JIKO (Jurnal Informatika dan Komputer) Accredited KEMDIKTISAINTEK, No.0173/C3/DT.05.00/2025
Vol. 8, No. 3, Desemberl 2025, pp. 161-172 p-ISSN: 2614-8897
DOI: 10.33387/jiko.v8i3.10083 e-ISSN: 2656-1948

SENTIMENT ANALYSIS OF PUBLIC HEALTH APP REVIEWS USING INDOBERT AND XLM-ROBERTA: A STUDY ON SATUSEHAT MOBILE APP

Dimas Ananda[1], Indra Budi[2], Aris Budi Santoso[3], Ali Adil Qureshi[4]

[1,2,3] Faculty of Computer Science, Universitas Indonesia, Depok, 16424, Indonesia
[4]Department of Computer Science, Khwaja Fareed University of Engineering & Information Technology, Pakistan
Email: [*1] dimas.ananda21@ui.ac.id , [2]indra@cs.ui.ac.id, [3]aris.budi@ui.ac.id , [4]shah87061@gmail.com

(Received: 12 June 2025, Revised: 30 June 2025, Accepted: 17 October 2025)

Abstract

Sentiment analysis is a key method for deriving insights from user-generated content, particularly in evaluating public satisfaction with digital health services. This study conducts a comparative analysis of sentiment polarity classification models on 34,178 Indonesian-language reviews from SATUSEHAT Mobile, a national health application by the Indonesian Ministry of Health. The dataset was manually annotated into positive, neutral, and negative classes. Three model categories were evaluated: classical machine learning (Support Vector Machine, XGBoost), baseline neural networks (Multilayer Perceptron, Convolutional Neural Network), and pretrained transformer-based models (IndoBERT, XLM-RoBERTa). All models were trained using stratified 5-fold cross-validation and tested on a held-out set. Results show that transformer-based models significantly outperform others in all metrics. IndoBERT achieved the highest weighted F1-score (0.8555), followed closely by XLM-RoBERTa (0.8552). Despite the similar average performance, XLM-RoBERTa exhibited the lowest performance variance across folds, making it the most stable and effective model overall. Statistical validation using Friedman and Nemenyi tests confirmed these differences as significant. However, all models struggled with neutral sentiment detection due to data imbalance. Although computationally more expensive than IndoBERT, XLM-RoBERTa offers superior robustness for sentiment classification in Indonesian health-related text. These findings support the integration of transformer-based sentiment monitoring into public health dashboards to enable timely, data-driven service improvements.

Keywords: *Sentiment analysis, IndoBERT, XLM-RoBERTa, SATUSEHAT, machine learning*

**Corresponding Author: Dimas Ananda*

1. INTRODUCTION

Mobile health (mHealth) applications have become pivotal in modern public healthcare delivery, especially in the post–COVID-19 era. These platforms enable digital vaccination records, symptom tracking, and teleconsultations, contributing to more efficient and accessible healthcare systems. In Indonesia, the Ministry of Health's SATUSEHAT Mobile App serves as a centralized digital health ecosystem, connecting millions of citizens and healthcare providers through integrated health data services.

Evaluating user feedback on such public service applications is essential for maintaining service quality and informing policy refinement. Sentiment analysis an application of natural language processing (NLP) enables the automated classification of opinions into categories such as positive, neutral, or negative. This technique allows policymakers and system developers to identify public perceptions, detect recurring problems, and improve responsiveness to user concerns.

Earlier research often employed classical machine learning methods such as Support Vector Machines (SVM) and Extreme Gradient Boosting (XGBoost), which offer strong interpretability and reliable performance [1], [2]. However, these models rely heavily on manually engineered features and struggle to capture the nuanced linguistic patterns and code-mixed expressions common in Indonesian mobile app reviews. Neural architectures such as Multilayer Perceptrons (MLP) and Convolutional Neural Networks (CNN) have improved this by learning representations directly from text [3], yet their

capacity to model long-range contextual dependencies remains limited.

Recent advances in transformer-based pretrained language models, such as IndoBERT [4] and XLM-RoBERTa (XLM-R) [5], have dramatically enhanced sentiment classification performance. These models are trained on extensive Indonesian and multilingual corpora, allowing them to effectively understand both syntactic and semantic contexts. In particular, XLM-RoBERTa demonstrates superior generalization across multilingual and cross-domain tasks, as it is trained on a larger, more diverse dataset than IndoBERT, enabling better handling of mixed-language inputs frequently found in Indonesian user reviews. This makes XLM-RoBERTa especially robust for real-world sentiment analysis where users may blend English and Indonesian expressions.

Despite these advancements, most existing studies focus on single-model or descriptive sentiment evaluations in mHealth contexts. Structured comparisons across classical machine learning, baseline neural networks, and pretrained transformer models especially within Indonesian government health service platforms remain scarce. Moreover, while transformer-based models generally yield high overall accuracy, they often struggle with correctly identifying neutral sentiments. This occurs because neutral reviews typically contain limited emotional cues and overlap linguistically with both positive and negative expressions, making them more challenging to distinguish.

This study contributes to the field by conducting a comprehensive comparative evaluation of sentiment polarity classification on SATUSEHAT Mobile app reviews across three model families: (a) Classical Machine Learning Models: SVM and XGBoost, (b) Baseline Neural Network Models: MLP and CNN, and (c) Pretrained Transformer Models: IndoBERT and XLM-RoBERTa. The findings provide empirical evidence of model performance trade-offs, highlighting the strengths of transformer-based architectures in processing Indonesian and multilingual content, while also identifying areas such as neutral sentiment classification that warrant further methodological refinement. These insights contribute to both academic research in NLP and the practical improvement of Indonesia's digital public health services.

2. RESEARCH METHOD

This section outlines the methodology employed in conducting sentiment polarity classification on Indonesian public health app reviews. To ensure a systematic and reproducible approach, the study adopts a multi-stage workflow encompassing four main phases: Data Collection, Data Preparation, Modeling, and Evaluation. Each phase comprises several key steps as illustrated in the research workflow (see Fig. 1), and is described in detail in the subsequent subsections.

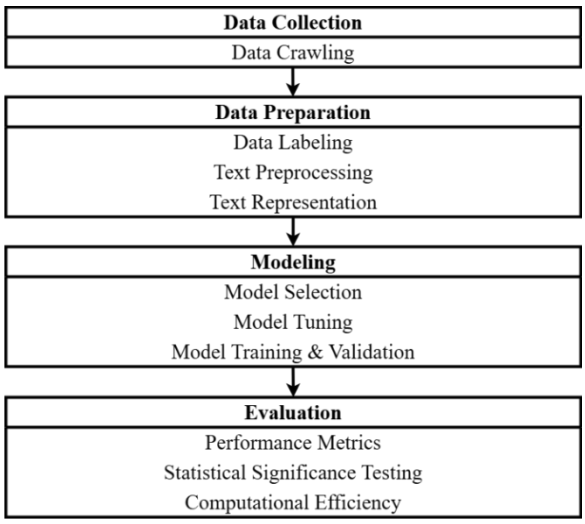

Figure 1. Research Methodology Workflow

The dataset used in this study comprises 34,178 user reviews of the SATUSEHAT Mobile application, collected from the Google Play Store using the google-play-scraper Python library. The crawler was configured to retrieve Indonesian-language reviews posted between March 1, 2023, and December 31, 2024. Only the review text and timestamp were retained for analysis, aligning with the study's focus on temporal sentiment dynamics. This automated data acquisition method has been widely adopted in mobile app feedback mining [16]. A monthly aggregation of review counts was visualized to illustrate the distribution of user engagement over time.

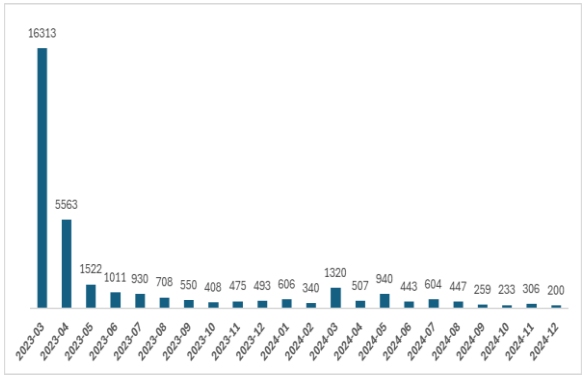

Figure. 2. Monthly distribution of SATUSEHAT app reviews from March 2023 to December 2024

Following the data collection stage, the dataset underwent preparation for supervised sentiment classification, beginning with a manual data labeling process. Each user review was annotated into one of three sentiment categories positive, neutral, or negative based on predefined polarity criteria reflecting user satisfaction with public health digital services. Labeling was conducted by two independent annotators over a period of 90 days, following a structured annotation guideline to ensure semantic consistency. The sentiment annotation scheme is summarized in Table 1.

Table 1. Sentiment Annotation Guidelines

Sentiment Class	Definition
Positive	Reviews indicating satisfaction, appreciation, or overall favorable response.
Neutral	Reviews containing factual or ambiguous content without clear sentiment.
Negative	Reviews indicating dissatisfaction, criticism, or reports of poor experience.

Manual annotation was chosen to ensure high-quality sentiment supervision, as accurate, context-aware labeling has been widely recognized as critical for achieving robust model performance on real-world unstructured data, particularly in the health domain where textual nuances influence interpretation [17], [18], [19].

After labeling, the dataset was subjected to standard text preprocessing steps to ensure data consistency and prepare the text for vectorization and model training. Preprocessing was guided by established NLP practices for Indonesian texts [20], [21], [22], [23] and included the following:

- Normalization and Cleaning, which involved lowercasing all text and removing punctuation, emojis, HTML tags, and special characters to reduce noise and ensure structural uniformity.
- Tokenization, where sentences were segmented into individual tokens to allow for discrete analysis by the learning algorithms.
- Stopword Removal, in which high-frequency but semantically uninformative words (e.g., "yang", "dan", "di") were excluded to improve signal quality.
- Stemming, performed using the Sastrawi algorithm a widely adopted stemmer for Bahasa Indonesia was applied to reduce inflected words to their root forms. This step simplified the vocabulary space and enhanced the generalizability of the model.

To better illustrate the preprocessing pipeline, Table 2. presents an example of stepwise transformation applied to a representative review.

Table 2. Example Of Stepwise Text Preprocessing Pipeline Applied To A Satusehat Review

Stage	Output Example
Original Text	Pelayanan lambat, aplikasi sering error pas buka fitur vaksin. Gimana ini?
Normalized & Cleaned	pelayanan lambat aplikasi sering error pas buka fitur vaksin gimana ini
Tokenized	[pelayanan, lambat, aplikasi, sering, error, pas, buka, fitur, vaksin, gimana, ini]
Stopword Removed	[pelayanan, lambat, aplikasi, sering, error, buka, fitur, vaksin]
Stemmed	[layan, lambat, aplikasi, sering, error, buka, fitur, vaksin]

This preprocessing pipeline was implemented prior to text representation, ensuring that input data was clean, token-consistent, and linguistically normalized before entering the subsequent modeling stage.

After the preprocessing stage, textual data was transformed into numerical representations suited for different model families. Three distinct representation strategies were adopted to align with the respective requirements of classical machine learning, baseline neural networks, and pretrained transformer-based models.

- TF-IDF Vectorization. For traditional classifiers such as Support Vector Machine (SVM) and XGBoost, this study employed Term Frequency–Inverse Document Frequency (TF-IDF) to quantify word importance within documents. Let t be a term, d a document, and "D" the entire corpus, the TF-IDF weight is calculated as:

$$\text{TF-IDF}(t,d) = \text{tf}(t,d) \times \log\left(\frac{N}{\text{df}(t)}\right)$$

  where "tf" (t,d) is the frequency of term t in document d, "df" (t) is the number of documents containing t, and N is the total number of documents. This method results in a sparse high-dimensional feature space, where discriminative terms carry higher weights in documents related to bugs or functionality issues [24].
- Integer Tokenization and Sequence Padding. For models such as Multilayer Perceptron (MLP) and Convolutional Neural Network (CNN), input reviews were first converted into sequences of integer indices based on a vocabulary map. These sequences were subsequently padded or truncated to a fixed length (100 tokens), ensuring input dimensional consistency across the batch. This representation preserves word order while enabling fast vectorized computation [25]
- Subword Tokenization and Attention Masking. For IndoBERT and XLM-RoBERTa, input sequences were processed using subword tokenization WordPiece and SentencePiece, respectively allowing for granular handling of rare or compound words. The text was then transformed into token IDs, attention masks, and segment embeddings, which serve as inputs to the transformer architecture. The maximum sequence length was capped at 128 tokens to balance context retention and computational efficiency [26].

The models were selected to represent three major classes of text classifiers: classical machine learning, baseline neural networks, and pretrained transformer-based models. Each model was chosen based on its theoretical foundation, empirical performance in prior studies, and compatibility with the structured output of the data preparation pipeline.

Support Vector Machine (SVM) was selected as a representative of classical linear classifiers due to its robustness in high-dimensional spaces and proven

performance in text classification tasks. SVM maximizes the margin between support vectors of opposing classes in a transformed feature space. With the use of TF-IDF features, SVM has demonstrated excellent performance for sparse input vectors [27] making it an ideal candidate for baseline comparisons in this study.

Extreme Gradient Boosting (XGBoost), a tree-based ensemble method, was included due to its superior performance in many structured and semi-structured data competitions. XGBoost enhances weak learners iteratively using a gradient descent optimization framework, regularization, and parallel processing, which collectively improve its generalization capability and computational efficiency [28]. In prior sentiment analysis research on short texts, XGBoost consistently outperformed shallow models, especially when feature sparsity and non-linearity are present.

Multilayer Perceptron (MLP) serves as the baseline for neural network models in this study. MLP is a fully connected feedforward network that learns non-linear decision boundaries via backpropagation and activation functions such as ReLU. Although MLP lacks spatial inductive bias, it offers a strong benchmark for understanding the performance lift contributed by deeper architectures.

Convolutional Neural Network (CNN) was incorporated to exploit the local spatial features within sequences of text tokens. Originally designed for image data, CNNs have shown competitive performance in sentence-level classification tasks through 1D convolutions and max-pooling operations [29]. CNN is especially effective in capturing n-gram patterns and is computationally more efficient than RNN-based architectures.

IndoBERT, a monolingual transformer model pretrained on a large Indonesian corpus, was adopted to capture rich contextual information at both word and sentence levels. IndoBERT uses the BERT architecture with WordPiece tokenization and is pretrained using Masked Language Modeling (MLM) objectives. Prior studies in Bahasa Indonesia sentiment classification have shown that IndoBERT outperforms both classical and standard deep learning models by a significant margin, particularly in domain-specific tasks [4].

XLM-RoBERTa, a multilingual transformer model based on RoBERTa architecture and trained on 100+ languages using the CommonCrawl dataset, was selected to test the transferability of multilingual contextual embeddings to the Indonesian healthcare domain. Unlike IndoBERT, XLM-RoBERTa uses SentencePiece tokenization and is optimized with a larger batch size and dynamic masking. It has demonstrated exceptional performance on cross-lingual benchmarks and low-resource language tasks [5] making it a relevant candidate for this study.

To maximize predictive performance across architectures, this study implemented tailored tuning strategies for each model family: hyperparameter tuning for classical machine learning and baseline neural network models, and fine-tuning for pretrained transformer-based models. Each model was tuned using parameter grids validated via stratified 5-fold cross-validation.

For classical machine learning models, Support Vector Machine (SVM) and XGBoost classifiers were tuned using grid search over selected hyperparameter spaces. SVM optimization involved the regularization strength parameter $C \in \{0.1, 1.0, 10.0\}$ and kernel types from {linear, radial basis function (RBF), sigmoid}, which influence margin flexibility and decision boundaries. For XGBoost, the tuning space included $n\_estimators \in \{100, 200, 300\}$ and $learning\_rate \in \{0.01, 0.1, 0.3\}$, enabling control over model complexity and convergence behavior. These settings are consistent with prior best practices in text classification tasks involving high-dimensional features.

For baseline neural networks models (MLP and CNN), the search space focused on architectural and training parameters. Experiments were conducted over $embedding\_dim \in \{50, 100, 150\}$, number of dense units or convolutional filters $units \in \{32, 64, 128\}$, and learning rates $\in \{1e{-}3, 5e{-}4, 1e{-}4\}$. All networks used the Adam optimizer, and early stopping was applied based on validation performance. The configuration aimed to balance expressive capacity and regularization for short-text sentiment input.

For pretrained transformer-based models (IndoBERT and XLM-RoBERTa), fine-tuning was conducted end-to-end on the sentiment-labeled corpus using the Hugging Face Transformers framework. The fine-tuning grid included learning rates $\in \{5e{-}5, 3e{-}5, 2e{-}5\}$, training epochs $\in \{3, 5\}$, and batch sizes $\in \{32, 128\}$ per device. During fine-tuning, all model weights were updated using gradient descent, and sequences were truncated to 128 tokens. Early stopping and validation-based model selection were also incorporated to prevent overfitting and ensure convergence. Table 3. summarizes the hyperparameter configurations and search spaces employed for each model category used in this study.

Table 3. Hyperparameter Configuration For Each Model

Model	Tuned Hyperparameters	Search Space
SVM	Regularization	{0.1, 1.0, 10.0}
	Kernel type	{linear, rbf, sigmoid}
XGBoost	Number of Trees (n_estimators)	{100, 200, 300}
	Learning Rate (learning_rate)	{0.01,0.1,0.3}
MLP / CNN	Hidden Units (units)	{32, 64, 128}
	Learning Rate	{1e−3, 5e−4, 1e−4}

Model	Tuned Hyperparameters	Search Space
IndoBERT / XLM-RoBERTa	Embedding Dimension (embedding_dim)	{50, 100, 150}
	Learning Rate	{5e−5, 3e−5, 2e−5}
	Batch Size	{32, 128}
	Epochs	{3, 5}

To ensure fair and reproducible evaluation across all models, the dataset was initially partitioned into two subsets: 70% for training and 30% for testing. The training set was used for both hyperparameter tuning and model fitting, while the testing set was withheld entirely for final evaluation to simulate real-world deployment performance. Within the training set, a stratified 5-fold cross-validation approach was applied. This technique preserved the class distribution across folds and enabled robust estimation of generalization performance, especially in the presence of imbalanced sentiment classes. The complete procedure is illustrated in Fig. 3 [30], which depicts the simulation of 5-fold training-validation cycles prior to final testing.

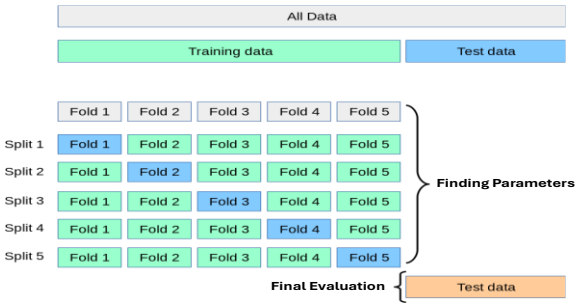

Figure 3. Simulation of 5-Fold Cross-Validation

The model training and validation procedures were executed on cloud-based infrastructure using Google Colab Pro. Classical machine learning models were trained using CPU resources, while baseline neural network models (MLP, CNN) leveraged the NVIDIA Tesla T4 GPU. For pretrained transformer-based models, such as IndoBERT and XLM-RoBERTa, a high-performance NVIDIA A100 GPU was utilized via runtime upgrade. The full computational environment specifications are summarized in Table 4.

Table 4. Computational Environment Specifications

Component	Specification
Platform	Google Colab Pro
CPU	Intel(R) Xeon(R) CPU @ 2.30GHz, 8 vCPU (4 cores, 2 threads/core)
CPU Architecture	x86_64, 64-bit
RAM	51 GB
Disk	235.7 GB Virtual Storage
GPU (for MLP, CNN)	NVIDIA Tesla T4 (CUDA), 15 GB VRAM
GPU (for Transformers)	NVIDIA A100 (CUDA), 40 GB VRAM (via runtime upgrade)
Virtualization	KVM (Kernel-based Virtual Machine)
L1/L2/L3 Cache	L1: 256 KB, L2: 1 MB, L3: 45 MB
Operating System	Ubuntu Linux (Kernel Virtualization)
Python Version	Python 3.11

EVALUATION

The evaluation phase of this study was structured to comprehensively assess model performance across three main dimensions: predictive performance metrics, statistical significance, and computational efficiency. This multi-faceted approach ensured not only that the models were accurate but also reliable and scalable for real-world application in Indonesian sentiment analysis.

1) Performance Metrics

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, which is a common challenge in sentiment classification tasks. It balances false positives and false negatives, offering a more informative view than accuracy alone when evaluating models with uneven class distributions.

Let TP, FP, FN, andTNrepresent true positives, false positives, false negatives, and true negatives, respectively. For each sentiment class c, the metrics are defined as follows:

- Precision (positive predictive value):
$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \quad (2)$$
- Recall (sensitivity):
$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \quad (3)$$
- F1-score (harmonic mean of precision and recall):
$$F1_c = 2 \cdot \frac{\text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \quad (4)$$
- Accuracy:
$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

The macro-average of each metric was calculated to ensure equal weighting across classes, which is particularly important for imbalanced datasets. In addition, to account for differences in class distribution and to more accurately reflect overall model performance in the presence of majority and minority classes, the weighted average of each metric was also computed. This allows a more reliable comparison between models by incorporating class frequency into the aggregation process [1].

2) Statistical Significance Testing

To assess whether observed performance differences among models were statistically significant, the Friedman test was employed. This non-

parametric test is suitable for comparing multiple classifiers over several datasets or folds [2]. Let R_ij denote the rank of the j-th model on the i-th fold. The Friedman statistic is computed as:

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_{c=1}^{k} \overline{R_j^2} - \frac{k(k+1)^2}{4}\right] \quad (6)$$

Where N = number of folds, k = number of models, and $(R_j)^{-}$ = average rank of model j. Following the Friedman test, the Nemenyi post-hoc test was conducted to identify pairwise differences among models. The critical difference (CD) is calculated as:

$$CD = q_\alpha \cdot \sqrt{\frac{k(k+1)}{6N}} \quad (7)$$

Where q_α is the critical value from the Studentized range distribution. This statistical evaluation framework enabled robust, fair, and replicable comparison across all modeling strategies used in this study.

3. RESULT AND DISCUSSION

This section presents the experimental results and provides a comprehensive analysis of model performance across various sentiment classification tasks. The discussion is structured to reflect key stages of the evaluation pipeline, beginning with an overview of the annotated data distribution, followed by detailed performance comparisons across six models representing three major learning paradigms: classical machine learning, baseline neural networks, and pretrained transformer-based architectures. Quantitative results are further examined using statistical significance testing and computational efficiency analysis, culminating in a discussion of the practical implications for real-world public health applications such as SATUSEHAT.

The labeled dataset used in this study comprises 34,178 user reviews of the SATUSEHAT Mobile application. Reviews were annotated into three sentiment categories: negative, neutral, and positive, following a standardized annotation protocol as described in the methodology. The resulting distribution reveals a substantial imbalance in sentiment classes. As visualized in Fig. 4, negative reviews dominate the dataset, accounting for 74.02% of all entries. Positive reviews comprise 21.58%, while only 4.41% are labeled as neutral.

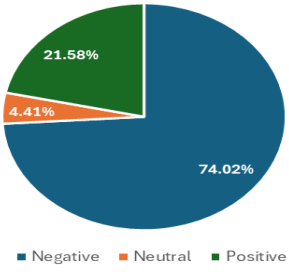

Figure 4. Sentiment label distribution of SATUSEHAT Mobile app reviews.

This distribution reflects the tendency of users to express dissatisfaction more vocally in public health platforms, particularly when encountering technical issues or unmet service expectations. The relatively small proportion of neutral reviews may suggest a lower likelihood of users submitting feedback unless they have a strongly polarized experience. The presence of such imbalance necessitates careful treatment in model development and evaluation, as it can bias predictions toward the majority class. Accordingly, macro-averaged metrics were employed throughout this study to ensure performance assessments remain class-balanced and reflective of true model generalizability across all sentiment types.

All models were trained and evaluated using a consistent experimental setup as previously outlined in the methodology. A 70:30 train–test split was applied to the labeled dataset to separate training and final evaluation stages. The training portion, comprising 70% of the labeled data, was then subjected to stratified 5-fold cross-validation to ensure class proportion consistency during model tuning and validation. The class distribution in the training and test sets closely mirrored the original dataset, as illustrated in Figure. 5.

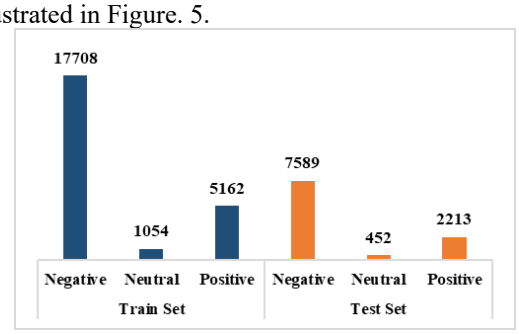

Figure 5. Label distribution across training and test sets after stratified 70:30 split.

Each model was evaluated under optimized hyperparameter configurations derived from either grid or randomized search, depending on the model category. To ensure fair and imbalanced-aware evaluation across sentiment classes, performance metrics including F1-score, accuracy, precision, and recall were computed using the weighted average method, which accounts for the proportion of instances in each class. The following tables present the selected best hyperparameters for each model (TABLE V. and summarize the corresponding cross-validation results (Table 5.), averaged over five folds.

Table 5. Summary Of Tuned Hyperparameters And Their Optimal Values

Model	Hyperparameters	Optimal Value
SVM	Regularization	10.0
	Kernel type	rbf
XGBoost	Number of Trees (n_estimators)	200

Model	Hyperparameters	Optimal Value
MLP	Learning Rate (learning_rate)	0.3
	Hidden Units (units)	32
	Learning Rate	0.001
	Embedding Dimension (embedding_dim)	100
CNN	Hidden Units (units)	32
	Learning Rate	0.001
	Embedding Dimension (embedding_dim)	100
IndoBERT	Learning Rate	2e-5
	Batch Size	32
	Epochs	3
XLM-RoBERTa	Learning Rate	2e−5
	Batch Size	128
	Epochs	5

Table 6. Average Performance Metrics From 5-Fold Cross-Validation

Model	F1-Score	Accuracy	Precision	Recall
SVM	0.8282 ± 0.0146	0.8166 ± 0.0097	0.8507 ± 0.0098	0.8507 ± 0.0098
XGBoost	0.8028 ± 0.0329	0.8096 ± 0.0160	0.8358 ± 0.0220	0.8358 ± 0.0220
MLP	0.8473 ± 0.0090	0.8355 ± 0.0152	0.8690 ± 0.0071	0.8690 ± 0.0071
CNN	0.8477 ± 0.0047	0.8295 ± 0.0050	0.8693 ± 0.0048	0.8693 ± 0.0048
IndoBERT	**0.8509 ± 0.0047**	**0.8396 ± 0.0043**	**0.8665 ± 0.0075**	**0.8665 ± 0.0075**
XLM-RoBERTa	**0.8564 ± 0.0029**	**0.8384 ± 0.0029**	**0.8785 ± 0.0029**	**0.8785 ± 0.0029**

Among the classical machine learning models, SVM demonstrated better performance than XGBoost across all metrics, particularly achieving an F1-score of 0.8282 ± 0.0146. For baseline neural network models, both MLP and CNN showed competitive results, with CNN slightly outperforming MLP in terms of recall and F1-score (0.8473 ± 0.0090). These results indicate that deeper architectures can generalize better on informal textual inputs, especially when combined with embedding-based representations.

However, the transformer-based models clearly surpassed the others in overall performance. XLM-RoBERTa attained the highest F1-score of 0.8564 ± 0.0029, while IndoBERT followed closely with 0.8509 ± 0.0047, reinforcing the strength of pretrained transformers in handling sentiment polarity classification for Indonesian user reviews. These models also consistently achieved top scores in precision, recall, and accuracy. Based on these cross-validation results, all models were then retrained using the best configuration on the full training set and evaluated on the held-out test set.

After the completion of hyperparameter tuning and cross-validation, each model was retrained on the full training set using the best-performing configuration. Final evaluation was conducted on the held-out 30% test set to assess the generalization capability of each model on unseen data. Table VII. and Fig. 6 presents the final performance metrics, including weighted F1-score, precision, recall, and accuracy for all six models across the three model families.

Table 7. Final Evaluation Results Of All Models On The Held-Out Test Set.

Model	F1 Score	Precision	Recall	Accuracy
SVM	0.8434	0.8281	0.8631	0.8631
XGBoost	0.8415	0.8294	0.8611	0.8611
MLP	0.8479	0.8362	0.8691	0.8691
CNN	0.8448	0.8257	0.8654	0.8654
IndoBERT	**0.8555**	**0.8400**	**0.8750**	**0.8750**
XLM-RoBERTa	**0.8552**	**0.8372**	**0.8772**	**0.8772**

The results consistently demonstrate that pretrained transformer-based models outperformed classical and baseline neural models across all evaluation metrics. IndoBERT achieved the highest F1-score (0.8555) and the best precision (0.8400), closely followed by XLM-RoBERTa with an F1-score of 0.8552. Among baseline neural networks, MLP slightly outperformed CNN (0.8479 vs. 0.8448), while in the classical model category, SVM slightly surpassed XGBoost, with both achieving competitive F1-scores above 0.84. These outcomes reaffirm the superior generalization ability of transformer-based models, as also reflected in the cross-validation results. However, class-wise performance reveals a persistent weakness in detecting neutral sentiments. All models including SVM, XGBoost, MLP, and CNN showed

near-zero F1-scores for the neutral class, likely due to the combination of semantic ambiguity and extreme label imbalance (neutral reviews comprise less than 5% of the dataset). In contrast, both negative and positive sentiments were more reliably captured, particularly by IndoBERT and XLM-RoBERTa, which consistently delivered strong performance across all classes. These results emphasize the importance of contextualized representations in handling nuanced sentiment expressions. Fig. 6 presents a detailed comparison of the F1-scores across sentiment classes for each model.

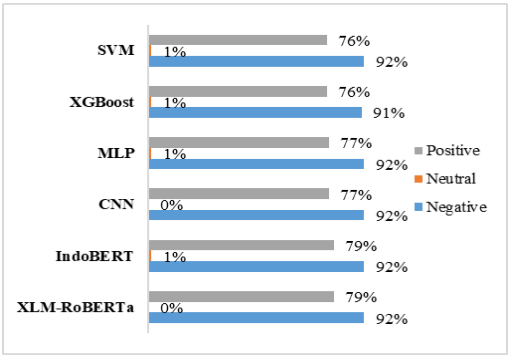

Figure 6. Class-wise F1-score comparison across all evaluated models.

To further illustrate this discrepancy, a confusion matrix of the best-performing model is provided in Fig. 7 and Fig. 8.

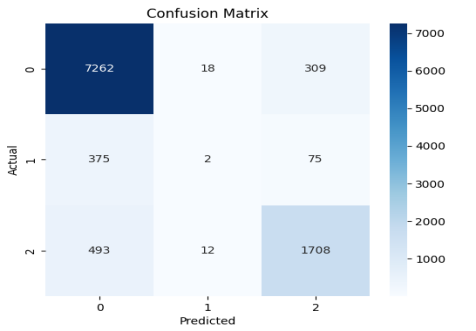

Figure 7. Confusion matrix of the IndoBERT model on the test set.

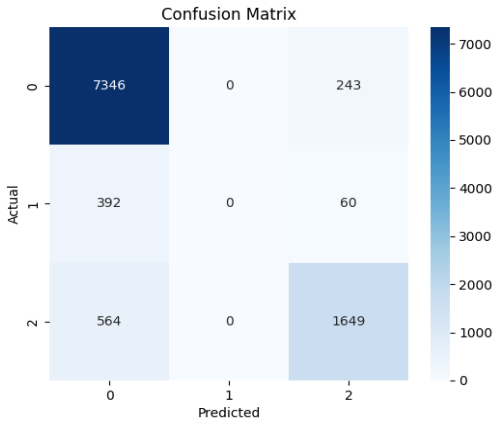

Figure 8. Confusion matrix of the XLM-RoBERTa model on the test set.

As illustrated in Fig. 7 and Fig. 8, both IndoBERT and XLM-RoBERTa exhibit strong classification performance, particularly in identifying negative (class 0) and positive (class 2) sentiments, as reflected by the dense diagonal patterns in their confusion matrices. However, both models consistently show reduced accuracy in detecting neutral sentiment (class 1), where the number of true positives is notably lower and misclassification into positive or negative categories is frequent.

This reduced performance for the neutral class can be attributed to several factors. First, neutral reviews often contain vague or context-dependent expressions that lack clear emotional cues, making them difficult to distinguish from mildly positive or negative comments. Second, class imbalance in the dataset where neutral samples are typically fewer or less diverse can lead to biased learning toward more dominant sentiment classes. Third, transformer-based models, despite their contextual understanding, tend to associate stronger attention weights with explicit sentiment indicators (e.g., “good,” “bad,” “problem,” “excellent”), which are less common in neutral statements. Consequently, the subtle linguistic patterns of neutrality are more prone to misclassification.

Nonetheless, the confusion matrices confirm that both transformer-based models maintain robust discriminative ability in classifying polarized sentiments, reinforcing their reliability in identifying strongly positive and negative user opinions. To statistically assess whether performance differences among the six evaluated models were significant, the Friedman test was employed as a non-parametric alternative to repeated-measures ANOVA. This test evaluates the null hypothesis H_0: there is no significant difference in performance rankings across models. The test was applied on F1-scores obtained from stratified 5-fold cross-validation. Table 8. summarizes the average F1-score ranks of all models across folds, which form the basis for the Friedman test analysis.

Table 8. Final Average Ranks Of Models

Model	Average Rank
XLM-RoBERTa	**1.0**
IndoBERT	**2.0**
MLP	3.2
CNN	4.2
XGBoost	5.2
SVM	5.4

Let N=5 denote the number of folds and k=6 the number of models. The Friedman statistic is computed as follows Equation (6). Substituting the values:

$$\chi_F^2 = \frac{12 \times 5}{6(6+1)}\left[89.08 - \frac{6 \times (7)^2}{4}\right] = 22.26$$

According to the chi-squared distribution, the critical value at α=0.05 for df=5 is approximately χ_0.05,5^2=11.07. Since 22.26 > 11.07, the null hypothesis H_0 (that all models perform equally) is rejected at the 95% confidence level. Furthermore, the corresponding p-value is 0.0005 which means p < 0.05, providing strong evidence that at least one model performs significantly differently from the others.

Following this result, a Nemenyi post-hoc test was applied to perform pairwise comparisons and identify which models differed significantly. The critical difference (CD) was computed using the Equation (1) :

$$CD = 2.728 \cdot \sqrt{\frac{6(6+1)}{6 \times 5}} \approx 3.2278$$

Where q_α is the critical value from the Studentized range distribution for k=6,α=0.05. These findings are visually presented in the critical difference diagram in Fig. 10, which illustrates the statistically significant differences between the models.

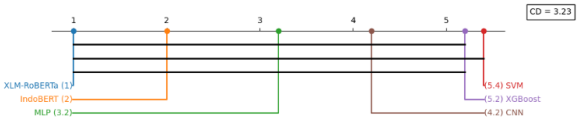

Figure 10. Critical Difference Diagram from the Nemenyi post-hoc test.

The average rank results from the Friedman test indicate that XLM-RoBERTa and IndoBERT outperform other models, with mean ranks of 1.0 and 2.0, respectively. The critical difference (CD) calculated from the Nemenyi post-hoc test is 3.23. Given that the rank difference between IndoBERT (2.0) and the next best model, MLP (3.2), is less than the CD, no statistically significant difference is observed between them. However, both transformer-based models (XLM-RoBERTa and IndoBERT) exhibit a statistically significant performance advantage over classical machine learning models (XGBoost and SVM), whose ranks (5.2 and 5.4) are well beyond the CD threshold. This result affirms the superior effectiveness of pretrained transformer models in sentiment polarity classification of Indonesian public health app reviews.

To complement the predictive performance analysis, this study also evaluates the computational efficiency of each model in terms of training time and prediction time per sample. These metrics are essential for assessing model practicality, particularly in real-world deployment scenarios with time or resource constraints. The training time reflects the total duration (in seconds) required to train each model on the full training set of 23,924 reviews, while the prediction time denotes the average time (in milliseconds) to infer a single sample from the test set (10,254 reviews). The summarized results are presented in Table 9.

Table 9. Train And Predicition Time Of Models.

Model	Train Time (s)	Prediction Time per Sample (ms)
SVM	132.14	1.002
XGBoost	9.24	0.010
MLP	37.47	0.144
CNN	26.80	0.118
IndoBERT	669.69	3.692
XLM-RoBERTa	1036.20	3.345

As shown in TABLE IX. XGBoost and SVM exhibit significantly lower training and prediction times, highlighting their computational efficiency for large-scale applications. Baseline neural network models, including MLP and CNN, require moderate time for training and inference, thus striking a balance between performance and computational cost. However, pretrained transformer-based models, namely IndoBERT and XLM-RoBERTa, incur the highest computational overhead both in training time, exceeding 600 seconds, and in prediction time, exceeding 3 milliseconds per sample. While these pre-trained transformer-based models achieve top-tier performance in sentiment classification, the substantial time costs associated with fine-tuning and inference require consideration in latency-sensitive or resource-constrained environments.

4. CONCLUSION

This study investigated the effectiveness of various machine learning approaches for sentiment polarity classification on Indonesian-language reviews of the SATUSEHAT Mobile application. Three distinct model families were evaluated: classical machine learning (SVM and XGBoost), baseline neural networks (MLP and CNN), and pretrained transformer-based models (IndoBERT and XLM-RoBERTa). Utilizing a carefully annotated dataset consisting of 34,178 reviews labeled into positive, neutral, and negative sentiments, the experimental results revealed that pretrained transformer-based models consistently outperform classical machine learning and baseline neural network models across all evaluation metrics. IndoBERT achieved the highest F1-score of 0.8555, closely followed by XLM-RoBERTa at 0.8552, demonstrating their robust capability to capture semantic nuances inherent in informal Indonesian text.

Addressing the first research question, "How do pretrained transformer-based models such as IndoBERT and XLM-RoBERTa compared to classical machine learning and baseline neural network approaches in classifying sentiment polarity in Indonesian public health app reviews?", this study employed statistical validation through Friedman and Nemenyi tests. The Friedman test indicated significant differences in performance across models (χ^2 = 22.2571, p < 0.05), with XLM-RoBERTa and IndoBERT significantly surpassing classical machine learning and baseline neural networks. Specifically, XLM-RoBERTa secured an average rank of 1.0,

illustrating a clear statistical superiority, while IndoBERT maintained strong performance at an average rank of 2.0, significantly outperforming SVM and XGBoost. Despite their computational overhead, these transformer-based models provided a decisive advantage in accurately classifying sentiment polarity, especially for nuanced and informal textual data.

Regarding the second research question, "How can sentiment analysis of user-generated reviews be leveraged to improve the quality and responsiveness of digital government health services?", this study highlights substantial practical implications. The findings suggest that integrating transformer-based sentiment classifiers such as IndoBERT and XLM-RoBERTa into real-time analytic dashboards of digital health platforms, like SATUSEHAT, can significantly enhance responsiveness and service quality. Such integration facilitates prompt detection of negative sentiments, allowing governmental health authorities to swiftly identify and rectify service issues, improving citizen engagement and fostering a more agile public health infrastructure.

Nevertheless, several limitations emerged from this research. The substantial imbalance in sentiment label distribution, predominantly skewed towards negative reviews (74.02%), significantly constrained models' ability to effectively classify neutral sentiments. Additionally, the dataset derived exclusively from Google Play Store reviews for one specific application potentially limits generalizability across diverse digital health contexts. These factors underscore the necessity for cautious interpretation and application of these findings in broader public health analytics contexts.

Future research should address these limitations by expanding datasets to encompass additional sources such as social media, health forums, and diverse app stores to enhance domain robustness and linguistic variability. Employing advanced data augmentation and rebalancing methods, such as SMOTE, backtranslation, or paraphrasing, could alleviate issues arising from imbalanced sentiment classes. Additionally, exploring larger language models (LLMs), including GPT-style architectures fine-tuned on Indonesian healthcare corpora, is highly recommended due to their superior contextual comprehension and scalability. Implementing these advanced models within real operational environments and validating them through collaborative evaluations with government stakeholders will further solidify their practical utility and help optimize public health digital services.

ACKNOWLEDGMENT

The authors would like to express their gratitude to the Ministry of Health of the Republic of Indonesia for providing the SATUSEHAT Mobile platform, which served as the basis for this study. Special thanks are extended to the annotators who contributed their time and expertise during the manual labeling process. This research was supported by the Faculty of Computer Science, Universitas Indonesia.*:*

5. REFERENCES

[1] E. Boiy and M. F. Moens, “A machine learning approach to sentiment analysis in multilingual web texts,” Inf Retr Boston, vol. 12, no. 5, pp. 526–558, Sep. 2009, doi: 10.1007/S10791-008-9070-Z/TABLES/14.

[2] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol. 13-17-August-2016, pp. 785–794, Mar. 2016, doi: 10.1145/2939672.2939785.

[3] Y. Zhang and B. C. Wallace, “A Sensitivity Analysis of (and Practitioners’ Guide to) Convolutional Neural Networks for Sentence Classification,” Oct. 2015, Accessed: Jun. 05, 2025. [Online]. Available: https://arxiv.org/pdf/1510.03820

[4] B. Wilie et al., “IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding,” Sep. 2020, Accessed: Jun. 05, 2025. [Online]. Available: https://arxiv.org/pdf/2009.05387

[5] A. Conneau et al., “Unsupervised Cross-lingual Representation Learning at Scale,” Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp. 8440–8451, 2020, doi: 10.18653/V1/2020.ACL-MAIN.747.

[6] H. Imaduddin, F. Y. A’la, and Y. S. Nugroho, “Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach,” International Journal of Advanced Computer Science and Applications, vol. 14, no. 8, pp. 113–117, 2023, doi: 10.14569/IJACSA.2023.0140813.

[7] N. Paramita and S. Noviarisanti, “Service Quality Analysis of Mhealth Services Using Text Mining Method : Alodokter and Halodoc,” International Journal of Management, Finance and Accounting, vol. 2, no. 2, pp. 1–21, Aug. 2021, doi: 10.33093/IJOMFA.2021.2.2.1.

[8] K. A. Safitri, D. Vita, W. Swasto, and A. Nurfikri, “Sentiment Analysis Telemedicine Apps Reviews Using NVIVO,” Proceedings 2022, Vol. 83, Page 4, vol. 83, no. 1, p. 4, Dec. 2022, doi: 10.3390/PROCEEDINGS2022083004.

[9] F. A. Alijoyo, S. Suhaerudin, and S. Meilia, “MEASURING THE USER EXPERIENCE OF THE SATUSEHAT APPLICATION WITH THE HEART METRICS METHOD APPROACH,” SIBATIK JOURNAL: Jurnal Ilmiah Bidang Sosial, Ekonomi, Budaya, Teknologi, Dan Pendidikan, vol. 3, no. 4, pp. 515–534, Mar. 2024, doi: 10.54443/SIBATIK.V3I4.1854.

[10] M. Clinton, T. Manullang, A. Z. Rakhman, H. Tantriawan, and A. Setiawan, “Comparative Analysis of CNN, Transformers, and Traditional ML for Classifying Online Gambling Spam

Comments in Indonesian," Journal of Applied Informatics and Computing, vol. 9, no. 3, pp. 592–602, Jun. 2025, doi: 10.30871/JAIC.V9I3.9468.

[11] 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 : jurnal dan penelitian teknik informatika, vol. 9, no. 2, pp. 872–883, May 2025, doi: 10.33395/SINKRON.V9I2.14723.

[12] H. M. Lee and Y. Sibaroni, "Comparison of IndoBERTweet and Support Vector Machine on Sentiment Analysis of Racing Circuit Construction in Indonesia," JURNAL MEDIA INFORMATIKA BUDIDARMA, vol. 7, no. 1, p. 99, Jan. 2023, doi: 10.30865/MIB.V7I1.5380.

[13] H. Jayadianti, W. Kaswidjanti, A. T. Utomo, S. Saifullah, F. A. Dwiyanto, and R. Drezewski, "Sentiment analysis of Indonesian reviews using fine-tuning IndoBERT and R-CNN," ILKOM Jurnal Ilmiah, vol. 14, no. 3, pp. 348–354, Dec. 2022, doi: 10.33096/ILKOM.V14I3.1505.348-354.

[14] S. Aras, M. Yusuf, R. Ruimassa, E. Agustinus, B. Wambrauw, and E. B. Palalangan, "Sentiment Analysis on Shopee Product Reviews Using IndoBERT," Journal of Information Systems and Informatics, vol. 6, no. 3, pp. 1616–1627, Sep. 2024, doi: 10.51519/JOURNALISI.V6I3.814.

[15] W. Wongso, D. Samuel Setiawan, S. Limcorn, A. Joyoadikusumo, and S. Wales, "NusaBERT: Teaching IndoBERT to be Multilingual and Multicultural," 2025. Accessed: Jun. 05, 2025. [Online]. Available: https://aclanthology.org/2025.sealp-1.2/

[16] C. H. Lin and U. Nuha, "Sentiment analysis of Indonesian datasets based on a hybrid deep-learning strategy," J Big Data, vol. 10, no. 1, pp. 1–19, Dec. 2023, doi: 10.1186/S40537-023-00782-9/TABLES/5.

[17] I. G. B. A. Budaya and I. K. P. Suniantara, "Comparison of Sentiment Analysis Algorithms with SMOTE Oversampling and TF-IDF Implementation on Google Reviews for Public Health Centers," MALCOM: Indonesian Journal of Machine Learning and Computer Science, vol. 4, no. 3, pp. 1077–1086, Jul. 2024, doi: 10.57152/MALCOM.V4I3.1459.

[18] S. Gohil, S. Vuik, and A. Darzi, "Sentiment Analysis of Health Care Tweets: Review of the Methods Used," JMIR Public Health Surveill, vol. 4, no. 2, Jan. 2018, doi: 10.2196/PUBLICHEALTH.5789.

[19] A. B. Nair, A. K., A. U., D. T. Jaison, A. V., and V. S. Anoop, "'Hey..! This medicine made me sick': Sentiment Analysis of User-Generated Drug Reviews using Machine Learning Techniques," Apr. 2024, Accessed: Jun. 06, 2025. [Online]. Available: https://arxiv.org/pdf/2404.13057

[20] A. W. Pradana and M. Hayaty, "The Effect of Stemming and Removal of Stopwords on the Accuracy of Sentiment Analysis on Indonesian-language Texts," Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, vol. 4, no. 4, pp. 375–380, Oct. 2019, doi: 10.22219/KINETIK.V4I4.912.

[21] M. H. R. Sofyan, A. Zulkifli, and R. Rasim, "Sentiment Analysis of Indonesian Presidential Candidate Before and After the Election," UltimaInfoSys: Jurnal Ilmu Sistem Informasi, vol. 15, no. 2, pp. 99–104, Dec. 2024, doi: 10.31937/SI.V15I2.3689.

[22] T. Rahman, F. E. M. Agustin, and N. F. Rozy, "Normalization of Unstructured Indonesian Tweet Text For Presidential Candidates Sentiment Analysis," 2019 7th International Conference on Cyber and IT Service Management, CITSM 2019, Nov. 2019, doi: 10.1109/CITSM47753.2019.8965324.

[23] S. K. Dirjen, P. Riset, D. Pengembangan, R. Dikti, S. Khomsah, and A. S. Aribowo, "Text-Preprocessing Model Youtube Comments in Indonesian," Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 4, no. 4, pp. 648–654, Aug. 2020, doi: 10.29207/RESTI.V4I4.2035.

[24] J. E. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries," 2003.

[25] X. Zhang, J. Zhao, and Y. Lecun, "Character-level Convolutional Networks for Text Classification," Adv Neural Inf Process Syst, vol. 2015-January, pp. 649–657, Sep. 2015, Accessed: Jun. 06, 2025. [Online]. Available: https://arxiv.org/pdf/1509.01626

[26] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, vol. 1, pp. 4171–4186, Oct. 2018, Accessed: Jun. 06, 2025. [Online]. Available: https://arxiv.org/pdf/1810.04805

[27] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features," pp. 137–142, 1998, doi: 10.1007/BFB0026683.

[28] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol. 13-17-August-2016, pp. 785–794, Mar. 2016, doi: 10.1145/2939672.2939785.

[29] Y. Kim, "Convolutional Neural Networks for Sentence Classification," EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, pp. 1746–1751, Aug. 2014, doi:

10.3115/v1/d14-1181.
[30] L. Jian, Z. Huang, J. Zhang, and Z. Hu, "Rapid Analysis of Cylindrical Bypass Flow Field Based on Deep Learning Model," IOP Conf Ser Earth Environ Sci, vol. 1037, no. 1, p. 012013, Jun. 2022, doi: 10.1088/1755-1315/1037/1/012013..