Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9879

Analysis and Prediction of MyTelkomsel Complaints Using IndoBERTweet and TimesFM-3 for Service Improvement Prioritization

Muhammad Syaikh Azka[1], Apriansyah Putra[2*], Rahmat Izwan Heroza [3], Ken Ditha Tania[4]
[1,2,3,4]Information System Department, Faculty of Computer Science, Universitas Sriwijaya, Indonesia
[1]azkamuhammadsyaikh@gmail.com, [2*]apriansyah@unsri.ac.id, [3]rahmatheroza@unsri.ac.id, [4]kenya.tania@email.com

open access

***Corresponding Author**

Article History:
Submitted: 22-10-2026
Accepted: 03-11-2026
Published: 07-11-2026

Keywords:
BERTopic; Complaint Forecasting; IndoBERTweet; MyTelkomsel; Sentiment Analysis; TimesFM-3.

Brilliance: Research of Artificial Intelligence is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

ABSTRACT

Application reviews contain useful feedback, but the growing volume of Indonesian-language reviews makes manual identification of complaints inefficient, and earlier MyTelkomsel studies stopped at sentiment classification. Objective: This study develops an integrated framework to identify the dominant technical complaints in MyTelkomsel reviews, forecast their trends, and prioritize service improvements. Methods: A total of 51,112 application-related Google Play reviews, published between November 2022 and September 2026, were kept after keyword-based relevance filtering. IndoBERTweet, fine-tuned on the SmSA dataset, performed sentiment analysis, and 44,937 negative reviews were grouped by BERTopic with K-Means into ten complaint categories. TimesFM-3 then produced zero-shot six-month complaint forecasting for each category, which was evaluated with a rolling-origin backtest and used to compute a priority score from forecast load and trend. Results: Login & Account Access, Loading/Performance Issues, and General Bugs/Errors accounted for 55.4% of the negative reviews. After two surge months were excluded, TimesFM-3 reached a weighted absolute percentage error of 54.65% and an 80% interval coverage of 84.42%. Loading/Performance Issues was the only high-priority category (score 74.5), and four categories were medium priority. Problems After Updates was the only medium-priority category with an increasing trend. Conclusion: The framework supports ranking and monitoring of complaint categories. Its moderate accuracy and wide intervals mean that the forecasts should guide priorities rather than predict exact monthly volumes.

INTRODUCTION

The rapid development of digital technologies has contributed to the growing volume of unstructured textual data, including user reviews on application distribution platforms such as the Google Play Store (Dąbrowski et al., 2022). Sentiment analysis using Natural Language Processing (NLP) offers a practical approach to automatically extracting public opinions from such data, enabling application providers to evaluate the quality of their digital services (Sarvestani, 2025).

MyTelkomsel is one application that receives a substantial volume of user reviews. As the official application of PT Telkomsel, it serves millions of customers who use it to manage their mobile services. These reviews reflect a wide range of user experiences, from satisfaction with application features to complaints about technical problems (Hardiansyah et al., 2024). Technical issues, including login failures, unsuccessful transactions, and slow application performance, can directly affect customers' ability to access services and therefore require appropriate prioritization. However, the increasing volume of user reviews makes manual identification of complaints inefficient. An automated machine learning approach is consequently needed to identify recurring complaint patterns and extract meaningful information from user feedback.

Sentiment analysis and topic extraction from application reviews present several challenges, particularly the prevalence of informal language, non-standard abbreviations, and imbalanced sentiment distributions (Dey et al., 2026). These characteristics require models capable of understanding Indonesian linguistic contexts, including the informal expressions commonly found in social media posts and user reviews. To address this challenge, this study employs IndoBERTweet (Koto et al., 2021), a language model pretrained on an Indonesian Twitter corpus. The model is fine-tuned on SmSA, a sentence-level sentiment dataset from the IndoNLU benchmark containing Indonesian comments and reviews collected from various online sources (Ahmadian et al., 2024; Jayadianti et al., 2022), to classify sentiment into three categories (positive, neutral, and negative) before being applied to MyTelkomsel reviews..

Sentiment classification alone, however, does not reveal the specific types of technical problems experienced by users. Therefore, this study employs BERTopic to group reviews into complaint categories through contextual embedding-based topic modeling (Janssens et al., 2025). For short-text datasets, the default HDBSCAN clustering algorithm in BERTopic may classify a substantial proportion of documents as outliers. To ensure that each review is

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

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9879

assigned to a cluster, K-Means is used as an alternative clustering algorithm (Li et al., 2025). The resulting topics are manually labeled based on their keywords and representative documents, following established practices in BERTopic-based research (Sağlam, 2026). Although this process provides an overview of existing complaint categories, it does not indicate how those complaints may evolve over time. To address this limitation, the study uses Google's TimesFM-3, a time-series foundation model based on a decoder-only Transformer architecture (Yang et al., 2025), to forecast complaint trends for each category.

Previous studies have investigated sentiment analysis of Indonesian-language application reviews using BERT-based models (Imaduddin et al., 2023), but many have primarily focused on sentiment classification. Research integrating sentiment classification, complaint topic extraction, and complaint trend forecasting for Indonesian-language telecommunications applications remains limited. Consequently, service improvement prioritization has yet to fully benefit from predictive insights derived from user reviews.

To address this research gap, this study proposes an integrated framework combining IndoBERTweet, BERTopic, and TimesFM-3. The study aims to (1) identify the dominant categories of technical complaints in MyTelkomsel reviews, (2) forecast the trends of individual complaint categories, and (3) establish service improvement priorities based on the resulting analyses. The main contribution of this research is a review analysis framework that extends beyond describing existing complaints by forecasting their potential trajectories, thereby providing data-driven insights to support service improvement decisions.

LITERATURE REVIEW

Previous studies on Indonesian-language application reviews support the selection of IndoBERTweet for sentiment classification. An earlier sentiment analysis study of MyTelkomsel reviews employed Naïve Bayes and Random Forest but focused exclusively on sentiment classification, without identifying complaint categories or forecasting their subsequent development (Rudi Sanjaya et al., 2026). In a study of Gojek application reviews, IndoBERTweet reportedly outperformed all evaluated IndoBERT configurations (Aqiilah et al., 2026). Similarly, a comparison of five NLP models applied to reviews of the IKD, BPJS Kesehatan Mobile, and MyPertamina applications found that IndoBERTweet achieved the best performance, while Naïve Bayes and SVM remained viable alternatives for environments without GPU support (Setiawan & Andriyani, 2026). Regarding training data, SmSA continues to be used in recent studies (Nasution et al., 2025) and has also been employed in the development of a hybrid IndoBERT model (Ahmadian et al., 2024), These studies support the use of SmSA as a fine-tuning dataset for three-class sentiment classification.

The use of BERTopic for complaint categorization is supported by previous research on topic modeling. BERTopic groups documents according to semantic similarity and, in its default configuration, does not require the number of topics to be specified in advance (Janssens et al., 2025). Its workflow involves generating document embeddings, reducing dimensionality using UMAP, clustering documents, and extracting representative keywords through class-based Term Frequency–Inverse Document Frequency (c-TF-IDF) (Egger & Yu, 2022). Replacing HDBSCAN with K-Means for short-text clustering, with the aim of assigning every review to a cluster, follows the approach described by Li et al. (Li et al., 2025). Since K-Means requires the number of clusters to be specified beforehand, this study evaluates multiple values of K before selecting the final configuration. The resulting clusters are subsequently labeled and grouped manually based on their keywords and representative documents, consistent with common practices in BERTopic-based studies (Sağlam, 2026).

Time-series foundation models offer the potential for zero-shot forecasting; however, empirical evidence regarding their reliability remains limited, warranting careful evaluation (Emmert-StreibFrank et al., 2026). TimesFM is a decoder-only model trained on a large corpus of time-series data and can generate forecasts for previously unseen series without additional model training (Emmert-StreibFrank et al., 2026). TimesFM-3 was released by Google Research on August 31, 2026, as a 330-million-parameter model trained on more than one trillion time points (*TimesFM-3: A Zero-Shot Foundation Model for Multivariate Forecasting*, n.d.). Its developers report that it ranks first among foundation models on the GIFT-Eval, FEV-Bench, and TIME benchmarks. However, these performance claims originate from the model developers and should therefore be interpreted in that context. In their study of telecommunications data from China, Yang et al. (2025) identified TimesFM as a promising forecasting approach because of its few-shot capabilities. They noted that ARIMA models require sufficient historical observations, whereas transformer-based models may overfit when trained on limited data. This issue is particularly relevant to the present study, which contains only 45 monthly observations for each complaint category. Nevertheless (Yang et al., 2025) reported forecasting failures for approximately 23% of resource pools at a 64-day horizon and 38% at a 128-day horizon. They attributed these limitations to distributional differences and the accumulation of forecasting errors over longer horizons, based on an earlier generation of TimesFM rather than TimesFM-33. These findings motivate the evaluation design adopted in this study, which employs rolling-origin backtesting across forecast horizons of one to six months and incorporates forecast intervals to assess predictive uncertainty.

Taken together, the reviewed studies indicate that previous sentiment analysis of MyTelkomsel reviews has not extended to complaint categorization and trend forecasting (Rudi Sanjaya et al., 2026), Meanwhile, the reviewed

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

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9879

application of TimesFM to telecommunications data focused on cloud resource forecasting rather than customer complaints (Yang et al., 2025). To the best of the authors' knowledge, no previous study has integrated sentiment classification, complaint categorization, and complaint-category trend forecasting for an Indonesian-language telecommunications application. This study addresses that gap by combining IndoBERTweet, BERTopic, and TimesFM-3 into an integrated framework for identifying technical complaints, forecasting their development over time, and supporting data-driven service improvement prioritization.

METHOD

This study employs an integrated text-mining and time-series forecasting framework to identify technical complaints in Indonesian-language reviews of the MyTelkomsel application and support service improvement prioritization. The framework combines IndoBERTweet for sentiment classification, BERTopic with K-Means clustering for complaint categorization, and TimesFM-3 for forecasting monthly complaint counts. The research workflow consists of eight consecutive stages: data collection, data preprocessing and relevance filtering, sentiment classification, topic modeling and cluster selection, complaint category consolidation, monthly time-series construction, forecasting and rolling-origin evaluation, and service improvement priority scoring.

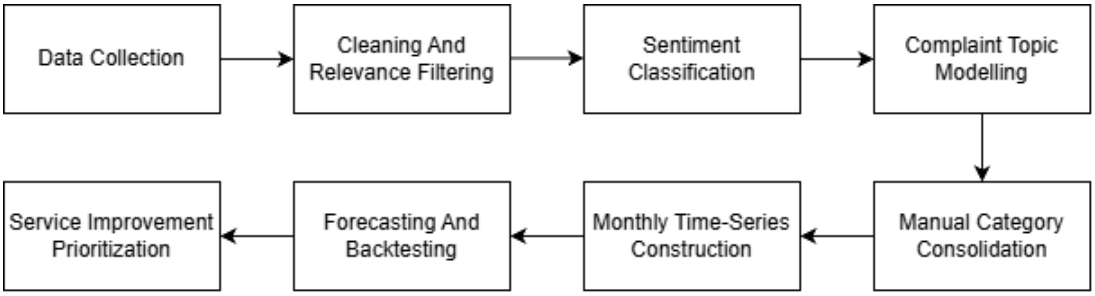


Figure 1. Research Methodology Workflow

Data Collection

The dataset was collected from the public Google Play Store page of the MyTelkomsel application, identified by the application ID com.telkomsel.telkomselcm. Data collection was performed using the google-play-scraper library in Python. The scraper was configured with Indonesian language and country settings (lang='id' and country='id') and retrieved reviews in descending chronological order using Sort.NEWEST. Each request retrieved up to 200 reviews, followed by a one-second pause before the next request. The collection process was configured to obtain 600,000 reviews, and the target number was reached.

The collected attributes consisted of review text, star rating, publication timestamp, and helpful-vote count. These attributes were stored in a Pandas DataFrame for subsequent analysis. The dataset covered reviews published between November 28, 2022, and September 27, 2026. Reviewer names and identifiers were not retained in the analysis dataset. The resulting raw dataset was subsequently processed to remove incomplete records and irrelevant review content.

Data Preprocessing and Relevance Filtering

A keyword-based relevance filter was subsequently applied to distinguish application-related reviews from those primarily concerning network services or pricing. The filter consisted of 26 application-related keywords, including login, OTP, error, bug, crash, loading, and force close, and 16 service- or price-related keywords, including sinyal (signal), jaringan (network), mahal (expensive), and kuota hilang (missing data quota).

Each review was assigned to one of four relevance groups: application-related, service-related, mixed, or unclear. Reviews containing at least one application-related keyword and no service- or price-related keyword were classified as application-related. Reviews matching service- or price-related keywords without application-related keywords were classified as service-related. Reviews matching both keyword lists were categorized as mixed, while those matching neither list were classified as unclear.

Table 1. Distribution of reviews after relevance filtering

Relevance category	Number of reviews	Percentage (%)
Application-related	51,112	10.35
Service-related	151,323	30.64
Mixed	0,299	2.06
Unclear	281,201	56.95
Total	493,806	100.00

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

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9879

Only the 51,112 application-related reviews were retained for the subsequent sentiment classification stage. The service-related, mixed, and unclear groups were excluded to maintain the study's focus on application-specific problems. Therefore, the resulting corpus represents reviews enriched for application-related issues rather than the entire population of MyTelkomsel reviews.

Sentiment Classification

Sentiment classification was performed using IndoBERTweet (indolem/indobertweet-base-uncased), a pretrained Indonesian language model based on social media text (Koto et al., 2021). The model was fine-tuned using the SmSA dataset from the IndoNLU benchmark (Purwarianti & Crisdayanti, 2019; Wilie et al., 2020) to classify reviews into three sentiment categories: positive, neutral, and negative.

The SmSA dataset consisted of 11,000 training documents and 1,260 validation documents. The original sentiment labels were mapped to numerical identifiers before being supplied to the model. Fine-tuning was performed using the Hugging Face Transformers library with the parameters summarized in Table 2.

Table 2. IndoBERTweet training configuration and validation performance

Parameter	Configuration
Pretrained model	indolem/indobertweet-base-uncased
Training dataset	SmSA training split
Number of training documents	11,000
Number of validation documents	1,260
Learning rate	2×10^{-5}
Training batch size	16
Validation batch size	16
Number of epochs	4
Weight decay	0.01
Maximum sequence length	128 tokens
Checkpoint selection	Highest validation macro-F1
Validation accuracy	94.44%
Validation macro-F1	92.76%

The selected checkpoint was applied to 51,112 cleaned MyTelkomsel reviews classified as application-related, using a batch size of 32. Each review was assigned the sentiment with the highest predicted probability, and its prediction score was retained. Reviews classified as negative were subsequently used for complaint categorization.

A separate manual assessment used 598 labelled MyTelkomsel reviews sampled across the three predicted sentiment classes. Population weights were applied to account for the sampling proportions, and 95% confidence intervals were estimated using bootstrap resampling. As the available annotations came from one annotator, inter-annotator agreement was not assesse.

Complaint Topic Modeling

Complaint topic modeling was performed on 44,937 negative MyTelkomsel reviews to identify recurring technical issues. The analysis employed BERTopic, an embedding-based topic modeling framework that represents topics using class-based Term Frequency–Inverse Document Frequency (c-TF-IDF) (Egger & Yu, 2022). Review texts were transformed into 384-dimensional embeddings using paraphrase-multilingual-MiniLM-L12-v2. The embeddings were reduced to five dimensions using Uniform Manifold Approximation and Projection (UMAP), with 15 neighbors, a minimum distance of 0.0, cosine distance, and a random state of 42. K-Means was used as the clustering algorithm within BERTopic, with 10 initializations and a random state of 42. Topic representations were extracted using unigram and bigram features, Indonesian stopwords supplemented with colloquial expressions, and a minimum document frequency of two.

Cluster selection was conducted in two stages. First, candidate values of K = 15, 20, 30, 40, 50, 60, and 80 were evaluated using the silhouette score, adjusted Rand index (ARI) stability between runs with random seeds 42 and 7, cluster sizes, and inertia. The silhouette score was calculated using a sample of 10,000 reviews. The screening results are presented in Table 3.

Table 3. K-Means candidate cluster screening results

Number of clusters (K)	Sillhouette score	ARI stability	Largest cluster (%)	Smallest cluster (%)	Inertia
15	0.406	0.995	14.5	0.09	57,576
20	0.404	0.971	9.8	0.09	43,608
30	0.403	0.828	7.8	0.09	29,409
40	0.388	0.832	5.6	0.04	21,660
50	0.384	0.766	5.7	0.04	17,462

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

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, Number 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9879

60	0.397	0.787	5.5	0.04	14,029
80	0.377	0.767	3.5	0.04	10,515

In the second stage, BERTopic configurations with K = 20, 30, 40, and 60 were compared using topic coherence and unigram diversity, complementary measures of topic quality discussed in previous comparative research (Janssens et al., 2025). Coherence was calculated using the C_v metric based on the top ten unigram keywords, while unigram diversity measured the proportion of unique words among the extracted topic keywords. The results are presented in Table 4.

Table 4. Evaluation of BERTopic configurations

K	Number of topics	Largest topic (%)	Topics below 0.5%	C_v coherence	Unigram diversity
20	20	9.8	4	0.519	0.485
30	30	7.8	5	0.546	0.450
40	40	5.6	7	0.585	0.439
60	60	5.5	12	0.598	0.413

The configuration with K = 60 was selected because it achieved the highest C_v coherence score (0.598) among the four configurations evaluated in the second stage. Although it produced 12 topics containing less than 0.5% of the reviews and had lower unigram diversity than the other candidates, it provided the highest coherence score in this evaluation. The resulting clusters were subsequently examined using representative keywords and review samples before being consolidated into ten broader complaint categories for service improvement prioritization.

Complaint Category Consolidation

Clusters obtained from the selected K-Means configuration were consolidated into ten broader complaint categories. This process was performed through a manually defined cluster-to-category mapping based on topic keywords, representative reviews, and sampled review texts. Each cluster was assigned to one category according to its dominant complaint pattern. This consolidation was intended to simplify the interpretation of recurring technical issues and support subsequent monthly trend analysis.

The mapping covered all 60 clusters, with each cluster assigned to exactly one of the ten categories. The resulting category distribution is presented in Table 5.

Table 5. Distribution of Consolidated Complaint Categories

No.	Complaint category	Number of reviews	Percentage (%)	Main complaint scope
1	Login & Account Access	8,809	19.6	Login failures, account access difficulties, and repeated logouts.
2	Loading/Performance Issues	8,503	18.9	Slow loading, application lag, and delayed responses.
3	General Bugs/Errors	7,605	16.9	General application errors, bugs, and unexpected behavior.
4	Mixed/General	5,727	12.7	Broad or overlapping complaints without one clearly dominant issue.
5	Problems After Updates	5,002	11.1	Complaints about degraded functionality or performance after application updates.
6	App Not Opening/Crashes	4,420	9.8	Failure to open the application, force closes, and crashes.
7	Transactions/Pulsa	2,150	4.8	Transaction failures, pulsa issues, and unsuccessful package purchases or payments.
8	OTP/SMS/Link Verification	1,780	4.0	Missing verification messages, OTP problems, and invalid or expired verification links.
9	UI/Display	582	1.3	Interface appearance, layout changes, and difficulties navigating the application.
10	Rewards/Stamp	359	0.8	Reward claims, stamp features, points, and related functionality.
	Total	44,937	100.0*	

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

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9879

The largest category was Login & Account Access, accounting for 8,809 reviews (19.6%), followed by Loading/Performance Issues with 8,503 reviews (18.9%) and General Bugs/Errors with 7,605 reviews (16.9%). Together, these three categories represented 55.4% of the negative reviews analyzed. Problems After Updates accounted for a further 11.1%, while App Not Opening/Crashes represented 9.8%.

The resulting categories provide a structured basis for aggregating complaints by month in the subsequent time-series analysis. However, the Mixed/General category reflects the presence of broad or overlapping complaint patterns that could not be represented adequately by a more specific category under the current mapping.

Monthly Time-Series Construction

The 44,937 negative reviews were aggregated by publication month into one monthly complaint-count series for each of the ten categories. Months without complaints were set to zero, giving a regular series from November 2022 to September 2026 (47 months). The panel total matched the number of negative reviews. Monthly aggregation was used because daily counts were sparse for the smallest categories (0.26 complaints per day for Rewards/Stamp and 0.42 for UI/Display).

Data completeness was audited using monthly aggregates of all 51,112 application-related reviews. A month was flagged as thin if its volume was below 25% of the median monthly volume (691 reviews) and as sparse if reviews appeared on fewer than 60% of its days. It was flagged as partial if collection began after the third day of the first month or if the last review was published three or more days before the end of the last month. Two edge months were flagged and removed: November 2022 (103 reviews on three days) and September 2026 (last review on September 27). No flagged month occurred inside the series, so no interpolation was needed. The final series covered December 2022 to August 2026 (45 months, 44,473 complaints).

December 2024 (7,002 reviews) and January 2025 (5,699 reviews) were far above the median of 691. Together they contained 10,735 negative reviews (24.1% of the cleaned series). Both months were retained in the model input without adjustment. Forecast accuracy was evaluated on all months and again excluding these two months as targets. Complaint volume and volatility varied widely across categories, with a coefficient of variation (CV) from 0.52 to 2.72 (Table 6).

Table 6. Descriptive Statistics of the Monthly Complaint Series (December 2022–August 2026)

Complaint category	Total	Mean	Median	Maximum (month)	CV	Zero months
Login & Account Access	8,693	193.2	110	1,357 (Aug 2024)	1.10	0
Loading/Performance Issues	8,413	187.0	113	739 (Jul 2025)	1.00	0
General Bugs/Errors	7,552	167.8	82	1,546 (Dec 2024)	1.83	0
Mixed/General	5,671	126.0	76	775 (Dec 2024)	1.11	0
Problems After Updates	4,952	110.0	62	838 (Dec 2024)	1.34	0
App Not Opening/Crashes	4,394	97.6	27	1,488 (Dec 2024)	2.72	0
Transactions/Pulsa	2,122	47.2	40	128 (Jan 2025)	0.52	0
OTP/SMS/Link Verification	1,746	38.8	30	133 (Oct 2024)	0.74	0
UI/Display	575	12.8	5	126 (May 2024)	1.91	4
Rewards/Stamp	355	7.9	5	39 (Dec 2022)	1.05	1

Forecasting and Rolling-Origin Evaluation

Monthly complaint counts were forecast with TimesFM-3 (google/timesfm-3.0-pytorch) using the timesfm Python package. The model was applied zero-shot, without fine-tuning on the MyTelkomsel data, and run on a CPU in Google Colab. Each of the ten category series was forecast separately as a univariate series with a six-month horizon. Forecasts were constrained to be non-negative, and the 80% prediction interval was taken from the 10th and 90th quantiles of the model output.

Forecast accuracy was evaluated with a rolling-origin backtest. The first origin used 18 months of context, and the context was extended by one month at each subsequent origin, so the model only saw observations available up to that origin. This gave 22 origins, with target months from June 2024 to August 2026, and 1,320 forecast-actual pairs across ten categories and six horizons. The evaluation was repeated excluding December 2024 and January 2025 as targets, which left 1,200 pairs.

Accuracy was measured with 5 metrics, given in (1)-(5) : mean absolute error (MAE), root mean squared error (RMSE), weighted absolute percentage error (WAPE), symmetric mean absolute percentage error (MAPE) and bias.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i| \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} \quad (2)$$

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

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9879

$$WAPE = \frac{\sum_{i=1}^{n}|e_i|}{\sum_{i=1}^{n}|y_i|} \times 100\% \quad (3)$$

$$sMAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{2|e_i|}{|y_i|+|\hat{y}_i|} \quad (4)$$

$$Bias = \frac{1}{n}\sum_{i=1}^{n} e_i \quad (5)$$

Where $e_i = \hat{y}_i - y_i$ is the forecast error for pair i, y_i is the actual complaint count, $\hat{y}_i$ is the forecast, and n is the number of forecast-actual pairs. A negative bias indicates underestimation.

WAPE and sMAPE were used instead of MAPE because some series contain zero or very small monthly counts, and pairs in which both the actual count and the forecast were zero were excluded from sMAPE. Interval quality was measured by the 80% coverage, the percentage of actual counts between the 10th and 90th quantile forecasts. Directional accuracy was the percentage of forecasts that matched the actual direction of change from the last observed month. Metrics were computed overall, by category, and by horizon. Finally, the full 45-month series was used as context to forecast September 2026 to February 2027.

Forecasting and Rolling-Origin Evaluation

Priorities were derived from the forecasts for September 2026 to February 2027. For each category k, the forecast load $\hat{y}_k$ was the mean of the six monthly point forecasts, and the baseline b_k was the mean of the last six observed months (March to August 2026). The forecast trend was the percentage change from the baseline (6) and was labeled increasing if above +10%, decreasing if below −10%, and stable otherwise.

$$\Delta_k = \frac{\hat{y}_k - b_k}{b_k} \times 100\% \quad (6)$$

The priority score P_k combined load V_k (7) and trend G_k (8), weighted by w = 0.6 and 1 − w, as given in (9).

$$V_k = \frac{\hat{y}_k}{M} \quad (7)$$

$$G_k = \frac{\Delta_k^c + 50}{100} \quad (8)$$

$$P_k = 100[wV_k + (1-w)G_k] \quad (9)$$

where $\hat{y}_{max}$ is the largest forecast load among the ten categories and Δ_k^c is Δ_k truncated to the range −50% to +50%. G_k was set to 0.5 when $\hat{y}_k$ or b_k was below 10 complaints per month. Categories were classified as high ($P_k \geq 60$), medium ($40 \leq P_k < 60$), or low ($P_k < 40$) priority.

Robustness was checked by recomputing the ranking with w from 0.4 to 0.8 and the trend with baselines of 3, 6, and 12 months. Categories whose backtest WAPE (excluding the two surge months) exceeded 70% were flagged as less reliable, and Mixed/General was flagged as heterogeneous. As a plausibility check, the partial September 2026 counts were scaled to a full month (× 30/27) and compared with the 80% interval for that month.

RESULT

Forecasting and Rolling-Origin Evaluation

The rolling-origin backtest produced 1,320 forecast-actual pairs. Across all pairs, TimesFM-3 obtained an MAE of 90.59, an RMSE of 229.95, a WAPE of 65.46%, and an sMAPE of 60.28%. The bias was −44.22 complaints per month, 79.62% of the actual counts fell inside the 80% interval, and directional accuracy was 61.52%. After excluding December 2024 and January 2025 as targets (1,200 pairs), MAE decreased to 53.86, RMSE to 114.80, WAPE to 54.65%, and sMAPE to 54.06%. The bias became −2.86, coverage increased to 84.42%, and directional accuracy to 64.75%.

Forecast error generally increased with the horizon. Excluding the surge months, WAPE increased from 49.01% for one-month-ahead forecasts to 59.85% for six-month-ahead forecasts, and sMAPE increased from 49.14% to 57.38%. Coverage ranged from 79.5% to 90.0% across horizons, and directional accuracy increased from 61.0% to 69.5%. With all months included, WAPE increased from 56.24% to 70.71%, and the bias grew from −32.27 to −49.44.

Table 7 shows the accuracy for each category. Excluding the surge months, WAPE was lowest for Transactions/Pulsa (32.3%), OTP/SMS/Link Verification (40.5%), and Login & Account Access (45.3%), and highest for Rewards/Stamp (79.8%), App Not Opening/Crashes (75.7%), and UI/Display (70.1%). Loading/Performance Issues had the lowest coverage (67.5%) and the largest negative bias (−50.3), while Problems After Updates had the largest

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

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9879

positive bias (+23.2). Excluding the surge months lowered WAPE in eight categories, with the largest reductions in General Bugs/Errors (78.3% to 63.6%), Problems After Updates (69.5% to 55.5%), and App Not Opening/Crashes (89.5% to 75.7%), and raised it slightly in UI/Display and Rewards/Stamp.

Table 7. Backtest Accuracy by Complaint Category

Complaint category	MAE	WAPE (%)	sMAPE (%)	Bias	80% coverage (%)	Direction (%)	WAPE, all months (%)
Login & Account Access	93.0	45.3	39.2	−3.2	92.5	52.5	47.6
Loading/Performance Issues	165.1	61.7	58.4	−50.3	67.5	61.7	64.6
General Bugs/Errors	96.3	63.6	55.8	1.1	75.0	66.7	78.3
Mixed/General	63.6	50.5	49.9	−2.8	90.0	62.5	62.8
Problems After Updates	43.8	55.5	53.5	23.2	88.3	73.3	69.5
App Not Opening/Crashes	29.2	75.7	61.0	7.8	89.2	68.3	89.5
Transactions/Pulsa	18.1	32.3	32.0	0.1	85.0	65.8	37.3
OTP/SMS/Link Verification	19.7	40.5	38.7	−5.5	81.7	63.3	42.3
UI/Display	4.8	70.1	69.7	1.9	90.0	69.2	66.0
Rewards/Stamp	5.1	79.8	82.8	−0.9	85.0	64.2	79.2
All categories	53.9	54.7	54.1	−2.9	84.4	64.8	65.5

Note. Values are for the 1,200 forecast-actual pairs (120 per category) that exclude December 2024 and January 2025 as targets, except the last column, which uses all 1,320 pairs. Bias is the mean of forecast minus actual (complaints per month). Direction is the directional accuracy

Forecast of Complaint Volumes

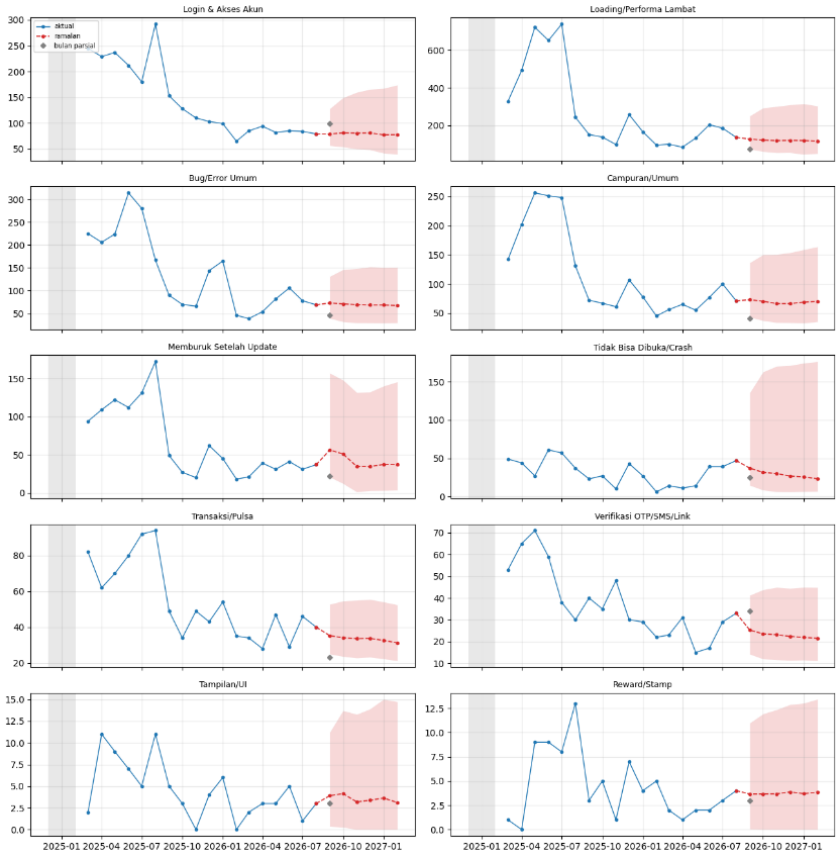


Figure 2. Observed monthly complaints (December 2022–August 2026) and TimesFM-3 forecasts for September 2026–February 2027 with 80% intervals

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

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9879

Figure 2 shows the observed series and the forecasts for September 2026 to February 2027. Loading/Performance Issues had the highest forecast, from 117.2 to 128.0 complaints per month, followed by Login & Account Access (77.3 to 81.1), General Bugs/Errors (67.6 to 73.2), and Mixed/General (66.2 to 72.9). The forecast for Problems After Updates decreased from 56.3 in September 2026 to 34.6 in November and December 2026, and that for App Not Opening/Crashes decreased from 37.0 to 23.3. Transactions/Pulsa and OTP/SMS/Link Verification decreased gradually, from 35.2 to 31.2 and from 25.2 to 21.5, while UI/Display and Rewards/Stamp stayed below five complaints per month. The 80% intervals were wide. For example, the interval for Loading/Performance Issues was 75 to 251 in September 2026 and 52 to 304 in February 2027, and the interval for Problems After Updates was 21 to 157 in September 2026.

Service Improvement Priorities

Table 8 and Figure 3 present the priority scores. Loading/Performance Issues was the only high-priority category (74.5), because it had the largest forecast load (25.6% of the total forecast) although its trend was decreasing (−13.8%). Four categories were medium priority: Login & Account Access (56.7), General Bugs/Errors (53.6), Mixed/General (53.2), and Problems After Updates (50.8). The latter was the only medium-priority category with an increasing trend (+25.4%). The five remaining categories were low priority, with scores from 21.8 to 36.8. UI/Display and Rewards/Stamp had forecasts below 10 complaints per month, so their increasing changes (+25.5% and +59.9%) were not scored. App Not Opening/Crashes, Rewards/Stamp, and UI/Display had a backtest WAPE above 70%, and Mixed/General was flagged as heterogeneous.

Table 8. Service Improvement Priorities Based on the TimesFM-3 Forecasts

Rank	Complaint Category	Actual	Forecast	80% interval	Change (%)	Trend	WAPE (%)	Score	Priority
1	Loading/Performance Issues	141.0	121.5	58–295	−13.8	Decreasing	61.7	74.5	High
2	Login & Account Access	84.8	79.4	48–157	−6.4	Stable	45.3	56.7	Medium
3	General Bugs/Errors	71.3	69.7	32–147	−2.2	Stable	63.6	53.6	Medium
4	Mixed/General	70.7	69.1	36–152	−2.3	Stable	50.5	53.2	Medium
5	Problems After Updates	33.3	41.8	7–142	+25.4	Increasing	55.5	50.8	Medium
6	App Not Opening/Crashes	27.3	29.0	8–165	+6.2	Stable	75.7	36.8	Low
7	Transactions/Pulsa	37.3	33.4	23–54	−10.5	Decreasing	32.3	32.3	Low
8	OTP/SMS/Link Verification	24.7	23.0	12–44	−6.9	Stable	40.5	28.6	Low
9	Rewards/Stamp	2.3	3.7	0–12	+59.9	Increasing	79.8	21.8	Low
10	UI/Display	2.8	3.6	0–14	+25.5	Increasing	70.1	21.8	Low

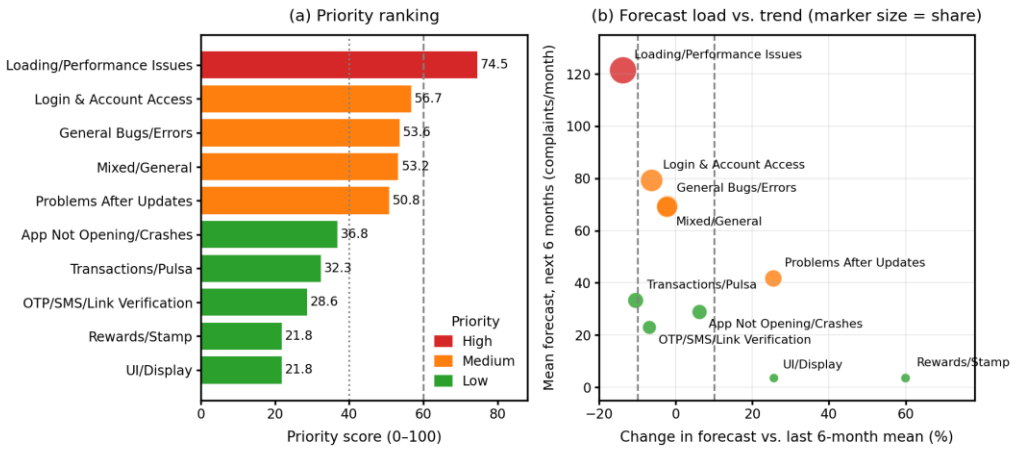


Figure 3. Priority ranking (a) and forecast load versus trend (b). Lines mark the priority thresholds of 40 and 60 in (a) and the ±10% trend thresholds in (b)

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

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9879

The ranking was sensitive to the volume weight for some categories. Loading/Performance Issues ranked first at every weight from 0.4 to 0.8, and the five lowest-ranked categories kept their positions. Problems After Updates ranked second at $w = 0.4$ and 0.5 and fifth at $w \geq 0.6$, while Login & Account Access, General Bugs/Errors, and Mixed/General each moved up one rank at $w \geq 0.6$. The trend direction was the same for the 3-, 6-, and 12-month baselines in five categories (Loading/Performance Issues, Problems After Updates, Transactions/Pulsa, Rewards/Stamp, and UI/Display) and changed in the other five. For App Not Opening/Crashes, the change was −30.4% with the 3-month baseline, +6.2% with the 6-month baseline, and +16.1% with the 12-month baseline.
After the partial September 2026 counts were scaled to a full month (×1.11), the scaled counts fell within the 80% interval of the September 2026 forecast for all ten categories. For example, the scaled counts were 82.2 for Loading/Performance Issues (interval 75 to 251) and 110.0 for Login & Account Access (interval 57 to 128).

DISCUSSION

The results show that MyTelkomsel complaints are dominated by access and performance problems. Login & Account Access, Loading/Performance Issues, and General Bugs/Errors made up 55.4% of the 44,937 negative reviews, while Rewards/Stamp and UI/Display were below 2%. This pattern partly reflects the relevance filter, which kept reviews with application-related terms such as login, error, and crash, so the share of negative reviews (87.9%) does not describe overall user satisfaction. The Mixed/General category (12.7%) shows the cost of K-Means, which assigns every review to a cluster, including vague ones (Li et al., 2025).

TimesFM-3 reached moderate accuracy without fine-tuning on only 45 monthly observations. WAPE was 65.46% with all months and 54.65% after excluding December 2024 and January 2025. The model underestimated this surge (bias −44.22 versus −2.86), which is expected because a zero-shot model sees only past counts. The 80% intervals covered 79.62% and 84.42% of the actual counts, close to the nominal level, but they were wide. Error also increased with the horizon, from 49.01% at one month to 59.85% at six months, which agrees with Yang et al. (2025). WAPE was higher for volatile series such as App Not Opening/Crashes (75.7%) than for stable ones such as Transactions/Pulsa (32.3%).

Loading/Performance Issues was the only high-priority category because of its large forecast load (25.6% of the total), although its trend was decreasing. Its backtest bias was negative (−50.3), so its priority is unlikely to be overstated, and it ranked first at every volume weight from 0.4 to 0.8. Problems After Updates was the only medium-priority category with an increasing trend (+25.4%), but its rank changed with the weight and its interval was wide, so it is better treated as a watch item. All high- and medium-priority categories had a WAPE below 70% (Table 8), so the reliability limits do not change the top of the ranking.

An earlier MyTelkomsel study classified sentiment only (Rudi Sanjaya et al., 2026). This study adds complaint categorization and category-level forecasting. IndoBERTweet follows earlier findings that it outperformed other compared models on Indonesian application reviews (Aqiilah et al., 2026; Setiawan & Andriyani, 2026). Yang et al. (2025) applied TimesFM to telecommunications cloud resources, whereas this study applies TimesFM-3 to complaint counts. A direct accuracy comparison is not possible because the data and metrics differ.

CONCLUSION

This study developed an integrated framework that combines IndoBERTweet, BERTopic with K-Means, and TimesFM-3 to identify MyTelkomsel complaints, forecast their development, and support service improvement prioritization. A total of 51,112 application-related reviews were classified by sentiment. The 44,937 negative reviews were grouped into ten complaint categories and aggregated into monthly series from December 2022 to August 2026.

Login & Account Access, Loading/Performance Issues, and General Bugs/Errors were the dominant complaint categories, accounting for 55.4% of the negative reviews. After two surge months were excluded, TimesFM-3 achieved a WAPE of 54.65% and an 80% interval coverage of 84.42% in the rolling-origin backtest. The forecasts are therefore suitable for ranking and monitoring complaint categories, not for predicting exact monthly volumes. Loading/Performance Issues was the only high-priority category (74.5), and four categories were medium priority. Problems After Updates was the only medium-priority category with an increasing trend.

The main contribution of this study is a review analysis framework that goes beyond describing complaints. It forecasts the trajectory of each complaint category and ranks the categories with an explicit reliability check. Future studies can compare TimesFM-3 with baseline models, add release and incident information, and validate the forecasts when actual data for September 2026 to February 2027 become available.

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

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9879

REFERENCES

Ahmadian, H., Abidin, T. F., Riza, H., & Muchtar, K. (2024). Hybrid Models for Emotion Classification and Sentiment Analysis in Indonesian Language. *Applied Computational Intelligence and Soft Computing*, *2024*. https://doi.org/10.1155/2024/2826773

Aqiilah, I. N., Saptono, R., & Syaifuddin, A. (2026). Sentence-Level Sentiment Analysis of Indonesian App Reviews Using IndoBERTweet. *Journal of Computing Theories and Applications*, *4*(1), 74–91. https://doi.org/10.62411/JCTA.16240

Dąbrowski, J., Letier, E., Perini, A., & Susi, A. (2022). Analysing app reviews for software engineering: a systematic literature review. *Empirical Software Engineering 2021 27:2*, *27*(2), 43-. https://doi.org/10.1007/S10664-021-10065-7

Dey, T., Zhang, Z., & Stefanidis, K. (2026). Improving Sentiment Analysis of Mobile App Reviews through Emoji and Slang Normalization. *Proceedings - 2026 IEEE 42nd International Conference on Data Engineering Workshops, ICDEW 2026*, 85–92. https://doi.org/10.1109/ICDEW71238.2026.00016

Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, *7*, 886498. https://doi.org/10.3389/FSOC.2022.886498/TEXT

Emmert-StreibFrank, TripathiShailesh, & Yli-HarjaOlli. (2026). Time Series Forecasting Methods: A Systematic Review from Classical Models to Deep Learning. *ACM Computing Surveys*, *58*(14), 1–39. https://doi.org/10.1145/3828164

Hardiansyah, D., Aziz, A., & Hasibuan, M. S. (2024). The Classification Method is Used for Sentiment Analysis in My Telkomsel. *International Journal of Artificial Intelegence Research*, *8*(2). http://ijair.id

Imaduddin, H., A'la, F. Y., & Nugroho, Y. S. (2023). Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach. *International Journal of Advanced Computer Science and Applications*, *14*(8), 113–117. https://doi.org/10.14569/IJACSA.2023.0140813

Janssens, W., Bogaert, M., & Van Den Poel, D. (2025). A Comparative Analysis of Topic Reduction Techniques for BERTopic. *IEEE Access*, *13*, 204087–204103. https://doi.org/10.1109/ACCESS.2025.3638956

Jayadianti, H., Kaswidjanti, W., Utomo, A. T., Saifullah, S., Felix, ;, Dwiyanto, A., & Drezewski, R. (2022). Sentiment analysis of Indonesian reviews using fine-tuning IndoBERT and R-CNN. *ILKOM Jurnal Ilmiah*, *14*(3), 348–354. https://doi.org/10.33096/ilkom.v14i3.1505.348-354

Koto, F., Lau, J. H., & Baldwin, T. (2021). IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 10660–10668. https://doi.org/10.18653/V1/2021.EMNLP-MAIN.833

Li, J., Lee, B., & Kim, J. (2025). Analyzing the Asymmetric Effects of COVID-19 on Hotel Selection Attributes and Customer Satisfaction Through AIPA. *Sustainability 2025, Vol. 17, Page 8546*, *17*(19), 8546. https://doi.org/10.3390/SU17198546

Nasution, A. H., Onan, A., Murakami, Y., Monika, W., & Hanafiah, A. (2025). Benchmarking Open-Source Large Language Models for Sentiment and Emotion Classification in Indonesian Tweets. *IEEE Access*, *13*, 94009–94025. https://doi.org/10.1109/ACCESS.2025.3574629

Rudi Sanjaya, M., Khoiriah, A., Heroza, R. I., Putra, B. W., Stik, K., & Palembang, I. (2026). Analysis of User Reviews for The Mytelkomsel App Using Naïve Bayes and Random Forest Methods. *Jurnal Testing Dan Implementasi Sistem Informasi*, *4*(1), 32–48. https://doi.org/10.55583/JTISI.V4I1.2220

Sağlam, M. H. (2026). From Delivery Delays to AI-Mediated Escalation Failures: A BERTopic Analysis of Complaints About Risk and Trust in E-Commerce Marketplaces (2019–2025). *Journal of Theoretical and Applied Electronic Commerce Research* , *21*(4), 116. https://doi.org/10.3390/JTAER21040116/S1

Sarvestani, M. (2025). An NLP-Based Framework for Sentiment and Topic Analysis of Citizen Feedback on U.K. Government Mobile Applications. *IEEE Access*, *13*, 210360–210377. https://doi.org/10.1109/ACCESS.2025.3641669

Setiawan, I., & Andriyani, W. (2026). Perbandingan Kinerja dan Efisiensi Model NLP pada Analisis Sentimen Ulasan Aplikasi Layanan Publik Digital. *Jurnal Sistem Komputer Dan Informatika (JSON)*, *7*(4), 1505–1517. https://doi.org/10.30865/JSON.V7I4.9774

TimesFM-3: A zero-shot foundation model for multivariate forecasting. (n.d.). Retrieved October 18, 2026, from https://research.google/blog/timesfm-3-a-zero-shot-foundation-model-for-multivariate-forecasting/

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding. *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL-IJCNLP 2020*, 843–857. https://doi.org/10.18653/V1/2020.AACL-MAIN.85

Yang, X., Zheng, Q., Zhu, X., Luo, M., Hou, Z., Zhang, J., & Lan, Y. (2025). DimAug-TimesFM: Dimension

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

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9879

Augmentation for Long-Term Cloud Demand Forecasting in Few-Shot Scenarios. *Applied Sciences 2025, Vol. 15, Page 3450*, *15*(7), 3450. https://doi.org/10.3390/APP15073450

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