ICIC Express Letters
Part B: Applications
Volume **13**, Number **10**, October **2022**

 ISSN 2185-2766
pp. **1093–1100**

SHORT MESSAGE SERVICE (SMS) SPAM FILTERING USING DEEP LEARNING IN BAHASA INDONESIA

Dimas Ramdhan, Henry Lucky, Ade Putera Kemala
and Andry Chowanda

Computer Science Department, School of Computer Science
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
{ dimas.ramdhan; henry.lucky; ade.kemala }@binus.ac.id; achowanda@binus.edu

Received January 2022; accepted April 2022

Abstract. *This paper proposes the exploration of deep learning architectures and algorithms to model a spam classifier. The model proposed in this research improved the performance of the spam classifier model from the previous work. Furthermore, this research proposes models trained using Transformer-based models such as IndoBERT, Multilingual BERT, and XLM. This research also proposed a model trained with Gated Recurrent Unit (GRU) with the IndoBERT model. Due to the nature and the size of the dataset, data augmentation techniques such as Easy Data Augmentation (EDA) and nlpaug were implemented to increase the data quality. The results indicate that deep learning techniques proposed in this paper have significantly increased the accuracy of the spam classification model compared to the previous work. EDA technique also increases the accuracy of the base models. The highest accuracy of 99.35% is reached using the EDA technique and IndoBERT model.*
Keywords: Short message service, Spam, Deep learning, Natural language processing

1. **Introduction.** With a population of 270.2 million people, Indonesia is currently the fourth most populous country globally. With a large population, in 2021, smartphone users in Indonesia have reached 199.2 million, making Indonesia the fourth-largest smartphone market in the world after China, India, and the United States. This growth is also predicted to increase until 2026 [1]. From these data, it can be seen that 73.3% of the Indonesian population uses smartphones. [2] stated that the number of smartphone users in Indonesia from 2017 to 2020 has increased. This number is also forecasted to continue to grow until 2026. With a surprisingly extraordinary number of smartphone users in 2020, one of the largest telecommunications providers in Indonesia managed to reap a profit of 75.82 trillion rupiah. The profits are obtained from data plans, the Internet, and Short Message Service (SMS) used by 163 million users.

From the data, SMS is one factor that must be considered because it also affects a significant source of income. Unfortunately, with the huge number of SMS users, SMS media is often misused as a medium of fraud against phone users. Fraud often happens by sending massive and random SMS up to ten thousand per day to all users and becoming spam SMS for many people. Spam SMS has several types: venture capital offers, investment, promotion, and fraud. The Indonesian Telecommunications Regulatory Agency (BRTI) also intervened to resolve this issue. BRTI will require all cellular operators to provide a feature that allows users to reject or receive particular SMS broadcast offers from operators or their partners. In addition, BRTI hopes that all communication service providers can provide features for customers to no longer receive SMS that they consider

DOI: 10.24507/icicelb.13.10.1093

disturbing. The features in question are opt-in and opt-out. This feature will allow users to choose whether they want to receive similar SMS or not [3].

Machine and deep learning can be implemented to build a spam classifier model. The main problem with deep learning is the availability of the dataset, particularly datasets in languages other than English, such as Bahasa. The data augmentation method has been proven to overcome the lack of datasets and improve accuracy in previous research [4]. Hence, this research aims to explore several deep learning architectures and data augmentation techniques to build spam classifier models for SMS. A similar previous study conducted by Theodorus et al. using machine learning resulted in 94% accuracy [5]. The results of our paper show that using deep learning can improve the performance of the spam filtering model up to 99.35% accuracy. In addition, we found that a simple data augmentation technique that produces more data gives better performance than a more complex technique that produces fewer data.

Our contributions in this paper are as the following: 1) we apply the state-of-the-art deep learning approaches in Natural Language Processing (NLP) to building Indonesian SMS spam filtering model, thus increasing the performance of the model from the previous study; 2) we explore several data augmentation techniques to increase the number of data and analyze their effects on the spam filtering model.

This research is divided into several sections. The second section discusses the previous research. The third section discusses the datasets and the methods used in the research: data pre-processing, data augmentation, and deep learning techniques. The fourth section discusses the result of this study. The last section is Section 5, and this section will show and discuss the research result.

2. **Related Works.** Several different methods have been proposed and implemented to filter spam from SMS in the past few years. The traditional machine learning approach was successful, and popular techniques were used to deal with SMS spam filtering tasks. A study done by Fernandes et al. [6] implemented Optimum-Path Forest-based (OPF) classifiers to filter out spam. The study compares OPF with K-Nearest Neighbor (KNN), Support-Vector Machine (SVM), and Artificial Neural Networks with Multilayer Perceptrons (ANN-MLP). The results show that training the OPF classifiers requires less time and resources than other algorithms. OPF perfectly identified ham messages but misidentified half of the spam messages as ham. The OPF classifier accuracy is higher than other algorithms except for SVM. However, SVM requires significantly more resources to classify messages compared to OPF. Another study done by Luo et al. [7] proposed a spam detection model with machine learning algorithms; they also explored Logistic Regression (LR), KNN, and Decision Tree (DT) to classify spam and ham messages. The result shows that LR is the best classifier for classifying spam and ham messages. However, the dataset's quality is not good with highly imbalanced data and no indication of undersampling or other techniques to normalize the data. Theodorus et al. [5] proposed a spam classifier supporting the Indonesian language. The models were trained with Multinomial Naive Bayes (MNB), Multinomial Logistic Regression (MLR), SVM, KNN, DT, Stochastic Gradient Descent (SGD), XGBoost (XGB), and Random Forest (RF). In addition, they also implement Bag of Words and Term Frequencies to extract the text features. Due to the substantial imbalance classes in the dataset, they also performed an undersampling method on the data. The model was also tested with an external dataset with different language styles from the one they trained. The best result is gained from the experiments where the model uses a selection from the dataset with an equal ratio for each class. The result suggests that RF has the best results with 10-fold cross-validation, while MLR and XGB both produced the best results with the external testing dataset.

While the machine learning approaches produced good results, researchers are commencing using the deep learning model for filtering spam in SMS with the advance of the

machine learning field. For example, work done by Raj et al. [8] implemented Long Short-Term Memory (LSTM) more efficiently. Furthermore, by implementing Word2Vec as the word embedding, the proposed LSTM model gains better results than the baseline algorithms, including KNN, SVM, NB, RF, and DT. Another work in deep learning was done by Roy et al. [9]. They implemented Convolutional Neural Network (CNN) and LSTM architectures for SMS spam filtering. They also compared several traditional machine learning algorithms, such as DT, LR, RF, SGD, and Gradient Boosting (GB). The experiments conducted by the authors indicated that CNN and LSTM performed much better than the traditional machine learning approaches tested in terms of SMS spam filtering. With the popularity of Transformer-based architecture in NLP [10], several works have been done by implementing the Transformer-based model. For example, the work done by Liu et al. [11] modified the Transformer architecture to apply it to the SMS spam detection task. Furthermore, they evaluated their model by comparing it with several other SMS spam detection approaches on the SMS and Twitter datasets. The experimental results show that the proposed model performs better on both datasets compared to traditional machine learning and other deep learning algorithms, such as LR, NB, RF, SVM, LSTM, and CNN-LSTM [12].

To further explore the possibility of Transformer-based architecture, we propose the use of a Transformer-based model, Bidirectional Encoders Representation from Transformer (BERT) [13], which is known to be exceptionally good for any NLP tasks, for SMS classification task using Indonesian dataset. Furthermore, we also propose the joint architecture between BERT and GRU for experimental research.

Four models will be used in this research. The first one is using the IndoBERT model, a pre-trained BERT model that uses a huge Indonesian language corpus to train it [14]. We will also use the multilingual BERT model BERT-ML [13] and XLM [15]. Lastly, we propose combining the IndoBERT model with the bi-GRU layer as a classifier.

3. Methodology.

The methodology proposed in this research comprises three major phases: dataset gathering, dataset pre-processing, and model training using a deep learning model.

3.1. Dataset.

The dataset used in this paper is gathered from the previous study [5]. The dataset contains 4,125 Indonesian messages gathered from SMS with three classes: ham, spam, and promo. There is also a second batch of the dataset containing 1,260 messages. The second batch dataset is the under-sampled variant of the original dataset with a balanced amount of data from each class (420 messages in each class). The detail of the dataset is shown in Table 1. In the first batch, the Ham class is the majority class with 2,323 messages, followed by the Promo class with 1,382 messages. The Spam class is the minority class with 420 messages in the dataset.

Table 1. Detail of dataset class distribution

	Ham	**Promo**	**Spam**
1st batch	2,323	1,382	420
2nd batch	420	420	420

3.2. Data pre-processing.

Deep learning methods require a massive amount of data to achieve the best result; however, the amount of data in the dataset for the Indonesian spam filtering task is relatively small. Therefore, several augmentation techniques were proposed to enrich the data. Data augmentation is a technique used to increase data by adding slightly modified copies of already existing data or newly created synthetic data from existing data. We consider two techniques in this paper, Easy Data Augmentation (EDA) [16] and nlpaug [17].

3.2.1. *Easy data augmentation.* EDA is one of the augmentation techniques implemented in this research to help improve text classification performance. This technique is inspired by a computer vision approach that has increased the model's accuracy by using augmented data. EDA has several powerful operations consisting of Synonym Replacement (SR), Random Insertion (RI), Random Swap (RS), and Random Deletion (RD). SR's function randomly selects n number of words from the sentence that does not stop the word and replaces each of these words with one of their randomly selected synonyms. RI functions as one of the operations to randomly search for synonyms of random words in a sentence that is not a stop word and then enter the synonym words into sentences with random positions. This operation is executed n times. Operation RS is used to randomly select two words in a sentence and then swap their positions, and this operation is also executed n times. Finally, the RD operation aims to eliminate randomly every word selected in the sentence with a probability of p. This paper uses the recommended parameters as a reference in performing data augmentation. Three variables consist of Ntrain, α, and naug. Ntrain is the number of training samples, α is the percent of words in a sentence changed by each augmentation, naug is the number of generated augmented sentences per original sentence. These three variables are used as a reference in choosing how much data can be augmented. Table 2 illustrates the total amount of the final data augmented using EDA. This method generated five times the total of original data.

Table 2. Before and after data augmentation with EDA and nlpaug

Batch	Before	EDA After	nlpaug After
1st batch	4,125	20,625	8,250
2nd batch	1,260	6,300	2,520

3.2.2. *nlpaug.* nlpaug [17] is a python library that can generate synthetic data to increase the number of datasets. This library can process both text data and audio data types. Among the available methods in the library, we choose to use the contextual word embedding method. This method utilizes the embedding generated by the BERT pre-training model. Then the input word will be replaced with a word with a similar embedding vector value to the input word. The BERT pre-trained model used to generate the embedding was 'bert-base-multilingual-uncased' [13]. The hyperparameter used in this process is an aug_p value of 0.1, meaning that 10% of the total words will be changed. The amount of augmented data can be seen in Table 2. Due to hardware limitations, it is decided only to generate twice the amount of original data.

3.3. **Deep learning technique.** This paper focuses on using deep learning algorithms for classifying SMS messages, particularly using the Transformer-based models. BERT [13] is exceptionally popular and heavily used by several researchers due to its excellent performance in multiple NLP tasks. Currently, there are two large-scale Indonesian BERT models [14,18].

While Koto et al. [18] pre-trained their IndoBERT mostly with news datasets, Wilie et al. (2020) pre-trained their IndoBERT with more general data, such as Wikipedia, webpage articles, Twitter, and other sources. Hence, the IndoBERT pre-trained model from Wilie et al. [14] was implemented as the model is more general and can suit the dataset used in this research with a similar language style with their pre-training data.

This paper also considers multilingual models such as BERT-ML [13] and XLM [15] for experimental research because of their excellent performance and comparison to IndoBERT. In addition, this research proposed an architecture using the IndoBERT with Gated Recurrent Unit (GRU) to see whether it could boost IndoBERT's performance.

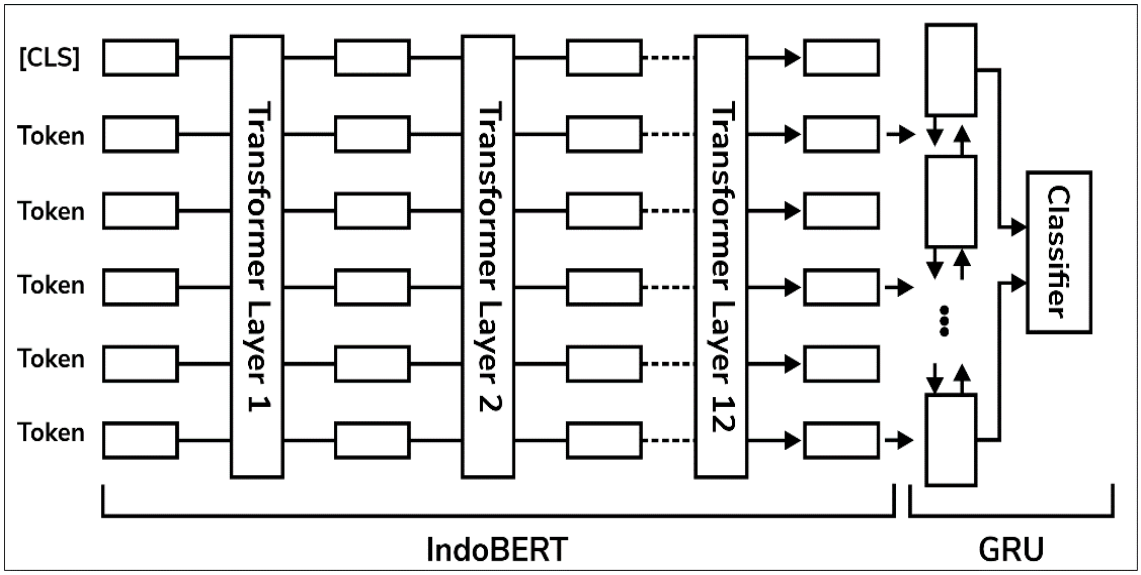

FIGURE 1. Architecture of IndoBERT + GRU

The IndoBERT + GRU architecture can be seen in Figure 1. A bi-directional layer variant of GRU was implemented in this research to increase the final model performance.

For training IndoBERT, BERT-ML and XLM, we used Adam as an optimizer with a learning rate of 2e-5 and batch size of 16. We trained the models for five epochs. Meanwhile, for IndoBERT + GRU, we freeze the IndoBERT weight and update the weight for GRU and the classifier. We trained the model with a learning rate of 2e-05 and a batch size of 16. As for the GRU layer, we configure it with a double bi-directional GRU layer, 256 hidden dimensions, and trained for five epochs.

4. **Result and Discussion.** Table 3 illustrates the results of the experiments. In overview, the model trained with IndoBERT architecture using the EDA augmentation method performed better than the other models. The highest accuracy (99.35%) was achieved by the model trained with 10-fold cross-validation using the second batch dataset.

TABLE 3. Result and evaluation

Model	**Augmentation type**	**Average 10-fold cross validation accuracy**		**Train test split 80 : 20**	
		1st batch	**2nd batch**	**1st batch**	**2nd batch**
IndoBERT	Normal	94.79%	99.05%	94.30%	98.55%
	EDA	96.57%	**99.35%**	96.14%	**99.28%**
	nlpaug	95.76%	99.25%	96.54%	99.00%
BERT-ML	Normal	94.72%	98.84%	93.58%	97.88%
	EDA	**96.62%**	98.85%	**96.80%**	99.20%
	nlpaug	95.67%	99.26%	95.69%	97.42%
XLM	Normal	94.52%	98.57%	93.33%	98.15%
	EDA	96.60%	99.04%	96.46%	99.20%
	nlpaug	95.21%	98.57%	96.06%	98.01%
IndoBERT + GRU	Normal	94.35%	90.39%	92.84%	94.04%
	EDA	96.09%	98.31%	96.50%	97.69%
	nlpaug	94.95%	94.12%	95.45%	93.05%

The second-best accuracy was achieved by the model trained with a second batch dataset that is split to 80 : 20 ratio using IndoBERT model and EDA augmentation method (99.28%). Thus, the model proposed in this research achieved better performance compared to the model proposed in the previous work [5]. The previous work achieved the highest accuracy of 94.62% in the first batch dataset and 95% in the second batch dataset using the RF algorithm. Deep learning algorithms, especially the Transformer-based, are known for their superiority over traditional machine learning algorithms. There is also no significant difference in performance between the average 10-fold cross-validation method and the 80 : 20 train-test split, which means the dataset is good enough for stable training on deep learning models. We also observe that the second batch of the dataset could boost the performance of the proposed models as it has a balanced class distribution.

4.1. **Model comparison.** Table 3 also demonstrates the results of all four Transformer-based architectures implemented in this research. The models trained with the k-fold cross-validation method resulted in better performance in overall results. The best model trained with the k-fold cross-validation method was achieved by the one trained with BERT-ML (96.62%) for the first batch of the dataset and the IndoBERT (99.35%) the second batch of the dataset. The best model trained with the 80 : 20 train-test split model was achieved by the one trained with the BERT-ML model with EDA (96.80%) for the first batch of the dataset and the IndoBERT model with EDA (99.28%) for the second batch of the dataset. Overall, for the first batch of the dataset, the BERT-ML model performs better than the other models with EDA. In contrast, the IndoBERT model is superior to the others with EDA for the second batch of the dataset.

Despite the use of augmentation, the IndoBERT model successfully beat other models in the normal setting, which is expected, as it has been pre-trained on the same language as our SMS dataset. Also, on average, the model that has the most stable performance for all scenarios is IndoBERT, with an average accuracy of 97.38%.

4.2. **The effect of data augmentation.** Figure 2 illustrates the average model accuracy based on the type of dataset used. According to the results, the augmentation method improves model performance for all test scenarios. Based on Figure 2, there is a trend of increasing accuracy for all models starting from the standard (non-augmented) data, augmented data using the nlpaug method, and augmented data using the EDA method. nlpaug uses the Multilingual BERT model for its augmentation operation. In contrast, the EDA method uses Indonesian WordNet for its operations, namely synonym replacement, insertion, swap, and deletion, which are able to create noises in training data, considering the SMS dataset contains messages that are alike.

Also, even though EDA uses simple operations, the unilingual Indonesian WordNet is more tuned for our dataset, which contains Indonesian SMS, than the Multilingual BERT used by nlpaug. Furthermore, the amount of dataset produced by EDA is significantly higher than the amount of dataset produced by nlpaug, which also makes the model with EDA achieve higher accuracy than nlpaug. Frankly speaking, the amount of augmented dataset could be tuned and equated for each augmentation method. However, this paper used the default parameters, thus producing a different amount of augmented dataset.

4.3. **IndoBERT + GRU performance.** Based on the results in Figure 2, which showed the averaged performance of every model according to every augmentation method, the IndoBERT + GRU model has a lower performance than the other three models. The low performance is caused by the IndoBERT model's parameters which are frozen, and the model is only used to generate embedding of input tokens. In addition, the GRU model that is trained only consists of one bidirectional layer. Hence, it can be concluded that the use of GRU in the IndoBERT model does not give better results in SMS classification.

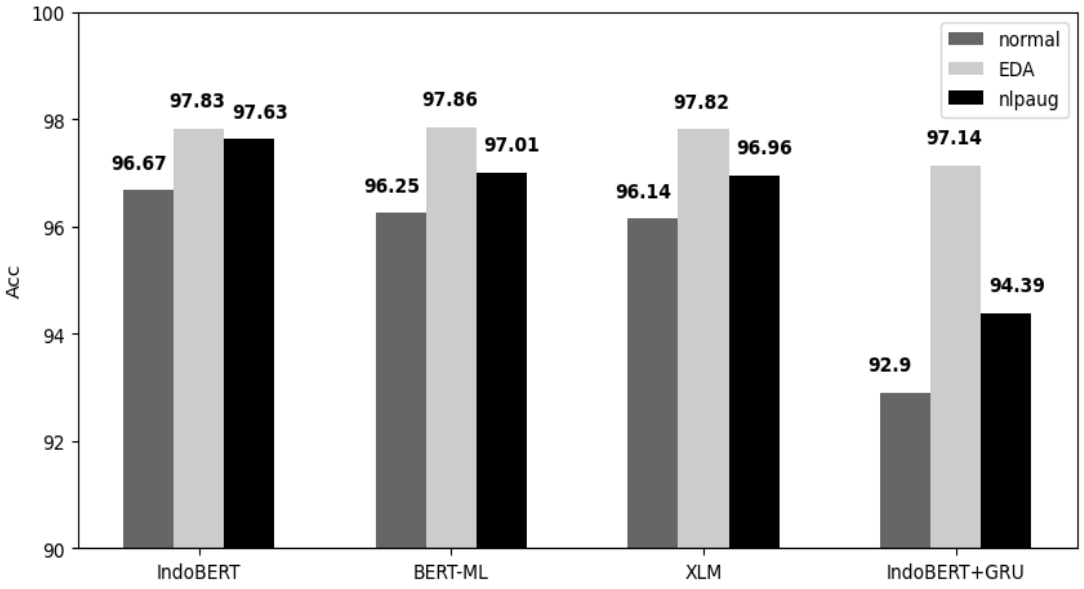

FIGURE 2. Result comparison on averaged accuracy across different schemes

4.4. **Comparison.** Table 4 shows the comparison of our best result with previous research's best result [1]. We compare our best models, IndoBERT and BERT-ML with the EDA augmentation method, with the previous research best method, namely RF and MLR, without any data augmentation. It is shown that all our results outperformed the previous research in every testing scenario.

TABLE 4. Result comparison

Model	Augmentation type	Average 10-fold cross validation accuracy		Train test split	
		1st batch	2nd batch	1st batch	2nd batch
IndoBERT	EDA	96.57%	**99.35%**	96.14%	**99.28%**
BERT-ML	EDA	**96.62%**	98.85%	**96.80%**	99.20%
RF	Normal	94.62%	95%	77.25%	82.67%
MLR	Normal	94.57%	94.6%	78.39%	83.90%

5. **Conclusions.** Four architectures, three dataset settings, two batches of datasets, and two data split settings were explored in this research, resulting in 48 combinations of settings. The best accuracy was achieved by the model trained with 10-fold cross-validation in IndoBERT architecture with EDA (99.35%) and the second batch dataset. The second-best accuracy was achieved by the model trained with an 80 : 20 train test split dataset in the second batch dataset with IndoBERT architecture with EDA (99.28%). The results achieved by these experiments are superior to the previous work used as the baseline in this research. In addition, the models trained with the deep learning architecture could produce better performance than the machine learning method. Dataset augmentation methods can also improve performance. More deep learning architecture can be explored to build the best spam classifier from SMS or social media for future research direction. Moreover, more datasets can be collected from SMS and social media to enhance the classifier's performance. As this paper used the default parameter and did not investigate the different parameters in augmentation methods, it causes differences in the results of data augmentation. The EDA technique was able to generate five times the amount of

the original dataset, while the nlpaug technique was only able to generate two times the dataset. Future research can also investigate the parameter in the augmentation methods.

REFERENCES

[1] S. Gadde, A. Lakshmanarao and S. Satyanarayana, SMS spam detection using machine learning and deep learning techniques, *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Chengdu, China, pp.358-362, DOI: 10.1109/ICACCS51430.2021.9441783, 2021.

[2] H. Nurhayati-Wolff, *Data, Internet and SMS revenue of PT Telkom Indonesia Tbk from financial year 2017 to 2020*, https://www.statista.com/statistics/750868/telkom-indonesia-data-internet-and-sms-revenue-indonesia/, 2021.

[3] L. Hasibuan, *SMS Spam Rises, BRTI Requests This to Telkomsel, XL, Indosat*, https://www.cnbcindonesia.com/tech/20201002103826-37-191138/sms-spam-marak-brti-minta-ini-ke-telkomsel-xl-indosat-cs, Accessed on September 09, 2021.

[4] A. Harras, A. Tsuji, S. Karungaru and K. Terada, Enhanced vehicle classification using transfer learning and a novel duplication-based data augmentation technique, *International Journal of Innovative Computing, Information and Control*, vol.17, no.6, pp.2201-2216, DOI: 10.24507/ijicic.17.06.2201, 2021.

[5] A. Theodorus, T. K. Prasetyo, R. Hartono and D. Suhartono, Short message service (SMS) spam filtering using machine learning in Bahasa Indonesia, *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, Surabaya, Indonesia, pp.199-202, DOI: 10.1109/EIConCIT50028.2021.9431859, 2021.

[6] D. Fernandes, K. A. P. Da Costa, T. A. Almeida and J. P. Papa, SMS spam filtering through optimum-path forest-based classifiers, *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, Miami, FL, USA, pp.133-137, DOI: 10.1109/ICMLA.2015.71, 2015.

[7] G. Luo, S. Nazir, H. U. Khan and A. U. Haq, Spam detection approach for secure mobile message communication using machine learning algorithms, *Security and Communication Networks*, vol.2020, 8873639, DOI: 10.1155/2020/8873639, 2020.

[8] H. Raj, W. Yao, S. K. Banbhrani and S. P. Dino, LSTM based short message service (SMS) modeling for spam classification, *ACM International Conference Proceeding Series*, pp.76-80, DOI: 10.1145/3231884.3231895, 2018.

[9] P. K. Roy, J. P. Singh and S. Banerjee, Deep learning to filter SMS spam, *Future Generation Computer Systems*, vol.102, pp.524-533, DOI: 10.1016/j.future.2019.09.001, 2020.

[10] A. Vaswani et al., Attention is all you need, *arXiv.org*, arXiv: 1706.03762, 2017.

[11] X. Liu, H. Lu and A. Nayak, A spam transformer model for SMS spam detection, *IEEE Access*, vol.9, pp.80253-80263, DOI: 10.1109/ACCESS.2021.3081479, 2021.

[12] A. Ghourabi, M. A. Mahmood and Q. M. Alzubi, A hybrid CNN-LSTM model for SMS spam detection in Arabic and English messages, *Future Internet*, vol.12, no.9, 156, DOI: 10.3390/fi12090156, 2020.

[13] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, *arXiv.org*, arXiv: 1810.04805, 2019.

[14] B. Wilie et al., IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding, *Proc. of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Suzhou, China, pp.843-857, 2020.

[15] S. Ruder, A. Søgaard and I. Vulić, Unsupervised cross-lingual representation learning, *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, Florence, Italy, pp.31-38, DOI: 10.18653/v1/p19-4007, 2019.

[16] J. Wei and K. Zou, EDA: Easy data augmentation techniques for boosting performance on text classification tasks, *Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, pp.6382-6388, DOI: 10.18653/v1/d19-1670, 2019.

[17] M. Ciolino, D. Noever and J. Kalin, Multilingual augmenter: The model chooses, *arXiv.org*, arXiv: 2102.09708, 2021.

[18] F. Koto, A. Rahimi, J. H. Lau and T. Baldwin, IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP, *Proc. of the 28th International Conference on Computational Linguistics*, Barcelona, Spain, pp.757-770, DOI: 10.18653/v1/2020.coling-main.66, 2020.