Proceeding of the International Conference on Electrical Engineering and Informatics
Volume. 1 Nomor. 1, 2024
Hal 109-119
Available Online at: https://prosiding.arteii.or.id/index.php/ICEEI

Emotion Classification in Sundanese Text Using LSTM and BERT Models

Mahazzam Afrad [1], Fauzi Irfan Syaputra [2], Gilang Fibarkah [3*], Tectonia Nurul Silvani[4]

[1,2,3,4], Information System, Telkom University Purwokerto, Banyumas, Indonesia

Email: [1] mahazama@telkomuniversity.ac.id, [2] fauipang@student.telkomuniversity.ac.id ,
[3]gilangfibarkah@student.telkomuniversity.ac.id , [4] tectonians@student.telkomuniversity.ac.id*

Abstract. The Sundanese language, once spoken by 48 million individuals, has experienced a significant decline in speakers, losing 2 million in the past decade. This decline is attributed to weakened intergenerational transmission and the dominance of more widely used languages. The challenges in developing Natural Language Processing (NLP) tools for Sundanese stem from the lack of annotated corpora, trained language models, and adequate processing tools, complicating efforts to preserve and enhance the language's usability. This research aims to address these challenges by implementing emotion classification in Sundanese text using Long Short-Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT) models. The study utilizes a dataset of annotated Sundanese tweets, applying preprocessing techniques such as cleansing, stopword removal, stemming, and tokenization to prepare the data for analysis. The results indicate that the BERT model significantly outperforms the LSTM model, achieving an accuracy of approximately 80% compared to LSTM's 70%. These findings highlight the potential of advanced NLP techniques in enhancing the understanding of emotional nuances in Sundanese communication and contribute to the revitalization of the language in the digital age.

Keywords BERT, Emotion Classification, LTSM, Sundanese Language.

1. INTRODUCTION

Sundanese, which was previously spoken by 48 million people, has lost 2 million speakers in the last 10 years, indicating a significant change in the number of speakers. (Aranditio, 2023). The reduced usage of the Sundanese language is influenced by factors such as the weakening of intergenerational transmission and the dominance of more widely used languages (Sudarma et al., 2018) (Hananto, 2022). Sundanese is categorized as a low-resource language in the field of Natural Language Processing (NLP) (Cahyawijaya et al., 2021). As part of artificial intelligence, NLP enables machines to interact with human language, both in text and speech, to understand, analyze, and generate natural language (Badawi, 2021). However, the development of NLP for the Sundanese language is hindered by challenges such as the lack of annotated text corpora, specifically trained language models, and adequate processing tools (Sakinah, Ramadhan, & Hartono, 2024). These limitations complicate efforts to preserve and enhance the usability of the Sundanese language. The scarcity of digital content in Sundanese, coupled with the declining interest among younger generations to use it in daily conversations and social media, further obstructs its preservation. Therefore, developing NLP-based technologies is essential to address these challenges and support the revitalization of the Sundanese language. Emotion

Received: October 30, 2024; Revised: November 30, 2024; Accepted: Desember 26, 2024; Published: Desember 31, 2024;

classification, for instance, can provide deeper insights into emotional nuances in Sundanese communication, enriching the digital content ecosystem.

The emotional state of an individual, commonly referred to as emotion, is dynamic and can change depending on time, situations, and individual characteristics (Afrad, 2024). Human emotions, which are complex and influenced by context, culture, and personal experiences, require advanced systems to classify various emotions in a machine-understandable form. In this regard, classifying emotions in the Sundanese language presents unique challenges due to the limitations of available data and resources. Nevertheless, emotion classification plays a crucial role in processing text that has been trained to identify emotions in new text, enabling the recognition of emotional patterns in Sundanese-language communication..

This research aims to address the challenges in classifying emotions in Sundanese text using two distinct methods: Long Short-Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT). The dataset utilized in this study was obtained from previous research (Putra, Wasmanson, Harmini, & Utama, 2020), consisting of annotated Sundanese text with emotions such as joy, fear, anger, and sadness. The collected data undergoes multiple pre-processing steps to prepare the text for analysis, including converting text to lowercase, removing numbers, punctuation, and excess spaces. Additionally, the process involves cosine similarity calculations with standardized words, tokenization, word replacement based on a dictionary, and stopword removal to enhance the quality and relevance of the text analysis. The LSTM and BERT methods are expected to yield optimal results in classifying emotions in Sundanese text. Evaluation metrics such as accuracy, precision, recall, and F1-Score are used to assess model effectiveness, while comparing the two methods provides insights into their respective strengths and limitations.

2. LITERATURE REVIEW

Emotion classification in text data has been a growing field of research, particularly for underrepresented languages such as Sundanese. The study by (Putra et al., 2020) addresses this gap by evaluating machine learning algorithms for emotion classification in Sundanese text, focusing on emotions such as fear, joy, anger, and sadness. Using a dataset of 2518 annotated Sundanese tweets, the authors employed preprocessing techniques like case folding, tokenization, and stopword removal, followed by feature extraction using TF-

IDF. The results showed that SVM achieved the highest accuracy (95%), establishing a crucial foundation for the development of emotion classification systems for Sundanese. Advanced deep learning methods such as Long Short-Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT) provide more sophisticated approaches. LSTM captures sequential patterns in text data, while BERT offers contextual word embeddings, allowing a deeper understanding of emotional context. This study builds pon the dataset of (Putra et al., 2020) but includes additional preprocessing steps to enhance data quality, such as converting text to lowercase, removing numbers and punctuation, tokenization, word standardization using cosine similarity, and stopword elimination. These steps ensure that the analyzed text is more relevant and clean, which helps the model classify emotions more accurately. While (Putra et al., 2020) research provides a solid foundation, it leaves gaps in terms of result generalization, dataset limitations, and the application of more advanced technologies. Building on BERT's contextual embeddings and LSTM's sequential pattern recognition, this research contributes to more effective emotion classification methods for Sundanese.

Besides (Putra et al., 2020) other studies have contributed significantly to emotion recognition in Indonesian text. One recent study by (William, Kenny, & Chowanda, 2024) evaluated emotion recognition on Indonesian text using a dataset annotated with six basic emotions from Ekman (happy, fear, anger, disgust, sadness, surprise) plus a neutral label. The dataset consists of 7,629 tweets after resolving annotator disagreements. The study used three models for emotion classification: original IndoBERT, fine-tuned IndoBERT, and Bidirectional Long Short-Term Memory (Bi-LSTM) networks. The results showed that the fine-tuned IndoBERT model achieved the highest accuracy (92.3%), outperforming the original IndoBERT (90.7%) and Bi-LSTM (84.0%). (William et al., 2024) research also highlighted challenges in classifying neutral emotions and provided a publicly accessible dataset for emotion recognition in Indonesian text. The findings suggest that fine-tuning pre-trained models can significantly improve accuracy, especially for languages with limited NLP resources like Indonesian. This research is relevant for enriching the literature on emotion classification in local languages, including Sundanese. Based on the findings, a combined approach leveraging BERT and LSTM's strengths could be a potential strategy for improving emotion classification accuracy in languages with complex structures and limited datasets. Additionally, adding neutral or mixed emotion labels to Sundanese datasets could be a next step toward capturing a broader spectrum of emotions.

Another related study by (Sudarma et al., 2018) on emotion detection in Indonesian text analyzed emotions in tweets from Transjakarta and Commuter Line users. The dataset included 20,395 tweets from Commuter Line and 10,649 from Transjakarta. The methods employed combined various word embeddings (BERT, Word2Vec, GloVe) and classification models (BiLSTM, LSTM, CNN) across nine test scenarios. The results indicated that the BERT-BiLSTM model performed the best, with the highest accuracy (89.20%) on the combined dataset, outperforming other models in precision, recall, and F1-score. These findings highlight the effectiveness of integrating the contextual understanding of BERT with the sequential processing capabilities of BiLSTM for emotion detection in Indonesian text. This research makes an important contribution by showing the potential integration of transformer-based models like BERT and sequential networks like BiLSTM in emotion analysis. The findings are also relevant for other local languages, including Sundanese, which shares similar structural and contextual characteristics

3. METHODS

This research aims to classify emotions in Sundanese text using LSTM and BERT based approaches. Figure 1 provides an overview of the emotion classification process, including dataset collection, preprocessing steps, and model training. The process begins with the collection of a dataset from the study conducted by (Putra et al., 2020). This dataset contains 2518 Sundanese tweets that have been previously annotated with approriaate emotion labels. The emotion labels used in this dataset, as explained in Table 1, include categories such as joy, anger, sad, and fear.

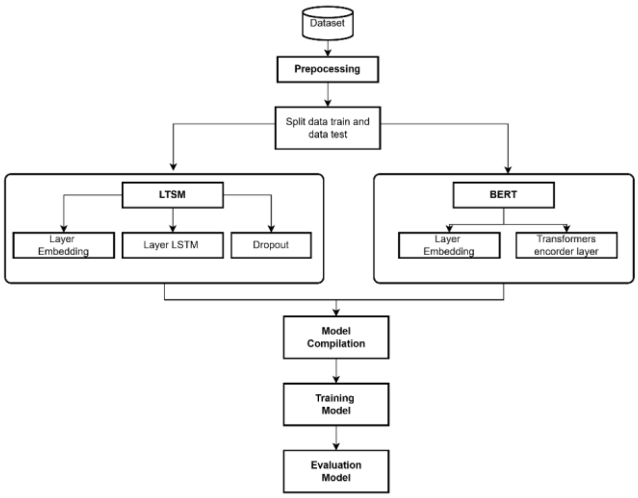

Figure 1. Overview of the emotion classification process

Table 1. Sample Dataset

label	data
anger	sok geura leungit atuh sia teh corona, matak gelo yeuh aing unggal poe gogoleran
sadness	nya prihatin siih tp kedah kitu at pak Bozz hapuntn abi tos kmh wntu
joy	Bingah pisan patepang sareng pangerasa. Sing katampi kalayan pinuh midulur...
fear	asa hariwang kieu.. lalakon hirup teh asa nyorangan.. asa ieu mah..

The data preprocessing stage involved several steps, including case folding, cleansing, stopword, stemming, and tokenizing. Initially, the next was converted to lowercase to ensure uniformity. Cleansing was applied to remove any unnecessary characters or noise (Sharou, Li, & Specia, 2021). Stemming was then conducted to reduce words to their root forms (Jabbar et al., 2023). Stopwords, which are common words to that do not contribute significantly to the meaning, were removed to enhance the focus on meaningful content(Kaur, 2018). Finally, tokenization was performed to break down the text into individual words or tokens, facilitating futher analysis (Wang et al., 2024). These steps are illustrated in Table 2, showing the transformation of the next before and after each preprocessing step.

Table 2. Before after preprocessing steps

	Before	After
lowercase	Nu katoel katuhu nu nyerina kenca, goblog wasitna	nu katoel katuhu nu nyerina kenca, goblog wasitna
cleansing	Nu katoel katuhu nu nyerina kenca, goblog wasitna	nu katoel katuhu nu nyerina kenca goblog wasitna
stemming	Nu katoel katuhu nu nyerina kenca, goblog wasitna	nu boga katoel katuhu nu boga nyerina kenca goblog wasitna
stopwords	Nu katoel katuhu nu nyerina kenca, goblog wasitna	boga katoel katuhu boga nyerina kenca wasitna
tokenizing	Nu katoel katuhu nu nyerina kenca, goblog wasitna	['boga', 'katoel', 'katuhu', 'boga', 'nyerina', 'kenca', 'wasitna']

After the prepocessing stage, the next step involved splitting the dataset into training and testing subsets. The distribution of the dataset, summarized in Table 3, was performed using an 80:20 ratio, allocating 80% of the data for training the model and reserving the remaining 20% for testing its performance.

Table 3. Distribution dataset

Ratio	Data Train	Data Test
80:20	2014	504

The focus moves into the LSTM model, starting with the embedding layer which plays crucial role in transforming input data into a suitable format for processing. In this layer, the input_dim parameter is set to the size of the word_index, which represents the total vocabulary and enables feature extraction from the input text (Hutagaol & Arifin, 2024). The input_dim parameter is set to the size of the word_index, representing the total vocabulary, allowing the model to learn meaningful word embeddings during training. Key parameters like input_dim, output_dim, and input_length are crucial for proper configuration. The LSTM layer is configured with arguments such as return_sequences, and kernel_regularizer. The return_sequences argument control whether the output includes the entire sequence or only the final state. The kernel_regularizzer applies penalties to the layer’s weights to enhance generalization and minimize overffiting, while droput serves a similar purpose by randomly setting a fraction of input units to zero during training, thereby preventing reliance neurons and improving the model’s robustness.

The layer embedding in BERT model using parameter word_embedding, position embedding, token_type_embedding. In this research, 12 transformer encoders layers are used, each containing multi-head self-attention and feed-forward sub-layers, which work together to capture complex contextual relationships within the input text.

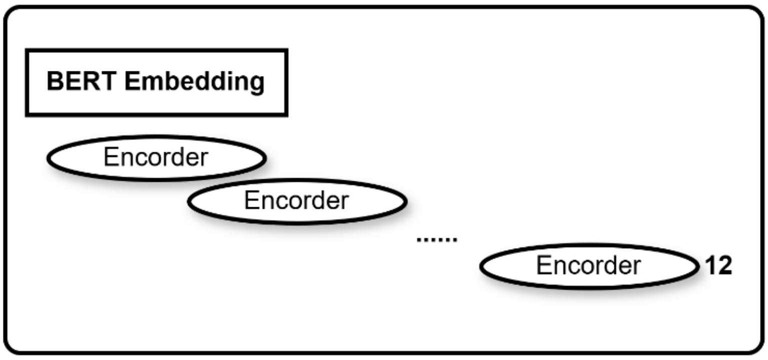

Figure 2. BERT layer

4. RESULTS

The confusion matrix presented in Figure 3 evaluates the performance o f the LTSM classification model accross four classes (0,1,2, and 3). The diagonal elements (73,71,84, and 66) represent the number of correctly predicted instances for each respective class, indicating the model’s accuracy in these cases. However, the off-diagonal values highliht missclassifications, where the model incorrectly predicted an instance to belong to a different class.

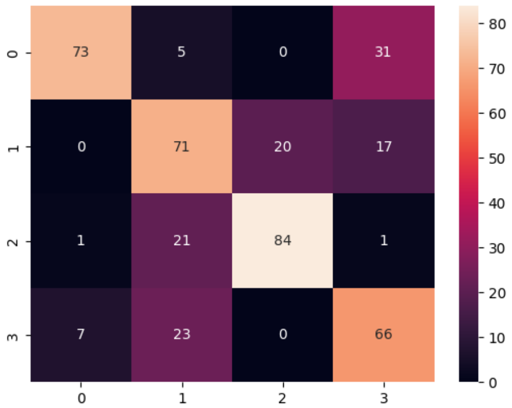

Figure 3. Confusion mats LSTM

For example, class 0 shows 31 instances misclassified as class 3, and class 1 has 20 instances misclassified as class 2. Additionally, class 3 exhibits 23 instances misclassified into class 1. These misclassifications suggest areas where the model's performance could be improved, particularly in differentiating between classes with similar characteristics. While the model demonstrates strong performance in class 2 with the highest correctly classified instances (84), further analysis and fine-tuning may be required to enhance the overall accuracy and reduce classification errors, particularly between classes 0, 1, and 3.

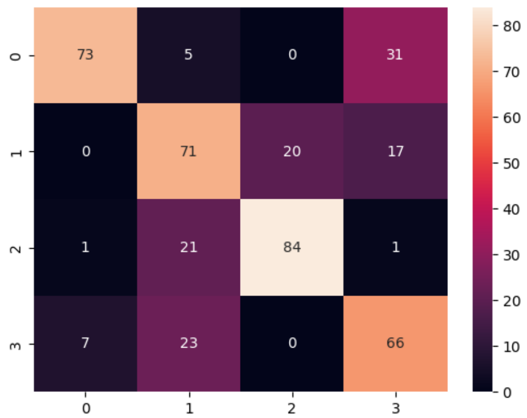

Figure 4. Confusion mats report

The confusion matrix report in Figure 4 evaluates the LSTM model's performance across four classes (0, 1, 2, and 3) using precision, recall, and F1-score metrics. Class 2 shows the best performance with 84 correct predictions and minimal misclassification, while classes 0, 1, and 3 exhibit higher error rates. Notably, class 1 has 20 misclassifications into class 2, and class 0 is often misclassified into class 3, with 31 incorrect predictions. These results suggest the model performs well for class 2 but requires further refinement to improve differentiation between the remaining classes.

	precision	recall	f1-score	support
0	0.90	0.67	0.77	109
1	0.59	0.66	0.62	108
2	0.81	0.79	0.80	107
3	0.57	0.69	0.63	96
accuracy			0.70	420
macro avg	0.72	0.70	0.70	420
weighted avg	0.72	0.70	0.71	420

Figure 5. Classification report metrics LTSM

Figure 5 presents the classification report metrics for the LSTM model, highlighting the precision, recall, F1-score, and support for each class. The results indicate that Class 0 achieved the highest precision and F1-score, demonstrating the model's effectiveness in identifying this category. Conversely, Class 1 exhibited the lowest performance metrics, suggesting potential areas for improvement in the model's classification capabilities.

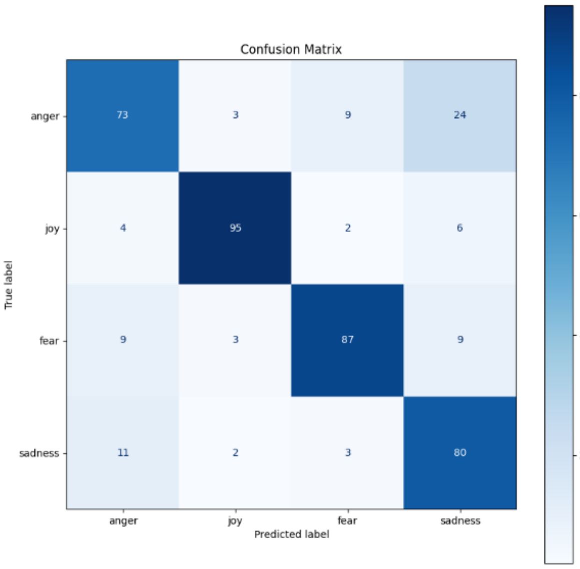

Figure 6. Confusion matrix BERT

Figure 6 illustrates the confusion matrix for the BERT model, providing a detailed overview of the model's classification performance across different emotion categories.

The matrix reveals that the model accurately predicted the "joy" class with a high true positive rate of 95, while the "anger" and "sadness" classes exhibited some misclassifications. Notably, the confusion matrix highlights areas for improvement, particularly in reducing the misclassification of "anger" and "sadness," which may enhance the overall performance of the model.

	precision	recall	f1-score	support
0	0.75	0.67	0.71	109
1	0.92	0.89	0.90	107
2	0.86	0.81	0.83	108
3	0.67	0.83	0.74	96
accuracy			0.80	420
macro avg	0.80	0.80	0.80	420
weighted avg	0.81	0.80	0.80	420

Figure 7. Classification report

The classification report or the BERT model, as shown in Figure 7, indicates a robust performance across all emotion categories, with an overall accuracy of 80%. Notably, Class 1 achieved the highest precision and F1-score, reflecting the model's effectiveness in accurately identifying this category. However, Class 3 demonstrated a lower precision of 0.67, suggesting that further optimization may be necessary to enhance the model's ability to classify this emotion accurately.

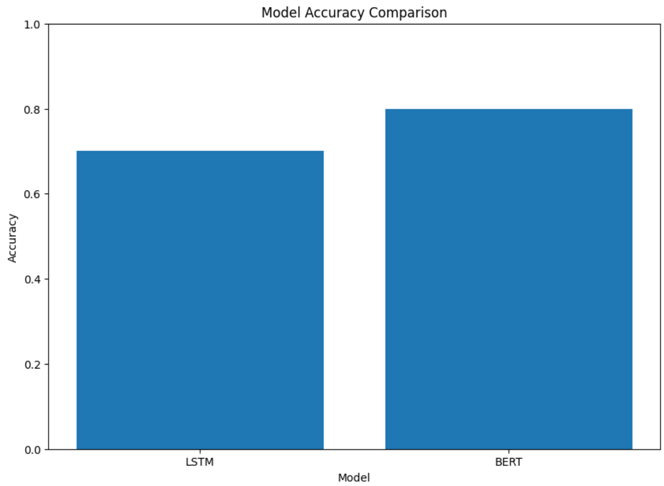

Figure 8. Comparison model accuracy

The performance of two models LSTM and BERT was evaluated, revealing that model LSTM achieved an accuracy of approximately 70% and model BERT significantly outperformed with an accuracy of approximately 80%, underscroing capability of BERT in handling complex language tasks.

5. DISCUSSION

This research compared the performance of LSTM and BERT models for emotion classification in Sundanese text. The results reveal that BERT significantly outperforms LSTM, achieving an accuracy of approximately 80%, while LSTM reached around 70%. This demonstrates the superior capability of BERT in handling complex language tasks, particularly in emotion classification. These findings highlight the potential of BERT in enhancing NLP systems for languages with limited resources and deepening the understanding of emotional nuances in Sundanese communication.

6. CONCLUSION

In this section, the author presents brief conclusions derived from the research results, along with suggestions for advanced researchers or general readers. The conclusion may review the main points of the paper but should not replicate the abstract. Additionally, the author should identify the major flaws and limitations of the study, which may affect the validity of the findings and raise questions from readers. These limitations require critical judgment and interpretation of their impact. The author should address the question: Is this a problem related to error, method, validity, or other factors

7. LIMITATION

This research faced challenges in hyperparameter tuning, which resulted in suboptimal model accuracy. The process of selecting the appropriate hyperparameters is crucial, as it directly impacts the model's performance. In future research, it is recommended to explore a more systematic approach to hyperparameter optimization, such as using grid search or Bayesian optimization techniques. Additionally, modifying the model architecture, including the layering and the number of units in each layer, could further enhance performance. By implementing these strategies, subsequent studies may achieve higher accuracy and better generalization of the model across various datasets.

8. REFERENCES

Afrad, M. (2024). Utilization of principal component analysis to improve emotion classification performance in text using artificial neural networks. Journal of Applied Intelligent System, 9(1), 8–18.

Aranditio, S. (2023). Para penutur bahasa daerah berguguran.

Badawi, A. (2021). The effectiveness of natural language processing (NLP) as a processing solution and semantic improvement. International Journal of Economic, Technology and Social Sciences (Injects, 2(1), 36–44.

Cahyawijaya, S., Winata, G. I., Wilie, B., Vincentio, K., Li, X., Kuncoro, A., … Fung, P. (2021). IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. EMNLP 2021 - Conference on Empirical Methods in Natural Language Processing, Proceedings, 8875–8898.

Hananto, A. (2022). Bahasa Sunda dan urgensi perlindungan bahasa daerah.

Hutagaol, Y. R., & Arifin, Y. (2024). Semantic-based email spam classification using BERT method. Journal of Information Technology and Computer Science (INTECOMS, 7(5), 1823–1836.

Jabbar, A., Iqbal, S., Tamimy, M. I., Rehman, A., Bahaj, S. A., & Saba, T. (2023). An analytical analysis of text stemming methodologies in information retrieval and natural language processing systems. IEEE Access, 11(December), 133681–133702.

Kaur, J. (2018). Stopwords removal and its algorithms based on different methods. International Journal of Advanced Research in Computer Science, 9(5), 81–88.

Putra, O. V., Wasmanson, F. M., Harmini, T., & Utama, S. N. (2020). Sundanese Twitter dataset for emotion classification. CENIM 2020 - Proceeding: International Conference on Computer Engineering, Network, and Intelligent Multimedia 2020, 391–395.

Sakinah, M. A., Ramadhan, T. I., & Hartono, R. (2024). Neural machine translation untuk bahasa Sunda loma – Sunda halus menggunakan long short term memory. Jurnal Komputer Antartika, 2(1), 26–34.

Sharou, K. Al, Li, Z., & Specia, L. (2021). Towards a better understanding of noise in natural language processing. International Conference Recent Advances in Natural Language Processing, RANLP, 53–62.

Sudarma, T. F. D., Wahya, Citraresmana, E., Indira, D., Muhtadin, T., & Lyra, H. M. (2018). Upaya pemertahanan bahasa-budaya Sunda di tengah pengaruh globalisasi. Jurnal Pengabdian Kepada Masyarakat, 2(12), 1–6.

Wang, D., Li, Y., Jiang, J., Ding, Z., Jiang, G., Liang, J., & Yang, D. (2024). Tokenization matters! Degrading large language models through challenging their tokenization. ArXiv Preprint ArXi, 1–17. Retrieved from http://arxiv.org/abs/2405.17067

William, S., Kenny, & Chowanda, A. (2024). Emotion recognition Indonesian language from Twitter using IndoBERT and Bi-LSTM. Communications in Mathematical Biology and Neuroscience, 2024, 1–15.