p-ISSN: 2252-3006
e-ISSN: 2685-2411

Comprehensive Analysis of Teacher Teaching Performance Through Sentiment and POS Tagging

Putu Agung Ananta Wijaya[a1], I Made Dika Anggara[b2], I Komang Hendra Trinium Jaya[a3], Gede Indrawan[a4], I Made Agus Oka Gunawan[c5]

[a]Universitas Pendidikan Ganesha, S2 Ilmu Komputer, Singaraja
[b]Universitas Pendidikan Ganesha, S2 Teknologi Pendidikan, Singaraja
[c]Universitas Tabanan, Tabanan
e-mail: [1]putu.agung.ananta@gmail.com, [2]dika.tugas@gmail.com, [3]komangjaya453@gmail.com, [4]gindrawan@undiksha.ac.id, [5]agusokagunawan@gmail.com

Abstrak

Penelitian ini bertujuan untuk mengembangkan sistem analisis sentimen dan POS (Part of Speech) pada asesmen guru oleh siswa dalam bahasa Indonesia. Sistem ini mengklasifikasikan sentimen asesmen menjadi kelas positif dan negatif serta mengidentifikasi aspek pembahasan asesmen. Hasil pengujian menunjukkan bahwa model w11wo/indonesian-roberta-base-sentiment-classifier memberikan performa terbaik dengan akurasi 0,87, serta precision, recall, dan F1-Score tertinggi untuk kelas positif. Model crypter70/IndoBERT-Sentiment-Analysis menempati posisi kedua dengan akurasi 0,84, namun kurang optimal dalam mendeteksi sentimen negatif. Sementara itu, model mdhugol/indonesia-bert-sentiment-classification menunjukkan performa paling rendah dengan akurasi 0,80, terutama dalam prediksi sentimen negatif.

Kata kunci: *analisis sentimen, pos tagging, asesmen guru*

Abstract

This study aims to develop a sentiment analysis and POS (Part of Speech) system for student assessments of teachers in Indonesia. The system classifies assessment sentiment into positive and negative classes while identifying the topic of discussion. Testing results show that the w11wo/indonesian-roberta-base-sentiment-classifier model provides the best performance, achieving an accuracy of 0.87, with the highest precision, recall, and F1-Score for the positive class. The crypter70/IndoBERT-Sentiment-Analysis model ranks second with an accuracy of 0.84 but performs less optimally in detecting negative sentiment. Meanwhile, the mdhugol/indonesia-bert-sentiment classification model has the lowest performance with an accuracy of 0.80, particularly in predicting negative sentiment.

Keywords: *sentiment analysis, pos tagging, teacher assessment*

1. Introduction

Evaluation is an important component in education, parallel to the learning process itself. Evaluation plays a role in collecting, analyzing, and interpreting information to determine the achievement of learning objectives by students. In addition to helping teachers design more effective teaching strategies, evaluation also functions as a reflection tool for students to assess their understanding of the material being taught. A good evaluation system will be able to provide an overview of the quality of learning which in turn will be able to help teachers plan learning strategies[1]. However, traditional evaluation methods often have limitations because they only focus on the final results, such as grades or scores, without providing in-depth feedback regarding the actual learning process. This makes it difficult for teachers and students to identify specific aspects that require attention.

In education, pedagogy plays a central role as it encompasses the methods and principles guiding the teaching and learning process. Pedagogy emphasizes not only what is taught but also how it is taught, considering factors such as classroom interactions, student motivation, and the adaptation of teaching approaches to meet diverse learner needs[2]. This foundational concept underlines the importance of evaluation as a tool to improve the pedagogical process by providing actionable insights[3].

In the context of modern education, sentiment analysis has emerged as an innovative approach to assessing teacher performance and the learning process. It uses natural language processing (NLP) techniques to evaluate verbal or written feedback from students, such as surveys, essays, or comments on social media. In this way, educators can understand how students perceive their learning experience, including interactions with teachers and their responses to the material being taught. Compared to traditional, static evaluation methods, sentiment analysis provides more dynamic and real-time insights, allowing teachers to adjust teaching strategies according to students' needs.

Sentiment analysis has been widely used in various sectors, such as e-commerce and social media, to assess public opinion on products and services. In e-commerce, sentiment analysis helps companies understand product reviews and customer satisfaction. Research [4], [5], [6] shows how sentiment analysis combined with machine learning algorithms can show the demographics of a product. This information can then be used by business owners to determine the next business steps.

Social media is a rich data field in public sentiment. Recent research by [7], [8], [9] states that sentiment analysis can be used to understand public opinion on a variety of topics, from politics to commercial products. Data from Twitter, Facebook, and Instagram are analyzed to assess user reactions to emerging issues.

Several studies in education [10], [11], [12] show the effectiveness of sentiment analysis in evaluating teacher performance. However, research that specifically combines sentiment analysis with Part-of-speech (POS) Tagging techniques is still very limited. POS Tagging-based aspect detection has the potential to enrich sentiment analysis by identifying words related to key aspects of learning, such as material delivery, student motivation, and classroom interactions. Using this technique, teachers can gain more specific insights into areas that need improvement and aspects of learning that are already effective.

This research gap opens up opportunities for further development in the field of educational evaluation, especially in combining sentiment analysis and POS Tagging. With this technique, student feedback can be automatically categorized into relevant aspects of learning, providing a clearer picture of the student experience and areas that need improvement. The application of this method allows for richer pedagogical evaluation and focuses on continuous improvement of the learning process.

2. Research Method / Proposed Method

In this study, we propose a Telegram Bot-based sentiment analysis system that allows students to provide feedback to teachers after the learning process. This system is designed to collect sentiment data in Indonesian, which is then analyzed using 3 pre-trained models specifically for the case of sentiment analysis in Indonesian. This system consists of 4 main processes, namely: data collection, data preprocessing, sentiment classification, and POS tagging. Details of the processes that occur in the system can be seen in Figure 1.

p-ISSN: 2252-3006
e-ISSN: 2685-2411

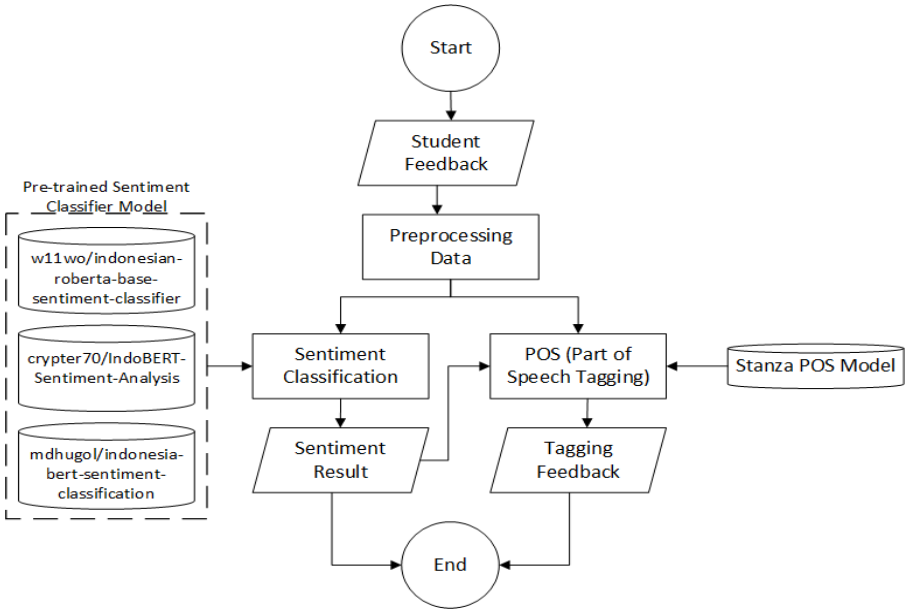

Figure 1. Flowchart System

In addition, we use the Part-of-Speech (POS) Tagging method to extract important aspects of student feedback, such as interaction, material delivery, classroom management, and learning reflection. With this approach, we can identify positive or negative sentiments related to each of these aspects.

The analysis process begins by collecting text data from the Telegram Bot, which is then preprocessed as in Figure 2 to remove irrelevant text elements. Next, a pre-trained model is used to classify student sentiment. The system then applies POS Tagging to extract pedagogical aspects from the text, allowing teachers to understand the feedback in more detail. With this aspect detection, the sentiment analysis results not only indicate whether the feedback is positive or negative but also provide information on specific aspects that need to be improved or maintained, helping teachers in a deeper learning reflection process.

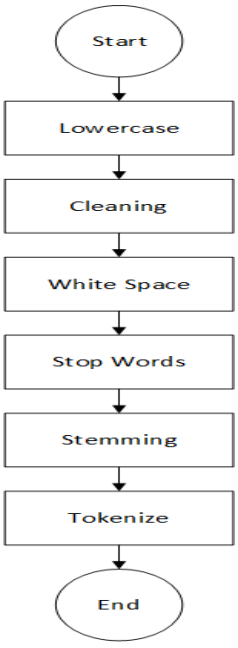

Figure 2. Preprocessing Data

p-ISSN: 2252-3006
e-ISSN: 2685-2411

3. Literature Study

3.1 Teacher Evaluation

Evaluation conducted by teachers on students is an important process in learning to determine the extent to which instructional objectives are achieved. Teachers use various evaluation methods, both tests and non-tests, to assess students' ability to understand the material. This evaluation is not only to assess knowledge but also serves as a means to provide feedback to students so that they can improve and enhance their learning outcomes. Appropriate evaluation also helps in identifying students' learning difficulties so that teachers can provide more appropriate assistance. Evaluations conducted by teachers provide quantitative values that are the basis for decision-making about student achievement[13]. This evaluation also aims to determine the effectiveness of the teaching methods used, and whether the curriculum implemented has succeeded in achieving learning objectives. Teachers can use this evaluation to measure the effectiveness of teaching methods, provide appropriate placement for students, and determine their graduation based on predetermined criteria.

3.2 Sentiment Analysis

Sentiment analysis is a technique in Natural Language Processing (NLP) that aims to identify, extract, and categorize opinions or emotions contained in text. This computational study includes people's perceptions, judgments, behaviors, and feelings towards things, people, events, issues, and their qualities[14]. This technique is commonly used to analyze responses in various domains, such as product reviews, customer surveys, or social media, to understand users' attitudes toward a topic

3.3 Part of Speech Tagging

Part of Speech Tagging is a technique in Natural Language Processing that is used to tag each word in a text with grammatical categories, such as nouns, verbs, adjectives, and so on. This process is very important for understanding the sentence structure and context of words in a text. In educational applications, POS Tagging can be used to detect keywords related to key aspects of learning, such as interactions between teachers and students, delivery of materials, and student motivation[15]. By using POS Tagging, student feedback can be analyzed in more detail and depth.

3.4 Confusion Matrix

Confusion Matrix is a tool used to evaluate the performance of a classification model in machine learning. This matrix measures how well the model predicts data into the correct category and includes four main elements: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN)[16]. Confusion Matrix provides the basis for calculating other evaluation metrics, such as accuracy, precision, recall, and F1 score, which help assess the reliability of the model.

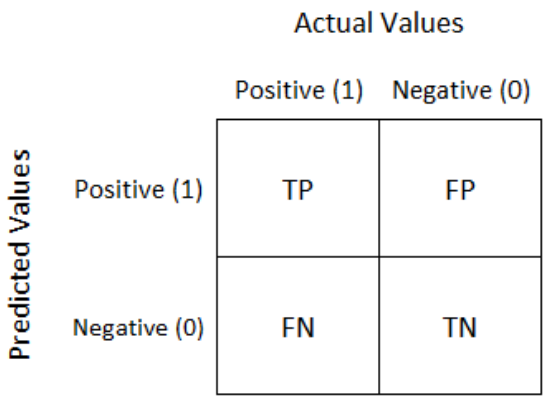

Figure 3. Confusion Matrix Table

p-ISSN: 2252-3006
e-ISSN: 2685-2411

Accuracy: The proportion of correct predictions out of all predictions.

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

Precision: The proportion of correct positive predictions out of all positive predictions.

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

Recall (Sensitivity or True Positive Rate): The proportion of positive data that is correctly detected.

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

F1 Score: The harmonic mean of precision and recall.

$$Recall = 2 \times \frac{Precision \times Recall}{Precision + Recall} \dots (4)$$

4. Result and Discussion

4.1. Collecting Student Feedback

In this study, data were collected through the Telegram Bot facility, where 53 SMAN 2 Denpasar students provided text feedback on the performance of three teachers. Each student provided their assessment through a chat sent directly to a Telegram bot designed to collect their responses. The total number of data collected was 159 chats, consisting of feedback on various aspects of teaching.

Pre-trained model was utilized in this study, eliminating the need for the training process from scratch. Pre-trained models, such as BERT[17], have already been trained on extensive datasets, allowing them to capture complex patterns and representations effectively. The decision not to perform training was driven by several factors: first, training a model from scratch requires substantial computational resources, including powerful hardware and significant time investment. Second, pre-trained models have demonstrated excellent performance across a wide range of tasks through transfer learning, enabling their adaptation to specific domains with minimal fine-tuning. Finally, leveraging pre-trained models reduces the dependency on large labeled datasets, which are often costly and time-consuming to prepare, as these models already generalize well from their extensive training on diverse corpora. These advantages make pre-trained models a practical and efficient choice for achieving high performance without the need for a resource-intensive training process.

As a result, the 159 data points collected for this study were used exclusively as testing data across three pre-trained models. This approach allowed for a comprehensive evaluation of the models' performance in processing the data without requiring a dedicated training phase. By employing pre-trained models, the study focused on assessing their ability to generalize and adapt effectively to the given dataset while maximizing resource efficiency and minimizing effort on dataset preparation.

Figure 4. Assessment Input Process

p-ISSN: 2252-3006
e-ISSN: 2685-2411

During the data collection phase, it was important to ensure that students could easily access the bot and provide their feedback without any hassle. Therefore, the bot was designed with a simple and easy-to-use interface as shown in Figure 4. Each student feedback includes an opinion expressed in text form which is then used in sentiment analysis. The bot also facilitates the storage and grouping of feedback based on the identity of the teacher being assessed. In addition, the data collection process ensured that all students were allowed to provide anonymous feedback, thus encouraging students to be more honest and open in providing their assessments.

4.2. Data Preprocessing

Once the feedback is received, the next stage is data preprocessing. This process is very important to ensure that the text data obtained in raw format is converted into a format that is ready for analysis. The preprocessing steps include:

- Lowercasing: Changes all text to lowercase for consistency.
- Tokenization: Breaking text into individual words or tokens.
- Stopwords Removal: Removes words that do not provide significant value in the analysis, such as “and”, “or”, “or in”.
- Stemming/Lemmatization: Changing words to their base form to reduce unnecessary word variations.

4.3. Sentiment Classification

After the data goes through the preprocessing stage, the sentiment classification process is carried out. In the diagram, there are three pre-trained models used for sentiment classification on student feedback:

- w11wo/indonesian-roberta-base-sentiment-classifier

  This model is based on RoBERTa, a variant of the transformer known for its natural language processing capabilities. RoBERTa is a development of the BERT model, with a focus on improving performance using better pretraining techniques. From the classification results, the precision for the negative class is 0.67 while for the positive class, it is 0.94. Recall for the negative class is 0.79 and 0.9 for the positive class. The F1 calculation for the negative class is 0.72, and the positive class has an F1-Score of 0.92. And for the accuracy of class recognition, it is 87%

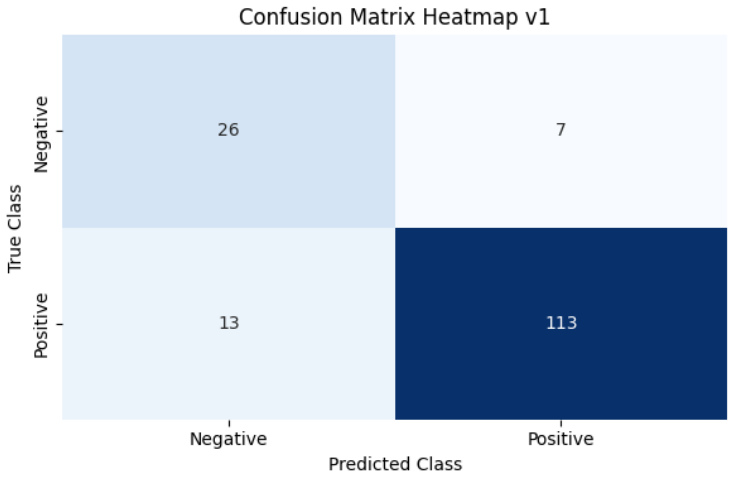

Figure 5. Indonesian RoBERTa Base Sentiment Classifier Confusion Matrix

- crypter70/IndoBERT-Sentiment-Analysis

  This model is based on IndoBERT, a BERT model that has been trained using data in Indonesian. IndoBERT is a modified version of BERT adapted for Indonesians. From the classification results, the precision for the negative class is 0.61 while for the

p-ISSN: 2252-3006
e-ISSN: 2685-2411

positive class, it is 0.89. The recall calculation for the negative class is 0.58 and 0.9 for the positive class. The F1 calculation for the negative class is 0.59, and the positive class has an F1-Score of 0.9. And for the accuracy of class recognition, it is 84%.

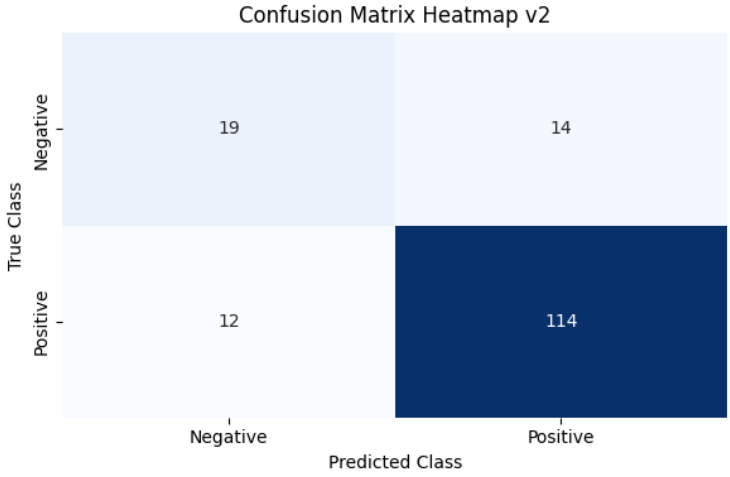

Figure 6. IndoBERT Sentiment Analysis Confusion Matrix

- mdhugol/indonesia-bert-sentiment-classification

  This model is based on BERT (Bidirectional Encoder Representations from Transformers), which is adapted for the task of sentiment classification in Indonesian. In the third experiment, the precision value was obtained for the negative class 0.51 while for the positive class, it was 0.89. The recall calculation for the negative class is 0.58 and 0.86 for the positive class. The harmonic value of precision and recall or F1 for the negative class is 0.54, and the positive class has an F1-Score of 0.87. And for the ratio of correct predictions or accuracy of class recognition, it is 80%

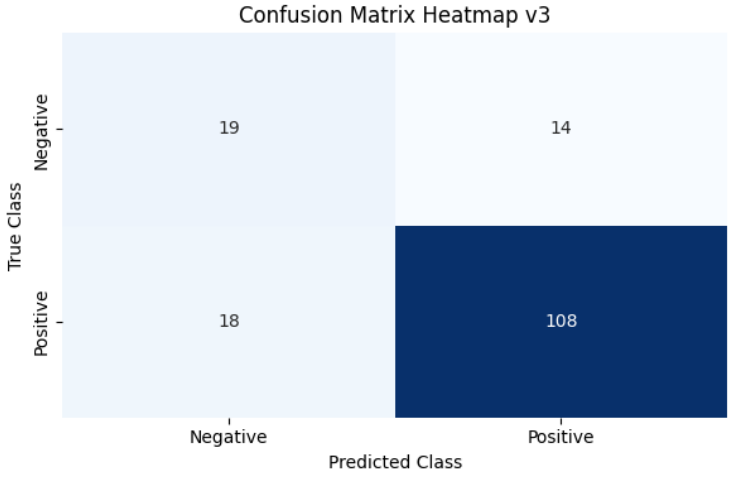

Figure 7. Indonesia BERT Sentiment Classification Confusion Matrix

These three models are pre-trained specifically for Indonesians. They are used to analyze text and categorize sentiment into several classes, such as positive, negative, or neutral, based on student responses. The output of this stage is the sentiment result, which contains the classification of each feedback.

4.4. POS Tagging (part of Speech Tagging)

The next step is Part of Speech (POS) Tagging, which aims to extract aspects of the feedback text. In this process, the POS model from Stanza POS Model is used, an NLP toolkit that can detect word types such as nouns, adjectives, verbs, etc. With this POS information, we

p-ISSN: 2252-3006
e-ISSN: 2685-2411

can tag feedback based on certain relevant aspects, such as teacher interaction, learning materials, or classroom management.

4.5. Tagging Feedback

After the POS tagging process, each student's feedback will be tagged based on the aspects found. This allows for a deeper analysis of what aspects are assessed by students, both from a positive and negative perspective. For example, if there are adjectives in the feedback that lead to the quality of interaction or delivery of material, the system will mark that section as an important aspect to pay attention to.

The Telegram Bot-based teacher assessment system can display the results of each teacher's assessment analysis using the "/graph_guru" command. Each analysis will display 4 figures, namely: sentiment distribution, negative aspect word cloud, positive aspect word cloud, and aspect evaluation based on sentiment.

Figures 8 - 11 are examples of graphs generated by the system after completing the analysis of student feedback. From this graph, we can see that Positive Sentiment dominates at 75.5% (120 instances), while Negative Sentiment is only 24.5% (39 instances). This shows that most students give positive evaluations of teachers, which can be interpreted as the teaching and pedagogical approaches used by teachers are considered effective by the majority of students. The material aspect is the most frequently mentioned aspect, with more positive sentiment than negative. This could indicate that students appreciate the delivery of the material, but there are still some criticisms or aspects that need improvement.

These results were achieved using the w11wo/indonesian-roberta-base-sentiment-classifier model, which demonstrated the best performance in sentiment classification for this task.

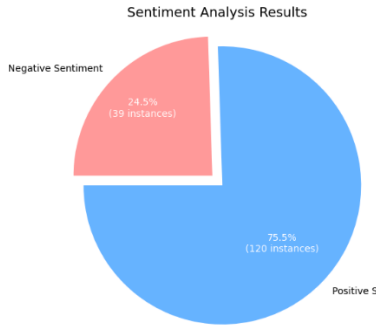

Figure 8. Sentiment Distribution

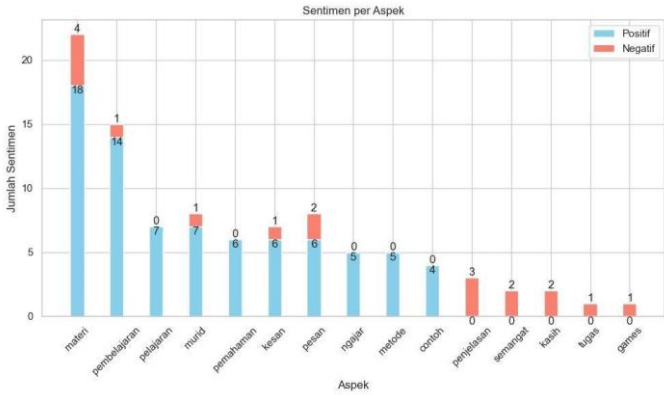

Figure 9. Aspect Evaluation Based on Sentiment

From the negative sentiment of Word Cloud, words such as explanation, enthusiasm, love, and task emerged. This could indicate criticism of the explanation given, lack of motivation, or tasks that were considered less appropriate. These aspects could be areas that require more attention in teacher pedagogical evaluation.

p-ISSN: 2252-3006
e-ISSN: 2685-2411

Figure 10. Word Cloud of Positive Aspect

Figure 11. Word Cloud of Negative Aspect

5. Conclusion

Based on the results of the sentiment classification that has been carried out, the w11wo/indonesian-roberta-base-sentiment-classifier model provides the best performance with an accuracy of 0.87, better precision and recall in predicting positive and negative classes, and the highest F1-Score for the positive class. The crypter70/IndoBERT-Sentiment-Analysis model is in second place with an accuracy of 0.84 but has a lower performance in detecting negative classes. Meanwhile, the mdhugol/indonesia-bert-sentiment-classification model shows the lowest performance with an accuracy of 0.80, and the greatest difficulty in predicting negative classes, although it is still quite good at handling positive classes.

References

[1] Nadya Putri Mtd, Muhammad Ikhsan Butarbutar, Sri Apulina Br Sinulingga, Jelita Ramadhani Marpaung, and Rosa Marshanda Harahap, “Pentingnya Evaluasi Dalam Pembelajaran Dan Akibat Memanipulasinya,” *Dewantara J. Pendidik. Sos. Hum.*, vol. 2, no. 1, pp. 249–261, Mar. 2023, doi: 10.30640/dewantara.v2i1.722.

[2] A. S. Munna and A. Kalam, “Teaching and learning process to enhance teaching effectiveness: a literature review”.

[3] L. Davies, D. Newton, and L. Newton, “Teachers’ Pedagogies and Strategies of Engagement”.

[4] Ruba Alnusyan *et al.*, “Hybrid Approach for User Reviews’ Text Analysis and Visualization: A Case Study of Amazon User Reviews,” *Int. J. Interact. Mob. Technol.*, vol. 16, no. 08, pp. 79–93, Apr. 2022, doi: 10.3991/ijim.v16i08.30169.

[5] Rafeef A. Hameed, Wael J. Abed, and Ahmed T. Sadiq, “Evaluation of Hotel Performance with Sentiment Analysis by Deep Learning Techniques,” *Int. J. Interact. Mob. Technol.*, vol. 17, no. 09, pp. 70–87, May 2023, doi: 10.3991/ijim.v17i09.38755.

[6] Mohamed Ashraf Fouad Abdelfattah *et al.*, “A Sentiment Analysis Tool for Determining the Promotional Success of Fashion Images on Instagram,” *Int. J. Interact. Mob. Technol. Ijim*, vol. 11, no. 2, pp. 66–73, Apr. 2017, doi: 10.3991/ijim.v11i2.6563.

[7] Shahzad Qaiser *et al.*, “Sentiment Analysis of Impact of Technology on Employment from Text on Twitter,” *Int. J. Interact. Mob. Technol. Ijim*, vol. 14, no. 7, pp. 88–103, May 2020, doi: 10.3991/ijim.v14i07.10600.

[8] Atif M. Gattan and Atif M. Gattan, “Deep Learning Technique of Sentiment Analysis for Twitter Database,” *Int. J. Interact. Mob. Technol.*, vol. 16, no. 01, pp. 184–193, Jan. 2022, doi: 10.3991/ijim.v16i01.27575.

[9] S. M. Abd-Alhalem, H. A. Ali, N. F. Soliman, A. D. Algarni, and H. S. Marie, “Advancing E-Commerce Authenticity: A Novel Fusion Approach Based on Deep Learning and Aspect Features for Detecting False Reviews,” *IEEE Access*, vol. 12, pp. 116055–116070, 2024, doi: 10.1109/ACCESS.2024.3435916.

[10] Senanu Okuboyejo, Senanu Okuboyejo, S. Okuboyejo, Ooreofe Koyejo, O. Koyejo, and Ooreofe Koyejo, “Examining Users’ Concerns while Using Mobile Learning Apps,” *Int. J. Interact. Mob. Technol. Ijim*, vol. 15, no. 15, pp. 47–58, Aug. 2021, doi: 10.3991/ijim.v15i15.22345.

[11] K. Nimala and R. Jebakumar, “RETRACTED ARTICLE: Sentiment topic emotion model on students feedback for educational benefits and practices,” *Behav. Inf. Technol.*, vol. 40, no. 3, pp. 311–319, Feb. 2021, doi: 10.1080/0144929X.2019.1687756.
[12] C. Pong-inwong and W. Songpan, “Sentiment analysis in teaching evaluations using sentiment phrase pattern matching (SPPM) based on association mining,” *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 8, pp. 2177–2186, Aug. 2019, doi: 10.1007/s13042-018-0800-2.
[13] S. Kusaeri, *Pengukuran dan Penilaian Pendidikan*, 1st ed., vol. 1. Jogyakarta: Graha Ilmu, 2012.
[14] B. Liu, *Sentiment Analysis and Opinion Mining*, 1st ed. US: Springer Cham, 2022.
[15] D. Jurafsky and J. H. Martin, *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd ed. England: Prentice Hall, 2020.
[16] F. Paraijun, R. N. Aziza, and D. Kuswardani, “Implementasi Algoritma Convolutional Neural Network Dalam Mengklasifikasi Kesegaran Buah Berdasarkan Citra Buah,” *KILAT*, vol. 11, no. 1, pp. 1–9, Apr. 2022, doi: 10.33322/kilat.v10i2.1458.
[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” May 24, 2019, *arXiv*: arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805.