JSCE (Journal of System and Computer Engineering)
Vol.7, No.4, October 2026, pp. 462-472
ISSN (online): 2723-1240
DOI : https://doi.org/10.61628/jsce.v7i4.2872

*Research Article Open Access (CC–BY-SA)* ■

# Comparing Four Classifiers for Imbalanced Computer Engineering Student Feedback

**Kharis Hudaiby Hanif [1,a,*]; Novita Ranti Muntiari [2,b] ; Awang Pradana [3,c]; Luthfi Muzakkii [4,d]**

*[1,3,4]Department of Computer Engineering, University of Borneo Tarakan, North Kalimantan, Indonesia*
*[2]Department of Health Promotion, Politeknik Kaltara, North Kalimantan, Indonesia*
*[a][hudaiby21@borneo.ac.id];[b][novitarantimuntiari@gmail.com];[c][awang.pradana@borneo.ac.id];[d][luthfimuzakkii2340304010@gmail.com]*
** Corresponding author*

**Abstract**

*Open-ended course evaluation comments provide rich evidence about learning experiences, but extreme class imbalance can make automated sentiment results misleading. This study compares Decision Tree, Support Vector Machine (SVM), Random Forest, and Gradient Boosting on Computer Engineering student comments using TF-IDF features and training-only SMOTE. The initial dataset contained 1,075 comments: 1,065 neutral, five negative, and five positive; 415 complete observations remained for modeling. Before SMOTE, Random Forest and SVM achieved 98.06 percent accuracy, while Decision Tree and Gradient Boosting achieved 97.58 percent. After SMOTE, SVM achieved 98.55 percent accuracy and the highest macro F1-score of 66.42 percent; Gradient Boosting, Random Forest, and Decision Tree obtained 49.59, 33.01, and 32.76 percent, respectively. Class-wise analysis showed that SVM and Gradient Boosting detected the single negative test comment, but none of the models detected the three positive comments. The contribution is a mechanism-oriented diagnosis of SMOTE under extreme low-resource minority conditions, showing that synthetic balancing can change classifier sensitivity without creating new linguistic diversity. Macro metrics, confusion matrices, and human-validated labels are therefore essential for reliable course-evaluation analytics.*

**Keywords**—sentiment analysis, student feedback, SVM, SMOTE, class imbalance

## 1. Introduction

Learning quality cannot be evaluated solely through numerical scores. Open-ended student comments often contain information about the clarity of learning materials, classroom interaction, assignments, laboratory activities, schedules, and supporting facilities. Because such comments are unstructured and written in varied language, manual review becomes increasingly inefficient as the number of responses grows. Sentiment analysis offers a computational approach for organizing these comments into categories that can support academic interpretation and formative evaluation (Grimalt-Alvaro & Usart, 2024).

Machine learning remains widely used for educational feedback analysis because it can learn relationships between numerical text representations and sentiment labels. Once comments are transformed into features, classification models can identify patterns associated with different sentiment classes. Dake and Gyimah (2023) showed that traditional classifiers, including SVM, Decision Tree, and Random Forest, remain relevant for qualitative student responses when preprocessing and model evaluation are carefully designed. Kasumba and Neumann (2024) further demonstrated the value of domain-specific supervised learning for sentiment in large computer-science classes, while Deshpande et al. (2025) reported different performance profiles among five classifiers applied to TF-IDF representations of 5,000 engineering-student feedback responses.

Recent studies have reported different results across algorithms and datasets. Putra et al. (2023) applied SVM to 1,791 EDOM comments and reported 82% accuracy with a 65% F-measure. Muntiari et al. (2023) compared Logistic Regression, SVM, and Gradient Boosting on student-comment sentiment data and reported the highest accuracy for Gradient Boosting. Karfindo et al. (2023) combined Random Forest and SMOTE for campus-service feedback and reported improved minority-class recognition after balancing. In Computer Engineering feedback, Hanif et al. (2025) compared Decision Tree and SVM, while Hanif et al. (2026) expanded the comparison to Decision Tree, Random Forest, and SVM. Sidik et al. (2025) also found that SMOTE changed the performance of SVM and Random Forest on student-satisfaction comments. Together, these findings indicate that algorithm rankings depend on class composition, feature representation, labeling, and the amount of genuine minority evidence.

The sentiment labels in this study were generated from a compound score using a rule-based polarity mapping similar to recent VADER-based student-feedback analysis (Mihai, 2024). This approach makes it possible to assign positive, neutral, and negative labels automatically. However, Indonesian student comments may contain abbreviations, academic terms, conversational expressions, and mixed vocabulary that are not fully represented in a general-purpose sentiment lexicon. Recent educational-feedback research also notes that lexicon-based approaches can have difficulty capturing contextual nuance (Dervenis et al., 2024). Consequently, the resulting labels should be interpreted as computational labels rather than substitutes for human judgment.

Class imbalance is another central issue. A model can appear highly accurate when most observations belong to a single class even though its ability to recognize minority classes is weak. Recent SMOTE research shows that synthetic interpolation can reduce numerical imbalance, but its effectiveness depends on the distribution and neighborhood structure of the original minority samples. Poorly supported neighborhoods may generate overlap, ambiguous synthetic points, or redundant regions rather than genuinely new information (Yang et al., 2022; Kosolwattana et al., 2023; Elreedy et al., 2024). Boudegzdame et al. (2024) further showed that oversampling can introduce synthetic artifacts that classifiers may learn instead of the underlying minority pattern. In sentiment classification, SMOTE should therefore be interpreted together with precision, recall, F1-score, and confusion matrices rather than accuracy alone (Hutami et al., 2022; Sediatmoko et al., 2024; Singgalen, 2024a).

Despite this literature, two research gaps remain. First, closely related educational-feedback studies compare subsets of classical classifiers or report aggregate performance after balancing, but they do not diagnose how single-tree partitioning, bagged trees, large-margin classification, and sequential boosting respond to the same extreme low-resource minority condition (Hanif et al., 2025, 2026; Sidik et al., 2025). Second, prior SMOTE-based sentiment studies commonly evaluate datasets with substantially more minority evidence, so they do not characterize the failure regime in which a sentiment class is represented by only a handful of genuine comments. In the present dataset, neutral comments account for 99.07% of the original observations, while the positive and negative classes contain only five comments each. Under this condition, high overall accuracy can coexist with complete failure on a minority class.

The novelty of this study is therefore an empirical diagnostic contribution rather than the proposal of a new classification algorithm. Under one controlled TF-IDF representation, train-test setting, and training-only SMOTE procedure, four learners with different mechanisms—single-tree partitioning, bagged trees, large-margin classification, and sequential boosting—are compared before and after balancing. The scientific contribution is threefold: (1) to quantify how these mechanisms respond differently to the same extreme imbalance condition; (2) to explain the limits of synthetic interpolation when only a few genuine minority comments define the TF-IDF neighborhood; and (3) to show, through class-wise errors, why balancing the number of training vectors does not necessarily balance the linguistic information available to the classifier. This mechanism-oriented diagnosis goes beyond simply adding Gradient Boosting to an existing comparison.

## 2. Method

### 2.1 Research Stages

The research stages follow the processing sequence used in the analysis notebook so that the relationship between raw data, feature construction, model training, and evaluation can be traced consistently. The overall workflow is shown in Figure 1.

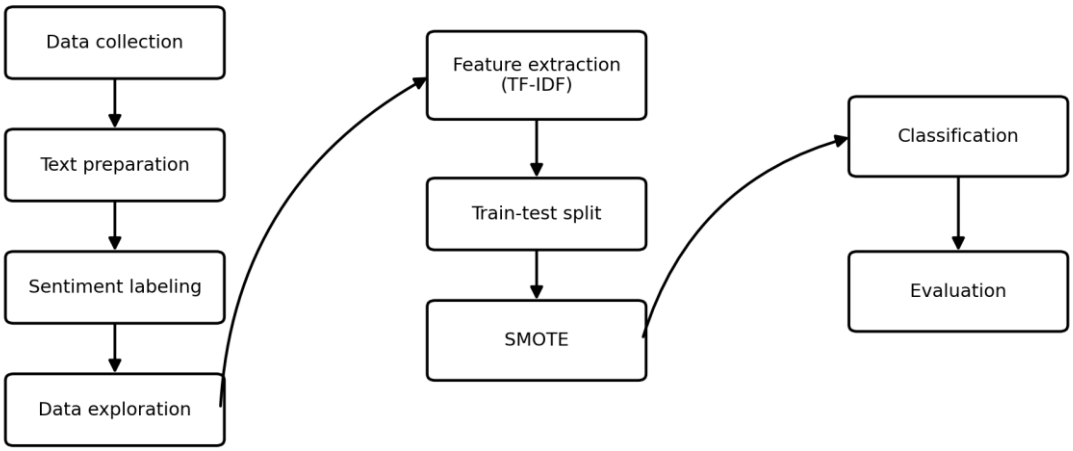



**Figure 1. Research workflow.**

As shown in Figure 1, the study begins with the collection of course-evaluation comments. The text is then prepared and assigned positive, neutral, or negative sentiment labels. Exploratory analysis is used to examine the class distribution before TF-IDF converts the comments into numerical features. The resulting data are divided into training and test sets, and SMOTE is applied only to the training set to reduce class imbalance. Four classifiers are then trained, and their performance is compared using accuracy, precision, recall, F1-score, and confusion matrices.

### 2.2 Data Collection

The data consist of comments submitted by students in the Computer Engineering Study Program through course evaluations. The initial dataset contains 1,075 comments and serves as the source for all subsequent steps, from sentiment labeling to model evaluation.

### 2.3 Text Preparation and Sentiment Labeling

The modeling process uses the cleaned comment field, Komentar_baru, together with Compound_Score and Sentiments. After these attributes were selected, rows with incomplete values were removed, leaving 415 complete observations. This step ensured that each record contained text, a polarity score, and a sentiment label that could be processed consistently.

Sentiment labels were determined from the compound score. Following the VADER-style thresholds used in recent student-feedback sentiment analysis, scores greater than or equal to 0.05 were mapped to positive sentiment, scores less than or equal to -0.05 were mapped to negative sentiment, and values between these thresholds were categorized as neutral (Mihai, 2024). Because the labels were generated automatically, they were treated as experimental computational labels rather than a human-annotated gold standard.

### 2.4 Class Distribution Exploration

Exploratory analysis was conducted by calculating the number and proportion of observations in each sentiment class before modeling. This step was necessary to quantify the degree of imbalance and to determine whether a minority-class handling strategy was required.

### 2.5 TF-IDF Feature Extraction

Text comments cannot be processed directly by the classifiers, so each document was transformed into a numerical vector using Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF assigns greater weight to terms that are frequent in a document but relatively uncommon in the overall corpus. The general form is shown in Equation (1).

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) \times \log(N / \text{DF}(t)) \quad (1)$$

In Equation (1), TF(t,d) is the frequency of term t in document d, DF(t) is the number of documents containing the term, and N is the total number of documents. The resulting TF-IDF vectors were used as input features for all four classifiers.

### 2.6 Train-Test Split

The complete data were divided into training and test sets before class balancing. The confusion-matrix outputs contained 207 test observations; with 415 complete records, the reconstructed experimental split consisted of 208 training observations and 207 test observations, or approximately 50:50. The same test set was used to compare all algorithms on an equivalent basis.

### 2.7 Class Balancing with SMOTE

SMOTE was used because the sentiment distribution was strongly dominated by the neutral class. The method increases minority-class representation by interpolating synthetic observations between neighboring minority samples in feature space. SMOTE was applied only to the training set after the split, while the test set retained its original distribution to reduce information leakage. Recent work emphasizes that SMOTE depends on local minority geometry; when genuine minority samples are very few, synthetic observations may be redundant, ambiguous, or overlap with majority regions rather than provide new structure (Yang et al., 2022; Kosolwattana et al., 2023; Elreedy et al., 2024). This limitation is especially relevant for sparse TF-IDF text vectors.

### 2.8 Classification Algorithms

The classification stage used supervised machine learning in which the models learned the relationship between TF-IDF features and sentiment labels. Four algorithms were evaluated under the same data scenario: Decision Tree, Support Vector Machine (SVM), Random Forest, and Gradient Boosting.

Decision Tree constructs hierarchical decision rules by selecting features that best separate the target classes. The model is easy to interpret but can be sensitive to changes in the training data and may overfit. SVM searches for a separating boundary with a large margin between classes and is well suited to high-dimensional sparse text features. Random Forest combines multiple decision trees trained on different samples and feature subsets, improving stability through aggregation. Gradient Boosting builds learners sequentially so that each new learner attempts to reduce the errors remaining from previous learners. These different learning mechanisms provide a useful basis for comparing model behavior on imbalanced text data.

### 2.9 Model Evaluation

Model performance was assessed using confusion matrices together with accuracy, precision, recall, and F1-score. A confusion matrix compares the actual and predicted classes, making it possible to examine errors separately for positive, neutral, and negative comments. For each class, true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values were used to calculate the metrics in Equations (2)-(5).

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (2)$$

$$\text{Precision} = TP / (TP + FP) \quad (3)$$

$$\text{Recall} = TP / (TP + FN) \quad (4)$$

$$\text{F1-score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (5)$$

Because the class sizes were highly unequal, precision, recall, and F1-score were also summarized using macro averaging, which gives the same weight to each class. The models were therefore compared by considering overall accuracy together with macro precision, macro recall,

macro F1-score, and class-specific errors in the confusion matrix. This evaluation design is particularly important for the present dataset because a high accuracy can be obtained by predicting the neutral majority correctly while missing rare positive or negative comments.

## 3. Results and Discussion

### 3.1 Sentiment Distribution and Data Completeness

The initial dataset contained 1,075 student comments. The computational labeling procedure produced 1,065 neutral comments, five negative comments, and five positive comments. Thus, the neutral class represented 99.07% of all observations, while the negative and positive classes each accounted for only 0.47%. This distribution indicates extreme class imbalance and creates a strong risk that accuracy will be dominated by correct predictions of the neutral class. The initial distribution is summarized in Table 1.

**Table 1. Sentiment distribution in the initial dataset.**

| Sentiment | Number | Percentage |
|---|---|---|
| **Neutral** | 1,065 | 99.07% |
| **Negative** | 5 | 0.47% |
| **Positive** | 5 | 0.47% |
| **Total** | 1,075 | 100% |

The same distribution is visualized in Figure 2. After Komentar_baru, Compound_Score, and Sentiments were selected and incomplete rows were removed, 415 observations remained for modeling.

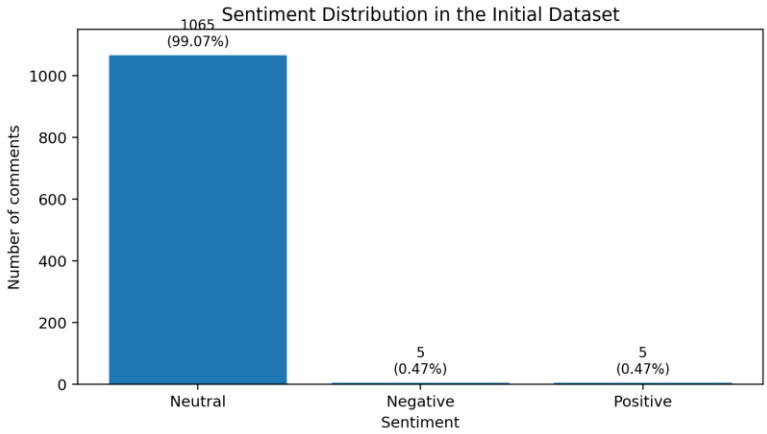



**Figure 2. Sentiment distribution in the initial dataset.**

### 3.2 Model Performance Before SMOTE

Before class balancing, all four algorithms produced accuracy values above 97%. Random Forest and SVM each reached 98.06%, whereas Decision Tree and Gradient Boosting reached 97.58%. As shown in Table 2, the differences in overall accuracy were small, but the extreme dominance of neutral observations means these values cannot be interpreted independently of class-level errors.

**Table 2. Model accuracy before SMOTE.**

| Algorithm | Accuracy before SMOTE |
|---|---|
| **Random Forest** | 98.06% |
| **Decision Tree** | 97.58% |
| **Support Vector Machine** | 98.06% |
| **Gradient Boosting** | 97.58% |

Figure 3 presents the same comparison graphically. The confusion matrices explain why high accuracy did not necessarily indicate balanced recognition. Random Forest and SVM correctly classified 203 neutral comments but assigned one negative and three positive comments to the

neutral class. Decision Tree and Gradient Boosting recognized one negative comment, but two neutral comments were predicted as negative and all three positive comments were still predicted as neutral. Therefore, none of the models correctly identified a positive test comment before SMOTE.

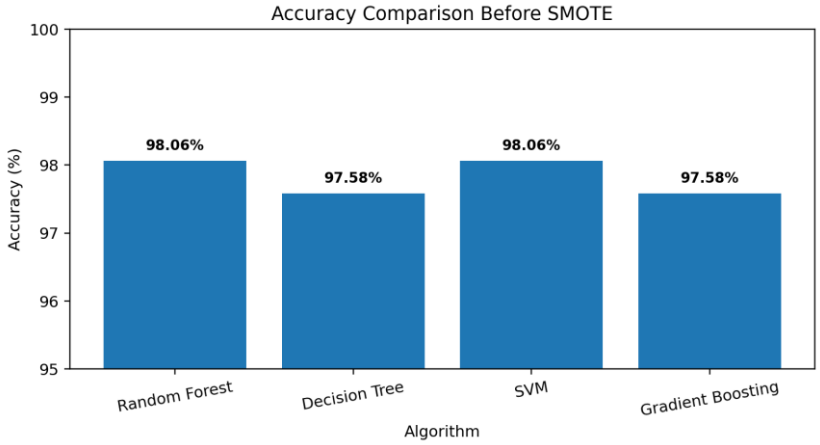



**Figure 3. Accuracy comparison of four algorithms before SMOTE.**

### 3.3 Model Performance After SMOTE

After SMOTE was applied to the training data, differences among the algorithms became clearer when macro-averaged metrics were considered. SVM achieved 98.55% accuracy, 66.18% macro precision, 66.67% macro recall, and 66.42% macro F1-score. This was the highest macro F1-score among the four models, but the class-wise analysis in Section 3.4 shows that the improvement was not uniform across all sentiment classes because the positive class remained undetected.

Gradient Boosting achieved a macro F1-score of 49.59% with 97.58% accuracy. Random Forest achieved 98.07% accuracy but only 33.01% macro F1-score because its predictions remained concentrated on the neutral class. Decision Tree recorded 96.62% accuracy and 32.76% macro F1-score. Table 3 summarizes the results.

**Table 3. Performance of four algorithms after SMOTE.**

| Algorithm | Accuracy | Macro Precision | Macro Recall | Macro F1-score |
|---|---|---|---|---|
| **SVM** | 98.55% | 66.18% | 66.67% | 66.42% |
| **Gradient Boosting** | 97.58% | 43.95% | 66.34% | 49.59% |
| **Random Forest** | 98.07% | 32.69% | 33.33% | 33.01% |
| **Decision Tree** | 96.62% | 32.68% | 32.84% | 32.76% |

The comparison in Table 3 is illustrated in Figure 4. The large difference between accuracy and macro-averaged measures shows that overall correctness remained strongly influenced by the neutral majority class.

### 3.4 Class-wise Performance of SVM

The post-SMOTE SVM confusion matrix contains 207 test observations: one negative, 203 neutral, and three positive. SVM correctly classified the single negative comment and all 203 neutral comments, while all three positive comments were predicted as neutral. Although precision and recall for the negative class appear to be 100%, this result is based on only one negative test observation and therefore does not demonstrate stable performance for that class. For the positive class, precision and recall were both 0% because no positive prediction was correct.

The difference between 98.55% accuracy and 66.42% macro F1-score demonstrates the effect of class imbalance. Accuracy was driven primarily by successful neutral predictions, whereas macro F1-score gave equal importance to the three classes. Figure 5 shows the SVM confusion matrix after SMOTE, and Table 4 compares the main error patterns of all four models.

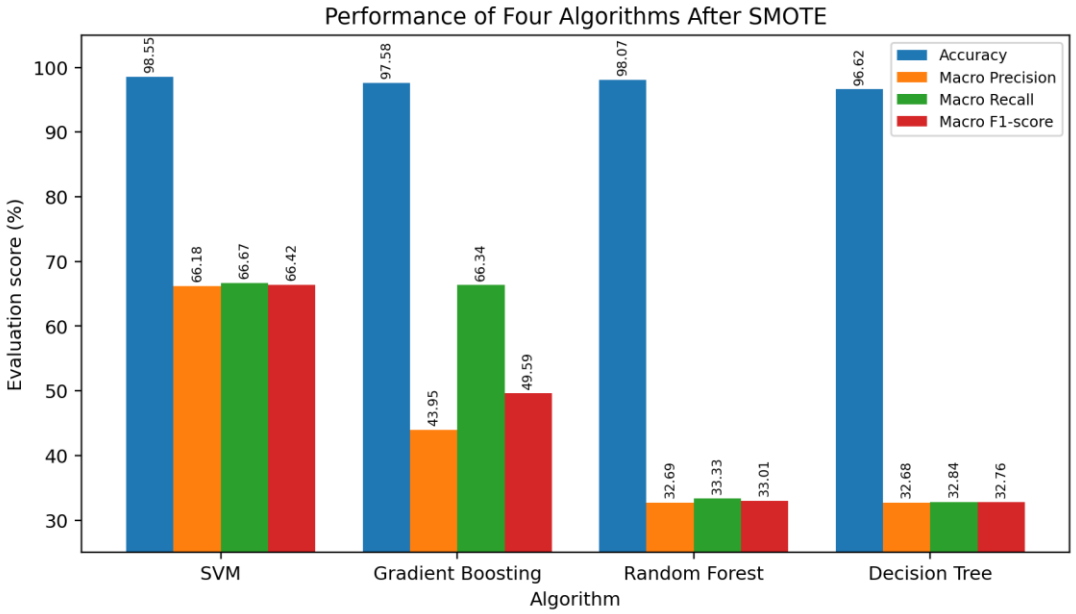



**Figure 4. Performance comparison of four algorithms after SMOTE.**

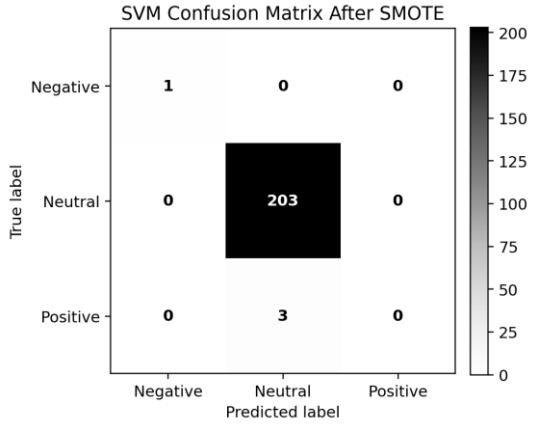



**Figure 5. SVM confusion matrix after SMOTE**

**Summary of confusion matrices after SMOTE view Table 4.**

**Table 4. Summary of confusion matrices after SMOTE.**

| Model | Negative correct | Neutral correct | Positive correct | Main error pattern |
|---|---|---|---|---|
| **SVM + SMOTE** | 1/1 | 203/203 | 0/3 | 3 positive -> neutral |
| **Gradient Boosting + SMOTE** | 1/1 | 201/203 | 0/3 | 2 neutral -> negative; 3 positive -> neutral |
| **Random Forest + SMOTE** | 0/1 | 203/203 | 0/3 | 1 negative and 3 positive -> neutral |
| **Decision Tree + SMOTE** | 0/1 | 200/203 | 0/3 | 1 negative -> neutral; 3 neutral -> negative; 3 positive -> neutral |

### 3.5 Comparative Interpretation and Mechanistic Discussion

The post-SMOTE results show why macro metrics must be read together with class-wise errors. SVM achieved the highest macro F1-score in this experiment (66.42%), but this value came from a specific pattern: it correctly recognized the single negative test comment and all 203 neutral comments, while positive recall remained 0%. In high-dimensional sparse TF-IDF space, a linear margin-based classifier can respond to oversampling when additional minority vectors

shift the effective separating boundary. Recent studies likewise report gains for SVM after SMOTE, although the magnitude depends on the dataset (Hutami et al., 2022; Andriyani et al., 2025). Sidik et al. (2025) reported improved student-comment classification after SMOTE, and Adyanata et al. (2026) found SVM with SMOTE to be the strongest classical configuration on a moderately imbalanced Indonesian sentiment dataset. In the present experiment, however, the benefit was limited to the negative class and should not be interpreted as stable three-class recognition.

Gradient Boosting displayed a different response. It also recognized the single negative test comment, but two neutral comments were predicted as negative, producing a macro recall of 66.34% together with a lower macro precision of 43.95% and a macro F1-score of 49.59%. This pattern is consistent with the sequential mechanism of boosting: later learners concentrate on errors left by earlier learners, so an enlarged synthetic minority region can receive increasing attention during training. The same mechanism can improve sensitivity to a rare class while also creating local decision regions that absorb nearby neutral vectors. Muntiari et al. (2023) reported higher overall accuracy for Gradient Boosting than SVM on another student-comment dataset, whereas SVM is stronger in the present study. The difference indicates that the relative performance of boosting and margin-based classifiers is dataset-dependent, particularly when minority support is extremely small.

Random Forest, in contrast, continued to predict the neutral class for all four minority test observations: one negative and three positive comments. Bagging improves stability by aggregating many trees, but each tree is still built from sampled observations and feature subsets. When genuine minority evidence is extremely sparse, discriminative minority-term patterns may appear too inconsistently across trees to influence the majority vote. SMOTE increases the number of minority vectors, but those vectors remain interpolations of the same limited TF-IDF directions and do not create new semantic patterns. This differs from Karfindo et al. (2023) and Sidik et al. (2025), who observed stronger Random Forest behavior after balancing on larger student-feedback datasets, and from Deshpande et al. (2025), where Random Forest performed strongly on 5,000 engineering feedback responses. The contrast supports the interpretation that minority sample diversity, not only sample count, constrains ensemble learning here.

Decision Tree produced the lowest post-SMOTE macro F1-score (32.76%). A single tree forms hard threshold rules on individual feature dimensions and is therefore sensitive to small changes in the available training patterns. With only a few genuine minority observations, synthetic TF-IDF points can encourage narrow splits that fit interpolated neighborhoods without capturing recurring linguistic patterns. Recent SMOTE studies describe related risks, including overlap, low-quality synthetic observations, and distributional distortion when minority neighborhoods are weakly supported (Yang et al., 2022; Kosolwattana et al., 2023; Elreedy et al., 2024). Boudegzdame et al. (2024) additionally warn that a learner can respond to oversampling artifacts rather than the intended minority structure. These mechanisms help explain why SMOTE did not improve Decision Tree in the same way as SVM.

The persistent failure to detect positive sentiment is explained most directly by the number of genuine positive examples. The original corpus contains only five positive comments, and the SVM test confusion matrix contains three positive observations. Therefore, even if all five positive comments survived the completeness filter, the training partition could contain no more than two genuine positive comments, and possibly fewer. SMOTE can generate many synthetic vectors from those positive points, but interpolation in TF-IDF space only fills geometric locations between the available lexical vectors; it does not invent new approval terms, phrases, or contexts that were absent from the original positive comments. This limitation matters because frequency-based representations such as TF-IDF can miss semantic subtleties in student feedback (Tuan et al., 2025). Widayanti and Kasih (2026) similarly found that TF-IDF with SMOTE-SVM remained limited relative to contextual representations on imbalanced Indonesian sentiment data. If the three positive test comments use different lexical cues or contextual constructions, synthetic

interpolation cannot supply the missing linguistic evidence, so the classifiers can still place them in the neutral region.

A second factor is the relationship between computational labels and text features. The sentiment labels were assigned automatically from a compound score rather than verified by human annotators. As noted in Section 2.3, Indonesian course-evaluation comments may contain abbreviations, domain-specific terminology, conversational language, or mixed vocabulary. Dervenis et al. (2024) reported that lexicon-based analysis of student feedback can have difficulty capturing contextual nuance, while Mihai (2024) demonstrates the usefulness of VADER-style scoring for teaching feedback but does not eliminate domain-dependence. A positive label can therefore be triggered by a limited lexical cue even when the remaining TF-IDF pattern resembles the dominant neutral class. Under extreme data scarcity, this possible label-feature mismatch becomes more influential because there are too few genuine positive examples for the classifier to learn a robust alternative pattern.

Overall, the findings are partly consistent with recent literature and partly different in an informative way. Studies using SVM with SMOTE frequently report gains in minority-class performance or competitive macro F1-scores (Hutami et al., 2022; Andriyani et al., 2025; Adyanata et al., 2026), and the present results also show a selective gain for SVM. However, the broader improvements reported for SVM and Random Forest on student-feedback or moderately imbalanced sentiment datasets (Karfindo et al., 2023; Sidik et al., 2025) were not reproduced for the positive class here. Widayanti and Kasih (2026) also show that imbalance handling with TF-IDF and SVM can remain constrained by the representation itself. The key difference is that balancing counts is not equivalent to balancing information: when the minority class contains almost no genuine linguistic diversity, synthetic interpolation can densify a small region without expanding the vocabulary or contexts represented by that class. This extreme low-resource condition is the main scientific insight of the comparison and explains why macro F1-score and confusion-matrix analysis are more informative than accuracy alone in this dataset.

### 3.6 Implications for Course Evaluation

In practical course evaluation, sentiment classification can serve as an initial screening mechanism for large volumes of student comments. Program administrators could use the predicted groups to prioritize comments for further reading and then relate them to issues such as learning materials, assignments, laboratory activities, schedules, networks, and facilities. This approach can reduce the workload of manual review as the volume of feedback increases.

However, automated sentiment labels should not be used as the sole basis for evaluating lecturers or courses. Classification errors remain possible, particularly for classes represented by very few observations. The model output is therefore more appropriate as supporting information that should be interpreted together with quantitative evaluation scores, course context, and human review.

## 4. Conclusions

The comparison of four algorithms under the same TF-IDF and training-only SMOTE setting shows that SVM achieved the highest macro F1-score in this experiment. Before SMOTE, SVM and Random Forest each achieved 98.06% accuracy, while Decision Tree and Gradient Boosting achieved 97.58%. After SMOTE, SVM reached 98.55% accuracy, 66.18% macro precision, 66.67% macro recall, and 66.42% macro F1-score. Gradient Boosting achieved a 49.59% macro F1-score, followed by Random Forest at 33.01% and Decision Tree at 32.76%. More importantly, the class-wise analysis showed that none of the models correctly recognized the three positive test comments. The scientific contribution of the study is therefore the diagnosis of model-specific responses to SMOTE under an extreme low-resource minority condition: SVM and Gradient Boosting became sensitive to the negative class, whereas Random Forest and Decision Tree remained largely dominated by neutral predictions, and synthetic balancing did not create sufficient information to recognize the positive class. The results confirm that high accuracy is insufficient for severely imbalanced sentiment data and that synthetic sample counts should not

be treated as a substitute for genuine linguistic diversity. Future work should collect substantially more authentic positive and negative comments, validate sentiment labels with multiple human annotators, report the exact minority counts at each split, evaluate stratified cross-validation and cost-sensitive or class-weighted alternatives, and compare text-aware augmentation or contextual language-model representations with TF-IDF. Aspect-based analysis could also be added to group comments into themes such as materials, assignments, laboratory activities, schedules, networks, and facilities.

**Acknowledgements**

The author would like to express the highest gratitude to Universitas Borneo Tarakan for providing the financial support that made this research possible. Special thanks are also extended to the Department of Computer Engineering (Teknik Komputer) for the academic guidance, facilities, and continuous support throughout the completion of this study.

**References**

Adyanata, I. K., Gunadi, I. G. A., & Sunarya, I. M. G. (2026). Comparison of SMOTE, class weighting, and classical machine learning models on the ID-SMSA Indonesian stock market dataset. Journal of System and Computer Engineering, 7(3), 229–240. https://doi.org/10.61628/jsce.v7i3.2688

Andriyani, D., Faqih, A., & Permana, S. E. (2025). The effect of SMOTE application on Support Vector Machine performance in sentiment classification on imbalanced datasets. Journal of Artificial Intelligence and Engineering Applications, 4(2), 752–757. https://doi.org/10.59934/jaiea.v4i2.742

Boudegzdame, N., Sedki, K., Tspora, R., & Lamy, J.-B. (2024). SMOTE: Are we learning to classify or to detect synthetic data? In Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART) (Vol. 3, pp. 283–290). SciTePress. https://doi.org/10.5220/0012325300003636

Dake, D. K., & Gyimah, E. (2023). Using sentiment analysis to evaluate qualitative students' responses. Education and Information Technologies, 28, 4629–4647. https://doi.org/10.1007/s10639-022-11349-1

Dervenis, C., Kanakis, G., & Fitsilis, P. (2024). Sentiment analysis of student feedback: A comparative study employing lexicon and machine learning techniques. Studies in Educational Evaluation, 83, 101406. https://doi.org/10.1016/j.stueduc.2024.101406

Deshpande, S. B., Tangod, K. K., Srinivasaiah, S. H., Alahmadi, A. A., Alwetaishi, M., Michael, G. K. O., & Rajendran, S. (2025). Elevating educational insights: Sentiment analysis of faculty feedback using advanced machine learning models. Advances in Continuous and Discrete Models, 2025, 89. https://doi.org/10.1186/s13662-025-03933-9

Elreedy, D., Atiya, A. F., & Kamalov, F. (2024). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. Machine Learning, 113, 4903–4923. https://doi.org/10.1007/s10994-022-06296-4

Grimalt-Álvaro, C., & Usart, M. (2024). Sentiment analysis for formative assessment in higher education: A systematic literature review. Journal of Computing in Higher Education, 36, 647–682. https://doi.org/10.1007/s12528-023-09370-5

Hanif, K. H., Fadllullah, A., Muntiari, N. R., & Fahrezi, I. A. (2025). A comparative sentiment analysis of Computer Engineering student feedback using Decision Trees and SVM. Jurnal Inotera, 10(1), 71–82. https://doi.org/10.31572/inotera.Vol10.Iss1.2025.ID436

Hanif, K. H., Muntiari, N. R., Harto, D., & Wiranata, D. S. (2026). Perbandingan analisis sentimen komentar mahasiswa Prodi Teknik Komputer menggunakan algoritma Decision Tree, Support Vector Machine (SVM) dan Random Forest. INSECT (Informatics and Security): Jurnal Teknik Informatika, 12(1), 24–33. https://doi.org/10.33506/insect.v12i01.5144

Hutami, W. P., Wijayanto, H., & Sulvianti, I. D. (2022). Penerapan Support Vector Machine dengan SMOTE untuk klasifikasi sentimen pemberitaan Omnibus Law pada situs CNNIndonesia.com. Xplore: Journal of Statistics, 11(1), 26–35. https://doi.org/10.29244/xplore.v11i1.852

Karfindo, Turaina, R., & Saputra, R. (2023). Optimalisasi klasifikasi umpan balik mahasiswa terhadap layanan kampus dengan sinergi Random Forest dan SMOTE. Jurnal Nasional Komputasi dan Teknologi Informasi, 6(6), 820–827. https://doi.org/10.32672/jnkti.v6i6.7269

Kasumba, R., & Neumann, M. (2024). Practical sentiment analysis for education: The power of student crowdsourcing. Proceedings of the AAAI Conference on Artificial Intelligence, 38(21), 23110–23118. https://doi.org/10.1609/aaai.v38i21.30356

Kosolwattana, T., Liu, C., Hu, R., Han, S., Chen, H., & Lin, Y. (2023). A self-inspected adaptive SMOTE algorithm (SASMOTE) for highly imbalanced data classification in healthcare. BioData Mining, 16, 15. https://doi.org/10.1186/s13040-023-00330-4

Mihai, G. (2024). Sentiment analysis of students' feedback using the VADER model for teaching process assessment. Economics and Applied Informatics, 30(3), 196–203. https://doi.org/10.35219/eai15840409445

Muntiari, N. R., Hanif, K. H., & Nisa, I. C. (2023). Perbandingan algoritma regresi logistik, Support Vector Machine, dan Gradient Boosting pada analisis sentimen data komentar siswa. Jurnal Ilmu Komputer dan Teknologi, 4(2), 27–32. https://doi.org/10.35960/ikomti.v4i2.1286

Putra, A. N. C., Haviana, S. F. C., & Badie'ah. (2023). Sentimen analisis komentar mahasiswa EDOM dengan metode Support Vector Machine (SVM). Seminar Riset Mahasiswa - Computer & Electrical (SERIMA-CE), 1(1), 260–264.

Sediatmoko, N. S., Nataliani, Y., & Suryady, I. (2024). Sentiment analysis of customer review using classification algorithms and SMOTE for handling imbalanced class. Indonesian Journal of Information Systems, 7(1), 38–52. https://doi.org/10.24002/ijis.v7i1.8879

Sidik, P., Sunarya, I. M. G., & Gunadi, I. G. A. (2025). Comparison of Random Forest and Support Vector Machine methods in sentiment analysis of student satisfaction questionnaire comments at ITB STIKOM Bali. Journal of Applied Informatics and Computing, 9(3), 794–802. https://doi.org/10.30871/jaic.v9i3.9617

Singgalen, Y. A. (2024a). Performance evaluation of SVM with synthetic minority over-sampling technique in sentiment classification. Jurnal Mantik, 8(1), 326–336. https://doi.org/10.35335/mantik.v8i1.5077

Singgalen, Y. A. (2024b). Comparative analysis of DT and SVM model performance with SMOTE in sentiment classification. KLIK: Kajian Ilmiah Informatika dan Komputer, 4(5), 2485–2494. https://doi.org/10.30865/klik.v4i5.1828

Tuan, T. A., Loan, D. T. T., & Kittiphattanabawon, N. (2025). Understanding students' sentiment from feedback with a new feature selection and semantics networks. International Journal of Data and Network Science, 9(1), 253–266. https://doi.org/10.5267/j.ijdns.2024.7.010

Widayanti, R., & Kasih, F. C. (2026). Comparative analysis of baseline IndoBERT, class-weighted IndoBERT, and SMOTE with Support Vector Machine for handling imbalanced sentiment classification in Indonesian. Jurnal Teknik Informatika (JUTIF), 7(3), 2857–2875. https://doi.org/10.52436/1.jutif.2026.7.3.5692

Yang, W., Pan, C., & Zhang, Y. (2022). An oversampling method for imbalanced data based on spatial distribution of minority samples SD-KMSMOTE. Scientific Reports, 12, 16820. https://doi.org/10.1038/s41598-022-21046-1