Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.1, February 2026, pp. 548~557
e-ISSN: 2548-6861

Comparative Analysis of IndoBERT and Classic Machine Learning Models for Sentiment Classification of Education Policy on Social Media X

Gabriella Fani Suciarti Medantoro[1]*, Muljono[2]*
* Department Informatic Engineering, Dian Nuswantoro University, Semarang, Indonesia
111202214441@mhs.dinus.ac.id [1], muljono@dsn.dinus.ac.id[2]

Article Info

Article history:

Received 2025-11-08
Revised 2025-12-29
Accepted 2026-01-07

Keyword:

Sentyment Analysist,
Social Media X,
Machine Learning,
Implicit,
Education.

ABSTRACT

Leadership changes provide an opportunity for new education policies, generating complex public opinions on social media X that often contain implicit sentiments like satire, making automated analysis challenging. This study aims to address this challenge by conducting a comparative analysis to evaluate the effectiveness of the IndoBERT model in capturing nuanced, implicit sentiments compared to traditional machine learning classifiers (SVM, Naïve Bayes, Logistic Regression, KNN, and Random Forest). This research utilized a dataset of Indonesian-language tweets, collected via crawling. Data was pre-processed (cleaning, case folding, etc.) and labeled (positive/negative) using a hybrid Lexicon-LLM approach. The TF-IDF technique was used for feature extraction for the machine learning models, while IndoBERT used its internal tokenization. Models were evaluated using accuracy, precision, recall, and F1-score. The results showed that the IndoBERT model performed best with an accuracy score of 97%, significantly outperforming the other best machine learning models, namely Random Forest 95% and SVM 95%. This study concludes that the IndoBERT model is a superior and more robust solution for analyzing nuanced public sentiment on educational policies, demonstrating a greater ability to understand complex context and implicit language compared to traditional TF-IDF-based methods.

I. INTRODUCTION

Entering 2024, the national leadership transition in Indonesia has sparked a massive public discourse on the direction of state policy, with the education sector emerging as one of the primary focuses [1]. The period of national leadership transition in 2024, particularly between December 2023 and December 2024, which covers the campaign period to the early phase of the new administration, is not merely a change of authority figures, but a crucial moment for public evaluation of the sustainability of strategic programs such as the Merdeka Curriculum and sensitive issues related to the accessibility of education costs (UKT) and teacher welfare. The social media platform X (formerly Twitter) serves as the primary arena for this discourse, where millions of opinions are expressed in real time [2]. With Indonesia ranking as the fourth-largest user base of X globally, reaching 24.45 million users by April 2024, the platform has become an exceptionally rich data source for capturing public aspirations and views on the dynamics of national education.

However, automatically analyzing these millions of raw opinions presents significant technical challenges. Public opinion on the X platform is often expressed not just literally, but also through implicit language, sarcasm, and context-specific slang [3]. Implicit sentiment is defined as the expression of opinion that does not directly contain polarity adjectives (such as ‘bad’ or ‘disappointed’), but still carries emotional weight through contextual understanding or the use of metaphors. This phenomenon, often referred to as Post-level Implicit Sentiment Analysis (PISA), is particularly prevalent in social media discussions where users convey criticism through irony or satire [4]. Conventional sentiment classification models that rely solely on word-matching (such as TF-IDF) often fail to capture

these nuances, thus potentially misclassifying satirical criticism as positive sentiment, or vice versa [5]. Given this abundant volume of textual data, this study analyzes public sentiment regarding the state of education in Indonesia. By leveraging Sentiment Analysis, a sub-field of text mining and Natural Language Processing (NLP), this research classifies public opinions into positive or negative polarities.

Previous research has specifically highlighted the challenges in analyzing implicit sentiment, where meaning is not expressed literally. A study by Zhang et al. [6] demonstrated that many approaches (including basic deep learning) are still 'weak in capturing content-aware information,' especially when users express feelings through 'innuendo.' Their study affirmed the need for methods that can 'recognize context-aware information' to overcome the limitations of lexicon-based methods. Corroborating this, Li et al. [7] specifically investigated 'Implicit Aspect-Level Sentiment Analysis' and found that 'omitted expressions' significantly increase the difficulty of semantic understanding. Their research concluded that this problem requires 'deeper contextual understanding,' which they addressed using generative models (T5) and Graph Neural Networks.

Previous research has also consistently affirmed the effectiveness of Indonesian-specific pre-trained models, such as IndoBERT, in extracting public opinion from digital text data. A study by Mubaraq & Maharani [8] analyzed sentiment on climate change issues on Twitter using IndoBERT. Through hyperparameter fine-tuning, their research achieved an F1-Score of 95.6%, demonstrating the model's superiority in mapping public sentiment on social issues. IndoBERT's performance has also been validated in various other specific domains. Hidayat & Pramudita [9] analyzed sentiment towards post-pandemic online learning and achieved 87% accuracy (89% F1-score), proving IndoBERT's capability in education-based text classification. Novandian et al. [10] extended IndoBERT's application to detect cyberbullying, achieving an exceptional accuracy of 96.7%, confirming the model's ability to handle complex and sensitive classification tasks. Nonetheless, performance can vary; a study by Hakim et al. [11] on sentiment towards the Whoosh High-Speed Railway on platform X recorded evaluation metrics of 78%, indicating the model's sensitivity to data variations in the infrastructure domain.

These studies collectively underscore the significant potential of IndoBERT across various domains (education, transportation, financial services, and social issues). Nevertheless, several research gaps remain unaddressed. Most of this research focuses on binary (positive/negative) or single-label classification and often struggles to identify implicit sentiments, such as satire or sarcasm, which require a deeper contextual understanding. Furthermore, there is a limited number of studies that comprehensively compare the performance of advanced deep learning models like IndoBERT against more traditional machine learning models (e.g., Logistic Regression, Naïve Bayes, SVM, K-NN, and Random Forest) on the same dataset to measure their relative performance advantages.

The objective of this research is to conduct a comparative analysis to evaluate the performance of the IndoBERT Deep Learning model against classic Machine Learning models. To ensure high-quality ground truth labels and address potential bias, this study implements a hybrid labeling technique (Lexicon-GPT) with manual validation. It also aims to test the extent to which the IndoBERT method can serve as a solution to overcome the challenges of detecting implicit and satirical sentiment within public opinion on education policy.

II. METHOD

This phase contains the complete stages of the research. It begins with the process of crawling the dataset from X, followed by preprocessing and labeling the dataset. The prepared dataset is then divided into three parts: training, validation, and testing, with an 80:10:10 ratio. Subsequently, modeling is performed, along with an evaluation of each model. Finally, the best-performing model is tested using new data, as illustrated in the proposed method in Figure 1.

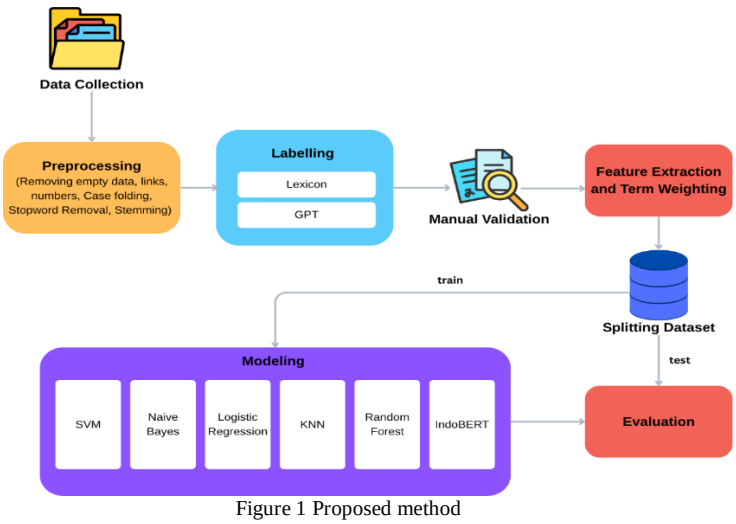

Figure 1 Proposed method

A. Data Collection

The data collection technique in this research used a crawling method. This method is the process of collecting data from the internet using a web crawler with the 'tweet-harverst' tool. In this stage, data was collected via the X application's API, using the assistance of Google Colab for the crawling process.

The data was taken from short message posts on the X application, which are better known as tweets. A tweet contains 280 words, allowing its users to share text, images, videos, links, and others in a short message format [12]. From the crawling process, 3,247 data entries were successfully obtained. All collected data was stored in tabular format, including information attributes such as *full_text*, *created_at*, *username*, and other interaction metrics for further processing in the pre-processing stage.

B. Preprocessing

This is the data preparation stage, which prepares raw data for the modeling phase. This stage involves cleaning the text of non-informative elements using regular expressions (regex), which includes removing empty data rows, URL links, numbers, symbols, punctuation marks, and other non-alphabetic characters, removing common words that have no significant meaning (stopwords), and converting all letters in the text to lowercase (case folding). In addition, this stage also involves removing links and numbers, replacing sensitive information with secure substitute values, and converting words to their base form (stemming). Apart from preparing the data, this process also improves the quality of the data to be processed. This decision was made to maintain the integrity and original characteristics of the language used by the community on social media X, so that the nuances of public expression regarding education policy are preserved in their original context.

C. Labeling

After the data preparation step is completed, the next step is the labeling of each data point. This labeling process is crucial for determining the class of each tweet. The labels are divided into two categories: 'pos' for positive and 'neg' for negative.

This study employs a hybrid labeling technique on the dataset, an approach that combines two different labeling methods:

1) *Lexicon:* A labeling technique that uses a dictionary as a linguistic source for the sentiment classification of each opinion [13].
2) *GPT:* This labeling utilizes a large language model (LLM), such as GPT, which automatically generates labels for the opinions [14].

The use of GPT aims to capture contextual nuances that cannot be detected by lexicon-based methods, especially in tweets containing satire, sarcasm, or hidden feelings that are often found in discussions of education policy. This hybrid tagging technique also aims to ensure that the tagging results in the dataset are accurate and precise.

D. Manual Validation

To ensure data label accuracy, the researcher(s) conducted manual validation on the automatically generated labels. This stage involved manually reviewing a sample of the data that had been labeled by the lexicon and GPT combination, matching the system-assigned sentiment labels ('positive' or 'negative') with the original context of the text, and correcting any labels that were deemed incorrect.

E. Feature Extraction and Term Weighting

This research utilizes BoW (Bag of Words) as feature extraction and TF-IDF (Term Frequency-Inverse Document Frequency) as word weighting. Data, which is still in a textual (discrete) format, needs to be converted into a numerical (continuous) representation to be processed by the algorithms. This stage transforms the collection of words into feature vectors that can be systematically measured and analyzed.

The BoW (Bag of Words) method works by constructing a vocabulary that contains all unique words from the dataset. Each tweet is then converted into a vector, where each element represents the frequency of occurrence of a word from that vocabulary.

To refine the feature representation, word weighting is performed using TF-IDF (Term Frequency-Inverse Document Frequency). This method assigns a higher weight to words that appear frequently within a single tweet but are rare across other tweets. Consequently, words considered more informative are given a greater value. The final output of this stage is the TF-IDF matrix.

F. Splitting Dataset

The dataset was divided into three main subsets: 80% was allocated as training data for the models, while the remainder served for evaluation. For traditional Machine Learning models, a commonly used proportion is 80% training data and 20% testing data. However, in the context of Transformer-based Deep Learning models such as IndoBERT, a more stringent splitting scheme was applied for effective optimization: 80% training data, 10% validation data, and 10% testing data.

G. Modeling

This study utilizes several types of classification models from Machine Learning and Transformer-based Deep Learning, including:

1) Support Vector Machine (SVM): This is included in Machine Learning classification under the supervised learning category. It is designed to process data in both linear and non-linear forms. SVM works by finding the best separating hyperplane that can distinguish between two

different classes *[15]*. In this study, SVM was configured using a kernel linear, the decision function for the separating hyperlane in SVM is formulated as follows:

$$f(x) = w \cdot x + b \quad (1)$$

Keterangan:
f(x): decision function.
w : weight vector, which determines the orientation of the hyperplane.
x : feature vector of the input data.
b : bias, which shifts the hyperplane from the origin.

2) *Naïve Bayes:* A statistical classification used to predict the probability of class membership. This classifier also provides the probability that a data point belongs to each possible class. In sentiment analysis, this algorithm is highly effective due to its high computational efficiency and its ability to handle large dimensions of text data. This study uses a variant of Multinomial Naïve Bayes, which is specifically optimized for data with discrete word frequencies resulting from TF-IDF or Bag of Words feature extraction. The basis for Naïve Bayes calculations using Bayes' Theorem is formulated as follows [16]:

$$P(y \mid X) = \frac{P(X \mid y)P(y)}{P(X)} \quad (2)$$

Description:
$P(y \mid X)$: The probability of class *y* given the observation data *X=(x1,x2,...,xn)*
$P(X \mid y)$: The probability of the observation *X* given that the class is *y*
P(y) : prior probability of class y.
P(X) : probability of the observation data X.

3) *Logistic Regression:* A classification model that fundamentally assesses the relationship between independent variables and a binary dependent variable [17]. This model can be extended to classify data into two or more classes. The probability that a data point belongs to a particular class is calculated using the logistic function. In this study, Logistic Regression was configured with the parameter max_iter=1000.

$$P(y = 1 \mid x) = \frac{e^{g(x)}}{1 + e^{g(x)}} = \frac{exp[g(x)]}{1 + exp[g(x)]} \quad (3)$$

Description:
$P(y = 1 \mid x)$: probability that the output is class 1 (positive) given the input *x*.
g(x) : predicted value.
e : euler's number

4) K-Nearest Neighbor: A machine learning algorithm that classifies objects based on the training data points that are closest in distance to a particular object *[18]*. This algorithm works by finding a number of *k* nearest neighbors of a new data point and determining the class label based on the majority vote of those neighbors. In this study, a value of k=5 was set to balance the smoothness of the decision boundary and sensitivity to noise in the text data. To determine the nearest neighbors, KNN measures the similarity between data points using the Euclidean Distance metric. The distance between two data points in an *n*-dimensional feature space is calculated using the following formula:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (4)$$

Description:
d(p,q) : euclidean distance between two data points, *p* and *q*.
p,q : data points in an n-dimensional feature space.
n : number of features (dimensions).
p_i, q_i : value of the *i*-th feature of data points *p* and *q*

5) *Random Forest:* This model is an ensemble (combination) of many decision trees, designed to achieve more stable and accurate predictions [19]. Random Forest addresses the correlation between decision trees, which can lead to overfitting, by implementing two forms of randomization: random sample selection and random feature selection. In this study, the Random Forest model was configured with the parameter n_estimators=100.

6) *IndoBERT:* Indonesian Bidirectional Encoder Representation from Transformer (IndoBERT) is a BERT transformer architecture model that was created in the Indonesian language [20]. This study uses the IndoBERT configuration, with IndoBERT model='indobert-base-p2' Hyperparameters Learning Rate=2e-5, Batch Size=32 Optimizer using AdamW. Epochs=20 (Early Stopping active).

H. Evaluation Model

In this research context, evaluation values (metrics) aim to identify the model's performance. The evaluation method in this study uses the confusion matrix. The confusion matrix itself is a table that provides a comparison between the predicted results and the actual results.

III. RESULTS AND DISCUSSION

A. Data Collection

The data collection phase, conducted from December 2023 to December 2024, yielded a total of 3,247 tweet entries. This specific timeframe is strategically significant as it captures public discourse during a period of national leadership transition, where fundamental issues such as curriculum changes, tuition fees (UKT), and teacher welfare became primary topics of debate on social media platform X. While the keyword 'pendidikan' (education) successfully gathered a substantial volume of data, an initial analysis of the raw dataset revealed significant challenges regarding data quality. As illustrated in Table I, the raw dataset was highly heterogeneous and contained a notable amount of noise.

TABLE I
SAMPLES OF RAW CRAWLED DATA AND CHARACTERIZATION ANALYSIS

full_text	Characteristic Analysis
Lagian sjk kapan jml org di sosmed itu punya andil gede terkait masalah pendidikan? Mau real life atau sosmed udah berisik banget soal ukt soal perbaikan pendidikan dr sistem sampe kesejahteraab guru pengajar dosen apa pernah didengerin??? Terdesak? Kasih atensi aja gak!	Substantive Opinion: Contains sharp systemic criticism regarding UKT and teacher welfare. High informational value for policy analysis.
Dukungan Kampus untuk Rizky Ridho: Tetap Menyala Capt!. Kapten Timnas Indonesia U-23 Rizky Ridho bakal absen membela Garuda Muda pada babak perebutan peringkat ketiga Piala Asia U-23 https://t.co/l59Nfb9jfi #PendidikanKesehatan #rizkyridho #timnasu23 via @beritajatimcom	Institutional News (Peripheral): While related to an educational institution, this tweet is primarily sports-related and lacks sentiment toward education policy. This represents 'thematic noise' for policy analysis.
@Cacaalagi @convomfs Biaya pendidikan udh langsung kepotong saat uang turun ke ATM mahasiswa	Specific Grievance: Reflects personal financial anxiety related to education administration.

The analysis of Table I demonstrates that the raw data encompasses more than just policy aspirations. Sample No. 2, for instance, shows that the keyword 'pendidikan' can capture institutional news that does not necessarily reflect public sentiment on policy. Beyond the primary *full_text* attribute, metadata such as *favorite_count*, *retweet_count*, and *reply_count* were also extracted to verify the level of public engagement with these varying types of content.

B. Preprocessing

The preprocessing stage effectively transformed unstructured tweet data into a more refined and consistent text corpus. This process is critical given that social media data from X (formerly Twitter) typically exhibits high levels of noise, including excessive punctuation and non-uniform capitalization. A comprehensive comparison of the results before and after preprocessing is presented in Table II.

TABLE II
COMPARISON OF PREPROCESSING RESULTS

Before preprocessing	*After preprocessing*
Lagian sjk kapan jml org di sosmed itu punya andil gede terkait masalah pendidikan? Mau real life atau sosmed udah berisik banget soal ukt soal perbaikan pendidikan dr sistem sampe kesejahteraab guru pengajar dosen apa pernah didengerin??? Terdesak? Kasih atensi aja gak!	sjk jml org sosmed andil gede kait didik real life sosmed udah berisik banget ukt baik didik dr sistem sampe kesejahteraab guru ajar dosen didengerin desak kasih atensi aja gak

Based on the results presented in Table II, several key points highlight the effectiveness of the preprocessing stage*:*

1) Noise Reduction: The removal of excessive punctuation (e.g., '???' and '!') and non-alphabetic characters was successfully executed. This process eliminates redundant feature ambiguity that is unnecessary for the model, allowing the analytical focus to shift entirely toward the lexical substance.

2) Feature Consistency through Case Folding: Uniformity achieved via lowercase conversion ensures that identical words with different capitalizations (e.g., 'Pendidikan' and 'pendidikan') are not treated as distinct features. This standardization directly enhances computational efficiency and reduces the dimensionality of the feature space.

3) Preservation of Social Media Linguistic Characteristics: Consistent with the research design established in Chapter II, abbreviations and slang—such as 'sjk' (sejak/since), 'jml' (jumlah/amount), and 'sosmed' (social media)—were intentionally preserved. Analysis indicates that maintaining this originality is vital for capturing the 'authentic voice' of users on the X platform.

4) Stemming Effectiveness: The conversion of affixed words into their root forms (e.g., transforming 'perbaikan' into 'baik') successfully reduced word variance. Consequently, the frequency of base words increased, which significantly benefits classical machine learning models utilizing TF-IDF weighting by allowing for more accurate recognition of sentiment patterns.

C. Labeling

The data labeling stage resulted in a distribution of 2,736 negative (neg) and 511 positive (pos) sentiment entries, as visualized in Figure 2. This high prevalence of negative sentiment highlights a significant critical trend in public discourse regarding education policies during the observed period. To ensure the integrity of this ground truth, a hybrid approach was implemented, significantly improving accuracy over traditional single-method labeling.

While the initial lexicon-based labeling provided a baseline, it proved insufficient for capturing the nuanced language of platform X, often failing to recognize sarcastic or context-dependent sentiments. By integrating a GPT-2 based Large Language Model (LLM), this study successfully resolved instances of implicit sentiment. The LLM's architecture, designed to extract meaning from the entire sentence structure rather than isolated tokens, directly addresses the research challenge of being 'weak in catching content-aware information.'

To further mitigate automated bias and ensure the highest reliability, a final manual validation was performed on all entries, resulting in the correction of 197 labels. This three-tier verification process—Lexicon, LLM, and Human

Expert—ensures that the dataset serves as a robust foundation for training the subsequent classification models, particularly in distinguishing between explicit praise and complex institutional criticism.

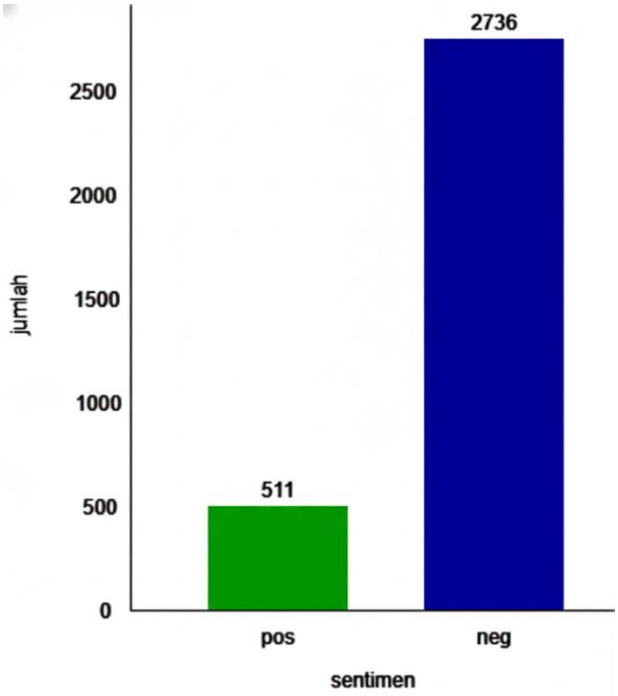

Figure 2 result labeling lexicon and gpt

TABLE III

EXAMPLES OF IMPLICIT AND SARCASTIC SPEECH

Text	***Label***	***Analysis***
balapan sama kalender pendidikan	Neg	Implicit: Reflects systemic pressure/stress. Lexicons might fail as no 'bad' words are present.
urus satu guru republik indonesia pgri perintah mudah ubah kurikulum didik sekolahsekolah	Neg	Sarcastic/Critical: Criticizes the frequency of curriculum changes.
please jgn jd hit tweet curhat takut siang aja alhamdulillah udh hasil selamat guys mudah percaya udh ketemu bg pendidikanpekerjaan oke ya guys asli tdk jamin	Pos	Explicit: Clear expressions of gratitude and success in education/career paths.

To ensure the integrity of the ground truth, a final manual validation was performed. Out of 3,247 automated labels, 197 labels were manually corrected. These corrections primarily involved nuanced sarcasm where even the LLM occasionally showed bias or ambiguity. This rigorous three-tier process (Lexicon, GPT, Manual) ensures that the models are trained on highly reliable data, directly mitigating the 'automated bias' concerns raised in previous studies.

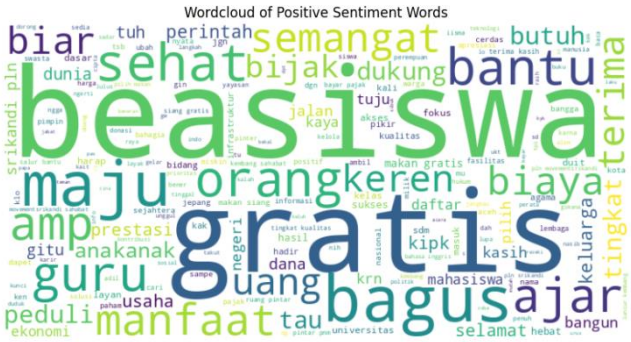

Figure 3 Word cloud positive sentiment

Figure 4 word cloud negative sentiment

The visualization of sentiment through Word Clouds, as depicted in Figures 3 and 4, provides a clear illustration of the distinct thematic focus within public discourse. Within the positive sentiment spectrum (Figure 3), the dominance of terms such as 'beasiswa' (scholarship), 'gratis' (free), and 'maju' (progress) indicates that public satisfaction is inextricably linked to educational accessibility and robust financial support. Conversely, the negative sentiment cloud (Figure 4) is characterized by the prominence of words like 'uang' (money), 'mahal' (expensive), 'biaya' (cost), and 'masalah' (problem). These findings confirm that economic barriers specifically the UKT (Tuition Fee) controversies highlighted stand as the primary drivers of public dissatisfaction. Ultimately, this strong thematic alignment between the empirical data and real-world policy challenges provides significant external validity to the dataset, ensuring it accurately reflects the socio-political climate and the genuine grievances of the public regarding the education system.

D. Manual Validation

Following the completion of the automated hybrid labeling process, a manual validation stage was conducted. This phase was essential to measure the accuracy of the automated outputs and to ensure the reliability of the ground truth before it was utilized for model training. Manual validation serves as a critical measure to mitigate 'automated bias,' particularly given the nature of social media data on platform X, which is often characterized by informal language, abbreviations, and implicit sentiments.

Based on the manual validation of the entire dataset

(3,247 entries), label discrepancies were identified in 197 data entries. Quantitatively, the initial automated labeling identified 511 entries as having positive sentiment. However, after a thorough re-examination by the researcher, the actual count of positive sentiment entries increased to 645. This shift represents a two-way adjustment (negative to positive and vice versa) to ensure that each tweet was classified based on its substantive context rather than the mere presence of certain keywords.

The necessity of these manual corrections highlights a critical finding regarding the limitations of automated language models in handling local dialects and complex sentence structures. Specific examples of these label corrections are presented in Table IV.

TABLE IV
COMPARATIVE ANALYSIS OF AUTOMATED AND MANUAL LABELLING CORRECTIONS

Preprocessed Tweet Text	***Automated Label***	***Manual Validation***	***Correction Analysis***
digi kipk biaya didik jt biaya hidup dasar klasternya semester jumlah brp dapat digi kipk dal jt semester jtbulan jt semester hasil jtbln	Neg	Pos	Technical Correction: The system detected the repeated word 'biaya' (cost) as an indicator of financial grievance. However, contextually, the tweet is informative regarding scholarship (KIP-K) details, thus reclassified as positive.
moga orang fakir miskin kurang biaya lanjut didik dapet kipk hidup beranta	Neg	Pos	Contextual Correction: The words 'fakir miskin' (the poor) and 'kurang biaya' (lack of funds) triggered a negative lexicon match. However, the full sentence expresses a hope/prayer for education aid, representing a positive public aspiration.

The analysis in Table IV reveals that automated methods often exhibit a bias toward words with literal negative connotations such as 'cost,' 'poor,' or 'lack' failing to grasp the user's underlying intent. This confirms that in education policy discourse, keywords related to economic barriers frequently reflect the dissemination of aid information or public aspirations rather than dissatisfaction. Through this manual refinement, the study achieved high contextual precision, resulting in a final distribution of 2,602 negative and 645 positive labels. This robust ground truth provides a solid foundation for evaluating the performance of classic Machine Learning models against IndoBERT's transformer architecture in handling complex linguistic nuances.

E. Feature Extraction and Term Weighting

The feature extraction process converted the 3,247 cleaned data entries (documents) into a numerical representation. From this entire corpus, a total of 9,766 unique terms were identified, forming the vocabulary. The weight for each term was then calculated using the TF-IDF scheme, resulting in a feature vector matrix with dimensions of (3247, 9766). This matrix was subsequently used as the input data for training and testing the classic machine learning models (SVM, Random Forest, Naïve Bayes, LR, and KNN).

It must be emphasized that this TF-IDF feature matrix was only used for the classic machine learning models. The IndoBERT model, which is based on the Transformer architecture, does not utilize this matrix. Instead, IndoBERT applies its own advanced sub-word tokenization (WordPiece) and internal embedding mechanisms, allowing it to process and understand sentence context directly from the raw text data.

F. Splitting Data

Before entering the modeling phase, the dataset was divided into several subsets to ensure the model performance evaluation was conducted objectively on unseen data. In accordance with the methodological design, two splitting schemes were applied, tailored to the model architecture: for the Machine Learning models (SVM, Naïve Bayes, Logistic Regression, Random Forest, and KNN), an 80% (2,598 data points) training and 20% (649 data points) testing ratio was used. Meanwhile, for the Deep Learning IndoBERT model, the split was 80% (2,597 data points) training, 10% (325 data points) validation, and 10% (325 data points) testing.

The selection of these ratios was based on common practices that have been proven effective and are considered standard in similar research for producing reliable evaluations. The allocation of a dedicated validation set for the IndoBERT model is a critical step in deep learning architectures, allowing for iterative monitoring of the training process and the prevention of overfitting. Thus, this split ensures that each model is evaluated using the most appropriate framework, thereby making the performance comparison results fair and valid.

G. Modeling

This section discusses the most critical phase, the previously processed and split dataset is used to train and test several classification models. The selected models include SVM, Naïve Bayes, Logistic Regression, Random Forest, KNN, and IndoBERT. Subsequently, the performance of these six methods will be evaluated and compared in Table V.

TABLE V
ACCURACY, PRECISION, RECALL, AND F1-SCORE TEST RESULTS

Model	Feature Extraction	Accuracy	Precision	Recall	F1-Score
IndoBERT	**WordPiece**	**0,9754**	**0,9400**	**0,9700**	**0,9500**
Random Forest	BoW	0,9569	0,9111	0,9359	0,9229
SVM	BoW	0,9569	0,9205	0,9205	0,9205
Logistic Regression	BoW	0,9354	0,9091	0,6667	0,7692
KNN	BoW	0,8815	0,8182	0,3429	0,4832
Naive Bayes	TF-IDF	0,8354	0,4912	0,5333	0,5114

As shown in Table V, *t*he experimental results demonstrate that the locally fine-tuned IndoBERT model achieved the highest performance across all metrics, with an Accuracy of 97.54% and an F1-Score of 0.9500. Unlike the classic models that rely on frequency-based representations (BoW or TF-IDF), IndoBERT's transformer-based architecture allows for a deeper understanding of bidirectional context. By fine-tuning the model on the specific discourse of Indonesian education policy, it successfully captured nuanced sentiments, including irony and domain-specific terminology (e.g., 'UKT,' 'KIP-K,' 'kurikulum'), which traditional models often misinterpreted.

Among traditional models, Random Forest and SVM with BoW provided competitive results with an F1-Score of 0.92, proving that raw word frequency is more effective than TF-IDF in capturing explicit sentiment in this dataset. Conversely, KNN and Naive Bayes experienced a significant decline in performance; although KNN recorded a high accuracy of 0.8815, the low Recall value of 0.3429 and F1-Score of 0.4832 indicate a strong bias towards the majority class (negative). The failure of classical models to detect positive sentiment underscores the importance of using F1-Score as the primary metric and confirms IndoBERT as the most robust solution for handling class imbalance in public policy sentiment analysis.

H. Evaluation

In this evaluation phase, the models' performance is analyzed to determine the most optimal one. Based on the test results presented in Table III in the previous sub-chapter, IndoBERT, Random Forest, and SVM were identified as the three top-performing models. To understand why these performance differences exist and to examine the error characteristics of each model, Figure 5 presents the confusion matrices for all six tested models. This confusion matrix analysis, particularly focusing on the top three models, will be used to visually dissect the counts of True Positives, True Negatives, and prediction errors (False Positives and False Negatives).

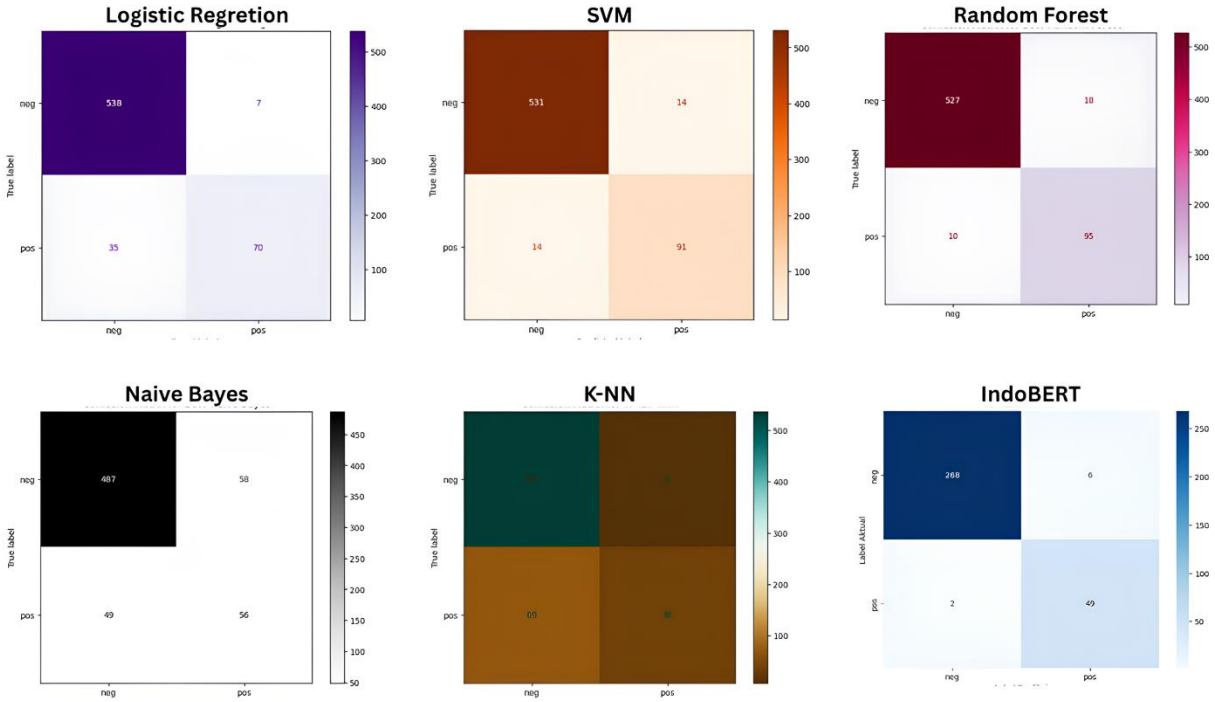

Figure 5 confusion matrix results in the model

The evaluation analysis is continued by examining the training process stability of the IndoBERT model, the best-performing model. This validation is crucial to ensure that the 97% accuracy score was achieved optimally and was not accompanied by overfitting.

Figures 6 and 7 provide visualizations of these training dynamics. Figure 7 plots the Learning Curve, which compares the training accuracy with the validation accuracy, while Figure 8 plots the Loss Curve, which shows the

comparison between the training loss and the validation loss at each epoch.

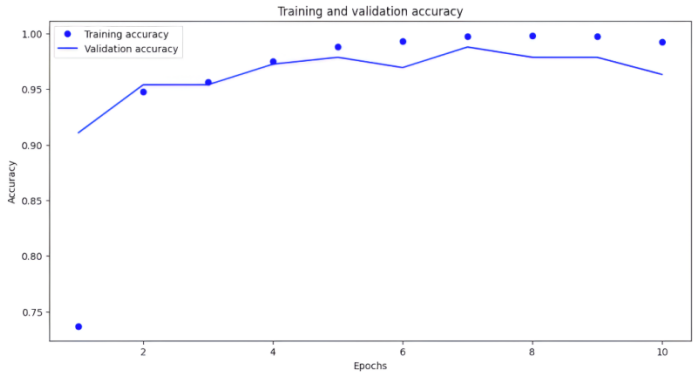

Figure 6 Accuracy learning curve

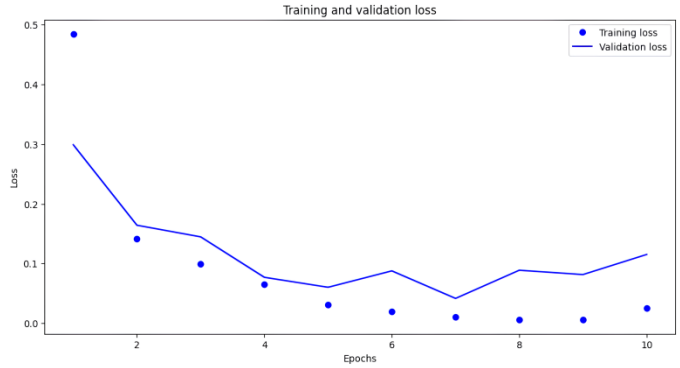

Figure 7 Loss learning curve

The performance analysis of IndoBERT as the best model is continued by validating its training process stability via Learning Curves. Figure 6 presents the Training Accuracy curve and the Validation Accuracy curve over 10 epochs. It is clearly observed that the validation accuracy curve maintains a very high value (above 95%) and moves in tandem with the training accuracy curve without any significant disparity. This harmonious movement between the two curves is a strong indication that the IndoBERT model successfully learned the data patterns effectively and was able to generalize that knowledge to previously unseen data (validation data), thus proving the model did not experience overfitting.

The model's stability is reinforced by the analysis of Figure 7, which displays the training Loss curve and the validation Loss curve. Since the initial epoch, a sharp decrease in loss occurred on both curves, indicating that the model learned quickly and efficiently. Both curves then stabilized at a very low value approaching zero in the subsequent epochs. This convergence and rapid stabilization at a low loss value further confirms that the hyperparameters and architecture of IndoBERT were optimized appropriately. The results of this Learning Curve analysis collectively validate the reliability and robustness of the IndoBERT model in handling and classifying complex sentiment within the education policy dataset.

IV. Conclusion

Based on the research results and discussion, it can be concluded that the IndoBERT method was successfully applied as a highly effective solution to address the challenges of classifying sentiment related to education policy that contains implicit meaning. This superior performance, evidenced by achieving the highest accuracy of 97% significantly surpassing other classic machine learning models demonstrates that IndoBERT's Transformer architecture is capable of deeply capturing the context and linguistic nuances within public opinion. Thus, this research successfully fulfills its objective of providing a reliable and accurate classification model.

Although the dataset used was imbalanced, this condition was intentionally maintained because the collected data was rich in sarcasm and implicit meanings that are sensitive to oversampling manipulations like SMOTE. As a recommendation for future research, it is suggested to expand the dataset variation on different education policy themes and to develop more advanced classification models. For example, implementing multi-label classification to identify specific types of implicit sentiment (such as sarcasm, irony, or slang) rather than just binary (positive/negative) labels.

References

[1] A. P. Putra, “Pemerintah, DPR, dan Penyelenggara Sepakati Pemilu Serentak 14 Februari 2024,” Kementerian Pendayagunaan Aparatur Negara dan Reformasi Birokrasi. Diakses: 6 Oktober 2025. [Daring]. Tersedia pada: https://menpan.go.id/site/berita-terkini/berita-daerah/pemerintah-dpr-dan-penyelenggara-sepakati-pemilu-serentak-14-februari-2024

[2] triya.andriyani, “Media Sosial jadi Sarana Penyampaian Pesan dan Kritik Sosial Kalangan Anak Muda,” Universitas Gadjah Mada. Diakses: 6 Oktober 2025. [Daring]. Tersedia pada: https://ugm.ac.id/id/berita/medis-sosial-jadi-sarana-penyampaian-pesan-dan-kritik-sosial-kalangan-anak-muda/

[3] I. Z. Hayati, R. Herdiana, dan S. Mulyani, “Gaya Bahasa Sindiran dalam Kolom Komentar Twitter Akun @tanyakanrl,” *Diksatrasia J. Ilm. Pendidik. Bhs. Dan Sastra Indones.*, vol. 8, no. 2, hlm. 556, Agu 2024, doi: 10.25157/diksatrasia.v8i2.15123.

[4] Z. Li, Y. Zou, C. Zhang, Q. Zhang, dan Z. Wei, “Learning Implicit Sentiment in Aspect-based Sentiment Analysis with Supervised Contrastive Pre-Training,” dalam *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, hlm. 246–256. doi: 10.18653/v1/2021.emnlp-main.22.

[5] A. O. Thakare, N. R. Soora, L. Jena, A. R. Singh, A. P. H, dan R. Pachlor, “Hate Speech Detection in Social Media Data Using Big Data Analytics*,” dalam *2025 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)*, Bhubaneswar, India: IEEE, Mei 2025, hlm. 1–9. doi: 10.1109/ASSIC64892.2025.11158561.

[6] Q. Zhang, X. Zhu, J. L. Zhao, dan L. Liang, “Discovering signals of platform failure risks from customer sentiment: the case of online P2P lending,” *Ind. Manag. Data Syst.*, vol. 122, no. 3, hlm. 666–681, Mar 2022, doi: 10.1108/IMDS-05-2021-0308.

[7] X. Li, X. Wang, C. Yao, dan Y. Li, “Graph-enhanced implicit aspect-level sentiment analysis based on multi-prompt fusion,” *Sci. Rep.*, vol. 15, no. 1, hlm. 17460, Mei 2025, doi: 10.1038/s41598-025-02609-4.

[8] M. F. Mubaraq dan W. Maharani, "Sentiment Analysis on Twitter Social Media towards Climate Change on Indonesia Using IndoBERT Model," *J. MEDIA Inform. BUDIDARMA*, vol. 6, no. 4, hlm. 2426, Okt 2022, doi: 10.30865/mib.v6i4.4570.

[9] M. N. Hidayat dan R. Pramudita, "Analisis Sentimen Terhadap Pembelajaran Secara Daring Pasca Pandemi Covid-19 Menggunakan Metode IndoBERT," *Inf. Manag. Educ. Prof. J. Inf. Manag.*, vol. 8, no. 2, hlm. 161, Jan 2024, doi: 10.51211/imbi.v8i2.2719.

[10] Y. D. Novandian *dkk.*, "IndoBERT-based Indonesian Cyberbullying Detection with Multi-stage Labeling," dalam *2024 International Seminar on Application for Technology of Information and Communication (iSemantic)*, Semarang, Indonesia: IEEE, Sep 2024, hlm. 515–521. doi: 10.1109/isemantic63362.2024.10762553.

[11] G. Hakim, T. N. Fatyanosa, dan A. W. Widodo, "Analisis Sentimen Masyarakat terhadap Kereta Cepat Whoosh pada Platform X menggunakan IndoBERT".

[12] A. I. Mu'alifah dan . S., "Self Disclosure Pada Pengguna Media Sosial Twitter (Studi Kualitatif Self Disclosure Pada Pengguna Media Sosial Twitter)," *J. SIGNAL*, vol. 11, no. 1, hlm. 01–14, Apr 2023, doi: 10.33603/signal.v11i1.7510.

[13] Y. Fauziah, B. Yuwono, dan A. S. Aribowo, "Lexicon Based Sentiment Analysis in Indonesia Languages : A Systematic Literature Review," *RSF Conf. Ser. Eng. Technol.*, vol. 1, no. 1, hlm. 363–367, Des 2021, doi: 10.31098/cset.v1i1.397.

[14] G. Colavito, F. Lanubile, N. Novielli, dan L. Quaranta, "Leveraging GPT-like LLMs to Automate Issue Labeling," dalam *Proceedings of the 21st International Conference on Mining Software Repositories*, Lisbon Portugal: ACM, Apr 2024, hlm. 469–480. doi: 10.1145/3643991.3644903.

[15] S. R. K. W. Tommy Rustandi, D. Suhaedi, dan Y. Pemanasari, "Pemetaan Hyperplane Pada Support Vector Machine," *Bdg. Conf. Ser. Math.*, vol. 3, no. 2, hlm. 109–119, Agu 2023, doi: 10.29313/bcsm.v3i2.8187.

[16] A. N. Sihananto dan H. Maulana, "Studi Literatur Tentang Performa Naïve Bayes Dalam Klasifikasi Data," *Pros. Semin. Nas. Inform. Bela Negara*, vol. 2, hlm. 132–135, Nov 2021, doi: 10.33005/santika.v2i0.134.

[17] E. Roflin, F. Riana, E. Munarsih, Pariyana, dan I. A. Liberty, *Regresi Logistik Biner dan Multinomial*. PT Nasya Expanding Management, 2023. [Daring]. Tersedia pada: https://books.google.co.id/books?id=FOi3EAAAQBAJ&lpg=PR1&ots=jrivgTSzA0&lr&hl=id&pg=PR4#v=onepage&q&f=false

[18] S. Zhang, X. Li, M. Zong, X. Zhu, dan D. Cheng, "Learning *k* for kNN Classification," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 3, hlm. 1–19, Mei 2017, doi: 10.1145/2990508.

[19] M. W. Nugroho, "Analisis Performa Algoritma Random Forest dalam Mengatasi Overfitting pada Model Prediksi," *J. JTIK J. Teknol. Inf. Dan Komun.*, vol. 9, no. 4, hlm. 1562–1571, Okt 2025, doi: 10.35870/jtik.v9i4.4236.

[20] D. Sebastian, H. D. Purnomo, dan I. Sembiring, "BERT for Natural Language Processing in Bahasa Indonesia," dalam *2022 2nd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*, Bandung, Indonesia: IEEE, Des 2022, hlm. 204–209. doi: 10.1109/ICICyTA57421.2022.10038230.