Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach

Helmi Imaduddin[1], Fiddin Yusfida A'la[2], Yusuf Sulistyo Nugroho[3]
Department of Informatics, Universitas Muhammadiyah Surakarta, Indonesia, Surakarta, Indonesia[1, 3]
Department of Informatics Engineering, Universitas Sebelas Maret, Surakarta, Indonesia[2]

***Abstract*—The rapid growth of application development has made applications an integral part of people's lives, offering solutions to societal problems. Health service applications have gained popularity due to their convenience in accessing information on diseases, health, and medicine. However, many of these applications disappoint users with limited features, slow response times, and usability challenges. Therefore, this research focuses on developing a sentiment analysis system to assess user satisfaction with health service applications. The study aims to create a sentiment analysis model using reviews from health service applications on the Google Play Store, including Halodoc, Alodokter, and klikdokter. The dataset comprises 9.310 reviews, with 4.950 positive and 4.360 negative reviews. The IndoBERT pre-training method, a transfer learning model, is employed for sentiment analysis, leveraging its superior context representation. The study achieves impressive results with an accuracy score of 96%, precision of 95%, recall of 96%, and an F1-score of 95%. These findings underscore the significance of sentiment analysis in evaluating user satisfaction with health service applications. By utilizing the IndoBERT pre-training method, this research provides valuable insights into the strengths and weaknesses of health service applications on the Google Play Store, contributing to the enhancement of user experiences.**

***Keywords*—Application; healthcare; IndoBERT; sentiment analysis**

I. INTRODUCTION

In the present age of digital advancements, a myriad of applications has emerged to cater to diverse facets of human life, spanning across desktops, tablets, and smartphones. This surge in demand has created lucrative business prospects, resulting in the proliferation of mobile applications aimed at resolving everyday challenges [1]. Regrettably, not all applications boast commendable features and functionalities, including the healthcare applications readily available on the Google Play Store. Consequently, there arises a pressing need for a system capable of comprehensively analyzing application reviews to enhance their overall performance. While opinions and ratings serve as primary means for gathering feedback on an app's usability, ratings alone may not consistently provide reliable insights [2]. Furthermore, ratings fail to offer a comprehensive understanding to improve the user experience aspect. Thus, the examination of customer reviews becomes crucial for gaining deeper insights and understanding [3]. User experience entails the intricate narrative surrounding a user's interaction with the app, while opinions delve into their underlying thoughts and emotions. Users possess the freedom to express their evaluations in various textual forms, resulting in a less structured review dataset, which in turn poses greater challenges in handling and analysis [4].

Sentiment analysis, commonly referred to as opinion mining, is a technique that aims to classify user sentiment based on polarity [5]. It encompasses a wide array of objectives, methodologies, and types of analytics. In the domain of sentiment analysis, three main methodologies are employed: machine learning (ML), hybrid learning, and lexicon-based approaches [6]. Among these, supervised learning emerges as the most popular and widely utilized ML approach. This methodology involves training the model using labeled data to predict outputs, while also incorporating additional unlabeled inputs for enhanced performance [7].

In the context of sentiment analysis, it is important to acknowledge that certain languages, such as English and Chinese, benefit from being considered high-resource languages, as they have readily available datasets accessible to the academic community. However, the majority of languages face challenges due to limited data collection and a lack of published research, including Indonesian [8]. Previous research on sentiment analysis of Indonesian text has explored the efficacy of machine learning models like Support Vector Machine (SVM) and Naïve Bayes, demonstrating their effectiveness in addressing this issue [9]. Nevertheless, the integration of pre-training using language models has emerged as a promising approach across various natural language processing tasks [10], [11]. One significant drawback of conventional language models is their unidirectional nature, which imposes limitations on the available architecture for pre-training. To overcome this limitation, a novel technique called Bidirectional Encoder Representations from Transformers (BERT) has been proposed to enhance the fine-tuning-based approach [12].

A number of previous studies have explored sentiment analysis in the context of Indonesian language, employing various approaches ranging from traditional machine learning classifiers to deep learning-based algorithms such as IndoBERT. For instance, sentiment analysis using random forest algorithms demonstrated promising results, achieving an average out-of-bag (OOB) score of 0.829 [13]. Similarly, research focused on emoticons and emoticon categories utilized classification-based machine learning algorithms like naïve Bayes and support vector machines [14]. Sarcasm data classification was also conducted using random forest classifiers, naïve Bayes, and support vector machines [14]. Furthermore, Word2Vec was employed as an alternative to hand-crafted features for sentiment analysis of hotel reviews in

Indonesian, with the conclusion that optimal accuracy can be achieved by simultaneously increasing vector dimensions and the amount of data [15]. IndoBERT, which outperforms both multi-Indonesian lingual BERT and Bert, has demonstrated superior data processing capabilities. The research involved data collection, data preprocessing, and fine-tuning of IndoBERT. Hoax detection classification was completed using pre-trained BERT models, with multilingual BERT for general purposes and IndoBERT specifically tailored for Indonesian. The fine-tuned IndoBERT model, trained on an Indonesian monolingual corpus, exhibited enhanced performance compared to the original BERT and improved multilingualism [16].

However, performing this analysis manually is quite difficult; hence we propose performing Indonesian language analysis using the IndoBERT algorithm, which was built exclusively to evaluate Indonesian language material. The primary objective of this research is to develop a sentiment analysis system for healthcare application reviews on the Google Play Store using the IndoBERT approach. Additionally, the system aims to assist users in selecting health service applications that offer optimal functionality and facilities. The proposed methodology involves leveraging IndoBERT as a pre-training model, renowned for its effectiveness in processing Indonesian language data. The data utilized in this research is sourced from the Google Play Store, making it a novel and previously unexplored area of investigation.

II. METHODOLOGY

The research conducted in this study encompasses several stages. Firstly, review data was collected by scraping the Google Play website, followed by manual labeling of the obtained data. The labeled data was then preprocessed to ensure cleanliness and suitability for classification. The dataset was subsequently divided into three parts: training data, validation data, and testing data, with a distribution ratio of 70:10:20. The next step involved creating a classification model using IndoBERT, and adjusting hyperparameters to optimize its performance. Lastly, the model was evaluated using the testing data, employing various parameters such as accuracy, precision, recall, and F1-score. For a comprehensive overview of the research design, please refer to Fig. 1.

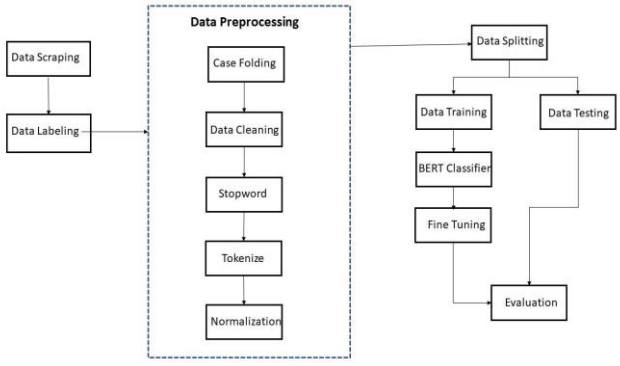

Fig. 1. Research design.

A. Data Scraping and Labeling

The dataset utilized in this research was obtained from the Google Play website. Data collection was performed using scraping techniques, employing the Python programming language and the Google-play-scraper library. The authors specifically collected review data from healthcare applications such as Alodokter, Halodoc, and Klikdokter, amassing a total of 9.310 user reviews in September 2022. To avoid the biases data, we removed any personally identifiable information (PII) from the reviews.

As the data was initially unlabeled, a manual data labeling process was conducted to facilitate the subsequent classification task; it is a very important process because the deep learning model will learn from the pattern of the given dataset.

The dataset was divided into two classes: positive and negative class. Following the completion of the labeling process, the dataset comprised 4.950 positive reviews and 4.360 negative reviews, so the dataset used is quite balanced. Detailed information regarding the datasets used in this study can be found in Table I.

TABLE I. DATASET

No	Class	Data
1	Positive	**4.950**
2	Negative	**4.360**
	Total	**9.310**

B. Preprocessing Data

Preprocessing is a crucial step in converting raw data into a format suitable for classification input [17]. This process involves five stages, namely case folding, data cleaning, stopword removal, tokenization, and normalization [18]. Case folding denotes a textual transformation operation that converts all capital letters within a string to a lowercase, to render the comparison and processing of text more consistent. Data cleansing constitutes a critical preprocessing step for natural language processing (NLP) pipelines. NLP involves manipulating and analyzing human language; hence the quality of the input data can substantially influence the performance and efficacy of NLP systems. This process aims to prepare the classification input by eliminating unwanted elements. Various actions are performed during data cleaning, such as removing unique characters, usernames, hashtags, punctuation, emojis, and excessive spaces, resulting in a dataset containing only words.

The next phase is known as the stopword process and entails removing common words that appear frequently but have no significant meaning. In computational linguistics and textual data analysis pipelines, function words considered uninformative are frequently used. These terms, known as stopwords, are eliminated prior to subsequent processes because they contribute marginally to the semantic content. Typically, they contain high-frequency words such as "the," "is," "and," "a," "an," "in," "of," etc. The precise stopword lexicon varies based on the objective of natural language processing and the examined language. The rationale behind the eradication of stopwords is the reduction of the dimensionality of textual data, which can accelerate processing

and improve the efficacy of specific natural language processing techniques, such as text categorization and information extraction. By removing these common terms, the emphasis transfers to more informative content words that can provide more meaningful distinctions. To accomplish this, a stop-list dictionary containing these words is compiled. Therefore, data containing stop list terms are removed to improve sentiment analysis performance.

Following this, the tokenization process is applied to convert each data entry into individual tokens, where each token typically corresponds to a single word [19], [20]. Lastly, data normalization is performed to address the presence of non-standard words commonly found in Google Play reviews. This conversion of non-standard words into standard ones is essential for improving classification performance

C. Data Splitting

After the data enters preprocessing, it will be divided into three parts: training data, validation data, and testing data [21]. Each part serves a different function: data training is used to create models, data validation reduces overfitting, and data testing evaluates the created models. The proportion of data sharing is 70% for the training set, 20% for the validation set, and 10% for the test set.

D. BERT Fine Tuning

BERT is a pre-training model that has undergone extensive training on a vast amount of data. The process of creating a BERT model involves two key steps: pre-training and fine-tuning. During pre-training, the model is trained on unlabeled data using various pre-training tasks. Subsequently, the BERT model is fine-tuned using labeled data from downstream tasks, starting with the pre-trained parameters. Despite commencing with the same pre-trained parameters, each downstream task results in a well-tuned model [12]. Fig. 2 illustrates the fine-tuning process on the pre-trained BERT model. The versatility of the BERT technique has been demonstrated through various studies aimed at addressing research gaps. For instance, researchers have successfully improved accuracy with transformer-based models when dealing with large, complex documents [22]. Additionally, BERT-based text classification has been enhanced by incorporating additional sentences and domain knowledge [23]. Notably, the impact of these improvements has been particularly evident in high-resource languages like English.

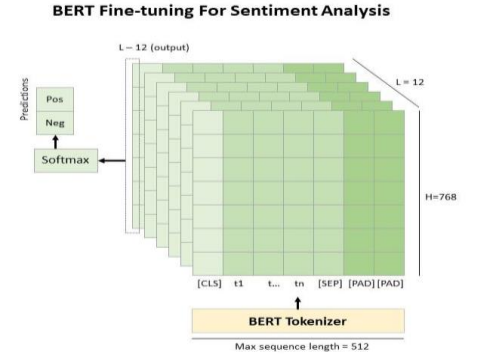

Fig. 2. BERT fine tuning.

To train pre-training models effectively, it is crucial to use specific languages. Since our dataset consists of user reviews in Indonesian, it necessitates pre-training models tailored for the Indonesian language. As a result, the BERT-based model has undergone significant enhancements, leading to the development of IndoBERT [24]. This improved version is built upon the Indonesian vocabulary, achieved by modifying the Huggingface framework. IndoBERT has been meticulously trained on an extensive dataset comprising over 220 million words, sourced from various Indonesian platforms, including Indonesian Wikipedia, news articles from Kompas, Tempo, Liputan6, and Korpus Web Indonesia. The training process involved running IndoBERT through 2.4 million steps or 180 epochs, taking approximately two months to complete. The positive attributes of IndoBERT motivated us to utilize this model for classifying Indonesian app reviews. For this study, we specifically employed "IndoBERT-base-p1," which represents one variant of the IndoBERT model [8].

E. Evaluation

Evaluation is a technique used to determine a model's classification aptitude. This study's model evaluation employs a confusion matrix that generates true positive (TP), false positive (FP), false negative (FN), and true negative (TN) values. Multiple metrics, including accuracy, sensitivity, specificity, and precision, as well as the F1-Score, are employed to evaluate the implemented model. The accuracy formula has been shown in equation (1), the recall formula has been shown in equation (2), the precision formula has been shown in equation (3), and the F1-Score formula has been shown in equation (4).

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

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$F1-Score = \frac{2\ x\ (Precision\ x\ recall)}{recall+precision} \quad (4)$$

III. Result and Discussion

This research utilizes IndoBERT transfer learning, a technique that leverages a pre-trained model to address new problems of a similar nature. In this study, we employ IndoBERT as the pre-trained model, which stands for Indonesia Bidirectional Encoder Representations from Transformers, built using the PyTorch framework. IndoBERT is a transformers-based model, derived from Bert Base with 12 hidden layers, tailored specifically for monolingual Indonesian language tasks [25].

The investigation utilized a dataset comprising 9.310 samples, which were categorized into three subsets: training, validation, and test data. Fig. 3 illustrates the data labeling, with 4.950 samples carrying a positive label and 4.360 samples carrying a negative label. It is evident that positive or "good" reviews constituted 53.2% of the data, while negative reviews accounted for 46.8%.

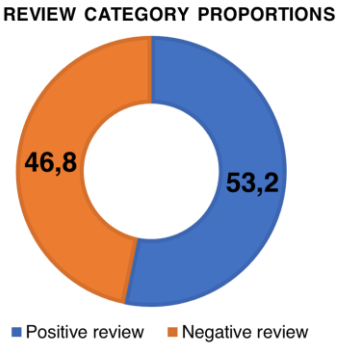

Fig. 3. Review category proportions.

The author employed 10 epochs for the training process. Observing the results, it becomes evident that utilizing 10 epochs yields commendable accuracy, depicted by a noticeable upward trend in the curve. Additionally, this study utilized a learning rate of 1e-6, as a parameter for Adam's optimizer. This parameter was chosen based on experimental induction. It is worth noting that the appropriate learning rate varies from case to case, as learning rates that are too large or too small can lead to suboptimal solutions. The learning rate typically ranges from 0 to 1, with higher values facilitating faster training but not necessarily guaranteeing more optimal results. Therefore, careful selection of the learning rate value is vital to achieve the best possible outcomes. Fig. 4 illustrates the training results against the dataset, showcasing the performance curve in relation to the chosen parameters.

According to Fig. 4, the curve exhibits a pronounced upward trend to the right, suggesting a well-trained model. Additionally, the model trained with the new dataset undergoes evaluation to determine its performance against the dataset. To assess the model's effectiveness, a confusion matrix is employed in this study. The evaluation of the model against the testing data is depicted in Fig. 5.

According to the observations from Fig. 3, the model demonstrates excellent predictive capabilities. Notably, the values for true positive and true negative are significantly higher than those for false positive and false negative. The study's results indicate an impressive accuracy score of 96%, an F1-Score of 95%, a Recall of 96%, and a Precision of 95%.

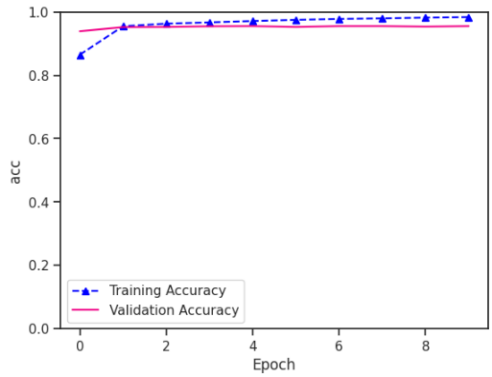

Fig. 4. Training history.

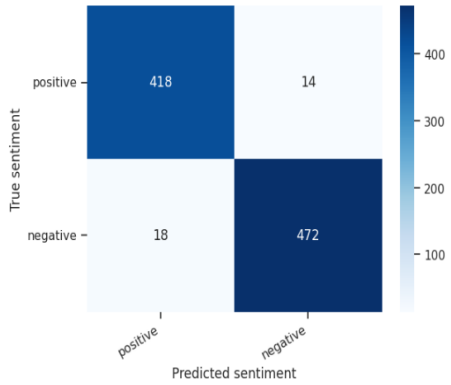

Fig. 5. Confusion matrix.

This study employs a different methodology than previous studies, such as Pandesenda et al., who conducted sentiment analysis on Alodokter data extracted from the Google Play Store in 2020 [26]. This procedure employs Fast Large-Margin, which yields an accuracy of 92.33%. Mehta et al., using Bidirectional LSTM, identify healthcare sentiment analysis from Twitter data with an accuracy of 80.88% in a separate study [27]. A comprehensive overview of these comparisons can be found in Table II.

TABLE II. COMPARISON WITH OTHER STUDIES

No	Researchers	Method	Accuracy
1	Pandesenda et al.,	Fast Large-Margin	92.33%
2	Mehta et al.,	Bidirectional LSTM	80.88%
3	**Our Study**	**IndoBERT-base-p1**	**96%**

The reasons for specific projections in the context of sentiment analysis are critical for various reasons. (1) Sentiment analysis provides interpretability, bridging the gap between the model's sophisticated computations and human perception of emotion. (2) Users and stakeholders have a right to know why certain decisions are being made, especially when those decisions impact their experiences or choices. (3) Domain experts can provide insights into why certain linguistic patterns may carry particular sentiment connotations in the given language or culture. The limitation of this research is that the model was trained using IndoBERT, which is specifically designed for Bahasa Indonesia content and has not been tested with other languages.

IV. CONCLUSION

In conclusion, this study focused on conducting sentiment analysis of Indonesian text using the transfer learning technique with the IndoBERT pre-trained model. The research was based on a dataset containing 9.310 reviews, each labeled as either positive or negative. During the training process, 10 epochs were used along with Adam's optimizer, employing a learning rate of 1e-6. The evaluation of the model yielded impressive results, with a high accuracy score of 96%, an F1-Score of 95%, a Recall of 96%, and a Precision of 95%. These findings underscore the effectiveness of transfer learning with IndoBERT as a robust approach for sentiment analysis of Indonesian text. If the dataset used increases, with reference to the current high accuracy value, there is a possibility that the performance will decrease but not significantly.

By contributing to the advancement of natural language processing research for the Indonesian language, this study holds significant value. The applications of this technique are diverse and can prove beneficial in areas like opinion mining, social media analysis, and market research. To further enhance the model's capabilities, future research may explore parameter optimization and evaluation with larger and more diverse datasets, thereby increasing its generalizability.

ACKNOWLEDGMENT

The author expresses gratitude to the Universitas Muhammadiyah Surakarta for providing research support, enabling the completion of this study. This research is fully funded by Riset Muhammadiyah (RisetMu).

REFERENCES

[1] R. Alturki and V. Gay, "Usability Attributes for Mobile Applications: A Systematic Review," 2019, pp. 53–62. doi: 10.1007/978-3-319-99966-1_5.

[2] [M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA: ACM, Aug. 2004, pp. 168–177. doi: 10.1145/1014052.1014073.

[3] K. S. Nugroho, A. Y. Sukmadewa, H. Wuswilahaken DW, F. A. Bachtiar, and N. Yudistira, "BERT Fine-Tuning for Sentiment Analysis on Indonesian Mobile Apps Reviews," in *6th International Conference on Sustainable Information Engineering and Technology 2021*, New York, NY, USA: ACM, Sep. 2021, pp. 258–264. doi: 10.1145/3479645.3479679.

[4] M. Hassenzahl, "Experience Design: Technology for All the Right Reasons," Morgan & Claypool Publishers.

[5] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008, doi: 10.1561/1500000011.

[6] A. Ligthart, C. Catal, and B. Tekinerdogan, "Systematic reviews in sentiment analysis: a tertiary study," *Artif Intell Rev*, vol. 54, no. 7, pp. 4997–5053, Oct. 2021, doi: 10.1007/s10462-021-09973-3.

[7] S. Sah, "Machine Learning: A Review of Learning Types," pp. 1–7, 2020.

[8] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," Sep. 2020.

[9] F. Y. A'la, "Indonesian Sentiment Analysis towards MyPertamina Application Reviews by Utilizing Machine Learning Algorithms," *Journal of Informatics Information System Software Engineering and Applications (INISTA)*, vol. 5, no. 1, pp. 80–91, 2022.

[10] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," pp. 1–12, 2018.

[11] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 328–339. doi: 10.18653/v1/P18-1031.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[13] M. A. Fauzi, "Random Forest Approach for Sentiment Analysis in Indonesian Language," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 12, no. 1, p. 46, Oct. 2018, doi: 10.11591/ijeecs.v12.i1.pp46-50.

[14] D. Alita, S. Priyanta, and N. Rokhman, "Analysis of Emoticon and Sarcasm Effect on Sentiment Analysis of Indonesian Language on Twitter," *Journal of Information Systems Engineering and Business Intelligence*, vol. 5, no. 2, p. 100, Oct. 2019, doi: 10.20473/jisebi.5.2.100-109.

[15] R. P. Nawangsari, R. Kusumaningrum, and A. Wibowo, "Word2Vec for Indonesian Sentiment Analysis towards Hotel Reviews: An Evaluation Study," *Procedia Comput Sci*, vol. 157, pp. 360–366, 2019, doi: 10.1016/j.procs.2019.08.178.

[16] L. H. Suadaa, I. Santoso, and A. T. B. Panjaitan, "Transfer Learning of Pre-trained Transformers for Covid-19 Hoax Detection in Indonesian Language," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 15, no. 3, p. 317, Jul. 2021, doi: 10.22146/ijccs.66205.

[17] M. L. L. Wijerathne, L. A. Melgar, M. Hori, T. Ichimura, and S. Tanaka, "HPC Enhanced Large Urban Area Evacuation Simulations with Vision based Autonomously Navigating Multi Agents," *Procedia Comput Sci*, vol. 18, pp. 1515–1524, 2013, doi: 10.1016/j.procs.2013.05.319.

[18] R. Kusumaningrum, I. Z. Nisa, R. P. Nawangsari, and A. Wibowo, "Sentiment analysis of Indonesian hotel reviews: from classical machine learning to deep learning," *International Journal of Advances in Intelligent Informatics*, vol. 7, no. 3, p. 292, Nov. 2021, doi: 10.26555/ijain.v7i3.737.

[19] N. Bahrawi, "Sentiment Analysis Using Random Forest Algorithm-Online Social Media Based," *Journal of Information Technology and Its Utilization*, vol. 2, no. 2, p. 29, Dec. 2019, doi: 10.30818/jitu.2.2.2695.

[20] F. Y. A'la, Hartatik, N. Firdaus, M. A. Safi'ie, and B. K. Riasti, "A Comprehensive Analysis of Twitter Data: A Case Study of Tourism in Indonesia," in *2022 1st International Conference on Smart Technology, Applied Informatics, and Engineering (APICS)*, IEEE, Aug. 2022, pp. 85–89. doi: 10.1109/APICS56469.2022.9918757.

[21] Merfat. M. Altawaier and S. Tiun, "Comparison of Machine Learning Approaches on Arabic Twitter Sentiment Analysis," *Int J Adv Sci Eng Inf Technol*, vol. 6, no. 6, p. 1067, Dec. 2016, doi: 10.18517/ijaseit.6.6.1456.

[22] C. Liao, T. Maniar, S. N, and A. Sharma, "Techniques to Improve Q&A Accuracy with Transformer-based models on Large Complex Documents," pp. 1–8, 2020.

[23] S. Yu, J. Su, and D. Luo, "Improving BERT-Based Text Classification With Auxiliary Sentence and Domain Knowledge," *IEEE Access*, vol. 7, pp. 176600–176612, 2019, doi: 10.1109/ACCESS.2019.2953990.

[24] 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 *Proceedings of the 28th International Conference on Computational Linguistics*, Stroudsburg, PA, USA: International Committee on Computational Linguistics, 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.

[25] S. L. Sariwening and Azhari, "IndoBERT: Transformer-based Model for Indonesian Language Understanding," in *Master Thesis*, Yogyakarta, 2020.

[26] I. Pandesenda, R. R. Yana, E. A. Sukma, A. Yahya, P. Widharto, and A. N. Hidayanto, "Sentiment Analysis of Service Quality of Online Healthcare Platform Using Fast Large-Margin," in 2020 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), IEEE, Nov. 2020, pp. 121–125. doi: 10.1109/ICIMCIS51567.2020.9354295.

[27] A. Mehta, S. Virkar, J. Khatri, R. Thakur, and A. Dalvi, "Artificial Intelligence Powered Chatbot for Mental Healthcare based on Sentiment Analysis," in 2022 5th International Conference on Advances in Science and Technology (ICAST), IEEE, Dec. 2022, pp. 185–189. doi: 10.1109/ICAST55766.2022.10039548.