THE 1ST JAMBI INTERNATIONAL CONFERENCE ON ENGINEERING, SCIENCE AND TECHNOLOGY (#1 JICEST)

Sentiment Analysis Model of Verbal Violence Behaviour on Twitter by Using Indo-BERT

Nurjoko[1], Admi Syarif [*,2,] R.Z. Abdul Aziz[3] and Ardian Ulvan[4]

**Corresponding author*
*ORCHID IDs: https://orcid.org/0000-0003-3316-0388

[1)] PhD Student, Department of Computer Science, Faculty of Mathematics and Sciences, Lampung University, Indonesia, 35145
[2)]Department of Computer Science, Faculty of Mathematics and Sciences, Lampung University, Indonesia, 35145
[3)]Department of Information Engineering, Faculty of Computer Science, Institute Informatics dan Business Darmajaya, Lampung, Indonesia
[4)]Department of Electrical Engineering, Faculty of Engineering, Lampung University, Indonesia, 35145

email: admi.syarif@fmipa.unila.ac.id

Abstract

The emergence of verbal violence on the social media platform Twitter has become an increasingly concerning issue in recent years. Verbal violence encompasses various forms of communication that are degrading, insulting, derogatory, and threatening, often harming individuals or groups. The primary issue addressed in this study is the identification and understanding of patterns, trends, and sentiments related to verbal violence behaviour in the context of social media, particularly on Twitter. Another challenge lies in the efficient efforts to gather and label a significant amount of relevant datasets. This research aims to develop sentiment analysis model of Verbal Violence Behaviour (VVB) on Twitter by using Indo-BERT. The accuracy of the model will be compared with those of BERT. This research commences with data collection through web crawling, utilizing the category of verbal violence behaviour as a reference. Labelling datasets is carried out through a combination of manual and automated methods using a semi-supervised learning approach. This process involves self-learning, where unlabelled data is automatically labelled using a pre-trained model. Datasets are categorized into positive sentiment, negative sentiment, and neutral sentiment. The Indo-BERT Model is employed as the analytical framework. The evaluation of results is conducted by implementing a confusion matrix. Findings from the experiments indicate that the model exhibits a stronger capability in processing the Indonesian.

Keywords: Sentiment Analysis, Verbal Violence Behaviour (VVB), Sosial Media, Twitter

1. INTRODUCTION

Social media has become one of the vital means for disseminating information and conveying messages to the public. Twitter, as one of the popular social media platforms, is used by many to communicate, share opinions, engage in discussions, and express views and thoughts on various social and political issues. However, the use of Twitter often involves negative content such as verbal abuse. Verbal abuse can trigger emotions and cause discomfort to those involved [1]. Therefore, it is essential to conduct sentiment analysis on Twitter content, especially related to verbal abuse.

Verbal violence behaviour often occurs in the form of spoken communication involving insults, derogatory language, or humiliation [2]. This encompasses behaviours such as mental abuse, blaming, negative labelling, or baseless accusations. Verbal violence behaviour also includes all forms of speech that are demeaning, offensive, threatening, and inappropriate, such as tarnishing someone's reputation, insulting religion, provocation, or even spreading false information[3]. Verbal violence behaviour on social media can lead to various serious problems for individuals and society. Its impacts include psychological issues, discomfort, cyberbullying, mental health and well-being concerns, legal violations, social divisions, and privacy breaches. One common problem arising from verbal violence behaviour on social media is cyberbullying. This is easier to perpetrate because perpetrators don't need to confront their victims directly. They can easily engage in intimidating actions through the internet or smartphones without seeing the consequences on others. [4] Cyberbullying often occurs on social networking platforms like Facebook, Twitter, and Instagram. It has a psychological impact, particularly on children, as victims of cyberbullying can be reached by anyone, anytime.

One effective method for sentiment analysis is the BERT model (Bidirectional Encoder Representations from Transformers), which is one of the latest natural language processing (NLP) techniques and has proven to perform language processing tasks accurately and efficiently [10]. In Indonesia, the INDOBERT model, designed specifically for the Indonesian language, has been developed. This model can effectively process the Indonesian language and analyze sentiments in Twitter content written in Indonesian. To ensure the accuracy and effectiveness of the INDOBERT model in analyzing sentiment in content containing verbal abuse on Twitter, an evaluation will be performed using the confusion matrix method. Previous research conducted [11] employed the Convolutional Neural Network (CNN) method to classify sentiment in Indonesian Twitter data and achieved an 81.74% classification accuracy [11]. However, this study focuses on the use of the BERT method and the INDOBERT model. Additionally, it specifically emphasizes sentiment classification related to verbal abuse. Previous research on sentiment analysis in the Indonesian social media context is limited, with few studies concentrating on verbal abuse. For example used the BERT model to analyzing sentiment regarding bullying on Twitter and achieved a high accuracy of 81%. However, that research did not specifically address verbal abuse. Therefore, this study aims to perform sentiment analysis on content in Indonesian that contains verbal abuse on Twitter using the INDOBERT model. The performance of the INDOBERT model will be measured and evaluated with a confusion matrix. It is expected that this research will contribute new insights into the use of the INDOBERT model for sentiment analysis of content containing verbal violence behaviour on Twitter.

Research in the field of NLP or Natural Language Processing using the Indonesian language is not yet widely available due to the lack of accessible data sources, even though Indonesian is spoken by around 199 million people and ranks 11th as one of the most commonly used

languages in the world in 2017 [5]. Therefore, a pre-trained model called Indo-BERT was developed [6]. Indo-BERT is a variation of BERT (Bidirectional Encoder Representations from Transformers) following the BERT-Base model (uncased) with 12 tasks, having a corpus of over 220 million words sourced primarily from Indonesian Wikipedia (74 million words), Kompas, Tempo, Artikel Liputan6 (total 55 million words), and Indonesia Web Corpus (90 million words). Indo-BERT was first introduced in a paper in September 2020, with data collection from Indo4B, covering contextual pre-training on publicly available texts, blogs, news, and websites [7]. [8] Experiments show that INDOBERT achieves state-of-the-art performance over most of the tasks in INDOLEM. Also achieved the highest accuracy compared to the use of KNN, SVM, and Naïve Bayes methods [9] In the first experiment, it was revealed that KNN and SVM could only classify data into the majority class, and Naive Bayes could classify the minority class but had lower accuracy compared to Indo-BERT, with 0.3391 in class 1, 0.0863 in class 2, and 0.2368 in class 3.

This research will use the Indo-BERT method to conduct sentiment classification analysis of verbal violence behaviour on Twitter social media. By considering features such as tweets and data labels, it is expected to achieve precise and accurate sentiment classification of verbal violence behaviour. Data will be selectively collected from user-generated Twitter content. Twitter was chosen as the data source due to its status as the second-largest social media platform in the world after Facebook, and Indonesia ranks as the fifth-largest Twitter user base globally, following Brazil [12]. The performance of the Indo-BERT model will be measured and evaluated using a confusion matrix. The results of this research will contribute new insights into sentiment analysis of verbal violence behaviour in the Indonesian language on Twitter. It is anticipated that this research will provide a better understanding of the effectiveness of using the INDOBERT model for sentiment analysis of content containing verbal abuse on social media.

2. RESEARCH METHODOLOGY

2.1.Research Flow

The stages of this research consist of four main steps, starting with data gathering from twitter, data labelling, data preprocessing, model classification, and evaluation. Figure 1 represents the system design diagram used in the study.

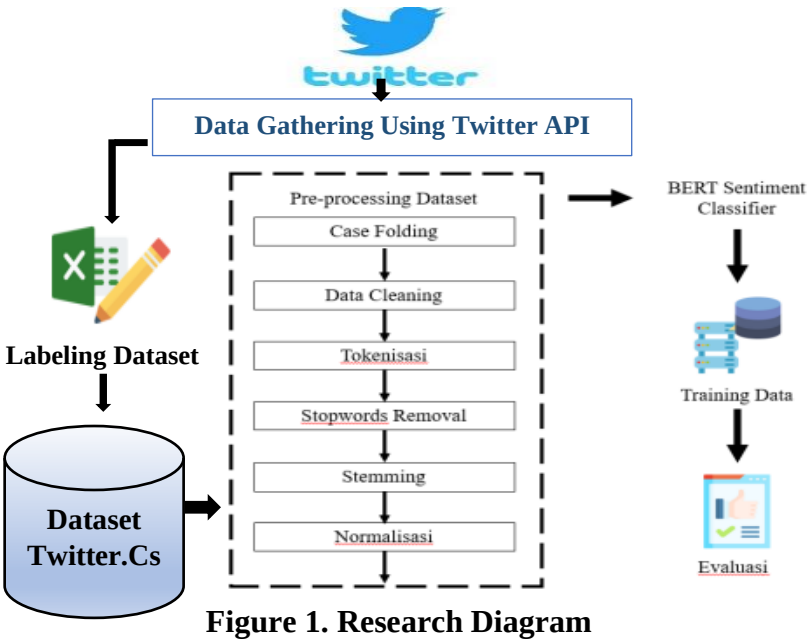

Figure 1. Research Diagram

2.2 Data Gathering

Data collection from Twitter was carried out using the crawling method using the Python library. Crawling was performed with keyword categories containing the meaning of verbal abuse, within the time range of March 01, 2023, to June 01, 2023. The total data obtained amounted to 44,467 entries, distributed based on the specified search keywords as shown in Figure 2. Documentation of the crawling results was saved in CSV file format.

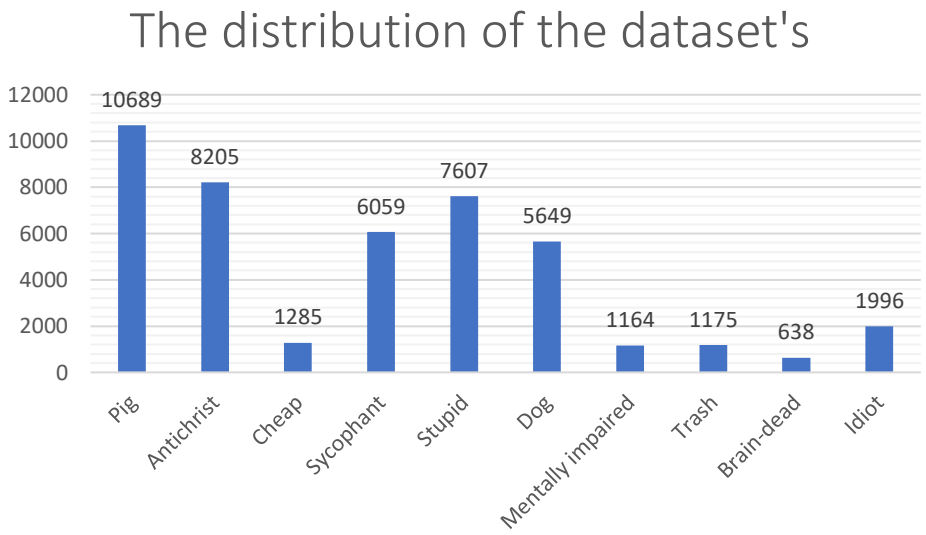

Figure 2. The distribution of the dataset's quantity

2.3 Data Labelling

The data labelling process is carried out manually and automatically. Manual labelling is performed on sampled data from the entire dataset. These data are then labelled by three different annotators according to the following criteria: 1) Data with negative sentiment are labelled as "2", 2) Data with positive sentiment are labelled as "1", and 3) Data with neutral sentiment are labelled as "0".

The Automatic Data Labelling process is carried out using one of the approaches in the Semi-supervised Learning process for sentiment analysis. In this method, data that is unlabelled will be automatically labelled using a model that has been pre-trained. This process leverages the sentiment predictions learned by the model from labelled data. The implementation of this technique aims to use data efficiently and increase the amount of labelled data for training the sentiment analysis model.

2.3 Data Preprocessing

Before data is used to train the Supervised Learning model, it is important to undergo data preprocessing steps to ensure data quality and consistency. In this phase, stop word removal and slang word replacement are performed using an NLP library to eliminate irrelevant words and replace slang words with their appropriate standard forms. Additionally, stemming is carried out using the Sastrawi stemmer to transform words in the dataset into their basic or root forms, addressing variations of similar words, so that words with the same root are considered as a single entity. The results of data preprocessing can be seen in Figure 3. The number of datasets

has decreased to 40,756 from the previous total of 44,467 data, which is due to the data cleansing process in this data preprocessing stage.

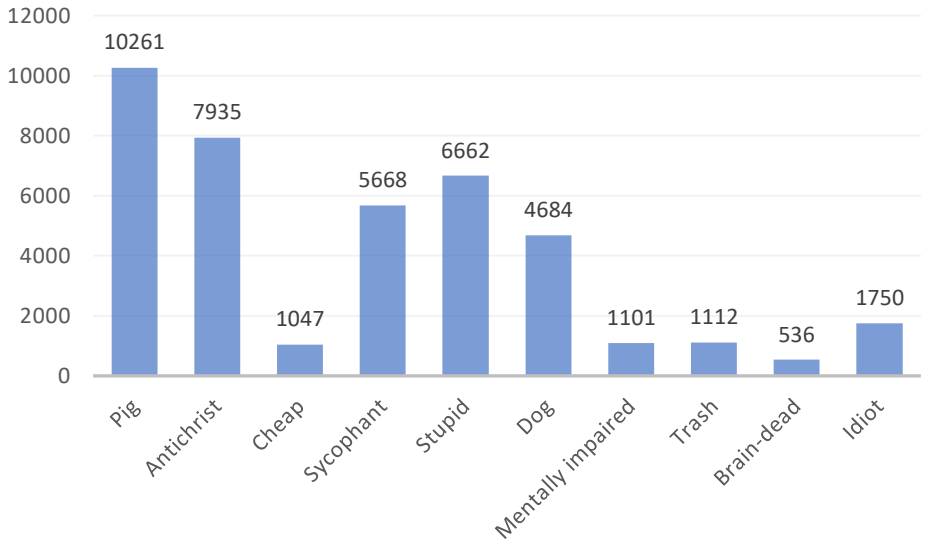

Figure 3. The distribution of the dataset's after preprocessing data

2.4 Model Classification

The researcher adopted the INDOBERT model as the analytical framework. This decision is based on the fact that INDOBERT has been trained on over 220 million Indonesian language words. After the initial data preprocessing stage, the dataset will be transformed into a format that is compatible with INDOBERT, which involves representing words in a vector format using the INDOBERT Tokenizer tool. The next step involves fine-tuning the hyperparameters, where the pre-trained INDOBERT model is adjusted to perform sentiment classification tasks. With the BERT method, a language model can read in both directions simultaneously. BERT is designed to help computers understand the meaning of ambiguous language in text by using the surrounding text to establish context through two pre-trained BERT tasks, Masked Language Model (MLM) and Next Sentence Prediction (NSP).

2.5 Evaluation

The evaluation stage is designed to assess the results of supervised learning-based labelling and sentiment analysis of the sentences within the dataset. When analyzing verbal abuse sentiment using the INDOBERT model, a confusion matrix can be employed to assess the model's performance in categorizing comments as positive, neutral, or negative. The highest accuracy value obtained during the previous training process will serve as the model's accuracy. To obtain predictions from the model, the confusion matrix is utilized, as depicted in Table 1.

Tabel 1. Confussion Matrix

Predicted Class		True Class		
		Positif	Netral	Negatif
	Positif	True Positif (TP)	False Positif (FNt)	False Positif (FP)
	Netral	False Netral (FNt)	True Netral (TNt)	False Netral (FNt)
	Negatif	False Negatif (FN)	False Negatif (FNt)	True Negatif) (TN)

Four indicators are generated by the Confusion Matrix, namely Accuracy, Precision, Recall, and F1Score or F-Measure. In this evaluation calculation, we still take into account the values as explained earlier. The equations for the confusion matrix can be observed in formulas (1) - (4).

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

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{F1 Score} = \frac{2 \text{ x Recall x Precision}}{\text{Recall+Precision}} \quad (4)$$

3. RESULT AND DISCUSSION

The next step involves building a model for classifying verbal violence behaviour using the Indo-BERT method. The Indo-BERT method itself has two mechanisms, which are the encoder and decoder. The encoder is responsible for reading the input, while the decoder generates predictions. Additionally, because Indo-BERT can read text bidirectionally, it allows the model to learn from the surrounding context.

This research was conducted using 40,756 datasets obtained from crawling Twitter. The experiments were carried out three times. To achieve optimal results, the number of epochs was set to 10 times. The accuracy results of the three experiments showed accuracy rates of 70%, 72%, and 71% respectively as shown in the figure 4, figure 5, and figure 6. The decrease in performance may be due to the tendency of overfitting caused by a high number of epochs. Overfitting occurs when the model becomes too focused on the training data and loses the ability to generalize to unseen data. This phenomenon results in less accurate and less consistent predictive abilities.

Test loss 0.8332937853225809 accuracy 0.7051867219917012

	precision	recall	f1-score	support
netral	0.55	0.49	0.52	1261
positif	0.65	0.63	0.64	676
negatif	0.78	0.82	0.80	2883
accuracy			0.71	4820
macro avg	0.66	0.65	0.65	4820
weighted avg	0.70	0.71	0.70	4820

Figure 4. The accuracy rates of the experiments 1

Test loss 1.9955765079608223 accuracy 0.7211618257261411

	precision	recall	f1-score	support
Netral	0.57	0.53	0.55	1261
Positif	0.66	0.62	0.64	676
Negatif	0.79	0.83	0.81	2883
accuracy			0.72	4820
macro avg	0.68	0.66	0.67	4820
weighted avg	0.72	0.72	0.72	4820

Figure 5. The accuracy rates of the experiments 2

Test loss 0.8364062545117953 accuracy 0.7197095435684647

	precision	recall	f1-score	support
netral	0.59	0.50	0.54	1261
positif	0.64	0.64	0.64	676
negatif	0.78	0.84	0.81	2883
accuracy			0.72	4820
macro avg	0.67	0.66	0.66	4820
weighted avg	0.71	0.72	0.71	4820

Figure 6. The accuracy rates of the experiments 3

In this study, a comparison was made between the BERT Uncased model, which is the base model of Indo-BERT and has excellent performance in sentiment analysis and classification. For comparison, experiments were conducted using the same data and parameters with the BERT model. Figure 7 shows the results obtained from the BERT model, with an accuracy rate of 69%, which is lower than the accuracy obtained by the Indo-BERT model, as shown in Figure 5, with an accuracy rate of 72%.

Test loss 1.4093390743414693 accuracy 0.6929460580912863

	precision	recall	f1-score	support
Netral	0.54	0.51	0.53	1261
Positif	0.56	0.66	0.61	676
Negatif	0.79	0.78	0.79	2883
accuracy			0.69	4820
macro avg	0.63	0.65	0.64	4820
weighted avg	0.70	0.69	0.69	4820

Figure 7. The accuracy rates of the experiments BERT

This indicates that INDOBERT works better for analyzing Indonesian sentiment. In addition, looking at the confusion matrix in Figure 9, the BERT model correctly predicted 2246 negative sentiments, 447 positive sentiments, and 647 neutral sentiments. In contrast, INDOBERT achieved 2413 negative sentiments, 430 positive sentiments, and 626 neutral sentiments, as shown in Figure 10. The training performance curve is shown in the figure 8

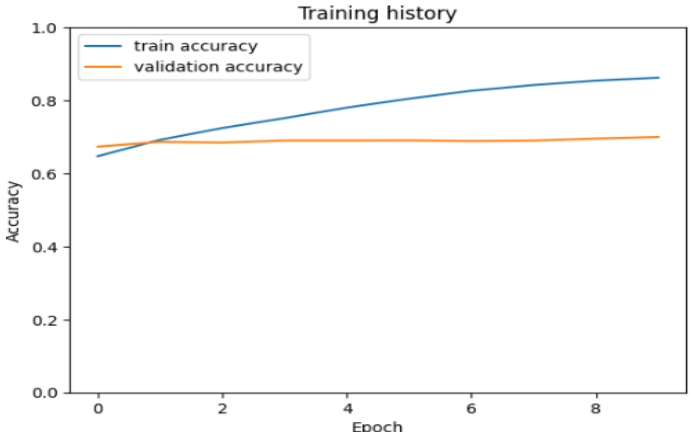

Figure 8. The training performance curve

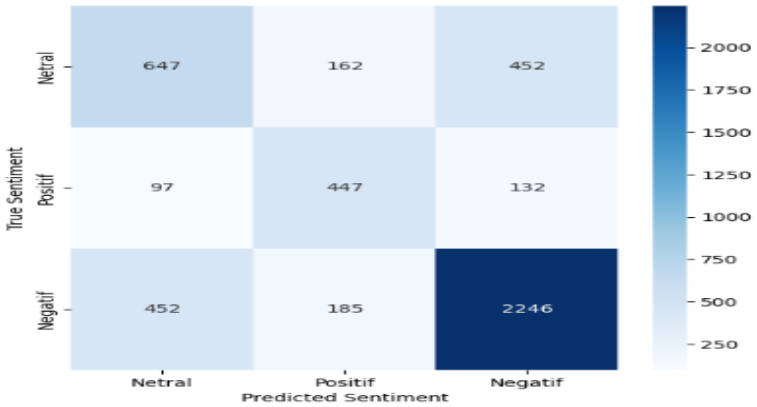

Figure 9. Confusion matrix BERT

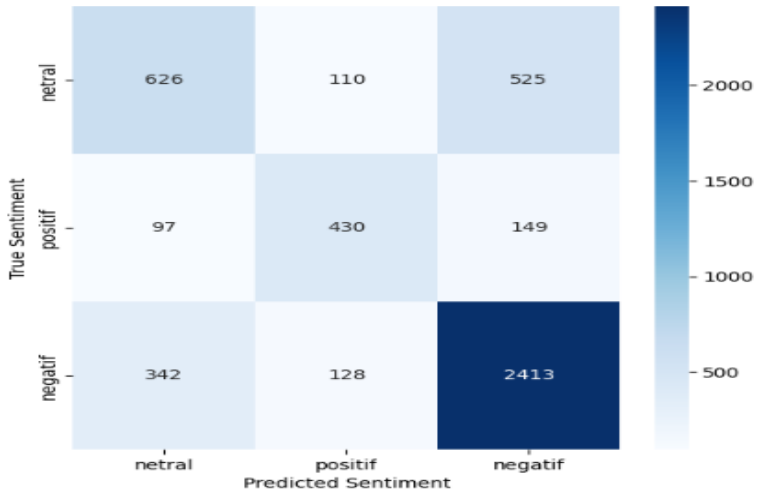

Figure 10. Confusion matrix Indo-BERT

4. CONCLUSION

Based on the testing results of sentiment analysis related to verbal abuse in the Indonesian language using the INDOBERT model, it is evident that Labelling with Supervised Learning is effective on balanced data. The labelling process using the Supervised Learning method is effective for data with a balanced distribution. This enables the model to learn sentiment patterns effectively in each category. Success in sentiment analysis heavily relies on careful data preprocessing and management steps. Evaluation of the INDOBERT model demonstrates its strong ability in sentiment analysis in the Indonesian language with an accuracy of 72%, compared to the BERT model with an accuracy of 69%. Despite some challenges in specific sentiment categories, this model consistently delivers stable performance. The model's reliability in understanding sentiment patterns is notable, as INDOBERT can discern everyday language sentiment patterns effectively. Based on the evaluation results, the INDOBERT model holds substantial potential for supporting various applications, such as public opinion monitoring, risk analysis, and social media content understanding.

REFERENCES

[1] "KEKERASAN_VERBAL_VERBAL_ABUSE_DI_ERA_DIGITAL_SEBAG.pdf."

[2] D. Wahdiyati and R. Dwi Putra, "Kekerasan Verbal dalam Konten Gaming di Youtube (Analisis Isi Kualitatif Konten Ulasan Permainan Online Maincraft dan Mobile Legend pada Akun Youtube Miuveox dan Brandonkent Everything)," *J. Indones. Sos. Teknol.*, vol. 3, no. 02, pp. 203–218, Feb. 2022, doi: 10.36418/jist.v3i2.358.

[3] N. Indrayati and L. Ph, "Gambaran Verbal Abuse Orangtua pada Anak Usia Sekolah," *J. Ilmu Keperawatan Anak*, vol. 2, no. 1, p. 9, May 2019, doi: 10.32584/jika.v2i1.220.

[4] A. Hammond, H. Smucrova, A. Tennant, Y. Prior, and M. A. M. Gignac, "Work Transitions Index -Czech version.," 2023, doi: 10.13140/RG.2.2.22376.24325.

[5] H. Weissbart and A. E. Martin, "The Structure and Statistics of Language jointly shape Cross-frequency Dynamics during Spoken Language Comprehension," Neuroscience, preprint, Oct. 2023. doi: 10.1101/2023.10.06.561087.

[6] A. Collart, "Ten years of linguistic diversity in language processing conferences".

[7] E. Smyrnova-Trybulska and E. Smyrnova-Trybulska, *E-learning in the Transformation of Education in Digital Society*, 1st ed., vol. 14. in E-learning, vol. 14. STUDIO NOA, 2022. doi: 10.34916/el.2022.14.

[8] 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*, Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.

[9] A. Lakhani, V. Upadhyay, and J. Fiaidhi, "Aspect Based Sentiment Analysis - Twitter".

[10] D. J. M. Pasaribu, K. Kusrini, and S. Sudarmawan, "Peningkatan Akurasi Klasifikasi Sentimen Ulasan Makanan Amazon dengan Bidirectional LSTM dan Bert Embedding," *Inspir. J. Teknol. Inf. Dan Komun.*, vol. 10, no. 1, Jun. 2020, doi: 10.35585/inspir.v10i1.2568.

[11] E. Y. Hidayat, R. W. Hardiansyah, and A. Affandy, "Analisis Sentimen Twitter untuk Menilai Opini Terhadap Perusahaan Publik Menggunakan Algoritma Deep Neural Network," *J. Nas. Teknol. Dan Sist. Inf.*, vol. 7, no. 2, pp. 108–118, Sep. 2021, doi: 10.25077/TEKNOSI.v7i2.2021.108-118.

[12] T. A. S. Rohmah and W. Maharani, "Personality Detection on Twitter Social Media Using IndoBERT Method," *Build. Inform. Technol. Sci. BITS*, vol. 4, no. 2, pp. 448–453, Sep. 2022, doi: 10.47065/bits.v4i2.1895.