Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.4, August 2026, pp. 3262~3269
e-ISSN: 2548-6861

Comparative Analysis Of Automatic Labeling With Cohen's Kappa Validation For Cyberbullying On Tiktok Using Pre-Trained Transformers

David Rian Prabowo [1]*, Khoiriya Latifah [2], Nugroho Dwi Saputro [3]***
* Department of Informatics, Faculty of Engineering and Informatics, Universitas PGRI Semarang, Semarang, Indonesia
22676001@upgris.ac.id [1], khoiriyalatifah@upgris.ac.id [2], nugputra@upgris.ac.id [3]

Article Info

Article history:

Received 2026-05-26
Revised 2026-07-24
Accepted 2026-08-08

Keyword:

Cyberbullying,
InSet Lexicon,
TikTok,
Transformer,
Zero Shot Classification.

ABSTRACT

The rapid growth of TikTok users in Indonesia increases the risk of cyberbullying, making manual content moderation inefficient. This study compares two automatic labeling methods, InSet Lexicon and Zero-Shot Classification, to address the problem of limited labeled data in cyberbullying detection. A total of 5,864 comment data were collected through scraping techniques and processed through comprehensive text preprocessing stages, including slang normalization. To evaluate labeling reliability, a validation test using Cohen's Kappa Score metric was conducted against a human annotated Gold Standard of 1,141 comments. The results show that Zero-Shot Classification achieves high reliability with a Kappa score of 0.9132 (almost perfect agreement), outperforming InSet Lexicon which drops to 0.2585 (fair agreement) due to lexical rigidity and high false positives on casual slang. The automatically labeled datasets were balanced using Random Oversampling (for the Zero-Shot Classification labeling scenario) and split via a stratified 80:10:10 ratio to fine-tune IndoBERT and RoBERTa. On independent test data, IndoBERT trained on Zero-Shot labels delivers the best performance, reaching an Accuracy and F1-Score of 91.10%, outperforming RoBERTa under the same scenario (85.83%). Conversely, training on InSet Lexicon labels reduces performance, limiting IndoBERT to an 86.94% F1-Score and RoBERTa to 80.21%. This study concludes that using Zero-Shot Classification and fine-tuning IndoBERT is more optimal for application to informal social media comment moderation systems.

I. INTRODUCTION

The advancement of digital technology has brought significant changes to the dynamics of social interaction, with social media becoming the primary space for the global community to communicate, share content, and express themselves. TikTok, as one of the most popular social media platforms among teenagers and adults, provides an interaction medium based on short videos and comments. Data obtained from DataReportal at the end of 2025 recorded that active TikTok users in Indonesia had reached more than 180 million people [1]. However, this ease of access and massive connectivity also increases the potential for deviant behavior, such as cyberbullying involving acts of aggression, humiliation, or threats online [2]. The impact of cyberbullying is very serious, ranging from mental health problems such as depression and anxiety, decreased academic achievement, to extreme cases such as suicide attempts [3]. Given the volume of comments on TikTok produced massively in seconds, manual content moderation efforts become very inefficient and time-consuming. Therefore, an intelligent automatic detection system capable of understanding the context of human language is needed to create a safer digital space [2].

One of the fundamental challenges in building an artificial intelligence-based detection system is the availability of large, accurate labeled datasets. The manual labeling process by human experts (ground truth) is often a

major obstacle because it requires very high costs and time [4]. As an innovative solution, this study adopts two automatic labeling approaches: the InSet Lexicon-based approach and Zero-Shot Classification. The first approach used for automatic labeling is InSet Lexicon, a sentiment lexicon specifically developed for Indonesian to calculate the polarity weight of words in a sentence and then determine sentiment based on the calculated sentence weighting. InSet Lexicon was chosen due to its customization focusing on the context of the local Indonesian language [5], where the words were collected from the Twitter platform by Fajri Koto and Gema Y. Rahmaningtyas [6]. Although previous evaluations showed satisfactory performance with an accuracy rate of 65.78%, this lexicon-based method has its own challenges in capturing more complex contextual meanings on social media. As an innovative comparison to the lexicon method, this study implements Zero-Shot Classification using the MoritzLaurer/mDeBERTa-v3-base-mnli-xnli model. Zero-Shot Classification allows text classification into specific categories without requiring specific training data on that domain, instead utilizing broad semantic knowledge acquired during the pre-training phase. This model was chosen for its ability to perform Natural Language Inference (NLI) in more than 100 languages, including Indonesian. With optimal processing speed, this model offers high efficiency for automatic labeling processes on large-scale datasets [7].

To ensure the reliability of the automation results, a labeling agreement measurement phase was conducted using the Cohen's Kappa Score algorithm. This testing is crucial to evaluate the stability of the labeling function and measure the extent of the system's consistency in classifying data into appropriate categories [8]. The use of Cohen's Kappa Score is considered superior to simple percentage agreement because this metric mathematically considers the factor of agreement that may occur by chance agreement. Based on the classification standards proposed by Landis and Koch [9] for categorical data, values above 0.81 are designated as almost perfect agreement. This standard becomes a methodological validity parameter to ensure that the resulting labels have a very high level of reliability before being used as training data.

In this study, the validated dataset was used to train and evaluate two state-of-the-art Pre-trained Transformers architectures, namely indobenchmark/indobert-base-p1 and w11wo/indonesian-roberta-base-sentiment-classifier. The indobenchmark/indobert-base-p1 model was chosen because it is built on the IndoBERT architecture, which has been specifically optimized using a giant-scale corpus called Indo4B containing approximately 4 billion words. Because this corpus also includes data from various social media, this model is proven capable of handling the characteristics of colloquial Indonesian text that is full of slang and non-standard terms. The robustness of this basic IndoBERT architecture is demonstrated in sentiment analysis evaluation, where this model achieved very high performance metrics (average macro F1-score), reaching 94.94% [10]. Meanwhile, the w11wo/indonesian-roberta-base-sentiment-classifier model was chosen because it is a derivative of the RoBERTa architecture known for its robustness. The RoBERTa architecture was developed by optimizing training methods rigorously, such as training models longer with much larger batch sizes, and applying dynamic masking techniques to training data. This fundamental optimization makes RoBERTa-based architectures have very strong and robust classification performance. On Indonesian language processing benchmarks, RoBERTa-based architectures have proven to have very strong sentiment classification performance with an F1-Score of 92% [11].

Several previous studies have proven the effectiveness of various methods in text labeling and classification. In the Transformer-based model category, Rizkia et al. [12] through systematic comparison found that Zero-Shot Classification-based automatic labeling using IndoBERTweet achieved the highest performance with an F1-Score of 0.9802. Meanwhile, Laurer et al. [13] offered a solution to annotation data scarcity through the mDeBERTa-v3-base-mnli-xnli model fine-tuned with the Natural Language Inference (NLI) dataset. This model is capable of performing multilingual zero-shot classification in 100 languages with cross-lingual accuracy reaching 79.8% to 80.8%, making it an efficient State-of-the-Art (SOTA) standard without dependence on language-specific training data [14]. As a comparison, dictionary-based (lexicon) methods continue to show their relevance. Fathoni et al. [15] proved that the combination of InSet Lexicon with Support Vector Machine (SVM) produced an average accuracy of 87.83%. Regarding dataset integrity, Riccosan et al. [16] reported a Cohen's Kappa value of 0.5679 on social media emotion labeling due to label ambiguity. A higher value was achieved by Prastiawan et al. [17] in analyzing the response to the Free Nutritious Meal (MBG) program with a Cohen's Kappa score of 0.71. The effectiveness of other Transformer variants was also validated by Wijaya et al. [18] using the indonesian-roberta-base-sentiment-classifier model in evaluating teacher teaching performance, achieving 87% accuracy. On the other hand, Jayanti and Rohman [2] tested the indobert-base-p1 model for cyberbullying detection on TikTok with 70.66% accuracy and an ROC-AUC score of 0.7969, demonstrating strong discriminatory ability in informal language.

Based on the literature review, research that comprehensively evaluates two automatic labeling approaches InSet Lexicon vs Zero-Shot Classification validated by Cohen's Kappa and measures their impact on the performance of two Transformer architectures IndoBERT and RoBERTa in an end-to-end manner in the context of cyberbullying detection on TikTok with informal Indonesian language characteristics is still not available. This study aims to: (1) Comparing the automatic labeling performance between InSet Lexicon and Zero-Shot Classification using Cohen's Kappa metric against the Gold Standard of rigorous human annotation results; (2) evaluate the impact of these two labeling methods on the classification performance of IndoBERT and RoBERTa

models in detecting cyberbullying on TikTok; and (3) identify the most optimal combination of labeling method and Transformer architecture for informal Indonesian text data. By integrating methodological validation through the Cohen's Kappa metric, this research is expected to provide recommendations for the most effective and reliable labeling strategies to handle non-formal language characteristics in the Indonesian social media ecosystem.

II. METHOD

This study integrates the CRISP-DM methodology with the NLP Project Lifecycle to develop a reliable cyberbullying detection system. The methodology focuses primarily on the use of Automatic Labeling InSet Lexicon and Zero-Shot Classification, validated using Cohen's Kappa, as well as the fine-tuning of Transformer models IndoBERT and RoBERTa. The complete research process is illustrated in Figure 1.

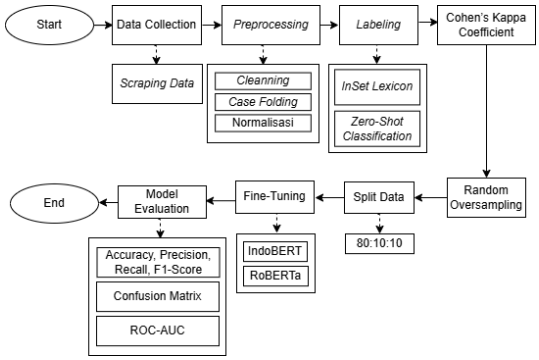

Figure 1. Research Method

A. Data Collection

Data collection was conducted using a scraping technique on the TikTok app via Apify. This process utilized the TikTok Comments Scraper actor, which runs on the Apify cloud platform to automatically retrieve data [19]. A total of 5,864 user comments in Indonesian that potentially contained cyberbullying were successfully collected. The scraped data was then extracted from the Apify dataset and saved in CSV format for further processing.

To maintain user privacy and comply with platform rules, ethical data filtering was performed. Data collection was limited to the comments section of public posts. After the data was downloaded, a complete anonymization process was performed: all personal information such as account names (usernames), user IDs, profile URLs, and direct mentions (@) were removed from the dataset. The final dataset retained only the pure comment text, devoid of any account owner identities.

B. Text Preprocessing

Text preprocessing was conducted to eliminate linguistic noise and improve data quality before labeling and for model training [20]. The preprocessing pipeline included several key steps:

- Data Cleaning: Removal of irrelevant symbols, numbers, mentions, hashtags, URLs, emojis, and punctuation marks.
- Case Folding & Normalization: All text is converted to lowercase, followed by the normalization of repeated characters.
- Slang Normalization: Conversion of non-standard terms or TikTok-specific slang into formal Indonesian vocabulary.

This study did not use stemming or stopwords, consistent with the findings of Khairani et al. [21] that the Transformer model yields more optimal results without these steps. Table 1 illustrates an example of the preprocessing results.

TABLE I
EXAMPLE OF TEXT PREPROCESSING RESULTS

Preprocessing	Text
Data	@raya_fams: sumpah suka banget pas nerya nari, luwes banget gerakannya terus positif vibes
Case Folding & Data Cleaning	sumpah suka banget pas nerya nari luwes banget gerakannya terus positif vibes
Slang Normalization	sumpah suka banget pas nerya nari luwes banget gerakannya terus positif suasana

C. Labeling

This study uses a binary classification, in which data is grouped into two classes: Cyberbullying (1) and Non-Cyberbullying (0). Comments are classified as Cyberbullying if they contain abusive language, harassment, hate speech, or direct insults that attack a person personally. Meanwhile, comments outside these categories, such as casual interactions, ordinary criticism, general discussions, or expressions of praise, are classified into the Non-Cyberbullying class.

Data Automatic labeling of 5,864 comment data points was performed using two different methods:

1. InSet Lexicon: An approach based on an Indonesian sentiment lexicon developed by Koto et al. [6]. Labeling is performed by identifying words in the comments that match lexicon entries, then calculating a total score based on the weight of the word values. Sentences with a final score below the threshold of 0.5 are automatically labeled as cyberbullying (1), and the rest as non-cyberbullying (0).
2. Zero-Shot Classification: An approach based on the pre-trained MoritzLaurer/mDeBERTa-v3-base-mnli-xnli model, using the Natural Language Inference (NLI) mechanism to predict labels without additional training [7]. The label parameters used are `candidate_labels = ["bullying", "non-bullying"].` This process is run with a single-label configuration, utilizing GPU acceleration on a CUDA 0 device.

D. Cohen's Kappa Score

To test the reliability of the automatic labeling results, a sample of 1,141 comment entries was selected to serve as the test data (gold standard). This sample was then manually

labeled by humans to serve as the ground truth. Cohen's Kappa coefficient (κ) was calculated to assess the consistency between the machine labels and the human labels using the formula:

$$\kappa = \frac{Po - Pe}{1 - Pe} \quad (1)$$

Where Po is the observed percentage of agreement and e and Pe is is the percentage of agreement that occurs by chance. [23].

E. Random Over Sampling & Split Data

The results of the automatic labeling revealed distinct data characteristics. The Zero-Shot Classification approach yielded a class imbalanced dataset, comprising 4,099 non-bullying data points (70%) and 1,765 bullying data points (30%). This issue was addressed by applying Random Oversampling (ROS) to the training data to balance the number of data points across both classes (totaling 8,198 data points) [24]. Meanwhile, in the InSet Lexicon track, the data distribution is relatively balanced with 3,480 non-bullying data points (59%) and 2,384 bullying data points (41%), so no oversampling process is required.

The datasets in both scenarios were then split using the Stratified Sampling technique with a 80:10:10 ratio and locked using the parameter random_state=42 [25]. For the InSet Lexicon track, 4,691 training data points, 586 validation data points, and 587 test data points were obtained. In the Zero-Shot Classification track (post-oversampling), 6,558 training data, 820 validation data, and 820 test data were obtained. All test data are independent and are not included in the model training phase [26].

F. Fine Tuning Model

Fine tuning was performed on two Transformer architectures: IndoBERT (indobenchmark/indobert-base-p1) and RoBERTa (w11wo/indonesian-roberta-base-sentiment-classifier) [10], [11]. Both models were selected to conduct a comparative analysis of the effectiveness of Transformer architectures in classifying cyberbullying comments on TikTok, as well as to evaluate the impact of data labeling characteristics on the models' final performance. The modeling process began with tokenization using WordPiece for IndoBERT and Byte-Pair Encoding (BPE) for RoBERTa [14], [27]. This was followed by converting tokens into numerical vectors, padding to equalize input length, and applying an attention mask to direct the model's focus on original tokens [28].

Fine tuning was performed on Google Colab using the Transformers library (Hugging Face). Hyperparameter optimization included adjusting the learning rate, batch size, and dropout rate. The hyperparameters used during training are presented in Table 2.

TABLE 2
HYPERPARAMETER CONFIGURATION

Parameter	Value
Learning Rate	1×10^{-5}
Batch Size	16
Max Epochs	10
Dropout Rate	0.3
Weight Decay	0.05
Label Smoothing	0.1
Early Stopping	Patience 2

Early stopping was applied to prevent overfitting by monitoring validation loss [29]. Labels were mapped binarily: 1 for bullying and 0 for non-bullying.

G. Evaluation Model

The evaluation phase of this study was structured to comprehensively assess model performance using test data [30]. The effectiveness of each model was evaluated using standard classification metrics: accuracy, precision, recall, and F1-score. Among these, the F1-score was adopted as the primary evaluation metric due to its robustness in handling class imbalance.

Let TP, FP, FN, and TN represent true positives, false positives, false negatives, and true negatives, respectively [31]. The metrics are defined as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$\text{Precission} = \frac{TP}{TP+FP} \quad (3)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

$$\text{F1-score} = 2* \frac{\text{Precission} * \text{Recall}}{\text{Precission}+\text{Recall}} \quad (5)$$

To compare the four scenarios (combinations of InSet Lexicon and Zero-Shot Classification with IndoBERT and RoBERTa), macro-average and weighted-average values were calculated based on the classification report to ensure fair evaluation across both classes [32].

In addition, ROC-AUC Curve analysis was conducted to assess the model's discriminatory power [33] between bullying and non-bullying classes.

III. RESULTS AND DISCUSSION

A. Labeling Results and Cohen's Kappa Validation

Automatic labeling of 5,864 TikTok comments was performed using two methods: InSet Lexicon and Zero-Shot Classification, with the categories non-bullying (0) and bullying (1). The label distributions generated by the two methods are shown in Figure 2.

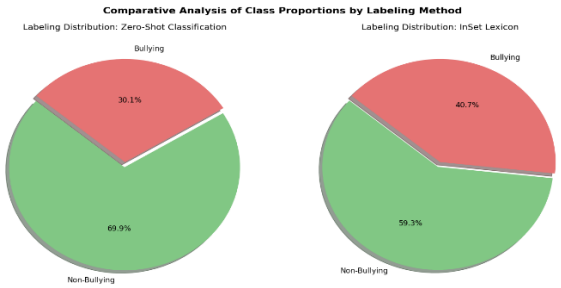

Figure 2. Sentiment Distribution of Labeling Results

As shown in Figure 2, the Zero-Shot Classification tends to be more selective, with non-bullying labels dominating at 69.9% (4,099 data) compared to bullying at 30.1% (1,765 data). Meanwhile, InSet Lexicon produces a relatively balanced distribution, with non-bullying posts accounting for 59.3% (3,480 data) and bullying posts for 40.7% (2,384 data). This difference in distribution stems from the fundamental logic of the two algorithms, Zero-Shot Classification employs a Transformer-based semantic context approach that is more selective in identifying bullying, whereas InSet Lexicon relies on the literal accumulation of negative word weights.

To illustrate the classification differences between the two methods, Table 3 presents a comparison of scores for several comments.

TABLE 3
EXAMPLE OF SCORE COMPARISON BY COMMENT

Cleaned Comment	Polarity Score	Confidence Score	Label InSet	Label ZSC
lebay dasarrr alay	9	0.8608	1	1
nevan saya emang selalu ganteng	-3	0.7487	1	0
maklum bapaknya kan banyak hutangnya jadi terpaksa anaknya cari yg gratisan	-12	0.622	1	0
sumpah suka banget pas nerya nari, luwes banget gerakannya terus positif suasana	9	0.8302	0	0
anak abah emang aneh sdm rendah semua	-5	0.6	1	1

When tested using the Gold Standard dataset of 1,141 human-labeled comments, the results were starkly different. The Zero-Shot Classification method achieved a Cohen's Kappa value of 0.9132 (Almost Perfect Agreement) according to the Landis and Koch classification [9]. Meanwhile, InSet Lexicon dropped sharply to 0.2585 (Fair Agreement).

The low Kappa value for the InSet Lexicon is due to this dictionary method frequently producing false positives, causing the number of bullying classifications to balloon to 2,384 data points. Because it rigidly matches words one to one without understanding sentence structure, InSet immediately labels comments containing negative-sounding words like "gila" "mati" or "aura" as cyberbullying. However, in the informal language of TikTok, these words are often used as hyperbole, jokes, or even compliments (e.g., "Sumpah gila luwes banget.") As a result, the accuracy of the lexicon's labels does not align with human interpretation.

B. Model Performance Comparison

Model performance was evaluated across four scenarios, each combining two labeling methods (InSet Lexicon and Zero-Shot Classification) with two Transformer architectures (IndoBERT and RoBERTa). All models were evaluated using independent test data with the metrics accuracy, precision, recall, and F1-score. The evaluation results are presented in Table 4.

TABLE 4
MODEL PERFORMANCE EVALUATION RESULTS

Scenario	Accuracy	Precission	Recall	F1-Score
IndoBERT ZSC	0.9110	0.9110	0.9110	0.9110
RoBERTa ZSC	0.8585	0.8607	0.8585	0.8583
IndoBERT InSet	0.8739	0.8694	0.8694	0.8694
RoBERTa InSet	0.8109	0.8061	0.7992	0.8021

As shown in Table 4, the IndoBERT model using the Zero-Shot Classification labeling scenario demonstrated the highest performance, with accuracy and F1-score reaching 91.10%. This result is significantly higher than that of the InSet Lexicon scenario on the same model, which only reached 86.94%. This finding demonstrates that the Zero-Shot Classification-based automatic labeling method (mDeBERTa-v3) is far more effective at capturing linguistic context on the TikTok platform compared to the lexicon-based method (InSet Lexicon).

The RoBERTa model demonstrated stable performance but still lagged behind IndoBERT in both scenarios, with a highest score of 85.83% in the Zero-Shot Classification scenario. RoBERTa's low performance in the InSet Lexicon scenario 80.21% reinforces the indication that this model requires a more contextual labeled dataset such as that generated by Zero-Shot Classification to function optimally. IndoBERT's advantage is influenced by its pre-training dataset (Indo4B), which is relevant to the characteristics of informal social media language in Indonesia, making the model more adaptive to slang vocabulary and non-standard sentence structures commonly found on TikTok.

To evaluate the model's discriminative ability in distinguishing between bullying and non-bullying classes, a Receiver Operating Characteristic (ROC) curve analysis was

conducted. Figure 3 presents a comparison of the ROC curves for the four scenarios along with their respective Area Under the Curve (AUC) values.

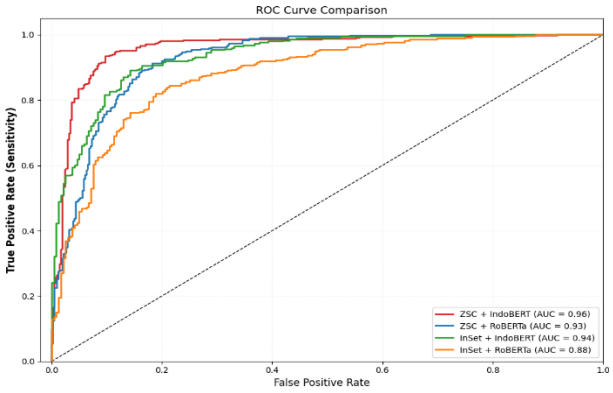

Figure 3. ROC Curve Comparing the Four Scenarios

As shown in Figure 3, the highest AUC value was achieved by the Zero-Shot Classification with IndoBERT scenario at 0.96, followed by InSet Lexicon with IndoBERT (0.94), Zero-Shot Classification with RoBERTa (0.93), and the lowest by InSet Lexicon with RoBERTa (0.88). The AUC value approaching 1 in the Zero-Shot Classification with IndoBERT scenario confirms that this model has excellent ability in distinguishing between bullying and non-bullying comments. The higher AUC values in the scenarios using IndoBERT (0.96 and 0.94) compared to RoBERTa (0.93 and 0.88) further reinforce the finding that the IndoBERT architecture is superior for the task of detecting cyberbullying in Indonesian text.

C. Confusion Matrix Analysis and Training Curve

To gain a deeper understanding of the classification performance characteristics, we visualized the confusion matrix for the best model, namely IndoBERT with Zero-Shot Classification labeling. The results are presented in Figure 4.

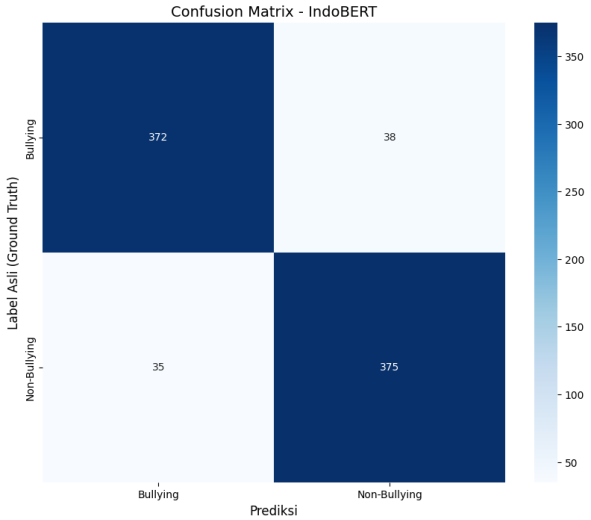

Figure 4. Confusion Matrix IndoBERT and Zero Shot Classification

Based on the confusion matrix in Figure 4, out of the 820 test data utilized there were 372 True Positives (correct classifications for the bullying class), while there were 375 True Negatives (correct classifications for the non-bullying class). There were 35 False Positives (non-bullying data incorrectly classified as bullying). False Negatives (bullying data classified as non-bullying) were recorded at 38 data points. In total, the model made 747 correct predictions (372 + 375) and 73 misclassifications (35 + 38). The exact accuracy rate calculated directly from this matrix is 747/820 = 0.9110, which directly matches the performance scores reported in Table 4.

An error analysis of the 73 misclassified comments reveals distinct structural patterns and system limitations based on informal social media linguistics. The False Positives, totaling 35 instances, occurred primarily when the model struggled to differentiate heavy profanity or local slurs from non-malicious boundaries. In these cases, teenagers frequently used aggressive terms for friendly banter, joking contexts, or expressions of intense shock on TikTok, causing the model to over-interpret the tokens and incorrectly flag them as actual cyberbullying. On the other hand, the 38 False Negatives happened when the model missed indirect cyberbullying strategies. These toxic comments heavily relied on subtle ironies, structural sarcasm, or regional cultural analogies without using explicit slurs, allowing the abusive intent to pass through the Transformer layers undetected.

The IndoBERT model training process was monitored via loss and accuracy curves to ensure the model learned effectively and remained stable. Figure 5 and 6 presents the training loss, validation loss, and validation accuracy curves over 10 epochs.

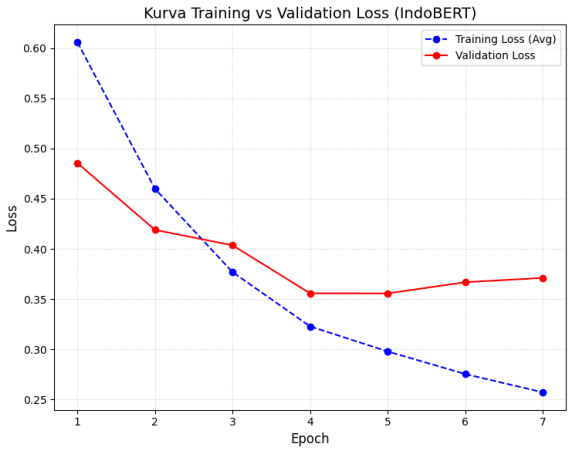

Figure 5. Training vs Validation Loss Curve (IndoBERT)

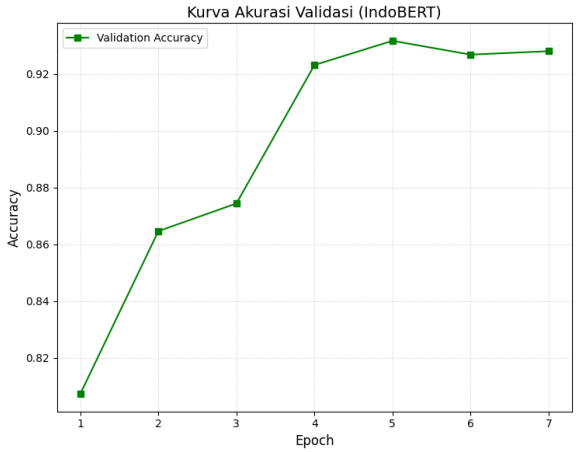

Figure 6. Validation Accuracy Curve (IndoBERT)

According to Figure 5, the training loss shows a steady decrease from 0.606 at epoch 1 to 0.257 at epoch 7. Concurrently, the validation loss drops from 0.485 at epoch 1 to its lowest points of 0.356 at epochs 4 and 5, before stabilizing at 0.371 by epoch 7. The parallel movement of both curves without any sudden spikes confirms that the model did not suffer from overfitting. In line with this trend, Figure 6 shows that the validation accuracy increases progressively from 0.807 at epoch 1 to its peak of 0.931 at epoch 5, locking its final performance at 0.928 at epoch 7. This upward trajectory confirms that the fine-tuning process successfully optimized the model parameters to recognize cyberbullying text patterns effectively.

Computational overhead metrics were tracked transparently during the execution phase across all scenarios using a single Tesla T4 GPU on Google Colab. During the automated annotation stage, the evaluation revealed a substantial efficiency gap between the two approaches. The Zero-Shot Classification pipeline required 5 minutes and 58 seconds to process the total population data at an average speed of 19.43 it/s. In contrast, the string-matching mechanics of the InSet Lexicon executed via CPU almost instantly, completing the entire dataset in just 2 seconds at a rate of 1,956.40 it/s.

During the fine-tuning phase, the layer-freezing strategy for encoder parameters successfully optimized the training timeline. Several configurations triggered early stopping constraints before reaching the maximum 10-epoch limit as the validation loss stabilized. The exact empirical runtime records are compiled in Table 5.

TABLE 5
COMPUTATIONAL EFFICIENCY AND TRAINING DURATION SUMMARY

Scenario	Total Trained Rows	Stopped Epoch	Training Time	Evaluation Status
IndoBERT ZSC	6.556 (After ROS)	Epoch 7/10	6m 8s	Early Stopping
RoBERTa ZSC	6.556 (After ROS)	Epoch 10/10	10m 5s	Full Epoch
IndoBERT InSet	4.691	Epoch 7/10	14m 48s	Early Stopping
RoBERTa InSet	4.691	Epoch 9/10	19m 3s	Early Stopping

Table 5 indicates that even though the InSet Lexicon pathway had fewer data instances (4,691 rows), its training process per iteration ran slower at 2.31 it/s compared to the Zero-Shot pathway. Within the same dataset brackets, the IndoBERT architecture consistently demonstrated faster convergence times and shorter total training runtimes than RoBERTa. This empirical baseline confirms the structural efficiency of the localized language model when handling text classification tasks in Indonesian.

IV. CONCLUSION

Based on the results of the research and discussions conducted, it can be concluded that the automatic labeling method using Zero-Shot Classification is significantly more reliable than the InSet Lexicon for detecting cyberbullying in TikTok comments. This is evidenced by an Cohen's Kappa score of 0.9132 (almost perfect agreement) compared to 0.2585 (fair agreement). The IndoBERT model, trained using data labeled via Zero-Shot Classification, achieved the best performance with an Accuracy and F1-Score of 91.10%, outperforming the RoBERTa model in the same scenario (85.83%) as well as the IndoBERT model labeled with the InSet Lexicon (86.94%). An AUC value of 0.96 further underscores the model's superiority.

The practical implications of this study indicate that the sentence-context-based labeling approach via Zero-Shot Classification is more suitable for social media data types, and the IndoBERT model developed with a local Indonesian language corpus is more sensitive to recognizing the characteristics of TikTok's informal language. This system is recommended for the development of automated comment moderation tools on digital platforms.

For future research, it is suggested to expand the scope of data sources to include other social media platforms to assess the model's generalization ability, conduct statistical significance tests (such as the McNemar test), and evaluate the system's speed performance (inference latency) in real-time comment filtering scenarios (real-time).

REFERENCES

[1] "Digital 2026: Indonesia," DataReportal – Global Digital Insights. Accessed: Apr. 23, 2026. [Online]. Available: https://datareportal.com/reports/digital-2026-indonesia

[2] H. D. Jayanti and A. Rohman, "Cyberbullying Detection in Indonesian TikTok Comments Using IndoBERT with Fairness Evaluation," *J. Inf. Syst. Inform.*, vol. 8, no. 1, pp. 907–927, Mar. 2026, doi: 10.63158/journalisi.v8i1.1448.

[3] E. Asalnaije, Y. Bete, M. A. Manikin, R. A. Labu, S. A. D. Tira, and Y. P. Lian, "Bentuk-Bentuk Cyberbullying Di Indonesia,"

Innov. J. Soc. Sci. Res., vol. 4, no. 4, pp. 6465–6473, Jul. 2024, doi: 10.31004/innovative.v4i4.12471.

[4] F. Husain, H. Alostad, and H. Omar, "Q8SentiLabeler Automated Labeling System for Arabic Sentiment Analysis," ResearchGate. Accessed: Apr. 23, 2026. [Online]. Available: https://www.researchgate.net/figure/Q8SentiLabeler-System-Architecture_fig4_378081726

[5] H. Firda, P. Putra, N. R. Oktadini, P. E. Sevtiyuni, and A. Meiriza, "Comparison of Rating-based and Inset Lexicon-based Labeling in Sentiment Analysis using SVM (Case Study: GoBiz Application Reviews on Google Play Store)," *SISTEMASI*, vol. 14, no. 2, p. 516, Mar. 2025, doi: 10.32520/stmsi.v14i2.4795.

[6] F. Koto and G. Y. Rahmaningtyas, "Inset Lexicon: Evaluation of A Word List for Indonesian Sentiment Analysis in Microblogs," in *Proc. Int. Conf. Asian Lang. Process. (IALP)*, 2017, pp. 391–394. doi: 10.1109/IALP.2017.8300625.

[7] P. He, J. Gao, and W. Chen, "DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing," Mar. 24, 2023, *arXiv*: arXiv:2111.09543. doi: 10.48550/arXiv.2111.09543.

[8] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960, doi: 10.1177/001316446002000104.

[9] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, Mar. 1977.

[10] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," Oct. 08, 2020, *arXiv*: arXiv:2009.05387. doi: 10.48550/arXiv.2009.05387.

[11] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 26, 2019, *arXiv*: arXiv:1907.11692. doi: 10.48550/arXiv.1907.11692.

[12] A. S. Rizkia, W. Wufron, and F. F. Roji, "Analisis Sentimen Coretax: Perbandingan Pelabelan Data Manual, Transformers-Based, dan Lexicon-Based pada Performa IndoBERT: Sentiment Analysis of Coretax: A Comparison of Manual, Transformers-Based, and Lexicon-Based Data Labeling on IndoBERT Performance," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 5, no. 3, pp. 1037–1048, Jul. 2025, doi: 10.57152/malcom.v5i3.2151.

[13] "MoritzLaurer/mDeBERTa-v3-base-mnli-xnli · Hugging Face." Accessed: May 04, 2026. [Online]. Available: https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-mnli-xnli

[14] T. Wolf *et al.*, "Transformers: State-of-the-art Natural Language Processing," Jul. 14, 2020, *arXiv*: arXiv:1910.03771. doi: 10.48550/arXiv.1910.03771.

[15] Muhammad Fernanda Naufal Fathoni, Eva Yulia Puspaningrum, and Andreas Nugroho Sihananto, "Perbandingan Performa Labeling Lexicon InSet dan VADER pada Analisa Sentimen Rohingya di Aplikasi X dengan SVM," *Modem J. Inform. Dan Sains Teknol.*, vol. 2, no. 3, pp. 62–76, Jul. 2024, doi: 10.62951/modem.v2i3.112.

[16] null Riccosan, K. E. Saputra, G. D. Pratama, and A. Chowanda, "Emotion dataset from Indonesian public opinion," *Data Brief*, vol. 43, p. 108465, Aug. 2022, doi: 10.1016/j.dib.2022.108465.

[17] K. H. Prastiawan and D. Yuniarto, "Analisis Sentimen Publik terhadap Program Makan Bergizi Gratis dengan Algoritma Naive Bayes," *RIGGS J. Artif. Intell. Digit. Bus.*, vol. 4, no. 4, pp. 5412–5419, Dec. 2025, doi: 10.31004/riggs.v4i4.3652.

[18] P. A. A. Wijaya, I. M. Dika Anggara, I. K. Hendra Trinium Jaya, G. Indrawan, and I. M. Agus Oka Gunawan, "Comprehensive Analysis of Teacher Teaching Performance Through Sentiment and POS Tagging," *J. Ilm. Merpati Menara Penelit. Akad. Teknol. Inf.*, vol. 12, no. 3, p. 147, Dec. 2024, doi: 10.24843/JIM.2024.v12.i03.p02.

[19] R. Erama, "Pemanfaatan Platform Cloud Google Colab Untuk Scraping Komentar Tiktok Pada Konten Gorontalo sebagai Dasar Analisis Respons Warganet," *J. Appl. Eng. Sci.*, vol. 1, no. 2, pp. 124–134, Dec. 2025, doi: 10.65177/jaes.v1i2.38.

[20] A. F. Setyawan and R. I. Nugraha, "Optimizing GPT And IndoBERT for sentiment analysis and consumer trend prediction on Lazada product reviews," *JIKO J. Inform. Dan Komput.*, vol. 8, no. 2, pp. 113–120, Jul. 2025, doi: 10.33387/jiko.v8i2.10066.

[21] U. Khairani, V. Mutiawani, and H. Ahmadian, "Pengaruh Tahapan Preprocessing Terhadap Model Indobert Dan Indobertweet Untuk Mendeteksi Emosi Pada Komentar Akun Berita Instagram," *J. Teknol. Inf. Dan Ilmu Komput.*, vol. 11, no. 4, pp. 887–894, Aug. 2024, doi: 10.25126/jtiik.1148315.

[22] N. Jain, H. Suh, S. Adeyinka, L. Roseman, and A. Allsop, "Multi-LLM Thematic Analysis with Dual Reliability Metrics: Combining Cohen's Kappa and Semantic Similarity for Qualitative Research Validation," Feb. 14, 2026, *arXiv*: arXiv:2512.20352. doi: 10.48550/arXiv.2512.20352.

[23] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochem. Medica*, vol. 22, no. 3, pp. 276–282, Oct. 2012.

[24] D. R. Alfinsyah and B. P. Hartato, "Evaluating the Impact of Random Over Sampling on IndoBERT Performance for Indonesian Sentiment Analysis," vol. 9, no. 6, pp. 3270–3282, 2025, doi: https://doi.org/10.30871/jaic.v9i6.11488.

[25] Z. Bami, A. Behnampour, A. Bora, and H. Doosti, "A New Flexible Train-Test Split Algorithm, an approach for choosing among the Hold-out, K-fold cross-validation, and Hold-out iteration," Jan. 01, 2026, *arXiv*: arXiv:2501.06492. doi: 10.48550/arXiv.2501.06492.

[26] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?," Feb. 05, 2020, *arXiv*: arXiv:1905.05583. doi: 10.48550/arXiv.1905.05583.

[27] A. Vaswani *et al.*, "Attention Is All You Need," Aug. 02, 2023, *arXiv*: arXiv:1706.03762. doi: 10.48550/arXiv.1706.03762.

[28] A. M. Putri, W. K. Nofa, and D. A. P. Hapsari, "Penerapan Metode BERT Untuk Analisis Sentimen Ulasan Pengguna Aplikasi Segari Di Google Play Store," vol. 4, no. 1, 2025, doi: https://doi.org/10.56127/juit.v4i1.1902.

[29] M. Jazzar and T. Duridi, "A Comprehensive Review of Machine Learning and Deep Learning Techniques for Cyberbullying Detection | Request PDF," in *ResearchGate*. doi: 10.1007/978-981-96-2182-8_1.

[30] S. S. Sabrina, D. F. Shiddieq, and F. F. Roji, "Comparative Analysis of SVM and BERT for Sentiment and Sarcasm Detection in the Boycott of Israeli Products on Platform X," *Sinkron*, vol. 9, no. 2, pp. 872–883, May 2025, doi: 10.33395/sinkron.v9i2.14723.

[31] D. Ananda, I. Budi, A. B. Santoso, and A. A. Qureshi, "Sentiment analysis of public health app reviews using IndoBERT and XLM-RoBERTa: A study on SATUSEHAT mobile app," *JIKO J. Inform. Dan Komput.*, vol. 8, no. 3, pp. 161–172, Nov. 2025, doi: 10.33387/jiko.v8i3.10083.

[32] E. Boiy and M.-F. Moens, "A machine learning approach to sentiment analysis in multilingual Web texts," *Inf. Retr.*, vol. 12, no. 5, pp. 526–558, Oct. 2009, doi: 10.1007/s10791-008-9070-z.

[33] Arif Fitra Setyawan, Amelia Devi Putri Ariyanto, Fari Katul Fikriah, and Rozaq Isnaini Nugraha, "Analisis Sentimen Ulasan iPhone di Amazon Menggunakan Model Deep Learning BERT Berbasis Transformer," *Elkom J. Elektron. Dan Komput.*, vol. 17, no. 2, pp. 447–452, Dec. 2024, doi: 10.51903/elkom.v17i2.2150.