Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9805

IndoBERTweet-Based Sentiment Analysis of AIESEC Indonesia's Instagram Comments for Brand Image Assessment

An'nur Aisyah Ghaidah Senjaya Mangkunegara[1], Apriansyah Putra[2*], Ken Ditha Tania[3], Mira Afrina[4]
[1,2,3,4]Department of Information Systems, Faculty of Computer Science, Universitas Sriwijaya, South Sumatera, Indonesia.
[1]aisyahghaidah@gmail.com, [2]apriansyah@unsri.ac.id, [3]kenya.tania@gmail.com, [4]miraafrina81@gmail.com

***Corresponding Author**

Article History:
Submitted: 16-10-2026
Accepted: 27-10-2026
Published: 02-11-2026

Keywords:
AIESEC;brand image; IndoBERTweet; Instagram; sentiment analysis; social media

Brilliance: Research of Artificial Intelligence is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

ABSTRACT

Social media platforms such as Instagram have become strategic channels through which organizations shape their brand image, and public comments on official accounts often reflect collective perceptions that are difficult to interpret manually because of their volume and informal language style. AIESEC Indonesia, the national chapter of the world's largest youth-run international non-governmental organization, relies heavily on its official Instagram account to communicate its leadership-development programs and cross-cultural exchange opportunities to Indonesian youth, yet no prior study has systematically examined how comment sentiment reflects the organization's brand image at the national level. This study aims to identify the sentiment tendencies of public comments on AIESEC Indonesia's Instagram posts, to evaluate the performance of IndoBERTweet in classifying these Indonesian-language comments through sentiment analysis, and to interpret the resulting sentiment distribution as a brand image indicator. A descriptive quantitative approach was applied, in which comments were collected through crawling, cleaned and normalized, manually labeled into positive, neutral, and negative categories, and classified using a fine-tuned IndoBERTweet model evaluated with accuracy, precision, recall, and F1-score. Results from 755 comments show 60.7% positive, 32.6% neutral, and 6.8% negative sentiment. IndoBERTweet achieved 83.3% accuracy and a 0.813 macro F1-score, outperforming the baseline by improving negative sentiment detection. Conclusively, AIESEC maintains a highly favorable brand image, and IndoBERTweet can effectively automate feedback monitoring. These findings are expected to provide an evidence-based understanding of public perception toward AIESEC Indonesia on social media.

INTRODUCTION

Social media has evolved from a mere interpersonal communication tool into a strategic instrument for shaping an organization's brand image. Instagram, as a visually driven platform with a large global user base, is widely used by both nonprofit and commercial organizations to reach audiences, communicate organizational values, and build engagement with the public. Through photo, video, and caption content, organizations not only share information about their activities but also actively construct how audiences perceive their identity and reputation.

AIESEC is a nonpolitical, nonprofit international youth organization founded in 1948 and currently active in more than 120 countries and territories, focusing on youth leadership development through cross-cultural exchange programs and social impact projects. AIESEC Indonesia, its national chapter, connects this global mission to Indonesian students and young professionals through internship placements, volunteering projects, and leadership-development programs delivered at the national and local-committee level. As an organization run entirely by students and young people, AIESEC Indonesia depends heavily on social media, particularly its official national Instagram account, to maintain visibility, credibility, and appeal among prospective members, partners, and the wider Indonesian public. Understanding how this Indonesian audience responds to content posted on the account is therefore relevant to examine scientifically, since their engagement reflects perceptions of the organization they interact with directly.

Public response to social media posts is commonly expressed through comments, which contain opinions, emotions, and evaluations, whether positive, negative, or neutral, toward the content or the organization itself. When analyzed systematically, this collection of opinions can serve as a quantitative indicator of public perception toward brand image. However, the large volume and unstructured nature of comments make manual analysis impractical, requiring a computational approach based on natural language processing (NLP), namely sentiment analysis.

A particular challenge arises when the analyzed text consists of Indonesian-language social media comments, which tend to be informal, contain code-mixing, abbreviations, and slang that general-purpose language models do not

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9805

always handle optimally. Koto, Lau, and Baldwin (2021) addressed this challenge by developing IndoBERTweet, a pretrained language model built specifically for informal Indonesian Twitter text through domain-specific vocabulary initialization. Although IndoBERTweet was pretrained for Indonesian Twitter text, its domain-specific vocabulary and ability to capture informal Indonesian social-media language make it a relevant candidate for evaluating sentiment in similarly informal Instagram comments.

Despite this potential, existing research applying IndoBERTweet has largely focused on the platform X (Twitter) and on sociopolitical topics, while its application to Instagram comments and to brand image measurement at the level of a national youth-organization chapter such as AIESEC Indonesia remains largely unexplored. This study addresses that gap by applying IndoBERTweet to classify the sentiment of public comments on AIESEC Indonesia's Instagram content, with the explicit objectives of (1) identifying the sentiment tendencies of these comments, (2) evaluating IndoBERTweet's classification performance, and (3) interpreting the results as an indicator of the organization's brand image.

LITERATURE REVIEW

Brand image refers to the perceptions and beliefs the public holds toward an entity, shaped by the strength, favorability, and uniqueness of the associations stored in their memory (Dzyabura & Peres, 2021). In the digital era, brand image is no longer shaped solely by an organization's own content (owned media) but is co-constructed by public responses on social media, such as comments and shares (earned media). Zulfiah, Fadhilah, and Lukitaningsih (2024) found that social media marketing significantly influences brand image through brand awareness, confirming that audience interaction on social platforms is a measurable component of brand perception. In a related context, a study of a coffee shop's Instagram presence demonstrated that social media culture actively contributes to the construction of a brand's public image (Utoyo & Karlina, 2023), reinforcing the relevance of using user-generated comments as an empirical indicator of public sentiment relevant to brand image on Instagram.

Sentiment analysis is a branch of NLP that identifies and classifies opinions expressed in text into positive, negative, or neutral categories. Transformer-based deep learning approaches, particularly BERT-derived models, have become the dominant method for this task because of their ability to capture bidirectional semantic context, outperforming lexicon-based and classical machine learning approaches on informal, context-dependent text (Koto et al., 2021). Ihalauw, Lapatta, Nugraha, Wirdayanti, and Lamasitudju (2025) applied IndoBERT combined with SMOTE data balancing to analyze public sentiment toward a new government's early performance on platform X, showing that transformer-based models handle the linguistic noise of social media more effectively than lexicon-based methods. Similarly, a comparative study on electric-vehicle discussions found that IndoBERTweet consistently outperformed the general-domain IndoBERT model on informal social media text, confirming the advantage of domain-specific pretraining (Dewi, 2025).

IndoBERTweet was introduced by Koto et al. (2021) as the first large-scale pretrained language model built specifically for Indonesian Twitter text, developed by extending a monolingually trained IndoBERT model with additive domain-specific vocabulary initialized through subword embedding averaging. Trained on more than 409 million tokens collected from Indonesian tweets between December 2019 and December 2020, the model achieved superior performance across seven Twitter-based evaluation datasets compared to models without domain adaptation. In an applied study, Hizkia Imanuel (2026) used IndoBERTweet with a pseudo-labeling strategy to classify public sentiment toward a government nutrition program on platform X, reporting an accuracy above 97%, which illustrates the model's strong performance on real-world, noisy Indonesian social media data. Marpaung and Tarigan (2026) further compared IndoBERT and Indonesian RoBERTa Base for sentiment classification, finding that class-weighting and hyperparameter optimization significantly affect model stability, a methodological consideration relevant to the fine-tuning procedure adopted in the present study.

Taken together, the reviewed literature shows that brand image research on Instagram has largely relied on survey-based methods, while IndoBERTweet-based sentiment analysis has mainly been applied to platform X and sociopolitical topics. The intersection of these two research streams, that is, using a domain-specific transformer model to analyze Instagram comments as an empirical indicator of a national youth-organization chapter's brand image, appears to remain underexplored, particularly in the context of Instagram comments and national youth organizations such as AIESEC Indonesia.

Table 1. Comparison of Previous Studies

Research	Method	Dataset/Topic	Result/Accuracy	Limitation/Gap
Dzyabura & Peres (2021)	Visual elicitation (online brand collages) + machine learning	4,743 collages from 1,851 respondents on 303 U.S. brands	Brand perception is an associative network; the favorability, strength, and uniqueness of associations position a brand.	Covers commercial U.S. brands; no sentiment analysis of Instagram comments.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9805

Zulfiah et al. (2024)	Survey + SEM-PLS	Social media is related to brand image	Social media marketing significantly influenced brand image.	Survey-based and does not analyze user-generated Instagram comments using sentiment analysis.
Utoyo & Karlina (2023)	Quantitative questionnaire	Tuku Coffee Shop Instagram and brand image	Examined recognition, reputation, affinity, and domain; recognition had the highest mean.	Uses questionnaire responses rather than computational analysis of Instagram comments.
Koto et al. (2021)	IndoBERTweet	Indonesian Twitter corpus (>409 million tokens); seven Twitter-based evaluation datasets.	Strong performance across seven Twitter-based datasets.	Developed for Twitter; application to Instagram comments and brand-image assessment remains outside its original scope.
Ihalauw et al. (2025)	IndoBERT + SMOTE	15,027 posts on X about government performance	87% accuracy; SMOTE was used to address class imbalance.	Uses platform X and focuses on government performance rather than organizational brand image on Instagram.
Dewi (2025)	IndoBERTweet vs. IndoBERT	Indonesian social-media text on electric vehicles	IndoBERTweet: 82.40% accuracy, F1 82.39%.	Focuses on electric-vehicle discussions rather than Instagram-based organizational brand perception.

Based on the previous studies, existing research has examined brand image through social media and has demonstrated the effectiveness of transformer-based models, particularly IndoBERTweet, for Indonesian social-media text. However, these studies have primarily relied on survey-based brand-image assessment or focused on Twitter/X and topic-specific sentiment analysis. The application of IndoBERTweet to Indonesian Instagram comments for assessing sentiment related to an organization's brand image remains underexplored. Therefore, this study applies IndoBERTweet to AIESEC Indonesia's Instagram comments to identify sentiment tendencies and examine their relevance as an empirical indicator of brand image.

METHOD

This study employed a descriptive quantitative approach using sentiment analysis based on a pretrained language model, following the methodological precedent established by Koto et al. (2021) for domain-specific transformer-based sentiment classification and adapted from the fine-tuning pipeline demonstrated by prior IndoBERTweet applications on Indonesian social media data (Hizkia Imanuel, 2026). A descriptive design was chosen because the study sought to characterize the existing distribution of public sentiment without manipulating any research variable.

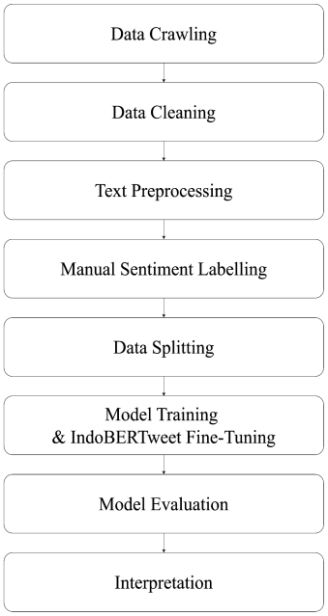


Figure 1. Research Stages

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9805

Figure 1 summarizes the research stages. The flowchart runs from top to bottom through eight sequential stages, and the output of each stage becomes the input of the next. In data crawling, public comments were collected from the official Instagram account (1,142 comments from 334 posts). Data cleaning and text preprocessing then removed noise and standardized the text, leaving 829 comments from 183 posts. In manual sentiment labelling, annotators assigned each comment a positive, neutral, or negative label, which produced the final dataset of 755 comments from 176 posts. Data splitting divided the labeled comments into training, validation, and test subsets (Table 2). Model training and IndoBERTweet fine-tuning used the training subset, and the validation subset was used to select hyperparameters (Table 3). Model evaluation measured the classification performance of the final model on the held-out test set (Tables 4 and 5 and Figure 5). The last stage, interpretation, relates the sentiment distribution and evaluation results to brand image, as presented in the Discussion.

The flowchart reads from top to bottom. Each box is one stage of the study, and each arrow shows that the output of one stage is passed to the next stage.

Public comments were collected from AIESEC Indonesia's non collaborated posts on their official Instagram account during January 2025 to September 2026 using Apify, a web scraping and data extraction platform. A total of 1142 comments from 334 posts were obtained initially. After a thorough process of data crawling, cleaning, and text preprocessing, the data turns out to be 829 comments from 183 posts. The data is later taken into a manual annotating process, resulting in the total quantity of data that will be used for this research is 755 comments from 176 posts.

Following data collection, comments underwent a sequential preprocessing procedure. The text was first case-folded to lowercase, after which URLs, hashtags, mentions, emoticons, and irrelevant non-alphabetic characters were removed through a cleaning step. Non-standard or slang expressions were then normalized to their standard Indonesian forms using an Indonesian slang-normalization dictionary, and the cleaned text was tokenized using IndoBERTweet's native WordPiece tokenizer to ensure vocabulary compatibility with the pretrained model.

Sentiment labels, namely positive, neutral, and negative, were subsequently assigned through manual annotation performed by two independent annotators, with inter-annotator agreement measured using Cohen's Kappa coefficient to ensure labeling consistency. The labeled dataset was then divided into training, validation, and test subsets. Classification was performed by fine-tuning the pretrained IndoBERTweet model, that is, retraining its final classification layer on the labeled dataset using the Hugging Face Transformers library, with the learning rate, batch size, number of epochs, and optimizer determined through hyperparameter tuning.

Model performance was measured, tested, and evaluated using four standard classification metrics, accuracy, precision, recall, and F1-score, computed from the confusion matrix generated on the held-out test set. The resulting sentiment distribution and evaluation scores were subsequently interpreted in relation to the brand image framework discussed in the Literature Review to address the study's three stated objectives.

RESULT

The final dataset contains 755 labeled comments from 176 Instagram posts. Two annotators labeled the comments independently, and The inter-annotator agreement reached a Cohen's Kappa of 0.772, indicating substantial agreement according to standard interpretation thresholds (0.61–0.80). A total of 74 comments were excluded during adjudication. The cleaned comments are short. The mean length is 6.0 words and the median is 4 words. Comments of three words or fewer make up 45.7% of the dataset. Positive comments are the shortest (mean 4.6 words, median 3). Neutral comments average 7.5 words (median 5.5). Negative comments are the longest, with a mean of 10.9 words (median 7).

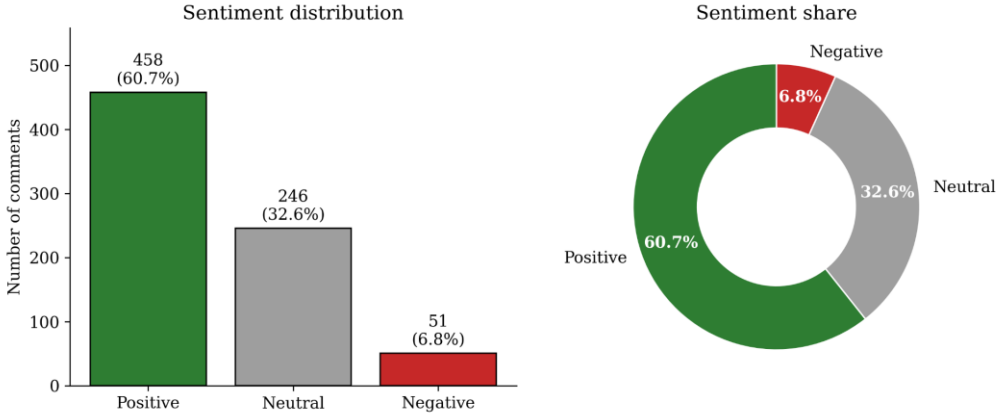


Figure 2. Sentiment distribution of the 755 labeled comments on AIESEC Indonesia's Instagram posts.
The bar chart (left) shows the number of comments per class, and the donut chart (right) shows the share of each class.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9805

Figure 2 shows the sentiment distribution. Positive comments total 458 (60.7%), neutral comments total 246 (32.6%), and negative comments total 51 (6.8%). The ratio of positive to negative comments is 9.0 to 1. The share of positive comments minus the share of negative comments gives a net sentiment of 0.54.

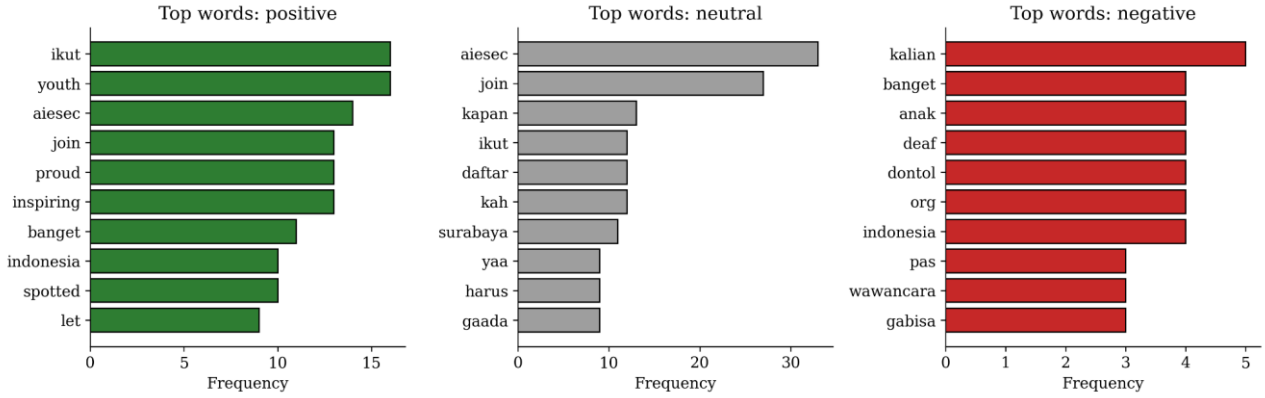


Figure 3. Ten most frequent words in positive, neutral, and negative comments (cleaned text, stopwords removed). The horizontal axis shows word frequency. Note that the axis scales differ between panels.

Figure 3 lists the ten most frequent words in each class after common stopwords were removed. Positive comments often contain *ikut, youth, join, proud, and inspiring*. Neutral comments are dominated by *aiesec, join, kapan, daftar, and kah*, together with city names such as *Surabaya*. Negative comments contain *kalian, banget, anak, org, and wawancara*. The negative words have low frequencies (three to five occurrences each) because the negative class is small.

The 176 posts received between one and 15 comments each. Fifty-five posts received a single comment. Twenty-four posts received at least one negative comment. Negative comments are concentrated. Three posts account for 23 of the 51 negative comments (45.1%). One post alone received 11 negative comments out of 14. Figure 4 shows the monthly volume. Comment volume peaked in May 2025 at about 92 comments, and positive comments dominated most months. The share of negative comments was highest in June 2025, with about 17 of 40 comments. Neutral comments formed the majority in September 2026, with about 31 of 38 comments.

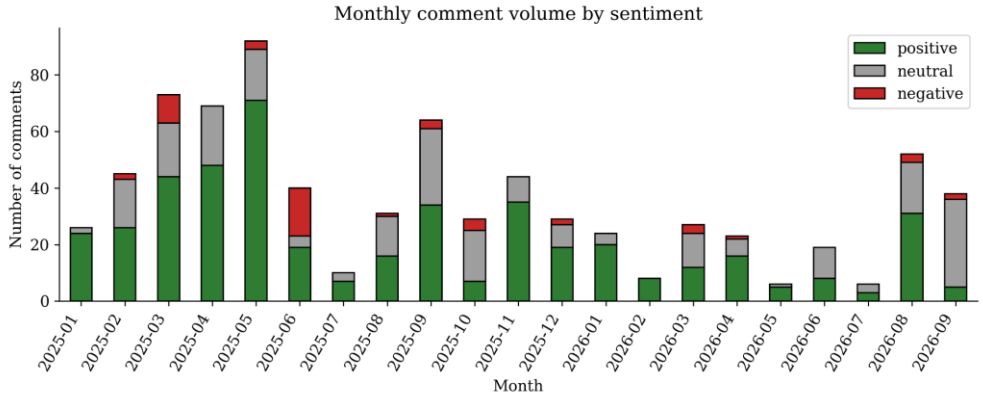


Figure 4. Monthly number of comments by sentiment class, January 2025 to September 2026. Bars are stacked, so bar height shows total comment volume.

The data were split in a stratified manner into 528 training, 113 validation, and 114 test comments (Table 2). Six combinations of learning rate and batch size were compared on the validation set (Table 3). The best setting was a learning rate of 3e-5 with a batch size of 16. It reached a validation macro F1-score of 0.891 at epoch 5. The weakest setting (2e-5 with a batch size of 32) reached 0.782.

Table 2. Number of Comments per Sentiment Class in Each Data Subset

Sentiment	Train	Validation	Test	Total	Share (%)
Negative	36	7	8	51	6.8
Neutral	172	37	37	246	32.6
Positive	320	69	69	458	60.7
Total	528	113	114	755	100

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9805

Table 3. Hyperparameter Search Results on the Validation Set

Learning rate	Batch size	Best epoch	Validation macro F1-score
3.00E-05	16	5	0.891
5.00E-05	16	4	0.877
2.00E-05	16	4	0.859
5.00E-05	32	6	0.845
3.00E-05	32	4	0.814
2.00E-05	32	5	0.782

Note. Validation macro F1-score is the unweighted mean of the F1-scores of the negative, neutral, and positive classes, computed on the validation set (113 comments). Best epoch is the epoch at which a setting reached its highest validation macro F1-score. Rows are ordered from the highest to the lowest score, so the first row is the selected setting.

Table 3 shows how the choice of hyperparameters affected validation performance. Macro F1-score ranged from 0.782 to 0.891, a difference of 0.109. Batch size had the clearest effect. At each of the three learning rates, a batch size of 16 scored higher than a batch size of 32 (0.891 versus 0.814 at 3e-5, 0.877 versus 0.845 at 5e-5, and 0.859 versus 0.782 at 2e-5). The mean macro F1-score was 0.876 for a batch size of 16 and 0.814 for a batch size of 32. With 528 training comments, a batch size of 16 gives about 33 weight updates per epoch, compared with about 17 for a batch size of 32, which may help the model learn from a small dataset. The effect of the learning rate was less consistent. With a batch size of 16, 3e-5 performed best, followed by 5e-5 and 2e-5. With a batch size of 32, the order was 5e-5, 3e-5, and 2e-5. The lowest learning rate (2e-5) gave the lowest score at both batch sizes. The best epoch fell between epochs 4 and 6 for all settings. The setting with a learning rate of 3e-5 and a batch size of 16 was therefore selected for the final model. Because the validation set contains only seven negative comments (Table 2), differences of a few hundredths between neighboring settings should be read with caution.

Table 4 compares the fine-tuned IndoBERTweet with the TF-IDF and SVM baseline on the held-out test set. Both models reached an accuracy of 0.833. IndoBERTweet obtained a higher macro F1-score (0.813 versus 0.737) and a higher macro recall (0.797 versus 0.695). The difference comes from the negative class. Recall for negative comments rose from 0.375 to 0.750, and the F1-score rose from 0.545 to 0.800. The baseline reached a precision of 1.000 on negative comments, but it predicted only three of them. IndoBERTweet had slightly lower F1-scores than the baseline on neutral comments (0.767 versus 0.779) and positive comments (0.871 versus 0.886).

Table 4. Test Set Performance of the Baseline and IndoBERTweet

Class	P (SVM)	R (SVM)	F1 (SVM)	P (IBT)	R (IBT)	F1 (IBT)	Support
Negative	1	0.375	0.545	0.857	0.75	0.8	8
Neutral	0.75	0.811	0.779	0.778	0.757	0.767	37
Positive	0.873	0.899	0.886	0.859	0.884	0.871	69
Macro average	0.874	0.695	0.737	0.831	0.797	0.813	114
Weighted average	0.842	0.833	0.827	0.833	0.833	0.833	114

Note. P = precision, R = recall, F1 = F1-score, SVM = TF-IDF with linear SVM baseline, IBT = fine-tuned IndoBERTweet. Accuracy was 0.833 for both models.

The model classified 95 of 114 test comments correctly. Of the 19 errors, 9 were neutral comments predicted as positive and 7 were positive comments predicted as neutral. Only three errors involved the negative class: one negative comment predicted as neutral, one negative comment predicted as positive, and one positive comment predicted as negative.

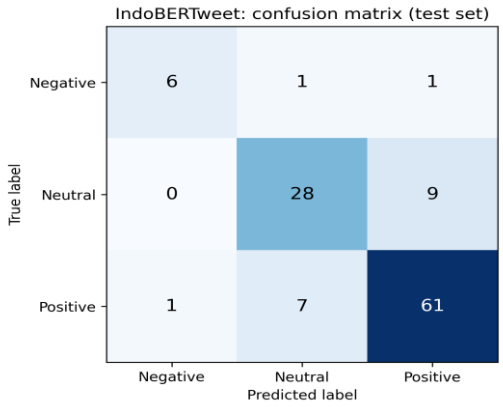


Figure 5. Confusion matrix of the fine-tuned IndoBERTweet on the test set (114 comments).
Rows show the true label, columns show the predicted label, and each cell shows the number of comments.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9805

Training was repeated with three random seeds (Table 5). Accuracy stayed at 0.833. The macro F1-score ranged from 0.748 to 0.813, with a mean of 0.785 and a standard deviation of 0.034. The F1-score for the negative class ranged from 0.571 to 0.800.

Table 5. Test Set Results of IndoBERTweet Across Three Random Seeds

Seed	Accuracy	Macro F1-score	Negative F1-score
42	0.833	0.813	0.8
43	0.833	0.795	0.737
44	0.833	0.748	0.571
Mean ± SD	0.833 ± 0.000	0.785 ± 0.034	0.703 ± 0.118

Several patterns appear among the 19 misclassified comments. Many are short questions about joining. Examples are *'How can i register?'* (labeled positive, predicted neutral) and *'gimana cara join nya'* (labeled neutral, predicted positive). Some are location requests, such as *'Denpasar?'* (labeled neutral, predicted positive). Two are long promotional-style comments that were labeled positive but predicted neutral. Four are short English comments, for example *'Can international candidates join ?'* and *'i already regist!'*.

DISCUSSION

The sentiment distribution points to a favorable brand image among the audience that comments on the account. Positive comments outnumber negative ones nine to one, and the net sentiment is 0.54. Brand image depends on the favorability of the associations that the public holds toward an entity (Dzyabura & Peres, 2021). Words such as proud and inspiring suggest favorable associations. Neutral comments (32.6%) mostly ask when and how to register or whether a city is included. These questions show interest in joining, so comments may work as a first step in the recruitment funnel. This agrees with studies that link social media activity to brand image (Utoyo & Karlina, 2023; Zulfiah et al., 2024). The positive comments are short, with a median of three words, and the word spotted appears often. Many of them seem to be tags or expressions of support rather than detailed evaluations. The favorable picture therefore describes engaged followers, not the whole public.

Negative sentiment is small, but it is not evenly spread. Three posts hold 45.1% of the negative comments, and one post received 11 negative comments out of 14. In these 11 comments, the authors criticize the interview experience, describe the organization as elitist, and question English-language gatekeeping. This reading rests on a small number of comments and should be treated as indicative. The pattern suggests that reputational risk for AIESEC Indonesia is driven by events rather than constant. A monitoring system should therefore flag sudden rises of negative comments on single posts, not only the overall average. Negative comments are also longer (mean 10.9 words) than positive ones. This may reflect that critics give reasons for their complaints.

IndoBERTweet reached 83.3% accuracy and a 0.813 macro F1-score. This accuracy is close to the 82.40% that Dewi (2025) reported for IndoBERTweet on electric-vehicle discussions. Higher figures appear in other studies, such as 94% in Chairat, Rizal, and Himawan (2025) and above 97% in Hizkia Imanuel (2026). Those studies built their labels with a lexicon or with pseudo-labeling, so the labels came from rules or from a model. A classifier can learn such labeling logic more easily than the judgment of human annotators. The scores are therefore not directly comparable with the manually labeled data of this study.

Both used class balancing, through balanced class weights in the SVM and a weighted loss in IndoBERTweet, so the gain on the negative class is not explained by weighting alone. Class weighting is a common remedy when one class is small (Marpaung & Tarigan, 2026). The advantage of IndoBERTweet appears only in the minority class. This margin is smaller than the gap between TF-IDF-based models and IndoBERTweet reported by Chairat et al. (2025). One possible reason is the very short text and the small training set (528 comments), where word frequencies already carry much of the signal. For feedback monitoring, the gain in negative recall matters most. The baseline found three of eight negative comments, whereas IndoBERTweet found six of eight.

Sixteen of the 19 errors lie between the neutral and positive classes. Sidauruk and Herowati (2026) also found that neutral sentiment is harder to separate than the other classes. Nugroho, Zyen, and Widiastuti (2025) showed that the choice of model changes how well neutral comments are recognized. The error analysis points to a likely source of confusion. Short questions about joining sit on the border between interest (positive) and a request for information (neutral). Similar questions received different labels, for example 'How can i register?' as positive and 'gimana cara join nya' as neutral. This suggests annotation ambiguity, and a more detailed labeling guideline is needed. The model also sees only the comment and not the post caption. A short comment such as 'Denpasar?' cannot be interpreted without that context.

Four of the 19 errors are English comments. IndoBERTweet was pretrained on Indonesian tweets (Koto et al., 2021), so English text lies outside most of its training data. Khairani, Mutiawani, and Ahmadian (2024) applied IndoBERTweet to Instagram news comments for emotion detection and reported 92.54% accuracy. IndoBERTweet has also been applied to TikTok reviews (Setiawan, Lhaksmana, & Bunyamin, 2023). These results suggest that a change of platform is not the main obstacle. The language mix of the comments may matter more. This explanation needs a test

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9805

with a separate English subset.

The macro F1-score of 0.813 comes from the seed that was also used in the hyperparameter search. The three-seed mean is 0.785, and the negative-class F1-score ranged from 0.571 to 0.800. With only eight negative comments in the test set, one error changes recall by 12.5 points. The mean value is therefore the safer figure to report.

This study has several limits. First, it covers one account and only the comments, so the findings describe commenters who chose to react. Comments written by AIESEC accounts were removed, and 74 comments were excluded during adjudication. Second, the dataset is small (755 comments), and the test set holds only 114 comments with 8 negative ones. Small and imbalanced data are a common concern in the sentiment analysis literature (Jim et al., 2024; Mao, Liu, & Zhang, 2024). Third, the split was random at the comment level, so comments from the same post can appear in different subsets. Fourth, two annotators labeled the data, and the neutral and positive boundary was ambiguous. Fifth, the study did not compare IndoBERTweet with the general-domain IndoBERT, so the benefit of Twitter-specific pretraining could not be isolated. Sixth, the observation period includes a short period of criticism that may not represent a typical month.

CONCLUSION

This study analyzed 755 public comments on AIESEC Indonesia's Instagram posts with a fine-tuned IndoBERTweet model. In answer to the first objective, the comments are mostly positive (60.7%), followed by neutral (32.6%) and negative (6.8%). Negative comments are concentrated in a few posts. In answer to the second objective, IndoBERTweet reached 83.3% accuracy and a 0.813 macro F1-score on the test set. It matched the TF-IDF and SVM baseline in accuracy and improved the macro F1-score from 0.737 by raising negative recall from 0.375 to 0.750. The three-seed mean macro F1-score was 0.785. In answer to the third objective, the nine-to-one ratio of positive to negative comments indicates a favorable brand image among the people who comment on the account.

The contribution of this study is an evaluation of IndoBERTweet on Indonesian Instagram comments for brand image assessment, a setting that the reviewed studies did not cover. The results are limited by the small test set, the single account, and the ambiguity between neutral and positive labels. Future work should enlarge the dataset, use post captions as context, compare IndoBERTweet with the general-domain IndoBERT, test a model that handles English comments, and track negative comments over time as an early warning of brand risk.

REFERENCES

Chairat, A. S. N., Rizal, R., & Himawan, H. (2025). Comparative sentiment analysis of YouTube comments on Indonesia's electric vehicle incentive policy using TF-IDF and IndoBERTweet models. Jurnal Teknik Informatika (JUTIF), 6(6), 5988–6002. https://doi.org/10.52436/1.jutif.2025.6.6.5499

Dewi, B. E. S. (2025). Sentiment analysis model on electric vehicles using IndoBERTweet and IndoBERT algorithm. Antivirus: Jurnal Ilmiah Teknik Informatika, 19(2), 192–202. https://doi.org/10.35457/w5r3g517

Dzyabura, D., & Peres, R. (2021). Visual elicitation of brand perception. Journal of Marketing, 85(4), 44–66. https://doi.org/10.1177/0022242921996661

Hizkia Imanuel, V. (2026). Analisis sentimen publik terhadap Program Makan Bergizi Gratis (MBG) di platform X menggunakan algoritma IndoBERTweet [Bachelor's thesis, Universitas Multimedia Nusantara]. https://kc.umn.ac.id/id/eprint/44384/

Ihalauw, S. A., Lapatta, N. T., Nugraha, D. W., Wirdayanti, & Lamasitudju, C. A. (2025). Analisis sentimen terhadap kinerja awal pemerintahan menggunakan IndoBERT dan SMOTE pada media sosial X. Jurnal Algoritma, 22(2). https://doi.org/10.33364/algoritma/v.22-2.2957

Jim, J. R., Talukder, M. A. R., Malakar, P., Kabir, M. M., Nur, K., & Mridha, M. F. (2024). Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review. Natural Language Processing Journal, 6, 100059. https://doi.org/10.1016/j.nlp.2024.100059

Khairani, U., Mutiawani, V., & Ahmadian, H. (2024). Pengaruh tahapan preprocessing terhadap model IndoBERT dan IndoBERTweet untuk mendeteksi emosi pada komentar akun berita Instagram. Jurnal Teknologi Informasi dan Ilmu Komputer, 11(4), 887–894. https://doi.org/10.25126/jtiik.1148315

Koto, F., Lau, J. H., & Baldwin, T. (2021). IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 10660–10668). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.833

Mao, Y., Liu, Q., & Zhang, Y. (2024). Sentiment analysis methods, applications, and challenges: A systematic literature review. Journal of King Saud University - Computer and Information Sciences, 36(4), 102048. https://doi.org/10.1016/j.jksuci.2024.102048

Marpaung, E. E. O., & Tarigan, M. (2026). Perbandingan model IndoBERT dan Indonesian RoBERTa Base dalam analisis sentimen pengguna Twitter terhadap Timnas Indonesia di Kualifikasi Piala Dunia 2026. JATI (Jurnal Mahasiswa Teknik Informatika), 10(1), 30–37. https://doi.org/10.36040/jati.v10i1.16570

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9805

Nugroho, M. B., Zyen, A. K., & Widiastuti, N. A. (2025). Multiclass sentiment analysis of electric vehicle incentive policies using IndoBERT and DeBERTa algorithms. Journal of Applied Informatics and Computing, 9(3), 910–919. https://doi.org/10.30871/jaic.v9i3.9511

Setiawan, J. C., Lhaksmana, K. M., & Bunyamin, B. (2023). Sentiment analysis of Indonesian TikTok review using LSTM and IndoBERTweet algorithm. Jurnal Ilmiah Penelitian dan Pembelajaran Informatika, 8(3), 774–780. https://doi.org/10.29100/jipi.v8i3.3911

Sidauruk, V. E., & Herowati, W. (2026). IndoBERT-based sentiment analysis of political discourse on platform X: The case of Prabowo-Gibran administration. Journal of Applied Informatics and Computing, 10(1), 673–683. https://doi.org/10.30871/jaic.v10i1.11586

Utoyo, A. W., & Karlina, N. (2023). Examining social media culture as the Tuku Coffee Shop's brand image on Instagram. E3S Web of Conferences, 426, 02006. https://doi.org/10.1051/e3sconf/202342602006

Zulfiah, E., Fadhilah, M., & Lukitaningsih, A. (2024). Analisis pemasaran media sosial terhadap sikap merek melalui kesadaran merek dan citra merek pada institusi pendidikan. Ideguru, 10(1). https://doi.org/10.51169/ideguru.v10i1.1615

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.