

Text Classification of Public Complaint Validity With Deep Learning Approaches

Ignatius Wisnu Prayogo¹, Yulia Kurniawati¹, Isnina Eva Hidayati¹, Amelia Khairunnisa¹, Yova Ruldeviyani¹

¹ Information Technology Study Program, Faculty of Computer Science, Universitas Indonesia, Depok, Jawa Barat 16424, Indonesia

[Submitted: 21 August 2025, Revised: 22 January 2026, Accepted: 16 March 2026]

Corresponding Author: Ignatius Wisnu Prayogo (ignatius.wisnu41@ui.ac.id)

ABSTRACT — Mobile apps Jakarta Kini (JAKI) recorded 173,327 public complaints in 2023, accounted for 91.37% of all public complaints to the DKI Jakarta Provincial Government. Reports submitted to the Cepat Respon Masyarakat system must be handled in accordance with the service level agreement (SLA) regarding handling time. Currently, the process of validating incoming reports is still done manually by officers, taking more than 30 minutes per report. In the same year, 15,634 complaints were recorded as unclear or invalid. This led to a decrease in the performance of local government agency, impacting 13% of them did not achieve a 100% SLA in 2023. This study aimed to automate the validity classification of public complaints to distinguish between valid and invalid reports. The study utilized a dataset of 2,000 reports and employed deep learning models, including Indonesian version of bidirectional encoder representations from transformers (IndoBERT) and multilingual BERT (mBERT), and to compare their performance against traditional machine learning baselines, including term frequency-inverse document frequency (TF-IDF) + extreme gradient boosting (XGBoost), naïve Bayes, and support vector machine (SVM) using a 5-fold cross-validation scheme. The results showed that the IndoBERT model could classify valid or invalid reports with an average accuracy of 88.8%, which was higher than other models. The implementation of this method has proven to increase the efficiency of report validation time with computation time of 6 minutes for 300 reports, thus helping government agencies achieve their SLA targets and contributing to research on the effectiveness of BERT in public complaint classification.

KEYWORDS — Text Classification, Deep Learning, Fine-Tuning, IndoBERT, JakartaKini.

I. INTRODUCTION

The Special Capital Region of Jakarta (DKI Jakarta) is the most densely populated province in Indonesia, with a projected population of 10.68 million in 2024 [1]. As the main center of economic and governmental activities, the city faces a high volume of citizen interactions with public institutions. One important instrument for measuring the quality of public services is the public satisfaction index (*indeks kepuasan Masyarakat*, IKM), which reflects the extent to which government agencies can deliver services that meet community expectations. To improve this index, the Provincial Government of DKI Jakarta has developed a Citizen Relationship Management (CzRM) application managed by Jakarta Smart City (JSC). This application integrates 13 official complaint channels and serves as the primary platform for handling public reports [2]. In 2023 alone, the CzRM system received a total of 173,327 public complaints, with 91.37% submitted through the Jakarta Kini (JAKI) mobile application. To ensure effective service delivery, Jakarta Regional Secretary Decree No. 99 of 2022 established detailed regulations regarding business processes and the service level agreement (SLA) for handling complaints [3].

The SLA sets strict timelines for complaint resolution, and it is directly linked to the performance evaluations and incentives for civil servants. However, monitoring and evaluation data in 2023 revealed that 13% of local government agencies failed to meet the 100% SLA target. Furthermore, the IKM survey in the same year indicated that response times for handling complaints remained below a score of 80, which is under the expected standard [3]. One of the main challenges contributing to this shortfall lies in the verification process of

public complaints. Reports must be validated to determine whether they are valid or invalid before further follow-up can occur. At present, this process is still conducted manually by officers, requiring more than 30 minutes per report. In 2023, as many as 15,634 reports were classified as unclear or invalid, which not only increased the workload of government personnel but also reduced the efficiency of complaint resolution [4]. The high volume of reports combined with the manual validation process has created a significant bottleneck in achieving SLA targets and in maintaining public trust in government services.

The challenge of validation is not only administrative but also strategic. Accurate identification of invalid reports is essential to prevent misallocation of resources, avoid redundant fieldwork, and ensure that agencies prioritize legitimate issues. Prolonged reliance on manual validation risks perpetuating systemic inefficiencies and constraining the government's capacity to respond promptly in critical situations. For instance, delays in addressing public infrastructure problems may lead to far-reaching social consequences, while unresolved cases may negatively impact overall satisfaction with public services.

In response to these challenges, the integration of advanced computational methods has become increasingly relevant. Natural language processing (NLP), a subfield of artificial intelligence, provides tools for automating the analysis of textual data, enabling systems to classify, interpret, and prioritize large volumes of complaints more effectively. In recent years, the adoption of deep learning has further enhanced the capacity of NLP models to capture linguistic and contextual nuances. Within this framework, bidirectional encoder representations from transformers (BERT) and its variants have

emerged as state-of-the-art methods for text classification tasks.

The application of transformer-based models specifically for validating Indonesian public complaints remains underexplored. Therefore, this study aimed to automate the validity classification of public complaints to distinguish between actionable reports and invalid noise. The study explored the use of Indonesian version for BERT (IndoBERT) [5] and compared its performance with multilingual BERT (mBERT), a multilingual model trained on data from multiple languages [6] as well as with traditional text classification approaches using term frequency-inverse document frequency (TF-IDF) combined with extreme gradient boosting (XGBoost) [7], support vector machine (SVM) and naïve Bayes. The main contribution of this research is the empirical evaluation of these models to provide a robust pre-validation filter. This automated approach significantly reduces manual processing time, assisting government agencies in meeting SLA targets and improving overall service quality.

II. RELATED WORKS

Text processing within NLP and deep learning involves analyzing and manipulating textual data to extract meaningful information. Deep learning enhances this process by capturing complex semantic and contextual relationships, enabling tasks such as text classification, sentiment analysis, and machine learning [8]. As a research foundation, this study reviewed the latest developments in the application of NLP in public services. Several prior studies have applied NLP to analyze public complaint reports through text classification.

One study utilized 56,766 reports from the JakLapor channel in the Jakarta Kini (JAKI) application, focusing on five complaint categories: disturbance of peace and order, roads, illegal parking, trees, and garbage. The methodology involved text preprocessing, feature extraction using count vectorizer, TF-IDF, N-Gram, and latent semantic analysis, followed by machine learning classification with SVM, random forest (RF), XGBoost, and AdaBoost. The results indicated that SVM and RF achieved the highest accuracy of 90% with TF-IDF, while XGBoost with LSA recorded at 86%, and AdaBoost had the lowest at 81%. Although RF was identified as an optimal method for balancing accuracy and runtime, the study remained limited to five complaint categories and applied default parameters without hyperparameter tuning, leaving space for improvement through more advanced approaches such as deep learning [9].

Another study analyzed 6,321 complaint reports from residents of Tangerang City, aiming to classify text collected from social media into formal public complaints [10]. The researchers proposed two text preprocessing approaches: context-based text preprocessing (Con-TP), which considers semantic context, and corpus-based text processing (Cor-TP), which relies on lexical resources. Both approaches were followed by feature extraction and machine learning classification. The results showed that Con-TP achieved the highest accuracy of 90.85%, outperforming Cor-TP in handling noisy and unstructured text. Despite these promising results, the study faced challenges related to dataset imbalance, the limited number of complaint categories, and the reliance on expert-based manual labeling, which constrained scalability for larger datasets.

A separate study proposed an NLP-based model to automate the classification of textual complaints, with experiments conducted in the banking/financial sector and a local dataset related to education services. The datasets included 10,000 local complaints divided into 10 categories and over 120,000 complaints from the consumer complaint database. Preprocessing involved tokenization, stop word removal, punctuation cleaning, stemming, and lemmatization, followed by feature extraction using count vectorizer and TF-IDF. Several classifiers were tested, including SVM, RF, logistic regression, multinomial naïve Bayes, and k-nearest neighbor (KNN). On the local dataset, accuracies ranged from 72% to 80%. Meanwhile, on the consumer complaint dataset, the models achieved higher performance, with logistic regression reaching 86% accuracy using TF-IDF. The study concluded that while SVM and logistic regression provided the most stable results, challenges remained in addressing data imbalance and misclassified short-text complaints [11].

In Portugal, a study explored the automation of routing public service complaints to the appropriate departments using machine learning [12]. The dataset consisted of 4,459 summarized complaints across 17 categories, which presented challenges due to its small size and highly imbalanced distribution. To address this, the researchers applied preprocessing steps such as lowercasing, normalization, removal of numbers and special characters, and feature extraction using TF-IDF. They also experimented with data augmentation through translation into multiple languages to increase training samples for underrepresented classes. Various models were compared, including naïve Bayes, KNN, SVM, XGBoost, and mBERT. Results showed that while traditional models like SVM performed strongly in the first-stage of classification, with mBERT achieving an accuracy of 87% with an F-score of 88% when the augmented dataset was used, highlighting the importance of balancing data for improving deep learning performance.

Another study investigated hoax news detection in the Indonesian language by comparing IndoBERT, SVM, and naïve Bayes [13]. The dataset consisted of 2,000 news headlines, balanced between 1,000 hoax and 1,000 non-hoax articles, collected from trusted Indonesian news portals such as Detik, Liputan 6, Kompas, and hoax repositories like Turnbackhoax and Cek Fakta. Preprocessing included text cleaning and normalization, followed by model training and evaluation using a 70:30 train-test split. Results showed that IndoBERT achieved the highest accuracy of 90%, outperforming naïve Bayes (80.25%) and SVM (73.42%). The study highlighted IndoBERT's advantage in capturing contextual relationships between words, enabling it to detect complex hoax patterns more effectively than traditional machine learning models.

From the reviewed studies, it is evident that most research focuses on multiclass categorization rather than validity classification. It is crucial to distinguish that the complaint validation task is fundamentally different from conventional topic classification or sentiment analysis, as it deals with distinguishing actionable content from “noise” based on unique linguistic features. While previous works have successfully applied machine learning for categorization, the specific challenge of identifying invalid reports in the Indonesian context using transformer-based models has not been adequately addressed. This study fills this gap by introducing a validation framework that specifically leverages IndoBERT's

capability to capture the nuances of informal text and slang often found in citizen reports.

III. METHODOLOGY

This study adopted a quantitative approach to evaluate the validity of public complaint reports using the 2023 CRM JAKI dataset. The overall research workflow began with data collection, where raw reports were gathered to form the initial dataset. This set then underwent a crucial data preprocessing stage to clean and standardize the text, making it suitable for analysis. Following this, the prepared data were partitioned into three distinct subsets: a data train set for teaching the models, a data validation set for hyperparameter tuning, and a data test set for the final unbiased evaluation, following established procedures [14], [15]. With the data prepared, the model development phase proceeded along two parallel streams. The first stream involved fine-tuning the pretrained IndoBERT and mBERT models, while the second focused on model training for the traditional TF-IDF + XGBoost, SVM and naïve Bayes approach, both utilizing the training data. The outcome of this phase was a Trained classification model for each method. Finally, a comprehensive model evaluation was conducted by assessing the performance of each trained model against the unseen test data. The results from this evaluation form the basis of the result and discussion section, where the models' effectiveness is analyzed and compared.

A. DATA COLLECTION

The dataset used in this study consisted of public complaint reports submitted through the JakLapor feature of the JAKI application in 2023. The data were formally obtained from JSC for research purposes related to public service complaints. Prior to use, the raw reports underwent an initial examination to ensure relevance and consistency with the research objectives. This process included filtering out duplicate entries, incomplete records, and irrelevant submissions that did not meet the criteria for complaint validation. The remaining data were then cleaned and organized into a structured format suitable for subsequent preprocessing, feature extraction, and model training and evaluation.

In JakLapor, each report typically contains several components, including a short title or caption, a more detailed textual description, and, in many cases, an image attachment provided by the user to illustrate the issue being reported. Additional metadata such as complaint category, location information, submission date, and status of handling are also recorded to support the management and follow-up process [9]. For this study, the primary focus was placed on the textual description of complaints, as this element provides the most relevant linguistic information for natural language processing and classification.

B. DATA PREPROCESSING

Data preprocessing is a crucial initial step in preparing textual data for machine learning and NLP analysis. This stage involves cleaning, transforming, and organizing the raw reports to improve the quality and effectiveness of model training. Proper preprocessing reduces noise and inconsistencies, enabling the models to learn more efficiently and accurately [16]. In this study, several preprocessing steps were applied sequentially. The process began with noise removal, where irrelevant elements such as punctuation, numbers, special characters, emojis, URLs, and meaningless words were eliminated to enhance accuracy. Following this step, case

folding was performed to convert all text into lowercase letters, ensuring uniformity in data representation. Next, normalization was conducted to standardize informal or slang language into formal equivalents, which is particularly important as complaint texts often contain nonstandard terms. The text was then tokenized into individual tokens, either words or sub-words, to facilitate further analysis. After tokenization, stop word removal was applied to filter out frequently used words (e.g., “di,” “adalah,” “dan”) that carry little semantic value, allowing the focus to shift toward more informative terms. Finally, stemming was carried out to reduce words to their root or base forms by removing suffixes, thereby minimizing variations in word forms and improving the consistency of text representation during analysis [15].

C. MODEL TRAINING

In this study, fine-tuning was conducted on both the IndoBERT and mBERT models, alongside the application of a traditional machine learning approach using TF-IDF combined with XGBoost, which has demonstrated strong performance in text classification tasks [9]. Fine-tuning is a deep learning process in which a pretrained model is further trained on a specific task or dataset. The goal is to adapt the previously learned features to better suit the new task, thereby improving performance without the need to train the model from scratch [17]. In this research, fine-tuning is expected to optimize the classification task for determining whether a report is valid or invalid, thus enabling appropriate follow-up actions.

BERT is a deep learning model designed to understand the contextual meaning of words within a sentence by considering both the preceding and following words. This bidirectional approach allows BERT to capture linguistic context more effectively than previous models, which typically process text in a unidirectional manner [18]. BERT has been extended into a multilingual version known as Multilingual BERT (mBERT), which can handle multiple languages [19].

mBERT has been pretrained on 104 languages, including Indonesian. It has been optimized to generalize across languages, enabling it to perform effectively on cross-lingual tasks. Its multilingual capabilities, combined with its training on Indonesian texts, make mBERT a powerful tool for applications that require contextual understanding in linguistically diverse environments [15]. IndoBERT, as a specialized variant of BERT, pretrained on 220 million Indonesian words. IndoBERT proves to be a highly effective model that leverages its deep understanding of contextual nuances to consistently outperform both multilingual transformers and traditional machine learning approaches in accuracy, robustness, and computational efficiency for Indonesian-language text analysis [6].

TF-IDF method is a widely used term-weighting technique in text classification. It represents the importance of a word within a document relative to a collection of documents [20]. XGBoost is an ensemble model based on decision trees, developed from gradient boosting decision trees (GBDT), and is known for its strengths in regularization and computational efficiency [21]. The combination of TF-IDF and XGBoost has been shown to produce strong classification performance in detecting categories within textual data [20], [21].

This study expanded the baseline comparison by including naïve Bayes and SVM alongside XGBoost. Naïve Bayes and SVM are widely used as baseline algorithms for classification

tasks due to their distinct characteristics and performance metrics. Both algorithms have been evaluated in various domains. These traditional algorithms provide a clear benchmark to assess the performance of the proposed deep learning models [9], [13].

The implementation of all five models was carried out using Python, utilized specific pretrained model architectures available on the Hugging Face repository. IndoBERT employed the *indobenchmark/indobert-base-p2* version with 124M parameters and for the multilingual baseline used *bert-base-multilingual-cased*. The computational experiments were conducted on the Google Colab Pro platform. Each model was evaluated using k-fold cross-validation with $k = 5$. K-fold cross-validation is highly beneficial in NLP tasks. This approach is particularly useful in model evaluation and selection, as it ensures consistent model performance across different datasets and helps reduce the risk of overfitting [22]. K-fold cross-validation is employed to address the issues of bias and variance by partitioning the dataset into k subsets, which are then used iteratively as training and validation sets. This method ensures that each data instance is utilized for both training and validation, thereby providing a more stable and reliable performance evaluation that is not dependent on a particular data split [23].

D. MODEL EVALUATION

To assess model performance, several standard evaluation metrics were used. Specifically, this study employed recall, accuracy, precision, and F1 score. These metrics were calculated based on the comparison between true positives (TP), where the model correctly predicts the positive class; and true negatives (TN), where the model correctly predicts the negative class. Additionally, these metrics also accounted for the model's errors: false positives (FP), when the model incorrectly identifies a negative instance as positive; and false negatives (FN), when the model fails to detect a positive instance. These four components are commonly visualized using a confusion matrix, which is designed to highlight where the model makes correct predictions and where it fails [24]. The confusion matrix is a tabular visualization that compares predicted class labels to actual class labels for all data instances. It is widely used in the evaluation of deep learning models, particularly in classification tasks. The rows of the matrix represent the actual class labels, while the columns represent the predicted class labels [25].

The model's performance was evaluated using four commonly employed metrics: accuracy, precision, recall, and F1 score, as presented in (1)–(4). Accuracy measures how often the model correctly predicts the overall data. A value closer to 1 indicates better performance, while a value closer to 0 suggests poor predictive ability. Precision evaluates the proportion of correctly predicted positive cases out of all predicted positives. Recall reflects the proportion of actual positives that are correctly identified by the model. Finally, the F1 score represents the harmonic mean of precision and recall, balancing the trade-off between the two [26].

$$\text{Accuracy: } \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision: } \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall: } \frac{TP}{TP+FN} \quad (3)$$

$$\text{F1 Score: } 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

IV. RESULTS AND DISCUSSION

A. DATA COLLECTION

This study utilized secondary data of public complaints, originally submitted through the JakLapor reporting system and managed by JSC. The data were delivered to the researchers in tabular format and was publicly available. It included attributes such as report descriptions, report categories, and report statuses (completed, rejected, and received), with a total of 20,772 entries. The textual descriptions are partially redacted, particularly location details, to protect the personal data of the complainants. Sample entries from the dataset are shown in Table I.

In the classification process, the complaint reports were categorized into two labels: “*selesai*” representing valid reports and “*ditolak*” for invalid ones. A “*selesai*” label indicates that the issue has been addressed and resolved by the local government agencies, while a “*ditolak*” label indicates that the report was deemed invalid following manual validation by authorized officers. The final dataset was prepared through a two-stage process. First, reports labeled as “*diterima*” were excluded as they had not yet undergone validation. Second, the remaining data were processed to identify and remove any duplicate entries. This preparation pipeline yielded a final dataset consisting of 15,991 unique reports for the study.

A significant imbalance was identified within the dataset, as reports labeled as rejected constituted only approximately 1.38%. Such a severe imbalance can lead to a biased model that tends to ignore the minority class, resulting in poor predictive performance for those instances. To address this issue, an undersampling technique was employed. This method involved selectively reducing the number of majority class instances to create a more balanced class distribution of approximately 80% valid and 20% invalid reports [27]. This specific ratio was chosen as a compromise to maintain a degree of real-world data representation while ensuring the model could learn the minority class patterns effectively [28]. Finally, the balanced dataset used for training and evaluation in this study comprised 2,000 data.

B. DATA PREPROCESSING

The dataset from the JakLapor feature underwent several preprocessing stages to refine the complaint texts before model training. The process began with noise removal, which eliminated unnecessary elements such as punctuation, filler words like “*tuh*” and masked location identifiers. This step was followed by case folding, ensuring all words were in lowercase for consistency. In the normalization stage, variations such as “*di perbaiki*” were standardized into a single form “*diperbaiki*,” thereby reducing inconsistencies in word usage.

Next, the text was segmented through tokenization, which separated words into individual units. Stopword removal then filtered out common function words such as “*di*” that carried little semantic value. Finally, stemming reduced words to their base form, for instance “*diperbaiki*” became “*perbaiki*” and “*lalulintas*” became “*lintas*.” Through these sequential steps, the original lengthy and noisy report was transformed into a concise, structured text that preserved only the essential meaning. Table II presents transformation in detail, showing how each stage progressively simplified the input to make it more suitable for feature extraction and classification.

TABLE I
SAMPLE OF AVAILABLE DATASET

Report Description	Report Category	Report Label
<i>Permasalahan: batang atau ranting pohon mengenai kabel, berpotensi rubuh dan membahayakan kabel listrik. mohon di pangkas bagian yg terkena atau menjuntai di kabel listrik PLN.</i>	<i>Pohon</i>	<i>Selesai</i>
<i>Lokasi *** MASKED ***</i>		
<i>Permasalahan: Pohon Beringin sudah tinggi + ranting-ranting nya membahayakan kabel² di saat angin kencang mohon pihak berwenang agar tebang pohon beringin tsb demi keamanan & kenyamanan warga sekitar</i>	<i>Pohon</i>	<i>Selesai</i>
<i>Lokasi *** MASKED ***</i>		
<i>Permasalahan: Saluran air seperti pada gambar mengalami kerusakan, dimohon untuk tindak lanjutnya</i>	<i>Jalan</i>	<i>Selesai</i>
<i>Lokasi *** MASKED ***</i>		
<i>Permasalahan: Bersih ya pak, enak dipandang untuk kita yang ingin melihat alam</i>	<i>Taman</i>	<i>Ditolak</i>
<i>Lokasi *** MASKED ***</i>		

C. TRAINING MODELS

Following the preprocessing phase, the model training was initiated using the prepared training and validation datasets. The fine-tuning of the IndoBERT and mBERT models was performed with a specific hyperparameter configuration determined through manual tuning strategy. This strategy is primarily to accommodate computational resource limitations on Google Colab and mitigate overfitting on relatively small dataset [29]. Final configuration consisted of 3 epochs, a batch size of 16, and a weight decay of 0.01. The number of epochs was limited to three to reduce training time while still allowing the models to converge without significant overfitting. A relatively small batch size of 16 was selected to provide more stable updates and better adaptation to the dataset size. Meanwhile, the weight decay value of 0.01 was applied as a regularization factor to maintain balanced model performance. To further ensure consistency and generalization, a 5-fold cross-validation method was employed, allowing the evaluation to capture variations across different subsets of the data.

The performance outcomes of each model are summarized in Table III. IndoBERT consistently demonstrated better results across all evaluation metrics, showing both high accuracy and balanced precision and recall. Its stability across folds indicates that the model was able to generalize well to different subsets of the data, minimizing the risk of overfitting and confirming its robustness for real-world application. The ability of IndoBERT to capture subtle linguistic variations and contextual meaning in Indonesian complaint texts made it particularly effective in distinguishing between valid and invalid reports.

The mBERT model also showed competitive performance, with results relatively close to IndoBERT's. While slightly

TABLE II
TEXT PREPROCESSING STAGES

Stages	Text
Raw text	<i>Permasalahan: lampu merah mati tuh tolong di perbaiki lampu lalu lintas Lokasi *** MASKED ***</i>
Noise removal	<i>Permasalahan lampu merah mati tuh tolong di perbaiki lampu lalu lintas Lokasi MASKED</i>
Case folding	<i>permasalahan lampu merah mati tuh tolong di perbaiki lampu lalu lintas lokasi masked</i>
Normalization	<i>permasalahan lampu merah mati itu tolong di perbaiki lampu lalu lintas lokasi masked</i>
Tokenization	<i>permasalahan, lampu, merah, mati, itu, tolong, di, perbaiki, lampu, lalu, lintas, lokasi, masked</i>
Stopword removal	<i>lampu merah mati tolong perbaiki lampu lintas</i>
Stemming	<i>lampu merah mati tolong perbaiki lampu lintas</i>

TABLE III
K-FOLD CROSS VALIDATION RESULTS

Model	Fold	Accuracy	Precision	Recall	F1 Score
IndoBERT	1	0.907	0.908	0.907	0.9075
	2	0.880	0.873	0.880	0.8765
	3	0.887	0.882	0.887	0.8845
	4	0.887	0.884	0.887	0.8855
	5	0.870	0.869	0.870	0.8695
mBERT	1	0.887	0.892	0.887	0.8895
	2	0.883	0.883	0.883	0.8830
	3	0.890	0.890	0.890	0.8900
	4	0.883	0.878	0.883	0.8805
	5	0.867	0.863	0.867	0.8650
TF-IDF + XGBoost	1	0.873	0.867	0.873	0.8700
	2	0.883	0.879	0.883	0.8810
	3	0.860	0.850	0.860	0.8550
	4	0.867	0.859	0.867	0.8630
	5	0.863	0.857	0.863	0.8600
Naïve Bayes	1	0.853	0.866	0.853	0.8595
	2	0.846	0.871	0.846	0.8585
	3	0.860	0.880	0.860	0.8700
	4	0.840	0.854	0.840	0.8470
	5	0.843	0.869	0.843	0.8560
SVM	1	0.866	0.858	0.866	0.8620
	2	0.870	0.862	0.870	0.8660
	3	0.846	0.835	0.846	0.8405
	4	0.856	0.845	0.856	0.8505
	5	0.853	0.842	0.853	0.8475

lower overall, mBERT maintained stable scores across folds and confirmed the strength of multilingual transformer-based approaches in handling classification tasks on Indonesian texts. In contrast, the traditional machine learning approach TF-IDF + XGBoost, naïve Bayes and SVM model produced weaker performance compared to the BERT-based models and exhibited greater variability in evaluation metrics. Its results suggested that traditional feature-based methods struggled to capture semantic richness and contextual nuances, particularly when faced with unstructured or noisy complaint data. Taken together, the findings in Table III emphasize that BERT-based approaches significantly outperform conventional machine

learning models, with IndoBERT emerging as the most reliable and effective method for this classification task.

D. MODEL EVALUATION

To gain deeper insights into model performance, the confusion matrices for the three top-performing models (IndoBERT, mBERT, and TF-IDF + XGBoost) are presented in Figure 1, Figure 2, and Figure 3. Although SVM and naïve Bayes were included as baselines, this detailed error analysis focuses on the highest-accuracy models to highlight the most significant classification patterns. From these matrices, it can be concluded that the primary challenge across all evaluated approaches lies in handling the classification of invalid reports. While the top models generally performed well in recognizing valid reports, they consistently struggled when distinguishing invalid ones. A common pattern observed was the tendency to misclassify a considerable number of invalid reports as valid, suggesting that the features or linguistic cues present in invalid submissions are either too similar to valid complaints or too limited for the models, even the deep learning ones, to capture effectively. This issue reflects the inherent complexity of invalid reports, which often contain ambiguous, incomplete, or less structured expressions compared to valid ones.

Among the three, IndoBERT demonstrated the strongest sensitivity in recognizing valid reports, showing the fewest errors when distinguishing them from invalid ones. However, even with its superior performance, IndoBERT still exhibited noticeable difficulty in capturing the characteristics of invalid reports. Similar patterns were observed in mBERT and TF-IDF + XGBoost, though with slightly more frequent misclassifications, suggesting that the issue is not limited to model architecture but also relates to the nature of the dataset itself. Invalid reports often contain short, ambiguous, or context-free text, making them harder to distinguish, especially when the majority of the dataset consists of well-structured valid complaints.

These findings imply that while BERT-based models offer clear advantages, improving specificity particularly the correct identification of invalid reports remains the key challenge. For practical implementation in public complaint management systems, misclassifying invalid reports as valid can reduce the system's efficiency, as resources may be wasted on processing irrelevant submissions. This highlights the need for more balanced classification performance so that the system can better support accurate decision-making and ensure that attention is directed toward truly valid complaints.

These recurring misclassifications raise the possibility that the difficulty is not solely a limitation of the models but could also be influenced by the dataset itself. For instance, it needs to be examined whether the labeling of invalid complaints was carried out consistently, or whether some reports contain characteristics that overlap with valid submissions. In addition, invalid reports may vary widely in terms of style and detail, making them more challenging to represent during training. Such factors may contribute to the imbalance observed in classification performance and highlight the need for further investigation to better understand the root causes of this pattern.

Based on the overall evaluation, IndoBERT demonstrated the strongest performance compared to the other models. It achieved more stable accuracy across folds and maintained a better trade-off between accuracy and recall, resulting in the highest F1 score. This consistency indicates IndoBERT's

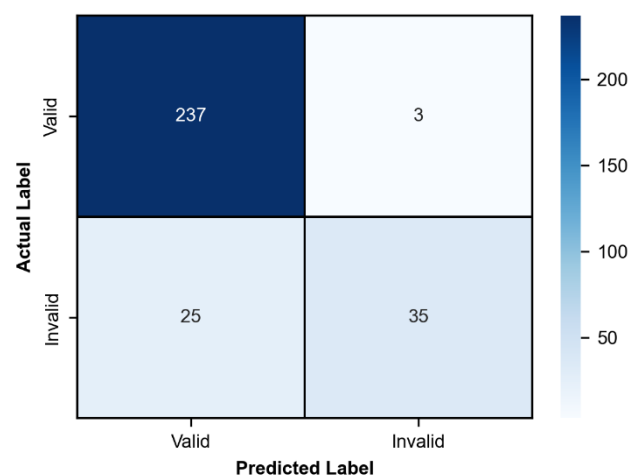


Figure 1. Confusion matrix of the IndoBERT with fold = 1.

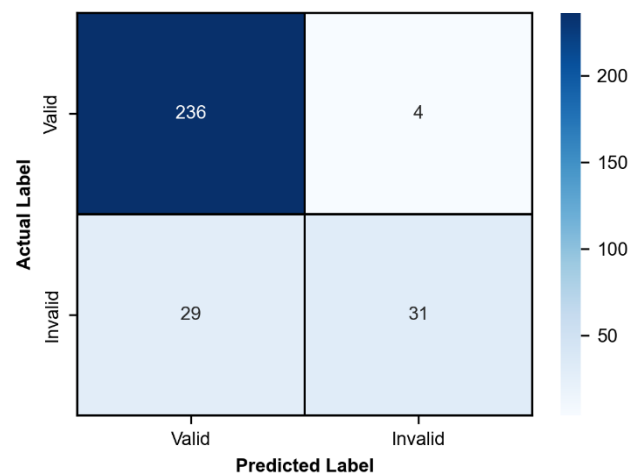


Figure 2. Confusion matrix of mBERT with fold = 3.

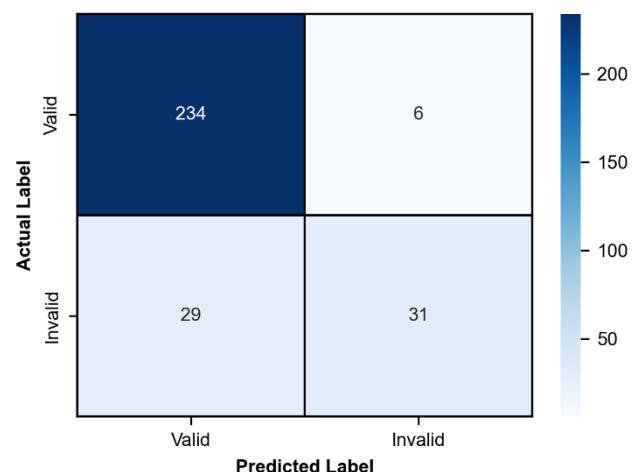


Figure 3. Confusion matrix of TF-IDF + XGBoost with fold = 2.

robustness in handling Indonesian public complaint texts, as shown in Figure 4.

When compared with previous studies, the findings demonstrate that IndoBERT provides a distinct advantage in semantic understanding. For instance, prior research applied SVM and RF with TF-IDF for complaint topic categorization, achieving around 90% accuracy [9]. However, traditional feature extraction often struggles with the unstructured, telegraphic nature of invalid reports evaluated in this study. The deep bidirectional architecture of IndoBERT overcomes this

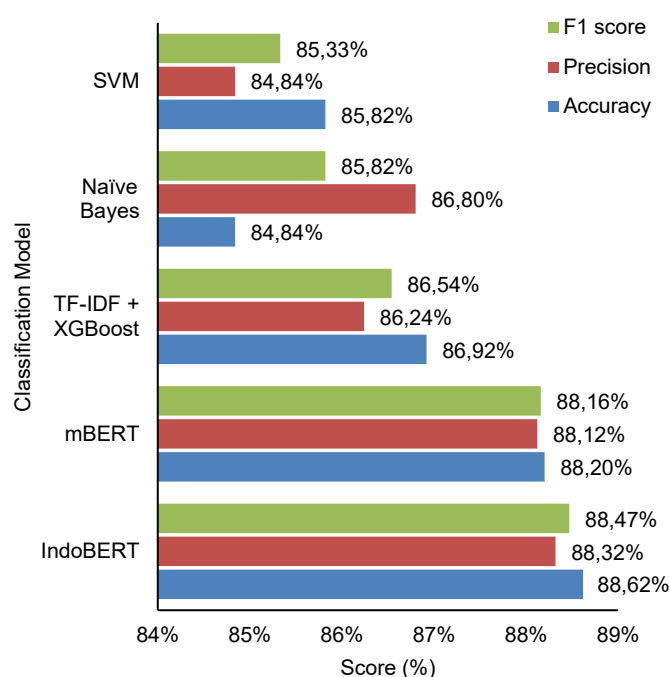


Figure 4. Model evaluation.

limitation by capturing contextual nuances more effectively than TF-IDF features. Furthermore, like the findings in hoax news detection, where IndoBERT outperformed SVM and naïve Bayes, the results of this study confirm that transformer-based models are significantly more robust when handling informal Indonesian text and slang [13]. The primary difference between the findings and previous studies lies in the task complexity: distinguishing operational noise from legitimate complaints requires high sensitivity to context-free ambiguity, a challenge where IndoBERT's pretrained linguistic representation proves highly effective, resulting in a superior balance of precision and recall compared to traditional baselines.

Further evaluation revealed IndoBERT's significant computational efficiency. Using the Google Colab Pro environment, the model classified 300 reports in approximately 6 minutes, averaging just 1.2 seconds per report. This performance starkly contrasts with the manual validation process, which requires over 30 minutes for a single report. To put this into perspective, validating 300 reports manually would demand over 150 hours of work, whereas the IndoBERT model completes the same task in a fraction of that time. This result confirms that IndoBERT not only delivers the most accurate classification for Indonesian public complaint texts but also offers exceptional efficiency.

This dramatic reduction in validation time suggests that implementing the IndoBERT model can lead to transformative practical benefits. By automating the initial validation stage, personnel in related agencies can be freed from a time-consuming administrative bottleneck, allowing them to focus their efforts on the more critical follow-up and resolution processes. This acceleration of the workflow is expected to directly contribute to achieving the SLA targets, an issue previously identified as a key challenge. Ultimately, ensuring faster and more accurate response times to public complaints is a crucial step toward improving the overall IKM.

Practical implication of deploying this model into the CzRM ecosystem is significant. By integrating the IndoBERT-

based classifier as a pre-validation filter in the manual validation by staff, the filter system can automatically flag high-probability invalid reports (e.g., spam or non-complaints). This allows human verifiers to prioritize their attention solely on valid complaints that require immediate action. Consequently, this prioritization strategy not only accelerates the resolution process but also reduces the cognitive load on government staff, ensuring consistent decision-making quality even during peak reporting periods.

E. ERROR ANALYSIS

Prediction errors in the IndoBERT model classification with two target classes (valid and invalid) can arise due to various factors. One major cause is textual ambiguity, such as the use of words with multiple meanings or unclear contexts, which makes it difficult for the model to distinguish whether a report should be classified as valid or invalid. Additionally, repeated reports submitted by the same user, which were manually marked as invalid by the officers, also contribute to classification errors.

Another common issue is the use of informal language structures in the reports, including regional dialects, slang, or typographical errors, which may hinder the model's ability to capture the intended context. In this study, the issue of slang was partially addressed by incorporating a slang-to-standard language conversion library during the preprocessing stage. An example of prediction error identification for report labels is shown in Table IV.

Another significant factor contributing to misclassification is the ambiguity caused by unstructured sentence styles often found in citizen reports. For instance, the report "*bangunan non rumah tinggal tanpa imb jalan*" (nonresidential building without permit road) was predicted as invalid despite being a valid complaint about a building violation. The lack of proper predicates and conjunctions obscures the semantic meaning, making it difficult for the model to distinguish between a legitimate report regarding a building without a permit (building construction permit) and a random string of keywords. This structural incoherence mimics the pattern of incoherent spam, leading the model to incorrectly reject the report.

To mitigate prediction errors caused by textual ambiguity or informal language structures, it is recommended to enrich the training dataset with more examples containing ambiguous terms or slang expressions. Furthermore, a more comprehensive lexicon or library for detecting slang, abbreviations, and colloquial expressions commonly used by Indonesian citizens is required to improve classification performance.

F. LIMITATIONS AND FUTURE WORK

The study is limited by several factors that warrant consideration. First, data privacy constraints restricted the use of personal information, allowing only general report text for analysis and potentially omitting valuable context. Second, the absence of standardized classification guidelines presents a significant challenge, as the manual labeling process is subject to officer discretion. This led to inconsistencies where repetitive reports or those expressing praise often marked as invalid by staff could have skewed the training data and influenced the model's learned patterns.

Furthermore, the dataset is temporally limited to a single year, which may not capture long-term trends. The undersampling technique used to balance the classes, resulted

TABLE IV
MISCLASSIFICATION OF INDOBERT

Report Text	Actual Label	Predicted Label
<i>bangunan non rumah tinggal tanpa imb jalan</i>	Valid	Invalid
<i>ini kompleks ruko jelas pasti ada parkir</i>	Valid	Invalid
<i>polisi asap mengganggu kenyamanan pernafasan</i>	Invalid	Valid
<i>fly over pesing bolongan tengah2 menghambat kendaraan arah indosiar</i>	Invalid	Valid

in discarding a large portion of majority class data, potentially losing some informational nuances. The current model might be less robust when encountering rare valid sentence structures that were not represented in the subsampled training set. Future research should explore alternative techniques such as synthetic minority over-sampling technique (SMOTE) or weighted loss functions to utilize the full dataset without discarding valuable information.

To build upon this research, future work is encouraged to develop a more holistic validation system. Consider that citizen complaint texts are often accompanied by supporting images. Future research is encouraged to incorporate image recognition features to further improve the system's classification accuracy. This advancement may lead to more robust and comprehensive complaint validation systems capable of better supporting public sector responsiveness.

V. CONCLUSION

This study demonstrates the significant potential of deep learning in enhancing both the efficiency and accuracy of citizen complaint validation, particularly through the implementation of the IndoBERT model. The proposed system automatically classified reports as valid or invalid, achieving a strong average accuracy of 88.8% and an F1 score of 88.5% in experimental evaluations using k-fold cross-validation. Operationally, the system validated 300 reports in just 6 minutes equivalent to approximately 1.2 seconds per report, a substantial reduction compared to the manual process that takes over 30 minutes per report. This improvement in processing time underscores its potential to help agencies meet SLA targets, improve service efficiency, and enhance public satisfaction in complaint follow-ups.

Despite these promising results, certain limitations remain that offer clear directions for future research. The model exhibited a bias toward the majority class, classification errors were observed in cases involving ambiguous or informal language and undersampling technique that removed majority of class data, which may limit its generalizability without further adaptation. As the study focused solely on text-based data, future work could integrate multimodal features, such as image recognition, to capture richer contextual information. Further research could also explore advanced data balancing strategies and compare IndoBERT with emerging transformer architectures to achieve even higher accuracy and generalization. Overall, this study provides strong empirical evidence for the effectiveness of transformer-based models for this task while identifying key areas for future advancement.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

AUTHORS' CONTRIBUTIONS

Conceptualization, Ignatius Wisnu Prayogo; methodology, Ignatius Wisnu Prayogo; software, Ignatius Wisnu Prayogo and Yulia Kurniawati; data acquisition, Ignatius Wisnu Prayogo and Isnina Eva Hidayati; data analysis, Yulia Kurniawati and Isnina Eva Hidayati; writing—original draft preparation, Ignatius Wisnu Prayogo; writing—reviewing and editing, Ignatius Wisnu Prayogo and Yova Ruldeviyani; visualization, Yulia Kurniawati and Amelia Khairunnisa; supervision, Yova Ruldeviyani.

REFERENCES

- [1] "Statistik Daerah Provinsi DKI Jakarta Tahun 2024," Badan Pusat Statistik Provinsi DKI Jakarta, 2024.
- [2] Bappeda Provinsi DKI Jakarta, "Rencana pembangunan daerah 2023-2026 Provinsi DKI Jakarta." Bappeda. Access date: 19-Aug-2025. [Online]. Available: <https://bappeda.jakarta.go.id/news/dokumen-rencana-pembangunan-daerah-tahun-2023-2026>
- [3] "Annual Report Jakarta Smart City 2023," Jakarta Smart City, 2024.
- [4] Jakarta Smart City, "Laporan Analisa CRM 2023," unpublished.
- [5] F. Koto, A. Rahimi, J.H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP," in *Proc. 28th Int. Conf. Comput. Linguist.*, 2020, pp. 757–770, doi: 10.18653/v1/2020.coling-main.66.
- [6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. N. Am. Chapter Assoc. Comput. Linguist.*, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
- [7] F. Lan, "Research on text similarity measurement hybrid algorithm with term semantic information and TF-IDF method," *Adv. Multimed.*, vol. 2022, no. 1, pp. 1–11, Apr. 2022, doi: 10.1155/2022/7923262.
- [8] Z. Xu, "Research on deep learning in natural language processing," *Adv. Comput. Commun.*, vol. 4, no. 3, pp. 196–200, Jul. 2023, doi: 10.26855/acc.2023.06.018.
- [9] S.M. Intani *et al.*, "Automating public complaint classification through JakLapor channel: A case study of Jakarta, Indonesia," in *2022 IEEE Int. Smart Cities Conf. (ISC2)*, 2022, pp. 1–6, doi: 10.1109/ISC255366.2022.9922346.
- [10] E.D. Madyatmadja, B.N. Yahya, and C. Wijaya, "Contextual text analytics framework for citizen report classification: A case study using the Indonesian language," *IEEE Access*, vol. 10, pp. 31432–31444, Mar. 2022, doi: 10.1109/ACCESS.2022.3158940.
- [11] Qurat-Ul-Ain, A. Shaikat, and U. Saif, "NLP based model for classification of complaints: Autonomous and intelligent system," in *2022 2nd Int. Conf. Digit. Futures Transform. Technol. (ICoDT2)*, 2022, pp. 1–6, doi: 10.1109/ICoDT255437.2022.9787456.
- [12] F. Caldeira, L. Nunes, and R. Ribeiro, "Classification of public administration complaints," in *Symp. Lang. Appl. Technol. (SLATE)*, 2022, pp. 1–12, doi: 10.4230/OASIS.SLATE.2022.9.
- [13] A. Rahmawati, A. Alamsyah, and A. Romadhony, "Hoax news detection analysis using IndoBERT deep learning methodology," in *2022 10th Int. Conf. Inf. Commun. Technol. (ICoICT)*, 2022, pp. 368–373, doi: 10.1109/ICoICT55009.2022.9914902.
- [14] F. Muftie and M. Haris, "IndoBERT based data augmentation for Indonesian text classification," in *2023 Int. Conf. Inf. Technol. Res. Innov. (ICITRI)*, 2023, pp. 128–132, doi: 10.1109/ICITRI59340.2023.10250061.
- [15] G.Z. Nabiilah, I.N. Alam, E.S. Purwanto, and M.F. Hidayat, "Indonesian multilabel classification using IndoBERT embedding and MBERT classification," *Int. J. Elect. Comput. Eng.*, vol. 14, no. 1, pp. 1071–1078, Feb. 2024, doi: 10.11591/ijece.v14i1.pp1071-1078.
- [16] C.P. Chai, "Comparison of text preprocessing methods," *Nat. Lang. Eng.*, vol. 29, no. 3, pp. 509–553, May 2023, doi: 10.1017/S1351324922000213.
- [17] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," *Proc. 56th Annu. Meet. Assoc. Comput. Linguist.*, 2018, pp. 328–339, doi: 10.18653/V1/P18-1031.
- [18] H. Niu *et al.*, "EHR-BERT: A BERT-based model for effective anomaly detection in electronic health records," *J. Biomed. Inform.*, vol. 150, pp. 1–11, Feb. 2024, doi: 10.1016/j.jbi.2024.104605.
- [19] L. Lin, "Multilingual text classification based on deep learning models," in *2023 IEEE 11th Jt. Int. Inf. Technol. Artif. Intell. Conf. (ITAIC)*, 2023, pp. 1202–1205, doi: 10.1109/ITAIC58329.2023.10409100.

- [20] Z. Qi, "The text classification of theft crime based on TF-IDF and XGBoost model," in *2020 IEEE Int. Conf. Artif. Intell. Comput. Appl. (ICAICA)*, 2020, pp. 1241–1246, doi: 10.1109/ICAICA50127.2020.9182555.
- [21] G. Ozogur, M.A. Erturk, Z.G. Aydin, and M.A. Aydin, "Android malware detection in bytecode level using TF-IDF and XGBoost," *Comput. J.*, vol. 66, no. 9, pp. 2317–2328, Sep. 2023, doi: 10.1093/comjnl/bxac198.
- [22] K.H. Liland, J. Skogholt, and U.G. Indahl, "A new formula for faster computation of the k-fold cross-validation and good regularisation parameter values in ridge regression," *IEEE Access*, vol. 12, pp. 17349–17368, Jan. 2024, doi: 10.1109/ACCESS.2024.3357097.
- [23] E. Sevinç, "An empowered AdaBoost algorithm implementation: A COVID-19 dataset study," *Comput. Ind. Eng.*, vol. 165, pp. 1–13, Mar. 2022, doi: 10.1016/j.cie.2021.107912.
- [24] N.A. Semary *et al.*, "Enhancing machine learning-based sentiment analysis through feature extraction techniques," *PLoS ONE*, vol. 19, no. 2, pp. 1–19, Feb. 2024, doi: 10.1371/journal.pone.0294968.
- [25] J. Görtler *et al.*, "Neo: Generalizing confusion matrix visualization to hierarchical and multi-output labels," in *CHI '22, Proc. 2022 CHI Conf. Hum. Factors Comput. Syst.*, 2022, pp. 1–13, doi: 10.1145/3491102.3501823.
- [26] S. Shila, F.T. Bayshakhy, and A. Sattar, "Detection and classification of road damage using deep learning approach with smartphone images," in *2023 14th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)*, 2023, pp. 1–7, doi: 10.1109/ICCCNT56998.2023.10307090.
- [27] F. Kamalov, A.F. Atiya, and D. Elreedy, "Partial resampling of imbalanced data," 2022, *arXiv:2207.04631*.
- [28] J.M. Johnson and T.M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, pp. 1–54, Mar. 2019, doi: 10.1186/s40537-019-0192-5.
- [29] M.A. Mutasodirin, R.E. Prasajo, A.F. Abka, and H. Rasyidi, "Simple hack for transformers against heavy long-text classification on a time- and memory-limited GPU service," in *2023 10th Int. Conf. Adv. Inform.: Concept Theory Appl. (ICAICTA)*, 2023, pp. 1–6, doi: 10.1109/ICAICTA59291.2023.10390269.