

Sentiment Analysis Using Machine Learning for Evaluating Public Perception to Support Accountability in Vocational Higher Education

Putu Satya Saputra^{1*}, I Putu Gede Abdi Sudiarmika², Ni Ketut Utami Dewi³

^{1,2,3} Politeknik Negeri Bali, Badung, Indonesia

ARTICLE INFO

Article history:

Initial submission
04-08-2026
Received in revised form
30-08-2026
Accepted 18-09-2026
Available online 30-09-2026

Keywords:

*sentiment analysis,
machine learning, public
perception, human
annotation, accountability*

DOI:

<https://doi.org/10.59356/smart-techno.v8i2.240>

ABSTRACT

This study examines public sentiment toward vocational higher education in Indonesia using 2,014 YouTube comments from 79 videos and evaluates Naïve Bayes, Support Vector Machine (SVM), and Random Forest classifiers against human-validated labels. A vocational higher-education lecturer and an Indonesian-language lecturer annotated all comments independently. They initially agreed on 1,958 comments (97.22%; Cohen's $\kappa = 0.945$), and resolved 56 disagreements by consensus. The final labels comprised 1,227 neutral (60.92%), 702 positive (34.86%), and 85 negative comments (4.22%). A stratified 80/20 split was created before any balancing; the 403-comment test set retained the natural distribution. On the training partition, unigram versus unigram, bigram features and no balancing, class weighting, or training-only undersampling were compared using repeated stratified five-fold cross-validation (two repeats). SVM with unigram TF-IDF and balanced class weights achieved the best cross-validated macro-F1 among the three selected model configurations (0.650 ± 0.054). On the untouched test set, SVM obtained 76.67% accuracy and 0.655 macro-F1, and recovered 6 of 17 negative comments (35.29% recall). Random Forest produced slightly higher accuracy (77.17%) but only 5.88% negative recall, while Naïve Bayes detected no negative cases. The results provide a reproducible screening pipeline that may complement institutional accountability mechanisms, but the limited recovery of negative feedback precludes its use as an autonomous early-warning system.

1. INTRODUCTION

Digital platforms provide continuously produced, unsolicited accounts of stakeholder questions, experiences, and evaluations. YouTube is particularly relevant because a video supplies topical context while its comment thread records public reactions. Such data are accessible and timely, but they are shaped by platform search, channel popularity, and self-selection and are not representative opinion polls (Chan et al., 2025; Efstratiou, 2025; Li & Ullah, 2025).

For vocational higher education, public perceptions matter because institutions are expected to demonstrate practical learning, industry alignment, graduate employability, and responsiveness to stakeholders. International TVET strategy likewise emphasizes relevance, inclusion, and successful transitions to work (UNESCO, 2020). Yet vocational pathways may be viewed as a second choice relative to academic education, making public communication and evidence-informed service review especially important (Wang & Wang, 2023). Social-media commentary can therefore offer a supplementary form of crowd-based scrutiny, although institutional accountability itself requires action, explanation, and formal governance rather than sentiment scores alone (Karunakaran et al., 2022).

Sentiment analysis can organize large comment collections into positive, negative, and neutral categories, but Indonesian social-media text is difficult to classify because it contains slang, abbreviations, code-mixing, emojis, nonstandard spelling, implicit criticism, and sarcasm. Indonesian NLP resources have improved, but task- and domain-specific labelled

data remain important for credible evaluation (Koto et al., 2020). Labels generated by a sentiment lexicon cannot, by themselves, serve as independent ground truth for testing supervised models because the resulting evaluation measures agreement with the lexicon rather than agreement with human interpretation.

This revision addresses that problem by replacing lexicon-generated targets with labels independently assigned by two human annotators to the complete 2,014-comment dataset. It also separates the natural test set before balancing, fits TF-IDF only within training operations, evaluates class-specific and macro-averaged metrics, and uses repeated stratified cross-validation with feature and balancing ablations. The contribution is not a novel classifier; it is a human-validated and leakage-controlled evaluation of classical models in the underexamined context of Indonesian vocational higher education, coupled with a cautious account of how social-media evidence may complement accountability practices.

2. LITERATURE REVIEW

Sentiment Analysis and Social-Media Evidence

Sentiment analysis assigns evaluative orientation to text through lexicon-based, supervised machine-learning, deep-learning, or hybrid approaches. Lexicons are transparent and inexpensive but are sensitive to coverage, context, negation, and domain-specific wording; supervised classifiers can learn corpus-specific regularities but depend on the validity of their labels (Alahmadi et al., 2025; Cero et al., 2024). Comparative studies therefore distinguish lexicon output from human ground truth and use independent annotations for final evaluation (Qi & Shabrina, 2023).

YouTube-based perception studies demonstrate the scale and contextual richness of comment data, but they also show that the sample is determined by query formulation, collection time, platform ranking, and engagement thresholds. Transparent reporting of queries, dates, selection rules, and comment retrieval settings is therefore necessary to establish the scope of the evidence rather than imply population representativeness (Chan et al., 2025; Efstratiou, 2025).

Indonesian Informal Language and Text Representation

TF-IDF remains a useful, interpretable baseline for moderate-sized labelled corpora. Unigrams represent individual lexical cues, while bigrams can preserve short combinations such as negation–adjective or intensifier–adjective sequences. However, preprocessing choices can materially change the information available to both traditional and neural classifiers; aggressive stopword removal or stemming may erase sentiment-bearing function words and informal variants (Siino et al., 2024).

For Indonesian comments, conservative normalization is preferable when polarity depends on forms such as ‘tidak’, ‘gak’, ‘banget’, or ‘tapi’. Slang mapping can consolidate common variants, while explicit emoji tokens retain affective cues that would otherwise be discarded. Code-mixed words and unusual spellings remain a challenge because a word-level sparse representation cannot fully recover discourse context, irony, or implicit evaluation. Contextual Indonesian language models are a logical extension, but strong baselines and human-validated data are required before their gains can be assessed (Koto et al., 2020).

Imbalance, Evaluation, and Accountability

Naturally occurring sentiment data are often imbalanced. Accuracy and weighted averages can be dominated by the majority class, so macro-F1 and class-specific precision, recall, and F1 must accompany aggregate results (Altalhan et al., 2025; Tharwat, 2021). Any undersampling or class weighting must be confined to training data, and all learned transformations must be fitted without access to the test set. Repeated cross-validation on the training partition provides evidence about variation across folds while preserving a final untouched test set (Yates et al., 2023).

An accountability-oriented application places special emphasis on negative recall because missed criticism cannot be triaged for review. Even so, a high-recall model would not demonstrate institutional accountability by itself. At most, aggregated sentiment patterns can inform agenda setting, communication, and targeted qualitative investigation; responsible action still requires human review, triangulation with surveys and service indicators, and transparent reporting of model limitations (Karunakaran et al., 2022).

3. METHOD

Research Design and Data Collection

The study followed the workflow in Figure 1: YouTube data collection and cleaning, independent human annotation, a frozen stratified split, training-only model selection, and evaluation on an untouched test partition. The YouTube Data API collection was executed on 28 August 2026. The 20 exact search queries were: 'kampus vokasi Indonesia'; 'pendidikan vokasi Indonesia'; 'kuliah vokasi'; 'review politeknik'; 'pengalaman kuliah di politeknik'; 'kampus politeknik Indonesia'; 'mahasiswa politeknik'; 'pelayanan akademik politeknik'; 'fasilitas kampus politeknik'; 'kehidupan mahasiswa politeknik'; 'pendapat mahasiswa tentang politeknik'; 'perbedaan vokasi dan sarjana'; 'lulusan vokasi kerja'; 'kampus vokasi terbaik Indonesia'; 'politeknik negeri Indonesia'; 'pengalaman kuliah di kampus vokasi'; 'review pendidikan vokasi Indonesia'; 'kelebihan dan kekurangan vokasi'; 'lulusan politeknik kerja'; and 'apakah vokasi bagus'. The API returned up to 20 videos per query, resulting in 296 unique videos.

A video was eligible when 'vokasi' or 'politeknik' occurred in its title, description, or channel name and its API metadata reported at least five comments. Eligible videos were ordered by comment count and view count, limited to three per channel, and the first 80 were requested. For each video, at most seven pages of 100 top-level comments were retrieved in reverse chronological order; replies were not collected. The export contained 4,479 comments. HTML entities and URLs were cleaned, blank items and duplicates with the same video ID and text were removed ($n = 4,333$), comments shorter than five tokens were excluded ($n = 3,264$), and a maximum of 120 comments per video was retained. The final corpus comprised 2,014 comments from 79 videos and 66 channels. Comments were published from 12 May 2016 to 27 August 2026; video dates ranged from 15 June 2015 to 4 June 2026. The mean comment length was 17.10 words (median = 11; range = 5–381). No automated language detector or separate spam classifier was applied; Indonesian-dominant and code-mixed items that passed the stated rules remained in the corpus.

Revised Human-Validated Sentiment Classification Workflow

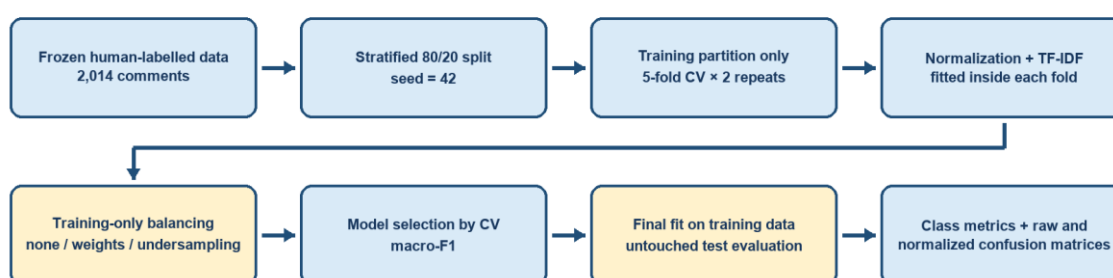


Figure 1. Human-validated, leakage-controlled model-development workflow

Independent Human Annotation and Adjudication

Two annotators independently labelled every comment: Annotator A was a vocational higher-education lecturer with domain familiarity, and Annotator B was an Indonesian-language lecturer with disciplinary expertise in Indonesian language. Both used the same written guideline and worked without access to lexicon scores, model predictions, or each

other's labels. Positive indicated explicit praise, satisfaction, support, hope, or beneficial evaluation; negative indicated explicit criticism, dissatisfaction, concern, rejection, or harmful evaluation; and neutral covered information, questions, mixed or indeterminate content without a dominant polarity. Context from the video title was available when necessary.

The annotators initially agreed on 1,958 of 2,014 items (97.22%). Cohen's kappa for the three nominal categories was 0.945, indicating very high chance-corrected agreement. All 56 disagreements were then reviewed jointly and resolved by consensus: 30 became neutral, 21 positive, and 5 negative. The resulting consensus label was frozen as the only target used for model training, selection, and evaluation.

Text Normalization and TF-IDF Representation

Normalization lowercased text, decoded HTML entities, replaced URLs, usernames, and numbers with stable tokens, reduced character repetitions of four or more to two characters, and removed remaining punctuation. Common variants were mapped to normalized forms: ga/gak/ngga/nggak/tdk→tidak, bgt/bgtt→banget, tp/tpi→tapi, krn→karena, dgn→dengan, yg→yang, sdh/udh→sudah, blm→belum, and klo/kalo→kalau. Positive, negative, and laughter emojis were converted to separate semantic tokens. Stopword removal and stemming were deliberately omitted to preserve negation, intensifiers, contrast markers, code-mixed words, and informal sentiment cues.

TF-IDF used lowercase=False after the custom normalizer, min_df=2, max_df=0.98, max_features=5,000, sublinear_tf=True, L2 normalization, and either unigram (1,1) or unigram–bigram (1,2) features. The vectorizer was an explicit pipeline step and was fitted separately inside every cross-validation fold; after configuration selection, it was refitted on the complete training partition and only then applied to the test comments.

Split, Models, Balancing, and Validation

The frozen dataset was divided with a stratified 80/20 split (random_state=42) before balancing or vectorization. Training contained 1,611 comments (68 negative, 981 neutral, 562 positive); the untouched test set contained 403 comments (17 negative, 246 neutral, 140 positive). Three classifiers were examined: Multinomial Naïve Bayes (alpha=1.0), linear SVM (LinearSVC, C=1.0), and Random Forest (200 trees, max_features='sqrt', min_samples_leaf=1, n_jobs=-1). The SVM class-weight condition used class_weight='balanced'; the Random Forest weight condition used class_weight='balanced_subsample'.

Within the training partition, 16 combinations compared unigram with unigram–bigram features and three balancing strategies: none, class weighting where supported, and neutral undersampling. Undersampling was performed independently inside each training fold by reducing neutral cases to the number of positive cases while retaining every negative case; validation folds and the final test set were never resampled. RepeatedStratifiedKFold used five folds and two repeats (10 validation estimates; random_state=42). The configuration with the highest mean macro-F1 was selected separately for each classifier, with negative recall and accuracy as tie-breakers. Each selected configuration was then fitted once to the full training partition and evaluated once on the untouched test set.

Reported measures are accuracy; macro-averaged precision, recall, and F1, which weight each class equally; weighted precision, recall, and F1, which weight by class support; and class-specific precision, recall, and F1. Raw and row-normalized confusion matrices show counts and within-class percentages. The analysis used Python 3.11.8, pandas 2.2.3, NumPy 2.2.5, and scikit-learn 1.6.1.

Ethics and Data Governance

The study analyzed publicly accessible, non-interventional YouTube comments and did not contact users. User display names and channel identifiers were excluded from the annotation and modeling datasets and were not used as features. Analysis files use coded comment IDs; results are reported only in aggregate, and no identifiable verbatim quotations

are reproduced. The raw export was kept separate from the derived analysis files. Public availability does not make the comments representative or eliminate contextual privacy risks, so negative predictions are intended for qualitative review rather than action against individuals. Ethical approval was granted by the Research Ethics Review Team of the Center for Research and Community Service (P3M), Politeknik Negeri Bali (Approval No. 21316/DST/PL8/AL.04/2026, dated 30 August 2026). The requirement for informed consent was waived because the study used anonymized publicly available data without interaction with users.

4. RESULT AND DISCUSSION

Human-Validated Sentiment Distribution

The consensus labels were predominantly neutral (60.92%), followed by positive (34.86%) and negative (4.22%) sentiment (Table 1 and Figure 2). Neutrality frequently represented informational statements, admissions or programme questions, and experience descriptions without a dominant evaluation. The distribution is less neutral-heavy than the earlier lexicon-generated labels, demonstrating why those pseudo-labels were unsuitable as an independent ground truth. The negative class nevertheless remained rare, making minority detection the central evaluation problem.

Table 1. Distribution of final consensus human labels

Sentiment	Count	Percentage
Neutral	1,227	60.92%
Positive	702	34.86%
Negative	85	4.22%

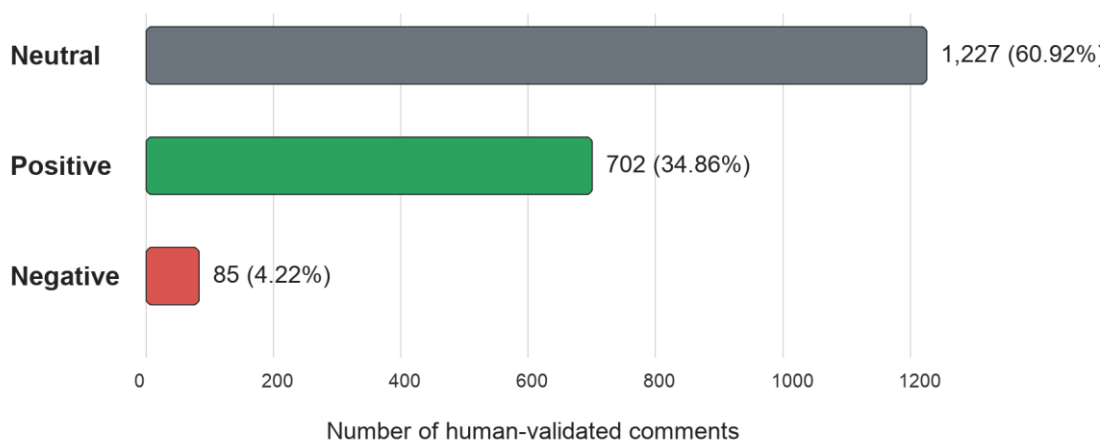


Figure 2. Distribution of human-validated sentiment labels

The 97.22% initial agreement and $\kappa = 0.945$ indicate that the three-category guideline was interpreted consistently by annotators with different expertise. Requiring consensus for every disagreement prevents the models from being evaluated against unresolved conflicts. Kappa should not be interpreted as proof that the labels are error-free, but it provides substantially stronger evidence than the lexicon-only procedure.

Cross-Validated Selection and Ablation

Table 2 reports the best configuration selected for each model using only the training partition. SVM with unigram features and balanced class weights achieved the highest macro-F1 (0.650 ± 0.054) and the highest mean negative recall (0.280 ± 0.110). Random Forest with no balancing achieved higher mean accuracy (0.815 ± 0.028) but a lower macro-F1 (0.574 ± 0.047) and near-zero negative recall (0.053 ± 0.051). Naïve Bayes selected training-only undersampling yet failed to identify negative validation examples.

Table 2. Best configuration per classifier in repeated stratified cross-validation

Model	Features	Balancing	Accuracy $\pm$ SD	Macro-F1 $\pm$ SD	Neg. recall $\pm$ SD
SVM	Unigram	Class weight	0.797 $\pm$ 0.028	0.650 $\pm$ 0.054	0.280 $\pm$ 0.110
Random Forest	Unigram	None	0.815 $\pm$ 0.028	0.574 $\pm$ 0.047	0.053 $\pm$ 0.051
Naïve Bayes	Unigram	Undersampling	0.782 $\pm$ 0.027	0.527 $\pm$ 0.019	0.000 $\pm$ 0.000

The ablation results do not support the assumption that more features or more balancing necessarily improve this corpus. For weighted SVM, adding bigrams changed mean macro-F1 from 0.650 to 0.647 and reduced negative recall from 0.280 to 0.235. Unweighted and undersampled unigram SVM obtained macro-F1 values of 0.611 and 0.615, respectively. For Random Forest, unigram/no balancing (0.574) exceeded the weighted (0.555) and undersampled (0.565) alternatives. For Naïve Bayes, unigram undersampling improved macro-F1 from 0.514 to 0.527 but did not recover negative instances. These results selected model-specific strategies rather than imposing one balancing rule on all classifiers.

Untouched Test-Set Performance

On the 403-comment test set, Random Forest produced the highest accuracy at 77.17%, only 0.50 percentage points above SVM (76.67%). This ordering reverses when each class receives equal weight: SVM achieved a macro-F1 of 0.655, compared with 0.544 for Random Forest and 0.490 for Naïve Bayes (Table 3). Weighted F1 tracked the majority classes and therefore presented a more favourable picture than macro-F1. SVM is treated as the primary model because it offers the best balance across classes and the best negative recall, not because it has the highest raw accuracy.

Table 3. Aggregate performance on the untouched natural-distribution test set

Model	Accuracy	Macro-P	Macro-R	Macro-F1	Weighted-F1
SVM	0.767	0.687	0.639	0.655	0.766
Random Forest	0.772	0.846	0.529	0.544	0.752
Naïve Bayes	0.727	0.479	0.512	0.490	0.716

Table 4. Class-specific performance on the untouched test set

Model	Class	Precision	Recall	F1	Support
SVM	Negative	0.545	0.353	0.429	17
SVM	Neutral	0.841	0.793	0.816	246
SVM	Positive	0.675	0.771	0.720	140
Random Forest	Negative	1.000	0.059	0.111	17
Random Forest	Neutral	0.774	0.907	0.835	246
Random Forest	Positive	0.763	0.621	0.685	140
Naïve Bayes	Negative	0.000	0.000	0.000	17
Naïve Bayes	Neutral	0.830	0.736	0.780	246
Naïve Bayes	Positive	0.605	0.800	0.689	140

The class-level results in Table 4 expose the practical difference. SVM correctly recovered 6 of 17 negative comments, whereas Random Forest recovered 1 and Naïve Bayes recovered none. SVM also correctly classified 195 of 246 neutral and 108 of 140 positive comments. The other models obtained competitive majority-class values while largely treating negative cases as neutral or positive. Consequently, the earlier conclusion that Random Forest was the strongest model based on 88.29% lexicon-label accuracy is withdrawn; that value measured reproduction of pseudo-labels and is not comparable to the present human-validated test result.

Confusion Matrices - SVM

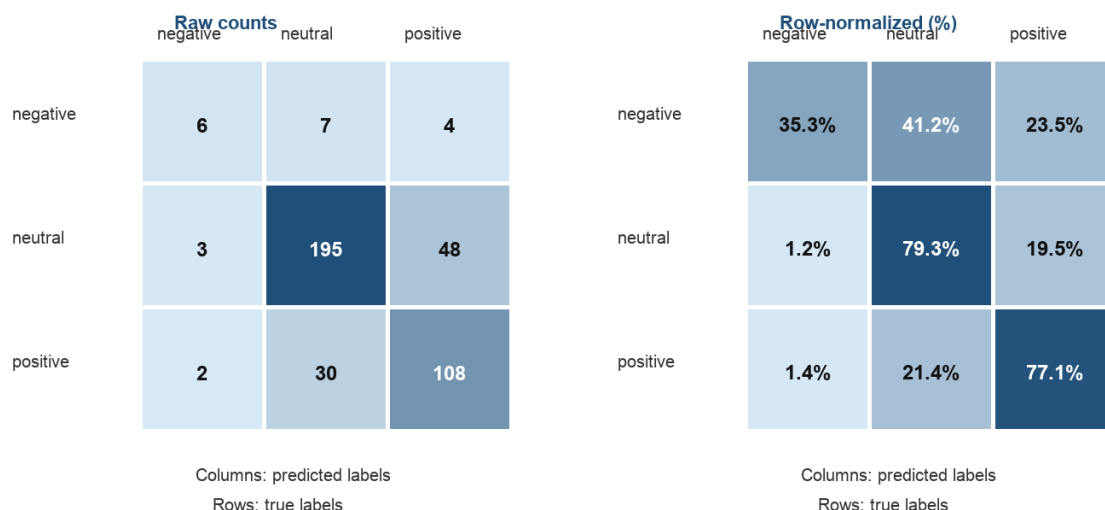


Figure 3. SVM confusion matrices: raw counts and row-normalized percentages

Confusion Matrices - Random Forest

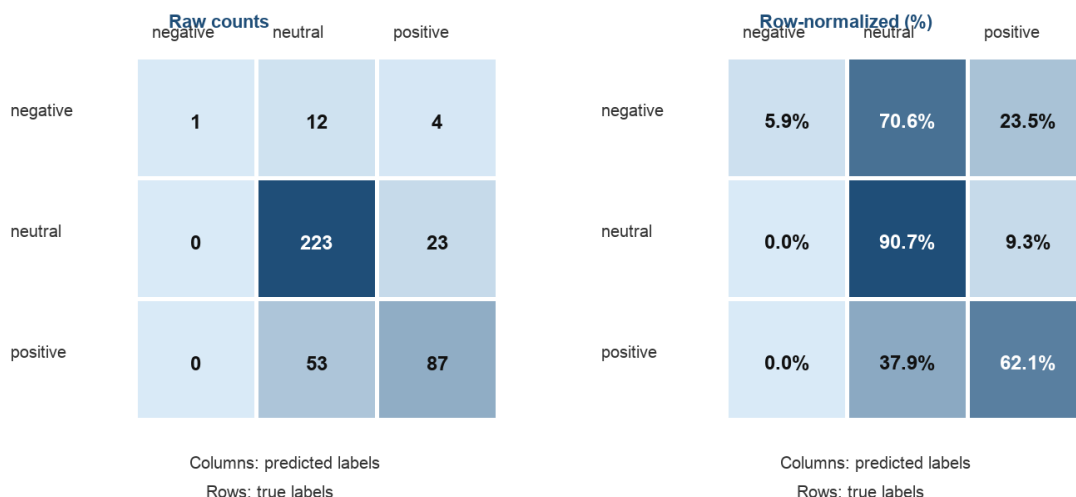


Figure 4. Random Forest confusion matrices: raw counts and row-normalized percentages

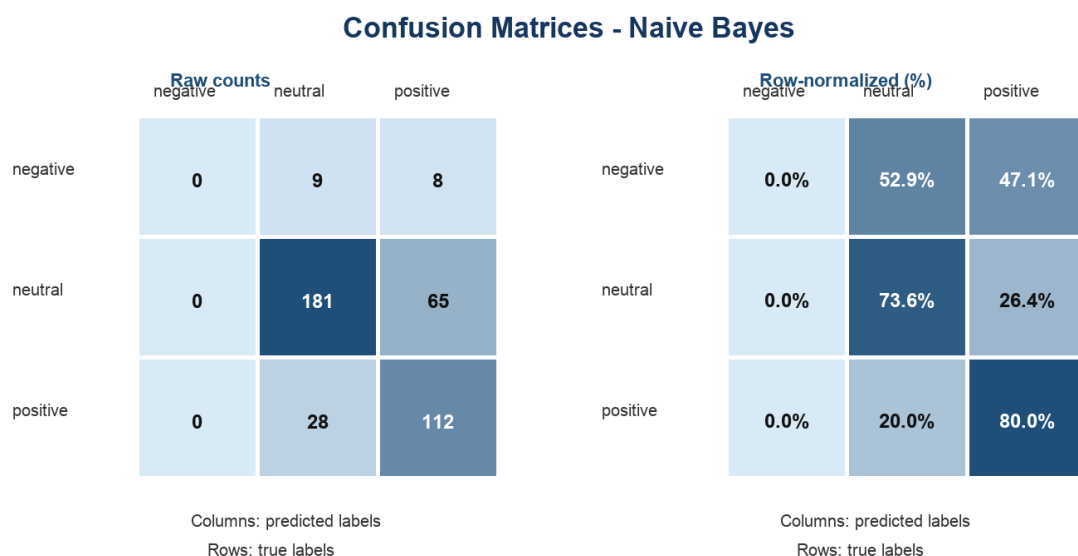


Figure 5. Naïve Bayes confusion matrices: raw counts and row-normalized percentages

Class-Specific Errors and Linguistic Interpretation

The SVM confusion matrix shows that 41.18% of negative comments were classified as neutral and 23.53% as positive. The 35.29% negative recall is an improvement over Random Forest but still means that 11 of 17 critical comments were missed. Rare-class scarcity is the most direct explanation: only 68 negative examples were available in the full training partition, and each cross-validation fold trained on fewer. Class weighting increased their influence but could not create missing linguistic variation.

Errors are also consistent with the limits of sparse word features. Criticism may be expressed indirectly through questions, comparisons, disappointment without explicit negative adjectives, or mixed praise and concern. Conservative normalization preserved common negators, intensifiers, contrast words, code-mixed tokens, and emojis, but it cannot reliably determine negation scope, sarcasm, or the evaluated educational aspect. Bigram features did not materially improve macro-F1, suggesting that the available sample was too small for the added sparse combinations to generalize.

Potential Contribution to Institutional Accountability

The results support a narrower claim than the original manuscript. The pipeline may complement existing accountability mechanisms by summarizing broad public commentary and prioritizing samples for human review. For example, institutions could monitor aggregate changes over time, compare model flags with complaints and survey themes, and route recurring issues to the responsible unit. The model does not measure institutional responsiveness, explain causes, or show that service quality improved.

SVM's 35.29% negative recall is insufficient for an autonomous early-warning system. A responsible use case is assisted screening: retain a broad comment sample, review all model-flagged negative cases, and also audit comments predicted as neutral to estimate what the system misses. Decisions should rely on qualitative reading and triangulation with formal evidence. Public reports should remain aggregated and should not identify, rank, or sanction individual commenters.

Limitations and Future Research

The evidence is limited to one platform, a query-based snapshot, and 79 videos from 66 channels. The Search API and engagement ordering may privilege popular or recent content, the newest top-level comments were retrieved, replies were excluded, and no independent spam or language detector was applied. The 2,014 comments therefore describe

the collected corpus rather than the Indonesian population or all vocational stakeholders. Comments after 28 August 2026 are not represented.

Although two expert annotators labelled the full dataset with high agreement, only 85 comments were negative and only 17 occurred in the test set. Percentage estimates for negative performance are thus sensitive to a small number of predictions. The study evaluates three broad sentiment categories rather than topics, targets, emotions, or aspects; it does not estimate uncertainty for operational decisions or test transfer to new institutions, time periods, or platforms.

Future research should purposefully collect more verified negative and mixed-sentiment examples, retain longitudinal and platform metadata, and create an external test set from new videos and dates. Aspect-based annotation could connect sentiment to teaching, facilities, administration, curriculum relevance, industry alignment, and graduate outcomes. Indonesian contextual language models may improve implicit sentiment detection (Minaee et al., 2022), but they should be compared under the same frozen split and human-labelled evaluation. Operational studies should finally test whether a human-in-the-loop dashboard actually improves response time, explanation, or service review rather than assuming accountability from classification performance.

5. CONCLUSION

This study re-evaluated Indonesian YouTube commentary on vocational higher education using complete human annotation and a leakage-controlled experimental design. Two independent annotators achieved 97.22% agreement and Cohen's $\kappa = 0.945$, and consensus produced 1,227 neutral, 702 positive, and 85 negative labels. Balancing and TF-IDF fitting were confined to training operations, model configurations were selected with repeated stratified five-fold cross-validation, and the final 403-comment test set retained the natural class distribution.

SVM with unigram TF-IDF and balanced class weights provided the strongest class-balanced result: 76.67% accuracy, 0.655 macro-F1, and 35.29% negative recall. Random Forest achieved 77.17% accuracy but recovered only 1 of 17 negative comments, while Naïve Bayes recovered none. These results replace the earlier lexicon-reproduction accuracy and demonstrate that minority-class performance is the decisive limitation. Sentiment analysis therefore has potential to complement—not replace—surveys, complaints, accreditation evidence, and human interpretation. Until negative detection is substantially improved and externally validated, the system should be treated as an assisted screening and research tool rather than an autonomous indicator of institutional accountability.

6. ACKNOWLEDGEMENT

The authors acknowledge Politeknik Negeri Bali for its institutional and administrative support. The authors also thank the vocational higher-education lecturer and the Indonesian-language lecturer who independently annotated all comments and participated in consensus adjudication.

7. REFERENCES

- Alahmadi, K., Alharbi, S., Chen, J., & Wang, X. (2025). Generalizing sentiment analysis: a review of progress, challenges, and emerging directions. *Social Network Analysis and Mining*, 15(1), 45. <https://doi.org/10.1007/s13278-025-01461-8>
- Altalhan, M., Algarni, A., & Turki-Hadj Alouane, M. (2025). Imbalanced Data Problem in Machine Learning: A Review. *IEEE Access*, 13, 13686–13699. <https://doi.org/10.1109/ACCESS.2025.3531662>
- Cero, I., Luo, J., & Falligant, J. M. (2024). Lexicon-Based Sentiment Analysis in Behavioral Research. *Perspectives on Behavior Science*, 47(1), 283–310. <https://doi.org/10.1007/s40614-023-00394-x>
- Chan, J., Wang, Y., Kok-Shun, B. V., & Woo, M. W. (2025). Exploring public perceptions of precision fermentation technology: A streamlined and labor-saving consumer perception analysis approach using YouTube data. *Future Foods*, 12, 100739. <https://doi.org/10.1016/j.fufo.2025.100739>

- Efstratiou, A. (2025). On YouTube Search API Use in Research. *Proceedings of the 2025 ACM Internet Measurement Conference*, 919–927. <https://doi.org/10.1145/3730567.3764492>
- Karunakaran, A., Orlikowski, W. J., & Scott, S. V. (2022). Crowd-Based Accountability: Examining How Social Media Commentary Reconfigures Organizational Accountability. *Organization Science*, 33(1), 170–193. <https://doi.org/10.1287/orsc.2021.1546>
- Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. *Proceedings of the 28th International Conference on Computational Linguistics*, 757–770. <https://doi.org/10.18653/v1/2020.coling-main.66>
- Li, F., & Ullah, A. (2025). Social media discourse as a window into energy transition: Analyzing public perception of electric vehicles on YouTube. *Energy Research & Social Science*, 129, 104402. <https://doi.org/10.1016/j.erss.2025.104402>
- Qi, Y., & Shabrina, Z. (2023). Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach. *Social Network Analysis and Mining*, 13(1), 31. <https://doi.org/10.1007/s13278-023-01030-x>
- Siino, M., Tinnirello, I., & La Cascia, M. (2024). Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. *Information Systems*, 121, 102342. <https://doi.org/10.1016/j.is.2023.102342>
- Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics*, 17(1), 168–192. <https://doi.org/10.1016/j.aci.2018.08.003>
- UNESCO. (2020). *Strategy for Technical and Vocational Education and Training (TVET)*. <https://unesdoc.unesco.org/>
- Wang, G., & Wang, Z. (2023). Vocational education: a poor second choice? A comparison of the labour market outcomes of academic and vocational graduates in China. *Oxford Review of Education*, 49(3), 408–427. <https://doi.org/10.1080/03054985.2022.2096583>
- Yates, L. A., Aandahl, Z., Richards, S. A., & Brook, B. W. (2023). Cross validation for model selection: A review with examples from ecology. *Ecological Monographs*, 93(1). <https://doi.org/10.1002/ecm.1557>