

LLM-Assisted Annotation in Comparative Sentiment and Trend Analysis of Indonesian K-12 Education on X

Abdurrohchim S. Wahyudi^{1*}, Rachmat Ramadhiansyah², Indra Budi³, Prabu Kresna Putra⁴,
Aris Budi Santoso⁵

^{1,2,3,5}Faculty of Computer Science, Universitas Indonesia, Jakarta, Indonesia

⁴National Research and Innovation Agency, Indonesia

¹abdurrohchim.s@ui.ac.id, ²rachmat.ramadhiansyah@ui.ac.id, ³indra@cs.ui.ac.id, ⁴prab003@brin.go.id,

⁵aris.budi@ui.ac.id

Submitted : Jun 21, 2026 | Accepted : Sep 5, 2026 | Published : Oct 5, 2026

Abstract: Indonesia was ranked 69th out of 81 participating nations in the 2022 PISA test results, reflecting persistent challenges in K-12 education quality. Public discourse on these challenges has grown substantially on social media, yet automated sentiment analysis of Indonesian education-related text remains difficult due to informal language and implicit sentiment expression. Prior work on Indonesian education sentiment has focused on short observation windows or a single policy, without comparing transformer-based models directly against traditional classifiers or scaling annotation with large language models. This study aims to compare traditional machine learning classifiers and transformer-based models for sentiment analysis of Indonesian K-12 education discourse on X, as well as the trend of sentiments longitudinally and key terms associated with different sentiments. The dataset comprises 38,957 tweets in the Indonesian language collected between January 2022 and October 2025 that were annotated via a multi-stage, large language model-assisted few-shot classification pipeline with self-consistency sampling tiebreaker, reaching inter-annotator agreement of 82.01% (Cohen's kappa = 0.672). This study evaluated the accuracy and macro F1 score of four traditional classifiers, including Logistic Regression, Support Vector Machine, Naïve Bayes, and Random Forest, alongside three transformer-based models, namely IndoBERTweet, IndoBERT-base-p1, and IndoBERT-large-p1. Transformer-based models performed better than traditional classifiers based on all metrics. IndoBERTweet reached the best performance with 0.884 accuracy and 0.859 macro F1, compared to Logistic Regression as the best traditional classifier achieved 0.768 and 0.714, respectively. Negative sentiment dominated public discourse, with “kurikulum merdeka” as the most frequent term across all sentiment categories.

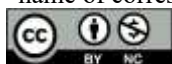
Keywords: K-12 Education; Large Language Model; Sentiment Analysis; Social Media; Transformer-Based Models

INTRODUCTION

K-12 education is important in developing human capital, increasing employment opportunities, and climbing the social ladder (Afandi et al., 2026). It is also essential in achieving the Indonesia Emas 2045 vision that aims to transform the nation into a developed country (Rosani et al., 2025). Education's reach extends to poverty too, as a panel analysis across Indonesia's 34 provinces found that longer average years of schooling correspond to lower poverty rates (Fauzan et al., 2026). The Indonesian government is constitutionally required to direct 20% of the state budget to education (Nazamuddin et al., 2026). Despite these efforts, the quality of education in Indonesia remains lagging. In the 2022 Program for International Student Assessment (PISA) results, Indonesia ranked 69th out of 81 countries (OECD, 2023). Indonesian educators have attributed this decline to factors including curriculum issues and resource limitations (Wijaya et al., 2024). Furthermore, uneven teacher quality and inadequate facilities remain significant challenges to the quality of education in Indonesia (Rosani et al., 2025).

Another effort from the Indonesian Government is shown by continuously revising the education curriculum at least 10 times. The most recent is the adoption of the Merdeka Curriculum that aims to prepare students with

*name of corresponding author



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

more practical skills and experiences to tackle the challenges in the industry (Hidayat et al., 2025). Traditional policy evaluation leans on outcome measures like enrollment or completion rates. Those numbers could miss how a policy is actually experienced by the people living through it (Bustillo et al., 2026). Sentiment drawn from public discourse fills that gap, giving policymakers a basis for revising a policy as concerns emerge (Sukhwal & Kankanhalli, 2022).

Social media, especially X (formerly Twitter) has become a common place for the public to express their opinion and perception freely (Smith-Mutegi et al., 2025). Indonesia has ranked as the 4th largest total active X users in the world, with a total of 24.45 million users (Medantoro & Muljono, 2026). With the massive Indonesian user base, X can facilitate gathering immediate and informal insights and observing sentiment changes over time (Wafda et al., 2025). By integrating sentiment analysis into the evaluation of the K-12 education system, the process of understanding and responding to the public's voices becomes more data-driven (Fitriyani et al., 2025).

Previous studies have explored public perceptions of Indonesian education from several angles, though each stops short somewhere. Kurnianingrum et al. (2023) and Fitriyani et al. (2025) captured general education sentiment on X, but only within a short window. Wafda et al. (2025) narrowed the lens to the Merdeka Curriculum alone, and Sandra et al. (2022) did the same for the Merdeka Campus program. None of them extended their annotation with large language models (LLMs) or tested a transformer directly against a traditional classifier. A multi-year view of how sentiment shifts across the wider K-12 education conversation remains untested too.

To address these gaps, this study looks at public sentiment about K-12 education quality in Indonesia using social media data by tracking how opinions change over time and comparing different sentiment analysis methods. Labels for the dataset come from three LLMs rather than manual annotation, allowing the labeling scale to match the dataset's size. The study aims to compare the performance of traditional machine learning (ML) and transformer-based pipelines for Indonesian-language sentiment classification, identify yearly sentiment trends from January 2022 to October 2025, and highlight key terms in each sentiment group. The results are intended to give the government in the education sector data-driven insights into public concerns and to provide the academic community with a direct comparison between traditional and transformer-based models for Indonesian sentiment analysis.

This study does not aim to evaluate or criticize any individual, institution, or specific policymaker. The analysis is limited to aggregated public discourse on X and should not be interpreted as representing the entire Indonesian population. All results are reported at the aggregate level to avoid exposing individual users or attributing sentiment to specific persons or institutions.

LITERATURE REVIEW

Sentiment analysis has become a common approach for measuring public response to government initiatives and policy implementation (Kurnianingrum et al., 2023; Sandra et al., 2022). Recent studies have begun to apply it specifically to Indonesian education policy, such as the Merdeka Curriculum (Herra et al., 2025; Wafda et al., 2025) and Merdeka Campus (Sandra et al., 2022). Analyzing this discourse presents multiple challenges, including class imbalance, informal phrasing, and sarcasm (Fitriyani et al., 2025). Traditional approaches relying on TF-IDF struggle with these linguistic complexities because they reduce tweets to word-frequency counts and discard the original sequence. Without that structural context, frequency-based models fail to capture meaning and often miss critical shifts in sentiment (Medantoro & Muljono, 2026). Transformer architectures overcome this limitation through bidirectional self-attention (Devlin et al., 2019). With that mechanism, the model reads the context of every word, which drives performance on ambiguous text.

Medantoro & Muljono (2026) found that a transformer-based model substantially outperformed all traditional classifiers on Indonesian education policy tweets, and Adhim & Cahyono (2025) confirmed the same advantage on Indonesian social media text on a different topic. Further improvements in the model's performance could be achieved by pretraining transformers on domain-specific data. Koto et al. (2021) showed that domain-specific pretraining on Indonesian Twitter data produced a model that outperformed the general-purpose baseline across multiple social media tasks. Model choice alone does not solve every problem in this setting. A model can post strong overall accuracy while still missing most of the minority-class cases, since accuracy weights every class by how often it appears. Recent studies have also addressed class imbalance in sentiment analysis by applying class weighting and oversampling, and by using macro F1-scores to more accurately evaluate the performance of minority classes (Sidharta et al., 2025).

The problem of subjective bias and the high cost of manual annotation has led to an increasing use of LLMs in the automatic labeling process. Although there have been promising results obtained while using open-source LLMs in zero-shot settings, their performance on non-English conversational text remains poor, as shown by a low recall score (Hidayatullah & Prawira, 2024). However, applying a few-shot prompting can help close that gap by providing the model concrete examples to anchor its decisions (Cahyawijaya et al., 2024). Another useful technique here is performing self-consistency sampling, which performs multiple independent runs and uses

majority voting to obtain more reliable predictions and reduce output variability (Niimi, 2025; Wu et al., 2024), but this technique is computationally too costly to utilize on the entire sample (Aggarwal et al., 2023).

Studies on Indonesian education sentiment have generally focused on specific topics or short observation periods. Moreover, previous studies tend to underreport the performance of minority classes when the distribution is skewed (Fitriyani et al., 2025; Sidharta et al., 2025). Comparing traditional and transformer models in predicting sentiment from Indonesian education policy texts is also relatively rare. This study aims to compare four traditional classifiers, such as Logistic Regression, SVM, Naïve Bayes, and Random Forest, with three transformer-based models, including general-purpose and domain-specific models. The pretrained LLMs will be used in the annotation stage, while class imbalance will be handled specifically within both pipelines.

METHOD

This study analyzed the public sentiment on K-12 education in Indonesia. The model is developed using Google Colab with 12 GB of RAM, 100 GB of temporary SSD storage, and GPU capabilities. As illustrated in Fig. 1, this study employs two modeling approaches for comparison, including a traditional ML pipeline combining TF-IDF feature extraction with multiple traditional classifiers and a transformer-based pipeline with fine-tuning IndoBERT variants. Fig. 1 displays the methodology used in this paper.

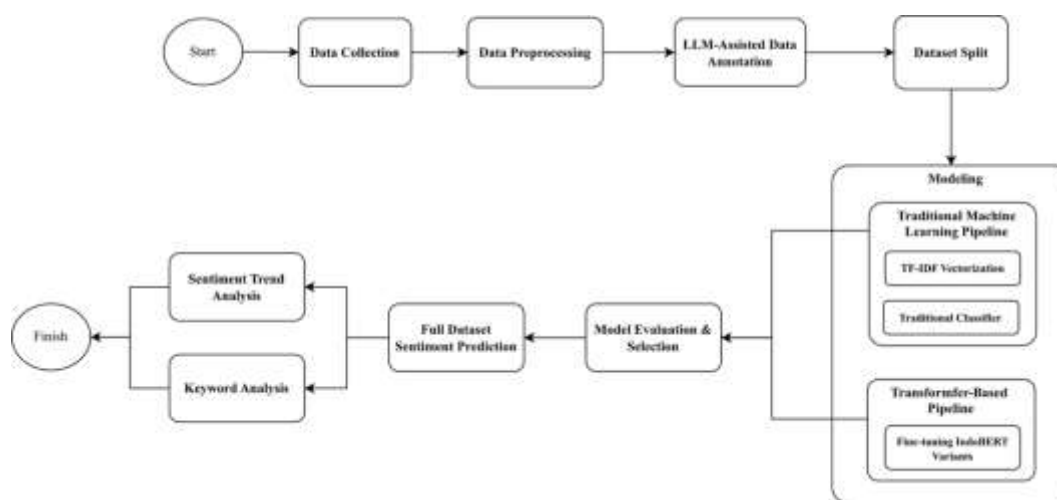


Fig. 1 Methodology

Data Collection

The collected dataset comprises tweets from X (Twitter). To gather this data, a Node.js library called tweet-harvest was used. Several keywords regarding K-12 education in Indonesia have been considered, including “sistem zonasi”, “kualitas pendidikan”, “kurikulum merdeka”, and “kompetensi guru”. The data collection was limited to tweets between January 2022 and October 2025 and ran in monthly batches. The number of tweets returned on each run can vary since the scraping tool was limited by the platform’s rate limit and how many matching tweets existed in each period. To keep the dataset relevant, we excluded tweets that were not in Bahasa Indonesia, along with duplicate, spam, and promotional content.

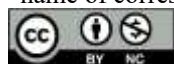
Data Preprocessing

There is noise present in the tweets written in the Indonesian language. URLs, user mentions (@username), hashtags (#pendidikan), and other non-alphabetic characters do not carry sentiment value, hence in the cleansing step, these elements were removed and the irregular whitespaces were standardized. Moreover, some of the tweets include expressions like “gk” which should be replaced by their respective formal expression like “tidak”. The preprocessor from IndoNLP was used to normalize these words. Finally, high-frequency function words (such as “dan”, “di”, and similar terms) were removed as they do not contribute to sentiment value but increase the dimensionality of the feature space. The Sastrawi library was utilized in this process.

Data Annotation

Data annotation comprised a few-shot LLM pipeline with multiple stages. First of all, two open-source models that can handle the language of Southeast Asia were chosen as primary annotators, namely Sailor-7B-Chat (Sail; Dou et al., 2024) and SeaLLMs-v3-7B-Chat (SeaLLM; Zhang et al., 2025). Both were pretrained on Southeast Asian languages and therefore could handle informal social media data better than general models. Then, they were fed with a few-shot prompt consisting of several pairs of examples along with their corresponding labels, and

*name of corresponding author



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

then they were asked to provide a label for each normalized tweet (positive, neutral, or negative). Agreement between the two primary annotators was quantified with Cohen's kappa and read against the benchmark bands of Landis & Koch (1977). When the two primary models produced different labels, the tweet was forwarded to Gemma2-9b-cpt-sahabatai-v1-instruct (Sahabat-AI; Sahabat-AI, 2024) to serve as a tiebreaker. The selection of this model as a tiebreaker is because it underwent pretraining with data from Indonesia, Java, and Sunda, providing a better understanding of language nuances in making decisions for ambiguous cases. Self-consistency sampling was performed using this model by generating three independent labels, and then the most frequently generated label was selected. The label was accepted only if the confidence score for it was above 0.6.

Annotation Reliability

Annotation used a random subsample of 10,000 tweets drawn from the full corpus that passed filtering. The size was set to fit the compute budget of the Google Colab environment, while still giving the classifiers enough training data. A manual review of 100 random tweets from that subset served as an author verification audit and was designed to catch class-wide mislabeling patterns rather than to estimate labeling accuracy. Ultimately, every label in this study traces back to the three LLMs alone, leaving a risk of shared bias across models trained on related data.

Dataset Split

The labeled dataset was split into training, validation, and test subsets using stratified sampling to keep class proportions consistent across all three splits. Training used 80% of the data, while validation and testing each received 10%. The model was learned from the training split only. Hyperparameter tuning was done through the validation data, whereas the test set remained untouched and was solely used once for final evaluation of the model.

Modeling Approaches

In this stage, two different computational approaches were employed, including a traditional ML pipeline and a transformer-based pipeline. In the traditional ML pipeline, fully preprocessed texts (including stop-word removal) were used in order to decrease noise and feature dimension. Firstly, the text was vectorized using TF-IDF, which consists of tokenization of the corpus and scoring of the word depending on the frequency of the term in a particular document and the rarity of the term in the corpus. Four classic algorithms were evaluated, such as Support Vector Machine (SVM), Logistic Regression, Random Forest, and Complement Naïve Bayes. To optimize performance, hyperparameter tuning was conducted using GridSearchCV with a 5-fold stratified cross-validation, optimized for the macro F1-score. To handle the imbalanced dataset, Random Oversampling was applied to the Naïve Bayes model, and balanced class weights were natively utilized for the SVM, Logistic Regression, and Random Forest classifiers.

For the transformer-based pipeline, three transformer-based models were fine-tuned, namely IndoBERTtweet-base-uncased (Koto et al., 2021), IndoBERT-base-p1, and IndoBERT-large-p1 (Wilie et al., 2020). The IndoBERT-base-p1 and IndoBERT-large-p1 models are general-purpose models trained on the 23.43 GB Indo4B text dataset. The difference between the two lies in their parameter counts, with 124.5 million and 335.2 million parameters, respectively. IndoBERTtweet is specifically designed for Indonesian Twitter discourse and has been pretrained on 409 million domain-specific word tokens. Unlike the traditional pipeline, the transformer-based approach used text that underwent only cleansing and normalization. The normalized text was encoded into contextual token embeddings using the models' respective tokenizers, with sequences truncated to a maximum length of 128 tokens. To handle the imbalanced data, a custom WeightedTrainer was implemented to override the standard loss computation. This custom trainer applies weighted cross-entropy utilizing balanced class weights, which were further augmented with a positive class weight multiplier to specifically penalize misclassifications of the positive class. Additionally, a label smoothing factor was applied to the models' cross-entropy loss functions. The training process utilized an early stopping mechanism with a patience of 2 epochs to prevent overfitting, saving the best model state based on the macro F1-score at the end of each epoch.

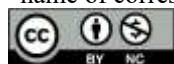
Model Evaluation Metrics

After passing through the training and modeling process, the effectiveness of the models was checked against the test dataset. The confusion matrix is used for assessment in the research. In a Confusion Matrix, rows represent actual classes, and columns represent predicted classes. As for indicators that will be measured, Accuracy, Precision, Recall, and F1-Scores (both weighted and macro) were used (Opitz, 2024). Accuracy can be measured as the ratio of correct predictions to the total number of predictions made (1).

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

Precision is the proportion of true positives among all predicted positives (2).

*name of corresponding author



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

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

Recall measures the proportion of actual positives that are correctly identified by the model (3).

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

F1-Score is the harmonic mean of precision and recall (4).

$$F1 - score = 2 \times \frac{precision \times recall}{precision+recall} \quad (4)$$

As the problem dataset contains an imbalance in classes, both the weighted and macro F1 scores will be utilized. The Weighted F1 score averages per-class scores weighted by the number of supports, which is the total number of true instances for each class. It represents how the model performs by considering the real distribution of instances. While the macro F1 score considers the predictions of all classes equally, it is more sensitive to misclassifying minority classes (Riyanto et al., 2023).

Sentiment Prediction

All cleaned tweets were labeled by the single best-performing model, chosen by macro F1. Using one model throughout kept the decision boundary consistent, since mixing classifiers would have introduced label artifacts at the boundaries between their respective training distributions. Those labeled tweets are then fed directly into the trend and key terms analyses.

Sentiment Trend and Key Terms Analysis

In the sentiment trend analysis, tweets were grouped by year and counted within each sentiment class to produce an annual distribution for the period from 2022 through October 2025. Tracking these counts year by year showed when the opinion shifted and how long each shift lasted. In the key terms analysis, each sentiment year subset was then scanned for bigram frequency. Bigrams capture co-occurring terms that single-word counts miss, such as “kurikulum merdeka”, which is more informative than either word alone. Comparing the top phrases across negative, neutral, and positive subsets can reveal which specific topics drove each sentiment.

RESULT

Data Collection and Preprocessing Result

The data collection yielded 60,856 tweets between 2022 and October 2025, of which 38,957 passed filtering. These formed the working dataset for all subsequent analyses. The filtered tweets were processed through cleansing, normalization, and stop-word removal to produce a standardized corpus. Table 1 traces that sequence on a single example tweet.

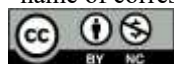
Table 1. Preprocessing Steps

Step	Tweet
Full Text	@username Riil. Ganti presiden ganti menteri ganti kurikulum. Ya gimana ga bingung. Kalo ada pr jg kdg ortu bingung mau ngajarinnya.
Text Cleansing	riil. ganti presiden ganti menteri ganti kurikulum. ya gimana ga bingung. kalo ada pr jg kdg ortu bingung mau ngajarinnya.
Normalization	riil. ganti presiden ganti menteri ganti kurikulum. ya bagaimana enggak bingung. kalau ada pr juga kadang orang tua bingung mau ngajarinnya.
Stopword Removal	riil. ganti presiden ganti menteri ganti kurikulum. bingung. pr kadang tua bingung ngajarinnya.

Data Annotation Results

A total of 10,000 data samples were randomly selected for labeling. The first stage was to put the samples into the SeaLLM and Sail models. As illustrated in Fig. 2a, the SeaLLM model produced sentiments of negative, neutral, and positive with proportions of 54.91%, 33.90%, and 11.19%, respectively. In comparison, the Sail model’s label output distributions (Fig. 2b) were 64.28% negative, 25.80% neutral, and 9.92% positive. The agreement rate for these models’ labels was 82.01%, with a Cohen’s kappa of 0.672, which falls in the substantial band. The 1,799 non-consensus labels from the SeaLLM and Sail models were then passed into the next labeling stage, adjudication. During the self-consistency sampling process using the Sahabat-AI model, all the non-consensus labels met the 0.6 threshold. A manual check of 100 samples drawn from the full 10,000 labeled dataset

*name of corresponding author



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

found no labeling errors. The final dataset distribution is shown in Fig. 2c, containing 59.47% negative, 28.97% neutral, and 11.56% positive tweets. This dataset was used for the data split and modeling stage.

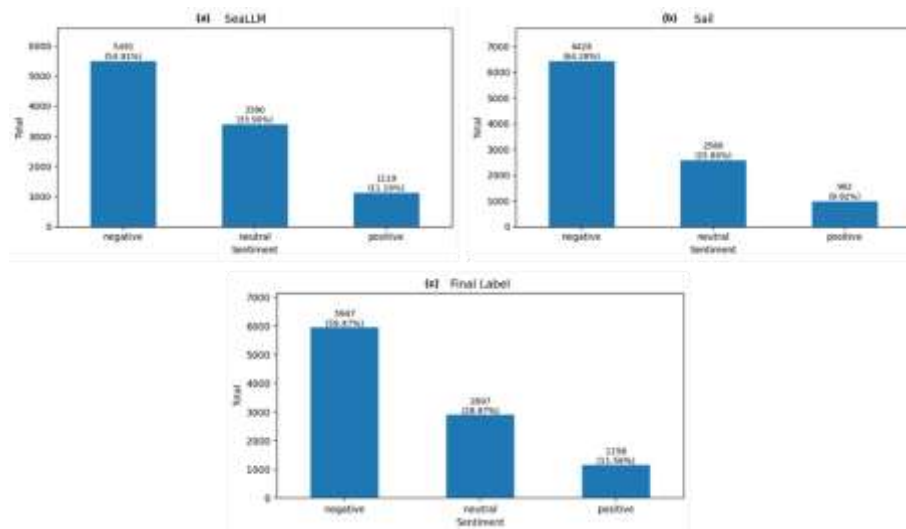


Fig. 2 Label distributions across annotation stages: (a) SeaLLM model, (b) Sail model, and (c) Final label

Model Performance and Comparative Analysis

Performance evaluation was conducted to compare the traditional classifiers against the fine-tuned transformer-based models. During the initial phase, a grid search cross-validation was executed across the four traditional classifiers to identify their optimal hyperparameter configurations based on the macro F1-score. The resulting best parameter tunings for each traditional model, including their optimal solvers, regularization parameters, and TF-IDF vectorizer settings, are presented in Table 2.

Table 2. Traditional Models' Best Parameter Tuning

Model	C	Kernel	Max Features	N-Gram	Solver	Alpha	Min. Sample Split	N-Estimator	Handling Imbalance
Logistic Regression	1	-	30000	(1, 1)	liblinear	-	-	-	Class Weight
SVM	1	rbf	10000	(1, 1)	-	-	-	-	Class Weight
Naïve Bayes (ComplementNB)	-	-	30000	(1, 2)	-	0.1	5	100	Random Sample
Random Forest	-	-	10000	(1, 1)	-	-	-	-	Class Weight

For the transformer-based pipeline, the IndoBERTweet, IndoBERT-base-p1, and IndoBERT-large-p1 models were fine-tuned utilizing the same optimal parameter configuration. This shared training configuration consisted of 5 epochs, a learning rate of $5e-5$, a label smoothing factor of 0.05, and a positive weight multiplier of 1.01 to address class imbalance. The evaluation results comparing the traditional and transformer-based models on the unseen testing dataset are presented in Table 3 using accuracy and F1-score metrics.

Table 3. Model Evaluation Comparison

Model	Accuracy	Macro F1	Weighted F1	F1 Neg	F1 Neu	F1 Pos
<i>Traditional Models</i>						
Logistic Regression	0.768	0.714	0.764	0.831	0.677	0.634
SVM	0.762	0.706	0.759	0.825	0.681	0.610
Naïve Bayes (ComplementNB)	0.734	0.671	0.737	0.813	0.660	0.541
Random Forest	0.733	0.653	0.722	0.811	0.612	0.537
<i>Transformer-Based Models</i>						
IndoBERTweet	0.884	0.859	0.884	0.922	0.829	0.825
IndoBERT-large-p1	0.882	0.852	0.883	0.921	0.839	0.796
IndoBERT-base-p1	0.842	0.806	0.841	0.891	0.773	0.754

The evaluation results indicate that all transformer-based models outperform all traditional models across all metrics. IndoBERTweet achieved an accuracy of 0.884 and a weighted F1-score of 0.884, closely followed by the

much larger IndoBERT-large-p1 (0.882 accuracy, 0.852 macro F1-score). In the traditional models' results, they all struggled to identify positive sentiment. Logistic Regression, as the best-performing traditional model, achieved an F1 score of 0.634 for the positive class, whereas the worst transformer-based model, IndoBERT-base-p1, achieved an F1 score of 0.754 for the positive class. The performance gap traces back to how each approach represents text. Transformer embeddings carry context across the entire sentence and enable the models to capture the sarcasm and informal phrasing of Indonesian tweets. In contrast, traditional classifiers count word frequencies and ignore the surrounding context.

The confusion matrices presented in Fig. 3 show that across all four traditional models, positive tweets were misclassified as negative more frequently than as neutral (32 instances for Logistic Regression, 28 for SVM, 22 for Naïve Bayes, and 48 for Random Forest). Neutral tweets exhibited a similar drift toward the negative class. This misclassification pattern likely stems from the frequency-based representation. Because a word-count model processes individual words without understanding the broader context, criticism expressed through negative terminology triggers a negative prediction even when the overall message remains supportive. Random Forest demonstrates this majority-class dominance most clearly. Alongside 527 correctly identified negative cases, the model incorrectly assigned the negative label to 113 neutral and 48 positive instances.

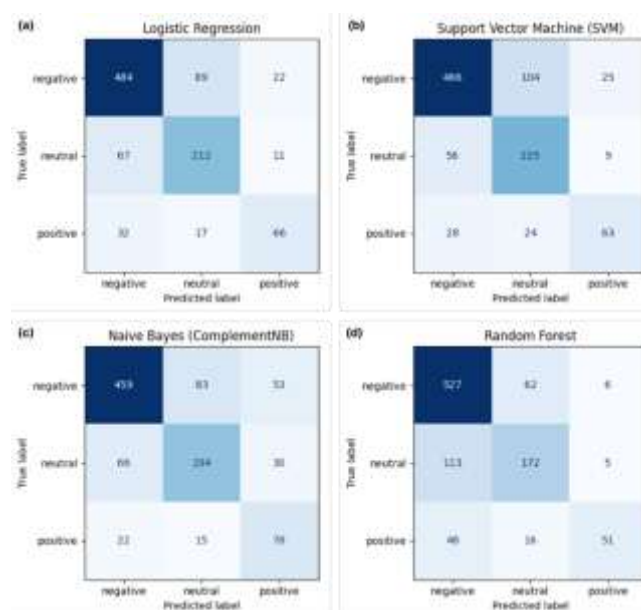


Fig. 3 Confusion matrices for traditional classifiers: (a) Logistic Regression, (b) Support Vector Machine, (c) Naïve Bayes, and (d) Random Forest

As illustrated in Fig. 4, the transformer models misclassify positive tweets as negative far less frequently. The misclassification rate drops from a range of 22 to 48 in the traditional models down to between 5 and 17 across the three transformers. IndoBERTweet and IndoBERT-large-p1 perform best by minimizing these negative misclassifications to just 5 and 6 instances, respectively. Both models shift most of their remaining positive-class errors into the neutral category. This classification pattern indicates that the models recognize the tweet is not negative but struggle to detect the full strength of the positive sentiment. In contrast, IndoBERT-base-p1 underperforms by retaining a stronger negative bias, sending 17 positive tweets into the negative class against 12 into neutral. A similar bias also affects the neutral class, where all three transformers still pull between 29 and 52 tweets into the negative category. Ultimately, these numbers demonstrate that transformers offer a clear improvement over traditional classifiers but cannot completely erase the tendency to assign instances to the majority class.

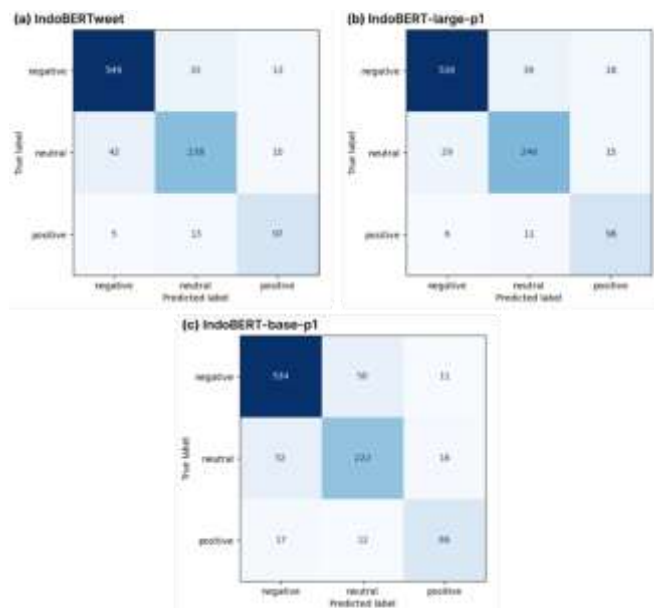


Fig. 4 Confusion matrices for transformer-based models: (a) IndoBERTweet, (b) IndoBERT-large-p1, and (c) IndoBERT-base-p1

Sentiment Trend Analysis

The distribution of tweets by their sentiment categories (negative, neutral, and positive) in each of the years from 2022 up to October 2025 is shown in Fig. 5. During the entire observation period, negative sentiments were prevalent throughout the social media discourse regarding K-12 education in Indonesia, trailed by neutral and positive sentiments. Sustained negativity across the period shows public dissatisfaction with education policies and their implementation.

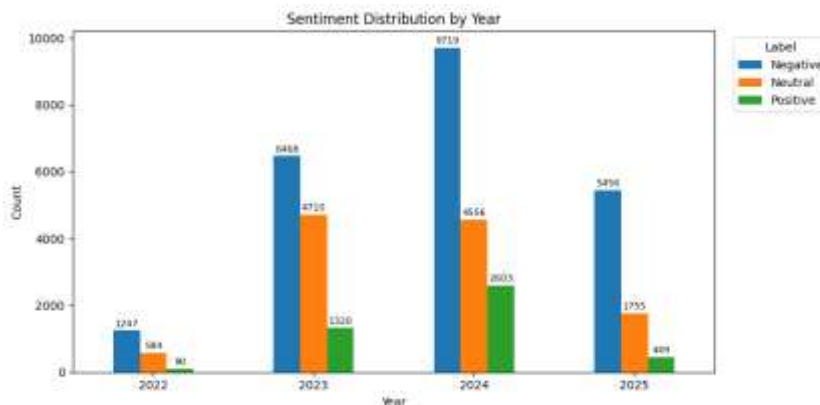
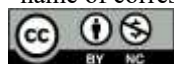


Fig. 5 Overall Sentiment Distribution by Year

Key Terms Analysis

Table 4 outlines the specific phrases shaping these discussions. "Kurikulum merdeka" appeared at the top of all three sentiment groups, but its negative class count of 11,388 mentions exceeds both its neutral and positive counts. That difference reflects how much friction the curriculum rollout generated. School enrollment policy drove another set of negative terms, namely "sistem zonasi" that appeared 4,555 times and "jalur zonasi" 1,423 times. Both point to the same ongoing disputes over zoning rules. Other negative terms were "skor pisa" (1,042 mentions) and "gaji guru" (969 mentions), showing that the public was also frustrated by how the education system performs overall and how poorly it compensates teachers.

*name of corresponding author



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

Table 4. Frequency of Key Terms by Sentiment Category (2022 to October 2025)

Negative Key Terms	Total	Neutral Key Terms	Total	Positive Key Terms	Total
kurikulum merdeka	11,388	kurikulum merdeka	7,721	kurikulum merdeka	3,679
sistem zonasi	4,555	sistem zonasi	1,098	transformasi pendidikan	832
jalur zonasi	1,423	jalur zonasi	1,010	jalur zonasi	432
skor pisa	1,042	implementasi kurikulum	743	sistem zonasi	323
gaji guru	969	kurikulum 2013	503	anak anak	220

Neutral tweets covered the same policy terms but in a more comparative means. “Kurikulum merdeka” led with 7,721 mentions, often appearing alongside “kurikulum 2013” as users weighed the two against each other rather than reacting emotionally to either. Positive tweets were fewer in number but more forward-looking in focus. “Transformasi pendidikan” appeared 832 times in positive tweets, suggesting that users who expressed support were responding to the broader direction of educational reform rather than to any specific policy or implementation detail.

DISCUSSIONS

As indicated by the data in Table 3, the traditional classifiers fell furthest behind on the minority sentiment classes where textual meaning relies on more factors than just word frequency. Adhim & Cahyono (2025) and Medantoro & Muljono (2026) reported the same pattern across different domains of Indonesian social media tasks, which implies that frequency-based representations carry a structural disadvantage. For comparison, Wafda et al. (2025) fine-tuned IndoBERT on 2,359 Merdeka Curriculum tweets and reached 85% accuracy on a dataset containing 31% positive instances. The accuracy reported in their research did not reach the 88.4% achieved in this study by IndoBERTweet on a highly imbalanced dataset with only 11.56% positive instances. Because both experiments utilized an IndoBERT variant, this performance difference suggests that dataset size and domain-specific pretraining are the key factors driving model performance.

The annotation phase of this study recorded an 82.01% label agreement rate, which closely aligns with the 85.0% agreement reported by Sidharta et al. (2025) on a binary sentiment analysis task. Their methodology relied on human adjudication to resolve label disagreements, contrasting directly with the LLM tiebreaker applied in this current research. Achieving comparable outcomes despite these completely distinct methodological approaches implies that the remaining gap in label agreement originates from the inherent complexity of the text itself instead of the annotation pipeline design.

IndoBERTweet outperformed IndoBERT-large-p1 by 0.7 points in macro F1 despite containing roughly one-third of the 335.2 million parameters found in the larger version. Koto et al. (2021) developed IndoBERTtweet by expanding the base vocabulary with a massive corpus of real Indonesian tweets to directly match the data source of this study. The broader Indo4B corpus used to pretrain IndoBERT-large-p1 leans heavily toward formal sources like news and Wikipedia (Wilie et al., 2020). The performance advantage of IndoBERTtweet becomes more apparent in the positive class. The model achieved an F1 score of 0.825 in this category compared to 0.796 for the larger model. The positive instances represent a minority class and provide very few training examples for IndoBERT-large-p1 to learn informal social media expressions. IndoBERTtweet successfully mitigates this data scarcity because its domain-specific pretraining already provides a vocabulary tailored to conversational Indonesian.

The observation period coincided with the implementation of the Merdeka Curriculum and followed the release of Indonesia’s 2022 PISA results (Wijaya et al., 2024). These developments sustained public attention on the education system’s performance and on policy-related issues. Negative posts consistently named specific policies rather than rejecting educational reform in general. The dominance of “transformasi pendidikan” in positive posts reflects that the public appears to accept the direction of reform while criticizing how it is being carried out. This research found that the Merdeka Curriculum was indeed a prominent topic discussed by Indonesians on social media, just like what the study conducted by Wafda et al. (2025) showed. “Sistem zonasi” ranked second only to “kurikulum merdeka” among the terms tracked in this study. Putri et al. (2023) found that opposition to zoning was not about the policy’s equity goal. Uneven zone boundaries and exclusion from preferred schools were the primary issues and drove some parents to falsify domicile documents. That kind of pressure is a plausible driver of the volume seen here. The appearance of “gaji guru” among negative terms is a structural condition beyond policies. Its presence in the top negative terms suggests that the public connects education quality to how the profession is resourced.

Two features of the data limit how far the findings can be generalized. X users cannot be assumed to be representative of the wider population, so the sentiment signal is more likely to capture the views of people who are already engaged with national policy debates than those of rural communities or lower-income groups affected by the same policies. The annotation pipeline also carries the tendencies of the three underlying LLMs, and cases where all three models agree on a wrong label will not be detected by the tiebreaker mechanism.

CONCLUSION

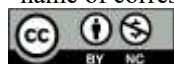
This study compared traditional classifiers and fine-tuned transformer-based models for sentiment analysis of the Indonesian K-12 education-related discussion on X (previously Twitter), posted from January 2022 until October 2025. The transformer-based models demonstrated better results than the traditional classifiers across all metrics. Specifically, IndoBERTweet showed the highest scores, scoring 0.884 accuracy and 0.859 macro F1 score, whereas the highest-scoring traditional classifier, which is logistic regression, scored 0.768 and 0.714 for those metrics, respectively. Most notable was the discrepancy between models' performance on minority classes, due to the lack of enough representational capabilities of TF-IDF-based models. Moreover, the results show that domain-specific pretraining plays a more important role than the model's size for social media sentiment classification. This finding points to representation quality driving performance more than scale.

Public criticism of K-12 education on Indonesian social media has been persistent and concentrated on specific policies rather than reform in general. Negative discourse centered on "kurikulum merdeka" and "sistem zonasi", while positive posts were attracted toward "transformasi pendidikan", which suggests the public broadly accepts the direction of change but takes issue with how it is being implemented. "Gaji guru" appeared among the top negative terms too. Teacher pay itself is derived from government policy, meaning its presence points to public concern over how the teaching profession is resourced, separate from curriculum or zoning specifically. For policymakers in the education sector, tracking sentiment around active policy names in real-time could serve as an early warning system for where implementation friction is concentrated.

Future work should address these gaps. Extending data collection to platforms with a broader demographic reach would test whether the sentiment patterns observed here hold outside the X user base. Additionally, an independent human double-annotation on a larger stratified sample would yield a reliable estimate of labeling accuracy and serve as a check on biases shared across the annotating LLMs. Applying aspect-based sentiment analysis to the same corpus would also allow opinions on curriculum to be separated from opinions on teacher pay or school zoning, preventing the grouping of all three into a single sentiment signal per tweet.

REFERENCES

- Adhim, N. F., & Cahyono, N. (2025). Optimization of IndoBERT for Sentiment Analysis of FOMO on Social Media Through Fine-Tuning and Hybrid Labeling. *Journal of Applied Informatics and Computing (JAIC)*, 9(6), 3786–3797. <https://doi.org/10.30871/jaic.v9i6.11686>
- Afandi, M. R., Siswanto, & Adalakun, N. O. (2026). Systematic Literature Review: The Role of Education in Reducing Economic Inequality and Poverty. *Sinergi International Journal of Education*, 4(1), 1–6. <https://doi.org/10.61194/education.v4i1.893>
- Aggarwal, P., Madaan, A., & Yang, Y. (2023). Let's Sample Step by Step: Adaptive-Consistency for Efficient Reasoning and Coding with LLMs. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 12375–12396. <https://doi.org/10.18653/v1/2023.emnlp-main.761>
- Bustillo, J. C. M., Arcos, J. R. D., MacEda, L. L., Elape, M. C., Cuartero, J. G., & Pojas, P. J. S. (2026). Multivariate Evaluation of the Free Higher Education Program Employing NLP-based Text Mining Techniques. *Proceedings of 9th International Conference on Inventive Computation Technologies, ICICT 2026*, 2111–2119. <https://doi.org/10.1109/ICICT68280.2026.11510879>
- Cahyawijaya, S., Lovenia, H., & Fung, P. (2024). LLMs Are Few-Shot In-Context Low-Resource Language Learners. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 405–433. <https://doi.org/10.18653/v1/2024.naacl-long.24>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- Dou, L., Liu, Q., Zeng, G., Guo, J., Zhou, J., Mao, X., Jin, Z., Lu, W., & Lin, M. (2024). Sailor: Open Language Models for South-East Asia. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 424–435. <https://doi.org/10.18653/v1/2024.emnlp-demo.45>
- Fauzan, E. N., Istikomah, N., Kurniawati, S., & Machmud, A. (2026). The Role of Human Capital Based on Education and Health against poverty in 34 provinces in Indonesia. *Dinasti International Journal of Education Management and Social Science*, 7, 2504–2515. <https://doi.org/10.38035/dijemss.v7i3>



- Fitriyani, V., Wibowo, F. A., Wulandhari, L. A., & Nabiilah, G. Z. (2025). A Comparison of Computational Approaches for Sentiment Analysis on Public Opinions About Education in Indonesia. *Proceedings of the 2025 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology, IAICT 2025*, 86–93. <https://doi.org/10.1109/IAICT65714.2025.11100431>
- Herra, N. P., Ramadhani, O. K., Mumtaz, N., Zakaria, D. M., & Maryamah. (2025). Sentiment and Social Network Analysis of X Social Media on the Implementation of the Merdeka Belajar Curriculum in Indonesia. *Journal of Advanced Technology and Multidiscipline*, 4(2), 68–79. <https://doi.org/10.20473/jatm.v4i2.71938>
- Hidayat, M. T., Suryadi, S., Latifannisa, N., Sari, S. N., & Rino, R. (2025). Evolution of The Education Curriculum in Indonesia. *Journal of Innovation in Educational and Cultural Research*, 6(2), 381–395. <https://doi.org/10.46843/jiecr.v6i2.1312>
- Hidayaturrahman, & Prawira, I. (2024). Leveraging Zero-Shot Learning in Large Language Models for Sentiment Analysis: A Comparative Study on the Indonesian Language. *Proceedings - 6th International Conference on Informatics, Multimedia, Cyber and Information System, ICIMCIS 2024*, 614–619. <https://doi.org/10.1109/ICIMCIS63449.2024.10956237>
- Koto, F., Lau, J. H., & Baldwin, T. (2021). INDOBERTWEET: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10660–10668. <https://doi.org/10.18653/v1/2021.emnlp-main.833>
- Kurnianingrum, D., Wijaya, A., Widjojo, R., Ratnapuri, C. I., Kartawinata, B. R., & Yustian, O. R. (2023). From Tweets to Insights: Analyzing Indonesian Public Sentiments on Education. *2023 International Conference on Informatics, Multimedia, Cyber and Information Systems, ICIMCIS 2023*, 644–647. <https://doi.org/10.1109/ICIMCIS60089.2023.10349070>
- Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159–174. <https://doi.org/10.2307/2529310>
- Medantoro, G. F. S., & Muljono. (2026). Comparative Analysis of IndoBERT and Classic Machine Learning Models for Sentiment Classification of Education Policy on Social Media X. *Journal of Applied Informatics and Computing (JAIC)*, 10(1), 548. <https://doi.org/10.30871/jaic.v10i1.11723>
- Nazamuddin, B. S., Syathi, P. B., Riswandi, R., Fitriyani, F., & Prayoga, A. E. (2026). The Dynamic Nexus between Human Capital and Poverty in Indonesia: Evidence from a Panel PMG-ARDL Approach. *Journal of the Knowledge Economy*, 17(2), 3794–3819. <https://doi.org/10.1007/s13132-025-02832-4>
- Niimi, J. (2025). A Simple Ensemble Strategy for LLM Inference: Towards More Stable Text Classification. *Natural Language Processing and Information Systems*, 189–199. https://doi.org/10.1007/978-3-031-97144-0_17
- OECD. (2023). *PISA 2022 Results (Volume I): The State of Learning and Equity in Education* (Vol. 1). OECD Publishing. <https://doi.org/https://doi.org/10.1787/53f23881-en>
- Opitz, J. (2024). A Closer Look at Classification Evaluation Metrics and a Critical Reflection of Common Evaluation Practice. *Transactions of the Association for Computational Linguistics*, 12, 820. https://doi.org/10.1162/tacl_a_00675
- Putri, F. P., Budiati, A. C., & Liestyasari, S. I. (2023). Konflik Sosial dalam Implementasi Kebijakan Sistem Zonasi. *Journal of Education Research*, 4(4), 2426–2438. <https://doi.org/10.37985/jer.v4i4.636>
- Riyanto, S., Sitanggang, I. S., Djatna, T., & Atikah, T. D. (2023). Comparative Analysis using Various Performance Metrics in Imbalanced Data for Multi-class Text Classification. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 14(6). <https://doi.org/10.14569/IJACSA.2023.01406116>
- Rosani, M., Dian Lestari, N., & Maria Valianti, R. (2025). Transformation of Education to Welcome the Golden Generation of Indonesia 2045. (*Jurnal Manajemen, Kepemimpinan, Dan Supervisi Pendidikan*), 10(1), 407–427. <https://doi.org/10.31851/jmksp.v10i1.18895>
- Sahabat-AI. (2024). *Gemma2 9B CPT Sahabat-AI v1 Instruct*. <https://huggingface.co/Sahabat-AI/gemma2-9b-cpt-sahabatai-v1-instruct>
- Sandra, L., Marcel, Gunarso, G., Fredicia, & Riruma, O. W. (2022). Are University Students Independent: Twitter Sentiment Analysis of Independent Learning in Independent Campus Using RoBERTa Base IndoLEM Sentiment Classifier Model. *2021 International Seminar on Machine Learning, Optimization, and Data Science, ISMODE 2021*, 249–253. <https://doi.org/10.1109/ISMODE53584.2022.9743110>
- Sidharta, S., Pranoto, H., Gasa, F. M., Kholis, N., & Ong, A. K. S. (2025). Analysis of Public Sentiment on the 17+8 People's Demands Issue Using IndoBERT and DistilBERT with LLM-Based Data Annotation. *2025 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, 575–580. <https://doi.org/10.1109/icimcis68501.2025.11327025>

- Smith-Mutegi, D., Mamo, Y., Kim, J., Crompton, H., & McConnell, M. (2025). Perceptions of STEM education and artificial intelligence: a Twitter (X) sentiment analysis. *International Journal of STEM Education*, 12(1). <https://doi.org/10.1186/s40594-025-00527-5>
- Sukhwal, P. C., & Kankanhalli, A. (2022). Determining containment policy impacts on public sentiment during the pandemic using social media data. *Proceedings of the National Academy of Sciences*. <https://doi.org/10.1073/pnas.2117292119>
- Wafda, A., Fudholi, D. H., & Nugraha, J. (2025). Aspect-Based Sentiment Analysis On Twitter Tweets About The Merdeka Curriculum Using IndoBERT. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, 10(3). <https://doi.org/10.33480/jitk.v10i3.5692>
- Wijaya, T. T., Hidayat, W., Hermita, N., Alim, J. A., & Talib, C. A. (2024). Exploring Contributing Factors To PISA 2022 Mathematics Achievement: Insights From Indonesian Teachers. *Infinity Journal*, 13(1), 139–156. <https://doi.org/10.22460/infinity.v13i1.p139-156>
- Wilie, B., Vincentio, K., Indra Winata, G., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding. *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 843–857. <https://doi.org/10.18653/v1/2020.aacl-main.85>
- Wu, J., Wang, X., & Jia, W. (2024). *Enhancing Text Annotation through Rationale-Driven Collaborative Few-Shot Prompting*. <http://arxiv.org/abs/2409.09615>
- Zhang, W., Chan, H. P., Zhao, Y., Aljunied, M., Wang, J., Liu, C., Deng, Y., Hu, Z., Xu, W., Chia, K., Li, X., & Bing, L. (2025). SeaLLMs 3: Open Foundation and Chat Multilingual Large Language Models for Southeast Asian Languages. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, 96–105. <https://doi.org/10.18653/v1/2025.naacl-demo.10>