Journal of Telecommunication, Electronic and Computer Engineering
ISSN: 2180 – 1843 e-ISSN: 2289-8131 Vol. 17 No. 1 (2025) 29-33
jtec.utem.edu.my
DOI: https://doi.org/10.54554/jtec.2025.17.01.004

Evaluation of Transformer-Based Models for Sentiment Analysis in Bahasa Malaysia

Mohd Asyraf Zulkalnain[1], A. R. Syafeeza[1]*, Wira Hidayat Mohd Saad[1] and Shahid Rahaman[2]
[1]*Fakulti Teknologi dan Kejuruteraan Elektronik dan Komputer, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia*
[2]*Department of Computer Science, University of Buner, Pakistan*

Article Info

Article history:
Received Nov 29th, 2024
Revised Jan 19th, 2025
Accepted Mar 12th, 2025
Published Mar 19th, 2025

Index Terms:
Transformer-based models
Sentiment analysis
Bahasa Malaysia
Natural Language Processing

Abstract

This study investigates the application of advanced Transformer-based models, namely BERT, DistilBERT, BERT-multilingual, ALBERT, and BERT-CNN, for sentiment analysis in Bahasa Malaysia, addressing unique challenges such as mixed-language usage and abbreviated expressions in social media text. Using the Malaya dataset to ensure linguistic diversity and domain coverage, the research incorporates robust preprocessing techniques, including synonym mapping and sentiment-aware tokenization, to enhance feature extraction. Through rigorous evaluation, BERT-CNN exhibits the best accuracy (96.3%), followed by BERT-multilingual (89.84%) and BERT (89.5%). DistilBERT and ALBERT delivered competitive performance (88.96% and 88.76%, respectively) while offering reduced computational requirements, highlighting the trade-offs between performance and efficiency. The study emphasizes optimized strategies for handling challenges in positive sentiment classification and demonstrates the efficacy of transformer architectures in nuanced sentiment detection for low-resource languages. These findings contribute to advancing Natural Language Processing (NLP) for scalable sentiment analysis across domains.

*Corresponding Author: syafeeza@utem.edu.my

I. INTRODUCTION

Natural Language Processing (NLP) has entered a transformative era with the advent of advanced transformer-based architecture for sentiment analysis. These models, including BERT, RoBERTa, and hybrid frameworks, leverage techniques such as attention mechanisms and contextual embeddings, pushing the boundaries of sentiment interpretation. Zhang et al. demonstrated that fine-tuned transformer models could outperform traditional tools by up to 35%, marking a pivotal shift in the field [1]. Subsequent advancements, such as the RoBERTa-GRU hybrid model proposed by Tan et al., further refined sentiment accuracy, achieving over 94% precision across datasets such as IMDb and Twitter [2], [3]. These developments emphasize how innovations like the attention mechanism dynamically weigh important features, enhancing contextual understanding [4].

Aspect-based sentiment analysis has also gained traction, with Sun et al. highlighting the importance of enhanced syntactic interaction for complex and fine-grained sentiment classification [5]. Language-specific enhancements have proven particularly valuable, as shown in Arabic sentiment analysis by El Karfi and Fkihi, while ensemble approaches combining multiple pre-trained models have achieved higher accuracy in diverse settings [6], [7]. Techniques such as sentiment-aware attention [8] and hybrid loss functions [9] continue to address challenges such as class imbalance and context dependency, improving model convergence and stability across varied datasets [10]. Additionally, transformer-based models have successfully applied to modern challenges, such as conversational AI and user-generated content analysis, exemplified by their use in analyzing ChatGPT-generated data [11].

While transformer-based models have significantly improved sentiment analysis for widely used languages, their application to low-resource languages such as Bahasa Malaysia remains underexplored [12], [13]. Sentiment analysis in Bahasa Malaysia presents unique challenges, including frequent mixed-language usage, colloquial expressions, and abbreviation-heavy text in social media contexts [14]. Existing techniques often lack precision in handling these nuances, particularly in detecting positive sentiments, due to data imbalances and computational inefficiencies [2].

This paper evaluates five transformer-based models for sentiment analysis in Bahasa Malaysia namely, BERT, DistilBERT, BERT-multilingual (mBERT), ALBERT, and BERT-CNN. These models were chosen for their proven capabilities in handling sentiment analysis across diverse

linguistic contexts. BERT is renowned for its bidirectional contextual understanding, while DistilBERT offers a lightweight alternative with lower computational costs, making it suitable for real-time applications [15], [16]. BERT-multilingual exhibits robust multilingual classification performance, whereas ALBERT's parameter-sharing mechanism reduces its memory footprint, making it a practical choice for resource-constrained environments [17], [18]. BERT-CNN combines BERT's contextual understanding with CNN's ability to identify local patterns, achieving superior performance in sentiment classification tasks [19].

By comparing these models on a dedicated Bahasa Malaysia dataset, this study provides insights into their accuracy, efficiency, and scalability for low-resource languages. It further explores preprocessing techniques and data augmentation strategies tailored to Bahasa Malaysia's linguistic features [20], [21]. The findings contribute to advancing NLP for low-resource languages, offering scalable solutions for sentiment analysis with implications for public feedback monitoring, business intelligence, and policymaking in Malaysia [22].

The remainder of this paper is organized as follows: Related theories and methodologies are presented in the next section, followed by results and discussion. Finally, the conclusion summarizes the overall work.

II. METHODOLOGY

This study evaluates the performance of transformer-based models, specifically BERT, DistilBERT, BERT-multilingual (mBERT), ALBERT and BERT-CNN in conducting sentiment analysis for Bahasa Malaysia. These models were fine-tuned using a carefully curated dataset of Bahasa Malaysia text social media text.

A. Dataset

The dataset, annotated into positive, negative, and neutral sentiment classes, is designed to capture the unique linguistic features of Bahasa Malaysia. The curated Malaya dataset, central to this research, encompasses mixed-language usage, colloquial expressions, and abbreviated forms, which are frequently found in social media text. Unlike conventional sentiment analysis datasets, the Malaya dataset integrates diverse text sources, ensuring a comprehensive representation of contemporary Malaysian communication styles [12]. The Malaya dataset is divided into training and testing sets with a ratio of 80:20, meaning 80% of the data is used for training, and 20% is used for testing. This dataset presents specific challenges, such as handling contextually dependent abbreviations, multilingual expressions, and imbalanced sentiment label distributions. To address these complexities, specialized preprocessing techniques and augmentation strategies were employed, establishing this dataset as a robust benchmark for sentiment analysis in low-resource languages.

Table 1
Total samples for each sentiment category

Label	Category	Data type	Total samples
Neutral	0	Train	66,563
		Test	16,641
Positive	1	Train	35,736
		Test	8,934
Negative	2	Train	46,330
		Test	11,583

Table 1 shows the total number of samples for each category. The data distribution indicates that the sentiment classes are imbalanced.

B. Preprocessing

To ensure data consistency and accuracy, special preprocessing steps were implemented, including data cleaning and tokenization, tailored to the linguistic nuances of Bahasa Malaysia. Synonym mapping and spelling correction were applied using tools from the Malaya library, following established methods in Malay sentiment analysis [14]. Data cleaning involved several text normalization operations to ensure that special characters or undesired symbols did not interfere with sentence context. This is because the special characters do not bring any meaning or context in the whole. Table 2 summarizes the text cleaning operations performed in this research.

The preprocessing steps were tailor to accommodate the linguistic nuances of Bahasa Malaysia, including the following:

- Text Cleaning: Removed non-ASCII characters, normalized special characters, and handling URLs.
- Tokenization: Employing WordPiece tokenization for compatibility with transformer models.
- Synonym Mapping: Using a manually created synonym dictionary (Table 5).
- Data Augmentation: Applying techniques like synonym replacement and random insertion to mitigate data imbalance. Synonym replacement was particularly effective in enhancing the diversity of the training data.

Table 1
Summary of text cleaning operations that are performed in this research

Text cleaning operations	Example sentence before	Example sentence after
Removes any non-ASCII characters and converts accented characters to their closest ASCII representation	*Mungkin cara beliau kelihatan sedikit cliché berbanding orang sekarang*	*Mungkin cara beliau kelihatan sedikit cliche berbanding sekarang*
Normalizes special Unidecode characters to their standard forms to ensure consistent representation	*Aku amat mencintaimu, wahai suamiku*	*Aku amat mencintaimu, wahai suamiku*
Handles URLs by replacing them with a space to remove them from the text	Saya telah melayari https://yahoo.com untuk mencari maklumat berguna	Saya telah melayari untuk mencari maklumat berguna
Splits the text into individual words and separates certain punctuation marks (such as '.', ',', '/') with spaces to treat them as separate tokens	Pasukan bola sepak dari … menunjukkan aksi / yang kurang memuaskan malam ini	[‘Pasukan’, ‘bola’, ‘sepak’, ‘dari’, ‘ ’, ‘menunjukkan’, ‘aksi’, ‘ ‘ , ‘yang’, ‘kurang’, ‘memuaskan’, ‘malam’, ‘ini’]
Removes any extra spaces and leading '@' symbols, often used in social media to mention users	*Kedai makan @cikkiah_nasilemak telah disita pagi tadi*	*Kedai makan telah disita pagi tadi*
Introduces random laughter expressions (like 'haha', 'hehe', etc.) in the text with a probability of 50%, making the text more expressive and natural	*Dia telah menari dengan penuh semangat dan baik*	*Dia telah menari dengan penuh semangat dan baik haha*

Tokenization is a process of breaking down a text or sentence into smaller units called tokens, which are then converted to IDs. Tokens typically consist of words or subwords that form the text and serve as the basic building blocks for NLP tasks. In the context of BERTt, tokens act as the standard format that BERT models can understand before training. Table 3 shows an example of tokenization. Transformers use three main types of tokenizers: WordPiece, SentencePiece and Byte-Pair Encoding (BPE). For each variation, pre-built tokenizers, such as BertTokenizer, AlbertTokenizer and DistilBertTokenizerFast are readily available.

Table 2
Example of Tokenization

Original sentence	Tokenized sentence
"Saya suka makan nasi"	["Saya", "suka", "makan", "nasi"]
"Cuaca sangat baik pagi ini"	["Cuaca", "sangat", "baik", "pagi", "ini"]
"Keputusan peperiksaan anak saya amat cemerlang"	["Keputusan","peperiksaan","anak", "saya", "amat", "cemerlang"]

Data augmentation is a technique used to expand the dataset by generating additional training samples, as highlighted by [3]. It is widely used in NLP, computer vision, and speech-related tasks, aiding in the development of robust models and improving classification performance. The Easy Data Augmented (EDA) approach, introduced in Boosting Performance on Text Classification Tasks, [21] suggests four different data augmentation techniques for text, as shown in Table 4.

Table 3
Data Augmentation strategies and their definitions

Data Augmentation Operation	Definition
Synonym Replacement (SR)	Choose random n words from sentence that are not stopwords. Replace each of the words with one of its synonyms chosen at random.
Random Insertion (RI)	Insert random word in sentence, that is not a stopword. Insert that synonym into random position.
Random Swap (RS)	Choose two words randomly in the sentence and swap their positions.
Random Deletion (RD)	Remove random words in sentence with probability 'p'.

For model optimization, synonym replacement and random insertion were applied to the dataset as data augmentation techniques. A manually created synonym dictionary was created to replace words with similar meaning in Bahasa Malaysia within the dataset. Table 5 provides an example of the synonym dictionary in Malay. The original text in the dataset was replaced with the synonym text, and the augmented data was added to the original dataset.

Table 4
Example of synonym dictionary in Malay

Original text	Synonym text
terjaga	terbangun
kuat	tabah
menderita	sengsara
mantap	padu
tolol	bangang
senang	mudah

The training process employed transfer learning with pre-trained weights from the HuggingFace library, allowing the models to efficiently adapt to the sentiment analysis task. The Adam optimizer was utilized with a learning rate scheduler, balancing convergence and mitigating overfitting. A multi-class Cross Entropy loss function was utilized for each model. For feature extraction and validation, k-fold cross-validation was applied, aligning with best practices recommendations by [7] to ensure robust model evaluation and reliability.

Experiments were conducted using Python with libraries such as PyTorch and Scikit-learn. The training was performed on Google Colab with Tesla K80 GPUs and a local GPU (GTX 4060 Ti), ensuring computational efficiency. Performance metrics included accuracy, precision, recall, F1-score, and computational efficiency. These metrics provided a comprehensive evaluation framework. They were systematically compared with traditional sentiment analysis techniques. The findings highlight the superiority of Transformer architectures in handling mixed-language and abbreviation-heavy texts in Bahasa Malaysia. Figure 1 illustrates the flowchart of the overall study.

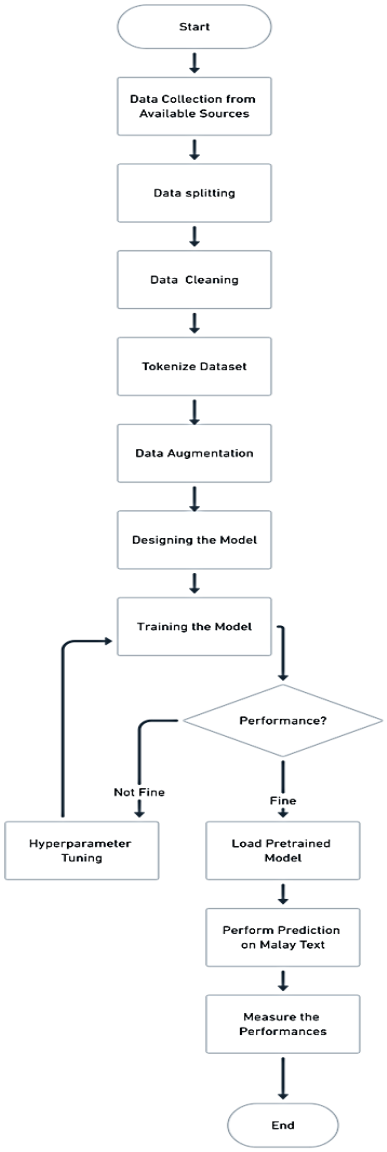

Figure 1. Flowchart of the overall work

III. RESULT AND ANALYSIS

Table 6 presents the performance of five Transformer-based models namely, BERT, DistilBERT, BERT-multilingual (mBERT), ALBERT, and BERT-CNN, evaluated on sentiment analysis tasks in Bahasa Malaysia. These models were assessed based on accuracy, F1 score, precision, and recall, with results highlighting their strengths and limitations. Among the five models, BERT-CNN achieved the highest accuracy of 96.30%, demonstrating its effectiveness in capturing contextual relationships. This model also attained an F1 score of 96.02%, precision of 95.13%, and recall of 97.03%. However, despite its superior overall performance, BERT-CNN struggled with positive sentiment classification, indicating an area for potential improvement.

The evaluation metrics were calculated as follows:

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

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

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

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

BERT-multilingual (mBERT) emerged as the second-best performer, achieving an accuracy of 89.84%, an F1 score of 89.34%, a precision of 89.37%, and a recall of 89.32%. mBERT demonstrated its strength in multilingual settings, benefiting from the diversity in its pretraining data. Similarly, the original BERT model performed competitively, with an accuracy of 89.5%, an F1 score of 91.13%, a precision of 96.7%, and a recall of 90.7%. While BERT's robust architecture effectively captured rich contextual features, it also exhibited challenges in accurately identifying positive sentiments.

DistilBERT offered a balance between efficiency and performance, achieving an accuracy of 88.96%, an F1 score of 89.73%, a precision of 88.85%, and a recall of 90.23%. As a lightweight version of BERT, DistilBERT's reduced computational requirements made it a strong candidate for resource-constrained applications, though its precision lagged slightly compared to its peers. ALBERT, designed with a lighter architecture to optimize memory usage, recorded the lowest performance metrics, with an accuracy of 88.76%, an F1 score of 88.55%, a precision of 89.66%, and a recall of 89.33%. Despite its lower performance, ALBERT remains a viable option for environments where computational efficiency is prioritized over slight reductions in accuracy.

The evaluation revealed clear trade-offs between model accuracy and computational efficiency. BERT-CNN is the most suitable model when high accuracy is critical, but its computational demands make it less practical for real-time applications. In contrast, DistilBERT and ALBERT provide efficient alternatives with acceptable performance for resource-limited scenarios. These findings emphasize the importance of selecting models based on specific use cases and computational constraints.

Overall, this study underscores the effectiveness of tailored preprocessing and data augmentation techniques in improving sentiment analysis performance for low-resource languages. The results demonstrate the potential of Transformer-based models to address linguistic nuances in Bahasa Malaysia, while also highlighting opportunities for further optimization, particularly in handling positive sentiment classifications. Future research should focus on addressing these limitations and exploring hybrid architectures that balance contextual understanding and computational efficiency.

Table 6
Performance of five transformer-based models: BERT, DistilBERT, BERT-multilingual, ALBERT and BERT-CNN

Model	Overall Accuracy (%)	F1 Score (Weighted, %)	Precision (Average, %)	Recall (Average, %)
BERT	89.5	91.13	96.7	90.7
DistilBERT	88.96	89.73	88.85	90.23
BERT-multilingual	89.84	89.34	89.37	89.32
ALBERT	88.76	88.55	89.66	89.33
BERT-CNN	96.30	96.02	95.13	97.03

IV. CONCLUSION

This study evaluated the efficacy of Transformer-based models, namely BERT, DistilBERT, BERT-multilingual, ALBERT, and BERT-CNN, for sentiment analysis in Bahasa Malaysia. Among these models, BERT-CNN achieved the highest accuracy at 96.30%, excelling in capturing contextual relationships but facing challenges with positive sentiment classification. BERT-multilingual and BERT also delivered strong performances, while DistilBERT and ALBERT demonstrated efficiency with lower computational costs but slightly reduced accuracy. These findings emphasize the trade-offs between accuracy and computational efficiency in sentiment analysis for low-resource languages.

The study highlighted the importance of tailored preprocessing techniques and data augmentation strategies in addressing linguistic nuances and imbalanced datasets. Future research should explore advanced hybrid architectures, such as combining lightweight models like DistilBERT with the contextual richness of BERT-CNN, to achieve both efficiency and accuracy. Expanding datasets to cover additional domains, including conversational AI and formal communications, will enhance the generalizability of sentiment analysis systems. Additionally, domain-specific multilingual models and the integration of sentiment-aware attention mechanisms could help address class imbalances and nuanced contexts more effectively.

Furthermore, emerging studies have introduced promising techniques in this field, such as hybrid transformer models with knowledge distillation [21], multilingual architectures for low-resource languages [17], and easy data augmentation techniques to mitigate data imbalance [20]. Incorporating these approaches into future research could significantly enhance sentiment analysis for languages like Bahasa Malaysia, paving the way for more scalable, efficient, and accurate NLP applications.

ACKNOWLEDGMENT

The authors would like to acknowledge Faculty of Electronics & Computer Technology and Engineering (FTKEK), Universiti Teknikal Malaysia Melaka (UTeM) for its support, particularly through the Machine Learning and

Signal Processing (MLSP) research group under the Centre for Telecommunication Research & Innovation (CeTRI).

CONFLICT OF INTEREST

Authors declare that there is no conflict of interests regarding the publication of the paper.

AUTHOR CONTRIBUTION

The authors confirm contribution to the paper as follows: study conception and design: Mohd Asyraf Zulkalnain; data collection: Mohd Asyraf Zulkalnain; analysis and interpretation of findings: Syafeeza Ahmad Radzi, Wira Hidayat Mohd Saad, Shahid Rahaman; draft manuscript preparation: Mohd Asyraf Zulkalnain, Syafeeza Ahmad Radzi. All authors had reviewed the findings and approved the final manuscript.

REFERENCES

[1] X. Zhang, J. Sun, Z. Huang, T. Li, and Y. Wang, "Fine-tuning transformer-based models for sentiment analysis," Natural Language Engineering, vol. 26, no. 4, pp. 456–470, 2020.

[2] M. H. Tan, L. F. Chua, and C. H. Loo, "RoBERTa-GRU hybrid model for sentiment analysis," Applied Sciences, vol. 12, no. 3, pp. 145–152, 2023.

[3] T. Liu, Y. Chen, X. Zhao, H. Zhu, and W. Zhang, "Transformer-based sentiment analysis for low-resource languages," Natural Language Processing Research, vol. 12, no. 1, pp. 45–59, 2020.

[4] A. Vaswani et al., "Attention is all you need," in Proc. of Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.

[5] Y. Sun, S. Wang, H. Li, X. Chen, and D. Chen, "Aspect-based sentiment analysis with enhanced syntactic interaction," in Proc. of EMNLP, 2020, pp. 29–40.

[6] M. El Karfi and S. Fkihi, "Ensemble approaches for Arabic sentiment analysis," Computational Intelligence and Neuroscience, vol. 2022, Art. no. 934875, 2022.

[7] S. Batra, R. Malhotra, P. Sharma, and A. Tiwari, "BERT-based sentiment analysis in software engineering," Journal of Software Engineering Research and Development, vol. 9, no. 2, pp. 15–26, 2021.

[8] X. Li, Z. Gao, M. Wang, and P. Liu, "Incorporating sentiment-aware attention into transformer models," Journal of Artificial Intelligence Research, vol. 65, pp. 303–318, 2020.

[9] Y. Huang, J. Zhou, C. Fang, S. Yang, and H. Wu, "Incorporating hybrid loss functions into sentiment analysis," IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 4, pp. 1345–1357, 2020.

[10] C. J. Yin, T. Mahmud, S. Lim, H. Ismail, and R. Tan, "Detecting sentiment in Malay short forms and internet slang using machine learning," Journal of Social Media Analytics, vol. 10, no. 1, pp. 100–110, 2021.

[11] A. Winardi, H. Kusuma, and A. Santoso, "Transformer algorithms for sentiment analysis on ChatGPT-generated tweets," Journal of Information Technology Engineering Research, vol. 18, no. 3, pp. 256–265, 2023.

[12] Z. Husein, "Malaya dataset," Accessed: Date: 1 Nov 2024, [Online.] Available: https://malaya-dataset.readthedocs.io.

[13] Q. Lu, "Fine-tuning transformer models for multilingual sentiment analysis," in Proc. of International NLP Conference, 2022, pp. 45–59.

[14] K. Chekima and R. Alfred, "RojakLex: A mixed-language lexicon for sentiment analysis," in Proc. of International Conference on Computational Linguistics, 2018, pp. 456–462.

[15] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171–4186.

[16] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2020.

[17] G. Manias, "Performance of multilingual BERT models on low-resource languages," in Proc. of International Conference on Multilingual NLP, 2023, pp. 12–18.

[18] Z. Lan et al., "ALBERT: A lite BERT for self-supervised learning of language representations," in Proc. of International Conference on Learning Representations, 2020.

[19] G. Kumar, "BERT-CNN hybrid models for enhanced sentiment analysis," Journal of Advanced Data Science, vol. 7, no. 2, pp. 85–92, 2024.

[20] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," arXiv preprint arXiv:1901.11196, 2019.

[21] Z. Gong, Z. Tan, and L. Zhang, "Hybrid transformer model for sentiment analysis using knowledge distillation and text augmentation," Frontiers in Psychology, vol. 13, pp. 1–12, 2022.

[22] M. Ismail, N. Razak, and Z. Ariffin, "A multistage sentiment classification model using Malaysian political ontology," International Journal of Advanced Computer Science and Applications, Vol. 12, No. 10, pp. 430-436, 2021.