Journal of Information Systems and Informatics

Vol. 6, No. 3, September 2024 e-ISSN: **2656-4882** p-ISSN: **2656-5935**

DOI: 10.51519/journalisi.v6i3.838 Published By DRPM-UBD

Indonesian Health Question Multi-Class Classification Based on Deep Learning

Wayan Oger Vihikan[1], I Nyoman Prayana Trisna[2]

[1,2]Information Technology Department, Udayana University, Bali, Indonesia
Email: [1]oger_vihikan@unud.ac.id, [2]prayana.trisna@unud.ac.id

Abstract

The health online forum is commonly used by Indonesian to ask questions related to diseases. A well-known example, Alodokter, has hundreds of thousands of health questions which are assigned to certain topics. Building a model to classify questions into a topic is important for better organization and faster response by relevant health professionals. This research experimented on 20 deep learning methods from RNN, CNN, and IndoBERT with different configurations to see the performance of each model when classifying questions into six different most common diseases that cause death in Indonesia. The results show the majority of the model can outperform the SVM as baseline. Bidirectional RNN such BiLSTM and BiGRU combined with CNN show a good metric score even though a certain version of the IndoBERT model generally outperforms all the other models.

Keywords: Health Question, Text Classification, Deep Learning, IndoBERT

1. INTRODUCTION

Indonesia is currently grappling with a significant burden of non-communicable diseases, which emerged as the leading cause of death among the population in 2019 [1]. This public health challenge is further complicated by the country's vast and diverse geography, which, along with disparities in healthcare access, exacerbates the difficulty of providing timely and effective medical care to all citizens. In response to these challenges, many Indonesians have turned to online health forums, such as Alodokter, to seek advice and information from verified health professionals. Alodokter is one of the most prominent health forums in Indonesia, hosting hundreds of thousands of health-related questions spanning thousands of topics. However, the sheer volume of inquiries necessitates an efficient system for organizing and categorizing these questions to ensure that users receive accurate and timely responses from relevant health experts.

The task of categorizing these health-related questions can be framed as a text classification problem, a common task in the field of natural language processing (NLP). Extensive research has been conducted on text classification, employing both traditional machine learning and more recent deep learning approaches. Applications of text classification are varied, covering areas such as sentiment analysis [2-4], legal document categorization [5, 6], and medical document classification [7, 8]. Specifically, in the realm of question classification, several studies have explored the use of Transformer-based methods and Bidirectional RNNs [9, 10]. Despite these advancements, there is limited research focused on the unique context of classifying health-related questions within the Indonesian language, particularly those from Alodokter's platform.

Previous research that did address Alodokter's question and answer data employed an ensemble method for classification [11]. However, this study primarily focused on classifying the content of the answers rather than the questions themselves. Moreover, it was restricted to a small subset of deep learning models, namely CNN, LSTM, and BERT, leaving out the exploration of a broader range of potentially more effective approaches. This limitation underscores a significant gap in the current research landscape, particularly in the context of classifying health-related questions in Indonesian, which is critical for improving the efficiency and accuracy of online health forums like Alodokter.

To address this gap, the present research aims to implement and evaluate multiple deep learning methods for the classification of Alodokter's community health questions. The goal is to categorize these questions into six predefined categories based solely on the question descriptions. The deep learning methods under consideration include LSTM, GRU, Bidirectional LSTM and GRU (BiLSTM and BiGRU), CNN, BiLSTM-CNN, BiGRU-CNN, and Transformer-based models pre-trained on Indonesian text, such as IndoBERT. By exploring various configurations of these models, this study seeks to identify the most effective approach for accurately classifying the questions.

In addition to implementing and comparing these deep learning models, this research will conduct a detailed text analysis to understand the causes of misclassification in the best-performing model. By identifying the sources of error, the study aims to provide insights that could inform further refinement of classification models, ultimately contributing to the development of more accurate and reliable systems for managing and responding to health-related inquiries on online platforms like Alodokter.

2. METHODS

The research follows four main processes as can be seen in Figure 1. The process is started with data preparation which includes tag selection, website scraping, and train-test data split. Then, it is followed by text pre-processing, model training and finally model evaluation.

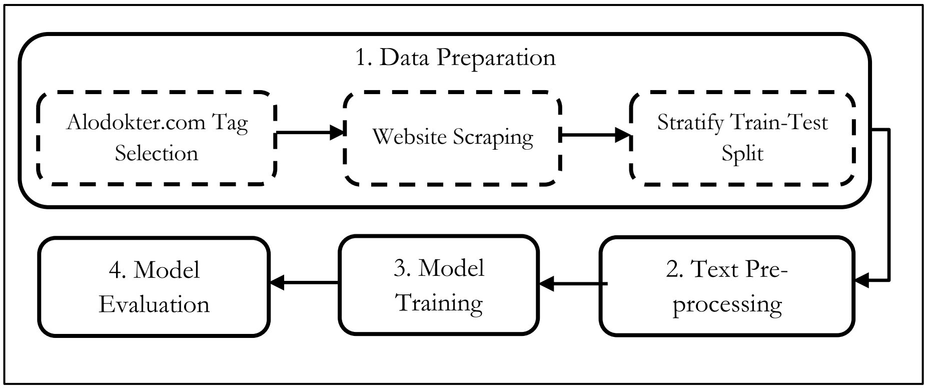

Figure 1. Steps of the research

2.1. Data Preparation

Data collection process starts from selecting which category of disease is going to be used in classification. The authors decided to use Top 10 Cause of Death in Indonesia data from WHO [1] since those diseases are potentially the most asked, and generally require faster response from the medical specialist. Those disease are Stroke, Ischaemic Heart Disease (IHD), Diabetes Mellitus (DM), Tuberculosis (TB), Cirrhosis of The Liver (CoL), Chronic Obstructive Pulmonary Disease (COPD), Diarrhoeal Diseases (DD), Hypertensive Heart Disease (HHD), Lower Respiratory Infections (LRI), and Neonatal Conditions.

Based on those ten diseases, related topics are selected manually. The question's URL under each topic dating up to 30th June 2024 are scraped from Alodokter community topic page. Once all question URLs are gathered, the duplicate URLs are removed because there are cases where one question belongs to multiple topics. Then, all of the question pages from the collected URL are scraped. The collected fields are question title, question description, question date, answer author, answer description, and answer date. The scraping process is done by using Selenium and BeautifulSoup Python libraries. The related tags with the English translation and number of samples of each disease can be seen in Table 1.

The Neonatal Conditions does not have a related topic, so it is excluded. There are only six diseases that have a number of samples higher than a thousand.

Therefore, only six categories of diseases are used as labels in this research, they are Stroke, TB, CoL, DD, HHD, and LRI. The text data used for this research is only the question description, hence the other fields are dropped. The data is then stratified split randomly into train and test dataset with 80% and 20% ratio respectively. In training the deep learning model, the train dataset is stratified split randomly again into new training dataset and validation dataset with ratio 90% and 10% respectively.

Tabel 1. Disease category and its related Alodokter topic

Disease Category	Related Topic	# of Sample
Stroke	Stroke	**1724**
IHD	Penyakit Jantung Koroner (Coronary Heart Disease) Serangan Jantung (Heart Attack)	766
DM	Diabetes Tipe 1 (Type 1 Diabetes) Diabetes Tipe 2 (Type 1 Diabetes)	395
TB	Tuberculosis	**6146**
CoL	Sirosis (Cirrhosis) Hepatitis B Hepatitis C	**1786**
COPD	Penyakit Paru Obstruktif Kronis (Chronic Obstructive Pulmonary Disease)	251
DD	Diare (Diarrhea)	**3363**
HHD	Hipertensi (Hypertension)	**2384**
LRI	Pneumonia Bronkitis (Bronchitis)	**2105**

2.2. Text Pre-processing

The text in the question description field is written by the user who asks the question. This leads to a word usage that is often informal, in the form of colloquial, abbreviated, grammatically incorrect, and contains typos. In addition to that, some of the text still contains HTML tags and Unicode characters. Therefore, the text cleaning process and word normalization are done consecutively.

The cleaning process removes the HTML tags, excessive whitespaces, excessive punctuations, and Unicode characters such as emoticons and non-Latin characters. Word normalization process is done to substitute common colloquial, abbreviated words, and typo to its normal form. An Indonesian colloquial dictionary [12] is used to aid the normalization of colloquial words. Additional custom dictionary is also created manually to handle abbreviated words, grammatically incorrect words, and typos with frequency higher than ten in the training and test dataset.

In the training process, a simple SVM algorithm with TF-IDF is used as the baseline for performance comparison. Therefore, the dataset for SVM is processed further. The punctuations and numeric characters are removed, stop words are removed except negation such as *tidak*, *bukan*, *jangan*, and *belum*, and lastly the word is stemmed using Sastrawi stemmer library [13-15].

2.3. Model Training

There are multiple algorithms used to create the question classification model. The Support Vector Machine (SVM) with Term Frequency – Inverse Document Frequency (TF-IDF) is used as the baseline. SVM has shown to perform well on text classification [2, 4, 16] and train faster compared to deep learning methods that are used in this research. The SVM is trained using grid search and cross validation to get the best hyper-parameters. TF-IDF vectorizer is used to transform the text into vectors with TF-IDF values from 1-gram and / or 2-gram.

The deep learning algorithms used in this research are from the Recurrent Neural Network (RNN) family, they are Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Bidirectional LSTM (BiLSTM), and Bidirectional GRU (BiGRU). LSTM is an improvement of vanilla RNN which prevents the vanishing gradient issue so it can propagate weight updates to all layers [17]. GRU is a simplification of LSTM that makes it less complex and more efficient than LSTM [17]. The bidirectional variants learn the forward and backward information of text which can improve model understanding on long text [18]. These methods have shown a good performance in text classification especially the bidirectional version [18-20]. Another well-known deep learning method that performs well in text classification [21, 22] and implemented in this research is Convolutional Neural Network (CNN). CNN utilizes convolutional layers and filters to process multiple neighboring words and get its pattern [17].

Both RNN and CNN use FastText as word embeddings. It has been trained on Indonesian Common Crawl and Wikipedia with a vector dimension of 300. A combination of RNN and CNN methods are also implemented, it is called BiLSTM-CNN and BiGRU-CNN. These two methods are connecting BiLSTM or BiGRU with CNN or in other words the output of BiLSTM or BiGRU is fed to the CNN as input [23]. The training configuration for the RNN, CNN, and its combination is similar, that is the maximum length of tokens is 150 and batch size is 64. For the RNNs, the number of hidden layers is two, each hidden layer has a hidden unit of 256, and probability for dropout layer is 0.5. CNN has a number of filter 100 and filter sizes of 2, 3, 4, 5. All these methods are trained at maximum 20 epochs where weight with the lowest validation loss is stored to be used for evaluation. The detailed configuration can be seen in Table 2.

Tabel 2. Deep learning model configurations

Deep Learning Model	Hyperparameter	Values
All	Token Length	150
	Batch Size	64
	Maximum Epoch	20
	Embedding	FastText [24] with dim. size: 300
LSTM and GRU	Hidden unit	2 layers × 256
	Dropout	0.5
BiLSTM and BiGRU	Hidden unit	2 directions × 2 layers × 256
	Dropout	0.5
CNN	Number of filters	100
	Filter sizes	2, 3, 4, 5
BiLSTM-CNN and BiGRU-CNN	Hidden unit	2 directions × 2 layers × 256
	Dropout	0.5
	Number of filters	100
	Filter sizes	2, 3, 4, 5

The last deep learning method that is carried out in this research is IndoBERT. The IndoBERT model is inspired by Bidirectional Encoder Representations from Transformer (BERT) [25, 26] which is based on Transformer. Transformer utilizes attention mechanisms to be able to understand the context of input text without recurrent mechanisms which solve the RNN issues such as parallel training [27]. There are two popular IndoBERT models that are fine-tuned in this research. IndoBERT^{B} which is pre-trained with 4 billion words from Indonesian text source [26] and IndoBERT^{F} which is pre-trained with 220 million words from Indonesian Wikipedia, web corpus and news articles [25]. Both IndoBERT models in this research have 12 transformer layers with dimension of 768. There are multiple experiments for the fine-tuning process by freezing certain Transformer layers of the model. Freezing the layer means that the layer is not being fine-tuned or left as it is. It may reduce the fine-tuning duration but keeps the model performance [28]. The layers that are frozen in this experiment is the first 2, 4, 6, 8 and 10 layers and is indicated by FnL label following the IndoBERT, where n is the first n layer to be frozen i.e. IndoBERT F4L means to freeze the first 4 layers. The detailed configuration can be seen in Table 3.

Tabel 3. IndoBERT model configurations

IndoBERT Model	Frozen Layer(s)
IndoBERT^B	No frozen layer
IndoBERT^B F2L	First 2 layers
IndoBERT^B F4L	First 4 layers
IndoBERT^B F6L	First 6 layers
IndoBERT^B F8L	First 8 layers
IndoBERT^B F10L	First 10 layers
IndoBERT^F	No frozen layer
IndoBERT^F F2L	First 2 layers
IndoBERT^F F4L	First 4 layers
IndoBERT^F F6L	First 6 layers
IndoBERT^F F8L	First 8 layers
IndoBERT^F F10L	First 10 layers

All the deep learning training and fine-tuning process is done using Pytorch library and runs on Nvidia L4 with 24GB VRAM while SVM implementation uses Scikit-learn library and runs on CPU.

2.4. Model Evaluation

The best weight of each model from the training process is loaded and the test dataset is fed to calculate the metric scores so that each model performance can be compared. The evaluation uses four metrics: accuracy, macro-precision, macro-recall, and macro-F1. Then, one model is selected to be analyzed based on its outcome on the test dataset.

3. RESULTS AND DISCUSSION

3.1 Model Performance

There are a total of 20 models being created for comparison purposes including the baseline. The performance metric for each model can be seen in Table 4. The SVM with TF-IDF as baseline shows a good performance and generally has only slightly lower scores than others. The RNN family and CNN models performance are slightly better generally than the baseline and their performances are similar in terms of accuracy except for LSTM. Within this category, BiGRU-CNN, BiLSTM, and BiLSTM-CNN show the best metric performance across accuracy, precision, recall and F1. This result is reasonable since with the bidirectional variant the model can learn the text context from left to right and right to left which improves the model comprehension. The combination with CNN helps the model further to understand the relation of the context within the neighboring vector based on the filter size.

IndoBERT^{B} model category shows slightly better model performance than the baseline and generally somewhat similar score than RNN and CNN model category. It can be seen that freezing the layer increases model performance in terms of accuracy and precision though insignificant. On the other hand, IndoBERT^{F} model category performance generally shows a minor increase than baseline but worse than the IndoBERT^{B} model category. This can be caused by the fact that the IndoBERT^{F} model is pre-trained with a significantly smaller amount of corpus which makes the model lack information of certain tokens and its context. Still, freezing the layer evidently increases the performance on certain metrics.

Tabel 4. Model performance on test dataset

Model	Accuracy	Precision	Recall	F1
SVM	0.854	0.849	0.840	0.844
LSTM	0.859	0.852	0.829	0.836
GRU	0.870	0.862	0.863	0.862
BiLSTM	0.873	**0.884**	0.844	0.857
BiGRU	0.871	0.865	0.854	0.859
CNN	0.876	0.873	0.869	0.871
BiLSTM-CNN	0.870	0.871	0.876	0.871
BiGRU-CNN	0.878	0.875	0.866	0.870
IndoBERT^{B}	0.879	0.876	**0.888**	**0.879**
IndoBERT^{B} F2L	**0.882**	0.876	0.880	0.878
IndoBERT^{B} F4L	0.875	0.868	0.878	0.872
IndoBERT^{B} F6L	0.881	0.875	0.869	0.871
IndoBERT^{B} F8L	0.880	0.878	0.862	0.869
IndoBERT^{B} F10L	0.874	0.872	0.856	0.863
IndoBERT^{F}	0.868	0.856	0.865	0.859
IndoBERT^{F} F2L	0.866	0.851	0.864	0.857
IndoBERT^{F} F4L	0.868	0.856	0.863	0.859
IndoBERT^{F} F6L	0.873	0.865	0.860	0.862
IndoBERT^{F} F8L	0.862	0.850	0.855	0.852
IndoBERT^{F} F10L	0.849	0.837	0.847	0.842

Even though most models only have small increases over the baseline, it is still an improvement, nonetheless. Overall, across all models, models with the highest accuracy, precision, recall, and F1 are IndoBERT^{B} F2L, BiLSTM, IndoBERT^{B} and IndoBERT^{B} respectively. This is inline with previous research [29-31] that uses the IndoBERT model for text classification and shows that it outperforms other deep learning models. Therefore, IndoBERT^{B} without frozen layer is used for test dataset text analysis.

Tabel 5. IndoBERT^{B} confusion matrix on test dataset

	CoL	DD	HHD	LRI	Stroke	TB
CoL	336	11	3	3	0	4

	CoL	DD	HHD	LRI	Stroke	TB
DD	0	665	4	0	1	3
HHD	3	6	425	8	29	6
LRI	0	10	5	339	0	67
Stroke	2	5	25	6	305	2
TB	9	15	4	192	1	1008

Tabel 5 shows the confusion matrix of the IndoBERT^{B} model. The most misclassified diseases are Tuberculosis and LRI. Tuberculosis disease is frequently misclassified as LRI disease and vice versa. Looking at a few wrongly classified samples in Table 6, misclassification happened because the question description for both labels are similar, and this happened for most of the misclassified samples in test dataset.

Tabel 6. Examples of misclassified test dataset

Original Text	Actual	Prediction
Halo Dok, Dok Saya Sudah 1 minggu Merasakan Sakit di Bagian dada Dan Kepala Saya Sering Pusing Dan Saya Sering Kalau Ludah Suka Keluar Darah Dok Penyakit Apa Yang Saya Alami Dok?? Mohon infonya Dok. (Hello doc, it's been a week I've felt hurt on my chest and my head often dizzy and often when I spit it contains blood, doc. What kind of disease am I having, doc? Please provide information doc.)	TB	LRI
Maaf dok saya ingin bertanya, saya masih berusia 18 th. Saya sudah mengalami banyak sakit di tubuh ini. Mulai dari berkunang" setelah itu kepala terasa berat, dada sesak yg mengakibatkan suara menjadi pendek, dan yg agak parahnya setelah bangun tidur sering mengeluarkan lendir yg cukup banyak dan mengandung sedikit darah kotor kehitaman. Jujur saya belum pernah cek" ke dokter karena malas namun saya ingin tahu kemungkinan sakit apa yg saya alami skrg ini? Terimakasih sebelumnya (Sorry doc, I want to ask, I'm 18 years old. I have experienced a lot of pain in my body. Starting from dizzy, then my head feels heavy, chest tightness which causes my voice to become short, and the worst thing is after waking up I often produce a lot of phlegm and contain a little bit of blood. Honestly, I've never checked it to a doctor because I'm lazy but I want to know what possible disease that I'm having currently? Thanks in advance.)	TB	LRI
Dok...sy baru berhenti merokok kurang lebih sdh 3 bulan, tapi akhir-akhir ini sy mengalami batuk berdahak kadang ada darahnya kenapa ya ? Terus obatnya apa ?	LRI	TB

Original Text	Actual	Prediction
(Doc, I just stopped smoking for around 3 months, but lately I have been coughing with sputum and sometimes it contains blood. Why is it? Then, what is the medicine?)		
Hai dok saya kahfi umur 23 tahun saya mengidap batuk yg cukup lama sekitar 1 bulan namun saya negatif TB dok dan sekarang batuk saya membuat saya kalau batuk keluar darah yang hanya terjadi setiap pagi setiap bangun tidur dan sekarang saya merasakan gejala radang dok. Dimohon jawabannya dok (Hi doc, I'm Kahfi, 23 years old. I have had a cough for a pretting long time, around 1 month but I'm a TB negative and now my cough produces blood which happened only in the morning after waking up from sleep and now I'm feeling an inflammation symptom doc. Please give me the answer doc.)	LRI	TB

Tuberculosis and Lower Respiratory Infection disease such as Pneumonia and Bronchitis have similar symptoms such as cough, producing sputum with blood, and fever [32, 33]. Therefore, it is reasonable that the model unable to differentiate between TB and LRI since the question description usually contains the symptoms that experienced by the user.

Considering the results above, IndoBERT^{B} as the best model has a big potential to be integrated with healthcare platforms such as the health online forums or telemedicine applications. It can streamline the process of obtaining medical advice and reducing the response time for the users. Additionally, in the future, the model can be used in a public health monitoring system by analyzing trends about the type of questions asked to identify emerging health concerns.

3.2 Discussion

The results from the evaluation of the 20 models demonstrate that while many of the models show incremental improvements over the baseline, the IndoBERTB model family, particularly the variant with two frozen layers (IndoBERTB F2L), consistently outperforms other models across multiple metrics. This outcome aligns with prior research that highlights the efficacy of the IndoBERT model in text classification tasks within the Indonesian language context [29-31]. The success of the IndoBERTB models can be attributed to their ability to capture rich contextual information from Indonesian text, which is critical for accurately categorizing complex health-related inquiries.

In contrast, the RNN and CNN models, including their bidirectional and hybrid variants like BiLSTM-CNN and BiGRU-CNN, also exhibit strong performance but do not surpass the accuracy and precision of the IndoBERTB models. The performance of these models can be linked to their architecture, where the bidirectional approach enhances the model's understanding by processing the input text in both directions. Additionally, the integration of CNN layers aids in capturing local patterns within the text, further boosting performance. However, despite these advantages, they fall slightly short compared to the IndoBERTB variants, which likely benefit from their pre-training on a large corpus of Indonesian text.

Interestingly, while the IndoBERTF models also demonstrate improvements over the baseline, they generally underperform compared to the IndoBERTB models. This discrepancy can be explained by the smaller corpus used in pre-training the IndoBERTF models, which may limit their ability to fully grasp the nuances of certain tokens and contexts within the health-related questions. The results suggest that while freezing certain layers can lead to performance gains, the quality and size of the pre-training corpus play a more significant role in achieving superior performance.

The confusion matrix analysis of the IndoBERTB model reveals specific challenges in distinguishing between Tuberculosis (TB) and Lower Respiratory Infections (LRI), such as Pneumonia and Bronchitis. The frequent misclassification between these categories can be attributed to the overlapping symptoms described in the questions, such as coughing, sputum with blood, and fever. These shared symptoms complicate the model's ability to differentiate between the two diseases accurately. A deeper text analysis of the misclassified examples indicates that the similarity in symptoms presented by users is the primary cause of the errors, highlighting the need for more nuanced model training or the incorporation of additional contextual information to improve classification accuracy.

Given the overall performance of the IndoBERTB model, it shows great potential for integration into healthcare platforms like online health forums or telemedicine applications. By automating the classification of health-related questions, the model could significantly reduce response times, thereby enhancing the user experience and ensuring timely medical advice. Additionally, the model's capabilities could be extended in the future to public health monitoring systems, where it could analyze trends in the types of questions being asked to identify and respond to emerging health concerns more proactively.

4. CONCLUSION

This research objective is to perform health questions classification from Alodokter's question and answer page using multiple deep learning methods and a baseline. Majority of the deep learning models can outperform SVM as baseline even though insignificant. The BiGRU-CNN, BiLSTM-CNN, and BiLSTM are some of the most stand out models in the RNN and CNN model category. IndoBERT^{B} has shown better performance than IndoBERT^{F} and freezing the layers of IndoBERT models while fine-tuning shows slight improvement in the metric score. Overall, IndoBERT^{B} is the best model in the experiment even though it is still having difficulties to differentiate two types of disease which have similar symptoms in the question description. Future research can improve the model performance by combining the text from question-and-answer description to give the model better context in training and making predictions.

REFERENCES

[1] Ministry of Health of the Republic of Indonesia, "Indonesia Health Profile 2019," Jakarta: Ministry of Health of the Republic of Indonesia, 2020.

[2] Y. A. Singgalen, "Sentiment Analysis on Customer Perception towards Products and Services of Restaurant in Labuan Bajo," *J. Inf. Syst. Inform.*, vol. 4, no. 3, pp. 511-523, 2022.

[3] P. R. A. Savitri, I. M. A. D. Suarjaya, and W. O. Vihikan, "Sentiment Analysis of X (Twitter) Comments on The Influence of South Korean Culture in Indonesia," *J. Inf. Syst. Inform.*, vol. 6, no. 2, pp. 979-991, 2024.

[4] P. A. Setiawati, I. M. A. D. Suarjaya, and I. N. P. Trisna, "Sentiment Analysis of Unemployment in Indonesia During and Post COVID-19 on X (Twitter) Using Naïve Bayes and Support Vector Machine," *J. Inf. Syst. Inform.*, vol. 6, no. 2, pp. 662-675, 2024.

[5] N. Limsopatham, "Effectively leveraging BERT for legal document classification," in *Proc. Nat. Legal Lang. Process. Workshop 2021*, 2021, pp. 210-216.

[6] W. O. Vihikan, M. Mistica, I. Levy, A. Christie, and T. Baldwin, "Automatic resolution of domain name disputes," in *Proc. Nat. Legal Lang. Process. Workshop 2021*, 2021, pp. 228-238.

[7] X. Li, M. Cui, J. Li, R. Bai, Z. Lu, and U. Aickelin, "A hybrid medical text classification framework: Integrating attentive rule construction and neural network," *Neurocomputing*, vol. 443, pp. 345-355, 2021.

[8] S. K. Prabhakar and D.-O. Won, "Medical text classification using hybrid deep learning models with multihead attention," *Comput. Intell. Neurosci.*, vol. 2021, no. 1, p. 9425655, 2021.

[9] N. Arif, S. Latif, and R. Latif, "Question Classification Using Universal Sentence Encoder and Deep Contextualized Transformer," in *Proc. 2021 14th Int. Conf. Develop. eSyst. Eng. (DeSE)*, 2021, pp. 206-211.

[10] D. Han, T. Tohti, and A. Hamdulla, "Attention-based transformer-BiGRU for question classification," *Information*, vol. 13, no. 5, p. 214, 2022.

[11] A. F. Abdillah, P. Putra, C. Bagus, S. Juanita, and D. Purwitasari, "Ensemble-based Methods for Multi-label Classification on Biomedical Question-Answer Data," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 8, no. 1, 2022.

[12] N. A. Salsabila, Y. A. Winatmoko, A. A. Septiandri, and A. Jamal, "Colloquial Indonesian Lexicon," in *Proc. 2018 Int. Conf. Asian Lang. Process. (IALP)*, 2018, pp. 226-229.

[13] J. Asian, *Effective techniques for Indonesian text retrieval*, Melbourne, Australia: RMIT University, 2007.

[14] A. Z. Arifin, I. Mahendra, and H. T. Ciptaningtyas, "Enhanced confix stripping stemmer and ants algorithm for classifying news document in Indonesian language," in *Proc. Int. Conf. Inf. Commun. Technol. Syst.*, 2009, vol. 5, pp. 149-158.

[15] A. D. Tahitoe and D. Purwitasari, "Implementasi modifikasi enhanced confix stripping stemmer untuk bahasa indonesia dengan metode corpus based stemming," *J. Ilm.*, vol. 12, no. 15, pp. 1-15, 2010.

[16] A. K. Darmawan, M. W. Al Wajieh, M. B. Setyawan, T. Yandi, and H. Hoiriyah, "Hoax news analysis for the Indonesian national capital relocation public policy with the support vector machine and random forest algorithms," *J. Inf. Syst. Inform.*, vol. 5, no. 1, pp. 150-173, 2023.

[17] M. Zulqarnain, A. K. Z. Alsaedi, R. Ghazali, M. G. Ghouse, W. Sharif, and N. A. Husaini, "A comparative analysis on question classification task based on deep learning approaches," *PeerJ Comput. Sci.*, vol. 7, p. e570, 2021.

[18] Y. Zhang and Z. Rao, "n-BiLSTM: BiLSTM with n-gram Features for Text Classification," in *Proc. 2020 IEEE 5th Inf. Technol. Mechatronics Eng. Conf. (ITOEC)*, 2020, pp. 1056-1059.

[19] A. A. Sharfuddin, M. N. Tihami, and M. S. Islam, "A deep recurrent neural network with bilstm model for sentiment classification," in *Proc. 2018 Int. Conf. Bangla Speech Lang. Process. (ICBSLP)*, 2018, pp. 1-4.

[20] L. Zhou and X. Bian, "Improved text sentiment classification method based on BiGRU-Attention," *J. Phys.: Conf. Ser.*, vol. 1345, no. 3, p. 032097, 2019.

[21] H. Wang, J. He, X. Zhang, and S. Liu, "A short text classification method based on N-gram and CNN," *Chin. J. Electron.*, vol. 29, no. 2, pp. 248-254, 2020.

[22] E. D. Ajik, G. N. Obunadike, and F. O. Echobu, "Fake News Detection Using Optimized CNN and LSTM Techniques," *J. Inf. Syst. Inform.*, vol. 5, no. 3, pp. 1044-1057, 2023.

[23] J. Zheng and L. Zheng, "A hybrid bidirectional recurrent convolutional neural network attention-based model for text classification," *IEEE Access*, vol. 7, pp. 106673-106685, 2019.

[24] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *Trans. Assoc. Comput. Linguistics*, vol. 5, p. 135, 2017.

[25] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: a benchmark dataset and pre-trained language model for Indonesian NLP," in *Proc. COLING 2020-28th Int. Conf. Comput. Linguistics*, 2020, pp. 757-770.

[26] B. Wilie et al., "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *Proc. 1st Conf. Asia-Pacific Chapter Assoc. Comput. Linguistics 10th Int. Joint Conf. Natural Lang. Process.*, 2020, pp. 843-857.

[27] A. Vaswani et al., "Attention is All You Need," presented at the *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017.

[28] A. Merchant, E. Rahimtoroghi, E. Pavlick, and I. Tenney, "What Happens To BERT Embeddings During Fine-tuning?," in *Proc. Third BlackboxNLP Workshop Analyzing Interpreting Neural Netw. NLP*, 2020, pp. 33-44.

[29] I. Budiman et al., "Classification Performance Comparison of BERT and IndoBERT on Self-Report of COVID-19 Status on Social Media," *J. Comput. Sci. Inst.*, vol. 30, pp. 61-67, 2024.

[30] S. Saadah, K. M. Auditama, A. A. Fattahila, F. I. Amorokhman, A. Aditsania, and A. A. Rohmawati, "Implementation of BERT, IndoBERT, and CNN-LSTM in Classifying Public Opinion About COVID-19 Vaccine in Indonesia," *J. RESTI (Rekayasa Sist. dan Teknol. Inform.)*, vol. 6, no. 4, pp. 648-655, 2022.

[31] M. I. K. Sinapoy, Y. Sibaroni, and S. S. Prasetyowati, "Comparison of LSTM and IndoBERT Method in Identifying Hoax On Twitter," *J. RESTI (Rekayasa Sist. dan Teknol. Inform.)*, vol. 7, no. 3, pp. 657-662, 2023.

[32] P. F. Wright and F. L. Marston, "The Detection of Respiratory Infections," *N. Engl. J. Med.*, vol. 282, no. 4, pp. 203-209, 1970.

[33] P. J. Barnes, "Mechanisms of Development of Multidrug-Resistant Tuberculosis," *Clin. Chest Med.*, vol. 30, no. 4, pp. 521-530, 2009.