Journal of Soft Computing Exploration
Homepage: htttps://shmpublisher.com/index.php/joscex
p-ISSN: 2746-7686
e-ISSN: 2746-0991

Comparative study of pre-trained RoBERTa sentiment models and zero-shot LLM on indonesian and english texts

Akmal Faiz Agiputra[1], Jumanto Unjung[2], Budi Prasetiyo[3], Nurrizky Arum Jatmiko[4]
[1,2,3,4]Department of Computer Science, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received December 26, 2025
Revised January 28, 2026
Accepted March 9, 2026

Keywords:

Sentiment analysis
RoBERTa
Zero-shot learning
Large language models
Multilingual text

ABSTRACT

The growth of user-generated content on social media has increased the need for effective sentiment analysis methods. Although fine-tuned transformer-based models and zero-shot large language models (LLMs) have both been applied to sentiment classification, comparisons across languages under unified evaluation settings remain limited. This study examines the trade-offs between task-specific fine-tuning and instruction-based zero-shot inference for multilingual sentiment classification. Experiments were conducted using two publicly available Twitter sentiment datasets in Indonesian and English, each annotated into three sentiment classes. Fine-tuned RoBERTa-based models were evaluated on full test sets, while all models, including a zero-shot LLM, were compared using an identical controlled subset. Performance was assessed using accuracy and macro-averaged precision, recall, and F1-score, with macro F1-score as the primary metric. The results show that fine-tuned RoBERTa-based models achieve stable and balanced performance across sentiment classes, with monolingual models consistently outperforming multilingual variants. Under controlled evaluation, zero-shot LLMs demonstrate competitive performance in English but remain less effective in Indonesian, indicating that their effectiveness is influenced by language resource availability. Overall, this study provides a controlled comparison of the strengths and limitations of fine-tuned and zero-shot approaches for multilingual sentiment classification.

Corresponding Author:

Akmal Faiz Agiputra,
Department of Computer Science,
Universitas Negeri Semarang,
Sekaran, Gunung Pati, Semarang City, Central Java 50229, Indonesia.
Email: faizagiputra@students.unnes.ac.id
https://doi.org/10.52465/joscex.v6i4.639

1. INTRODUCTION

The rapid growth of digital platforms such as social media, online forums, and product review websites has resulted in an unprecedented volume of unstructured textual data. These texts contain valuable information reflecting public opinions, attitudes, and emotions toward various topics, products, or events. However, manual analysis of such large-scale textual data is impractical and inefficient, making automated sentiment analysis an essential task in Natural Language Processing (NLP) to transform raw text into structured and actionable insights for decision-making across multiple domains [1], [2], [3], [4].

To realize this potential, the development of computational methods in NLP, particularly for automated sentiment analysis, has rapidly expanded over recent years [5], [6], [7], [8], [9]. Early approaches to sentiment analysis primarily relied on lexicon-based methods and traditional machine learning algorithms such as Support Vector Machine, Decision Tree, and Naive Bayes. While these methods were widely adopted and effective as baseline classifiers, they were heavily dependent on handcrafted features and bag-of-words representations, which limited their ability to capture contextual meaning and semantic nuances within text [10], [11]. Although sequential neural network-based methods have addressed this problem, they are hampered by inefficient sequential processing at training time and suffer from limitations in retaining information in long sentences [12], [13].

The introduction of the Transformer architecture marked a significant advancement in NLP by enabling parallel processing through self-attention mechanisms [12]. This innovation led to the emergence of pre-trained language models, notably Bidirectional Encoder Representations from Transformers (BERT), which introduced bidirectional contextual representations [14]. Further improvements were achieved by Robustly optimized BERT approach (RoBERTa), which optimized the training strategy by removing the Next Sentence Prediction objective, applying dynamic masking, and utilizing larger training corpora, resulting in stronger performance on various NLP classification tasks [15], [16]. In addition, recent Large Language Models (LLMs) offer zero-shot capabilities that enable sentiment classification without extensive task-specific fine-tuning, which is particularly advantageous in real-world scenarios involving limited or continuously evolving labeled data [17].

Several studies have reported strong performance of RoBERTa-based models for sentiment analysis across diverse domains and languages. By leveraging contextualized representations and task-specific fine-tuning, RoBERTa has been shown to effectively handle sentiment classification on both formal and informal text, indicating its robustness across different application settings [18], [19], [20], [21], [22], [23], [24].

In addition to RoBERTa, other transformer-based models, including monolingual and multilingual architectures, have been widely explored for sentiment analysis. Monolingual models tailored to specific languages have demonstrated strong capability in capturing language-specific semantics, while multilingual models provide broader cross-lingual generalization, particularly for low-resource languages. These studies suggest a trade-off between linguistic specificity and multilingual robustness in transformer-based sentiment analysis [25], [26], [27], [28], [29].

More recent studies have investigated the use of LLMs for sentiment analysis in zero-shot or instruction-based settings. Several studies indicate that LLMs can achieve competitive performance without task-specific fine-tuning, highlighting their flexibility in multilingual and dynamically evolving real-world scenarios [30], [31], [32]. However, other findings suggest that LLMs may still underperform compared to fine-tuned transformer-based models for certain tasks, such as aspect-based sentiment analysis [17].

Despite the extensive use of transformer-based models and the growing interest in zero-shot LLM approaches, existing studies generally evaluate these methods in isolation. Prior research often focuses on a single model family, a specific language, or a particular learning paradigm, making direct comparisons across different transformer-based approaches and LLMs difficult. Moreover, variations in datasets, class definitions, and evaluation protocols further limit the generalizability of conclusions drawn from previous studies.

Therefore, there remains a lack of comprehensive evaluation that jointly compares fine-tuned encoder-based transformer models and zero-shot Large Language Models under a unified experimental framework. In particular, systematic comparisons involving monolingual and multilingual transformer-based models across multiple languages for three-class sentiment classification are still limited. To address this gap, this study aims to compare the performance of fine-tuned RoBERTa-based models and a zero-shot LLM for sentiment classification in Indonesian and English, using consistent datasets and evaluation metrics. The results of this study are expected to provide clearer insights into the trade-offs between task-specific fine-tuning and zero-shot inference in multilingual sentiment analysis.

2. METHOD

Research Framework

This study follows a structured experimental workflow consisting of data collection, data preparation, model initialization and inference, and performance evaluation. The overall research framework is illustrated in Figure 1, which outlines the sequence from dataset collection to model performance evaluation and conclusion.

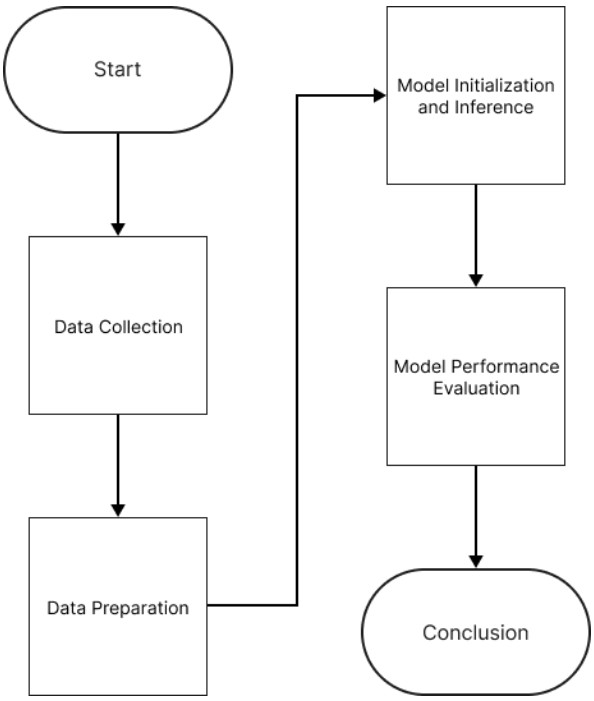

Figure 1. Research framework flow diagram

Datasets

This study utilizes two publicly available sentiment analysis datasets to represent Indonesian and English language scenarios. The Indonesian dataset is the Indonesian Twitter Sentiment Analysis Dataset–PPKM [33], while the English dataset is the Twitter US Airline Sentiment dataset [34]. The Indonesian dataset consists of 23,644 tweets, and the English dataset contains 14,640 tweets, both annotated into three sentiment classes: positive, neutral, and negative. Both datasets exhibit class imbalance, which reflects real-world sentiment distributions and motivates the use of class-wise evaluation metrics.

Data Pre-processing

No additional text normalization or manual preprocessing was applied. All input texts were processed directly using each model's built-in tokenizer. This allows the evaluation to focus on the models' native capability to handle noisy social media text.

Models

Three RoBERTa-based models were selected based on capacity equivalence (all base variants), public availability, and the absence of training data overlap with the evaluation datasets.

The w11wo/indonesian-roberta-base-sentiment-classifier represents a monolingual RoBERTa model for Indonesian and was fine-tuned using the SmSA dataset from IndoNLU. For English sentiment analysis, cardiffnlp/twitter-roberta-base-sentiment-latest was employed. This model was fine-tuned on the TweetEval benchmark, which contains Twitter-based sentiment datasets. While exact dataset overlap cannot be fully ruled out, no known direct overlap with the Twitter US Airline Sentiment dataset used in this study has been reported in the model documentation. In addition, clapAI/roberta-base-multilingual-sentiment was included as a multilingual RoBERTa-based model trained on cross-lingual sentiment datasets to evaluate generalization across both languages.

Differences in pre-training corpora and fine-tuning strategies among these publicly released models are treated as inherent characteristics rather than experimental variables. Accordingly, all models were evaluated in an as-is setting without additional re-training to ensure objective and fair comparison.

In addition to fine-tuned transformer models, an LLM, GPT-4o-mini, was evaluated in a zero-shot setting using API-based inference. Unlike RoBERTa-based models, the LLM was not fine-tuned for sentiment classification and relied solely on instruction-based prompting to infer sentiment labels. This configuration enables a direct comparison between task-specific fine-tuning and instruction-driven zero-shot inference for sentiment analysis in Indonesian and English.

To ensure methodological validity, two evaluation settings are defined in this study. First, fine-tuned RoBERTa-based models are evaluated on the full test sets to analyze large-scale performance and robustness. Second, a controlled subset evaluation is conducted in which all models, including the zero-shot LLM, are

evaluated on an identical stratified subset of samples. This separation allows both large-scale robustness analysis and fair cross-paradigm comparison.

Zero-shot LLM Evaluation

To enable a fair comparison between fine-tuned transformer models and a zero-shot LLM, all models were evaluated on an identical stratified subset with N = 500. The subset preserves the original class distribution of the datasets and was generated using a predefined random seed (42) to ensure that all models were evaluated on the same data samples and maintain reproducibility.

The zero-shot LLM (GPT-4o-mini) was evaluated using instruction-based prompting without any task-specific fine-tuning or in-context examples. The same subsets were also used to perform inference with all RoBERTa-based models, allowing direct comparison under identical data conditions.

Performance was evaluated using accuracy and macro-averaged precision, recall, and F1-score. Given the class imbalance and multi-class nature of the task, macro F1-score is used as the primary metric for comparative analysis.

Experimental Setup

All experiments were implemented in Python and executed in the Google Colab environment. RoBERTa-based models were evaluated using the Hugging Face Transformers library, while Zero-shot LLM inference was conducted via API calls within the same environment.

To ensure consistency, all RoBERTa-based models were evaluated on entire labeled datasets for each language without additional training or dataset splitting. The LLM evaluation was performed independently and did not involve dataset splitting or parameter optimization.

Evaluation Metrics

Model performance was evaluated using accuracy, macro-averaged precision, macro-averaged recall, and macro-averaged F1-score. Given the class imbalance present in both datasets, macro-averaged F1-score was used as the primary metric for performance comparison. In addition, confusion matrix analysis was used to examine class-level prediction behavior, particularly for minority sentiment classes.

For the zero-shot LLM, evaluation was conducted using the same quantitative metrics as the transformer-based models. This enables direct comparison between fine-tuned transformer models and instruction-based zero-shot inference under identical evaluation conditions. Evaluation was performed using instruction-based prompting, where the model was instructed to assign exactly one sentiment label (negative, neutral, or positive) to each input text without providing explanations.

3. RESULTS AND DISCUSSIONS

The results indicate that fine-tuned RoBERTa-based models achieve strong performance on both Indonesian and English sentiment classification tasks. On the Indonesian dataset, the monolingual RoBERTa model outperforms the multilingual model in terms of macro F1-score, highlighting the benefit of language-specific fine-tuning for informal social media text. A similar pattern is observed in the English dataset, where the English RoBERTa model demonstrates the highest overall performance. These results reflect large-scale evaluation of fine-tuned transformer models and are reported to provide contextual insight into their overall performance characteristics. A detailed comparison of model performance is presented in Table 1.

Table 1. Overall performance of fine-tuned RoBERTa-based Models

Model	Language	Accuracy	Macro Precision	Macro Recall	Macro F1-score
RoBERTa (Monolingual)	Indonesian	0.7624	0.6335	0.7640	0.6630
RoBERTa (Multilingual)	Indonesian	0.7353	0.6019	0.6993	0.6236
RoBERTa (Monolingual)	English	0.8100	0.7453	0.7825	0.7606
RoBERTa (Multilingual)	English	0.7480	0.6894	0.7367	0.7065

Class-wise Error Analysis Using Confusion Matrices

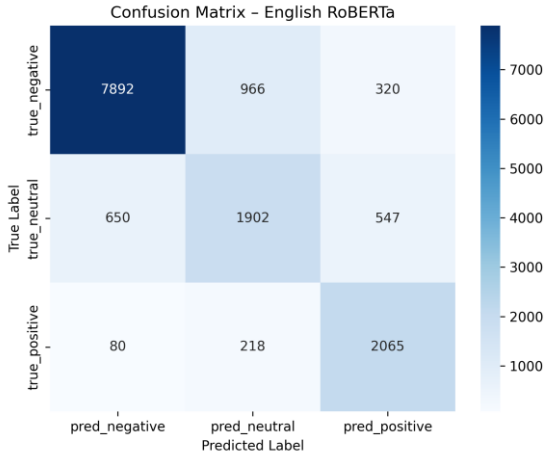

Figure 2. Confusion matrix English RoBERTa

Figure 2 presents the confusion matrix of the English RoBERTa model evaluated on the Twitter US Airline Sentiment dataset. It shows that the model achieves overall accuracy of 0.8100. However, with the macro F1-score of 0.7606, it becomes clear that the performance varies across different sentiment classes.

The negative class performs relatively well, with precision and recall reaching 0.9153 and 0.8599, respectively, indicating that most instances are correctly identified. Misclassifications occur primarily in the neutral class, while error against the positive class remains limited. This indicates that the model is effective in capturing patterns associated with this negative class.

In contrast, the neutral class shows comparatively lower performance, as reflected by its precision and recall of 0.6163 and 0.6137, respectively. A notable number of instances from this class were misclassified into negative and positive classes, possibly indicating ambiguity in the neutral sentiment. This shows that the model finds this class more challenging to distinguish.

The positive class performed moderately, with a precision of 0.7043 and a recall of 0.8739. Misclassification errors were primarily directed at neutral sentiment, which may indicate that subtle positive sentiment is still sometimes perceived as neutral.

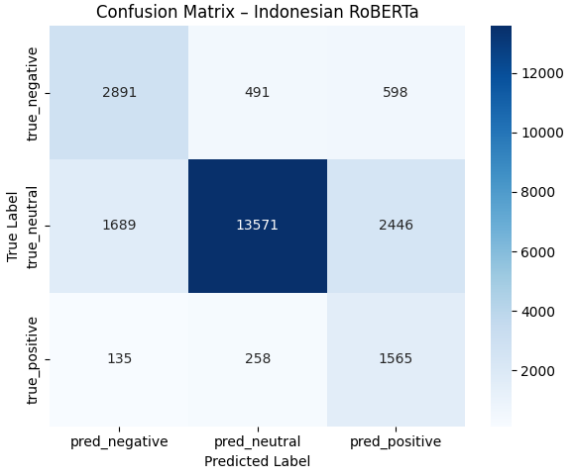

Figure 3. Confusion matrix Indonesian RoBERTa

Figure 3 illustrates the confusion matrix of the Indonesian RoBERTa model on the PPKM Twitter dataset. Compared to the previous matrix, the model achieves a slightly lower overall accuracy and macro F1-score of 0.7624 and 0.6630, respectively.

The negative class shows moderate performance, with precision at 0.6131 and a higher recall of 0.7264. This indicates that the model is able to capture most negative instances but also tends to classify neutral as negative. A portion of neutral instances are misclassified as negative, showing that neutral sentiment might have some degree of ambiguity in the sentiment.

For the neutral class, it performs decently, with a high precision of 0.9477 and a recall of 0.7665. For this label, the model shows that predictions of neutral sentiment are highly reliable, but with some labels still mixed with negative and positive labels. It indicates that some neutral expression may overlap with other sentiment categories.

Finally, for the positive class, this model had the lowest precision of 0.3396, with a fairly good recall of 0.7993. Although this model was able to capture most positive cases, it frequently classified non-positive cases as positive. A large number of negative and neutral labels were incorrectly predicted as positive, indicating that the model overpredicted the positive class.

Controlled Subset Evaluation Across Models

To enable a fair comparison across different modeling paradigms, all models were evaluated on an identical stratified subset of N = 500 samples, preserving the original class distribution of the datasets. Unlike the large-scale evaluation presented earlier, this controlled evaluation setting was specifically designed to ensure that fine-tuned transformer models and the zero-shot LLM were assessed under the same data conditions, using the same inference-only procedure and identical evaluation metrics.

Performance on the controlled subset was evaluated using accuracy and macro-averaged precision, recall, and F1-score. Given the presence of class imbalance and the multi-class nature of the task, macro F1-score is used as the primary metric for comparative analysis, as it better reflects balanced class-wise performance across sentiment categories. The results of this controlled comparison are summarized in Table 2.

Table 2. Performance comparison of fine-tuned and zero-shot LLM on identical controlled subset (N = 500)

Model	Language	Accuracy	Macro Precision	Macro Recall	Macro F1-score
RoBERTa (Monolingual)	Indonesian	0.748	0.6252	0.7417	0.6433
RoBERTa (Multilingual)	Indonesian	0.722	0.5949	0.7033	0.6152
GPT-4o-mini (Zero-shot)	Indonesian	0.682	0.5562	0.6722	0.5799
RoBERTa (Monolingual)	English	0.852	0.7987	0.8272	0.8110
RoBERTa (Multilingual)	English	0.794	0.7481	0.7772	0.7564
GPT-4o-mini (Zero-shot)	English	0.850	0.7936	0.8210	0.8056

Under identical data constraints, fine-tuned RoBERTa-based models continue to demonstrate more balanced class-wise performance compared to the zero-shot LLM, particularly on the Indonesian dataset. The monolingual and multilingual RoBERTa models achieve higher macro F1-scores than the LLM, showing that specific training to a particular language can make the model more reliable in classifying sentiment with that language compared to those trained on multiple languages.

For the English dataset, the performance gap between fine-tuned models and the zero-shot LLM is reduced, indicating that the LLM remains competitive with the monolingual model. This may be attributed to the extensive exposure of English data during pretraining, which enables the LLM to better capture sentiment patterns even without task-specific fine-tuning.

Discussions

This study compares fine-tuned RoBERTa-based models and a zero-shot LLM for multilingual sentiment classification under large-scale and controlled evaluation settings. Large-scale evaluation demonstrates that monolingual RoBERTa models achieve the strongest overall performance in both Indonesian and English datasets, highlighting the effectiveness of language-specific fine-tuning.

It can be seen that the monolingual model outperforms the multilingual model on both Indonesian and English, indicating that models trained specifically on a particular language have an advantage in this regard. Furthermore, it can also be observed in the controlled subset evaluation that the zero-shot LLM does not surpass either the monolingual or multilingual models on the Indonesian dataset. This suggests that zero-shot approaches may be less effective in lower-resource language settings.

In contrast, for the English dataset, the zero-shot LLM demonstrates competitive performance, even approaching the results of the monolingual model. This indicates that zero-shot LLMs can be more effective in high-resource language settings.

4. CONCLUSION

This study compares fine-tuned RoBERTa-based models and a zero-shot LLM for multilingual sentiment classification under both large-scale and controlled evaluation settings. The results show that the monolingual RoBERTa model consistently produces better performance than the multilingual model. In the controlled subset scenario, the monolingual model still outperforms both the multilingual model and the zero-shot LLM on the Indonesian dataset, indicating that the zero-shot approach is suboptimal for low-resource languages. In contrast, on the English dataset, the zero-shot LLM demonstrates competitive performance, approaching the monolingual model. This indicates that in addition to the monolingual model, the zero-shot LLM approach is also effective in high-resource languages. Overall, the monolingual model remains a more stable choice for multi-class sentiment analysis, followed by the zero-shot LLM which can be an alternative by considering the characteristics of the target language. Future research may explore approaches that combine the strengths of both paradigms, such as prompt optimization or lightweight fine-tuning, to improve the robustness of sentiment analysis systems in multilingual contexts.

DECLARATION OF AI USE

This manuscript was prepared with the assistance of AI-assisted tools, including ChatGPT (OpenAI), Gemini (Google), Claude (Anthropic), and Perplexity AI. These tools were used to support language refinement, structural organization, and drafting assistance.

All experimental design, data processing, model evaluation, and result analysis were conducted and critically reviewed by the authors. The authors take full responsibility for the accuracy, integrity, and originality of the content presented in this work.

REFERENCES

[1] K. L. Tan, C. P. Lee, K. M. Lim, and K. S. M. Anbananthen, "Sentiment Analysis With Ensemble Hybrid Deep Learning Model," *IEEE Access*, vol. 10, pp. 103694–103704, 2022, doi: 10.1109/ACCESS.2022.3210182.

[2] L. Yang, Y. Li, J. Wang, and R. Sherratt, "Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning," *IEEE Access*, vol. 8, p. 1, Jan. 2020, doi: 10.1109/ACCESS.2020.2969854.

[3] K. L. Tan, C. P. Lee, and K. M. Lim, "A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research," *Appl. Sci.*, vol. 13, no. 7, 2023, doi: 10.3390/app13074550.

[4] H. M. U. Ali, Q. Farooq, A. Imran, and K. El Hindi, "A systematic literature review on sentiment analysis techniques, challenges, and future trends," *Knowl. Inf. Syst.*, vol. 67, no. 5, pp. 3967–4034, 2025, doi: 10.1007/s10115-025-02365-x.

[5] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowledge-Based Syst.*, vol. 226, p. 107134, 2021, doi: https://doi.org/10.1016/j.knosys.2021.107134.

[6] J. Cui, Z. Wang, S.-B. Ho, and E. Cambria, "Survey on sentiment analysis: evolution of research methods and topics," *Artif. Intell. Rev.*, vol. 56, no. 8, pp. 8469–8510, 2023, doi: 10.1007/s10462-022-10386-z.

[7] N. V. Babu and E. G. M. Kanaga, "Sentiment Analysis in Social Media Data for Depression Detection Using Artificial Intelligence: A Review.," *SN Comput. Sci.*, vol. 3, no. 1, p. 74, 2022, doi: 10.1007/s42979-021-00958-1.

[8] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges.," *Multimed. Tools Appl.*, vol. 82, no. 3, pp. 3713–3744, 2023, doi: 10.1007/s11042-022-13428-4.

[9] Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: A systematic literature review," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 36, no. 4, p. 102048, 2024, doi: https://doi.org/10.1016/j.jksuci.2024.102048.

[10] T. Islam *et al.*, "Lexicon and Deep Learning-Based Approaches in Sentiment Analysis on Short Texts," *J. Comput. Commun.*, vol. 12, no. 1, 2024, doi: jcc.2024.121002.

[11] Q. Li *et al.*, "A Survey on Text Classification: From Traditional to Deep Learning," *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 2, Apr. 2022, doi: 10.1145/3495162.

[12] A. Vaswani *et al.*, "Attention Is All You Need," Jun. 2017.

[13] N. Sharma, A. B. M. S. Ali, and A. Kabir, "A review of sentiment analysis: tasks, applications, and deep learning techniques," *Int. J. Data Sci. Anal.*, vol. 19, pp. 351–388, Jul. 2024, doi: 10.1007/s41060-024-00594-x.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.04805

[15] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019, doi: https://doi.org/10.48550/arXiv.1907.11692.

[16] F. Barbieri, J. Camacho-Collados, L. Espinosa-Anke, and L. Neves, *TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification*. 2020. doi: 10.18653/v1/2020.findings-emnlp.148.

[17] C. Wu, B. Ma, Z. Zhang, N. Deng, Y. He, and Y. Xue, *Evaluating Zero-Shot Multilingual Aspect-Based Sentiment Analysis with Large Language Models*. 2024. doi: 10.48550/arXiv.2412.12564.

[18] M. E. Chatzimina, H. A. Papadaki, C. Pontikoglou, and M. Tsiknakis, "A Comparative Sentiment Analysis of Greek Clinical Conversations Using BERT, RoBERTa, GPT-2, and XLNet," *Bioengineering*, vol. 11, no. 6, 2024, doi: 10.3390/bioengineering11060521.

[19] H. U. Khan, A. Naz, F. K. Alarfaj, and N. Almusallam, "Analyzing student mental health with RoBERTa-Large: a sentiment analysis and data analytics approach," *Front. Big Data*, vol. Volume 8-2025, 2025, doi: 10.3389/fdata.2025.1615788.

[20] B. Paneru, B. Thapa, and B. Paneru, "Sentiment Analysis of Movie Reviews: A Flask Application Using CNN with RoBERTa Embeddings," *Syst. Soft Comput.*, 2025, [Online]. Available: https://api.semanticscholar.org/CorpusID:275661118

[21] B. Setiadi, E. Purwanto, and H. Permatasari, "Optimisasi Klasifikasi Sentimen Pada Review Hotel Bahasa Inggris Dengan Model Roberta Twitter ," *SINTECH (Science Inf. Technol. J.*, vol. 7, no. 2 SE-Articles, pp. 70–79, Aug. 2024, doi: 10.31598/sintechjournal.v7i2.1547.

[22] A. Jaya, "Analisis Sentimen Pandangan Public Profesi PNS (Pegawai Negeri Sipil) dari Twiter menerapkan indonesian Roberta

Base Sentiment Classifier," *Indones. J. Data Sci.*, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:259402547

[23] Z. Maryam, F. Rehman, K. Tariq, U. Ashraf, M. S. Shakil, and M. Yousif, "Sentiment Analysis on Social Media Posts Using Roberta: A Deep Learning Approach For Text Classification," *J. Comput. Biomed. Informatics*, vol. 9, no. 1, 2025.

[24] U. Sirisha and B. S. Chandana, "Aspect based Sentiment & Emotion Analysis with ROBERTa, LSTM," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 11, 2022, doi: 10.14569/IJACSA.2022.0131189.

[25] A. Maretta and A. Meiriza, "Aspect-Based Sentiment Analysis of Hospital Service Reviews Using Fine-Tuned IndoBERT," *J. Appl. Informatics Comput.*, vol. 9, pp. 2541–2551, Oct. 2025, doi: 10.30871/jaic.v9i5.10765.

[26] N. Nurhasiyah, R. Dwiyansaputra, S. I. Murpratiwi, and A. Aranta, "Analisis sentimen pengguna platform media sosial X pada topik pemilihan presiden 2024 menggunakan perbandingan model monolingual dan multilingual BERT," *J. Mhs. Tek. Inform.*, vol. 1, 9AD.

[27] Ardiansyah, A. Widagdo, K. Qodri, F. Saputro, and N. Putri, "Analisis sentimen terhadap pelayanan Kesehatan berdasarkan ulasan Google Maps menggunakan BERT," *J. FASILKOM*, vol. 13, pp. 326–333, Sep. 2023, doi: 10.37859/jf.v13i02.5170.

[28] M. Paramarta, R. Dwiyansaputra, and R. Rassy, "PERFORMANCE ANALYSIS OF MULTILINGUAL AND MONOLINGUAL MODELS IN PREDICTING INDONESIAN LANGUAGE EMOTION USING TWITTER DATASET," *J. Teknol. Informasi, Komputer, dan Apl. (JTIKA)*, vol. 7, pp. 237–246, Sep. 2025, doi: 10.29303/jtika.v7i2.482.

[29] Khen Dedes, Fatimatuzzahra, M. Hermansyah, A. B. Setiawan, R. P. Pradana, and A. F. M. Harvyanti, "BERT Sentimen: Fine-Tuning Multibahasa untuk Ulasan Bahasa Indonesia," *J. Komput. Teknol. Inf. Sist. Komput.*, vol. 4, no. 2 SE-Articles, pp. 1080–1084, Sep. 2025, doi: 10.62712/juktisi.v4i2.585.

[30] Z. Wang, Q. Xie, Z. Ding, Y. Feng, and R. Xia, *Is ChatGPT a Good Sentiment Analyzer? A Preliminary Study*. 2023. doi: 10.48550/arXiv.2304.04339.

[31] A. Haza Nasution, A. Onan, Y. Murakami, W. Monika, and A. Hanafiah, "Benchmarking Open-Source Large Language Models for Sentiment and Emotion Classification in Indonesian Tweets," *IEEE Access*, vol. PP, p. 1, Jan. 2025, doi: 10.1109/ACCESS.2025.3574629.

[32] I. Muhammad and M. Rospocher, "On Assessing the Performance of LLMs for Target-Level Sentiment Analysis in Financial News Headlines," *Algorithms*, vol. 18, no. 1, 2025, doi: 10.3390/a18010046.

[33] A. Widiarta, "Indonesian Twitter Sentiment Analysis Dataset-PPKM," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/anggapurnama/twitter-dataset-ppkm

[34] CrowdFlower, "Twitter US Airline Sentiment," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment

.