JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 8, No. 3, September 2023, Pp. 774-780

SENTIMENT ANALYSIS OF INDONESIAN TIKTOK REVIEW USING LSTM AND INDOBERTWEET ALGORITHM

Jerry C. Setiawan*[1)], Kemas M. Lhaksmana[2)], Bunyamin[3)]

1. School of Computing, Telkom University, Bandung, Indonesia
2. School of Computing, Telkom University, Bandung, Indonesia
3. School of Computing, Telkom University, Bandung, Indonesia

Article Info

Keywords: Sentiment Analysis; LSTM; IndoBERTweet; NLP; TikTok

Article history:
Received 7 March 2023
Revised 21 March 2023
Accepted 4 April 2023
Available online 1 September 2023

DOI : https://doi.org/10.29100/jipi.v8i3.3911

* Corresponding author.
Jerry Cahyo Setiawan
E-mail address:
jerrycahyosetiawan@student.telkomuniversity.ac.id

ABSTRACT

TikTok is currently the most popular app in the world and thus gets many reviews on the Google Play Store and other app marketplace platforms. These reviews are valuable user opinions that can be analyzed further for many purposes. Harnessing valuable analyses from these reviews can be obtained manually, which will be time-consuming and costly, or automatically with machine learning methods. This paper implements the latter with LSTM and IndoBERTweet, a derivative of BERT, using Indonesian vocabulary from Twitter post data. This research aims to determine the appropriate method to create a model that can automatically classify TikTok reviews into negative, neutral, and positive sentiments. The result demonstrates that IndoBERTweet outperforms the other, with an accuracy of 80%, whereas the LSTM accuracy is at 78%.

I. INTRODUCTION

THE social media is currently not only a place to communicate and seek entertainment but also a place of business [1]. One of the most popular social media is TikTok. In 2018 TikTok users increased rapidly compared to other platforms like YouTube and Facebook. TikTok users increased by 109%, while Facebook was 11% and YouTube was 27% [2]. TikTok has hundreds of millions of active users who download applications from the Google Play Store.

This application's popularity raises various negative, neutral, and positive sentiments in application reviews on the Google Play Store. Huge user reviews make it difficult for readers to classify and conclude the results of the applications necessary to analyze TikTok user reviews. The sentiment analysis method is suitable to handle this case. Sentiment analysis is a process of text data on opinion sentences to obtain information on positive, negative, and neutral sentiment tendencies. Classification is also used in sentiment analysis because classification is one method of grouping information in text mining [3].

The data text used in this study is Indonesian sentences, with the results of this study to help TikTok users in Indonesia assess whether this application is appropriate for interacting and seeking entertainment. The purpose of this research is also to add insight to TikTok developers in the Indonesian region in developing applications based on opinions on user reviews on the Google Play Store.

In implementing deep learning, the methods used for sentiment analysis are Long Short-Term Memory (LSTM) and IndoBERTweet. Sentiment analysis aims to classify opinions in sentences into positive, negative, and neutral labels. The initial stage is to analyze the text data that has passed the pre-processing of the data [4]. Related research has been done using the LSTM method for multi-aspect sentiment analysis on Indonesian TikTok reviews. BERT word embedding was used in the research to improve the LSTM performance. The result of the research showed that the method could make a model with the best performance for sentiment analysis on the business, content, and feature aspect. The dataset used in previous research amounts to 10,000 with predetermined vocabulary depending on the aspect [5]. Focus on this research only used standard LSTM without integration from another method to compare with the IndoBERTweet as a pre-trained model.

The comparison method is IndoBERTweet, a variant of the BERT algorithm as a deep learning model in natural language processing (NLP) [6]. IndoBERTweet is a pre-trained model using Indonesian vocabulary in Twitter tweets [7]. Tweets on Twitter have the same characteristics as TikTok reviews, which are informal sentences on the Google Play Store. The similarity of data characteristics is the reason for using these methods in conducting sentiment analysis of the TikTok application reviews on the Google Play Store. This paper implements the LSTM and IndoBERTweet algorithms to compare the sentiment analysis model performance to conclude user opinions automatically.

II. Research Method

This research has been done in passed some stages. The processes start with data crawling, preprocessing, splitting data, making sentiment models, model testing, and sentiment analysis evaluations. Fig. 1 is a workflow diagram describing the steps from the beginning to the aim of the research.

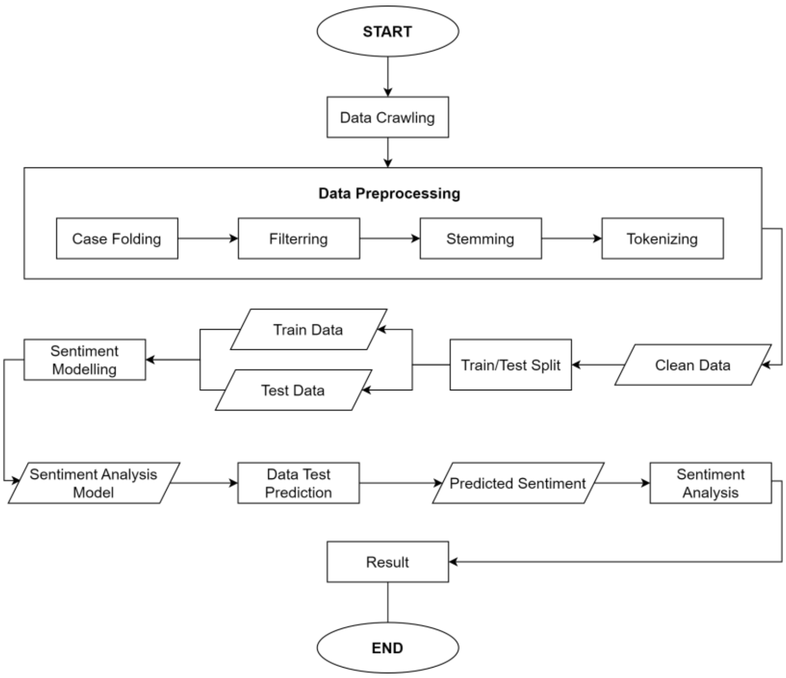


Fig. 1. Workflow of Indonesian TikTok review sentiment analysis using LSTM and IndoBERTweet

A. Dataset

This research uses data from Indonesian TikTok user reviews on Google Play Store, which are retrieved by the crawling method using the google-play-scrapper library. The data obtained from the crawling results is 50,000 review data. In this work, attribute selection is the first stage to creating a sentiment analysis model. An attribute in machine learning is known as a feature. The feature should be measurable and related to the observation [8].

We used the attribute content from the dataset as a feature to create sentiment analysis. The selected feature contains various sentiments from the Indonesian TikTok user feedback. It also has characteristics and noise, such as an emoticon, missing value, repeated character, link, and non-string variables. Those noises were solved by data text preprocessing with the natural language processing method.

B. Natural Language Processing

Natural language processing (NLP) is a branch of artificial intelligence that can help computers understand data text meaning like humans [8]. Natural language processing was used to preprocess data such as labeling, case folding, filtering, tokenizing, and stemming.

Labeling is a process of adding value by identifying each piece of data. It helps the model to learn data from the label [9]. In this work, a dataset containing text reviews from Indonesian TikTok users is classified into three labels: positive, neutral, and negative sentiment. The result of labeling shown in Table I.

TABLE I
LABELING RESULT

Data	Label
Bagus banget pokonya	Positive
Tidak bisa atur Volume suara	Negative
TikTok bales Dong, biar Emak aku bangga	Neutral

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 8, No. 3, September 2023, Pp. 774-780

The case folding stage was transforming labeled data containing uppercase into lowercase. This stage aims to make data only contain the letters 'a-z'. It removed other characters besides letters and was considered a delimiter. The result of case folding shown in Table II.

TABLE II
CASE FOLDING RESULT

Data	Result
Bagus banget pokonyaaa.	bagus banget pokonyaaa
Tidak bisa atur Volume suara	tidak bisa atur volume suara
TikTok bales Dong, biar Emak aku bangga	tiktok bales dong biar emak aku bangga

Filtering text has rules for removing noise data and using the stopword removal method. Stopword removal is a step to filter out words that contain little information and keep meaningful words [10]. In this work, we used a slang word dictionary from *colloquial-indonesian-lexicon* to be filtered slang words and used a stopword dictionary from dataset *stopwords-id-satya*. The result of filtering text, shown in Table III.

TABLE III
FILTERING TEXT RESULT

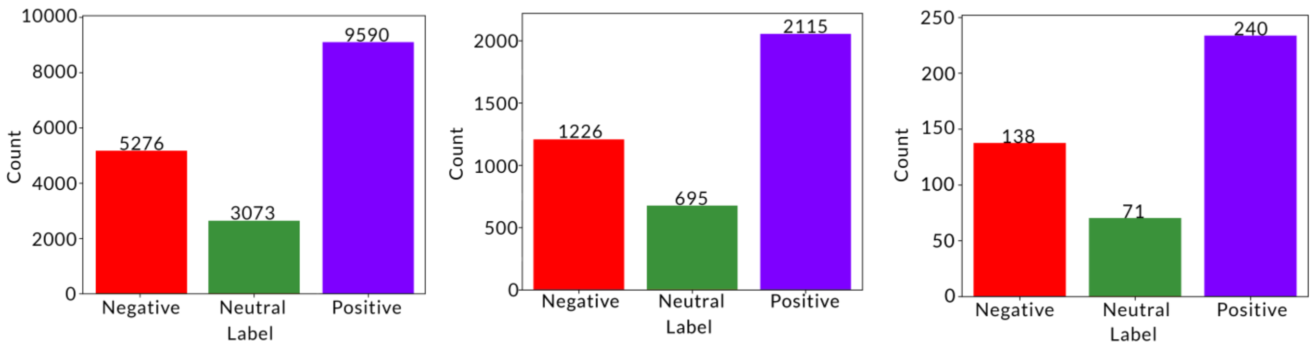
Data	Result
bagus banget pokonyaaa	bagus banget pokonya
tidak bisa atur volume suara	tidak bisa atur volume suara
tiktok bales dong biar emak aku bangga	balas agar ibu saya bangga

Tokenization in natural language processing separates text into tokens [11]. Tokens can be in the form of words, letters, or sub words (n-gram). In this work, we tokenized the dataset into a word. The tokenized word was used to the stemming process. Stemming generally refers to transforming a word in a text into base words [12], [13]. The filtered data were stemmed using a python library called Sastrawi, which makes the stemming process in this work more accessible. Sastrawi is a stemmer library that contains root Indonesian words [14]. To train the sentiment analysis model, we used the result of text preprocessing. Table IV shows the result of tokenized and stemmed review samples.

TABLE IV
TOKENIZING AND STEMMING RESULT

Data	Tokenized	Stemmed
bagus banget pokonya	['bagus', 'banget','poko']	bagus banget poko
tidak bisa atur volume suara	['tidak', 'bisa', 'atur', 'volume', 'suara']	tidak bisa atur volume suara
balas agar ibu saya bangga	['balas', 'agar', 'ibu', 'saya', 'bangga']	balas agar ibu saya bangga

In the text preprocessing stage, there was data reduction because data was filtered based on requirements for modeling. The preprocessed obtained 22,424 data. Those data are split into data training, data validation, and data testing to create LSTM and IndoBERTweet models. The amount of data is based on the label, as shown in Fig. 2. In the training process, the model learned from labeled data trained as input [15]. To evaluate early error while training using data validation and for evaluating learned model used data testing [16].



(a) (b) (c)

Fig. 2. Comparison of each label on (a) Data Train (b) Data Validation (c) Data Test

C. LSTM Model

Long short-term memory (LSTM) modifies the most popular recurrent neural network algorithm. The difference between LSTM and a standard feedforward neural network is that it has a feedback connection [17]. The purpose of the LSTM modification is to complement the shortcomings of the RNN algorithm in that it cannot make predictions based on information stored for a long time. LSTM is being able to store information for a long time and remove information that is no longer relevant.

Relevant research using the LSTM + Word2Vec method with data is an Indonesian review of the Traveloka application that got an excellent sentiment analysis model performance. With Word2Vec integration into previous research for pre-processing to help the system learn words as a vector to improve model performance [18]. Another relevant research is a multi-aspect sentiment analysis of TikTok apps review using LSTM + BERT word embedding. The research showed that LSTM combined with BERT word embedding could make high average accuracy for each prediction [5]. In this research, the model uses SoftMax as an activation function, with architecture in Fig. 3.

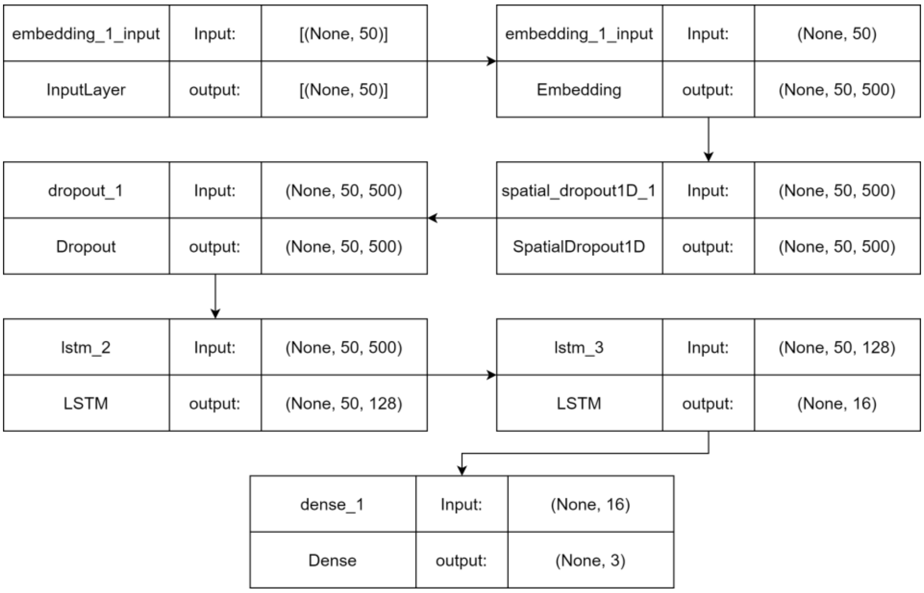


Fig. 3. LSTM Architecture

D. IndoBERTweet Model

IndoBERTweet is a pre-trained model enhanced by fine-tuning using tweets on Twitter [7]. Fine-tuning is a method to get the maximum performance limit from the standard model, with hyperparameters to improve the performance of deep learning models [19], [20]. The Indonesian TikTok review is used for fine-tuning in this research. We used 3 epochs and batch size value is 4 for hyperparameter. IndoBERTweet is an additional feature from IndoBERT, and it has architecture as in Fig. 4.

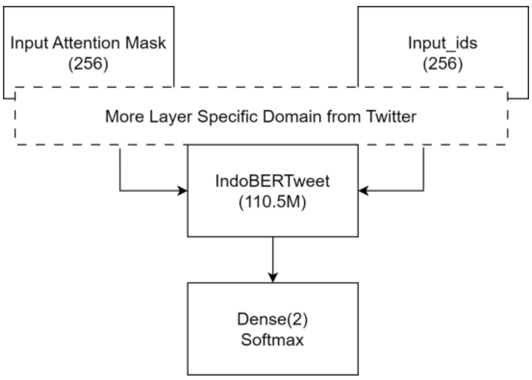


Fig. 4. IndoBERTweet Architecture

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 8, No. 3, September 2023, Pp. 774-780

In this research, the first stage of building the IndoBERTweet model is tokenizing data using the AutoTokenizer library from transformers with *indolem/indobertweet-base-uncased* as a parameter. The IndoBERTweet tokenizer has the same structure as BERT. There are [CLS] tokens as prefixes and [SEP] as closing tokens [21]. Table 5 shows the structure sample of the tokens. We used an encoded token to train the data for building the IndoBERTweet model.

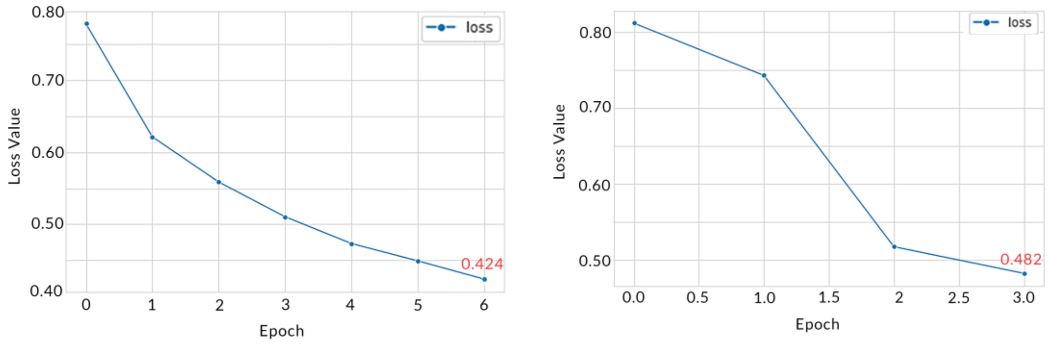
TABLE V
INDOBERTWEET TOKENIZER RESULT

Text	Tokenize	Encode Token
tidak bisa update	['[CLS]', 'tidak', 'bisa', 'update', '[SEP]']	[3, 8603, 1630, 1580, 4]
udah bagus ada fitur story foto	['[CLS]', 'udah', 'bagus', 'ada', 'fitur', 'story' 'foto', '[SEP]']	[3, 9988, 4839, 9988, 7180, 14098, 4724, 4]
Bintang lima terimakasih	['[CLS]', 'bintang, 'lima', 'terimakasih', '[SEP]']	[3, 3571, 2715, 17023, 4]

III. RESULT AND DISCUSSION

This research evaluated long short-term memory and the IndoBERTweet model using the loss function, classification report, and confusion matrix. We used data validation to evaluate the loss function on each epoch for the best model performance. The purpose of learning algorithms is to minimize the error from loss function while the training process [22].

Fig. 5(a) shows the LSTM with 42% loss data, and Fig. 5(b) shows the IndoBERTweet with 48% data. The LSTM training stopped in the sixth and IndoBERTweet in the third epoch because the loss validation did not change, while the loss kept reducing. It could prevent overfitting of the model.



(a) (b)

Fig. 5. The loss history of (a) LSTM (b) IndoBERTweet

A. *Classification Report*

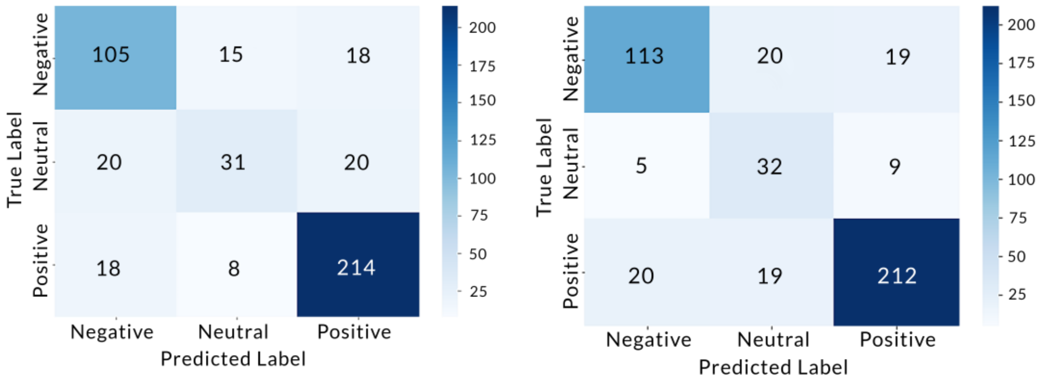
For the classification report, we used the same data test amount 449 by split dataset with a random state of 42 to evaluate each model. The classification report has some metrics for evaluating each model: precision, recall, f1-score, and accuracy. Precision is the ratio of true-positive data to the amount of positive predicted data, recall is the ratio of true-positive predictions to true-positive data, whereas f1-score is the harmonic mean of precision and recall [23], [24]. Accuracy is a measure that determines the value of the similarity between the prediction and the actual value measured [25]. Table 6 shows the evaluation of the LSTM and IndoBERTweet classification report comparison.

TABLE VI
CLASSIFICATION REPORT RESULT

Algorithm	Precision	Recall	F1-Score	Accuracy
LSTM	77%	78%	77%	78%
IndoBERTweet	82%	80%	80%	80%

B. *Confusion Matrix*

Confusion matrices are an evaluation method to calculate the classification process performance or evaluate the error in a classifier [26]. We used a data test to evaluate the model performance and visualized it on a confusion matrix with the axis predicted and true label, as in Fig. 6.



(a) (b)

Fig. 6. Confusion matrices of classifier for Indonesian TikTok reviews dataset comparison on (a) LSTM and (b) IndoBERTweet algorithm

Evaluation with a total of 449 data tests in Fig. 6(a) shows that the LSTM model can predict negative sentiment correctly with 105 data, incorrect predictions should be 20 data neutral, and 18 data should be positive. The model predicts neutral sentiment correctly with 30 data, incorrect predictions should be 15 data negative, and 8 data should be positive. For the positive sentiment, the model is better at predicting as many as 214 data correctly, incorrect predictions should be 20 data neutral, and 18 data should be negative sentiments. Fig. 6(b) displays the evaluation of the confusion matrix in the IndoBERTweet model. he was able to predict 113 data of negative sentiment correctly and incorrectly predicted five that should be neutral and 20 that should be positive. In neutral sentiment, the IndoBERTweet model can correctly predict 32 data, incorrect predictions 20 data should be the negative sentiment, and 19 data should be positive. The model can correctly predict 212 data for the positive sentiment, 19 incorrect data predictions should be the negative sentiment, and nine data should be neutral.

From these results, the standard LSTM model in this research was trained and tested with all random aspects on TikTok review from the different versions of the previous research dataset, and the model could reach 78% as the highest prediction accuracy. The previous research used LSTM integrated with Word2Vec and used three aspects with pre-determined vocabulary for data training. The LSTM model in previous research only predicts well on those aspects within 90% as the highest average accuracy. From the evaluation in previous research, it only works in 25% of all aspects of the dataset [5]. The evaluation result of IndoBERTweet on this research got the 80% as the highest model prediction accuracy to predict TikTok reviews. Relevant research using IndoBERTweet as a public opinion classifier about Covid-19 Vaccine in Indonesia was trained with the dataset from tweets on Twitter got 68% as the highest accuracy [27].

IV. CONCLUSION

In this work, the dataset from Indonesian TikTok user reviews was used to compare the performance of LSTM and the IndoBERTweet algorithm to make a sentiment analysis model. Models can classify automatically to help users quickly draw conclusions based on opinions from user reviews of TikTok. The evaluation showed that IndoBERTweet as a pre-trained model, achieved the best performance with an accuracy of 80%, while LSTM achieved an accuracy of 78%. Those evaluations were based on history training with validation data, classification reports, and confusion matrix as model prediction results. The sentiment analysis result showed that most Indonesian users give positive feedback to TikTok as a social media platform.

REFERENCES

[1] P. A. Permatasari, L. Linawati, and L. Jasa, “Survei Tentang Analisis Sentimen Pada Media Sosial,” *Majalah Ilmiah Teknologi Elektro*, vol. 20, no. 2, pp. 177–186, Dec. 2021, doi: 10.24843/mite.2021.v20i02.p01.

[2] L. Stappen, E. Cambria, and B. W. Schuller, “Sentiment Analysis and Topic Recognition in Video Transcriptions,” *Department: Affective Computing and Sentiment Analysis*, vol. 36, no. 2, pp. 88–95, Apr. 2021, doi: 10.1109/mis.2021.3062200.

[3] O. Somantri and D. Apriliani, “Support Vector Machine Berbasis Feature Selection Untuk Sentiment Analysis Kepuasan Pelanggan Terhadap Pelayanan Warung dan Restoran Kuliner Kota Tegal,” *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 5, no. 5, pp. 537–548, Oct. 2018, doi: 10.25126/jtiik.201855867.

[4] B. A. Rachid, H. Azza, and B. G. Henda, “Sentiment analysis approaches based on granularity levels,” in *WEBIST 2018 - Proceedings of the 14th International Conference on Web Information Systems and Technologies*, 2018, pp. 324–331. doi: 10.5220/0007187603240331.

[5] D. Wahyudi and Y. Sibaroni, “Deep Learning for Multi-Aspect Sentiment Analysis of TikTok App using the RNN-LSTM Method,” *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 1, pp. 169–177, Jun. 2022, doi: 10.47065/bits.v4i1.1665.

[6] R. Mas, R. W. Panca, K. Atmaja, and W. Yustanti, “Analisis Sentimen Customer Review Aplikasi Ruang Guru dengan Metode BERT (Bidirectional Encoder Representations from Transformers),” *JEISBI*, vol. 2, no. 3, p. 2021, Jul. 2021, [Online]. Available: ejournal.unesa.ac.id/index.php/JEISBI/article/view/41567

[7] F. Koto, J. H. Lau, and T. Baldwin, "IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," in *Association for Computational Linguistics*, 2021, pp. 10660–10668. doi: 10.18653/v1/2021.emnlp-main.833.
[8] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimed Tools Appl*, pp. 1–32, Jul. 2022, doi: 10.1007/s11042-022-13428-4.
[9] Q. Bi, K. E. Goodman, J. Kaminsky, and J. Lessler, "What Is Machine Learning: a Primer for the Epidemiologist Qifang," *Am J Epidemiol*, vol. 188, no. 12, pp. 2222–2239, Dec. 2019, doi: https://doi.org/10.1093/aje/kwz189.
[10] A. A. V. A. Jayaweera, Y. N. Senanayake, and P. S. Haddela, "Dynamic Stopword Removal for Sinhala Language," in *2019 National Information Technology Conference (NITC)*, Oct. 2019, pp. 1–6. doi: 10.1109/NITC48475.2019.9114476.
[11] S. Ahmadi, "A Tokenization System for the Kurdish Language," in *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, 2020, pp. 114–127. [Online]. Available: https://aclanthology.org/2020.vardial-1.11
[12] D. Khyani, S. B. S, N. N. M, and D. B. M, "An Interpretation of Lemmatization and Stemming in Natural Language Processing," *Journal of University of Shanghai for Science and Technology*, vol. 22, no. 10, pp. 350–357, Oct. 2021, [Online]. Available: https://www.researchgate.net/publication/348306833
[13] K. K. Purnamasari and I. S. Suwardi, "Rule-based Part of Speech Tagger for Indonesian Language," in *IOP Conference Series: Materials Science and Engineering*, 2018, vol. 407, no. 1, p. 012151. doi: 10.1088/1757-899X/407/1/012151.
[14] M. A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, "Improving Text Preprocessing for Student Complaint Document Classification Using Sastrawi," in *IOP Conference Series: Materials Science and Engineering*, 2020, vol. 874, no. 1, p. 012017. doi: 10.1088/1757-899X/874/1/012017.
[15] V. Nasteski, "An overview of the supervised machine learning methods," *Horizons. B*, vol. 4, pp. 51–62, Dec. 2017, doi: 10.20544/horizons.b.04.1.17.p05.
[16] E. Breck, N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich, "Data Validation for Machine Learning," in *Proceedings of Machine Learning and Systems 1*, 2019, pp. 334–347. [Online]. Available: https://proceedings.mlsys.org/paper/2019/file/5878a7ab84fb43402106c575658472fa-Paper.pdf
[17] Dr. G. S. N. Murthy, S. R. Allu, B. Andhavarapu, M. Bagadi, and M. Belusonti, "Text based Sentiment Analysis using LSTM," *International Journal of Engineering Research and*, vol. V9, no. 5, pp. 299–303, May 2020, doi: 10.17577/ijertv9is050290.
[18] P. F. Muhammad, R. Kusumaningrum, and A. Wibowo, "Sentiment Analysis Using Word2vec and Long Short-Term Memory (LSTM) for Indonesian Hotel Reviews," in *Procedia Computer Science*, 2021, vol. 179, no. 2020, pp. 728–735. doi: 10.1016/j.procs.2021.01.061.
[19] G. W. Anderson and D. J. Castaño, "Measures of fine tuning," *Physics Letters B*, vol. 347, no. 3–4, pp. 300–308, Mar. 1995, doi: 10.1016/0370-2693(95)00051-L.
[20] P. Izsak, M. Berchansky, and O. Levy, "How to Train BERT with an Academic Budget," in *Conference on Empirical Methods in Natural Language Processing, Proceedings*, Sep. 2021. doi: 10.18653/v1/2021.emnlp-main.831.
[21] 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 *ACM International Conference Proceeding Series*, 2021, pp. 258–264. doi: 10.1145/3479645.3479679.
[22] D. Rengasamy, M. Jafari, B. Rothwell, X. Chen, and G. P. Figueredo, "Deep learning with dynamically weighted loss function for sensor-based prognostics and health management," *Sensors (Switzerland)*, vol. 20, no. 3, Jan. 2020, doi: 10.3390/s20030723.
[23] N. Tatbul, T. J. Lee, S. Zdonik, M. Alam, and J. Gottschlich, "Precision and recall for time series," in *Advances in Neural Information Processing Systems*, Dec. 2018, pp. 1–11. doi: doi.org/10.48550/arXiv.1803.03639.
[24] P. A. Flach and M. Kull, "Precision-Recall-Gain curves: PR analysis done right," in *Advances in neural information processing systems 28*, 2015, pp. 1–9.
[25] D. K. Wright, "Accuracy vs. Precision: Understanding Potential Errors from Radiocarbon Dating on African Landscapes," *African Archaeological Review*, vol. 34, no. 3, pp. 303–319, Jun. 2017, doi: 10.1007/s10437-017-9257-z.
[26] E. Beauxis-aussalet and L. Hardman, "Visualization of Confusion Matrix for Non-Expert Users," in *IEEE Information Visualization (InfoVis 2014)*, 2014.
[27] S. Saadah, Kaenova Mahendra Auditama, Ananda Affan Fattahila, Fendi Irfan Amorokhman, Annisa Aditsania, and Aniq Atiqi Rohmawati, "Implementation of BERT, IndoBERT, and CNN-LSTM in Classifying Public Opinion about COVID-19 Vaccine in Indonesia," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 4, pp. 648–655, 2022, doi: 10.29207/resti.v6i4.4215.