CLASSIFICATION OF SPAM MAIL UTILIZING MACHINE LEARNING AND DEEP LEARNING TECHNIQUES

Bandar Alshawi (1), Amr Munshi (1), Majid Alotaibi (1), Ryan Alturki (2), Nasser Allheeib (3)*

(1) Department of Computer and Networks Engineering, Umm Al-Qura University, Makkah; (2) Department of Software Engineering, College of Computing, Umm Al-Qura University, Makkah; (3) Department of Information Systems, College of Computer and Information Sciences, King Saud University, Riyadh
Saudi Arabia

* Corresponding Author, e-mail: aaamunshi@uqu.edu.sa

Abstract: The Internet and social media networks usage has increased nowadays and become a prominent medium of communicating. Email is one of the professional reliable methods of communication. Automatic classifications of spam emails have become an area of interest. In order to detect spam emails, this study utilizes a dataset, including spam and non-spam emails. Various techniques are applied to obtain higher accuracy using machine learning techniques. NLP is also utilized for improvising accuracy using embeddings. For that, this work utilizes the BERT model, to achieve satisfactory detection of spam emails. Further, the results are compared with state-of-the-art methods, including, KNN, LSTM and Bi-LSTM. The results obtained by Bi-LSTM and LSTM were 97.94% and 86.02%, respectively. The presented methodology is promising in detecting spam emails due to the higher accuracy achieved.

Keywords: Spam mails; machine learning (ML); long short-term memory (LSTM); k-nearest neighbors (KNN); deep learning.

1. INTRODUCTION

Particularly when the emails are business-related and connected to those of an organization, filtering away spam-based emails from the inbox is thought to be a vital activity. The most difficult duty, however, is sorting through spam emails since these emails occasionally find new and inventive ways to display undesirable material [1]. Spam mails are something uncommon, and unwanted mail is sent in bulk. Deep learning techniques can be used in conjunction with machine learning and other computer algorithms to assist with these categorization tasks. Spam emails are categorized into different types, mainly it is used for commercial purposes. These emails are used in the form of advertisements in both business and marketing [2]. Mostly these emails do not harm. There are spam emails that are in the form of warnings about virus attacks and the user must aware not to click the email links in order to avoid those warnings. Email

spoofing is also coming under email spam. These emails cannot be suspected by the user as it uses company's logos and brand names and whether the email is spam or not is only verified by contacting the company [3-5].

In this email spam classification research, some of the techniques used are vectorization, tokenization, and lemmatization [6]. The K-Nearest Neighbors (KNN) classifier is implemented to detect spam emails [7]. Further, the study illustrates the application and results of deep learning and Machine Learning (ML) to efficiently categorize spam emails. The KNN classifier used demonstrates dependability and higher accuracy rates in handling positive samples in the data set.

The main contributions of this research to the field of email security can be summarized as:

- Presenting a thorough comparison of machine learning and deep learning techniques for spam email detection, offering insights into their comparative effectiveness.
- Development of reliable spam filtering systems demonstrating high accuracy rates, achieved through advanced methods such as Bidirectional Encoder Representations from Transformers (BERT) embeddings and Bidirectional Long Short-Term Memory (Bi-LSTM) models.
- Practical implications and applications of effective spam filtering contributing to enhanced user experience, productivity, and overall cybersecurity.

The remainder of the paper is organized as follows. Section two of this paper presents the related work, followed by explaining the methodology used for this research and also includes the working of the three models discussed here along with the evaluation metrics of the models as well in section three. section four details the experimental results of the research. It shows the figures and listings related to the code implications and its discussions. Finally, section five shows the conclusions and discussions drawn from the whole research.

2. RELATED WORK

According to Malhotra and Malik [8], every year unwanted emails like phishing and also spam emails cost a lot to the public and businesses, and a lot of models are developed for email spam classification. So, utilizing deep learning techniques and ML techniques that offer a better success rate and also the Natural Language Processing (NLP) helps improve the accuracy of the model, especially for the classification of email spam. Their study engrossed in the usefulness of word embedding for the classification of email spam. The majority of research shows that the best spam mail filter relies on models with high recall and superior testing accuracy [9]. It describes both the deep learning model's performance and dependability. BERT is one of these pre-trained models that are particularly adept at identifying text, such as spam emails, and works well with NLP to assist to get cutting-edge results [10, 11]. Large data sets are required for these spam mail categorization jobs, which increases the model memory requirements [12]. This requires the use of a model that offers superior memory patterns in order to meet the filter's data processing requirements. Therefore, when LSTM models are used, unnecessary data from the data may be removed and used better. This is due to the fact

that these models are created with an increasing number of hidden layers through which the input goes, making it simple [13].

Utilized the model pre-trained transformer and fine-tuned the BERT to implement the identification of spam emails from that non-spam emails (which are also called HAM) related tasks. Lin et al. [14] also discussed the working of these models, however, the research in [8] has implemented these models for the classification of email spam. The BERT employed attention layers to draw the text context into its viewpoints. The model baseline Deep Neural Network (DNN) consists of layers- Bi-LSTM and two stacked dense layers along with KNN and Naïve Bayes (NB) classifiers, the results are compared. Their research helped in obtaining higher accuracy results. The Area Under the Curve (AUC), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are evaluated by using ML techniques [15, 16]. The loss value and accuracy are obtained by implementing deep learning techniques.

In the research presented in [8], the BERT model was the best one which provided 98.67% accuracy and an F1 score of 98.66%. The BERT-related word embedding enhances the ability to identify spam emails when it is compared with the Keras word embedding, each word characterized by a unique integer which was utilized in Bi-LSTM that provided an accuracy of 96.43% and an F1 score of 96%.

In a recent study, related to spam email classification, [7] explained that finding email spam classification is a challenging task. Also explained that the prior models were proposed but still required to enhance accuracy, should take less training time, and must have less error rate. Their research is related to the classification of email as either spam or non-spam. To find the extreme values exterior to a particular range utilized the isolation forest and Density-Based Spatial Clustering of Applications with Noise (DBSCAN).

For choosing the features that are much more effective, the feature selection methods used were recursive feature elimination, heatmap, and chi-square [17]. The model proposed used ML and deep learning to obtain the comparative analysis. The utilized models were Random Forest (RF), multinomial NB, gradient boosting, and KNN to make known to ensemble approach in ML implementation [18]. Artificial Neural Network (ANN) Recurrent Neural Network (RNN) and Gradient Descent (GD) are implemented here as the application of deep learning in the work [19]. The ensemble approach is created to combine numerous classifiers' outcomes. The prediction accuracy reached through the ensemble methods is better when compared to the single classifier [20]. It was found that the ML technique works well when compared to the deep learning techniques mainly for tabular datasets to detect spam e-mails. Their research helped in reaching 100% accuracy for their proposed model and also for AUC equalling 100, MSE and RMSE error both are zero for ML implementation and for deep learning implementation the accuracy reached is 99% and loss value found to be 0.0165.

Deep learning networks and the power of NLP are thought to perform better, however, the success of these classification tasks depends on the presentation of the data [20]. Because they have been configured to function optimally with word embeddings, they are similarly assessed for relation classification, and the provided model is utilized herein to achieve enhanced accuracy. Hence, it is evident that no matter how crucial the data is, transformer-based models outperform the LSTMs. The research is focused on

comparing deep learning neural network models with ML models since both have been shown to operate well when the job is spam categorization. A more accurate model is provided, and the findings are analyzed. Most of the papers, [11, 12] out there uses homogeneous data from the same portal but in this case study we acquired data for NLP task from different portals and yet it has proven deep learning models performs very well compared to ML.

The accuracy of the model is improved by using deep learning approaches, ML techniques with a higher success rate, and NLP. This is especially true for the categorization of email spam [7, 8, 14]. Their research focused on the value of word embedding for categorizing email spam used the model-pretrained transformer and improved the BERT to accomplish tasks related to separating spam emails from non-spam emails (which are also referred to as HAM).

3. METHODOLOGY

Firstly, the spam data was collected, and the processing related to that of the data set was conducted such as concatenating the data and selecting the non-spam sample such that both classes in the output label are balanced. Initially, two datasets were taken and combined into one to balance it out. These datasets were taken from the Kaggle, which is an open-source platform. This constitutes the process of spam collection and cleaning. Then the data set is loaded using the Pandas library and visualizations are considered which will be explained in the following section.

When utilizing the BERT model, the similarity based on the cosine similarity was seen and conducted for the words such as “money, buy, guarantee”. The text processing techniques such as word lemmatization is considered along with removing the stop word removal, and tokenization as suggested by HaCohen-Kerner et al. [20]. Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer is used to implement the ML model KNN. The methods seen in Figure 1 and the methodological steps are described to obtain the results.

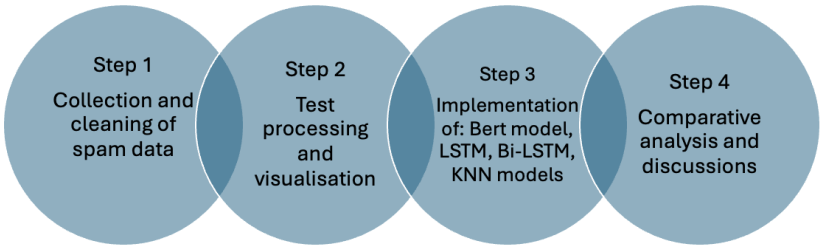

Figure 1. Steps conducted to gain the results.

3.1. Spam Data Collection and Cleaning

The data is collected from the open-source platform that is the Kaggle website [21, 22]. As the data contains the null values the data cleaning process is used. The data is cleaned in such a way it is ready for the implementation of the models. Two datasets

are taken and concatenated in such a way that two classes are balanced. This process takes place on the spam data.

3.2. Different Models Used to Detect the Email Spam

The BERT model is the framework used in NLP [23]. This model is used to understand the text by the computer. This model is trained to understand the text language and it finely tunes the data according to the user's needs, and this process is known as transfer learning. BERT model has an option in the voice search, but the challenging one is the words that with different meanings in the different sentences.

LSTM is the model generally used in the deep learning technique and it is a type of RNN. This LSTM is used for the improvement of the storage of the data and no loss of data as it has an efficient memory cell. The structure of LSTM (shown in Figure 2) has three gates: input, output, and forget gate. The input gate is for the input vectors and in the output gate, the memory cell is used for the effect on the outputs. The forget gate is used to remember past data.

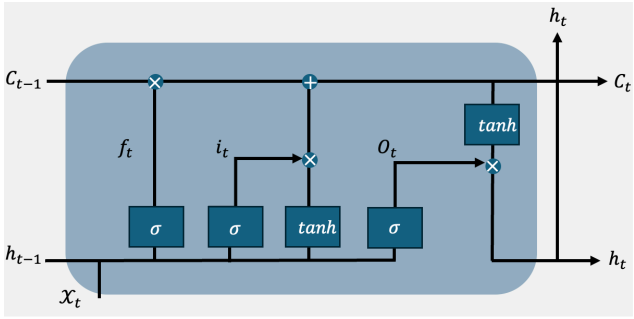

Figure 2. Structure of LSTM [24, p. 1676].

Bi-LSTM is the Bidirectional LSTM, and it is the extension of the LSTM which consists of two LSTMs. It consists of both the input layer and output layer as shown in Figure 3. One LSTM works in the forward direction and the other LSTM works in the backward direction. The forward direction is defined as the past to the future and the backward is for the future to the past.

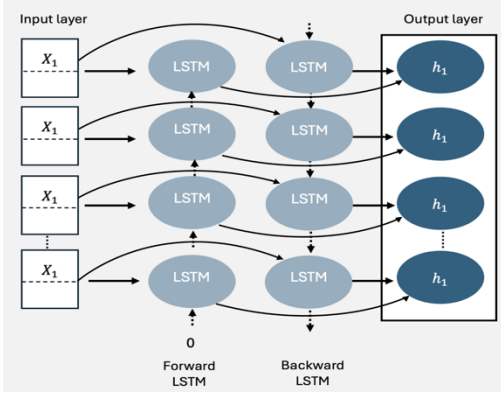

Figure 3. Structure and Process of Bi-LSTM [24].

KNN is the ML model that is used for the classification of the text and some question-answer type data and to detect the data. In terms of text classification, the KNN is a slow process while training the data. It is complicated to categorize the text and characters and to change in terms of numbers.

3.3. Evaluation Metrics

Confusion metrics are used for the classification of text in NLP. It is used to evaluate the classification model performance and in this N is the target class. As the name itself, it gets confused in the classification models. There are two classes present in the confusion metrics they are positive class and negative class. The true positive is the predictions are correct, the false positive is the predictions are wrong, the true negative is the prediction is correct, but the result is negative, and the false negative is the prediction is wrong and the result is negative.

1. Training Loss and Validation Loss

Training loss and validation loss are used in deep learning techniques. The training loss is defined as the errors obtained while training the data. The total errors are computed by sum and it is visualized in the form of a graph. Validation loss is the loss that occurs while the process of model implication. The errors obtained while training the data are calculated for each set and this loss is known as the validation loss.

2. Accuracy, Precision, and Recall

Accuracy is one of the evaluation metrics and it is used for the prediction of the models. The accuracy is defined as the correct prediction value to the total number of predictions as seen in (1).

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \quad (1)$$

where, TP = True positive,
TN = True negative,
FP = False positive,
FN = False negative.

Precision is the performance of the model which gives only positive predictions. Precision is defined as the division of a few positive predictions by the total number of predictions that are classified as positive as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \quad (2)$$

The recall is the number of positive samples that are detected, and it is defined as the dividend of true positive to the sum of the true positive and false negative:

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \quad (3)$$

4. EXPERIMENTAL RESULTS

The dataset was collected and cleaned from the open-source platform [21, 22]. The body and labels are renamed and concatenated to get the final data, such that no spam values are organized to be 3000 (in count). This new data was saved in the form of CSV and the details can be seen in Figure 4.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5928 entries, 0 to 5927
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   text    5928 non-null   object
 1   spam    5928 non-null   int64
dtypes: int64(1), object(1)
memory usage: 92.8+ KB
```

Figure 4. Information of the data

This data was considered for the implementation of the LSTM model, Bi-LSTM model as well as ML. Initially, there were no null values in the data set and the sequence of the number of words in the sentences was a higher density, and both the count of spam and not spam mail in the data set were also balanced. The 30 most common words were sorted out from the data set using the word count plot and the word "the" was at the top.

The word email is repeated a lot when compared to the free information and address, which are considered to be the important words in most spam emails, according to the word lemmatize, which was used to convert the data into short, and words before converting it into lowercase. This count plot is shown in Figure 5 and it shows the number of repeated words.

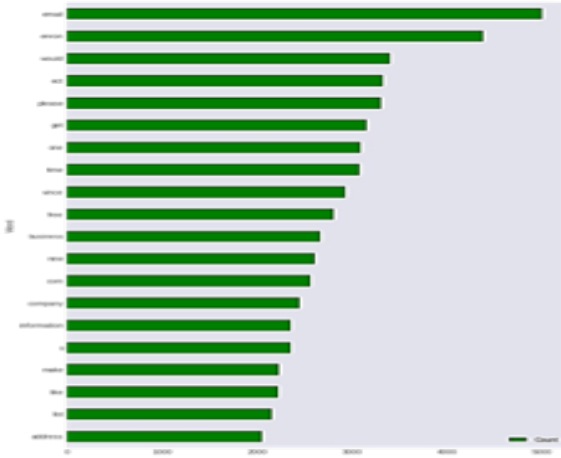

Figure 5. The most common words in the data.

The tokenizer was then initialized to match the email data, converting and creating the number sequences. The ML model was then implemented in the first phase of the project, which involved considering feature extraction using the count vectorizer and TF-IDF vectorizer, respectively. Predictions are made on the test set after the KNN model has been initialized and the data has been divided into training and testing so that the train size is 20 % of the whole data set. As a result, it was noted that the model had 97 % precision and 51 % of recall non-spam texts, while 66 % of precision was recorded

for spam-based texts, which explains the true positive labels in Figure 6. The use of the KNN model, however, has highlighted the fact that for the spam-based text, 98 % of recall was produced. The ML model's total testing accuracy was reported to be 74%. The receiver operating characteristics of the model are recorded as 89%. The K-fold cross-validation results obtained from KNN with 5-splits is 73.71%.

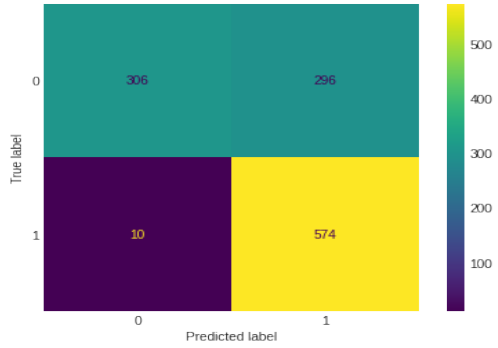

Figure 6. Confusion matrix of KNN classifier.

The snowball stemmer is used in this context with English so that cleaning is taken into account for the email text. Then it is applied to the data using Lambda functions, resulting in the storage of these texts as independent and dependent variables for LSTM and Bi-LSTM implementation. The maximum length of the sequence was initialized to be 50 and the test size is 20%. The word index size was 55,573 when tokenization was taken into account. For the training and testing sets, the pad sequence was used with the maximum length initialized above. The sequential model was developed using TensorFlow and Keras, and the monitor was initialized for the validation loss.

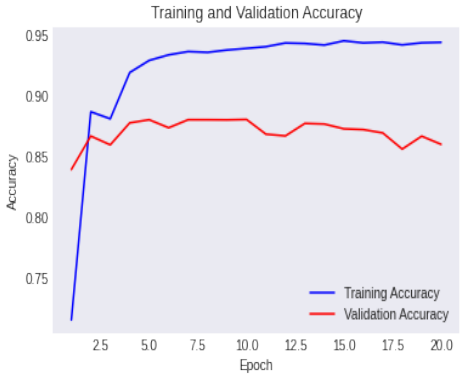

Figure 7. Accuracy graphs for the LSTM model.

Figure 8. Loss graphs for the LSTM model.

The rectified linear unit is employed as the activation function in the LSTM model, which was initially initialized with 50 as the input length and dropouts of 0.2. The validation accuracy and the training are tracked for 20 epochs while taking into account

the validation loss monitor and Adam Optimizer. Figure 7 and Figure 8 show the graphs for the validation and training accuracy as well as loss.

Similarly, the Bi-LSTM was also applied, over here sigmoid function is used for the activation and the maximum length is the similar as that of the LSTM. Over here the word size, the filter size and the loss function to binary cross entropy is initialized such that the matrix accuracy and validation accuracy are recorded. The model is fitted and trained for 10 epochs. The validation accuracy kept on increasing after the second epoch and the overall validation accuracy of a bidirectional LSTM applied for the spam classification was 97% as shown in Figure 9.

Similarly, the BERT model was initialized and evaluated based on the maximum accuracy, precision, and recall. The binary cross entropy was considered to be the function of loss along with the Adam optimizer. The testing accuracy increased from 83% to 86% from overall the epochs and the final procession value for the BERT model was 86% and 84% of the recall value was recorded as seen in Figure 10.

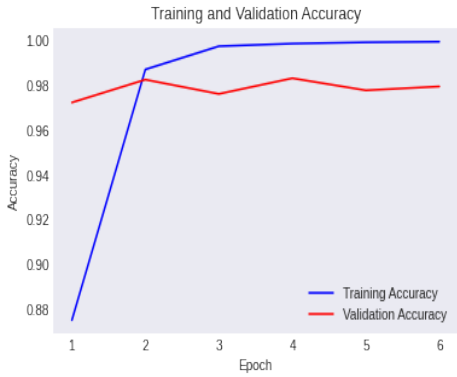

Figure 9. Accuracy graph for the Bi-LSTM model.

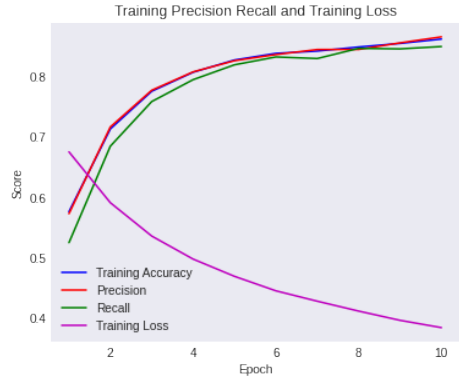

Figure 10. Performance evaluation for BERT model.

5. CONCLUSION AND DISCUSSION

The utilization of the Internet and social media networks for communication has significantly surged. Among these, email remains a firm and professional means of correspondence. As result, the classification of spam emails has gained considerable attention. For that, this work attempts to addresses the critical issue of spam mail utilizing a combination of machine learning and deep learning techniques. An advanced vectorization techniques (TF-IDF) was performed to enhance input data quality while attaining greater classification accuracy. Furthermore, investigation into leveraging NLP along with BERT for spam classification offered encouraging outcomes.

In this research BERT exhibited an extensive enhancement in accuracy, precision, and recall, underlining the prospective of transformer-based architectures in tackling complex text classification tasks. Additionally, the investigation simulations exposed substantial differences for different models' performance. Although well-known ML algorithm such as KNN demonstrated decent outcomes, the implementation of deep learning models such as LSTM along with Bi-LSTM delivered more promising results.

Bi-LSTM model demonstrated remarkable precision and recall values, thus emphasising its effectiveness in classifying among spam and non-spam emails. The Adam optimizer and the binary cross entropy were thought to be the function of loss. The final procession value for the BERT model was 86%, while the recall value was recorded at 84%. The testing accuracy rose from 83% to 86% across all epochs. Lastly, the validation accuracy score of Bi-LSTM is 97.94%, followed by 86.02% for LSTM, and 74.20% for KNN, while 98% for the recall. In this research ML and deep learning techniques were applied to classify email spams effectively, which can serve as an example for future research. Moreover, KNN showed high consistency dealing with positive samples with an overall performance of 98%.

The outcomes of this research imply that the utilization of a refined pre-processing mechanisms along with an advanced deep learning architectures could improve the spam mail classification systems enormously. However, further research is may be conducted to explore further avenues for enhancing model performance, such as incorporating ensemble methods and alternative deep learning architectures.

ACKNOWLEDGEMENT

The authors would like to thank the Researchers Supporting Project (No. RSPD2023R609), King Saud University, Riyadh, Saudi Arabia, for supporting this work.

REFERENCES

[1] M. Bassiouni, M. Ali, E. A. El-Dahshan. Ham and spam e-mails classification using machine learning techniques. *Journal of Applied Security Research*, vol.13, no.3, 2018, pp.315-331.

[2] G.A.Tran, Strutton.D. Comparing email and SNS users: Investigating e-services cape, customer reviews, trust, loyalty and E-WOM. *Journal of Retailing and Consumer Services*, vol.53, 2020, pp.101782.

[3] L. Gallo, A. Maiello, A. Botta, G. Ventre. 2 Years in the anti-phishing group of a large company. *Computers & Security*, vol.105, 2021, art.102259.

[4] I. Noninska, R. Romansky. Organization of technological structures for personal data protection. *International Journal on Information Technologies and Security*, vol.14, no.1, 2022, pp.97-106.

[5] R. Romansky. Digital age and personal data protection. *International Journal on Information Technologies and Security*, vol.14, no.3, 2022, pp.89-100.

[6] A. P. Rodrigues, R. Fernandes, A. Shetty, K. Lakshmanna, R.M. Shafi. Real-time twitter spam detection and sentiment analysis using machine learning and deep learning techniques. *Computational Intelligence and Neuroscience*, vol.2022, 2022.

[7] P. Charanarur, H. Jain, G. S. Rao, et al. *Machine-learning-based Spam mail detector*. *SN COMPUT. SCI*. vol.4, art. 858, 2023.

[8] P. Malhotra, S. Malik. *Spam email detection using machine learning and deep learning techniques*. Available at SSRN 4145123 (2022).

[9] P. K. Roy, J. P. Singh, S. Banerjee. Deep learning to filter SMS spam. *Future Generation Computer Systems*, vol.102, 2020, pp.524-533.

[10] K. Özçift, K. Akarsu, F. Yumuk, C. Söylemez. Advancing natural language processing (NLP) applications of morphologically rich languages with bidirectional encoder representations from transformers (BERT): an empirical case study for Turkish. *Automatika: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije*, vol. 62, no.2, 2021, pp.226-238.

[11] A. Misini, A. Kadriu, E. Canhasi. Authorship classification techniques: Bridging textual domains and languages. *International Journal on Information Technologies and Security*, vol.16, no.1, 2024, pp.27-38.

[12] V. S. Tida, S. Hsu. Universal Spam detection using transfer learning of BERT model. *arXiv preprint arXiv*, 2202.03480, 2022.

[13] S. Bouktif, A. Fiaz, A. Ouni, M. A. Serhani. Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches. *Energies*, vol.11, no.7, 2018, p.1636.

[14] J. Lin, R. Nogueira, and A. Yates. "*Pretrained transformers for text ranking: Bert and beyond*." Synthesis Lectures on Human Language Technologies 14, no. 4 (2021): 1-325.

[15] D. Jung, Y. Choi. Systematic review of machine learning applications in mining: Exploration, exploitation, and reclamation. *Minerals*, vol.11, no.2, 2021, p.148.

[16] A. Shahapurkar, S. F. Rodd. Efficient feature aware machine learning model for detecting fraudulent transaction in streaming environment. *International Journal on Information Technologies and Security*, vol.14, no.3, 2022, pp.3-14.

[17] F. Hossain, M. N. Uddin, R. K. Halder. Analysis of optimized machine learning and deep learning techniques for spam detection. In *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pp. 1-7. IEEE, 2021.

[18] Mohammed, R. Kora. An effective ensemble deep learning framework for text classification. *Journal of King Saud University-Computer and Information Sciences*, 2021.

[19] T. Young, D. Hazarika, S. Poria, Erik Cambria. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, vol.13, no.3, 2018, pp.55-75.

[20] Y. HaCohen-Kerner, D. Miller, Y. Yigal. The influence of preprocessing on text classification using a bag-of-words representation. *PloS one*, vol.15, no.5, 2020, e0232525.

[21] N. Bharathi, *Email spam dataset*. https://www.kaggle.com/datasets/nitishabharathi/email-spam-dataset (Visited on 26.10.2023)

[22] K.Veerakumar. *Spam Filter*. https://www.kaggle.com/datasets/karthickveerakumar/spam-filter. (Visited on 26.10.2023).

[23] N. Azzouza, K. Akli-Astouati, R. Ibrahim. Twitterbert: Framework for twitter sentiment analysis based on pre-trained language model representations. *International Conference of Reliable Information and Communication Technology*, pp. 428-437. Springer, Cham, 2019.

[24] B. Jang, M. Kim, G. Harerimana, S. U.Kang, J. W. Kim. Bi-LSTM model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism. *Applied Sciences*, vol.10, no.17, 2020.

Information about the authors:

Bandar Alshawi – is an Assistant Professor in the Department of Computer and Networks Engineering at Umm Al-Qura University. His interest in research includes artificial intelligence, wireless sensor networks, cyber security, and IoT.

Amr Munshi – is an Associate Professor with the Department of Computer and Networks Engineering, Umm Al-Qura University. His research interests include artificial intelligence, big data and smart grids.

Majid Alotaibi – is a Professor with the Department of Computer and Networks Engineering, Umm Al-Qura University. His research interests include Internet of Things, intelligent systems and networking.

Ryan Alturki –is an Associate Professor with the Department of Software Engineering at Umm Al-Qura University. His research interests include eHealth, mobile technologies, the Internet of Things, artificial intelligence, cloud computing, and cybersecurity.

Nasser Allheeib – is Assistant Professor at King Saud University, Saudi Arabia. He received his Ph.D. degree from the Clayton School of Information Technology, Monash University, Australia. His research interests include query processing, spatial databases and artificial intelligence.

Manuscript received on 07 April 2024