Received: 16 March 2022 | Revised: 21 April 2022 | Accepted: 22 April 2022 | Published online: 24 April 2022

RESEARCH ARTICLE

Journal of Computational and Cognitive Engineering
2023, Vol. 2(1) 5–9
DOI: 10.47852/bonviewJCCE2202192

BON VIEW PUBLISHING

Spam Detection Using Bidirectional Transformers and Machine Learning Classifier Algorithms

Yanhui Guo[1],*, Zelal Mustafaoglu[1] and Deepika Koundal[2]

[1]*Department of Computer Science, University of Illinois Springfield, USA*

[2]*Department of Systemics, University of Petroleum and Energy Studies, India*

Abstract: Spam email has accounted for a high percentage of email traffic and has created problems worldwide. The deep learning transformer model is an efficient tool in natural language processing. This study proposed an efficient spam detection approach using a pretrained bidirectional encoder representation from transformer (BERT) and machine learning algorithms to classify ham or spam emails. Email texts were fed into the BERT, and features obtained from the BERT outputs were used to represent the texts. Four classifier algorithms in machine learning were employed to classify the features of the text into ham or spam categories. The proposed model was tested using two public datasets in the experiments. The results of the evaluation metrics demonstrate that the logistic regression algorithm achieved the best classification performance in both datasets. They also justified the efficient ability of the proposed model in detecting spam emails.

Keywords: spam detection, transfer learning, transformer, BERT, classifier, machine learning

1. Introduction

Email is a popular communication mode. The number of email users is projected to grow to 4.6 billion in 2025, and around 306 billion emails were exchanged globally in 2020 (Statista Research Department, 2021). Individuals and organizations all over the world use email for both casual and formal correspondence and exchange significant amounts of data. However, this popularity comes with a problem: almost 85% of emails are spam emails that are unsolicited and often malicious, costing businesses $20.5 billion every year (Cveticanin, 2022). The remaining 15% of emails are legitimate and referred to as ham emails.

Spam emails take up memory and waste computing power. Advertisement emails make up 36% of spam emails, making it the most common spam category (Cveticanin, 2022). Other categories of spam emails, such as phishing, scams, fraud, and identity theft, have malicious intentions and potentially more dire consequences. These incidents may harm individuals and institutions both financially and personally.

As the number of email users continues to grow, spam emails are increasing in number, and attackers are becoming increasingly clever with their tricks. The variety in the content of spam emails and the intentions of attackers introduces complexity. It is significant to identify spam emails with high accuracy, and there are various detection tools and techniques. The existing spam detection models can be classified into computational models, machine learning (ML) models, and deep learning (DL) models.

Numerous computational models use the negative selection algorithm (NSA) for spam detection. Idris and Selamat (2014) proposed a spam email detection model that combined a particle swarm optimization (PSO) with a NSA to achieve stability, consistency, and accuracy. The real-value NSA is used to randomly create candidate detectors. Particle swarm optimization (PSO) improves the detector generation process, with the local best as the optimum solution of the system. A local outlier factor (LOF) works as a fitness function to obtain the best features. Experimental results showed that it has 91.22% accuracy and an F1 score of 74.95%, which is higher than the PSO model at 81.32% accuracy and an F1 score of 71.84% and the NSA model at 68.86% accuracy and an F1 score of 36.01%. The NSA–PSO model outperformed another model proposed by Idris et al. (2014), which achieved 80.66% accuracy and an F1 score of 69.76%.

ML models use a variety of approaches and algorithms for spam classification. Ahmed et al. (2022) surveyed ML techniques used in email and discussed their findings. The survey demonstrated that supervised learning was the most common approach with 57% of the analyzed studies implementing supervised models. In comparison, unsupervised learning came in second at 29% and reinforcement learning came last at 14%. Moreover, Naive Bayes (NB), logistic regression, and support vector machine (SVM) were the most commonly used algorithms, with SVM generally achieving the best performance out of the three. Karim et al. (2019) reviewed bioinspired spam detection algorithms and discussed some key insights derived from their analysis. Supervised learning approach takes a ratio of 67% in email spam detection methods compared to unsupervised learning with 19% and semisupervised learning with 14%. In addition, NB was the most common algorithm followed by SVM, random forest (RF), and artificial neural network (ANN).

Madhavan et al. (2021) validated the performance of k-nearest neighbor (KNN), NB, SVM, and rough set classifier for email spam detection. The NB algorithm achieved the best performance with

*Corresponding author: Yanhui Guo, Department of Computer Science, University of Illinois Springfield, USA. Email: yguo56@uis.edu

99.46% accuracy, 98.46% recall, and 99.66% precision. KNN achieved 96.20% accuracy, 97.14% recall, and 87% precision; SVM achieved 96.90% accuracy, 95% recall, and 93.12% precision; rough set classifier achieved 97.42% accuracy, 92.26% recall, and 98.70% precision. Olatunji (2017) proposed a model based on SVM, a statistical ML technique that can be used to model complex relationships among variables, which achieved an accuracy of 94.06%. Amjad and Gharehchopogh (2019) proposed a model that uses improved KNN to classify test samples, which selected the k closest training examples in feature space and combined with the scatter search algorithm and feature selection for spam detection. The model achieved 94.54% accuracy, 94.23% precision, 95.27% recall, and an F1 score of 94.74% compared to other models such as RF at 93.89% accuracy and decision tree at 91.71% accuracy. Wang et al. (2021) proposed a model that uses the LEP manifold learning algorithm to extract the feature and employed the SVM algorithm for classification. The model achieved 94.7% accuracy on the Enron-Spam dataset, 96.9% accuracy on the PU1 dataset, and 95.1% accuracy on the GenSpam dataset.

DL-based models have yielded state-of-the-art results for spam email detection. Sumathi and Pugalendhi (2020) proposed a DL model and employed the RF algorithm to calculate the features' attribute scores. The model achieved 88.59% accuracy, outperforming other classifier models such as KNN and SVM. Siddique et al. (2021) employed a long short-term memory (LSTM) and trained two ML models, NB and SVM, and a DL model, convolutional neural network (CNN), for comparison purposes. The LSTM model achieved 98.40% accuracy compared to NB at 98.00%, SVM at 97.50%, and CNN at 96.20%. In addition, the SVM model achieved 97% precision, 92.50% recall, and an F1 score of 95%; the NB model achieved 96.50% precision, 95% recall, and an F1 score of 96%. Magdy et al. (2022) applied an ANN model to create a 3-fold classifier. The model achieved a maximum accuracy of 99.57% and a maximum precision, recall, and F1 score of 99.68% on the Spambase dataset.

A novel approach to spam detection involves transformers, and DL models that use the self-attention mechanism to perform natural language processing (NLP) tasks. Transformers have a limited choice of architectures during pretraining. Bidirectional encoder representations from transformers (BERT) achieves greater performance by using a masked language model to use pretrained deep bidirectional representations (Devlin et al., 2019).

AbdulNabi and Yaseen (2021) developed a BERT-based spam detection method. The BERT-based model was compared to a bidirectional long short-term memory (BiLSTM) model and a KNN-based model and an NB-based model. Two datasets, Spambase from the UCI Machine Learning Repository and the Spam Filter Dataset from Kaggle, were used. The BERT-based model outperformed the other models with an accuracy of 98.67% and an F1 score of 98.66%. The BiLSTM model achieved 96.43% accuracy, the KNN model achieved 92.92% accuracy, and the NB model achieved 94.69% accuracy.

Tida and Hsu (2022) used a pretrained BERT uncased model. It contains three fully connected linear layers with batch normalization layers, four dropout layers, and the ReLU and the log softmax activation functions. The model attained 97% accuracy, 97% recall, and 95% precision on the combined dataset; its highest F1 score was 96.08%. The model achieved 97% accuracy on the Enron dataset, and an accuracy of 98% on the SpamAssassin, LingSpam, and SpamText datasets.

In this study, we propose a BERT-based model to classify ham or spam emails. The novel contributions are as follows:

- A spam detection model based on a pretrained BERT is constructed and trained on two different datasets.
- The performance of four classifier algorithms, SVM, KNN classifiers, RF, and logistic regression, is compared.
- The results are analyzed using three evaluation metrics: precision, recall, and F1 score.

The experimental results demonstrate that the logistic regression algorithm achieved the best classification performance with 97.86% precision, 97.83% recall, and an F1 score of 97.84% on the first dataset and 95.95% precision, 96% recall, and an F1 score of 95.92% on the second dataset.

The following parts are structured as follows: we introduce the model in Section 2, discuss our experimental results in Section 3, and conclude our study in Section 4.

2. Proposed Method

2.1. BERT

Recently, DL methods, particularly recurrent neural networks (RNNs), have been applied in NLP. Many models have been proposed to solve the NLP tasks such as language modeling, machine translation, and question answering.

A transformer was introduced in Vaswani et al. (2017) and has replaced RNN models for NLP problems.

A transformer improves the self-attention mechanism, differentially weighting the significance of the input. It has no convolutional and recurrent layers and uses a self-attention mechanism to build relationships between all words in a sentence Wolf et al. (2020).

BERT, one of the most popular transformer-based models, is an encoder stack of transformer structure and applies the bidirectional training of transformer to language modeling (Devlin et al., 2019).

BERT architectures have extensive feedforward networks and attention heads. It takes a classification (CLS) token and a sequence of words as input. Each layer uses self-attention and passes the result through a feedforward network to the next encoder. The output corresponding to the CLS token can be used for the classification task.

2.2. Feature extraction

Transfer learning uses the trained model to acquire knowledge for a specific application whereas pretrained models were usually trained using big datasets.

In this study, a pretrained BERT model produces word embedding from email texts, and they are then used as features to represent the texts for further processing.

2.3. Classification

After the features were obtained on the email text using a pretrained BERT model, spam detection becomes a classification problem, and a classifier in ML is used to solve it via classifying the feature vectors into spam or ham categories. Supervised classifiers were first trained on the feature sets, and then the tuned classifiers were employed to classify the unknown samples.

3. Result and Discussion

3.1. Dataset

This study used two publicly available datasets. After the preprocessing, the samples' content is used for training and testing.

The first dataset, the Enron-Spam dataset, was published by Androutsopoulos et al. (2006). The Enron dataset has 33,716 emails, including 17,171 spam mails and 16,545 ham mails.

The second dataset, the spam or not spam dataset, was published by Raftogiannis (2021). This dataset consists of 2,999 valid samples: 499 spam and 2,500 ham. The spam or not spam dataset is available on the Kaggle website.

3.2. Experiment platform and settings

All experiments were implemented using Python and taken on a server with an Intel Xeon processor with 128 GB memory and an NVIDIA Tesla K40 GPU having 12 GB memory.

Five-folder cross-validation experiments were taken to validate the performance of four famous supervised classifiers on two datasets.

3.3. Evaluation results

Three evaluation metrics, precision, recall, and F1 score, are defined to evaluate the classification performance of different classifier algorithms, which are defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

where TP is true positive where spam samples are predicted as spam, TN is true negative where ham samples are predicted as ham, FP is false positive where spam samples are wrongly predicated as ham, and FN is false negative where ham samples are wrongly predicted as spam.

A receiver operating characteristic curve (ROC) is employed to measure the diagnostic ability of a classifier with varied discrimination thresholds. It shows the relationship between the true-positive rate against the false-positive rate at various threshold values. The area under the ROC curve is named AUC and used to compare the classifiers. The higher the AUC, the better the classification ability the classifier has.

Four *famous classifier algorithms*, SVM, k-nearest neighbor classifiers, random forest, and logistic regression are used to classify spam emails and their performances are evaluated quantitatively. The metric results with average and standard deviation on dataset 1 are compared in Table 1 and the results on dataset 2 are compared in Table 2. We can see that the logistic regression algorithm attains the highest precision, F1 score, and AUC values on both datasets. SVM achieves the highest recall value on the second dataset. The ROCs for four algorithms are shown in Figures 1, 2, 3 and 4.

Figure 1
ROC of the SVM algorithm on two datasets

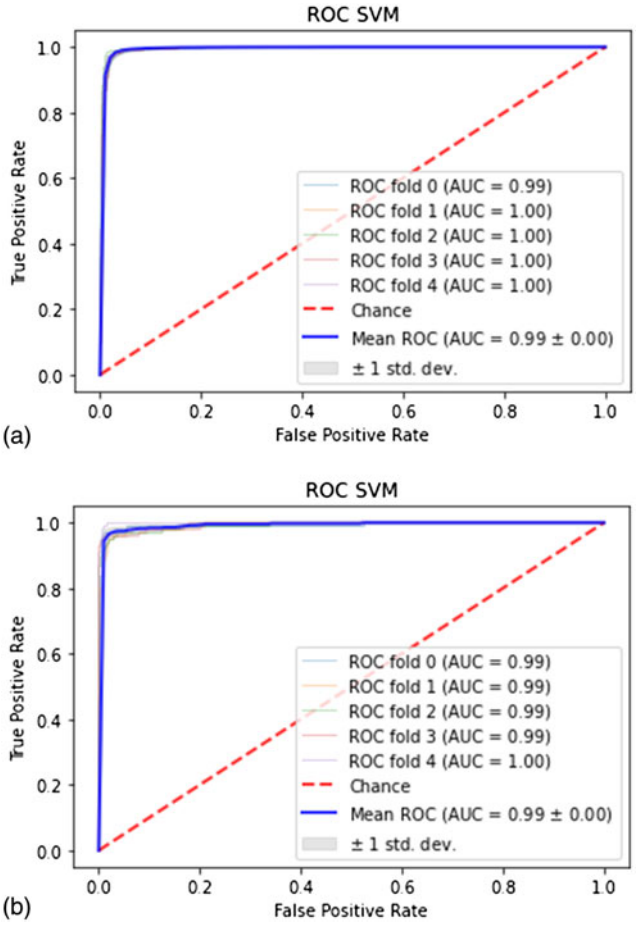


Table 1
Evaluation results for different machine learning algorithms on dataset 1

Algorithm	Precision	Recall	F1 score	AUC
SVM	0.9772 ± 0.0102	0.9769 ± 0.0102	0.9770 ± 0.0102	0.9964 ± 0.0028
Logistic Regression	**0.9786** ± 0.0081	**0.9783** ± 0.0081	**0.9784** ± 0.0081	**0.9971** ± 0.0024
Random forest	0.9639 ± 0.0204	0.9634 ± 0.0204	0.9635 ± 0.0204	0.9946 ± 0.006
KNN	0.9654 ± 0.0308	0.9637 ± 0.0343	0.964 ± 0.034	0.9905 ± 0.0069

Table 2
Evaluation results for different machine learning algorithms on dataset 2

Algorithm	Precision	Recall	F1 score	AUC
SVM	0.9553 ± 0.0278	**0.9656** ± 0.037	0.9596 ± 0.0127	0.9943 ± 0.0053
Logistic Regression	**0.9595** ± 0.0337	0.9600 ± 0.0338	**0.9592** ± 0.0232	**0.9950** ± 0.0045
Random forest	0.9591 ± 0.0365	0.8692 ± 0.0296	0.9064 ± 0.0284	0.9847 ± 0.0137
KNN	0.9372 ± 0.0545	0.9251 ± 0.0394	0.9307 ± 0.0437	0.9794 ± 0.0139

Figure 2
ROC of the logistic regression algorithm on two datasets

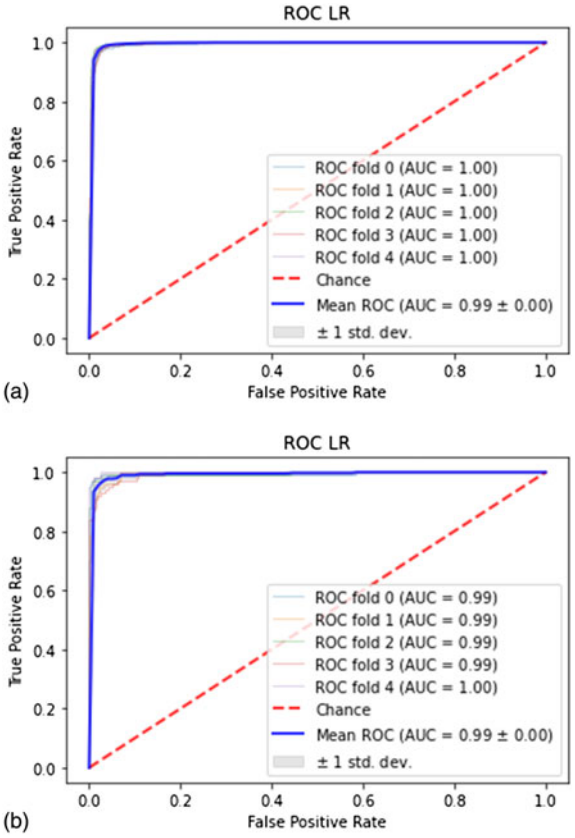


Figure 3
ROC of the random forest algorithm on two datasets

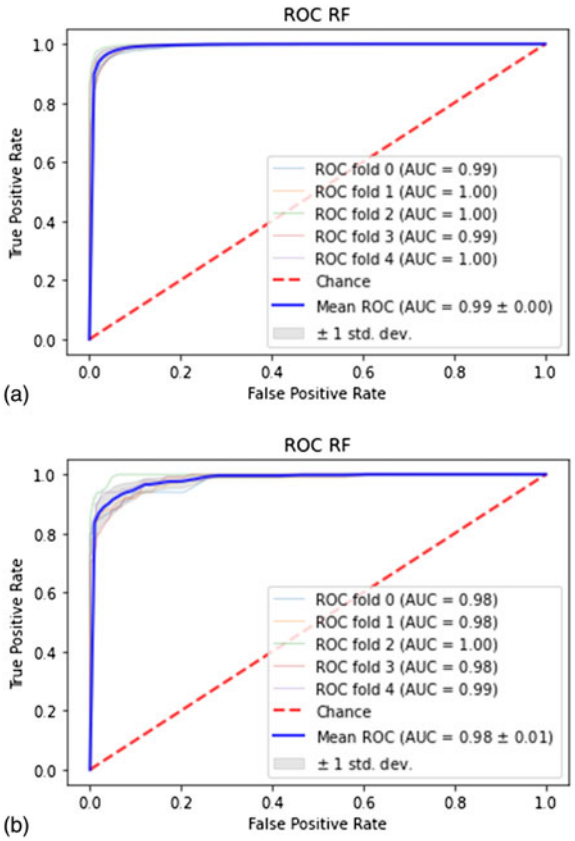
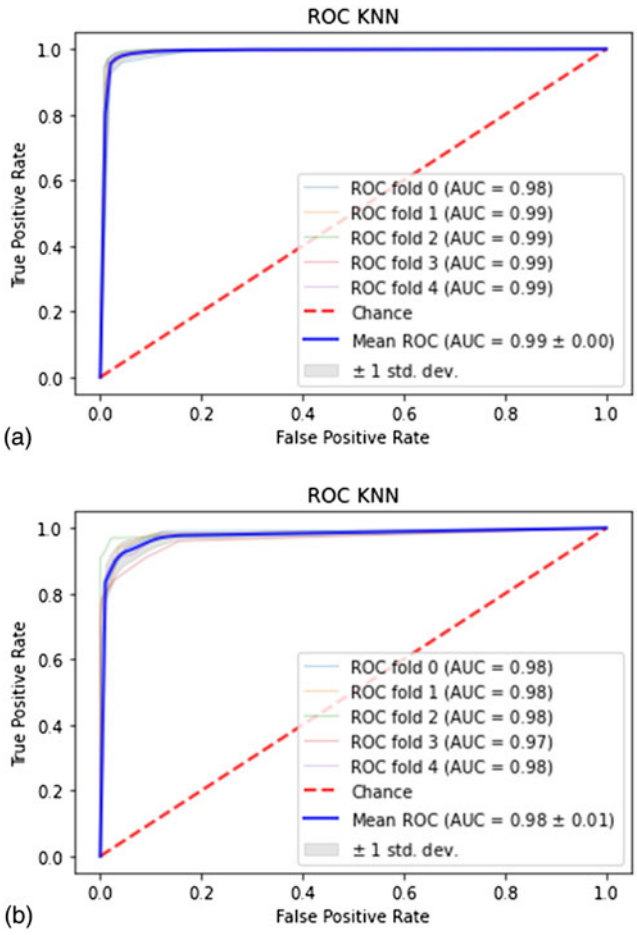


Figure 4
ROC of the k-nearest neighbor algorithm on two datasets



4. Conclusion

In this study, an efficient spam detection model was proposed based on a BERT model and supervised learning classifier to detect spam emails. Email texts were represented via the features obtained from the BERT outputs, and classifier algorithms in machine learning were employed to classify the feature vectors into ham or spam categories. The experimental results demonstrate that the logistic regression algorithm achieved the best classification performance in two publicly available datasets. To sum up, there is a promotion to use the BERT model and classifier in spam detection.

This study can be extended to various applications, e.g., spam messages detection in a mobile system and fake news detection in social media platforms. This study demonstrates the high ability of the BERT model to interpret text and provides salient features for future processing. Further research in combing more comprehensive layers inside the BERT is encouraged to further validate the proposed framework.

Conflicts of Interest

Yanhui Guo is an editorial board member for *Journal of Computational and Cognitive Engineering*, and was not involved in the editorial review or the decision to publish this article. The authors declare that they have no conflicts of interest to this work.

References

AbdulNabi, I., & Yaseen, Q. (2021). Spam email detection using deep learning techniques. *Procedia Computer Science*, *184*, 853–858. https://doi.org/10.1016/j.procs.2021.03.107

Ahmed, N., Amin, R., Aldabbas, H., Koundal, D., Alouffi, B., & Shah, T. (2022). Machine learning techniques for spam detection in email and IOT platforms: Analysis and research challenges. *Security and Communication Networks*, *2022*, 1–19. https://doi.org/10.1155/2022/1862888

Androutsopoulos, I., Metsis, V., & Paliouras, G. (2006). *The Enron-spam datasets*. Retrieved from: http://www2.aueb.gr/users/ion/data/enron-spam/

Amjad, S., & Gharehchopogh, F. S. (2019). A novel hybrid approach for email spam detection based on scatter search algorithm and k-nearest neighbors. *Journal of Advances in Computer Engineering and Technology*, *5*(3), 181–194. https://journals.srbiau.ac.ir/article_14397.html

Cveticanin, N. (2022). What's on the other side of your inbox - 20 spam statistics for 2022. https://dataprot.net/statistics/spam-statistics/

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *1*, 4171–4186. https://arxiv.org/abs/1810.04805

Idris, I., & Selamat, A. (2014). Improved email spam detection model with negative selection algorithm and particle swarm optimization. *Applied Soft Computing*, *22*, 11–27. https://doi.org/10.1016/j.asoc.2014.05.002

Idris, I., Selamat, A., & Omatu, S. (2014). Hybrid email spam detection model with negative selection algorithm and differential evolution. *Engineering Applications of Artificial Intelligence*, *28*, 97–110. https://doi.org/10.1016/j.engappai.2013.12.001

Karim, A., Azam, S., Shanmugam, B., Kannoorpatti, K., & Alazab, M. (2019). A comprehensive survey for intelligent spam email detection. *IEEE Access*, *7*, 168261–168295. https://doi.org/10.1109/access.2019.2954791

Madhavan, M. V., Pande, S., Umekar, P., Mahore, T., & Kalyankar, D. (2021). Comparative analysis of detection of email spam with the aid of machine learning approaches. *IOP Conference Series: Materials Science and Engineering*, *1022*(1), 012113. https://doi.org/10.1088/1757-899x/1022/1/012113

Magdy, S., Abouelseoud, Y., & Mikhail, M. (2022). Efficient spam and phishing emails filtering based on deep learning. *Computer Networks*, *206*, 108826. https://doi.org/10.1016/j.comnet.2022.108826

Olatunji, S. O. (2017). Improved email spam detection model based on support vector machines. *Neural Computing and Applications*, *31*(3), 691–699. https://doi.org/10.1007/s00521-017-3100-y

Raftogiannis, N. (2021). *Simple spam email classifier ~= 95%.* Retrieved from: https://www.kaggle.com/nikosraftogiannis/simple-spam-email-classifier-95

Siddique, Z. B., Khan, M. A., Din, I. U., Almogren, A., Mohiuddin, I., & Nazir, S. (2021). Machine learning-based detection of spam emails. *Scientific Programming*, *2021*, 1–11. https://doi.org/10.1155/2021/6508784

Sumathi, S., & Pugalendhi, G. K. (2020). Cognition based Spam Mail text analysis using combined approach of deep neural network classifier and random forest. *Journal of Ambient Intelligence and Humanized Computing*, *12*(6), 5721–5731. https://doi.org/10.1007/s12652-020-02087-8

Statista Research Department. (2021). *Number of e-mail users worldwide 2017-2025*. Retrieved from: https://www.statista.com/statistics/255080/number-of-e-mail-users-worldwide/

Tida, V. S., & Hsu, S. H. (2022). Universal spam detection using transfer learning of Bert model. In *Proceedings of the Annual Hawaii International Conference on System Sciences*. https://doi.org/10.24251/hicss.2022.921

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . , & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30.* https://doi.org/10.48550/arXiv.1706.03762

Wang, C., Li, Q., Ren, T. Y., Wang, X. H., & Guo, G. X. (2021). High efficiency spam filtering: A manifold learning-based approach. *Mathematical Problems in Engineering*, *2021*, 1–7. https://doi.org/10.1155/2021/2993877

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., . . . , & Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). https://doi.org/10.18653/v1/2020.emnlp-demos.6

How to Cite: Guo, Y., Mustafaoglu, Z., & Koundal, D. (2023). Spam Detection Using Bidirectional Transformers and Machine Learning Classifier Algorithms. *Journal of Computational and Cognitive Engineering* 2(1), 5–9, https://doi.org/10.47852/bonviewJCCE2202192