Vol. 8, No. 1, February 2026

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Sentiment Analysis of Indonesian Netizens toward Vasectomy on X Using the IndoBERT Model

Yelli Nur Alinda[1], Allsela Meiriza*[2], Dinna Yunika Hardiyanti[3]

[1,2]Information System, Faculty of Computer Science, Sriwijaya University, Palembang, Indonesia

[3]Accounting Computerization, Faculty of Computer Science, Sriwijaya University, Palembang, Indonesia

Received:
November 8, 2025
Revised:
January 15, 2026
Accepted:
January 30, 2026
Published:
February 13, 2026

Corresponding Author:
Author Name*:
Allsela Meiriza
Email*:
allsela@unsri.ac.id

DOI:
10.63158/journalisi.v8i1.1425

Abstract. Vasectomy-related conversations on X (Twitter) frequently generate polarized pro–contra debates that can shape public understanding of male contraception, yet evidence on Indonesian netizens' sentiment remains limited. This study maps and classifies sentiment toward vasectomy during April–June 2025 using a descriptive quantitative text-mining and NLP pipeline. After preprocessing (cleaning and deduplication), 9,817 posts were analyzed. Semi-supervised labeling was performed using the teacher model taufiqdp/indonesian-sentiment with confidence-based refinement, supported by a rule-based sarcasm_flag that identified 330 potentially sarcastic texts. A 20% manually verified GOLD subset (1,963 samples) served as ground truth, and IndoBERT (indolem/indobert-base-uncased) was fine-tuned with weighted cross-entropy and early stopping. Evaluation on the GOLD test set (n = 393) showed strong performance (accuracy = 0.8168; macro F1 = 0.8141), with most errors concentrated in short, ambiguous, or humor/sarcasm-leaning posts. Full-corpus predictions produced 3,957 negative, 3,520 positive, and 2,340 neutral texts, indicating a contested and polarized discourse with a slightly higher negative share. These findings support the need for evidence-based digital communication strategies to address misconceptions and stigma surrounding male contraception.

Keywords: Sentiment Analysis, IndoBERT, Semi-supervised Learning, Social Media Analysis, Vasectomy

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

1. INTRODUCTION

Indonesia's Family Planning Program positions male involvement as a key component of efforts to manage population growth and promote more equitable roles in family planning; however, evidence from practice indicates that men's participation particularly in the uptake of permanent contraception remains substantially lower than women's [1], [2]. Although vasectomy is clinically proven to be safe and effective and does not impair hormonal function or sexual performance, it is frequently viewed negatively because it is associated with reduced masculinity, perceived conflict with religious teachings, or equated with irreversible infertility [3], [4]. This tension between clinical evidence and social perceptions became more apparent in 2025, when a public figure's statement about vasectomy sparked widespread debate on X and rapidly made the topic viral, turning digital spaces into arenas for diverse expressions of emotion, support, rejection, humor, and sarcasm surrounding male contraception [5].

In this context, social media functions not only as a communication channel but also as a real-time mirror of public opinion dynamics that can amplify, normalize, or contest stigma and misinformation. Because these dynamics shape how vasectomy is interpreted and accepted, they need to be examined systematically so that policy framing and communication strategies particularly by institutions such as BKKBN can be aligned with evolving public discourse, emerging concerns, and the language actually used by netizens [6]. Without such alignment, program messaging risks missing the most salient barriers and narratives circulating in the public sphere.

Prior studies in Indonesia indicate that social media sentiment analysis can effectively capture public responses to health and policy-related issues [7], [8]. However, evidence focusing specifically on vasectomy discourse is still limited: existing Twitter/X-based studies are scarce, often rely on classical machine-learning methods, and are constrained by dataset scale and the linguistic variability of informal Indonesian text (e.g., slang, sarcasm, code-mixing, and non-standard spelling), which can weaken classification robustness [9]. Related research on other contraceptive topics (e.g., LARC) has reported a tendency toward negative sentiment, suggesting persistent skepticism that may also color male-contraception conversations online [10]. At the same time, pre-trained language models such as IndoBERT have shown strong performance for

Indonesian sentiment classification across diverse public-debate contexts [11], [12], yet IndoBERT-based deep learning that systematically maps Indonesian netizens' sentiment toward vasectomy on X—especially using a combined labeling approach to improve reliability—remains relatively underexplored.

Building on this gap, the present study uses public conversations on X during the period when the vasectomy issue gained heightened attention as the primary data source. The data are collected through crawling and subsequently processed using text mining and Natural Language Processing procedures so that initially unstructured posts can be analyzed systematically to identify sentiment tendencies and their distribution in the discourse [13]. To support reliable modeling under noisy social media conditions, the dataset is refined through tailored preprocessing and labeled using a semi-supervised scheme that combines model-based automatic labeling with manual verification on a GOLD subset to strengthen label quality and reduce error propagation [14]. The labeled corpus is then used to train and evaluate sentiment classification models—including an IndoBERT-based deep learning model—with performance assessed using relevant evaluation metrics for three-class (positive/negative/neutral) sentiment classification [15].

Through this design, the study aims to identify patterns of public sentiment toward vasectomy, assess IndoBERT's ability to classify sentiment in informal Indonesian texts, and extend prior work by broadening data coverage while strengthening the modeling and labeling approach [11], [12], [15]. Accordingly, this study contributes by systematically mapping Indonesian netizens' sentiment toward vasectomy discourse on X during a high-salience period [5], [13], improving label reliability through a semi-supervised labeling strategy that integrates teacher-model labeling with a manually verified GOLD subset [14], and providing empirical evidence to support more targeted public communication and educational strategies related to male contraception programs in Indonesia, in line with the evolving public discourse that institutions such as BKKBN need to address [6].

2. METHODS

This study adopts a descriptive quantitative approach, integrating text mining and Natural Language Processing (NLP) to classify public sentiment toward the vasectomy issue on X (Twitter) during the April–June 2025 period. A descriptive quantitative design was selected because the study aims to map sentiment distributions and discussion dynamics in a measurable manner, rather than to explore new constructs (exploratory) or to test causal relationships and hypotheses [16], [17].

Figure 1 presents the overall research methodology and summarizes six main blocks, namely Problem Identification, Data Collection, Preprocessing, Data Labeling, Modeling, and Evaluation. In this study, text mining mainly supports data collection and preprocessing, while NLP is applied in labeling and modeling through auto-labeling, tokenization, and IndoBERT fine-tuning, followed by evaluation using accuracy, macro F1-score, and confusion matrix. Presenting the workflow as a block-based pipeline is commonly used in sentiment analysis studies because it clearly delineates corpus construction, text quality improvement, and systematic model evaluation stages [18].

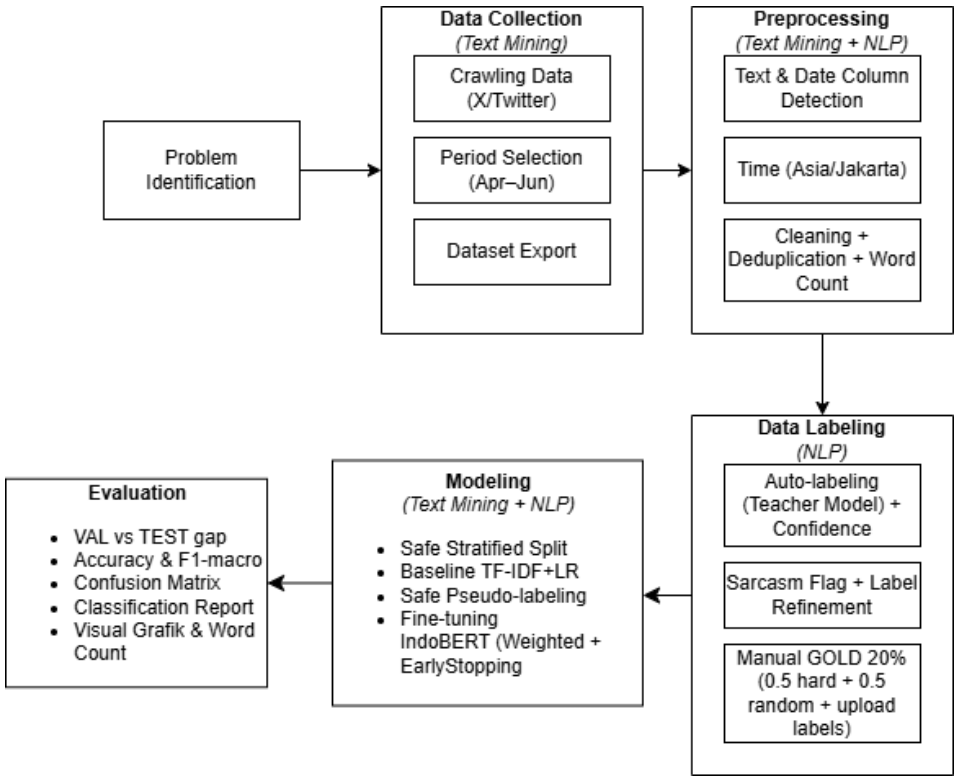

Figure 1. Research Methodology

2.1. Problem Identification

The vasectomy issue in Indonesia often generates public debate due to a tension between medical evidence emphasizing the procedure's safety and effectiveness and prevailing social perceptions shaped by stigma, masculinity narratives, and persistent misconceptions regarding infertility [19][20]. When the topic becomes salient and spreads rapidly on X, public conversations evolve quickly and exhibit diverse forms of expression, ranging from support and rejection to humor and sarcastic remarks, thereby requiring a computational approach to map sentiment tendencies in a measurable and consistent manner. Previous Twitter-based vasectomy sentiment studies also indicate that social media–driven sentiment mapping is relevant for describing public acceptance more objectively, while opportunities remain to improve both data scale and the adequacy of language representation [9].

Accordingly, this study aims to produce an updated sentiment mapping based on publicly available X data collected through crawling and analyzed using modern NLP methods, including the use of an Indonesian Transformer model (IndoBERT), which has been shown to capture informal and highly variable Indonesian language contexts effectively. In addition, the performance of three-class sentiment classification (positive, negative, neutral) is evaluated using commonly adopted sentiment analysis metrics to enable fair comparison with baseline approaches and prior studies, while also providing empirical evidence that may support the development of more targeted communication and educational strategies related to male contraception.

2.2. Data Collection

The data collection stage was conducted to obtain a raw text corpus for analysis. The dataset was collected by crawling public conversations on X using keywords relevant to the vasectomy issue. The crawling output was stored in tabular formats (.xlsx/.csv) containing the text, posting time, and additional metadata to support the analysis. The use of Twitter/X data to capture public opinion has been widely adopted because it reflects spontaneous public responses to policy-related issues.

In this study, crawling was performed using a third-party scraping tool executed in Google Colab (tweet-harvest via Node.js/npx), rather than the official Twitter/X API. The search query combined vasectomy-related keywords with X search operators, including

a language filter (lang:id) and date constraints (since–until), and the LATEST tab was used to prioritize newly emerging posts within the target period. To reduce duplicated content, the exported dataset was deduplicated using the unique tweet identifier (tweetId) before further processing.

To maintain a consistent discourse context, the dataset was filtered to the time window when the topic was newly emerging, namely from April 2025, which coincided with KDM's statement related to vasectomy, to June 2025, when the issue began to subside. Limiting the period to three months (one quarter) helps reduce the mixing of irrelevant issues, minimizes potential bias, and ensures that the analyzed data represent the viral phase of the vasectomy topic [21]. The final dataset was exported into spreadsheet formats (.xlsx/.csv) to facilitate auditing and replication. The export process was carried out within Google Colab to enable seamless integration with an automated Python-based pipeline. Ethical considerations were applied by restricting collection to publicly accessible posts and by avoiding any access to private or restricted content. To minimize privacy risks, identifiers such as usernames, profile links, and tweet IDs were not reported in the manuscript; findings are presented primarily in aggregated form, and any illustrative excerpts (if used) are anonymized and kept to the minimum necessary.

2.3. Data Preprocessing

Preprocessing is a critical stage in this study to transform raw crawled X data into cleaned text for NLP-based analysis. It ensures consistent structure, topic relevance, and reduced noise commonly found in social media content, such as spam, links, emojis, and non-standard words [22].

The pipeline runs automatically, covering text/time column detection, Asia/Jakarta time parsing and normalization, text cleaning, and deduplication. Automatic column detection enables the workflow to adapt to different dataset schemas (e.g., text, komentar, full_text; created_at, timestamp), improving cross-file compatibility and reproducibility with minimal manual adjustment. Time zone normalization is applied to keep temporal analyses consistent, particularly for weekly or monthly aggregation.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

After the structure is identified, the text is cleaned by removing URLs, mentions, hashtags, emojis, and excessive symbols. Non-standard words and repeated characters are normalized to improve lexical consistency. Deduplication is then applied to reduce bias from repeated or identical posts (e.g., reposts). The cleaned corpus is summarized using word-count statistics as an initial quality indicator. Overall, preprocessing improves input quality and helps the dataset better represent unique public opinions. Systematic preprocessing has been shown to enhance sentiment analysis performance in social media settings characterized by informal language, code-mixing, and high writing variation [14].

2.4. Data Labeling

The data labeling stage is a critical component of this research pipeline because it directly determines the quality of the sentiment classification model to be developed. Labeling was conducted using a semi-supervised strategy that combines automatic labeling based on a pre-trained teacher model with manual verification on a subset of the data. This approach was selected to balance efficiency in handling large-scale datasets with the reliability and accuracy of the resulting sentiment labels [23], [24].

Table 1. Data Labeling Parameters

Component	Method	Parameter (code)	Output
Auto-label (teacher)	Pre-trained Indonesian Transformer	TEACHER_MODEL='taufiqdp/indonesian-sentiment'; softmax -> confidence	auto_label_raw + auto_confidence
Sarcasm flag	Rule-based detection of humor/sarcasm cues	sarcasm_tokens={wkwk,haha,lol,anjir,...}; flag∈{0,1}	Marker for potentially sarcastic texts
Refinement label	Conservative rules for noise reduction	conf<0.60→netral; wc<5 & conf<0.80→netral; sarcasm=1 & conf<0.90→netral	auto_label stable
GOLD 20% template	Semi-directed sampling	20% (min 50); 50% hard: conf<0.80 atau wc<=5 atau sarcasm=1; 50% random	.xlsx template label_manual

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Component	Method	Parameter (code)	Output
Manual label validation	Label normalization & invalid checks	mapping negatif/netral/positif; remove empty/invalid labels	label_manual (GOLD)

Based on Table 1, the auto-labeling step produces initial sentiment labels along with confidence scores that indicate prediction certainty. The pipeline then adds a sarcasm flag to mark potentially ambiguous texts and applies a refinement procedure to reduce noise by stabilizing labels in higher-risk cases, such as low-confidence predictions, very short texts, or posts suspected to contain sarcasm. After automatic labeling, a 20% GOLD subset is constructed using a combination of hard-case sampling and random sampling so that manual verification covers both common patterns and more challenging examples. The manually verified labels are subsequently normalized and validated to ensure consistency across the three classes (negative, neutral, and positive). This process results in more stable labels for training and a more reliable ground truth for model validation and evaluation [15]. Although a sarcasm flag is included to reduce obvious ambiguity, sarcasm remains difficult to detect reliably in short social media texts; therefore, sarcasm handling is treated as a limitation that may still contribute to residual label noise and misclassification.

2.5. Sentiment Classification Modeling

The modeling stage constitutes the core of the study, as it aims to develop a system that can classify sentiment into three classes, namely positive, negative, and neutral, based on public conversations on X. To enable objective comparison, this study employs two approaches: a feature-based baseline model and a Transformer-based deep learning model [25]. The GOLD dataset is then split into training, validation, and test sets using stratified splitting to preserve class proportions, with a non-stratified fallback mechanism applied when the class distribution does not satisfy stratification requirements [26].

In the IndoNLU benchmark, IndoBERT and IndoBERT-lite are introduced as contextual representation models for Indonesian texts, designed to capture variations in vocabulary and writing styles, including informal forms commonly found in social media

data. IndoBERT-lite is intended for computational efficiency, whereas the main model provides a stronger and more stable contextual representation; therefore, it is preferred to maintain classification quality in a three-class setting [27]. Based on these considerations, this study selects the base IndoBERT model and adopts an efficient training configuration using truncation (max_length = 128), gradient accumulation, and early stopping in a GPU-enabled Google Colab environment, so that the fine-tuning process remains practical for iterative experimentation, as shown in Table 2.

Table 2. Model Classification Parameters

Parameter	Value	Notes
learning_rate	2e-5	Common setting for Transformer fine-tuning
num_train_epochs	8	With early stopping (patience = 2)
max_length	128	Truncation for efficiency
batch_size train	16 (GPU)	Number of samples processed per step
gradient_accumulation	2 (GPU)	Maintains a larger effective batch size
warmup_ratio	0.1	Stabilizes early training
weight_decay	0.01	Regularization
metric_for_best_model	f1_macro	Prioritizes balance across classes
loss	Weighted Cross-Entropy	Addresses class imbalance

After the dataset split, safe pseudo-labeling was applied by adding non-GOLD samples with high confidence and no sarcasm indication only to the training set, thereby increasing textual variation without affecting validation and testing. The main stage consisted of fine-tuning IndoBERT using the training configuration in Table 2, including weighted cross-entropy to address class imbalance [28] and early stopping to reduce overfitting when validation performance no longer improves [29]. Overall, this design yields a stable modeling pipeline that is adaptive to imbalanced data distributions and provides performance gains compared with the baseline approach.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

2.6. Model Evaluation

Model evaluation was conducted to quantitatively measure sentiment classification performance and to support result interpretation through corpus visualizations [12]. In this pipeline, evaluation metrics were computed by comparing model predictions against the ground-truth labels in the test set, thereby reflecting the model's generalization capability. In addition to aggregate metrics, the evaluation includes per-class summaries via a classification report and a confusion matrix visualization to examine systematic prediction errors across classes, as shown in Table 3.

Table 3. Main Metrics and Their Roles

Metric	Implementation	Interpretation
Accuracy	accuracy_score(labels, y_pred)	Measures the overall proportion of correct predictions
Precision (macro)	precision_recall_fscore_support(labels, y_pred, average="macro", zero_division=0)[0]	Measures the average prediction correctness across classes equally
Recall (macro)	precision_recall_fscore_support(labels, y_pred, average="macro", zero_division=0)[1]	Measures the model's ability to capture each class equally
F1 (macro)	precision_recall_fscore_support(labels, y_pred, average="macro", zero_division=0)[2]	Primary metric assessing the balance of precision and recall across classes
F1 (weighted)	precision_recall_fscore_support(labels, y_pred, average="weighted", zero_division=0)[2]	Complementary metric that accounts for class support (sample proportions)

Referring to Table 3, this study prioritizes macro-F1 as the main metric because the three sentiment classes are treated as equally important, preventing the evaluation from being dominated by a larger class. Accuracy is still reported as a general indicator, while macro-precision and macro-recall are used to assess the balance between

prediction correctness and coverage across classes. As a complement, weighted F1 is included to reflect overall performance while considering the class distribution in the dataset [30].

Table 4. Visualizations and Their Roles

Visualization	Implementation	Interpretation
Weekly tweet volume plot	df["__week__"]=df[DATE_COL].dt.to_period("W").astype(str) lalu weekly=df.groupby("__week__").size().reset_index(name="count") dan plt.plot(weekly["__week__"], weekly["count"], marker="o")	Shows the weekly tweet volume dynamics during the April–June period
Histogram word count	df["word_count"]=df["text_clean"].str.split().str.len() lalu plt.hist(df["word_count"].values, bins=50)	Shows text-length distribution to indicate corpus density and potential classification difficulty
Confusion matrix	cm=confusion_matrix(y_true, y_pred, labels=[0,1,2]) lalu plt.imshow(cm, cmap="Blues") dan anotasi nilai pada sel (fungsi plot_confmat_blue)	Visualizes misclassification patterns across classes (negative, neutral, positive) and identifies frequently confused classes

Referring to Table 4, the weekly visualization provides descriptive context by showing fluctuations in discussion intensity across the data collection period, while the word count histogram helps explain variations in text length, which may be associated with classification difficulty. The word cloud is used as an exploratory summary to highlight dominant words after cleaning, ensuring that the interpretation remains grounded in the most frequent lexical patterns within the corpus. The confusion matrix constitutes an important part of evaluation because it reveals the direction of prediction errors, enabling a more specific understanding of model weaknesses beyond average scores [31]. This study combines quantitative metrics and interpretive visualizations to produce more informative and publication-ready results. The combination of macro-F1, per-class summaries, and the confusion matrix provides a more representative assessment for

three-class classification, while descriptive corpus analysis helps connect model performance with the structure and characteristics of the analyzed data.

3. RESULTS AND DISCUSSION

This section presents the overall research results, covering the entire workflow from data collection, preprocessing, labeling, and sentiment classification modeling to model performance evaluation. All stages were designed as a structured pipeline to ensure reproducibility, and each output (tables/figures) can serve as empirical evidence for the subsequent discussion.

3.1. Data Collection

The research data were obtained by crawling public conversations on X (Twitter) using a Google Colab pipeline. Data collection was restricted to April–June 2025 to capture the phase in which the vasectomy issue was actively discussed while the discourse context remained relatively stable. After the downloaded tabular files (.xlsx/.csv) were loaded, 9,900 raw entries were retrieved for the selected window (prior to cleaning and deduplication).

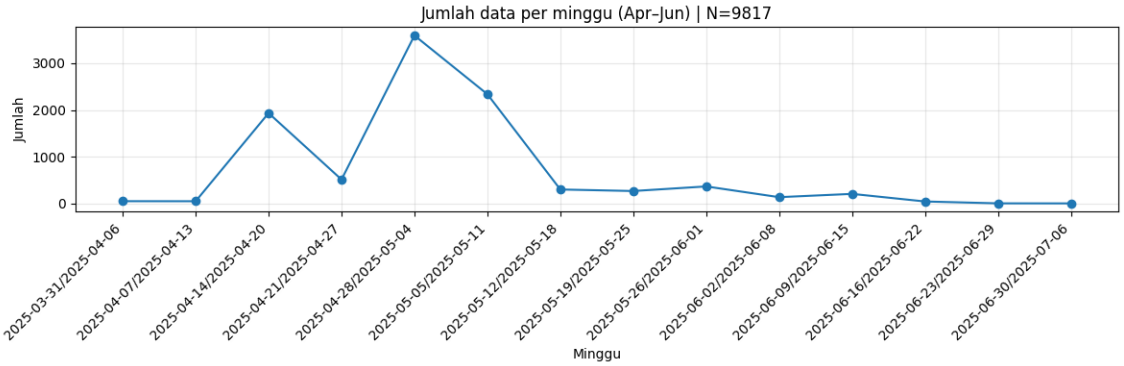

Figure 2. Comment Volume Visualization

Based on the weekly/monthly comment volume visualization as shown in Figure 2, the discussion dynamics were uneven across the observation period. In April, 3,351 posts were collected; the conversation peaked in May with 6,110 posts; and it declined sharply in June to 439 posts. This pattern suggests a strong viral phase in the middle of the period followed by a rapid decrease toward the end. Therefore, restricting the data collection window to three months is considered appropriate to reduce the risk of

contextual bias (e.g., the inclusion of unrelated issues outside the viral phase), maintain discourse homogeneity, and ensure that the analyzed data represent the most meaningful period of discussion intensity.

3.2. Data Preprocessing

The preprocessing stage was conducted to improve corpus quality before proceeding to labeling and modeling. The process included checking for empty entries, normalizing timestamp formats, cleaning social media text, and performing deduplication to reduce bias caused by repeated content (e.g., reposts/retweets or identical comments).

Table 5. Examples of Before vs. After Comment Cleaning

text_raw	text_clean
Metode kontrasepsi vasektomi bagi laki-laki aman dilakukan. Namun tidak banyak yang menjadi akseptor vasektomi. #Humaniora #AdadiKompas https://t.co/lRq3neM3wa	*metode kontrasepsi vasektomi bagi laki laki aman dilakukan namun tidak banyak yang menjadi akseptor vasektomi*
pentingnya mengerti pengunaan alat kontrasepsi. ada kondom ada morning pills ada kb ada vasektomi tinggal dipilih	*pentingnya mengerti pengunaan alat kontrasepsi ada kondom ada morning pills ada kb ada vasektomi tinggal dipilih*
Pro-Kontra Rencana Dedi Mulyadi Jadikan Vasektomi Sebagai Syarat Penerima Bansos https://t.co/MzORgrplck	*pro kontra rencana dedi mulyadi jadikan vasektomi sebagai syarat penerima bansos*
@DISSOSP3APPKB Kapan ya di Klaten ada program vasektomi gratis?	*kapan ya di klaten ada program vasektomi gratis*
Ada Program Vasektomi Gratis Ternyata Segini Biayanya Kalau Menggunakan BPJS Kesehatan https://t.co/hm857bwHyX	*ada program vasektomi gratis ternyata segini biayanya kalau menggunakan bpjs kesehatan*

Text cleaning was performed by removing URLs and mentions (@user), normalizing hashtags, removing emojis, filtering non-informative non-alphanumeric characters, and normalizing selected common slang forms to ensure more consistent lexical representations, as illustrated in the examples above.

Table 6. Preprocessing Summary

rows_before_nonempty	rows_after_nonempty	removed_empty_text	duplicates_removed	final_rows_after_dedup
9900	9900	0	83	9817

The preprocessing summary is presented in Table 6, with the following details: the number of entries before filtering empty text was 9,900, and after confirming non-empty text it remained 9,900 (removed_empty_text = 0). A total of 83 duplicates were identified (duplicates_removed = 83), resulting in 9,817 final entries after deduplication (final_rows_after_dedup = 9,817). This indicates that the corpus used in subsequent stages consists of 9,817 unique cleaned comments, which is more representative and reduces the influence of repeated texts on label distributions and model learning. Accordingly, preprocessing not only improves data cleanliness but also strengthens analytical validity by ensuring that each entry is more likely to reflect a distinct opinion.

3.3. Data Labeling

The labeling stage in this study applied a semi-supervised (weakly supervised) labeling approach by combining model-based auto-labeling with manual labeling as the reference dataset (GOLD). Auto-labeling was performed using a pre-trained Indonesian Transformer model, namely taufiqdp/indonesian-sentiment.

3.3.1. Auto-labeling (Teacher Model) and Refinement

Auto-labeling (TeacAuto-labeling was conducted using a pre-trained Indonesian Transformer (teacher model) that produces two main outputs: auto_label_raw (initial predicted label) and auto_confidence (prediction confidence score). The distribution of auto labelling as shown in Table 7. Auto-label DNext, refinement rules were applied to stabilize labels in cases that are more likely to be noisy (e.g., low-confidence predictions or very short texts). After refinement, the distribution remained negative 6,846, neutral

2,248, and positive 723, as reported in Table 7. This refinement step conservatively shifts a portion of risky cases from polar classes (negative/positive) toward neutral, which is reasonable because short or low-confidence texts are often ambiguous and are more prone to misclassification if forced into a polar category.istribution.

Table 7. Auto-label Distribution

Label	Count
Negative	6846
Neutral	2248
Positive	723

To help identify potentially difficult texts (e.g., humor, sarcasm, or laughter tokens such as "wkwk"), the pipeline also added a rule-based sarcasm_flag marker. The results in Table 8 show that 330 texts were flagged as potentially sarcastic (flag = 1), while 9,487 texts were not flagged (flag = 0). Although this is a rule-based indicator rather than a full sarcasm detector, it is useful for (i) enriching manual samples with more challenging cases and (ii) reducing overconfidence in automatic labels for texts that are pragmatically ambiguous. Table 8 shows that 330 texts were flagged as potentially sarcastic (sarcasm_flag = 1), while 9,487 texts were not flagged (sarcasm_flag = 0). Although this marker is rule-based rather than a full sarcasm detection model, it is useful for (i) enriching the manual subset with more challenging cases and (ii) reducing overconfidence in automatic labels for texts that are pragmatically ambiguous.

Talel 8. Sarcasm Flag Distribution

Sarcasm_flag	Count
0	9487
1	330

3.3.2. Manual Labeling (GOLD)

After auto-labeling, the study created a GOLD subset for manual labeling, amounting to 20% of the total data (1,963 comments). Sampling was semi-directed: a portion was selected randomly to represent typical data, while another portion was drawn from hard cases (e.g., lower confidence, very short texts, or sarcasm-flagged posts) so that

manual verification also covered examples that were most prone to error. The manual labeling distribution in Table 9 shows a composition of 798 negative, 671 positive, and 494 neutral samples out of 1,963 texts. This indicates that within the human-verified subset, negative and positive classes are relatively comparable in size, while neutral remains present but smaller. These GOLD labels were then used as the primary reference for train/validation/test splitting and for model evaluation, thereby improving the reliability of reported performance results.

Table 9. Manual Label Distribution (GOLD)

Label	Count
Negative	798
Positive	671
Neutral	494

3.4. Sentiment Classification Modeling

The modeling stage aims to develop a model that can classify sentiment into three classes (negative, neutral, and positive). In this study, evaluation was conducted under a strict protocol in which the validation and test sets were drawn only from the GOLD subset, ensuring that the reported performance metrics reflect the model's ability to generalize to human-validated labels.

3.4.1. IndoBERT Tokenization

Before training, the cleaned texts (text_clean) were processed using the IndoBERT tokenizer, which applies a subword-based WordPiece scheme, as shown in Table 10. The tokenization output typically includes special tokens such as [CLS] at the beginning and [SEP] at the end, and it may split a word into subwords (e.g., tokens prefixed with ##) when the word is not available as a single token. The tokenization table (columns n_tokens and tokens_preview) illustrates that a single sentence can produce multiple tokens representing word fragments, allowing the model to handle non-standard words, spelling variations, and emerging terms that commonly appear in social media text.

Table 10. IndoBERT Tokenization

text_clean	n_tokens	tokens_preview(<=30)
metode kontrasepsi vasektomi bagi laki laki aman dilakukan namun tidak banyak yang menjadi akseptor vasektomi	26	*[CLS] metode kon ##tra ##sep ##si vas ##ek ##tom ##i bagi laki laki aman dilakukan namun tidak banyak yang menjadi akseptor vas ##ek ##tom ##i [SEP]*
pentingnya mengerti pengunaan alat kontrasepsi ada kondom ada morning pills ada kb ada vasektomi tinggal dipilih	24	*[CLS] pentingnya mengerti pengunaan alat kon ##tra ##sep ##si ada kondom ada morning pills ada kb ada vas ##ek ##tom ##i tinggal dipilih [SEP]*
pro kontra rencana dedi mulyadi jadikan vasektomi sebagai syarat penerima bansos	20	*[CLS] pro kontra rencana dedi mul ##yad ##i jadikan vas ##ek ##tom ##i sebagai syarat pene ##rima bans ##os [SEP]*
kapan ya di klaten ada program vasektomi gratis	13	*[CLS] kapan ya di klaten ada program vas ##ek ##tom ##i gratis [SEP]*
ada program vasektomi gratis ternyata segini biayanya kalau menggunakan bpjs kesehatan	18	*[CLS] ada program vas ##ek ##tom ##i gratis ternyata segini biayanya kalau menggunakan bp ##js kese ##hatan [SEP]*

Table 10 reports WordPiece tokenization results for text_clean, including n_tokens (the number of tokens) and tokens_preview (a short token preview), with [CLS] added at the beginning and [SEP] at the end. Subword splitting (the ## prefix) enables IndoBERT to represent informal spelling and non-standard forms more robustly, while converting the text into a model-ready input format.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

3.4.2. Training Process and Early Stopping

The main model was trained by fine-tuning IndoBERT for a maximum of 8 epochs and equipped with early stopping (patience = 2), so training stops automatically when validation performance no longer improves. The training history records per-epoch evaluation results (e.g., Training Loss, Validation Loss, Accuracy, Macro F1, and related metrics), as summarized in Table 11.

Table 11. Evaluasi epoch Process

Epoch	Training Loss	Validation Loss	Accuracy	F1 Macro	F1 Weighted	Precision Macro	Recall Macro
1	0.652200	0.574588	0.700637	0.711266	0.700985	0.715992	0.719744
2	0.407800	0.465435	0.859873	0.859107	0.860050	0.871125	0.851956
3	0.325400	0.434778	0.821656	0.823953	0.822407	0.829008	0.820186
4	0.255200	0.551781	0.828025	0.828807	0.830479	0.859170	0.823095

In this case, training stopped at epoch 4 (rather than 8) because validation improvements were no longer consistent or did not exceed the best performance achieved in earlier epochs for multiple consecutive evaluations, in line with the patience mechanism. Therefore, stopping at epoch 4 indicates that the best-performing model was identified earlier, and continuing training could increase overfitting without meaningfully improving generalization.

3.5. Sentiment Classification

This section presents the final outputs of the sentiment classification process using IndoBERT, which was fine-tuned on the GOLD dataset (manual labels) and enriched with high-confidence pseudo-labeled samples in the training set. The results are reported through (1) the predicted label distribution for the full corpus (label_model), (2) quantitative evaluation on the GOLD test set via a classification report, and (3) error pattern visualization using a confusion matrix, enabling the model performance to be discussed in both measurable and interpretive terms.

Table 12. IndoBERT Predicted Label Distribution

Label	Count
Negative	3957
Positive	3520
Neutral	2340

Based on Table 12, IndoBERT predictions over the entire corpus indicate that the negative class is the largest category (3,957 samples), followed by a closely comparable positive class (3,520 samples), while neutral is the smallest group (2,340 samples). This pattern suggests that discussions on vasectomy during April–June 2025 more frequently contain expressions of rejection, criticism, or concern (negative), yet are also accompanied by a substantial level of supportive or defensive responses (positive), whereas informational or ambiguous responses (neutral) appear less frequently than the other two classes.

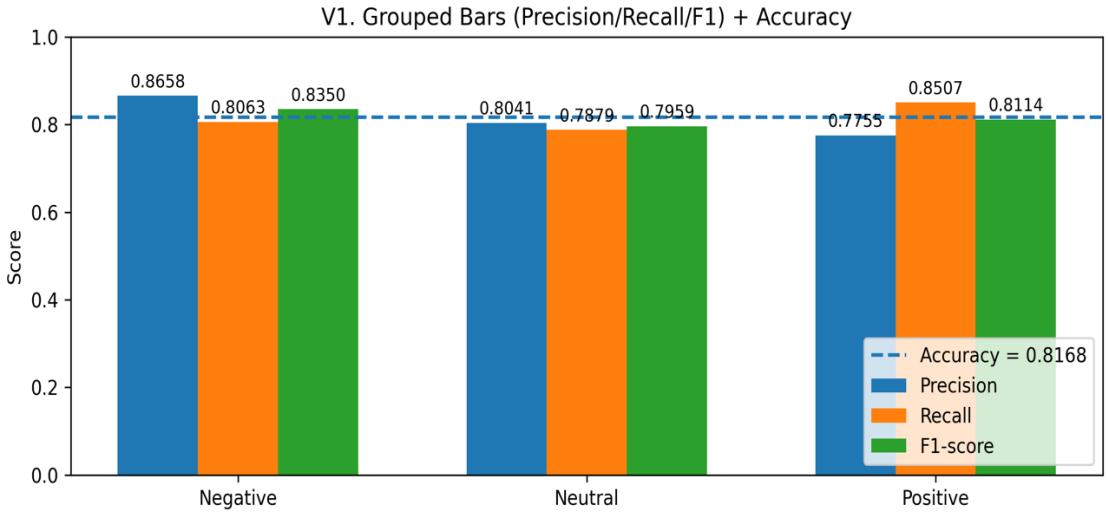

Figure 3. IndoBERT Evaluation Metrics

Quantitative evaluation on the GOLD-only test set shows that the model achieved an accuracy of 0.8168 (≈81.68%, commonly rounded to ~81%). At the class level, the negative class achieved a precision of 0.8658 and a recall of 0.8063 (F1 = 0.8350), indicating that negative predictions are relatively accurate and that most negative instances are successfully identified. The neutral class achieved an F1-score of 0.7959 (precision 0.8041, recall 0.7879), reflecting stable performance, although it remains

susceptible to confusion with other classes. The positive class achieved the highest recall (0.8507) but a lower precision (0.7755), indicating that the model captures positive instances well, yet some positive predictions still include samples from other classes. The macro-F1 of 0.8141 indicates relatively balanced performance across classes rather than performance driven by a single majority class, while the weighted F1 of 0.8171 reflects overall performance while accounting for class support. Confusion Matrix IndoBERT as shown in Figure 4.

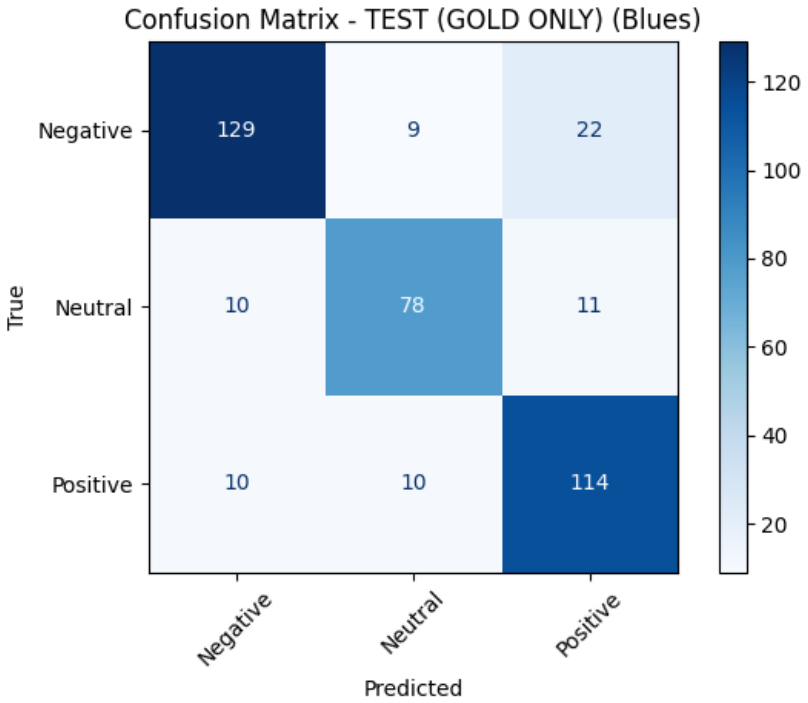

Figure 4. Confusion Matrix IndoBERT

A more detailed interpretation is provided by the GOLD test confusion matrix. For the negative class, the model correctly classified 129 samples, but misclassified 9 as neutral and 22 as positive. For the neutral class, 78 samples were correctly classified, while 10 were misclassified as negative and 11 as positive. For the positive class, the model correctly classified 114 samples, while 10 were misclassified as negative and 10 as neutral. This pattern suggests that the most prominent error occurs in negative-to-positive confusion (e.g., 22 negatives predicted as positive), which is common in social media text due to sarcasm, irony, and pragmatic ambiguity, where emotional intent is difficult to infer from surface-level wording alone.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Figure 5. Word Count IndoBERT

The word count visualization shows the distribution of text lengths (number of words) in each post after cleaning. Figure 5 helps describe typical corpus characteristics and contextual density, which may influence how effectively the model captures sentiment signals during classification.

The modeling results indicate that fine-tuning IndoBERT yields strong classification performance on the GOLD test set, with an accuracy of approximately 81% and a macro-F1 of approximately 0.81, while providing a consistent sentiment mapping of public discussions during April–June 2025. The classification report and confusion matrix further suggest that a portion of prediction errors concentrates in posts that are ambiguous, very short, or contain humor/sarcasm, where sentiment cues are less explicit and more easily confused across classes. Substantively, the full-corpus prediction distribution indicates a "tension" in public opinion: positive support remains visible and substantial, yet the overall discussion pattern is slightly more inclined toward negative sentiment. This tendency implies that the public discourse on vasectomy is not fully stable and continues to be influenced by resistance, stigma, and concerns, even as a segment of users expresses more rational and supportive views. These findings provide a basis for subsequent discussion, particularly in relating discourse dynamics, social media language characteristics, and the observed imbalance in sentiment tendencies during the study period.

3.6. Discussion

The weekly/monthly volume pattern across April–June 2025 shows a highly concentrated attention cycle that can be interpreted as a "viral window": discussion grew in April (3,351 posts), peaked sharply in May (6,110 posts), and then collapsed in June (439 posts). This steep rise–fall dynamic indicates that the discourse was temporally bounded and strongly event-driven, which supports the methodological decision to restrict analysis to a narrow window in order to preserve contextual consistency and reduce topic drift from unrelated issues [21]. In other words, the three-month restriction helps ensure that the analyzed posts represent a relatively coherent public conversation responding to the same salient trigger, rather than a mixture of different news cycles. This also aligns with the broader premise that social media functions as a real-time mirror of public opinion dynamics, where rapid bursts of attention can amplify both evidence-based information and emotionally charged reactions [6].

Within this viral window, the full-corpus sentiment mapping produced by the fine-tuned IndoBERT model suggests a slightly negative-leaning but clearly contested discourse: negative sentiment is the largest category (3,957), followed closely by positive sentiment (3,520), while neutral sentiment is smaller (2,340). This distribution implies polarization rather than uniform rejection. Substantively, such a split is consistent with the tension described in the background literature: although vasectomy is medically safe and does not impair hormones or sexual performance, the topic is frequently filtered through masculinity norms, stigma, perceived religious incompatibility, and misconceptions about infertility [3], [4], [19], [20]. The simultaneous visibility of positive sentiment indicates that supportive narratives—such as correcting myths, emphasizing shared responsibility, or defending male involvement—were also present and actively circulated during peak attention, consistent with prior evidence that online spaces can host both resistance and counter-responses to health and policy debates [7], [8]. Therefore, the sentiment landscape should be understood as an arena where competing frames coexist, rather than as a single dominant narrative about male contraception.

From a modeling standpoint, the evaluation on the GOLD-only test set indicates that IndoBERT provides strong and relatively balanced three-class performance for informal

Indonesian social media text, achieving an accuracy of 0.8168 and macro-F1 of 0.8141. Class-wise, the negative class achieves higher precision (0.8658) than recall (0.8063), suggesting that when the model predicts "negative," it is usually correct, but a portion of truly negative texts still spill into other classes. Conversely, the positive class shows the highest recall (0.8507) but lower precision (0.7755), implying that the model tends to "capture" positive instances broadly but occasionally absorbs borderline or pragmatically ambiguous texts that belong elsewhere. The neutral class remains stable (F1 = 0.7959), but the confusion matrix confirms that neutrality is often a buffer zone where ambiguous expressions are pulled toward negative or positive. These patterns are compatible with known challenges in Indonesian social media sentiment classification—especially where slang, short text length, code-mixing, and pragmatic cues drive meaning beyond literal wording—conditions under which Transformer models like IndoBERT generally outperform classical baselines but still face ambiguity limits [9], [11], [12], [27].

The confusion matrix provides a more concrete explanation of where ambiguity is most costly. While correct predictions are high for all classes (129 negative, 78 neutral, 114 positive), the largest cross-polar error occurs when negative texts are predicted as positive (22 cases), exceeding the reverse direction (10 positives predicted as negative). This asymmetry is plausible in vasectomy discourse on X, where sarcastic endorsement, mock praise, or humor can appear "positive" lexically while intending criticism (or vice versa), especially in short posts with compressed context. The study's pipeline anticipated this by introducing a sarcasm flag and conservative relabeling rules that shift high-risk cases toward neutral (e.g., low-confidence predictions, very short texts, and sarcasm-marked posts) [14]. However, the remaining errors indicate that sarcasm/irony is still difficult to capture reliably through rule-based cues alone, and that pragmatic ambiguity remains a key source of residual misclassification. In this sense, the observed error profile is not merely a technical artifact: it also signals that the public debate itself contains rhetorical forms (humor, satire, insinuation) that blur sentiment boundaries and can complicate interpretation for both humans and models.

Taken together, these findings imply two practical points for communication and policy. First, institutions such as BKKBN should avoid assuming that online discussion reflects a single stable attitude toward vasectomy; the near-balance between negative and

positive sentiment suggests a contested space where supportive education and stigmatizing narratives circulate simultaneously [6]. Second, the model's error hotspots highlight which types of content may be most vulnerable to misunderstanding and misinformation persistence: neutral/ambiguous posts, short statements, and humor/sarcasm-laden reactions. These are precisely the areas where public health communication may benefit from clearer framing, myth-focused clarification, and message designs that anticipate joking or ironic reframing while still reinforcing accurate information about safety, function, and the intended permanence of the method [3], [4]. Methodologically, the semi-supervised labeling strategy—combining teacher-model labeling with a manually verified GOLD subset—appears justified for scaling sentiment mapping while maintaining reliability, and the strong GOLD-only evaluation provides support for IndoBERT as a practical backbone model in Indonesian public-debate sentiment studies [14], [23], [24]. Future work could further reduce cross-polar errors by incorporating richer sarcasm detection, conversation-thread context, or stance/target-aware modeling so that the system distinguishes "support for vasectomy" from "support for coercive framing," which may be sentimentally positive but substantively controversial in policy terms.

4. CONCLUSION

This study mapped Indonesian netizens' sentiment toward vasectomy on X (Twitter) during April–June 2025 through a traceable end-to-end pipeline: data crawling in Google Colab, preprocessing (text cleaning and deduplication), semi-supervised labeling using the teacher model taufiqdp/indonesian-sentiment with confidence-based refinement and a rule-based sarcasm_flag, and supervised IndoBERT fine-tuning evaluated strictly on a manually verified GOLD subset. The sentiment distribution over 9,817 cleaned posts suggests a contested public conversation rather than a single dominant stance, with negative sentiment slightly leading (3,957) but positive sentiment remaining highly visible (3,520), and neutral content forming the smallest segment (2,340). This pattern indicates that misinformation, stigma, and resistance coexist alongside corrective and supportive messaging during the same high-attention period.

Model evaluation on the GOLD-only test set (n = 393) demonstrates that IndoBERT (indolem/indobert-base-uncased) delivers strong three-class performance (accuracy =

0.8168; macro F1 = 0.8141), supporting its practical use for Indonesian social media sentiment mapping under noisy language conditions. However, the remaining misclassifications concentrate in cross-polar and ambiguous texts, consistent with short-form discourse where sentiment is expressed through humor, sarcasm, or implied meaning—signals that are difficult to resolve reliably with a rule-based sarcasm_flag. Practically, these findings suggest that public communication strategies (e.g., by BKKBN and related stakeholders) should treat the vasectomy discourse as an actively contested space: messaging should not only disseminate clinical facts, but also anticipate ironic framing, address recurring misconceptions directly, and provide clear, culturally sensitive narratives that reduce ambiguity and improve interpretability in the formats people actually use online.

REFERENCES

[1] P. Sari, C. A. Febriani, and A. Farich, "Determinant Factors of Men's Participation as Family Planning Acceptors in Indonesia (2017 IDHS Data Analysis)," *J. Kesehat. Komunitas*, vol. 9, no. 1, pp. 138–148, 2023, doi: 10.25311/keskom.Vol9.Iss1.1306.

[2] N. M. Amanati, S. B. Musthofa, and A. Kusumawati, "Analysis of Factors Associated with Vasectomy Use in Karanganyar Village, Ngawi Regency, East Java (Analisis Faktor yang Berhubungan dengan Penggunaan Vasektomi di Desa Karanganyar Kabupaten Ngawi Jawa Timur)," *MKMI Media Kesehat. Masy.*, vol. 20, no. 2, pp. 91–98, 2021, doi: 10.14710/mkmi.20.2.91-98.

[3] F. Yang et al., "Review of Vasectomy Complications and Safety Concerns," *World J Men's Heal.*, vol. 39, no. 3, pp. 406–418, 2021, doi: 10.5534/wjmh.200073.

[4] E. S. Pallangyo, A. C. Msoka, S. Brownie, and E. Holroyd, "Religious beliefs, social pressure, and stigma: Rural women's perceptions and beliefs about vasectomy in Pwani, Tanzania," *PLoS One*, vol. 15, no. 3, 2020, doi: 10.1371/journal.pone.0230045.

[5] A. Fahrudin, N. Lisnarini, and G. Kurnia Dewi, "Policy Populism of West Java Governor Dedi Mulyadi (A Sentiment Analysis Study) (Populisme Kebijakan Gubernur Jawa Barat Dedi Mulyadi (Studi Analisis Sentimen))," *MASSIVE J. Ilmu Komun.*, vol. 5, no. 1, pp. 17–31, 2025, doi: 10.35842/massive.v5i1.175.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[6] E. A. Winanto, Z. Ali, P. A. Jusia, and Sharipuddin, "Sentiment Analysis of the Hashtag #KaburAjaDulu on Twitter Using a Lexicon-Based Method (Analisis Sentimen Terhadap Tagar Kabur Aja Dulu Di Twitter Menggunakan Metode Lexicon-Based)," *J. Process.*, vol. 20, no. 2, pp. 223–233, 2025, doi: 10.33998/processor.2025.20.2.2542.

[7] N. Putu Gita Naraswati, D. Cindy Rosmilda, D. Desinta, F. Khairi, R. Damaiyanti, and R. Nooraeni, "Twitter Public Sentiment Analysis Regarding Indonesia's COVID-19 Response Policy Using Naive Bayes Classification (Analisis Sentimen Publik dari Twitter Tentang Kebijakan Penanganan Covid-19 di Indonesia dengan Naive Bayes Classification)," *J. Sist. Inf.*, vol. 10, no. 1, pp. 228–238, 2021, doi: 10.32520/stmsi.v10i1.1179.

[8] M. F. Naufal and S. F. Kusuma, "Sentiment Analysis on Twitter Toward the Community Activity Restrictions Enforcement Policy (PPKM) Using Deep Learning (Analisis Sentimen pada Media Sosial Twitter Terhadap Kebijakan Pemberlakuan Pembatasan Kegiatan Masyarakat Berbasis Deep Learning)," *JEPIN (Jurnal Edukasi dan Penelit. Inform.)*, vol. 8, no. 1, pp. 44–49, 2022, doi: 10.26418/jp.v8i1.49951.

[9] J. S. Febrilliani and A. Wibowo, "Twitter Public Sentiment towards Vasectomy in Indonesia Using SVM and Naïve Bayes," *Res. Horiz.*, vol. 5, no. 4, pp. 1415–1424, 2025, doi: 10.54518/rh.5.4.2025.756.

[10] N. P. Sari, A. Munir, M. R. At, and M. Iskandar, "Twitter Sentiment Analysis of Long-Acting Reversible Contraceptives (LARC) Methods in Indonesia with Machine Learning Approach," in *Proceedings of the International Conference on Multidisciplinary Studies (ICoMSi 2023)*, Atlantis Press SARL, 2024, pp. 167–185. doi: 10.2991/978-2-38476-228-6_15.

[11] M. P. Firdaus and D. Trisnawarman, "Public Sentiment Analysis of the Public Housing Savings Program Using the IndoBERT Lite Model on YouTube Comments," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 5, no. 1, pp. 359–368, 2025, doi: 10.57152/malcom.v5i1.1744.

[12] V. D. Setiawan, D. U. Iswavigra, and E. Anggiratih, "Implementation of IndoBERT for Sentiment Analysis of the Constitutional Court's Decision Regarding the Minimum Age of Vice Presidential Candidates," *Sci. J. Informatics*, vol. 12, no. 3, pp. 397–406, 2025, doi: 10.15294/sji.v12i3.26360.

[13] A. Roihan, T. T. Atmojo, R. A. Wardoyo, and M. S. T. Saputra, "Sentiment Analysis of Twitter Data on the 2024 Indonesian Presidential Election Using BERT," *CCIT J.*, vol. 18, no. 1, pp. 39–45, 2024, doi: 10.33050/ccit.v18i1.3210.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[14] S. Khairunnisa, Adiwijaya, and S. Al Faraby, "The Effect of Text Preprocessing on Sentiment Analysis of Public Comments on Twitter (COVID-19 Pandemic Case Study) (Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19))," *J. Media Inform. Budidarma*, vol. 5, no. 2, pp. 406–414, 2021, doi: 10.30865/mib.v5i2.2835.

[15] N. F. Adhim and N. Cahyono, "Optimization of IndoBERT for Sentiment Analysis of FOMO on Social Media Through Fine-Tuning and Hybrid Labeling," *JAIC J. Appl. Informatics Comput.*, vol. 9, no. 6, pp. 3786–3797, 2025, doi: 10.30871/jaic.v9i6.11686.

[16] P. Slater and F. Hasson, "Quantitative Research Designs, Hierarchy of Evidence and Validity," *J. Psychiatr. Ment. Health Nurs.*, vol. 32, no. 3, pp. 656–660, 2024, doi: 10.1111/jpm.13135.

[17] H. D. Pradana, R. Rusijono, I. Y. Maureen, and E. Youhanita, "Artificial Intelligence in Learning Design: Acceptance, Perceived Effectiveness, and Barriers," *J. Penelit. dan Pengkaj. Ilmu Pendidik. e-Saintika*, vol. 9, no. 2, pp. 489–511, 2025, doi: 10.36312/e-saintika.v9i2.2688.

[18] D. Z. Abidin, L. Afuan, A. N. Toscany, and Nurhadi, "A Comprehensive Benchmarking Pipeline for Transformer-Based Sentiment Analysis using Cross-Validated Metrics," *J. Tek. Inform.*, vol. 6, no. 4, pp. 1797–1810, 2025, doi: 10.52436/1.jutif.2025.6.4.4894.

[19] J. A. Borrell, C. Gu, N. Ye, J. N. Mills, and J. J. Andino, "Comparing vasectomy techniques, recovery and complications: tips and tricks," *Int. J. Impot. Res.*, pp. 1–7, 2025, doi: 10.1038/s41443-025-01018-5.

[20] Y. N. D. W. D. Sehnur, "The Phenomenon of Masculine Panic Behind Male Contraception Programs (Fenomena Kepanikan Maskulin Dibalik Program Kontrasepsi Laki-Laki)," *LENTERA J. Gend. Child. Stud.*, vol. 4, no. 2, pp. 297–311, 2024, doi: 10.26740/lentera.v4i2.33455.

[21] A. Bechini, A. Bondielli, P. Ducange, F. Marcelloni, and A. Renda, "Addressing Event-Driven Concept Drift in Twitter Stream: A Stance Detection Application," *IEEE Access*, vol. 9, pp. 77758–77770, 2021, doi: 10.1109/ACCESS.2021.3083578.

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

[23] P. Ayuningtyas, S. Khomsah, and Sudianto, "Semi-Supervised Learning–Based Sentiment Labeling Using LSTM and GRU Algorithms (Pelabelan Sentimen Berbasis Semi-Supervised Learning menggunakan Algoritma LSTM dan GRU)," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 9, no. 3, pp. 217–229, 2024, doi: 10.14421/jiska.2024.9.3.217-229.

[24] Y. Yu, S. Zuo, H. Jiang, W. Ren, T. Zhao, and C. Zhang, "Fine-Tuning Pre-trained Language Model with Weak Supervision: A Contrastive-Regularized Self-Training Approach," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 1063–1077. doi: 10.18653/v1/2021.naacl-main.84.

[25] A. A. Qolbu, N. Fitriyati, and N. Inayah, "Performance of Naive Bayes, SVM, and IndoBERT for IndiHome Twitter Sentiment Analysis with Imbalanced-Data Handling Strategies (Performa Naïve Bayes, SVM, dan IndoBERT pada Analisis Sentimen Twitter IndiHome dengan Strategi Penanganan Data Tidak Seimbang)," *J. FOURIER*, vol. 814, no. 1, pp. 29–44, 2025, doi: 10.14421/fourier.2025.141.29-44.

[26] M. R. Manoppo et al., "Public Sentiment Analysis on Social Media Regarding Indonesia's 12% VAT Increase Using IndoBERT (Analisis Sentimen Publik Di Media Sosial Terhadap Kenaikan Ppn 12% Di Indonesia Menggunakan Indobert)," *J. Kecerdasan Buatan dan Teknol. Inf.*, vol. 4, no. 2, pp. 152–163, 2025, doi: 10.69916/jkbti.v4i2.322.

[27] W. A. Hidayat and V. R. S. Nastiti, "Performance Comparison of Pre-Trained IndoBERT-Base and IndoBERT-Lite for Sentiment Classification of TikTok Reviews of Tokopedia Seller Center Using IndoBERT (Perbandingan Kinerja Pre-Trained Indobert-Base Dan Indobert-Lite Pada Klasifikasi Sentimen Ulasan Tiktok Tokopedia Seller Center Dengan Model Indobert)," *JSil (Jurnal Sist. Informasi)*, vol. 11, no. 2, pp. 13–20, 2024, doi: 10.30656/jsii.v11i2.9168.

[28] D. K. Sumartha, "Implementation of IndoBERT for Sentiment Analysis of Public Opinion on the UKT Tuition-Fee Increase Policy in the Current Government Era (Implementasi Indobert Untuk Analisis Sentimen Opini Publik Terhadap Kebijakan Kenaikan UKT Di Era Pemerintahan)," *J. Inform. dan Tek. Elektro Terap.*, vol. 13, no. 3S1, pp. 867–875, 2025, doi: 10.23960/jitet.v13i3S1.7880.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[29] R. I. Perwira, V. A. Permadi, D. I. Purnamasari, and R. P. Agusdin, "Domain-Specific Fine-Tuning of IndoBERT for Aspect-Based Sentiment Analysis in Indonesian Travel User-Generated Content," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 11, no. 1, pp. 30–40, 2025, doi: 10.23960/jitet.v13i3S1.7880.

[30] A. Aljabar, B. M. Karomah, N. Tarisafitri, and J. Jeffry, "Sentiment Analysis in Indonesian's Presidential Election 2024 Using Transfomer (Distilbert-Base-Uncased)," *JSCE (Journal Syst. Comput. Eng.)*, vol. 6, no. 2, pp. 197–203, 2025, doi: 10.61628/jsce.v6i2.1867.

[31] A. S. Muliana, D. Lestarini, and S. P. Raflesia, "Analysis of Public Sentiment on Election Results using Naïve Bayes in Social Media X," *Sistemasi*, vol. 13, no. 6, pp. 2467–2478, 2024, doi: 10.32520/stmsi.v13i6.4592.