JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

Research Article
Collection: Artificial Intelligence and Emerging Technologies

Spam Email Detection using Transfer Learning of BERT Model

Ashiq Shazad[1*], Muhammad Naman Chaudhry[1], Muhammad Kamran Abid[1], and Naeem Aslam[1]

[1]NFC Institute of Engineering and Technology, Multan, 60000, Pakistan
[*]Corresponding Author: Ashiq Shahzad. Email: ash.shahzad32@gmail.com

Academic Editor: Salman Qadri **Published:** February 01, 2024

Abstract: Spam is the term for unsolicited and indiscriminate mass emails that are not wanted by the recipients and are often motivated by economic interests. Despite ethical concerns, many organizations persist in employing spam as a marketing tactic. Spam emails pose a significant challenge in today's digital landscape, potentially causing financial harm to businesses and annoyance to individual users. In order to address this issue, advances in natural language processing (NLP) have been applied to increase spam detection programs' accuracy. Specifically, efforts have been directed toward optimizing the performance of the already existing BERT (Bidirectional Encoder Representations from Transformers) transformer model. BERT utilizes attention mechanisms to contextualize the content of text data, enabling more effective discrimination between spam and non-spam (HAM) emails. The training of deep learning transformer models on text data through self-attention methods makes them significant. This dissertation explores the real-time classification of spam and ham emails using Google Bidirectional Encoder Representations from Transformers (BERT) base uncased models that have already been trained. The study trained several models with the goal of distinguishing between spam and ham emails using Enron datasets that were made publicly available. One of the models that was created performed well enough to classify emails with accuracy. Utilizing Enron datasets during the training phase allowed the model's hyperparameters to be adjusted for the best spam detection results. The same hyperparameters from our model were used to fine-tune the model. An F1-score in each model is at or above 0.9 when they are each using the appropriate dataset. 98% of the time was accurate overall, while the F1 score was 99%. The consequences and research results were examined. The study's findings demonstrated the effectiveness of the suggested strategy with remarkable performance metrics: 98% accuracy, 99% F1 score, 96% precision, and 99% recall or true positive rate (TPR). Furthermore, it was found that the true negative rate (TNR) was 73%, while the false positive rate (FPR) was 47%. The method's success demonstrates how well it can differentiate between emails that are spam and those that are not.

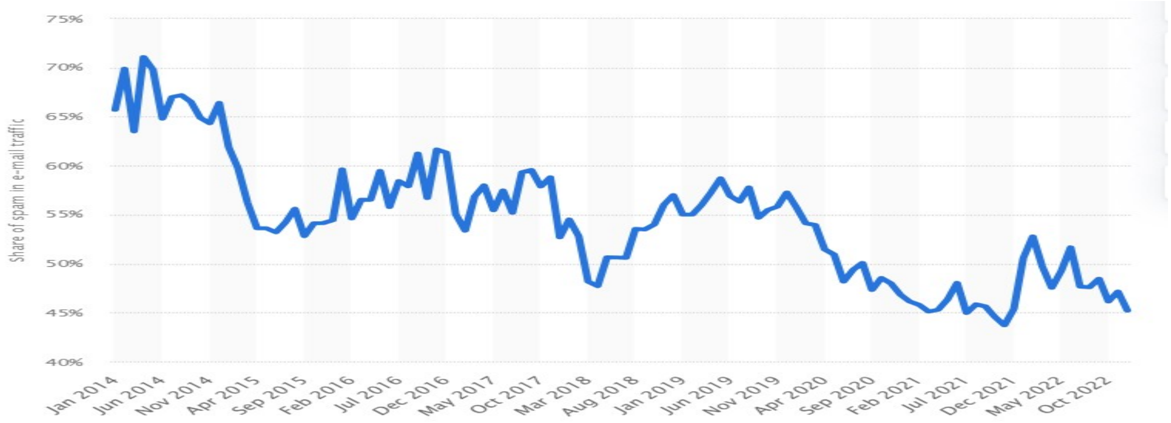
Keywords: Spam Email Detection; Transfer learning; BERT Model.

1. Introduction

Email spam has significantly increased along with the exponential growth of internet users, and it is becoming a tool used for immoral and unlawful activities. Phishing, deception, and behaviors. Spam takes up storage space and connection capacity, requires the user to read through unwanted mail, and wastes their time. They are able to rapidly and simultaneously deliver their unclear message to several email accounts. [1]. When an email is sent that is not requested, the recipient has not given permission to receive it. Since the previous ten years, using spam emails has become more and more common. On the internet, spam has grown to be really unfortunate. Spam interferes with users' capacity to make the best use of their time, storage, and network resources. Spam overload on computer networks negatively impacts a number of factors, such as memory on email servers, user productivity, CPU use, and bandwidth for communication [2].

Spam is becoming a bigger problem every year and accounts for more than 77% of all email traffic worldwide [3]. Furthermore, a sizable percentage of emails (45.37%) in December 2021 were categorized as spam. July 2021 saw the peak of the global spam volume between 2020 and 2021, with 283 billion out of 336.41 billion emails being classified as spam. Out of 105.67 billion emails sent worldwide in September 2021, almost 88.88 billion were spam emails [4]. In December 2022, spam accounted for more than 45 percent of all email traffic. Remarkably, with 29.82% of the world's total amount of unsolicited spam emails, Russia became the primary source in 2022. It is an indisputable fact that spam accounts for the great bulk of unwanted emails that advertisers send out on a daily basis. The United States was the country that produced the most spam emails per day as of January 2023.Although a lot of people think that emails with such content should go in the spam bin, marketing communications are usually innocuous, even if they annoy the recipient. [5].

A monthly observation of the global spam share of email traffic shows the global prevalence of spam together with an estimated quantity of unsolicited emails (see Figure 1). [5].



1.1

Figure 1. SPAM email traffic [5]

1.1 Background

Emails are a convenient means of communicating important information for certain consumers and company concerns. The practice of sending unwanted or promotional emails to a list of recipients via email is known as “email spam”. When an email is sent that is not requested, the recipient has not given permission to receive it. Since the previous ten years, using spam emails has become more and more common. On the internet, spam has grown to be really unfortunate. Spam wastes message speed and storage capacity [7].

Generally speaking, unsolicited messages sent by spammers via email are referred to as junk email or spam. The procedure gathers the email address from the internet and uses the username of the domain to send the message. In fact, a variety of techniques and technologies, including mail transfers, spoofing, botnets, open proxies, bulk mailing programs like mailers, and more, are used to generate spam for commercial purposes. Spam filtering is quite difficult for a variety of reasons. When it comes to dealing with spam emails, users face a variety of challenges, such as network congestion, storage capacity limitations, computing limitations that impair the effectiveness of email searches, time-consuming processes, and increased security vulnerabilities. Therefore, improving the security and efficacy of email filtering is crucial [8].

The benefits of using deep learning for natural language processing (NLP) jobs have been brought to light by a new player in the spam processing transformer space. [59]. The models' efficiency was improved, which resulted in a significant decrease in processing time. In the past, techniques like Gated Recurrent Units (GRU) and Long/Short Term Memory (LSTM) were frequently used for this. [9], and models based on recurrent neural networks (RNNs) [10] has to wait to get the previous time step data. As the model developed, it processed the data in a sequential manner without taking into account data points that had already been used, which made it difficult to capture long-range dependencies. Transformers, on the other hand, deal with this problem by parallelizing computations and including word position through position encoding. Moreover, they use a multiheaded self-attention mechanism to manage inputs and long-range relationships with skill. [11]. Many pre-trained models have been built on top of transformers, one notable example being Bidirectional Encoder Representations from Transformers (BERT).Google developed BERT, which has gained recognition for its simple design and exceptional usefulness. Transformers are typically designed with both encoder and decoder parts; however, BERT only uses the encoder part and discards the decoderIn the case of the BERT models, large datasets that comprise of numerous textual documents from the Wikipedia and Book Corpus were used [61]. Thus, these models often give two outputs. Tasks for which language translation is useful, such as speech tagging and name entity recognition, are generally among the first to be considered. That category of applications, which includes, sentence analysis, fake news identification, make use of the other output. This paper aims to establish a powerful anti-spam system that utilizes natural language processing (NLP) along with the pre-trained Bidirectional Encoder Representations from Transformers (BERT) model. BERT is a pre-trained model intended to sail the turbulent waters of Natural Language Understanding tasks which was brought by Google AI in 2018. The role of this is to assist the machine to understand the context of the texts inputs. BERT advances its effectiveness performance throug the transfer learning that ensures it uses its original training to adapt to the specific tasks. BERT is a flexible tool for NLP positions due to its the Transformer model foundation and adeptness in finding intricate linguistic subtleties. Modeling the sentence prediction and masked language modeling tasks have been the main two tasks on which BERT has been trained.

The model is provided with sets of sentences in the sentence order prediction task and the model objective is to decide whether one pair of sentences follows another. The language masking mechanism entails feeding sentences to the model that comprises of covered or hidden words; the model thus has to guess these concealed words in the input sentences. BERT's training dataset is diversified, and it includes all English Wikipedia articles and 11,038 books. This means BERT can learn the English linguistic environment with great accuracy. The phrase is shivered into the BERT model as input, and each word is then decomposed and re-entered into the model. For transforming a tokenized word, to create vectorized representation, BERT incorporates the encoder part of the model based on the transformer, a type of neural network [12].

BERT is distinguished from the previous RNNs like LSTM and the Transformers by their use of the encoders [13] or RNNs [14] [15]. This is a case of a sentence being processed as one unit the encoder processes all the inputs at the same time. Hence, BERT forms context for words by looking at both followed and following, whereas LSTM or RNN models, which take into account only previous inputs during processing, look at solely preceding inputs. Therefore, this difference is actually represented by a number value that serves as a vector output. Likewise, in the phrases "I need the apple product" and "I need the apple," the LSTM or RNN would assign the same value of each word. BERT has shown to be quick enough when used, but generates fresh vectors for every word, giving an increased speed in this case.

This study's objective was to create a reliable spam email detection system by utilizing the BERT basic uncased model that already existed. The Hugging Face Transformers collection contained resources that the researchers used to help with message classification. In order to identify the spam classification of incoming messages, they concentrated on extracting the second output from Google's pre-trained BERT basic uncased model [13]. Three major conclusions are highlighted in this studyInitially, researchers layered numerous layers atop the 768-length output vector to establish the ideal sequence length, learning rate, and model architecture. After that, they pre-processed hyperparameters such as learning rate and sequence length such that the pre-trained BERT uncased base model could be independently trained on four different datasets. This method expedited both the training procedure and the final model architecture selection. Then, using the Enron datasets, the suggested Bert model was tested with different tests and mini-batch sizes. Metrics including recall, precision, F1-score, and accuracy were employed. Using a publicly accessible dataset that contains both spam and ham emails, the system's efficacy is evaluated and contrasted with many state-of-the-art methods. Generally speaking, emails are categorized as spam and ham. The message that most people want to hear is ham. The user just needs to make sure that their mailbox contains just ham messages. Spam is any unsolicited email. Spam has developed into a powerful advertising strategy for reaching a wider audience with product information. A spam detector is an application that is intended to identify and stop unsolicited, undesired, and virus-filled emails from getting into a user's mailbox (shown in Figure 2).Separating spam from real (ham) emails is the main objective of spam filtration. This dissertation presents a novel method for applying Natural Language Processing (NLP) techniques to classify emails into categories such as spam and ham.

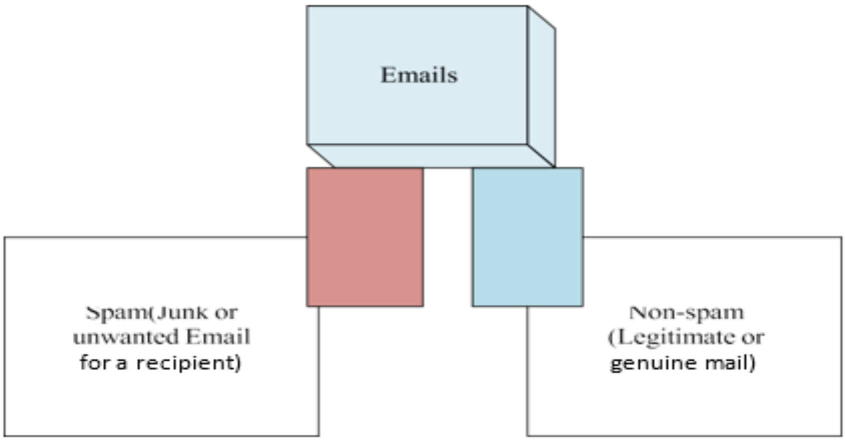

Emails
Spam(Junk or unwanted Email for a recipient)
Non-spam (Legitimate or genuine mail)

Figure 1. Email types (spam or non-spam)

2. Literature Review

Research projects by a multitude of academics have made significant contributions to the subject of spam identification. Machine learning is the process of teaching a computer system to carry out particular functions, such as classification and regression, on its own. Machine learning algorithms are given directives or guidelines, frequently in the form of algorithms, to solve issues by obtaining data attributes. When using machine learning techniques, programmers frequently need to manually extract features.

Nandan Parmar [17] Combining the integrated Naïve Bayes (NB) algorithm, a machine learning methodology, with Particle Swarm Optimization (PSO), a computational intelligence method, a novel framework is built for spam email identification. This method improves the parameters of the integrated NB model through PSO, which raises the overall classification accuracy. It does this by using the ling spam dataset for training. With a 7.75% increase in classification accuracy, the integrated NB and PSO strategy performs better than the individual NB method, according to the results. The efficacy of the suggested method is demonstrated by evaluation measures including recall, precision, and F1 measure, which achieve over 95% accuracy in comparison to conventional machine learning techniques.

Jáñez- Martin [18] Combining TF-IDF and NB techniques produced a TF-IDF and SVM model that had the fastest spam classification rate and an impressive F1-score of 95.39%.

S. Zubair etal [19] enhanced the spam detection on the websites under study by proposing framework for comment spam identification that uses a hybrid categorization technique. More ideal and adjusted features were chosen with the use of a weighting mechanism, and the suggested method's accuracy rose from 93 to roughly 96%. Prior research included fewer spam features and did not give weighting algorithms any special consideration.

Shuaib et al. [20] The whale optimization technique was employed in their study to choose features for email spam detection. The findings showed that the random forest algorithm classified emails with a precision higher than 95% after the feature selection step.

A system for spam email detection was developed by Kumaresan et al [21]. Feature selection involved the Cuckoo Technique where textual and visual traits were employed and a hybrid kernel-based support vector machine (HKSVM) classifier was then used for classification. The framework had a made an accuracy of about 94.11% which was notable to the existing spam detection approaches. It had a higher precision than the current spam detection technologies.

In the Study by Tsehay Admassu [22] One supervised learning algorithm was used upon the pre-processing of the dataset. The data repository of Kaggle contains the spam dataset featured in this study, which was composed of 4,601 samples with 58 descriptive features. Synthetic minority oversampling (SMOTE) was applied to reset the imbalance of the class in the original dataset. The method was employed in the RG model for classification which resulted in the prediction of spam emails with amazing 96.6% accuracy.

N. Varun et al. [23] Examined the level of performance of the random forest model in the spam filtering. The model of random forest was explored using test data. With the accuracy score of 95%, the results of running the trial have indicated that the random forest model is accurate. Furthermore, the study also completed a comparison between random forest model and clustering techniques, concluding that random forest model is better than clustering with regard to spam identify of emails.

O. Olatunji [24] proposed a spam detection model using support vector machines in (2019) as these are able to reach the best parameters and enhanced performance when tuning the systems. The experimental results demonstrated that the proposed model performed better than all theother methodsis that have been developed on the collective dataset employed in this study. 95.87% and 94.06% form the accuracies of 100% and 94.06% of the testing accuracy sets. The test has a 3.11% higher accuracy than the latest related study, with its 94.06% accuracy rate.

Without the need for human interaction, deep learning solves the given job autonomously by mimicking human brain activity [25]. A neural network with numerous layers and a multitude of parameters is used in deep learning. Using a few hyperparameters to give the architecture shape, deep learning automatically extracts features. A neural network with numerous layers and a multitude of parameters is used in deep learning. Because of this, deep learning approaches outperform their machine learning counterparts in the majority of areas.

Abdullah S [27] For the purpose of filtering spam emails, the effectiveness of several models was evaluated, including deep learning, decision trees (DT), support vector machines (SVM), Naïve Bayes (NB), random forests, and extreme boosting. The convolutional neural network (CNN) performed the best out of all of these models, properly classifying spam emails at a rate of 96.52%.

Safaa S.I. [28] proposed using feature selection in a genetic decision tree to identify spam emails. According to the experimental results, genetic decision trees perform better than conventional decision trees. Furthermore, a model for email spam filtering based on support vector machines was created [11]. The study put the established model for email spam filtering to the test. According to the experimental results, the constructed model attains a 94.06% accuracy score. The study does not compare the developed model with the existing machine learning model, even if the developed model achieves higher accuracy. The extant literature emphasizes that comparison studies concerning email spam detection mostly concentrate on a narrow range of machine learning approaches. Furthermore, accuracy is frequently the major factor used to assess supervised learning models in this field. Thus, the purpose of this study is to investigate the efficacy of seven supervised learning algorithms using a variety of performance metrics, such as F-score, receiver operating characteristic curve (ROC), accuracy, and precision.

Jie [29] The M-BERT multi-type bilingual language spam detection technique, which included image-based spam detection and attained an astounding accuracy rate of almost 96%, was examined in this study. In addition, M-BERT showed a respectable F1 score of 96% and a 96% Maximum Recall and Precision. This field of inquiry benefited further from Lee's work, which gathered phishing emails [30] A rate of 87% accuracy was demonstrated by using the Sophos AI-recommended CATBERT model. Other uses for the BERT model include sentiment analysis, deception detection, and the detection of fake news. Jie [31] Additionally, the research explored the field of unsupervised deep learning, which is a technology that has been proposed for the identification of fraudulent information on social media sites. Additionally, Barsever looked into this matter as well [32] suggested a model for lying detection that uses a novel generative adversarial network. In his academic paper, the author suggested a sentiment analysis framework that made use of BERT and convolutional neural networks (CNN) have attracted a lot of interest in recent studies. [33], Achieving 90.5% and 85.2% accuracy rates, respectively, indicates noteworthy performance levels for both models.

The effectiveness of using natural language processing (NLP) to identify phishing emails was tested by Egozi et al. [34] The features they collected included word counts, stopword counts, punctuation counts, and originality variables by processing samples of email content. An ensemble learning model, specifically based on a linear kernel Support Vector Machine (SVM), was trained using all 26 characteristics that were extracted. This model showed that it could correctly detect over 80% of phishing emails and 95% of real (ham) emails.

Natural language processing methods were used by Egozi and Verma [35] to identify phony emails. To distinguish between spam and legitimate emails, their algorithm selects 26 features using a feature selection method. With just 26 parameters, their method successfully distinguished between more than 95% of authentic emails and 80% of phishing emails.

A preview of their examination of previous research on content-based spam detection systems is shown in Table 1. With just 26 parameters, their method successfully distinguished between more than 95% of authentic emails and 80% of phishing emails.

Researchers have been working to make email a secure communication for the past few decades. One of the main components of a secure email platform is spam filtering. There have apparently been advancements in a number of study areas, yet there are still some unrealized possibilities. Classifying spam emails has been a popular field of research with the goal of tackling new problems. Much work has been done over the years to improve email usability for users using a variety of tactics and approaches.

Table 1. An assessment of state-of-the-art methods for identifying spam emails.

Author	Classifier	Accuracy	Data Set
Ashish Salunkhe [36]	Bilstm + Attention+Glove (100D)	90.25% achieved	Spam corpus spam corpus consisting of 1600 hotel reviews
	CNN+LSTM+Doc2Vec+TF-IDF	92.19 % achieved	Corpus Spam
Abdulhamid et al [37]	Diverse Machine Learning Techniques	Achieved Accuracy of 94.2%	UCI Machine learning Repository
Pooja Malhotra [38]	Different Algorithms for Machine Learning (Random Forest, Gradient Boosting, DT, Naïve Bayes, Logistic Regression, and Dense Sequential Model)	Random Forest has Achieved highest accuracy of 96.8% among all machine learning classifiers	Enron Dataset
		The F1 measure of Bi-LSTM is 96%, and its highest accuracy is 96.5%. 94% accuracy and 94% recall in precision	
Sharma and Bhardwaj [39]	For spam mail detection (SMD), a novel method combining a variety of machine learning approaches—including feature selection, data preprocessing, dataset organization, and a special hybrid bagging technique—is utilized. The techniques include Naïve Bayes and the J8 decision from four models.	Attained 87.5% accuracy rate.	Ling spam dataset

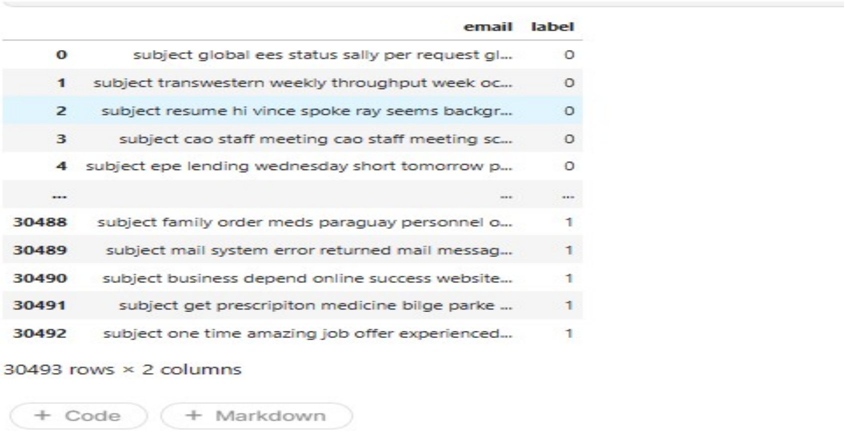
Ali Hosseinalipour [40]	An innovative method for spam identification suggests an improved method that makes use of the horse herd metaheuristic Optimization Algorithm (HOA). This strategy has various advantages over other approaches like K nearest neighbors, naïve Bayesian, multilayer perceptron, and support vector machines neighbors, and gray wolf optimization	Achieved an accuracy of 96%	UCI data set
Nandhini S [41]	Several machine learning approaches are used to evaluate performance, such as Decision Trees, Support Vector Machines (SVM), Logistic Regression, K-nearest neighbors (KNN), and Naïve Bayes.	90%, 79%, and 93% accuracy rates were reached.	UCI Machine Learning Repository Spambase
K.lyyengar et.al [57]	• INB Integrated Navie Bayes • PSO Particle Swarm Optimization	With INB, the maximum accuracy attained is 95.5%.	Spam Base emails dataset
S. Suryawanshi et.al [58]	SVM utilizing KNN	With SVM, the maximum accuracy attained is 97.5%.	5674 Labelled Dataset

3. Materials and Methods

In this study, we provide our main contribution and our suggested method for spam detection using the BERT model and transfer learning. With the help of attention mechanisms, Google AI unveiled the pre-trained model BERT Transformer, which is able to understand the contextual relationships between words in phrases. The encoder in the model is in charge of encoding textual input, and the decoder in the model decides the output outcome according to the predetermined objective [42]. The Hugging Face Resource Kit. [44] Based on the developed API, the Simple Transformers framework is created. [43], Using the Simple Transformers framework, the bert-base-cased model was built. A dataset including 800 million words from BookCorpus and 2500 million words from the English Wikipedia was used to train it [45]. A predetermined set of hyperparameters, comprising a sequence length of 300, three epochs, a batch size of 32, a learning rate set at 4e-5, and optimization via the AdamW optimizer, were used throughout the model's training.

3.1 Datasets description

32,638 emails total in the sample, of which 16,544 are classified as spam and 16,094 as ham. This is a widely used dataset that provides almost all possible sample combinations and is considered a classic benchmark in spam categorization.



	email	label
0	subject global ees status sally per request gl...	0
1	subject transwestern weekly throughput week oc...	0
2	subject resume hi vince spoke ray seems backgr...	0
3	subject cao staff meeting cao staff meeting sc...	0
4	subject epe lending wednesday short tomorrow p...	0
...	...	...
30488	subject family order meds paraguay personnel o...	1
30489	subject mail system error returned mail messag...	1
30490	subject business depend online success website...	1
30491	subject get prescripiton medicine bilge parke ...	1
30492	subject one time amazing job offer experienced...	1

30493 rows × 2 columns

Figure 2. Dataset Description

3.2 Transformer Model

For deep learning problems, transformers are an advanced neural network design; BERT stands for bidirectional encoder representations from transformers. Both input and output pieces are linked in this model, and dynamic weightings are assigned based on how they relate to each other contextually a process known as attention in natural language processing (NLP). One unique property of BERT is its capacity to process text in both left-to-right and right-to-left directions simultaneously. This is not the case with traditional language models, which handle text inputs in a sequential manner. BERT uses attention-based bidirectional learning, which it learned through masked language modeling and next sentence prediction challenges, to understand subtle contextual differences between words in a phrase. The BERT transformer model, created by Google AI, is made up of two main parts: an encoder that processes text input and a decoder that determines the output result depending on the given aim [62]. In this instance, the Hugging Face Foundation's software development kit is used [64] Based on the main transformers library's architecture, the Simple Transformers library is a software interface [63], With the given parameters a sequence length of 300, training across three epochsa 32-person batch, a 4e-5 learning rate, and Adam W optimization the BERT base cased model was constructed. The English Wikipedia corpus was used to train the model. The building's architecture BERT transformer model is visually displayed in Figure 7.

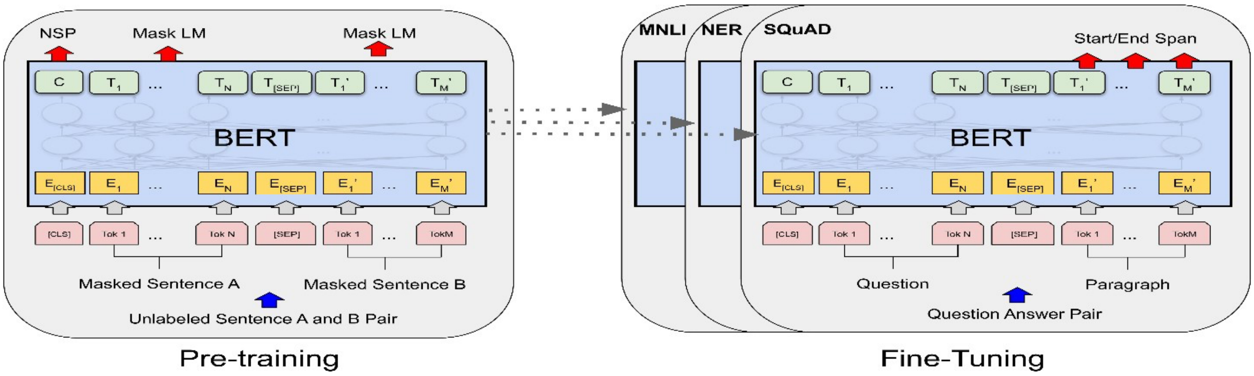


Figure 3. BERT Transformer Model

3.3 Data preprocessing and Model Selection Process

For this task, choosing a pre-trained model is crucial. Any work involving natural language processing is thought to require data preprocessing. However, depending on the model used to train the conditions, there are a few guidelines that must be observed when using pre-trained datasets. Applications for the BERT architecture are diverse and include question answering, text embedding generation, named entity recognition, and text classification. Because of its adaptability, it may be used to a wide range of natural language processing jobs and provides a reliable solution for a number of language-based applications. The BERT model, with its multiple variations, is the most suitable for classifying spam communications. The chosen model fits very nicely with our research objectives, especially in terms of spam detection. This version is ideal for our needs because it just has 12 encoders with 110 million settings. As opposed to transformers, BERT only considers the encoder component, discarding the decoder part (Figure 8). Each encoder, as shown in Figure 7, consists of feed-forward and self-attention neural networks, the same layers as its transformer counterpart. That's why BERT is viewed as a language model instead of a sequence-to-sequence model. The model's illustration of bidirectional processing allows it to learn from both ends of the input sequence, allowing for more accurate word predictions within the context.

A book corpus (800 million words) and unlabeled text corpora (2.5 million words) from Wikipedia were used to train the model. By varying the weights of the word representations that are obtained from intermediate layers during the training phase, our system is able to determine if the input sample is spam or ham. Later in the model, additional neural network layers are added to improve classifier.

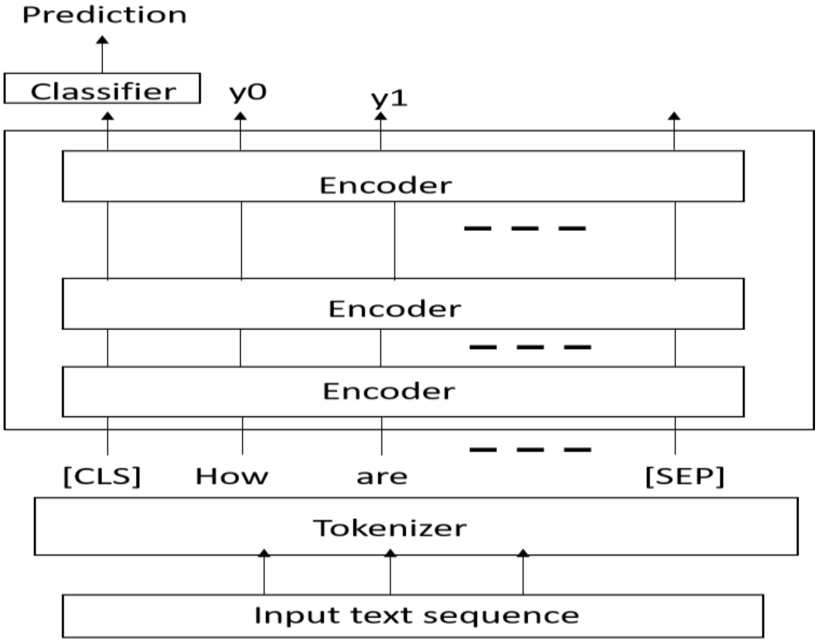


Figure 5. The BERT architecture

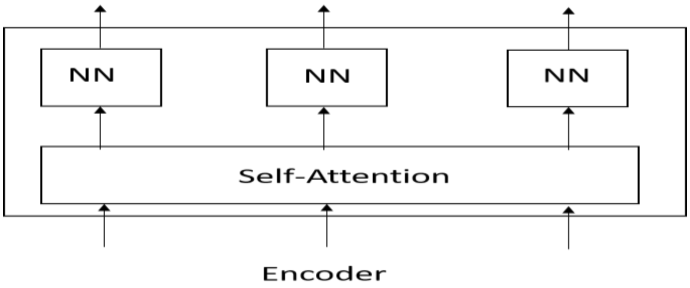


Figure 6. An encoder inside structure [61]

3.4 Final modeling

The finished model, which is represented in Figure 7, was created by carefully adjusting the hyperparameters. Three completely connected linear layers, along with batch normalization, dropout, and particular activation functions, are all part of the design. Making use of the [CLS] token side output improves the final model's ability to detect spam. An input vector with a length of 768 is fed into the linear layer of this finished model, which consists of 175 neurons. This linear layer takes an input vector of 768 and outputs a vector length of 175 to create a shape of (768, 175). A dropout layer with a factor of 0.1 is used to reduce overfitting by ignoring 10% of the linear layer's neuron outputs. By lowering generalization error, the batch normalization layer speeds up training. The outputs of the batch normalization layer are processed using the ReLU activation function. Furthermore, the output of the batch normalization layer is followed by another dropout layer with a factor of 0.1.When it came to precision and recall levels that were almost at maximum, the model did the best. True negatives and false positives were reduced by the combination. The classification of a false positive as spam is different from that of a true negative, which is categorized as spam. It is advised to include dropout layers before and after the batch normalization layer in order to reduce differences between false-positive and true-negative results in the trained model utilizing the combined dataset. Incorporating dropout layers improves accuracy and F1-score metrics in addition to increasing precision and recall scores. Accuracy and F1-score are important metrics to evaluate the performance of the model. The model's accuracy measures how well it classifies sampled data into the appropriate classes, and the F1-score gives information about how the data samples are distributed.

After the whole process is repeated, a linear layer of size (100, 2) is added. It uses log SoftMax activation to determine if the input sample is spam or ham, producing binary results of 0 or 1, respectively. After the model's hyperparameters are carefully adjusted, its performance finally outperforms that of other configurations.

Hyperparameter optimization is used to determine how these layers and neurons are configured. To accelerate the convergence of the model, the weights are adjusted in this context using the Adam optimizer [39].

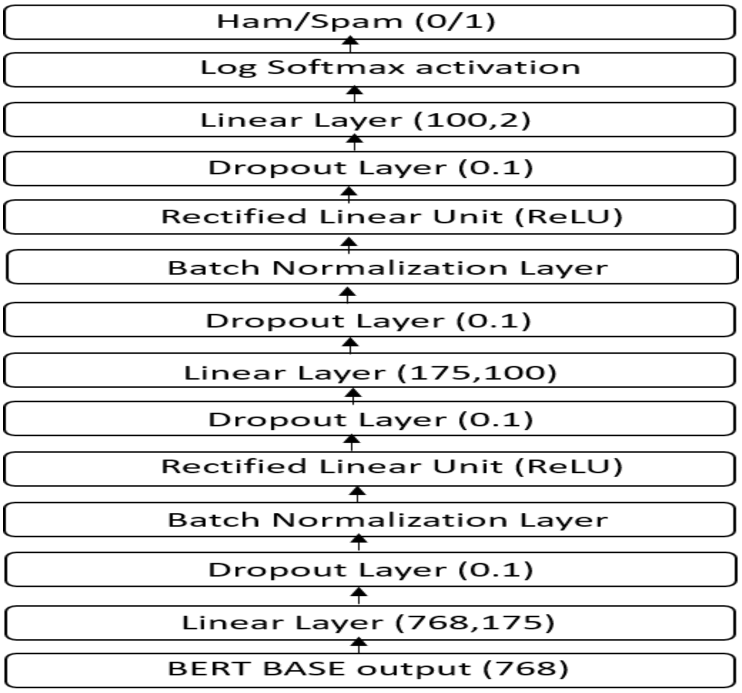


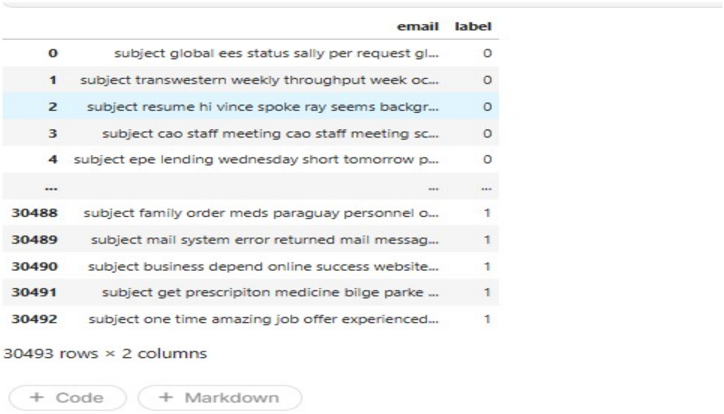
Figure 7. Final model architecture for classifier

4. Results and Discussion

This section explores the findings of a study that used transfer learning techniques to detect spam emails using a pre-trained BERT model. The Enron corpus dataset was used for the experimental process. The next section outlines the result of state-of-the-art performance in binary classification. After training our model with the training dataset for each corpus, we examined its performance with the evaluation dataset for that corpus, which produced the following results.

4.1 BERT Results

This section offering the classification results for the proposed Bert model. In this transfer learning was applied on a pre trained Bert model using a publicly accessible Enron corpus dataset. The dataset that is being used is called the Spam filter dataset and it comes from Kaggle. It is freely accessible [21] There are 6 deemed to be spam emails out of the 949 emails in the dataset. It is clear by examining the distribution of the HAM and SPAM classes that there is an imbalance in both datasets, with the HAM class being less common. In order to reduce bias toward the primary HAM class, randomly chosen spam samples from the Spam base and Spam filter datasets are combined to create a balanced training dataset that excludes duplicate entries. There are 949 SPAM samples and 930 HAM samples in this new collection. 10% of this dataset is set aside for validation, while the remaining 90% is used for training. Although a test set may be used, it is not required in this case. After exploring the data, the goal shifts to binary classification, where the job is to identify if the content should be categorized as HAM (0) or SPAM (1). Table Two presents information on how the training and testing dataset's HAM and SPAM classes are distributed.



	email	label
0	subject global ees status sally per request gl...	0
1	subject transwestern weekly throughput week oc...	0
2	subject resume hi vince spoke ray seems backgr...	0
3	subject cao staff meeting cao staff meeting sc...	0
4	subject epe lending wednesday short tomorrow p...	0
...	...	...
30488	subject family order meds paraguay personnel o...	1
30489	subject mail system error returned mail messag...	1
30490	subject business depend online success website...	1
30491	subject get prescripiton medicine bilge parke ...	1
30492	subject one time amazing job offer experienced...	1

30493 rows × 2 columns

Figure 4. Data Exploration Result

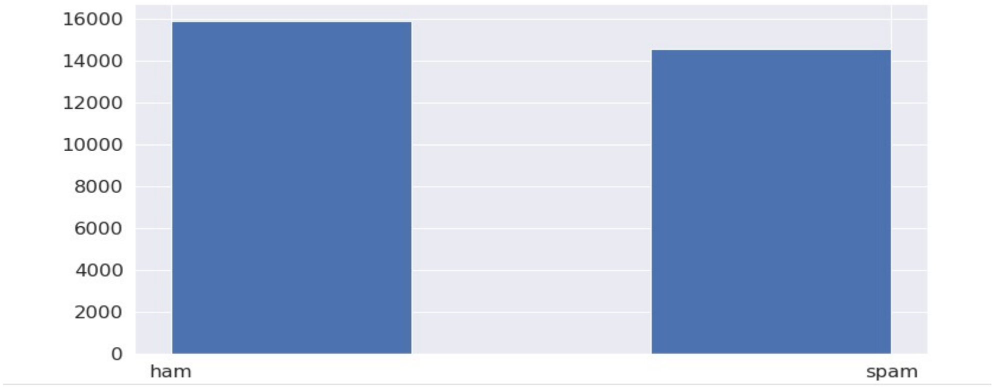


Figure 9. Email spam and ham counts from the Enron dataset

The effect of hyperparameters on model performance is shown in Figure 10, which emphasizes the importance of the learning rate in obtaining the best outcomes. The most efficient learning rate for the model can be found with the help of this visualization. A graph showing the model's loss during the training process is also shown in Figure 8. This graphical depiction enables for the evaluation of the model's convergence and performance over iterations and offers insightful information about the training procedure.

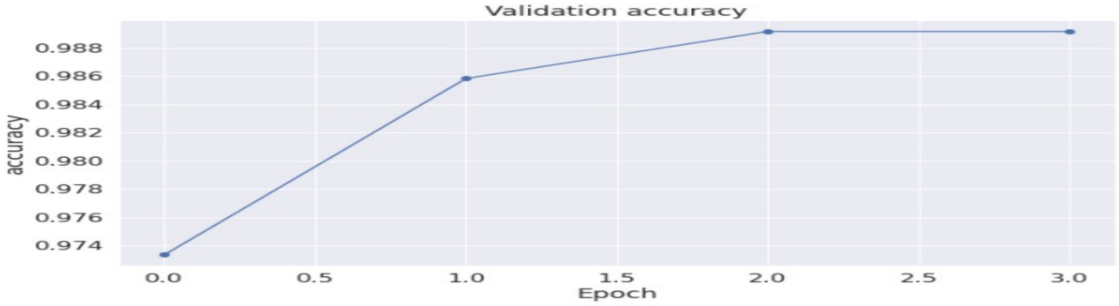


Figure 5. Validation accuracy

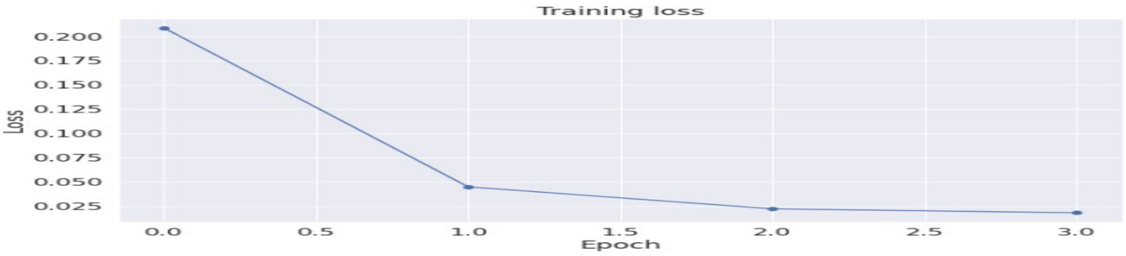


Figure 6. Training loss

A 90% training dataset and a 10% assessment dataset were separated out from the Enron corpus for training our model. The model's efficacy is revealed by the confusion matrix shown in Figure 5. Based on the results, we can conclude that our spam detection algorithm correctly classified 14 out of 19 spam emails and 929 out of 930 ham emails. Table 2 provides comprehensive metrics that show the overall performance of the model, such as the F1-scores for each class, Precision, and Recall.

4.2 Confusion Matrix

Confusion Matrix used for performance evaluation of spam emails detection. Moreover, various performance evaluation metrics are designed for performance evaluation. Confusion Matrix is one of the efficient tools for classification results. Figure 12 shows the result of confusion matrix for Spam Detection (Bert model). Total 949 data were used for assessment the performance of model. From which the model classified total 943 patches as truly classified and 6 patches are false classified.

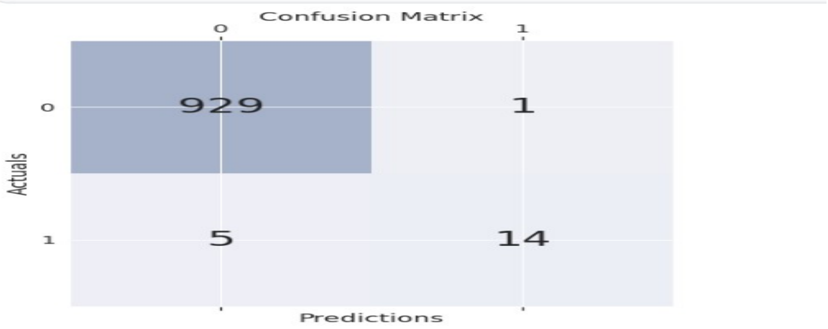


Figure 7. Confusion Matrix

Table 2. Confusion Matrix Detailed for SPAM Classification

	Class 0 Predicted	**Class 1 Predicted**
Class 0 (HAM) Actual	929	1

Class 1 (SPAM) Actual	5	14

4.3 Evaluation Metrics

The sophisticated Bert model demonstrated remarkable precision in spam email identification, with a remarkable 98% accuracy rate. Significant performance measures were also achieved by the suggested model, such as a high F1-Score value of 99%, Precision of 96%, Recall of 99%, True Positive Rate (TPR) of 99%, False Positive Rate (FPR) of 47%, True Negative Rate (TNR) of 73%, and a negligible False Negative Rate (FNR) of 0.00%. These outcomes highlight the model's effectiveness and dependability in correctly classifying spam emails. The following graph is illustrated in figure 13 which demonstrates the performance of the detail’s measurement values of all evaluation metrics of proposed Bert model in percentage (%). These achievable performance measures will help in the medical field for timely and accurate classification of Spam email detection.

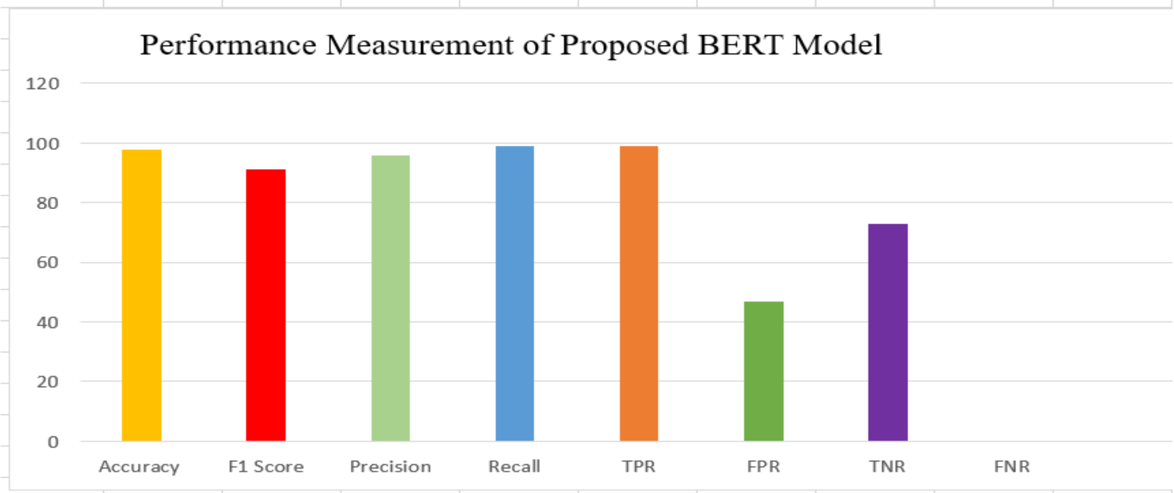


Figure 8. Performance Measurement Values of Proposed BERT Model

A thorough summary of the evaluation measures used to judge the Bert model's efficacy in spam email detection is given in Table 5. In this innovative method, it provides a range of quantitative measurements given as percentages (%) to evaluate the model's correctness and efficacy.

Table 3. Assessment of BERT Model Performance

Evaluation Metrics	Measurement Value (%)
Accuracy	98%
F1 Score	99%
Precision	96%
Recall or True Positive (TPR)	99%
False Positive Rate (FPR)	47%
True Negative Rate (TNR)	73%
False Negative Rate (FNR)	0.00%

4.4 Graphical Comparison of Various Existing Studies with Proposed BERT Model

In this subsection, a graphical comparison of various models with the proposed BERT model is illustrated. These graphs describe the comparison of existing studies based on literature and discussion.

The following are comparative graphs regarding precision and the F1-score metric are illustrated below: -

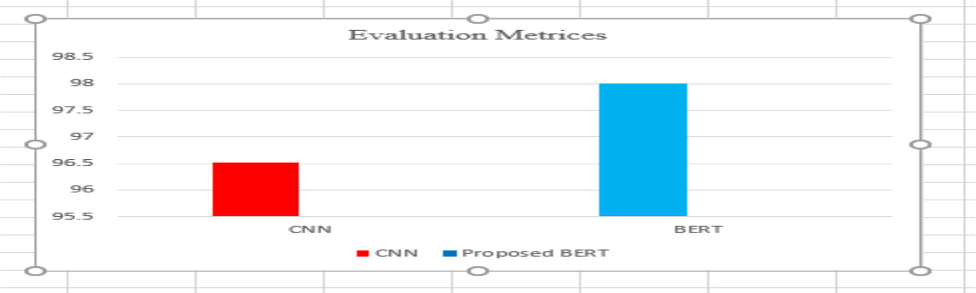


Figure 9. Comparison of CNN Architecture with Proposed BERT Model

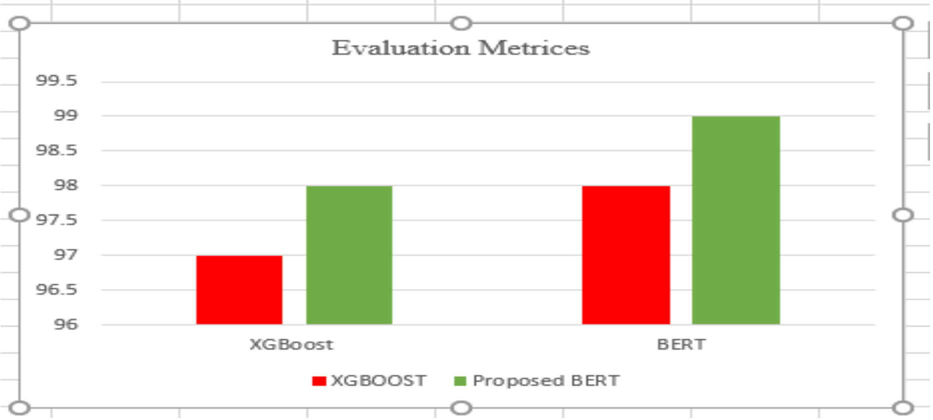


Figure 10. Comparison of XGBoost with Proposed BERT Model

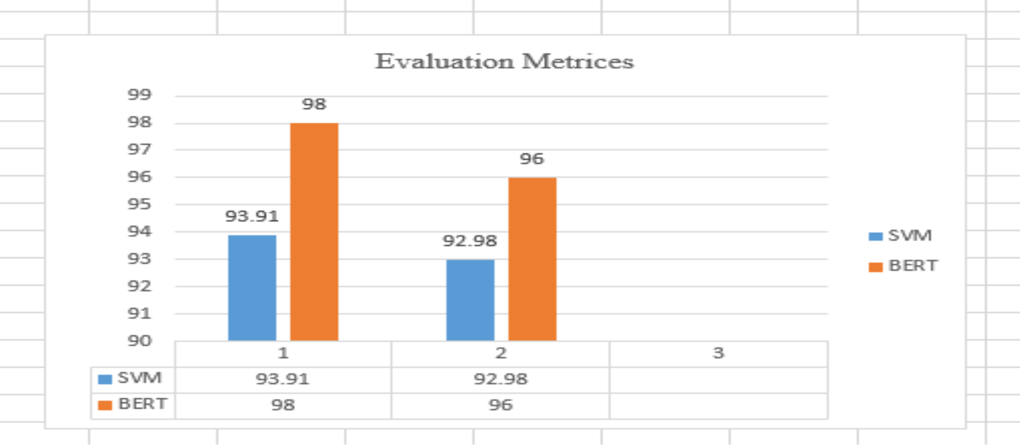


Figure 11. SVM and the Proposed BERT Model Comparison

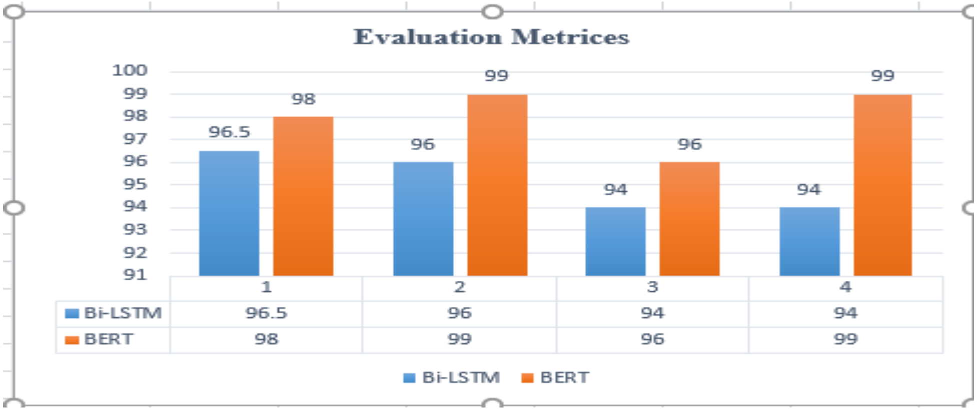
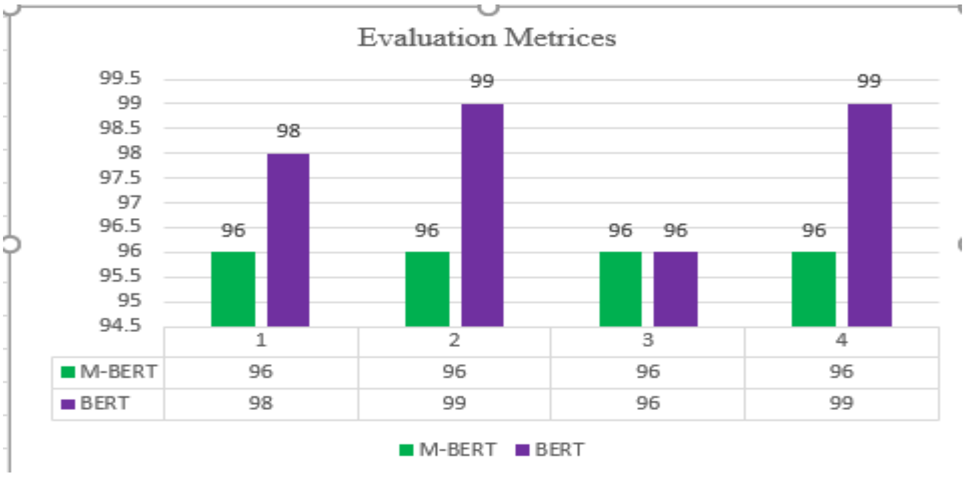


Figure 17. Comparison of Bi-LSTM with Proposed BERT Model



	1	2	3	4
🟩 M-BERT	96	96	96	96
🟪 BERT	98	99	96	99

Figure 18. Comparison of Genetic Algorithm with Proposed BERT Model

5. Conclusions

In order to address the discrepancies in results between datasets, this section suggests using transfer learning methods to identify spam emails using the BERT model. The created model shows promise for efficient real-time spam classification scenarios by utilizing transfer learning with BERT based on Enron datasets as input. Performance metrics including F1 score, precision, and recall are improved when transfer learning is used with the BERT model; these improvements outperform those of separately trained models on Enron datasets. To enhance the model's performance, more dropout layers are added both above and below batch normalization layers. The model's effectiveness in detecting spam emails is demonstrated by its incredible 98% accuracy and astounding 99% F1 score. Transformer models that are regularly released allow for the deployment of superior models for more accurate spam data identification with shorter training times. In real-time classification, over-trained models represent a noteworthy additional issue. The model does not link to other samples if it is trained on a single dataset. The model outperforms the overfitted model when more data samples are added. This theory can be used to a number of different scenarios, such as the identification of false news on social media sites, insufficient content filtering from internet sources, and so forth. When given more data samples, deep learning models perform more accurately. Research on deep learning that is primarily concentrated on a particular dataset may not yield the desired results. It is recommended that more research be done using multiple datasets to validate spam email detection with the BERT model through transfer learning. Examine how susceptible BERT-based models are to adversarial assaults in relation to spam detection. Provide a method to improve the resilience of the models. Continue the research by using a variety of languages to train and fine-tune BERT models in order to offer multilingual spam detection. Evaluate the models inter language transfer learning ability.

References

1. Dr.P.Selvarani ,A. Kanagalashmi, M.Keerthika M. Kathambari and Mancy Arokiya Mary, "A SURVEY OF EMAIL SPAM DETECTION USING ARTIFICAL INTELLIGENCE ", JETIR June 2023, Volume 10, Issue 6
2. Nikhil kumar, Sanket Sonowal and Nishant "Email Spam Detection using Machine Learning Algorithms", (ICIRCA-2020) IEEE Xplore Part Number: CFP20N67-ART; ISBN: 978-1-7281-5374-2
3. Visited on May 15,2017, Kaspersky Lab Spam Report, 2017, 2012, https://www.securelist.com/en/analysis/204792230/Spam_Report_April_2012.
4. Sefat E Rehman, Shofi Ullah, "Email Spam Detection using Bidirectional Long short Term Memory with Convolutional Neural Network " 2020 IEEE Region 10 Symposium (TENSYMP), 5-7 June 2020 DOI: 10.1109/TENSYMP50017.2020.9230769
5. Anurag Sinha, shubham, "A DETAILED STUDY ON EMAIL SPAM FILTERING TECHNIQUES", International Journal of Data Science and analytics October 2020
6. A. Sherstinsky, "Fundamantels of Recurrent Neural Network RNN and Long Short-Term Memory (LSTM) network," Phys. D Nonlinear Phenom, 2020.
7. J. Chung, "Gated Recurrent Neural Network on Sequence Modling arXiv": 1412.355v1 [cs. NE] 11 Dec 2014," Int. Conf Mach. Learn., 2015.
8. A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing System, 2017.
9. Jacob Devlin, Ming- Wei Chang, Kenton Lee, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" Proceedings of NAACL-HLT 2019, pages 4171-4186 Minneapolis, Minnesota, June 2-June 7,2019. C 2019 Association for Computational Linguistics
10. A.M.R. Thomas Wolf, Lysandre Debut, Victor Sanh Julien Chaumound, Clement Delangue
11. Anthony Moi, Pierric Cistac, Tim Result, Remi Louf, Morgin Funtowicz, Joe Davision, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine jernite, Julien Plu, Canwen Xu, Teven Le Scao, "HuggingFaces's Transformers: State of the art Natural Language Processing," arXiv:1910.
12. Nandan Parmar, Ankita Sharma, Harshita Jain, Dr, Amol K. Kadam, "Email Spam Detection using Naïve Bayes and Particle Swarm Optimization" March 2020, IJIRT Volume 6 Issue 10, pp. 7
13. F Janez-Martino, E. Fidalgo, S. Gonzalez- Martinez , and J. Velasco-Mata, "Classification of spam emails through hierarchical clustering and supervised learning " arXiv. 2020.
14. Asghar, M.Z., Ullah, A., Ahmad, S., Khan, A,: Opinion spam detection framework using hybrid classification scheme. Soft Comput. 24,3475-3498 (2020)
15. Shuaib, M., Adebayo, O.S., Osho, O., Idris, I., Alhasan, J.K., Rana, N.: Whale optimization algorithm-based email spam feature selection method using rotation forest algorithm for classifcation. SN Appl. Sci. 1(5), 390 (2019).
16. Kumaresan, T., Saravankumar, S., Balamurugan, R: Visual and textual features based email spam classification using S-Cuckoo search and hybrid kernel support vector machine. Clust. Comput. 22 (1), 33-46 (2019).
17. Tsehay Admassu Assegie, "Evaluation of supervised Learning Models for Automatic Spam Email Detection", Creative Commons Attribution 4.0 International License. July 27th, 2023 DOI:
18. Neel V., Pratap S., and Keshav A. Mail Spam Detection Using Clustering & Random Forest Algorithm. International Journal of Research in Advent Technology. 2019, 1(1), 190-192.
19. Olatunji S O, Improved email spam detection model based on support vector machines 2019, Neu. Comp.and App., 31, 691–99.
20. Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning: The MIT Press, 2016, 800 pp, ISBN: 0262035618, October 2017.
21. Abdullah S. Comparison of Deep and Traditional Learning Methods for Email Spam Filtering. International Journal of Advanced Computer Science and Applications. 2021; 12(1): 560-565.
22. Safaa S.I. Efficient E-Mail Spam Detection Strategy Using Genetic Decision Tree Processing with NLP Features. Hindawi Computational Intelligence and Neuroscience. 2022; 11): 1-16.

23. J. Cao and C.Lai, “A bilingual multi type spam detection model based on M-BERT,” in 2020 IEEE Global Communication Conference, GLOBECOM 2020- Proceedings,2020.
24. R.H. Younghoo Lee, Joshua Saxe, “CATBERT: CONTEXT-AWARE TINY BERT FOR DETECTING SOCIAL ENGINEERING EMAILS,” arxiv,2020.
25. J. Tao, X. Fang, and L. Zhou, “Unsupervised Deep Learning for Fake Content Detection in Social Media,” in Proceedings of the 54th Hawaii International Conference on System Sciences, 2021
26. D. Barsever, S. Singh, and E. Neftci, “Building a Better Lie Detector with BERT: The Difference Between Truth and Lies,” in Proceedings of the International Joint Conference on Neural Networks, 2020.
27. R. Man and K. Lin, “Sentiment analysis algorithm based on bert and convolutional neural network,” in Proceedings of IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers, IPEC 2021,2021.
28. G. Egozi and R. Verma, “Phishing email detection using robust nlp techniques,” in 2018 IEEE International Conference on Data Mining Workshops (ICDMW), IEEE, 2018, pp. 7–12.
29. Egozi G, Verma R “Phishing email detection using robust nlp techniques. In: IEEE international conference on data mining workshops (ICDMW). (2018)
30. Ashish Salunkhe, Attention-based Bidirectional LSTM for Deceptive Opinion Spam Classification, arXiv:2112.14789v1 [cs.CL] 29 Dec 2021.
31. Sharma P, Bhardwaj U (2018) Machine learning based spam e-mail detection. Int J Intell Eng Syst 11(3):1–10.
32. Pooja Malhotra, Sanjay Kumar Malik “Spam Email Detection using Machine Learning and Deep Learning Techniques” January 2022, SSRN Electronic Journal DOI:10.2139/ssrn.4145123
33. Shafi’i Muhammad Abdulhamid, M. S., Osho, O., Ismaila, I., & Alhassan, J. K. (2018). Comparative Analysis of Classification Algorithms for Email Spam Detection
34. Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, 2015. arXiv: 1506.06724 [cs.CV]
35. HuggingFace Transformers Library.” [Online]. Available: https://huggingface.co/transformers/quiktour.html.
36. O.Konur, “Adam Optimizer, Energy Education Science and Technology Part B:Social and Educational studies.2013”.
37. HuggingFace Transformers Library." [Online]. Available: https://huggingface.co/transformers/quicktour.html.
38. R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training recurrent neural networks,” in 30th International Conference on Machine Learning, ICML 2013.
39. N.S. Murthy, “Datasets and Dataloaders in PyTorch.” [online]. Available: https://medium.com/analyticsvidhya/datasets-and-dataloaders-in-pytorchb1066892b759
40. Mahmoud Jazzar, Rasheed F. Yousef and Derar Eleyan “ Evaluation of Machine Learning Techniques for Email Spam Classification “ in International Journal of Education and Management Engineering · August 2021 DOI: 10.5815/ijeme.2021.04.04
41. E. M. Bahgat, S. Rady, W. Gad, and I. F. Moawad, “Efficient email classification approach based on semantic methods,” Ain Shams Engineering Journal, vol. 9, no. 4, pp. 3259 – 3269, 2018
42. H. Faris, A. M. Al-Zoubi, A. A. Heidari, I. Aljarah, M. Mafarja, M. A. Hassonah, and H. Fujita, “An intelligent system for spam detection and identification of the most relevant features based on evolutionary Random Weight Networks,” Information Fusion, vol. 48, no. August 2018, pp. 67–83, Aug 2019.
43. Safaa S.I. Efficient E-Mail Spam Detection Strategy Using Genetic Decision Tree Processing with NLP Features. Hindawi Computational Intelligence and Neuroscience. 2022; 11): 1-16.
44. Sunday O.O. Improved email spam detection model based on support vector machines. Ismail. Neural Computing and Application. 2017, 31(1): 691–699.
45. Temidayo O.O. Hyperparameter Optimization of Ensemble Models for Spam Email Detection. Applied Science. 2023; 23(1): 1-17.

46. A. Iyengar, G. Kalpana, S. Kalyankumar, and S. GunaNandhini, “Integrated spam detection for multilingual emails,” in Proceedings of the 2017 International Conference on Information Communication and Embedded Systems (ICICES), pp. 1–4, IEEE, Chennai, India, February 2017.
47. S. Suryawanshi, A. Goswami, and P. Patil, “Email spam detection: an empirical comparative study of different ml and ensemble classifiers,” in Proceedings of the 2019 IEEE 9th International Conference on Advanced Computing (IACC), pp. 69–74, IEEE, Tiruchirappalli, India, Dec 2019.
48. T. Wolf et al., "Transformers: State-of-the-art natural language processing," arXiv. 2019.
49. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 2019.
50. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., “Huggingface’s trans-formers: State-of-the-art natural language processing,” ArXiv, arXiv–1910, 2019.
51. Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, 2015. arXiv: 1506.06724 [cs.CV].
52. Gupta, V., Mehta, A., Goel, A., Dixit, U., Pandey, A. C. (2019). Spam detection using ensemble learning. Harmony Search and Nature Inspired Optimization Algorithms, Advances in Intelligent Systems and Computing, 74, 661–668. 10.1007/978-981-13-0761-4_63.
53. C. E. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," arXiv. 2018.
54. A. De Brebisson and P. Vincent, “An exploration of softmax alternatives belonging to the spherical loss family”, in 4th international conference on learning representations , ICLR 2016-Conference Track Proceedings, 2016
55. K. Janocha and W. M. Czarnecki, "On loss functions for deep neural networks in classification," Schedae Informaticae, 2016.