Journal on Informatics Visualization and Social Computing
Volume 2, Issue 1, 33-47.
e_ISSN: E-ISSN 3123-7002
https://e-journal.gomit.id/jivsc
DOI:10.58723/jivsc.v2i1.123

Sentiment Analysis on the Indonesian Military Draft Bill in YouTube Comments Using a LSTM Model

Adib Raihan Ashidiq
University PGRI Yogyakarta, Indonesia

Fradika Anggara Putra
University PGRI Yogyakarta, Indonesia

Agil Febri Pradana
University PGRI Yogyakarta, Indonesia

Nurirwan Saputra*
University PGRI Yogyakarta, Indonesia

Article Info

Article history:

Received: April 10, 2026
Revised: May 20, 2026
Accepted: June 10, 2026

Keywords:

Sentiment Analysis;
RUU TNI;
Long Short-Term Memory (LSTM);
FastText Embedding;
Public Opinion Monitoring.

Abstract

Background: The legislative proposal for the revision of the Indonesian Military Draft Bill (RUU TNI) has incited significant public controversy, largely articulated in digital arenas like YouTube. This revision fuels widespread apprehension about the potential revival of the military's dual-function role (Dwifungsi ABRI) and its potential detriment to civilian supremacy.
Aims: This study aims to quantify and analyze public sentiment towards the RUU TNI, as expressed in Indonesian YouTube comments, using a Deep Learning approach combining FastText embedding and Long Short-Term Memory (LSTM) classification.
Methods: A dataset of 520 Indonesian comments was collected and manually labeled into three classes: positive, neutral, and negative. The methodology included comprehensive text preprocessing, feature extraction via FastText word embeddings (300 dimensions), and sentiment classification using the LSTM architecture. The model was rigorously evaluated using a confusion matrix across standard metrics, including accuracy, precision, recall, and F1-score.
Result: The dataset exhibited significant class imbalance, dominated by negative sentiment (42.31%). The optimal LSTM configuration, tested with Stratified K-Fold Cross Validation, yielded an accuracy of 42.39%, a precision of 43%, a recall of 47%, and an F1-score of 60%. Topic modeling via LDA revealed dominant themes criticizing the bill's quality, state corruption, and legislative integrity. The low accuracy, barely surpassing the baseline, suggests that the model struggled with the limited and imbalanced data.
Conclusion: Public sentiment regarding the RUU TNI is strongly negative and critical, reflecting deep concerns about democracy and state institutions. While the FastText-LSTM pipeline was established, its classification performance was severely constrained by the small, highly imbalanced dataset and the complex nature of political discourse. Future research must utilize advanced Transformer models (e.g., IndoBERT) and larger, balanced datasets to achieve meaningful classification performance on this critical socio-political domain.

To cite this article: Ashidiq et al.. (2026). Sentiment Analysis on the Indonesian Military Draft Bill in YouTube Comments Using a LSTM Model. *Journal on Informatics Visualization and Social Computing, 2*(1), 33-47. https://doi.org/10.58723/jivsc.v2i1.123

INTRODUCTION

Indonesia is one of the world's largest electoral democracies, and within its democratic framework the role of the armed forces is deliberately constrained to prevent interference in civilian affairs and to preserve the subordination of the military to civilian authority (Aspinall & Hicken, 2019; Berenschot & Aspinall, 2019). The public debate surrounding the proposed revision of the Indonesian Military Law (Rancangan Undang-Undang Tentara Nasional Indonesia, RUU TNI) has reopened long-standing concerns about the military's involvement in the civil domain. Critics contend that several proposed amendments could reopen the door to the military's dual-function role (Dwifungsi ABRI) and thereby threaten civilian supremacy (Agustino & Hikmawan, 2026; Djuyandi et al., 2025). The Indonesian Legal Aid Foundation (YLBHI), for example, has publicly opposed the revision, framing it as a regression in Indonesian democratic governance. Because such debates increasingly unfold in online spaces, systematically measuring how the public reacts to the bill is both timely and policy-relevant(Fernando et al., 2024).

*** Corresponding author:**
Nurirwan Saputra, Universitas PGRI Yogyakarta, Indonesia. nurirwan@upy.ac.id.

In the digital era, public opinion on socio-political issues is increasingly voiced through social media rather than conventional mass media alone (Yuniningsih et al., 2024). YouTube is a particularly influential venue: its comment sections let users express views openly and spontaneously, and prior work indicates that analyzing these comments can capture public perception effectively because of the platform's high participation and interaction rates (Wicaksono et al., 2024). YouTube comments therefore constitute a rich, near real-time source for mapping public sentiment toward sensitive national policies such as the RUU TNI (Brynjolfsson et al., 2024).

Sentiment analysis, a core task in Natural Language Processing (NLP), identifies and classifies the opinions or emotions expressed in text into polarities such as positive, negative, and neutral (Udayana et al., 2023). In the Indonesian context, sentiment analysis has been applied to product reviews, online news, and platforms such as Twitter/X to summarize opinion trends. Methodologically, classification approaches range from classical machine-learning algorithms such as Naïve Bayes (Putri et al., 2022; Saputra et al., 2022), Support Vector Machine (Saputra et al., 2015), Random Forest (Isnayni et al., 2024; Saputra et al., 2023), and Lazy K-Star (Saputra, 2019) to deep-learning models. Among the latter, the Long Short-Term Memory (LSTM) network (Hochreiter & Schmidhuber, 1997) is well suited to sequence tasks because its gating mechanism mitigates the vanishing-gradient problem of standard Recurrent Neural Networks (RNNs) and lets it model longer-range dependencies in text (Nidhi et al., 2024; Tripathi, 2021).

Two gaps motivate this study. First, work that specifically examines public opinion on the RUU TNI through Indonesian YouTube comments remains scarce, so the domain itself is under-documented(Sadat & Basir, 2025). Second, the deep-learning studies that motivate such work are usually trained on general domains and on much larger corpora, and the behavior of an LSTM classifier on a small, imbalanced, politically sensitive Indonesian dataset has not been documented in detail. This study addresses both gaps. It maps public sentiment and topics for the RUU TNI, and it reports how a FastText–LSTM pipeline behaves under these constrained data conditions(Airlangga & Adiningsih, 2023; Topić, 2026).

The primary objective is to quantify and characterize the sentiment polarity (positive, neutral, negative) and the dominant thematic topics surrounding the RUU TNI in Indonesian YouTube comments. The secondary objective is to build and rigorously evaluate a FastText–LSTM model for multi-class sentiment classification on this domain-specific dataset, and to analyze honestly the conditions under which it succeeds or fails.

The contributions of this study are threefold. (i) An empirical contribution: a manually labeled, domain-specific Indonesian dataset of 520 RUU TNI comments, together with a descriptive characterization of public opinion and its dominant topics. (ii) A methodological contribution: a reproducible end-to-end workflow that combines scraping, preprocessing, FastText embeddings, LSTM classification, and LDA topic modeling. (iii) A cautionary empirical finding: evidence that, without class-imbalance handling and a larger corpus, a FastText–LSTM classifier collapses onto the majority class on this task, an outcome that quantifies the data requirements for future work rather than claiming a competitive classifier.

METHOD

Research Design and Data Collection

The study follows a standard NLP pipeline, summarized in Fig. 1: data collection and labeling, text preprocessing, FastText feature extraction, LSTM classification with model evaluation, and LDA topic modeling.

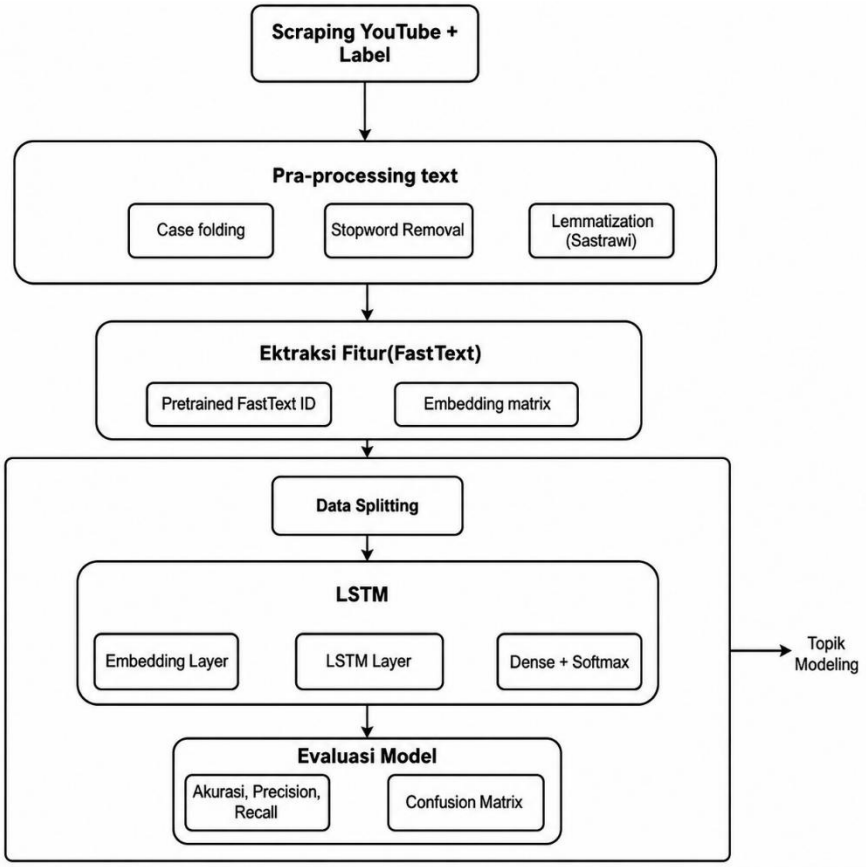

Figure 1. Overall workflow of the proposed sentiment-analysis method.

Dataset and sampling

Comments were collected from YouTube using the query "RUU TNI". Videos were chosen by purposive sampling on the basis of explicit topical relevance (the RUU TNI, the military's civil role, or Dwifungsi ABRI). For each selected video, all qualifying Indonesian-language comments were retrieved (total sampling), yielding 520 comments. Three trained student annotators independently assigned each comment to one of three sentiment classes: positive, neutral, or negative. Only textual content was processed, and user identities were not retained, preserving anonymity(Škoda & Guzoňová, 2026).

Text preprocessing

Each comment passed through four preprocessing stages (Laksana et al., 2022; Hakim et al., 2024). Cleaning removed non-essential characters, symbols, and formatting artifacts. Case folding converted all text to lowercase for consistency. Stopword removal eliminated common Indonesian function words that carry little sentiment information. Stemming reduced inflected words to their root form using an Indonesian stemmer. Table 3 gives a worked example of the four stages.

Feature extraction (FastText embedding)

Tokens were first mapped to integer indices, and sequences were padded to a fixed length to match the LSTM input format. Each index was then mapped to a 300-dimensional vector using the pre-trained Indonesian FastText model (cc.id.300.vec). FastText (Bojanowski et al., 2017) represents words through subword (character n-gram) information, which is advantageous for a morphologically rich language such as Indonesian and allows it to produce useful vectors even for out-of-vocabulary words (Nasution et al., 2024; Wirsching et al., 2025). Each comment was thus represented as a fixed-size embedding matrix serving as the LSTM input layer; sentiment labels were one-hot encoded.

LSTM classification model

The LSTM performs sequence classification. Its ability to retain long-range dependencies arises from three interacting gates: the forget gate (f_t), the input gate (i_t), and the output gate (o_t), defined in Equations (1)–(3) (Hochreiter & Schmidhuber, 1997; Tripathi, 2021):

$$f_t = \sigma(W\_f \cdot [h_{t-1}, x_t] + b\_f) \quad (1)$$

$$i_t = \sigma(W\_i \cdot [h_{t-1}, x_t] + b\_i) \quad (2)$$

$$o_t = \sigma(W\_o \cdot [h_{t-1}, x_t] + b\_o) \quad (3)$$

where σ is the sigmoid function, W denotes weight matrices, b denotes bias vectors, h_{t-1} is the previous hidden state, and x_t is the current input. The gating structure is illustrated in Fig. 2. The candidate cell state ($\tilde{c}_t$), the updated cell state (c_t), and the hidden state (h_t) are given by Equations (4)–(6):

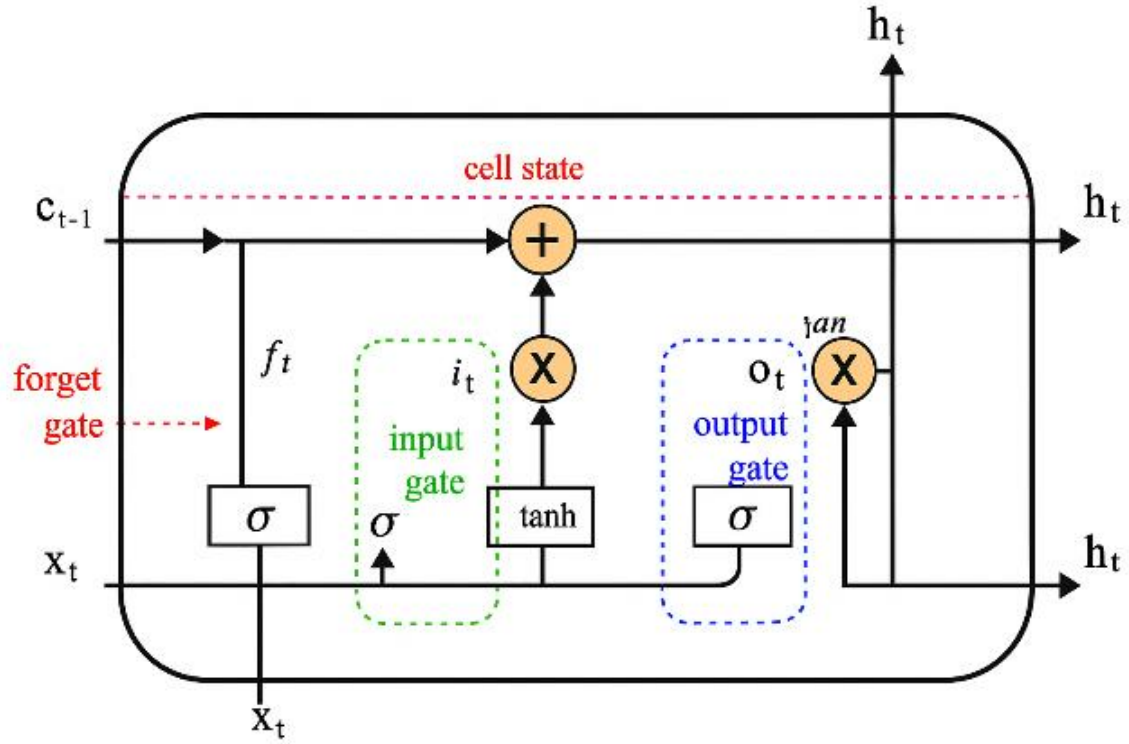

Figure 2. Internal architecture of an LSTM cell (forget, input, and output gates).

$$\tilde{c}_t = tanh(W\_c \cdot [h_{t-1}, x_t] + b\_c) \quad (4)$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t \quad (5)$$

$$h_t = o_t \cdot tanh(c_t) \quad (6)$$

Model evaluation and metrics

Model performance was quantified with a confusion matrix in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Table 1 shows the confusion-matrix structure for a single class treated as "positive" in a one-vs-rest formulation of the multi-class problem(Alzubaidi et al., 2021; Thakur et al., 2024).

Table 1. Confusion-matrix structure (one-vs-rest formulation).

	Predicted Positive	**Predicted Negative**
Actual Positive	TP	FN
Actual Negative	FP	TN

From the confusion matrix, accuracy, precision, recall, and F1-score were computed as in Equations (7)–(10). Accuracy is the overall proportion of correctly classified comments; precision is the proportion of predicted-positive comments that are truly positive; recall is the proportion of truly positive comments that are correctly retrieved; and F1-score is the harmonic mean of precision and recall, which is informative under class imbalance.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (7)$$

$$Precision = TP / (TP + FP) \quad (8)$$

$$Recall = TP / (TP + FN) \quad (9)$$

$$F1\text{-}score = 2 \cdot (Precision \cdot Recall) / (Precision + Recall) \quad (10)$$

Because the three classes are imbalanced, both macro-averaged and weighted-averaged metrics are reported, together with per-class precision, recall, and F1-score. Macro averaging weights each class equally and therefore penalizes failure on minority classes, whereas weighted averaging is dominated by the majority class; reporting both prevents a single aggregate figure from masking class-level collapse(Emekci & Roxas, 2026). Model variants (data-splitting schemes and configurations) are compared descriptively on the test partitions; no inferential statistical test is applied, consistent with the exploratory scope of the study.

Topic modeling (LDA)

To complement the sentiment analysis, Latent Dirichlet Allocation (LDA) was used to uncover latent themes in the corpus (Blei et al., 2003; Agus & Er, 2021; Karmila et al., 2022). LDA is an unsupervised generative model that identifies groups of words that co-occur in similar contexts. It requires two main choices: the number of topics and the number of representative keywords per topic. Each document is modeled as a mixture of topics, and each topic as a distribution over words; the fitted model thus provides a thematic summary of the corpus and groups comments by dominant theme(Ma et al., 2025).

RESULTS AND DISCUSSION

Result

Dataset

The dataset comprises 520 Indonesian YouTube comments about the RUU TNI, each labeled negative, neutral, or positive. As shown in Table 2, the corpus is imbalanced: 220 comments (42.31%) are negative, 161 (30.96%) are neutral, and 139 (26.73%) are positive.

Table 2. Dataset composition by sentiment class.

Sentiment	Count	Percentage
Positive	139	26.73%
Neutral	161	30.96%
Negative	220	42.31%
Total	**520**	**100.00%**

The distribution is visualized in Fig. 3. Negative comments clearly dominate, followed by neutral, with positive comments least frequent. This imbalance indicates that most sampled users express disapproval of, or concern about, the RUU TNI; it also foreshadows the classification difficulty analyzed in Section 3.5, because the majority-class share (42.31%) sets a strong baseline that any useful classifier must exceed.

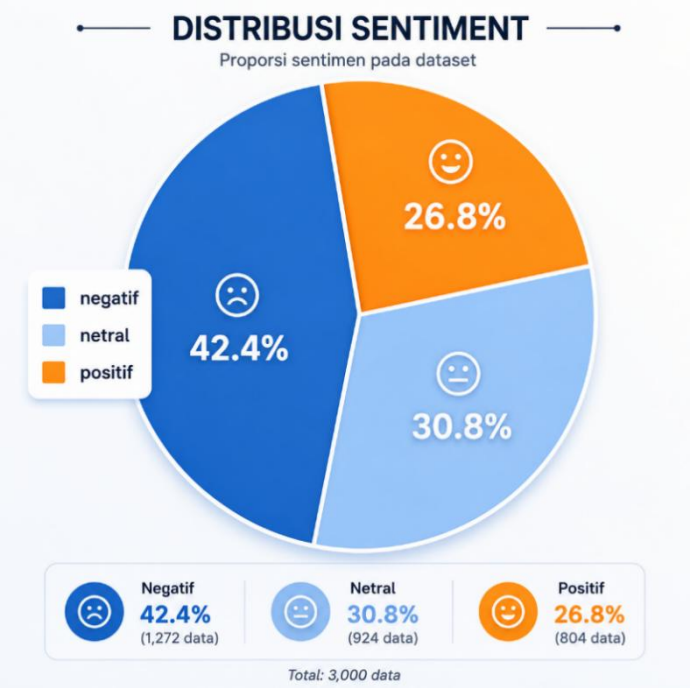

Figure 3. Sentiment distribution of YouTube comments on the RUU TNI.

Text Preprocessing

Before modeling, every comment was cleaned, case-folded, stopword-filtered, and stemmed. Table 3 illustrates the four stages on a representative comment: cleaning removes irrelevant punctuation and symbols; case folding standardizes capitalization; stopword removal discards uninformative function words such as "dan", "yang", and "di"; and stemming maps inflected forms to their roots (for example, "disahkan" → "sah" and "berharap" → "harap"). The process reduces noise and leaves a more compact, structured token sequence for the model.

Table 3. Example output at each text-preprocessing stage.

Stage	Input	Output
Cleaning	Kalau undang-undang gercep disahkan tapi UU ampas dan jangan berharap dprnya bubar	Kalau undang undang gercep disahkan tapi UU ampas dan jangan berharap dprnya bubar
Case folding	Kalau undang undang gercep disahkan tapi UU ampas dan jangan berharap dprnya bubar	kalau undang undang gercep disahkan tapi uu ampas dan jangan berharap dprnya bubar
Stopword removal	kalau, undang, undang, gercep, disahkan, tapi, uu, ampas, dan, jangan, berharap, dprnya, bubar	kalau, undang, gercep, disahkan, tapi, uu, ampas, jangan, berharap, dprnya, bubar
Stemming	kalau, undang, gercep, disahkan, tapi, uu, ampas, jangan, berharap, dprnya, bubar	kalau, undang, gercep, sah, tapi, uu, ampas, jangan, harap, dpr, bubar

FastText Feature Extraction

After preprocessing, tokens were converted to numerical form. Each comment was represented as a fixed-size matrix whose rows are 300-dimensional FastText vectors, ready for the LSTM. To inspect the embedding qualitatively, the vectors of several issue-related keywords were projected to two dimensions with Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE; van der Maaten & Hinton, 2008), as shown in Fig. 4.

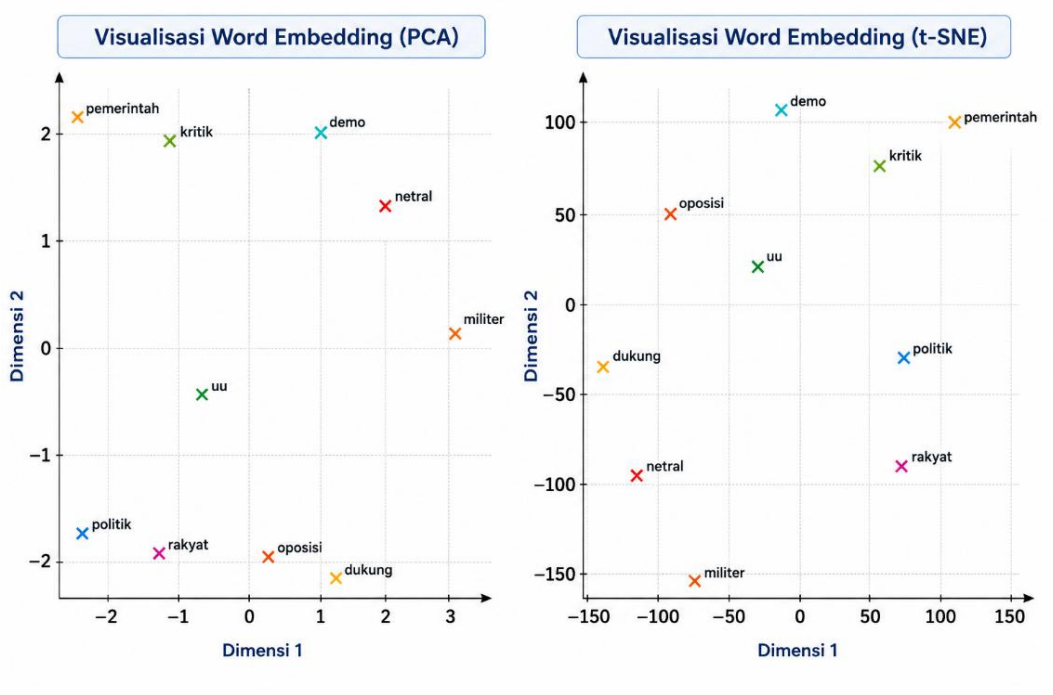

Figure 4. Two-dimensional projection of FastText word embeddings using PCA (left) and t-SNE (right).

In Fig. 4, semantically related terms tend to lie close together while unrelated terms are more distant, indicating that the FastText embedding captures basic semantic structure for the domain vocabulary. This provides a reasonable input representation for the LSTM; as the classification results below show, however, an informative embedding alone is not sufficient to overcome the constraints imposed by the small, imbalanced training set.

LSTM Classification Results

The FastText features were used to train the LSTM under three data-splitting schemes: hold-out (80:20), K-fold cross-validation (k = 5), and stratified K-fold cross-validation (k = 5). Stratification preserves the class proportions in every fold, which is important under imbalance. An overfitting model would show low training error but high generalization error and fail on unseen data (Girelli Consolaro et al., 2023); here, however, the dominant problem is underfitting to the minority classes rather than overfitting. The accuracies are summarized in Table 4.

Table 4. LSTM accuracy under the three data-splitting schemes.

Data split	Accuracy
Hold-out (80:20)	41.82%
K-fold cross-validation (k = 5)	42.40%
Stratified K-fold (k = 5)	42.39%

All three schemes yield accuracies of 41–42%, with the best value (42.40%) from K-fold. These values are statistically indistinguishable from the majority-class baseline of 42.31%, the accuracy obtained by always predicting “negative”. None of the configurations classifies better than a trivial constant predictor. This equivalence with the baseline, rather than a merely low accuracy, is the central result, and Section 3.5 examines it through the confusion matrix.

Figure 5 shows the training and validation accuracy across epochs for the stratified K-fold scheme. Both curves rise slightly in the first few epochs and then plateau close to 0.42–0.43, which matches the majority-class share. The small gap between training and validation accuracy shows that the model does not overfit. Instead it converges to a degenerate solution that predicts the majority class and cannot separate the minority classes.

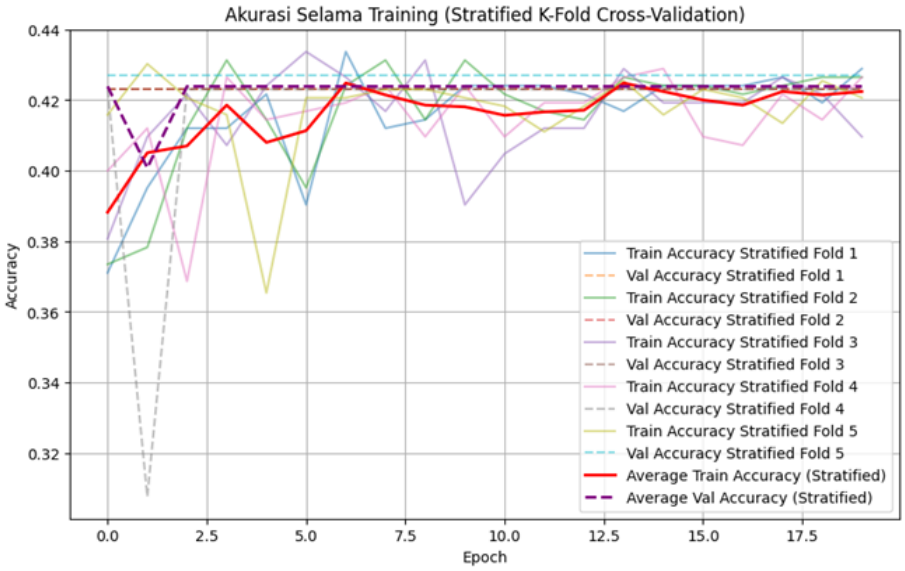

Figure 5. Training and validation accuracy per fold under stratified K-fold cross-validation.

Model Evaluation

Figure 6 shows the three-class confusion matrix (negative, neutral, positive). Almost every test comment, whether truly negative, neutral, or positive, is predicted as negative. None of the neutral comments and none of the positive comments are classified correctly, and the model recovers only the negative class.

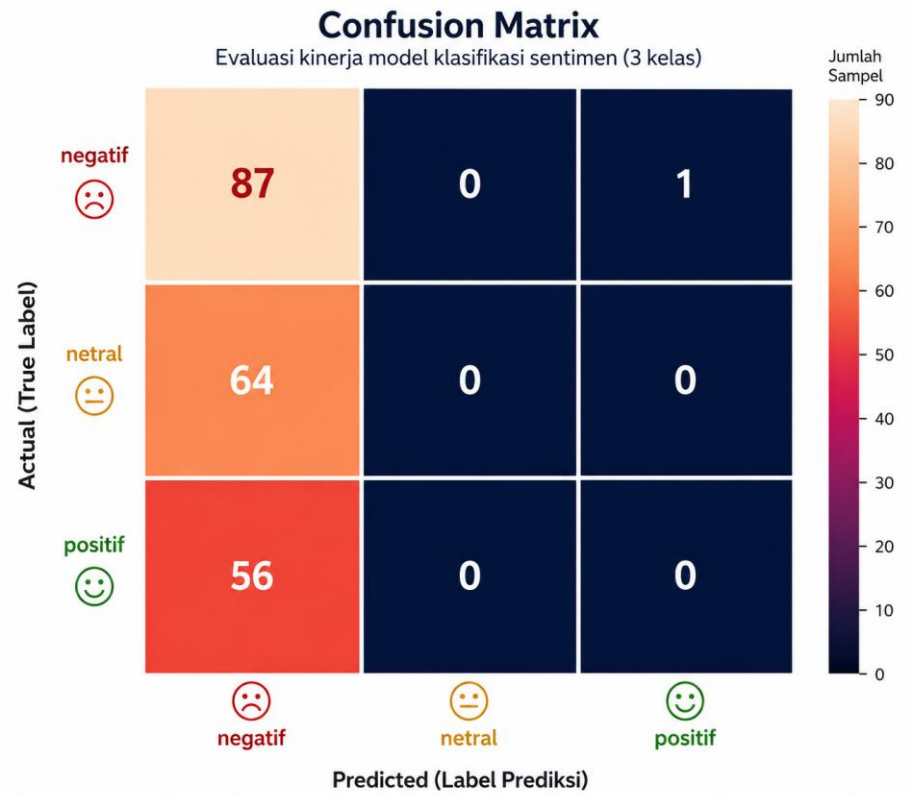

Figure 6. Confusion matrix of the LSTM classifier for the three sentiment classes.

The counts in Fig. 6 make the collapse explicit. Reading directly from the matrix, 87 of 88 negative comments are predicted negative (per-class recall ≈ 98.9%), while 0 of 64 neutral and 0 of 56 positive comments are correctly identified (per-class recall = 0% for both). Table 5 reports the per-class metrics computed from this matrix alongside macro and weighted averages.

Table 5. Per-class and averaged evaluation metrics computed from the confusion matrix in Fig. 6.

Class	Precision	Recall	F1-score	Support
Negative	42.0%	98.9%	59.0%	88
Neutral	0.0%	0.0%	0.0%	64
Positive	0.0%	0.0%	0.0%	56
Macro avg.	14.0%	33.0%	19.7%	208
Weighted avg.	17.8%	41.8%	25.0%	208
Accuracy	–	–	41.8%	208

These per-class figures explain why a single aggregate number is misleading. The negative class attains an F1-score of about 59% because the classifier predicts "negative" almost everywhere, whereas the neutral and positive classes attain an F1-score of 0%. A weighted or majority-driven average can therefore report an F1 near 60% even though the classifier has learned nothing about two of the three classes. The macro-averaged F1 of roughly 20% is the accurate summary of multi-class performance, and the accuracy of 41.8% reproduces the majority-class baseline.

Topic Modeling

LDA was fitted to the corpus with five topics; the top keywords and their weights are listed in Table 6. Although the classifier failed, the topic model still provides an informative, model-agnostic view of the discourse because it does not depend on the sentiment labels.

Table 6. Top LDA keywords (with weights) for each of the five topics.

Topic	Top keywords (weight)
Topic 0	koruptor (0.039), uu (0.038), ampas (0.032), aset (0.031), rakyat (0.023), jadi (0.021), nya (0.018), dpr (0.017), apa (0.017), takut (0.016)
Topic 1	rakyat (0.045), tni (0.022), banyak (0.019), negara (0.016), dpr (0.015), mahasiswa (0.014), nya (0.012), lebih (0.012), wakil (0.011), mana (0.011)
Topic 2	tni (0.026), dpr (0.022), pilih (0.019), negara (0.017), undang (0.017), kan (0.015), rakyat (0.015), siapa (0.015), apa (0.014), buat (0.014)
Topic 3	aset (0.041), ampas (0.036), rakyat (0.034), uu (0.033), dpr (0.031), tni (0.030), ruu (0.022), apa (0.016), buat (0.014), negara (0.013)
Topic 4	tni (0.034), jadi (0.022), rakyat (0.018), dpr (0.016), negara (0.016), bubar (0.016), indonesia (0.013), undang (0.012), korupsi (0.012), baru (0.011)

The topics are thematically coherent. Topic 0 ("koruptor", "uu", "ampas", "aset") expresses sharp criticism of the bill's quality and of corruption. Topics 1 and 2 ("rakyat", "tni", "negara", "dpr") center on the roles of the military, parliament, and citizens in the structure of power. Topics 3 and 4 ("aset", "ampas", "bubar", "korupsi", "indonesia") reflect fears about the bill's impact on state assets, institutional integrity, and the future of democracy.

Figure 7 shows the corpus-level weight of each topic. Topic 3 is most prominent, followed by Topics 0, 4, 1, and 2. The dominance of criticism directed at the bill's quality and its relation to state assets and the legislature is consistent with the strongly negative sentiment distribution in Section 3.1.

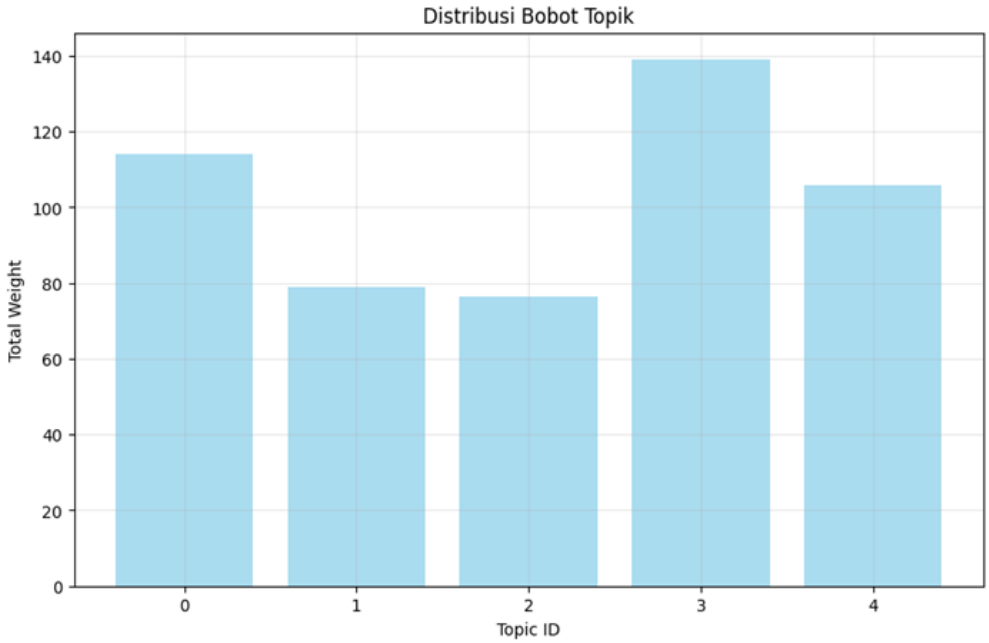

Figure 7. Corpus-level topic-weight distribution from LDA.

Discussion

The two findings should be interpreted separately. The first concerns public opinion. More than 40% of the comments are negative and only about a quarter are positive. Combined with the LDA themes (criticism of the "UU ampas", distrust of "koruptor", concern over "aset negara", and calls to disband or criticize the DPR), this indicates a public discourse dominated by distrust of the legislative process and political institutions. Because it does not depend on the classifier, this descriptive result is robust.

The second finding concerns the classifier. The FastText–LSTM model did not work: its accuracy equals the majority-class baseline, and the confusion matrix shows a complete collapse onto the negative class, with no correct predictions for the neutral and positive classes. The data conditions account for this outcome. LSTMs typically require thousands to tens of thousands of labeled examples to estimate their recurrent weights reliably. A corpus of 520 comments, split further for validation, provides too few minority-class examples for the network to learn a decision boundary, so gradient descent settles on the trivial majority-class solution. The strong class imbalance compounds the problem, and the linguistic complexity of political discourse, including sarcasm, irony, slang, and implicit stance, makes the minority classes especially hard to separate. The near-zero gap between training and validation accuracy in Fig. 5 confirms that the failure reflects underfitting rather than overfitting.

Relative to prior Indonesian sentiment studies that report high accuracy, this study's low performance is not contradictory but complementary: those studies typically use larger and more balanced datasets or binary tasks, whereas the present three-class, small, imbalanced, politically sensitive setting is substantially harder. The contribution is therefore to quantify that difficulty concretely and to establish the majority-class baseline that future models on this dataset must beat.

The findings carry several implications. For policymakers, the dominance of negative sentiment and the recurring critical themes can serve as an early, low-cost signal of public resistance in policy communication and in reconsidering the substance of the regulation, although the small, single-platform sample makes these signals indicative rather than representative. For civil-society actors, the analysis shows that YouTube comments can serve as a fast and inexpensive channel for monitoring public opinion. For NLP researchers, the results show that small, imbalanced political datasets require explicit imbalance handling, for example SMOTE (Chawla et al., 2002), and stronger contextual models such as IndoBERT (Wilie et al., 2020) rather than an off-the-shelf LSTM.

The study has clear limitations. The sample is small (520 comments) and drawn from a single platform in a single language, so it does not represent the full range of public opinion; other platforms (X, Facebook, Instagram) and regional languages are not covered. The corpus is imbalanced, which biases the classifier toward the majority class. Only the FastText–LSTM combination was explored; stronger contextual models (BERT, IndoBERT, or Transformer variants) were not tested. Finally,

pragmatic phenomena such as sarcasm and irony are likely mishandled, further depressing classification performance. Inter-annotator agreement was not reported, so label reliability cannot yet be assessed.

Implications

The findings have practical implications for policymakers, civil-society organizations, and researchers. For policymakers, the dominance of negative sentiment and the recurring critical themes may provide an early and relatively low-cost indication of public concerns regarding policy communication and legislative decision-making. However, because the dataset is limited to a relatively small sample from a single social-media platform, these findings should be interpreted as indicative signals rather than as representative measurements of the entire Indonesian population.

For civil-society organizations and public-opinion observers, the study demonstrates that YouTube comments can provide a rapid source of information for monitoring emerging public concerns surrounding socio-political issues. Combining sentiment analysis with topic modeling can help identify not only whether public responses are predominantly positive or negative but also the substantive issues underlying those responses.

For NLP researchers, the findings emphasize the importance of considering dataset size and class distribution before selecting deep-learning architectures. Small and imbalanced political datasets require explicit strategies for handling minority classes, such as class weighting, focal loss, or resampling techniques. The results also suggest that stronger contextual language models, particularly Indonesian language models such as IndoBERT, may provide more appropriate representations for complex political discourse than a conventional FastText–LSTM configuration.

Research Contribution

This study makes three primary contributions. First, it provides an empirical characterization of public opinion regarding the RUU TNI based on a manually labeled dataset of 520 Indonesian YouTube comments. The analysis identifies a predominance of negative sentiment and reveals several recurring themes associated with criticism of the bill, distrust of political institutions, and concerns about democratic and institutional consequences.

Second, the study proposes a reproducible end-to-end analytical workflow integrating data collection, manual sentiment labeling, text preprocessing, FastText embedding, LSTM classification, and LDA topic modeling. The integration of sentiment classification and topic analysis provides both quantitative and thematic perspectives on public discourse.

Third, the study contributes a cautionary empirical finding regarding the limitations of deep-learning classification under constrained data conditions. The results demonstrate that, without sufficient training data and appropriate class-imbalance handling, the FastText–LSTM model may converge toward the majority class and fail to distinguish minority sentiment categories. This negative result is itself a relevant contribution because it establishes an empirical baseline and provides evidence that future models applied to this domain must exceed.

Limitations

Several limitations should be considered when interpreting the findings. First, the dataset consists of only 520 comments collected from YouTube. Consequently, the findings cannot be generalized to represent the entire Indonesian population or the full diversity of online public opinion. Other major platforms, such as X, Facebook, and Instagram, as well as discussions expressed in regional languages, were not included.

Second, the dataset is imbalanced, with negative comments representing the largest class. This imbalance likely contributed to the classifier's tendency to predict the majority class and reduced its ability to identify neutral and positive comments. The study did not implement dedicated imbalance-handling techniques such as oversampling, class weighting, focal loss, or synthetic data generation.

Third, only the FastText–LSTM combination was evaluated for sentiment classification. More advanced contextual architectures, including Transformer-based models and IndoBERT, were not examined. Therefore, the results should not be interpreted as evidence that deep learning is generally ineffective for Indonesian political sentiment analysis, but rather that the specific model configuration used in this study was inadequate under the available data conditions.

Finally, political discourse frequently contains sarcasm, irony, slang, implicit stance, and other pragmatic expressions that may not be captured effectively by the present model. In addition, inter-

annotator agreement was not reported, which limits the ability to assess the reliability and consistency of the manual sentiment labels.

Suggestions

Future research should first expand the dataset substantially and improve the balance among positive, neutral, and negative sentiment classes. Data may be collected from multiple social-media platforms to capture a broader range of public discourse and improve the generalizability of the findings. A larger dataset would also provide sufficient training examples for more complex deep-learning architectures.

Second, future studies should implement explicit class-imbalance handling techniques, including class weighting, focal loss, undersampling, oversampling, or resampling methods such as SMOTE where appropriate. Model evaluation should continue to report per-class metrics and macro-averaged scores so that strong performance on the majority class does not obscure poor performance on minority classes.

Third, future work should compare the FastText–LSTM approach with stronger contextual models, such as BiLSTM with attention, BERT-based architectures, and IndoBERT. Comparative experiments using identical datasets and evaluation protocols would provide clearer evidence regarding the most suitable approach for Indonesian political sentiment analysis.

Finally, future studies should strengthen the reliability of the dataset and topic analysis by reporting inter-annotator agreement and topic-coherence measures. Incorporating aspect-based sentiment analysis and temporal analysis may also provide deeper insights into how specific policy issues and public attitudes evolve over time. These improvements could support the development of a more reliable and scalable public-opinion monitoring system for policy debates in Indonesia.

CONCLUSION

This study mapped public opinion on the RUU TNI in 520 Indonesian YouTube comments. The discourse in the sample is strongly negative (42.31% negative, 30.96% neutral, 26.73% positive), and LDA identified five recurring themes dominated by criticism of the bill's quality, distrust of the DPR, concern over state assets, and fear for democratic and institutional integrity. Public reaction to the bill in this sample is thus predominantly critical and apprehensive.

As a classifier, the FastText–LSTM pipeline did not succeed. Across hold-out, K-fold, and stratified K-fold schemes, accuracy remained at 41–42%, matching the majority-class baseline of 42.31%, and the confusion matrix shows a collapse onto the negative class with no correct predictions for the neutral or positive classes. A high aggregate F1-score reflects this majority-class collapse rather than balanced performance, since the macro-averaged F1 is only about 20%. The study therefore reports a cautionary baseline that quantifies how demanding the task is under small, imbalanced data.

The study contributes a labeled, domain-specific Indonesian dataset on the RUU TNI, a reproducible sentiment-plus-topic workflow, and empirical evidence that meaningful classification in this domain requires more and better-balanced data and stronger models. Future work should enlarge and balance the dataset, apply explicit class-imbalance techniques (class weighting, focal loss, or resampling such as SMOTE; Chawla et al., 2002), and adopt contextual models such as BiLSTM with attention or IndoBERT (Wilie et al., 2020). Extending the analysis to additional platforms would further support a practical public-opinion-monitoring system for policy debates in Indonesia. Reporting inter-annotator agreement and topic-coherence scores would strengthen the reliability of future studies on this dataset.

REFERENCES

Agus, N., & Er, S. (2021). Implementasi Latent Dirichlet Allocation (LDA) untuk klasterisasi cerita berbahasa Bali. Jurnal Teknologi Informasi dan Ilmu Komputer, 8(1), 127–134. https://doi.org/10.25126/jtiik.0813556

Agustino, L., & Hikmawan, M. D. (2026). How powerful are village heads? Clientelism in the reform era of Indonesia's democracy. *Frontiers in Political Science*, *8*. https://doi.org/10.3389/fpos.2026.1747452

Airlangga, U., & Adiningsih, T. E. (2023). CRYPTO TAXATION AND FINANCIAL REPORTING IN THE CONTEXT OF PANCASILA. *Jurnal Akuntansi Multiparadigma*, *14*(3).

https://doi.org/10.21776/ub.jamal.2023.14.3.34
Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A. Q., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal Of Big Data*, *8*(1), 53. https://doi.org/10.1186/s40537-021-00444-8
Aspinall, E., & Hicken, A. (2019). Guns for hire and enduring machines: clientelism beyond parties in Indonesia and the Philippines. *Democratization*, *27*(1), 137–156. https://doi.org/10.1080/13510347.2019.1590816
Berenschot, W., & Aspinall, E. (2019). How clientelism varies: comparing patronage democracies. *Democratization*, *27*(1), 1–19. https://doi.org/10.1080/13510347.2019.1645129
Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022. https://dl.acm.org/doi/10.5555/944919.944937
Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135–146. https://doi.org/10.1162/tacl_a_00051
Brynjolfsson, E., Li, D., & Raymond, L. (2024). Generative AI at Work. *The Quarterly Journal of Economics*, *140*(2), 889–942. https://doi.org/10.1093/qje/qjae044
Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953
Djuyandi, Y., Nugraha, K. P., Sugiharto, A., Jamri, M. H., & Ridzuan, A. R. (2025). Normalizing the illegal: public perceptions of vote-buying in West Bandung Regency and its democratic consequences. *Frontiers in Political Science*, *7*. https://doi.org/10.3389/fpos.2025.1630335
Emekci, H., & Roxas, D. Q. (2026). Dissecting Medical RAG: Why Reranking Matters More Than Complexity in Question Answering. *Black Sea Journal of Engineering and Science*, *9*(2), 549–561. https://doi.org/10.34248/bsengineering.1849342
Fernando, H., Larasati, Y. G., Abdullah, I., & Horáková, H. (2024). Leadership and the Money Politics Trap in Islamic Legal Thought: A Case Study of Indonesia as a Muslim-Majority Country. *Nurani Jurnal Kajian Syari Ah Dan Masyarakat*, *24*(1), 199–214. https://doi.org/10.19109/s79vb707
Girelli Consolaro, N., Shinde, S. S., Naseh, D., & Tarchi, D. (2023). Analysis and performance evaluation of transfer learning algorithms for 6G wireless networks. Electronics, 12(15), 3327. https://doi.org/10.3390/electronics12153327
Hakim, G., Fatyanosa, T. N., & Widodo, A. W. (2024). Analisis sentimen masyarakat terhadap kereta cepat Whoosh pada platform X menggunakan IndoBERT. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 8(10), 1–10. http://repository.ub.ac.id/id/eprint/235285
Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
Isnayni, B. N., Saputra, N., & Hastono, T. (2024). Sentiment analysis of coffee shop reviews using Random Forest classifier method. JTH: Journal of Technology and Health, 1(4), 233–244. https://doi.org/10.61677/jth.v2i2.152
Karmila, S., & Ardianti, V. I. (2022). Metode Latent Dirichlet Allocation untuk menentukan topik teks suatu berita. Jurnal Informatika dan Komputasi: Media Bahasan, Analisa dan Aplikasi, 16(1), 36–44. https://doi.org/10.56956/jiki.v16i01.100
Laksana, M. D. B., Karyawati, A. E., Putri, L. A. A. R., Santiyasa, I. W., Er, N. A. S., & Kadnyanan, I. G. A. G. A. (2022). Text summarization terhadap berita bahasa Indonesia menggunakan dual encoding. JELIKU (Jurnal Elektronik Ilmu Komputer Udayana), 11(2), 339–348. https://doi.org/10.24843/jlk.2022.v11.i02.p13
Ma, T., La, T.-T., Huu, L.-T. Le, Nguyen, M.-N., & Luu, K.-V. P. (2025). REBot: From RAG to CatRAG with Semantic Enrichment and Graph Routing. *Lecture Notes in Computer Science*, 435–447. https://doi.org/10.1007/978-981-95-4960-3_35
Nasution, N. A., Nababan, E. B., & Mawengkang, H. (2024). Comparing LSTM algorithm with word embedding: FastText and Word2Vec in Bahasa Batak–English translation. In 2024 12th International Conference on Information and Communication Technology (ICoICT) (pp. 306–313). IEEE. https://doi.org/10.1109/ICoICT61617.2024.10698481

Nidhi, Singh, O., Prince, & Garg, M. (2024). A lightweight LSTM framework for contextual sentiment classification. International Journal of Research – GRANTHAALAYAH, 12(6), 147–159. https://doi.org/10.29121/granthaalayah.v12.i6.2024.6095

Putri, S. B., Anisa, Y. N., & Saputra, N. (2022). Analisis sentimen film Kuliah Kerja Nyata (KKN) di Desa Penari menggunakan metode Naive Bayes. JuSiTik: Jurnal Sistem dan Teknologi Informasi Komunikasi, 5(2), 22–26. https://doi.org/10.32524/jusitik.v5i2.704

Sadat, A., & Basir, M. A. (2025). Clientelism, coalitions, and concessions: pragmatism in local elections and its democratic costs. *Frontiers in Political Science*, *7*. https://doi.org/10.3389/fpos.2025.1664786

Saputra, N. (2018). Analisis sentimen dengan preprocessing kata. Jurnal Dinamika Informatika, 7(1), 45–57. file:///C:/Users/USER/Downloads/nurirwan,+4.+Nurirwan+Saputra+(ANALISIS+SENTIMEN+DENGAN+PREPROCESSING+KATA).pdf

Saputra, N. (2019). Analisis sentimen dengan menggunakan metode klasifikasi Lazy K-Star. Seri Prosiding Seminar Nasional Dinamika Informatika, 1(1). https://scholar.google.com/citations?view_op=view_citation&hl=en&user=48NFQPAAAAAJ&cstart=20&pagesize=80&citation_for_view=48NFQPAAAAAJ:qjMakFHDy7sC

Saputra, N., Adji, T. B., & Permanasari, A. E. (2015). Analisis sentimen data Presiden Jokowi dengan preprocessing normalisasi dan stemming menggunakan metode Naive Bayes dan SVM. Jurnal Dinamika Informatika, 5(1), 1–12. https://scholar.google.com/citations?view_op=view_citation&hl=en&user=48NFQPAAAAAJ&citation_for_view=48NFQPAAAAAJ:u-x6o8ySG0sC

Saputra, N., Nurbagja, K., & Turiyan, T. (2022). Sentiment analysis of presidential candidates Anies Baswedan and Ganjar Pranowo using Naïve Bayes method. Jurnal Sisfotek Global, 12(2), 114–119. https://doi.org/10.38101/sisfotek.v12i2.552

Saputra, N., Riyadi, A., & Tentua, M. N. (2023). Sentiment analysis of COVID vaccination policy in Indonesia using Random Forest (pp. 205–209). https://doi.org/10.2991/978-94-6463-338-2_31

Škoda, M., & Guzoňová, V. (2026). Formal Harmonization, Persistent Accounting Uncertainty: Practitioner Evidence on Crypto-Asset Valuation After MiCA in Slovakia. *FinTech*, *5*(3), 65. https://doi.org/10.3390/fintech5030065

Thakur, A. S., Choudhary, K., Ramayapally, V. S., Vaidyanathan, S., & Hupkes, D. (2024). Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2406.12624

Topić, P. (2026). Digital Transformation of Tax Administration in Small EU Member States: Organisational Capacity, the Information Gap, and the Pre-DAC8 Baseline from Croatia (2020–2024). *Organizacija*, *59*(3), 223–240. https://doi.org/10.2478/orga-2026-0014

Tripathi, M. (2021). Sentiment analysis of Nepali COVID-19 tweets using NB, SVM and LSTM. Journal of Artificial Intelligence and Capsule Networks, 3(3), 151–168. https://doi.org/10.36548/jaicn.2021.3.001

Udayana, I. K. A. P. A. N., Mahendra, I. B. M., et al. (2023). Analisis sentimen opini berbahasa Indonesia pada sosial media menggunakan TF-IDF dan Support Vector Machine. JELIKU (Jurnal Elektronik Ilmu Komputer Udayana), 12(1), 45–52. https://garuda.kemdiktisaintek.go.id/documents/detail/3694249

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9, 2579–2605. https://jmlr.org/papers/v9/vandermaaten08a.html

Wicaksono, B., Rahmayanti, V., & Nastiti, S. (2024). Analisis sentimen dalam opini publik di channel YouTube Indonesia Lawyers Club tentang isu populer dengan menggunakan metode LSTM dan Bi-LSTM. Jurnal Algoritma, 21(2), 241–251. https://doi.org/10.33364/algoritma/v.21-2.1696

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (pp. 843–857). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.aacl-main.85

Wirsching, E. M., Rodriguez, P. L., Spirling, A., & Stewart, B. M. (2025). Multilanguage word embeddings for social scientists: Estimation, inference, and validation resources for 157 languages. Political Analysis, 33(2), 156–163. https://doi.org/10.1017/pan.2024.17

Yuniningsih, T., Harahap, T. K., Lestari, A. W., Herawati, A. R., Hertati, D., Djumiarti, T., Lituhayu, D., Mulyani, S., Suryani, D. A., Siswanto, D., Suyatno, M., Nirmala, R. J., Suwitri, S., Rahmawati, T., Mutmainah, N. F., Rostyaningsih, D., Khozin, M., Rahman, A. Z., Larasati, E., … Hartini. (2024). *Etika Administrasi Sektor Publik*.