	33062	88%	86%	-
[21]	Enron-Spam dataset	Support Vector Machine	33062	88%	81%	-
[21]	Enron-Spam dataset	XGBoot	33062	89%	87%	-
[21]	Enron-Spam dataset	Random Forest	33062	91%	92%	-
[35]	SMS Spam Collection	CNN-LSTM	8304	95%	87%	91 %
[36]	SMS spam collection dataset	Logistic Regression	5572	93%	86%	-
[36]	SMS spam collection dataset	K-Nearest neighbors	5572	80%	60%	-
[36]	SMS spam collection dataset	Decision Tree	5572	95%	86%	-
The proposed model	Enron-Spam dataset	BERT-deep neural network	33716	84%	99%	91%

The analysis result of the comparison presented in Table 4 clearly demonstrates that our newly proposed spam detection method, which combines dynamic embedding with a deep learning classifier, outperforms all previous related methods in accurately predicating spam. The proposed model shows its effectiveness by predicting the majority of spam cases, which is clearly shown by obtaining a significant 99% recall rate for spam instances. One of the main reasons was the advanced capabilities of BERT, which can learn intricate and context-rich word representations due to its pre-training on a vast corpus of text data, enabling it to detect subtle features.

The proposed model in this work achieved high generalization, making it an efficient model for implementation in real applications. However, it is worth noting that the work referenced in[17], [21], [35], and [36] achieved slightly higher precision than the model proposed in this work. This was

achieved despite using a smaller number of samples for testing, indicating a narrower variety in the dataset used in these related works. It is important to consider the impact of dataset variety on the model's generality when comparing different approaches.

6. Conclusions and Future Works

This research is conducted by combining of dynamic embedding with deep learning classifier in which BERT and artificial neural network layers are utilized in being applied to the Enron-Spam dataset. This combination led to a promising result. BERT model characteristics especially its efficacy and ability in capturing long-range dependencies, the contextual meaning of each word in a sentence, self-attention mechanisms, and bidirectional encoding are the main reason that make BERT suitable for complex contextual information capturing and in simplifying spam detection process. Additionally, it is noted that a less complex classifier or a reduced number of hidden layers can lead to higher detection accuracy. Conversely, increasing the number of layers may decrease accuracy due to the overfitting problem. The performance and accuracy of deep learning are improved as it is trained on larger datasets.

In the future, this proposed model can be considered an efficient and less complex model that can pave the way to handle big data and leverage cloud systems to ensure secure message transmission as a future avenue of this work; on the other hand, this framework can be applied in communication systems to detect fake text, mobile applications to enhance user security, and social media platforms to identify misleading information. Integrating advanced algorithms may improve the accuracy of content verification and enhance the integrity and security of communication systems.

Funding: “This research received no external funding”

Conflicts of Interest: “The authors declare no conflict of interest.”

References

[1] N. A. Saeejil et al., "Exploring Algorithmic Paradigms in Message Classification: Insights from the Enron E-mail Dataset," in *International Conference on Advances in Information Communication Technology & Computing*, 2024, pp. 27-40.

[2] C. N. Mohammed and A. M. Ahmed, "A semantic-based model with a hybrid feature engineering process for accurate spam detection," *Journal of Electrical Systems and Information Technology*, vol. 11, no. 1, p. 26, 2024.

[3] A. Hussain, A. Khatoon, A. Aslam, and M. A. Khosa, "A Comparative Performance Analysis of Machine Learning Models for Intrusion Detection Classification," *Journal of Cybersecurity*, vol. 6, 2024.

[4] D. Gupta, S. Dubey, and M. Mallik, "Foretelling the compressive strength of concrete using twin support vector regression," *International Journal of Information Technology*, pp. 1-18, 2024.

[5] S. Balamurugan, E. Gurumoorthi, P. Devi, and R. Maruthamuthu, "Impact of nutrients in food quality and safety by machine learning classifier using internet of things," *International Journal of Information Technology*, pp. 1-10, 2024.

[6] R. Cho, M. Zaman, K. T. Cho, and J. Hwang, "Investigating brain activity patterns during learning tasks through EEG and machine learning analysis," *International Journal of Information Technology*, pp. 1-8, 2024.

[7] S. Tared, L. Khaouane, S. Hanini, A. Khaouane, and M. Roubehie Fissa, "Enhancing lung cancer prediction through crow search, artificial bee colony algorithms, and support vector machine," *International Journal of Information Technology*, pp. 1-11, 2024.

[8] V. Chirchi, E. Chirchi, and K. E. Chirchi, "Pattern matching for the iris biometric recognition system uses KNN and fuzzy logic classifier techniques," *International Journal of Information Technology*, vol. 16, no. 5, pp. 2937-2944, 2024.

[9] D. Jayabalan and S. Elango, "ICE-VDOP: an integrated clustering and ensemble machine learning methods for an enhanced vector-borne disease outbreak prediction using climatic variables," *International Journal of Information Technology*, vol. 16, no. 4, pp. 2077-2088, 2024.

[10] S. Mondal, S. Ghosh, and A. Nag, "Brain stroke prediction model based on boosting and stacking ensemble approach," *International Journal of Information Technology*, vol. 16, no. 1, pp. 437-446, 2024.

[11] A. Qazi, N. Hasan, R. Mao, M. E. M. Abo, S. K. Dey, and G. Hardaker, "Machine Learning-Based Opinion Spam Detection: A Systematic Literature Review," *IEEE Access*, 2024.

[12] S. Xiao, R. Hao, G. Cheng, X. Xu, and T. Li, "EC-BERT: A BERT Language Model with Error Correction for Mandarin Chinese Speech Recognition," *Journal of Shanghai Jiaotong University (Science)*, pp. 1-7, 2024.

[13] A. M. M. Al Zoubi, "Spam Reviews Detection Models in Multilingual Contexts applying Sentiment Analysis, Metaheuristics, and Advanced Word Embedding," 2024.

[14] A. Singla, "Roberta and BERT: Revolutionizing Mental Healthcare through Natural Language," *Shodh Sagar Journal of Artificial Intelligence and Machine Learning*, vol. 1, no. 1, pp. 10-27, 2024.

[15] M. A. Uddin, M. N. Islam, L. Maglaras, H. Janicke, and I. H. Sarker, "ExplainableDetector: Exploring Transformer-based Language Modeling Approach for SMS Spam Detection with Explainability Analysis," *arXiv preprint arXiv:2405.08026*, 2024.

[16] K. S. Reddy and E. S. Reddy, "An Efficient Methodology to Detect Spam In Social Networking Sites," *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 15, no. 7, 2017.

[17] N. Ali, A. Fatima, H. Shahzadi, A. Ullah, and K. Polat, "Feature extraction aligned email classification based on imperative sentence selection through deep learning," *Journal of Artificial Intelligence and Systems*, vol. 3, no. 1, pp. 93-114, 2021.

[18] V. S. Tida and S. Hsu, "Universal spam detection using transfer learning of BERT model," *arXiv preprint arXiv:2202.03480*, 2022.

[19] O. Agboola, "Spam Detection Using Machine Learning and Deep Learning," Louisiana State University and Agricultural & Mechanical College, 2022.

[20] Y. Guo, Z. Mustafaoglu, and D. Koundal, "Spam detection using bidirectional transformers and machine learning classifier algorithms," *Journal of Computational and Cognitive Engineering*, vol. 2, no. 1, pp. 5-9, 2023.

[21] M. K. Islam, M. A. Al Amin, M. R. Islam, M. N. I. Mahbub, M. I. H. Showrov, and C. Kaushal, "Spam-detection with comparative analysis and spamming words extractions," in *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2021, pp. 1-9.

[22] V. Metsis, I. Androutsopoulos, and G. Paliouras, "Spam filtering with naive bayes-which naive bayes?," in *CEAS*, 2006, vol. 17, Mountain View, CA, pp. 28-69.

[23] A. P. Bhopale and A. Tiwari, "An Application of Transfer Learning: Fine-Tuning BERT for Spam Email Classification," in *Machine Learning and Big Data Analytics (Proceedings of International Conference on Machine Learning and Big Data Analytics (ICMLBDA) 2021)*, 2022, pp. 67-77.

[24] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, 2019, vol. 1, Minneapolis, Minnesota, p. 2.

[25] P. Tang and Y. Guan, "Log anomaly detection based on BERT," *Signal, Image and Video Processing*, pp. 1-11, 2024.

[26] F. Souza, R. Nogueira, and R. Lotufo, "BERT models for Brazilian Portuguese: Pretraining, evaluation and tokenization analysis," *Applied Soft Computing*, vol. 149, p. 110901, 2023.

[27] A. Vaswani, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.

[28] X. Luo, H. Ding, M. Tang, P. Gandhi, Z. Zhang, and Z. He, "Attention mechanism with BERT for content annotation and categorization of pregnancy-related questions on a community Q&A site," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2020, pp. 1077-1081.

[29] S. Lu, M. Wang, S. Liang, J. Lin, and Z. Wang, "Hardware accelerator for multi-head attention and position-wise feed-forward in the transformer," in *2020 IEEE 33rd International System-on-Chip Conference (SOCC)*, 2020, pp. 84-89.

[30] O. Galal, A. H. Abdel-Gawad, and M. Farouk, "Rethinking of BERT sentence embedding for text classification," *Neural Computing and Applications*, pp. 1-14, 2024.

[31] P. P. S. Bedi, M. Bala, and K. Sharma, "MLM: Masked Language Modeling Using Deep Learning for Efficient Summarization of Unstructured Data," in *International Conference on Data Analytics & Management*, 2023, pp. 339-347.

[32] S. Al-augby and K. Nermend, "Using Rule Text Mining Based Algorithm to Support the Stock Market Investment Decision," *Transformations in Business & Economics*, vol. 14, 2015.

[33] S. Kumar, J. R. Saini, and P. B. Bafna, "Identification of Malayalam Stop-Words, Stop-Stems and Stop-Lemmas Using NLP," in *IOT with Smart Systems: Proceedings of ICTIS 2022*, vol. 2, Springer, 2022, pp. 341-350.

[34] Z. Ch. Oleiwi, E. N. AlShemmary, and S. Al-Augby, "Developing hybrid CNN-GRU arrhythmia prediction models using fast Fourier transform on imbalanced ECG datasets," *Mathematical Modelling of Engineering Problems*, vol. 11, no. 2, pp. 413–429, Feb. 2024, doi:10.18280/mmep.110213.

[35] A. Ghourabi, M. A. Mahmood, and Q. M. Alzubi, "A hybrid CNN-LSTM model for SMS spam detection in arabic and english messages," *Future Internet*, vol. 12, no. 9, p. 156, 2020.

[36] L. GuangJun, S. Nazir, H. U. Khan, and A. U. Haq, "Spam detection approach for secure mobile message communication using machine learning algorithms," *Security and Communication Networks*, vol. 2020, no. 1, p. 8873639, 2020.