X Sentiment Analysis on Indonesia’s New Capital (IKN) Using TF-IDF+SVM and IndoBERT, and Its Policy-Monitoring Implications

Muhammad Riansyahputra[1]*, Abdul Aziz[2], Agung Purwanto[3]

[1,2,3]*Information System, UNDA University, Jl. Batu Berlian No.10, Mentawa Baru Hulu, Kec. Mentawa Baru Ketapang, Kabupaten Kotawaringin Timur, Kalimantan Tengah, Indonesia*

Keywords

Information; system; technology; software

Correspondence Email:
riankynn@gmail.com

Abstract

This study investigates public perceptions of relocating Indonesia’s national capital to Ibu Kota Nusantara (IKN) through sentiment analysis of Indonesian-language X data, with two target classes: positive and negative. We compare two complementary modeling routes to balance semantic capacity and operational reliability. The lexical route pairs TF-IDF with a linear Support Vector Machine (SVM), providing a lightweight, stable, and reproducible baseline. The contextual route employs IndoBERT, a transformer model tailored to Indonesian, designed to capture implicit meaning and long-range dependencies within sentences. Preprocessing follows contemporary Indonesian NLP practice Unicode normalization, lowercasing, removal of URLs, mentions, and hashtags, normalization of slang into standard forms, removal of numerals and stopwords, and compression of elongated characters to stabilize lexical signal and reduce tokenization artifacts. Because the test data are imbalanced (the positive class is larger), evaluation emphasizes macro-F1 and negative-class recall so that overall accuracy is not inflated by the majority class. Final runs show TF-IDF+SVM achieves Accuracy 0.8908 and macro-F1 0.8709 with negative-recall 0.839; IndoBERT achieves Accuracy 0.9488 and macro-F1 0.9389 with negative-recall 0.920. The recall gain reduces undetected criticism and strengthens the practical value of a social-listening dashboard for governance and environmental issues where early warning is crucial.

1. Introduction

Relocating the national capital to IKN is a strategic decision with fiscal, social, environmental, and governance implications. Social legitimacy for such a decision depends on accurate and timely readings of public sentiment, including not only explicit support but also hesitation and criticism. X, with its high-velocity and open conversational flows, provides a measurable stream of public discourse that can serve as an early sensor for policy-relevant signals (Zachlod et al. 2022)(Yuan et al. 2023). When social-listening outputs are connected to clear follow-up procedures such as clarifying information, convening public Q&A, or adjusting communication. Institutions can respond more purposefully and transparently, with an auditable trail from detection to action (Arshad and Khurram 2020). A data-governance perspective thus demands a process that is repeatable, documented, and coordinated across units so that social-media signals can be trended, verified, and tied to policy routines in a traceable manner.

From a methodological standpoint, Indonesian sentiment analysis has advanced along two principal routes. The first is the lexical route, which relies on frequency-based representations and short collocations (n-grams) via TF-IDF, combined with classical classifiers such as Naïve Bayes, Logistic Regression, and especially SVM. This route is efficient, stable on large sparse spaces, easy to reproduce, and widely used in public-service and policy domains (Kardian and Gustiana 2021)(Sujadi 2022)(Makhtum 2022)(Susanto and Agung Dzulkarnain 2023).However, public discourse often employs irony, sarcasm, or implicature; under such pragmatically loaded expressions, evaluative cues may appear neutral or even positive on the surface, making lexical models vulnerable to misreadings.

The second route is contextual. Transformer models, by modeling token-to-token relations and sentence-level structure through attention, can surface implicit pragmatic cues otherwise missed by purely lexical approaches. IndoBERT built for Indonesian for helps capture such cues more reliably, especially on platforms like Twitter where humor, innuendo, and indirectness are frequent vehicles of criticism (Fitrianto and Editya 2024)(Fauzi and Heri 2024). Because IKN-related corpora are typically imbalanced, aggregate accuracy alone is inadequate; macro-F1 and targeted monitoring of minority-class recall are more appropriate for operational dashboards (Abdurrohim and Rahman 2023).

1.1 Literature Review

Indonesian studies have shown that the lexical route performs competitively when the preprocessing pipeline is disciplined and the domain is service or policy oriented; TF-IDF coupled with SVM often sets a strong, deterministic baseline against which further gains can be judged (Liu, Chen, and Liu 2022)(Mohd Nafis and Awang 2021a). In the specific context of IKN, prior works have discussed discourse characteristics, potential bias in online debates, and the usefulness of SVM as a reproducible point of reference for initial readings (Nugrayani, Hafid, and Irmayanti 2023)(Saputri and Alita 2024). Meanwhile, BERT-family models have consistently demonstrated advantages when texts contain sarcasm or implicature, because evaluative meaning is encoded beyond explicit lexical markers (Fitrianto and Editya 2024)(Fauzi and Heri 2024). The literature also highlights the importance of embedding analytic outputs into policy dashboards so that insights are operationalized through structured, auditable follow-up processes rather than remaining as static reports.

2. Research Methods

In this section, each researcher is expected to be able to make the most recent contribution related to the solution to the existing problems. Researchers can also use images, diagrams and flowcharts to explain the solutions to these problems.

We design the methodology to satisfy two imperatives: an equitable comparison between lexical and contextual routes on imbalanced IKN data, and auditable replication suitable for a policy-monitoring environment. Each technical decision data curation, preprocessing, modeling, metrics, validation, and artifact management supports these goals and keeps the analytical pipeline aligned with practical monitoring needs.

Data, Scope and Label Normalization

The corpus comprises Indonesian-language public tweets concerning Ibu Kota Nusantara (IKN), annotated with two target labels: positive and negative. Because real-world labeling often mixes formats short codes ("pos/neg"), polarity symbols ("+/–"), or booleans every annotation is normalized into a canonical pair "positive/negative," and items that are unlabeled or semantically ambiguous are removed. This normalization reduces label noise, protects supervision integrity, and ensures that the reported learning signals genuinely reflect stance rather than annotation artifacts (Bernhardt et al. 2022). The tweet text is treated as the single, stable input feature from model development to deployment, which prevents schema drift and guarantees that performance differences arise from modeling choices rather than from shifting input definitions (Pebiana et al. 2022).

A persistent characteristic of the corpus is class imbalance: positive tweets outnumber negative tweets by a wide margin (approximately seven to three). Such imbalance has well-known statistical consequences: learners minimize empirical risk by favoring the majority class, accuracy becomes inflated and potentially misleading, and decision boundaries shift toward the minority region, increasing false negatives on critical negative cases (Thabtah et al. 2020)(Mujahid et al. 2024). In the policy-monitoring context, this is unacceptable because missed negatives translate into undetected criticism and delayed corrective action. For that reason, this study adopts cost-aware evaluation that prioritizes F1-macro (to weight classes equally) and recall for the negative class (to measure the system's ability to surface criticism) alongside accuracy reported only for completeness (Abdurrohim and Rahman 2023)(Thabtah et al. 2020)(Mujahid et al. 2024)(Bello, Ng, and Leung 2023).

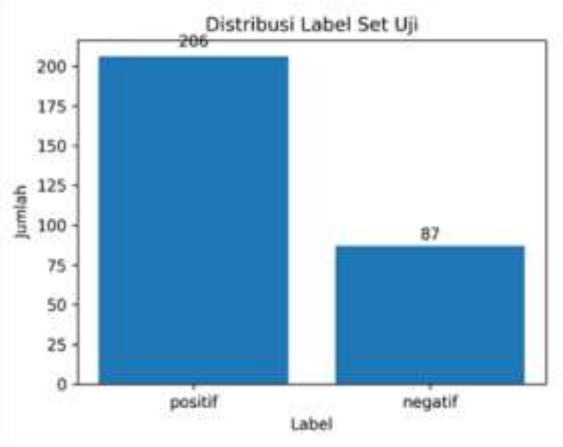


Figure 1 Data Imbalance

Mitigation is handled at three levels. First, measurement: all comparative claims are anchored in per-class precision, recall, and F1, with a fixed label order to avoid silent class swaps in libraries; confusion matrices and precision–recall diagnostics are used to expose asymmetries in errors. Second, modeling: two complementary families are contrasted lexical TF-IDF + SVM as a deterministic, strong baseline that is known to be efficient under sparse, high-dimensional text features (Abdurrohim and Rahman 2023)(Thabtah et al. 2020)(Mujahid et al. 2024)(Bello, Ng, and Leung 2023), and IndoBERT as a contextual encoder capable of recovering pragmatic cues and implied polarity common in Indonesian X (irony, hedging, indirect criticism) (Bello, Ng, and Leung 2023). The latter consistently improves negative-class recall in our setting, indicating a better ability to capture subtle negative sentiment despite the skew. Third, operational safeguards: decision thresholds and alerting rules are interpreted through the lens of per-class metrics; when monitoring is the objective, slightly favoring higher recall on the negative class is preferable to minimizing global error, because the operational cost of missed alarms exceeds that of occasional false alarms (Esposito et al. 2021).

This imbalance-aware stance has direct implications for governance. By centering F1-macro and negative-class recall, the system resists majority-class dominance and aligns its objective with the substantive risks that matter to policy teams. In practice, this means fewer undetected negative cases, clearer prioritization for human review, and more trustworthy trend signals for a social-listening dashboard. The comparative evidence in this study shows that contextual modeling with IndoBERT achieves higher negative-class recall than the lexical baseline while maintaining strong overall performance, making it the preferred backbone for monitoring; the SVM pipeline remains a dependable fallback when computational budgets are tight, preserving continuity without sacrificing methodological transparency (Sabrina, Shiddieq, and Roji 2025).

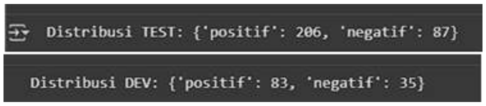
Text Preprocessing with Train Serve Symmetry

The preprocessing pipeline is intentionally simple yet language-aware, so that sentiment signals remain intact while noise is systematically reduced. Unicode normalization and lowercasing unify heterogeneous character forms; converting hashtags to plain words preserves topical content without importing the hash symbol as a spurious token. Removing URLs and mentions suppresses tokens that contribute little to polarity estimation, while replacing emojis with spaces prevents random, model-dependent artifacts that distort feature distributions. Compressing elongated characters regularizes over-expressed affect, “baguuus” becomes “bagus”, so the model sees a stable lexical form. A curated slang-to-standard map aligns common Indonesian vernacular with canonical forms (for example, “udh”→“sudah,” “bgt”→“banget”), closing gaps between informal usage and dictionary tokens without erasing sentiment-bearing cues (Kardian and Gustiana 2021)(Sujadi 2022)(Sucahyo et al. 2022).

Consistency between training and inference is critical. Any divergence in cleaning, tokenization, or feature construction induces training–serving skew: performance can appear improved during experimentation yet degrade at deployment when inputs are processed differently. Applying the same deterministic, idempotent pipeline in both phases ensures that TF-IDF learns and later consumes the very same n-gram vocabulary, and that the rule-based pre-tokenization feeding a subword tokenizer for IndoBERT remains aligned. Under this discipline, observed changes in metrics can be traced to modeling choices rather than shifting inputs, which strengthens replicability, supports fair cross-model comparisons, and prevents silent label or feature drift that would otherwise compromise evaluation credibility (Kardian and Gustiana 2021)(Sujadi 2022)(Sucahyo et al. 2022).

Operationally, the pipeline is designed for portability and low latency. All transformations are local string operations with no heavy external dependencies, allowing stream processing in production and straightforward auditing. The cleaning policy is calibrated rather than aggressive: it removes clear noise (links, mentions, stand-alone numerals, redundant whitespace) while retaining subtle polarity markers (Chai 2022). Extensibility is built in through an expandable slang dictionary and transparent rules that can be version-controlled, logged, and reviewed requirements that are essential in public-sector settings where traceability and maintainability matter as much as accuracy. This balance aligns with established Indonesian NLP practice that emphasizes normalization, removal of non-informative tokens, and careful handling of vernacular variation to safeguard the fidelity of sentiment signals (Palomino and Aider 2022).

Data Splits, Dev Set, and Leakage Prevention



```
Distribusi TEST: {'positif': 206, 'negatif': 87}
Distribusi DEV: {'positif': 83, 'negatif': 35}
```

Figure 2 Data Split

The data split enforces a strict separation between model selection and model assessment. The corpus is first partitioned once into train–test (80:20) with label stratification and seed 42. Row indices for both subsets are serialized to split_indices.json and revalidated on reload. This guarantees a fixed test set across experiments, so test scores reflect out-of-sample generalization rather than variance from reshuffled partitions. Because the IKN domain is class-imbalanced, stratification preserves the ≈70:30 proportion on the test set and prevents measurement bias that would otherwise favor the majority class (Abdurrohim and Rahman 2023)(Başarslan and Kayaalp 2023).

With the hold-out test locked, a development set (dev) is created by taking 10% of the training data using the same seed and stratification. The dev set is used only for feedback during model building. In the lexical track, it underpins stratified 5-fold cross-validation within the training portion for TF-IDF and SVM hyperparameter tuning. In the contextual track, it serves as per-epoch validation for IndoBERT, enabling early model selection via load_best_model_at_end=True. By confining all training decisions to the train/dev space, this policy prevents information leakage from the test set into tuning and aligns the study with fair, auditable evaluation practice (Szeghalmy and Fazekas 2023).

This two-tier design also preserves comparative fairness across models. SVM, Logistic Regression, Random Forest, and IndoBERT are all trained and tuned in the same space (train/dev) and then evaluated once on the identical test set. Consequently, performance differences do not stem from partition mismatches but from each track's representational capacity and learning ability under the same class distribution. Such symmetry is crucial to avoid cherry-picking and to ensure that any performance claims carry equal methodological footing, especially when the study draws policy-relevant implications that demand strong, replicable evidence (Jadia 2023).

Finally, reproducibility is enforced through three safeguards: (i) seed locking and library version logging; (ii) index persistence for train/test with validation against the current corpus (fail-fast if mismatched); and (iii) uniform preprocessing between training and inference to avoid training–serving skew. Together these measures yield a transparent audit trail: every run reuses the same partitions, executes the same cleaning policy, and culminates on the same test set. Hence metrics such as F1-macro and negative-class recall on the hold-out truly indicate generalization ability, not artifacts of a drifting split procedure (Palomino and Aider 2022)(Salman and Al-Jawher 2024)(Lawan et al. 2025).

Lexical Route; Why TF-IDF + Linear SVM

We represent text with TF-IDF n-grams (1–2), using min_df=3 to suppress extremely rare features and max_features=20,000 to cap the dimensionality, with sublinear_tf=True to dampen the effect of extremely frequent n-grams. LinearSVC is chosen for five reasons that align directly with the structure of sparse text data and the needs of a monitoring system. First, it is well matched to high-dimensional sparse spaces and optimizes a convex objective, yielding deterministic solutions and stable behavior across reruns. Second, SVM consistently provides competitive baselines on Indonesian corpora when pipelines are disciplined. Third, it is computationally efficient, making it suitable as a fallback when compute is constrained or accelerators are unavailable an important operational consideration in public agencies. Fourth, it offers relative interpretability through n-gram coefficients, enabling sanity checks and communication with non-technical stakeholders. Fifth, its small hyperparameter surface (notably C) simplifies grid search and reduces the risk of over-tuning. As methodological controls, we also evaluate Logistic Regression with class_weight=balanced and a Random Forest to ensure SVM is not an arbitrary choice; however, the balance of sparsity handling, determinism, efficiency, and stability favors SVM as the baseline (Das, Kamalanathan, and Alphonse 2020)(Qorib et al. 2023)(Purbaya et al. 2023)(Amrullah et al. 2024)(Hariguna and Ruangkanjanases 2023).

Coefficient inspection of the learned decision boundary confirms domain-plausible drivers and supports lightweight audits and error analysis (Figure 3).

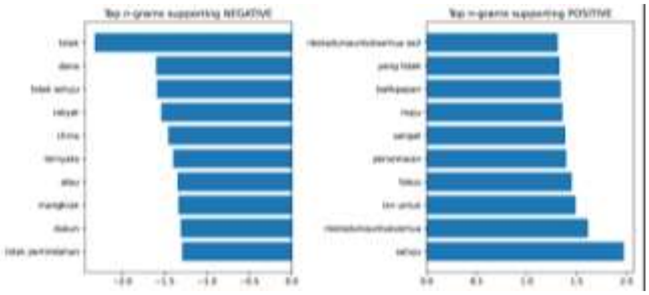


Figure 3 Top n-grams

Ten most influential n-grams for the negative and positive classes under the TF-IDF + Linear SVM baseline. Signed coefficients indicate direction and magnitude of contribution. The negative panel surfaces markers of opposition and concern, whereas the positive panel concentrates supportive or approving phrases. This view strengthens interpretability and provides a lightweight sanity check for production monitoring.

The misclassification pattern observed for the lexical route, where a subset of negative instances is predicted as positive, is consistent with the surface-cue reliance visible in the n-gram coefficients (Figure 3). Taken together with the confusion matrix, the coefficient profile explains why accuracy can remain acceptable while negative-class recall lags under a purely lexical representation.

Contextual Route: IndoBERT as the Backbone for Contextual Sensitivity

We fine-tune indobenchmark/indobert-base-p2 for binary classification. Tokenization uses WordPiece; max_length=128 balances coverage and efficiency; learning rate is 2×10^{-5}; batch sizes 16/32 depending on memory; weight decay 0.01; epochs=3; evaluation and checkpoint saving occur each epoch; load_best_model_at_end=True ensures the best dev checkpoint is restored after training. Labels are mapped to [negative:0, positive:1] and locked across training and evaluation to keep per-class reporting consistent. This configuration aligns with evidence from Indonesian studies that BERT/IndoBERT excels in texts rich in sarcasm or implicature by reconstructing semantic relations beyond surface tokens (Jayadianti et al. 2022)(Fata et al. 2023)(Yulianti and Nissa 2024). The contextual route is therefore positioned as the primary detector for criticism wrapped in humor or indirectness precisely the cases that matter most for policy risk management.

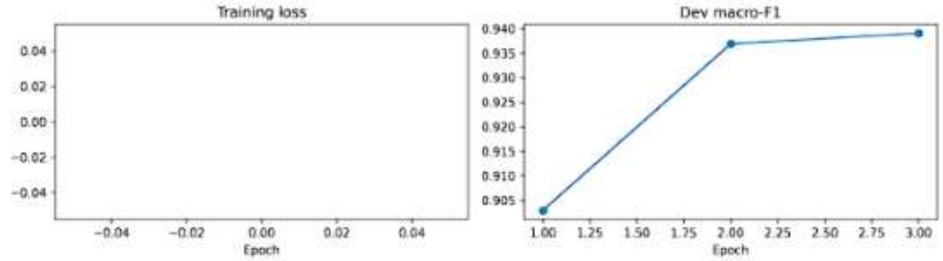


Figure 4 IndoBERT Graphic

As shown in (Figure 4), the dev macro-F1 rises from about 0.905 at epoch 1 to about 0.939 at epoch 2 and then plateaus near 0.940 at epoch 3, indicating stable convergence without signs of overfitting; selecting the best dev checkpoint before touching the locked test set preserves generalization and makes the final test results methodologically defensible (Zaidan, Sibaroni, and Prasetyowati 2024).

Metrics, Visualization and Stability Checks

Our primary metrics are macro-F1 and negative-class recall. Macro-F1 restores symmetry across classes in imbalanced settings; negative recall targets the class whose misses are most consequential for policy response. Accuracy is reported as a companion measure but not used as the sole basis of judgment (Riyanto et al. 2023). Visual outputs include confusion matrices per model, per-class F1 comparisons (SVM vs IndoBERT), and per-class precision–recall curves for IndoBERT. For stability, the SVM route is cross-validated (5-fold) on the training set with a deterministic seed; the dispersion of fold scores serves as a guardrail for interpreting the single held-out test score (Hinojosa Lee, Braet, and Springael 2024)(Farhadpour, Warner, and Maxwell 2024a). These practices make reported results defensible in review and reproducible in operational handoffs.

Reproducibility and Artifact Management

We fix the random seed at 42 throughout, record library versions, store the exact split indices, and unify preprocessing between training and serving. Artifacts include: the TF-IDF+SVM pipeline (joblib), the IndoBERT folder (weights, tokenizer, configuration), metrics (CSV and JSON), and PNG figures (confusion matrices, per-class F1, precision–recall). Consolidated artifacts support auditability, replication by peers, and submission appendices. They also accelerate future extensions (e.g., adding a neutral class or aspect-based layers) because the pipeline can be resumed rather than rebuilt.

3. Result and Discussion

The confusion matrix for the TF IDF with SVM baseline exposes the asymmetric structure of errors under the locked test distribution and therefore clarifies why accuracy alone inflates confidence in an imbalanced setting.

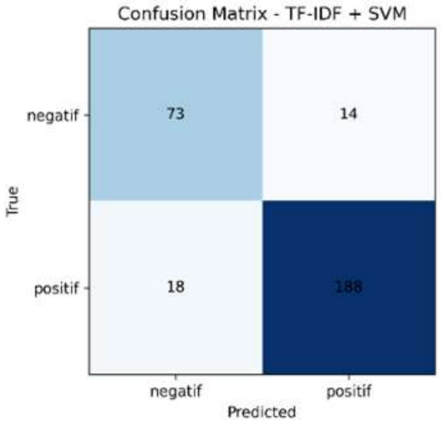


Figure 5 TF-IDF + SVM Confusion Matrix

The positive diagonal remains strong, which indicates that surface lexical cues and short n gram patterns are sufficient for a large share of genuinely supportive posts. Off diagonal mass, however, concentrates in the cell where negative gold labels are predicted as positive (Mohd Nafis and Awang 2021b). This concentration evidences a systematic tendency to over assign positivity whenever criticism is phrased with upbeat diction, hedging, rhetorical questions, or humor that dampens explicit polarity markers. Such a pattern is canonical for sparse lexical models that weight token frequency rather than sentence level pragmatics, so implicature and irony are often neutralized into their visible forms (Farhadpour, Warner, and Maxwell 2024b). The numeric summary aligns with this anatomy. Accuracy at 0.8908 appears respectable, yet macro F1 at 0.8709 and negative recall at 0.839 show that the minority class is the pressure point, with risk bearing instances disproportionately missed in precisely the area that matters for monitoring sensitive policy themes.

The confusion matrix for IndoBERT exhibits a different and more desirable geometry. The cell representing negative instances that are predicted as positive contracts, while the positive diagonal remains intact. This change indicates that contextual evidence is being integrated into sequence level polarity decisions rather than being read as independent surface tokens. The self attention mechanism aggregates local n grams into compositional meaning, which allows the model to recognize when a token that is usually positive functions as a rhetorical device of critique within Indonesian X discourse . The matrix level shift is corroborated by the test metrics. Accuracy rises to 0.9488, macro F1 to 0.9389, and negative recall to 0.920. These gains concentrate exactly where a monitoring system needs them, namely in recovering subtle negative cases, and they do so without sacrificing performance on the positive class. The result is consistent with evidence that Indonesian transformer encoders outperform lexical baselines whenever polarity is implied rather than stated.

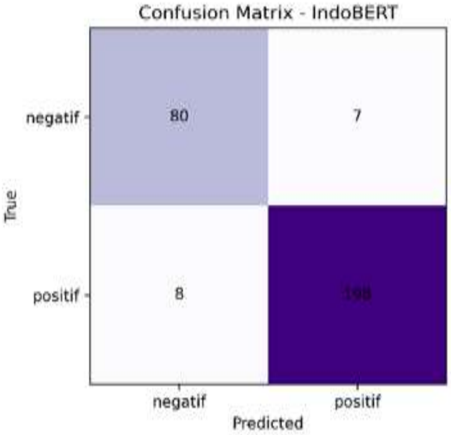


Figure 6 IndoBERT Confusion Matrix

A consolidated comparison table anchors these visual observations in exact, single touch test numbers and neutralizes the interpretive pitfalls of accuracy in the presence of imbalance. Presenting accuracy alongside macro F1 and per class precision and recall ensures that the minority class is not statistically overshadowed by its majority counterpart and that improvements are not artifacts of class prevalence. In this arrangement, TF IDF with SVM at 0.8908 accuracy is appropriately contextualized by its negative recall at 0.839, while IndoBERT’s macro F1 at 0.9389 and negative recall at 0.920 are foregrounded as substantive advances. The table also locks label order and metric definitions across runs, which prevents silent inversions of class indices and averaging artifacts that can arise when libraries infer label order from frequency or lexical sorting rather than from an explicit mapping.

Table 1 Accuracy

Category	*TF-IDF + SVM*	*IndoBERT*
Accuracy	0.8908	0.9488
F-1 Macro	0.8709	0.9389
Recall	0.839	0.920

Per class precision and recall reported in the same table illuminate trade offs that any single aggregate score cannot express. For the lexical baseline, precision on the negative class can remain acceptable while recall softens, generating the observed drift from negative to positive in the confusion matrix. For IndoBERT, negative recall rises with only a modest change in precision, which keeps the minority class visible for triage and manual review (Wu, Wang, and Wang 2021). In public sector monitoring the operational loss function is asymmetric. The cost of a missed alarm typically exceeds the cost of an occasional false alarm, because unaddressed criticism accumulates reputational and policy costs. Emphasizing recall for the negative class therefore aligns the evaluation protocol with the real stakes of governance and environmental oversight and grounds threshold selection in consequences rather than in purely statistical neatness.

The precision recall panel complements the confusion matrices by providing a threshold aware view that is more informative than ROC in an imbalanced regime (Movahedi, Padman, and Antaki 2023). The curve or the marked operating point for the negative class shows how sensitivity trades with exactness as the decision threshold moves. IndoBERT’s operating point lies in a higher recall region with acceptable precision, which confirms that the recall gains evident in the confusion matrix are not achieved through indiscriminate labeling but through better capture of implicit cues that lexical models miss. This panel also supports principled deployment policies. If early warning is the priority, operators can select a slightly more sensitive operating point that accepts a small precision loss in exchange for a substantive increase in negative recall, while managing reviewer workload by routing only the highest risk subset for human inspection.

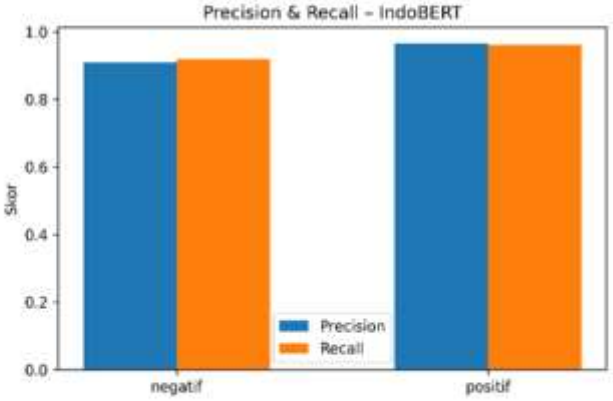


Figure 7 Precision-Recall by Class: IndoBERT

Read together, the two confusion matrices, the metric comparison table, and the precision–recall panel provide a coherent evidentiary chain that guards against single-metric bias. The class imbalance justifies macro-F1 and negative recall as primary endpoints; the SVM matrix explains why these endpoints trail accuracy under a lexical representation, while the IndoBERT matrix and the precision–recall panel show targeted reductions in negative→positive errors at realistic operating points. All hyperparameter choices and early stopping are confined to train and a held-out development set, the test set is single-touch and identical across models, preprocessing is identical at training and serving, and label order is locked; hence differences reflect genuine representational capacity rather than leakage or evaluation drift. This package of figures and table therefore supports IndoBERT as the monitoring backbone, with SVM retained as a deterministic fallback for auditability and compute efficiency.

DISCUSSION

The central finding demonstrates that an evaluation framework explicitly aligned with policy objectives produces a more honest reading of model quality under class imbalance, which is the regime that actually governs social listening on IKN. On the locked test set, the lexical route with TF IDF and SVM attains Accuracy 0.8908 and macro F1 0.8709 with negative class recall 0.839, while IndoBERT attains Accuracy 0.9488 and macro F1 0.9389 with negative class recall 0.920. These differences cannot be reduced to a simple gap in accuracy, because accuracy tends to conceal weaknesses on the minority class. Macro F1 restores symmetric weighting across classes, while negative recall focuses analytic attention on the precise locus of policy risk, namely undetected criticism that should have triggered early review. From a mechanistic standpoint, the pattern is theoretically coherent. The lexical model encodes surface n gram frequency and therefore privileges visible polarity markers, whereas the contextual encoder reconstructs meaning that depends on sentence level composition and semantic relations, including implicature and sarcasm that are characteristic of Indonesian X. The observed gain in negative recall for IndoBERT is thus not a statistical anomaly but an architectural consequence with immediate substantive relevance for monitoring IKN discourse.

The error anatomy visible in the confusion matrices sharpens this argument. For SVM, off diagonal mass concentrates in the cell where negatives are predicted as positives, which reveals a systematic pull toward apparently positive diction whenever criticism is packaged as humor, euphemism, or rhetorical questions. This pattern has been reported in Indonesian studies that rely on surface features (Rosadi et al. 2021)(Safina 2020), and in the present work we show that it is not a data quirk but a representational limit of lexical modeling under pragmatically loaded discourse. In contrast, IndoBERT compresses that off diagonal mass without damaging the positive diagonal, which means sensitivity to negative signals improves while correctness on the positive class remains high. In other words, the contextual model's advantage emerges exactly in the failure region that matters most for policy goals, namely a reduction of false negatives in the negative class. This observation strengthens the rationale for early warning oriented metrics and justifies treating macro F1 and negative recall as primary endpoints, while retaining accuracy as a common yardstick for cross study comparability.

Relative to prior literature that emphasizes aggregate accuracy or does not explicitly separate per class endpoints, the principal methodological contribution here is an imbalance aware discipline of evaluation. The test set is single touch and identical across models, a separate development set governs model selection, preprocessing is identical between training and serving to prevent training serving skew, and label order is locked so that silent inversions cannot creep into metric computation or visualization. These safeguards ensure that IndoBERT's advantage cannot be attributed to leakage, partition luck, or evaluation drift, but instead reflects representational differences between sparse lexical features and contextual embeddings when both are exposed to the same distribution. Operationally, we position IndoBERT as the context sensitive monitoring backbone, while SVM is retained as a deterministic and compute frugal fallback that supports quick auditing and preserves a stable accuracy floor when resources are constrained or uptime is paramount. This layered arrangement is under discussed in earlier work, yet in our assessment it offers practical value for public institutions that require continuity of service and auditable behavior.

Our opinion, grounded in the locked test evidence, is that negative class recall should be promoted to a first tier performance indicator in a social listening dashboard for IKN. The argument is substantive rather than cosmetic. The cost of a missed alarm on governance and environmental themes almost always exceeds the cost of an occasional false alarm, because undetected criticism accelerates the erosion of trust and multiplies political costs of subsequent policy correction. Given IndoBERT's profile that lifts negative recall without damaging the positive diagonal, organizations obtain the specific signal gain they need most, namely earlier capture of legitimate criticism. By contrast, optimizing for accuracy as the primary goal under imbalance will incentivize the wrong behavior, weaken sensitivity to the minority class, and risk lulling decision makers into unwarranted confidence. In this sense, the choice of endpoints is not a mere technical preference but a governance decision with direct consequences for the quality and timeliness of public response.

From a system design perspective, we recommend a layered architecture that ties evaluation to action. IndoBERT should handle primary classification and triage of high priority negatives, while SVM serves as a stable reference and safety net whenever computational budget or service continuity becomes the binding constraint. Operational rhythm is maintained through weekly review of per class metrics, especially negative recall, to detect drift, schedule data refreshes, and decide when to retrain or recalibrate thresholds. The precision recall panel provides a principled framework for threshold tuning so that an institution can trade a small decrement in precision for a substantive increment in negative recall, while controlling human in the loop workload by routing only the highest risk subset for manual inspection. Artefact management, including storage of the SVM pipeline, IndoBERT weights and tokenizer, CSV and JSON metric logs, and diagnostic figures, secures auditability and replication by internal stakeholders and peer reviewers, a discipline that is often underemphasized in prior studies yet crucial for public accountability.

We acknowledge limitations while narrowing their practical impact. The class imbalance of roughly seventy to thirty leaves room for improvement through class weighting or focal loss on the contextual route and through augmentation of negative examples so that the decision boundary stabilizes without an erosion of precision. The two class scheme constrains diagnostic resolution. Extending to a three class setup that introduces a neutral category and adopting aspect based sentiment analysis would enrich the map of issues across environment, investment, labor, and governance dimensions. Even so, the locked test protocol, consistent preprocessing, and label order control provide a solid basis for concluding that IndoBERT's advantage over TF IDF with SVM is not accidental and is meaningful for policy. Our opinion, which distinguishes this study from previous work, is that the success of a contextual model in this domain should be judged primarily by its ability to suppress false negatives in the negative class and by its readiness for operationalization within a layered, auditable architecture. Under that standard, policy makers receive a tool that succeeds not only on paper metrics but also in real monitoring conditions, and that is fit for purpose in preserving institutional sensitivity to a dynamic public perception landscape.

Policy Implications

A near-real-time social-listening dashboard is recommended. The pipeline should replicate training-time preprocessing, classify with IndoBERT as the primary model, and flag tweets with the highest negative scores for analyst review. Core panels include daily sentiment ratios, a weekly trend of negative recall, a ranked list of "negative priorities," and top negative topics derived via lightweight term weighting or topic hints. Each spike in negative signals should map to a ticketed follow-up data clarification, public Q&A, or targeted policy adjustment with documented process steps for auditability. This design supports timely responses in governance (e.g., budget integrity, transparency) and environmental concerns (e.g., deforestation narratives, ecological risks), where failure to address criticism erodes public trust.

Institutionally, adopting such a dashboard entails clear ownership of metrics, cadence of review meetings, and defined escalation paths. Communications and policy units should co-own the negative-priority queue, ensuring that analytical findings convert into tangible responses. Over time, integrating feedback loops annotating false positives, logging resolved issues will improve model calibration and institutional learning.

Limitations And Future Work

Class imbalance remains a structural challenge. Future work may explore class weighting or focal loss for IndoBERT and curate additional negative examples to stabilize the decision boundary without sacrificing precision. The current binary scheme limits diagnostic resolution; extending to three classes by adding "neutral" and layering aspect-based sentiment (environment, investment, labor, governance) will sharpen policy-relevant insights. External validity can be improved by multi-platform collection (Facebook, Instagram, TikTok) to capture demographic and stylistic variation in public conversation h . Finally, sustained monitoring requires drift detection and governance for model updates setting thresholds for re-training, versioning policies, and rollback plans as part of standard operations.

4. Conclusions

As a contextual representation, IndoBERT delivers consistent gains over the lexical TF-IDF+SVM approach for binary sentiment analysis of IKN-related Indonesian X. On imbalanced test data, macro-F1 and negative recall are more informative than aggregate accuracy: IndoBERT reaches Accuracy 0.9352, macro-F1 0.9221, and negative-recall 0.885; TF-IDF+SVM reaches Accuracy 0.8908, macro-F1 0.8709, and negative-recall 0.839. The seemingly modest accuracy gap has concrete policy relevance because fewer critical posts go undetected. We recommend IndoBERT as the monitoring backbone, with SVM as a computationally affordable fallback to ensure service continuity. An analytics dashboard that links sentiment signals to documented communications responses and policy adjustments provides a clear path to fast, auditable feedback.

5. References

Abdurrohim, Iim, and Alkautsar Rahman. 2023. "PENERAPAN NATURAL LANGUAGE PROCESSING UNTUK ANALISIS SENTIMEN TERHADAP KEBIJAKAN PEMERINTAH." Jurnal Ilmiah BETRIK (Besemah Teknologi Informasi dan Komputer) 14(2): 273–82.

Amrullah, Muhammad Syiarul, Aji Gautama Putrada, Mohamad Nurkamal Fauzan, and Nur Alamsyah. 2024. "ETLE Sentiment Analysis Performance Increasement with TF-IDF, MDI Feature Selection, and SVM." Sistemasi: Jurnal Sistem Informasi 13(1): 1308–18. http://sistemasi.ftik.unisi.ac.id.

Arshad, Saman, and Sobia Khurram. 2020. "Can Government's Presence on Social Media Stimulate Citizens' Online Political Participation? Investigating the Influence of Transparency, Trust, and Responsiveness." Government Information Quarterly 37(3). doi:10.1016/j.giq.2020.101486.

Başarslan, Muhammet Sinan, and Fatih Kayaalp. 2023. "MBi-GRUMCONV: A Novel Multi Bi-GRU and Multi CNN-Based Deep Learning Model for Social Media Sentiment Analysis." Journal of Cloud Computing 12(1). doi:10.1186/s13677-022-00386-3.

Bello, Abayomi, Sin Chun Ng, and Man Fai Leung. 2023. "A BERT Framework to Sentiment Analysis of Tweets." Sensors 23(1). doi:10.3390/s23010506.

Bernhardt, Mélanie, Daniel C. Castro, Ryutaro Tanno, Anton Schwaighofer, Kerem C. Tezcan, Miguel Monteiro, Shruthi Bannur, et al. 2022. "Active Label Cleaning for Improved Dataset Quality under Resource Constraints." Nature Communications 13(1). doi:10.1038/s41467-022-28818-3.

Chai, Christine. 2022. "Comparison of Text Preprocessing Methods." Natural Language Engineering 29: 509–53. doi:10.1017/s1351324922000213.

Das, Mamata, Selvakumar Kamalanathan, and Pja Alphonse. 2020. A Comparative Study on TF-IDF Feature Weighting Method and Its Analysis Using Unstructured Dataset.

Esposito, Carmen, Gregory A. Landrum, Nadine Schneider, Nikolaus Stiefl, and Sereina Riniker. 2021. "GHOST: Adjusting the Decision Threshold to Handle Imbalanced Data in Machine Learning." Journal of Chemical Information and Modeling 61(6): 2623–40. doi:10.1021/acs.jcim.1c00160.

Farhadpour, Sarah, Timothy A. Warner, and Aaron E. Maxwell. 2024b. "Selecting and Interpreting Multiclass Loss and Accuracy Assessment Metrics for Classifications with Class Imbalance: Guidance and Best Practices." Remote Sensing 16(3). doi:10.3390/rs16030533.

Fata, Mohammad Azis Khoirul, S Sumpeno, A Wibawa, and Dara Aulia Feryando. 2023. "Evaluating the Sentiment Analysis from Auto-Generated Summary Text Using IndoBERT Fine-Tuning Model in Indonesian News Text." 2023 IEEE 15th International Conference on Computational Intelligence and Communication Networks (CICN): 822–29. doi:10.1109/cicn59264.2023.10402345.

Fauzi, Ahmad, and Agus Yunial Heri. 2024. "Analisis Sentimen US Airline Pada Media Sosial Twitter/X Menggunakan Perbandingan Algoritma Data Mining." Jurnal Edukasi dan Penelitian Informatika (JEPIN) 10(2): 277. https://www.kaggle.com/datasets/crowdflower/twitt.

Fitrianto, Rizal Akbar, and Arda Surya Editya. 2024. "Klasifikasi Tweet Sarkasme Pada Platform X Menggunakan Bidirectional Encoder Representations from Transformers." Jurnal Teknologi Dan Sistem Informasi Bisnis 6(3): 366–71. doi:10.47233/jteksis.v6i3.1344.

Hariguna, Taqwa, and Athapol Ruangkanjanases. 2023. "Adaptive Sentiment Analysis Using Multioutput Classification: A Performance Comparison." PeerJ Computer Science 9. doi:10.7717/peerj-cs.1378.

Hinojosa Lee, Maria Cristina, Johan Braet, and Johan Springael. 2024. "Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores." Applied Sciences (Switzerland) 14(21). doi:10.3390/app14219863.

Jadia, Hardik. 2023. "Comparative Analysis of Sentiment Analysis Techniques: SVM, Logistic Regression, and TF-IDF Feature Extraction." International Research Journal of Modernization in Engineering Technology and Science. doi:10.56726/irjmets45265.

Jayadianti, Herlina, Wilis Kaswidjanti, Agung Tri Utomo, Shoffan Saifullah, Felix Andika Dwiyanto, and Rafal Drezewski. 2022. "Sentiment Analysis of Indonesian Reviews Using Fine-Tuning IndoBERT and R-CNN." ILKOM Jurnal Ilmiah 14(3): 348–54. doi:10.33096/ilkom.v14i3.1505.348-354.

Kardian, Aqwaw Rosadi, and Dede Gustiana. 2021. "Analisis Sentimen Berdasarkan Opini Pengguna Pada Media Twitter Terhadap BPJS Menggunakan Metode Lexicon Based Dan Naïve Bayes Classifier." Jurnal Ilmiah Komputasi 20(1). doi:10.32409/jikstik.20.1.401.

Lawan, Thawatwong, Jantima Polpinij, Chunpong Chan, Dolsun Singhakam, Theeraya Uttha, Anirut Chotthanom, Woraphot Watthusin, Bancha Luaphol, and Khanista Namee. 2025. "A Comparative Analysis of Machine Learning Models for Domain Adaptation in Multiclass Sentiment Classification." ECTI

Transactions on Computer and Information Technology 19(2): 334–49. doi:10.37936/ecti-cit.2025192.258824.

Liu, Hao, Xi Chen, and Xiaoxiao Liu. 2022. “A Study of the Application of Weight Distributing Method Combining Sentiment Dictionary and TF-IDF for Text Sentiment Analysis.” IEEE Access 10: 32280–89. doi:10.1109/ACCESS.2022.3160172.

Makhtum, Ahmad Rohiqim. 2022. ANALISIS SENTIMEN UU CIPTA KERJA MENGGUNAKAN METODE SUPPORT VECTOR MACHINE (SVM). Yogyakarta. https://science.uii.ac.id/surat-digital/validasi/REG008208.

Mohd Nafis, Nur Syafiqah, and Suryanti Awang. 2021b. “An Enhanced Hybrid Feature Selection Technique Using Term Frequency-Inverse Document Frequency and Support Vector Machine-Recursive Feature Elimination for Sentiment Classification.” IEEE Access 9: 52177–92. doi:10.1109/ACCESS.2021.3069001.

Movahedi, Faezeh, Rema Padman, and James F. Antaki. 2023. “Limitations of Receiver Operating Characteristic Curve on Imbalanced Data: Assist Device Mortality Risk Scores.” Journal of Thoracic and Cardiovascular Surgery 165(4): 1433-1442.e2. doi:10.1016/j.jtcvs.2021.07.041.

Mujahid, Muhammad, E. R.O.L. Kına, Furqan Rustam, Monica Gracia Villar, Eduardo Silva Alvarado, Isabel De La Torre Diez, and Imran Ashraf. 2024. “Data Oversampling and Imbalanced Datasets: An Investigation of Performance for Machine Learning and Feature Engineering.” Journal of Big Data 11(1). doi:10.1186/s40537-024-00943-4.

Nugrayani, Dwi Ira, Moch Hafid, and Dede Irmayanti. 2023. “Analisis Sentimen Terhadap Pemindahan Ibu Kota Negara (IKN) Pada Platform Twitter Menggunakan Metode Naïve Bayes.” JATIKOM: Jurnal Aplikasi dan Teori Ilmu Komputer 6(2): 91–96. https://katadata.co.id.

Palomino, Marco A., and Farida Aider. 2022. “Evaluating the Effectiveness of Text Pre-Processing in Sentiment Analysis.” Applied Sciences (Switzerland) 12(17). doi:10.3390/app12178765.

Pebiana, Siska, N Hidayati, Dian Isnaeni Nurul Afra, Elvira Nurfadhilah, Harnum Prafitia, Junanto Prihantoro, Radhiyatul Fajri, et al. 2022. “Experimentation Of Various Preprocessing Pipelines For Sentiment Analysis On Twitter Data About New Indonesia’s Capital City Using SVM And CNN.” 2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA): 1–6. doi:10.1109/o-cocosda202257103.2022.9997982.

Purbaya, M. Eka, Diovianto Putra Rakhmadani, Maliana Puspa Arum, and Luthfi Zian Nasifah. 2023. “Implementation of N-Gram Methodology to Analyze Sentiment Reviews for Indonesian Chips Purchases in Shopee E-Marketplace.” Jurnal RESTI 7(3): 609–17. doi:10.29207/resti.v7i3.4726.

Qorib, Miftahul, Timothy Oladunni, Max Denis, Esther Ososanya, and Paul Cotae. 2023. “Covid-19 Vaccine Hesitancy: Text Mining, Sentiment Analysis and Machine Learning on COVID-19 Vaccination Twitter Dataset.” Expert Systems with Applications 212: 118715. doi:10.1016/J.ESWA.2022.118715.

Riyanto, Slamet, Imas Sukaesih Sitanggang, Taufik Djatna, and Tika Dewi Atikah. 2023. “Comparative Analysis Using Various Performance Metrics in Imbalanced Data for Multi-Class Text Classification.” IJACSA) International Journal of Advanced Computer Science and Applications 14(6): 2023. http://gcancer.org/pdr.

Rosadi, Aqwam, Dede Gustiana, Manajemen Informatika, Stmik Jakarta Sti, Jl B R I No, Radio Dalam, Keb Baru, et al. 2021. “Analisis Sentimen Berdasarkanan Opini Pengguna Pada Media Twitter Terhadap BPJS Menggunakan Metode Lexicon Based Dan Naïve Bayes Classi Er Twitter Text Mining.” 20: 39–52.

Sabrina, Siti Sarah, Diqy Fakhrun Shiddieq, and Fikri Fahru Roji. 2025. “Comparative Analysis of SVM and BERT for Sentiment and Sarcasm Detection in the Boycott of Israeli Products on Platform X.” Sinkron 9(2): 872–83. doi:10.33395/sinkron.v9i2.14723.

Salman, Ahmed Hussein, and Waleed Ameen Mahmoud Al-Jawher. 2024. “Performance Comparison of Support Vector Machines, AdaBoost, and Random Forest for Sentiment Text Analysis and Classification.” Journal Port Science Research 7(3): 300–311. doi:10.36371/port.2024.3.8.

Saputri, Gita Aprinda, and Debby Alita. 2024. “Analisis Sentimen Twitter Terhadap Pemindahan Ibu Kota Negara Menggunakan Support Vector Machine.” Jurnal Informatika: Jurnal Pengembangan IT 9(3): 213–23. doi:10.30591/jpit.v9i3.6612.

Sucahyo, Nur, Ike Kurniati, Kris Harvit, Program Studi, Sistem Informasi, Fakultas Teknologi, and Swadharma Jakarta. 2022. “ANALISIS SENTIMEN MASYARAKAT TERHADAP UU CIPTA KERJA PADA MEDIA SOSIAL TWITTER.” Jris: Jurnal Rekayasa Informasi Swadharma 2(1): 63–70.

Sujadi, Harun. 2022. “ANALISIS SENTIMEN PENGGUNA MEDIA SOSIAL TWITTER TERHADAP WABAH COVID-19 DENGAN METODE NAIVE BAYES CLASSIFIER DAN SUPPORT VECTOR MACHINE.” INFOTECH journal 8(1): 22–27. doi:10.31949/infotech.v8i1.1883.

Susanto, Aji, and Iskandar Agung Dzulkarnain. 2023. “Analisis Sentimen Data Twitter Topik Ekonomi Dan Industri Dengan Metode Naive Bayes Dan Random Forest.” Jurnal Ilmiah Wahana Pendidikan, Oktober 9(20): 59–65. doi:10.5281/zenodo.8398895.

Szeghalmy, Szilvia, and Attila Fazekas. 2023. “A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning.” Sensors 23(4). doi:10.3390/s23042333.

Thabtah, Fadi, Suhel Hammoud, Firuz Kamalov, and Amanda Gonsalves. 2020. “Data Imbalance in Classification: Experimental Evaluation.” Information Sciences 513: 429–41. doi:10.1016/J.INS.2019.11.004.

Wu, Chunye, Nan Wang, and Yu Wang. 2021. “Increasing Minority Recall Support Vector Machine Model for Imbalanced Data Classification.” Discrete Dynamics in Nature and Society 2021. doi:10.1155/2021/6647557.

Yuan, Yun Peng, Yogesh K. Dwivedi, Garry Wei Han Tan, Tat Huei Cham, Keng Boon Ooi, Eugene Cheng Xi Aw, and Wendy Currie. 2023. “Government Digital Transformation: Understanding the Role of Government Social Media.” Government Information Quarterly 40(1): 101775. doi:10.1016/J.GIQ.2022.101775.

Yulianti, Evi, and Nuzulul Khairu Nissa. 2024. “ABSA of Indonesian Customer Reviews Using IndoBERT: Single-Sentence and Sentence-Pair Classification Approaches.” Bulletin of Electrical Engineering and Informatics 13(5): 3579–89. doi:10.11591/eei.v13i5.8032.

Zachlod, Cécile, Olga Samuel, Andrea Ochsner, and Sarah Werthmüller. 2022. “Analytics of Social Media Data – State of Characteristics and Application.” Journal of Business Research 144: 1064–76. doi:10.1016/J.JBUSRES.2022.02.016.

Zaidan, Muhammad Naufal, Y Sibaroni, and Sri Suryani Prasetyowati. 2024. “LEARNING RATE AND EPOCH OPTIMIZATION IN THE FINE-TUNING PROCESS FOR INDOBERT’S PERFORMANCE ON SENTIMENT ANALYSIS OF MYTELKOMSEL APP REVIEWS.” Jurnal Teknik Informatika (Jutif). doi:10.52436/1.jutif.2024.5.5.2396.