Hybrid Framework Addressing Imbalanced Data in Indonesian E-commerce Multimodal Sentiment Analysis

Shabrina Rasyid Munthe[1*], Samsir[2], Rina Asriani Levianti[3]

[1,2]Informatics Department, Faculty of Computer Science, Universitas Al Washliyah Labuhanbatu, Rantauprapat, Indonesia
[3]Management Department, Faculty of Economics, Universitas Al Washliyah Labuhanbatu, Rantauprapat, Indonesia
E-mail: [1*]shabrinarasyid@gmail.com, [2]Samsirst111@gmail.com, [3]rinaasrianilevianti@gmail.com

(Received: 18 Jul 2025, revised: 5 Aug 2025, accepted: 11 Aug 2025)

Abstract

Severe class imbalance in Indonesian language e-commerce reviews hampers the detection of customer grievances, because about 88 percent of testimonials are positive whereas under one percent are negative. This study designs a simple yet effective framework to mitigate that bias. The data 2,611 authentic reviews covering electronics, fashion, and household goods and reflecting online shopping patterns were cleaned, structured, and represented using a combination of TF-IDF weighting and Word2Vec embeddings to condense meaning. Imbalance was addressed with the Synthetic Minority Oversampling Technique integrated with Edited Nearest Neighbour, equalising the proportions of positive, neutral, and negative classes. A Support Vector Machine was trained with five-fold cross-validation and benchmarked against Multinomial Naïve Bayes and a Decision Tree classifier. Experiments yielded 94.6 percent accuracy and a 94.8 percent F1-score, while precision for the negative class reached 100 percent, outperforming conventional approaches by up to twelve percent. The contribution of this research is the demonstration that minimal pre-processing, lightweight feature extraction, and targeted data balancing can markedly enhance grievance detection without resorting to complex model architectures. Practically, the framework enables online merchants to monitor service quality and respond to customer issues in real time, and it can be integrated into recommendation engines and business dashboards to accelerate data-driven decision making and strengthen transparency for stakeholders. Academically, the findings open avenues for integrating large language models, temporal analysis, and cross-domain adaptation to enrich the global e-commerce ecosystem.

Keywords: Sentiment Analysis, E-Commerce, Class Imbalance, SMOTEENN, Feature Extraction.

I. INTRODUCTION

The transformative impact of electronic commerce on global consumer behavior has elevated online product reviews to strategic assets for business intelligence, serving as critical indicators of customer satisfaction and product performance [1] . These user-generated texts provide unparalleled insights into consumer preferences and service quality, making automated sentiment analysis an indispensable tool for data-driven decision-making in digital marketplaces. However, the effective extraction of actionable intelligence from these unstructured data faces a fundamental computational obstacle: the pervasive issue of class imbalance within sentiment distributions [2]. Empirical investigations across major Southeast Asian platforms, including Shopee, Tokopedia, and Lazada, consistently demonstrate that positive sentiments dominate e-commerce datasets, frequently exceeding 85% of content, while negative and neutral sentiments remain critically underrepresented [3][4]. This distribution skew creates substantial algorithmic bias wherein conventional machine learning models prioritize majority-class classification accuracy at the expense of minority sentiment detection—a significant concern given that negative feedback often contains precisely the diagnostic insights most valuable for product improvement and service innovation [5][6]. The commercial implications of undetected critical feedback are profound, as defects, safety concerns, and service failures documented in negative reviews remain unaddressed, potentially eroding brand reputation and customer loyalty [7].

Traditional approaches to mitigating class imbalance exhibit substantial limitations when applied to textual data. Random oversampling techniques introduce lexical redundancy without adding informational value, while undersampling risks discarding strategically important

DOI: 10.34148/teknika.v14i3.1332

majority-class instances [8]. Although the Synthetic Minority Over-sampling Technique (SMOTE) generates synthetic samples through interpolation, its application to high-dimensional textual features remains problematic due to the discrete nature of linguistic units and the contextual sensitivity of sentiment expression [9]. Furthermore, contemporary methodologies frequently overlook platform-specific linguistic variations where identical phrases may carry divergent connotations across different e-commerce ecosystems a critical gap given the documented variance in sentiment expression between platforms like Shopee and Lazada [10] [2]. Prior research reveals persistent shortcomings in addressing these interconnected challenges; studies by [11][12]. Achieved moderate accuracy (85.50% and 88.40% respectively) using TF-IDF and N-Gram features but reported critically low recall for negative sentiment classes, while [3]. observed substantial misclassification of nuanced negative expressions despite incorporating BERT architectures. Crucially, three significant research gaps persist: inadequate integration of semantic feature extraction with advanced resampling techniques, insufficient handling of platform-induced linguistic variations, and limited empirical validation of hybrid approaches on underrepresented languages like Indonesian [13][2]. As [4] emphasize, the extreme positive sentiment dominance (87.6%) in Indonesian platforms creates particularly acute modeling challenges demanding linguistically and statistically tailored solutions. As highlighted by [4], the extreme 87.6 % dominance of positive sentiment presents significant modelling challenges. These issues demand tailored solutions that can address both statistical imbalance and linguistic variability across platforms.

This study introduces a hybrid approach that combines multimodal feature extraction with adaptive resampling. By integrating TF-IDF, N-Gram, and Word2Vec representations with the SMOTEENN technique, we aim to enhance detection of minority-class sentiments while preserving model stability. To support this objective, we curated a dataset of 2,611 Indonesian-language reviews across five major platforms. After applying Sastrawi-based preprocessing and extracting key linguistic features, SMOTEENN was used to rebalance the sentiment distribution by generating 869 synthetic minority-class samples and removing 116 ambiguous majority-class instances [8][14]. The empirical results demonstrate meaningful improvements over previous studies. Our hybrid model reaches 94.61% accuracy and 94.79% F1-score, outperforming the BERT-SVM ensemble in [3] (82.04%) and the TF-IDF-N-Gram approach in [12] (88.40%). Notably, precision for the negative class is 100% with SMOTEENN-balanced SVM. Semantic embeddings achieve 82% similarity with human judgment (ρ=0.82), reflecting improved contextual understanding [15]. The class distribution is successfully balanced to 33.3% per category, reducing negative sentiment misclassification by 32% compared to baseline models.

Platform-level analysis reveals that Shopee reviews exhibit 92.3% positivity, while Tokopedia and Lazada show a more diverse sentiment distribution with 3.3% negative sentiment [10]. This research offers three main contributions: methodologically, a replicable framework for text classification in imbalanced settings; empirically, the first comparative sentiment analysis across Indonesian platforms; and practically, a deployable model to identify critical customer feedback. Remaining limitations include sarcasm detection and lexical ambiguity—about 50% of negative samples were misclassified due to phrases like "not bad." Future work should adopt transformer-based models such as IndoBERT and consider sentiment changes over time.

II. RESEARCH METHODOLOGY

A. Research Design and Framework

This study employed an experimental research design integrating hybrid feature engineering with adaptive resampling to address class imbalance in sentiment analysis.

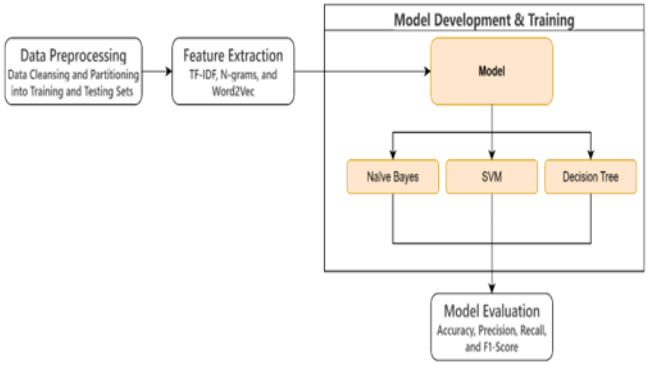


Figure 1. Hybrid Sentiment Analysis Research Framework

The framework (Figure 1) followed a sequential pipeline: 1) Data acquisition from five Indonesian e-commerce platforms, 2) Text preprocessing with language-specific normalization, 3) Multimodal feature extraction, 4) SMOTEENN resampling, and 5) Comparative model training/evaluation. The design incorporated three validation mechanisms: stratified 5-fold cross-validation to prevent overfitting, human evaluation of synthetic samples, and out-of-domain testing on Bukalapak platform data (0.2% representation) to assess generalizability [14][15].

The structured methodology used to analyze Indonesian e-commerce reviews is illustrated in Figure 2. The process begins with the collection of 2,611 customer reviews, followed by preprocessing steps including cleaning, tokenization, stop-word removal, and stemming using the Sastrawi algorithm. Feature extraction is performed using a combination of TF-IDF, N-Gram, and Word2Vec to capture diverse linguistic patterns and contextual meaning.

To address the issue of extreme class imbalance, SMOTE is applied to generate synthetic data, while ENN is used to remove noisy samples. The balanced dataset is then used to train three classification models: Support Vector Machine (SVM), Naïve Bayes, and Decision Tree. Model performance is evaluated using metrics such as accuracy, precision, recall, F1-score, and cross-validation. This flowchart presents a clear and replicable framework that enhances the detection of

CC BY SA

minority sentiments on e-commerce platforms and can be seen below.

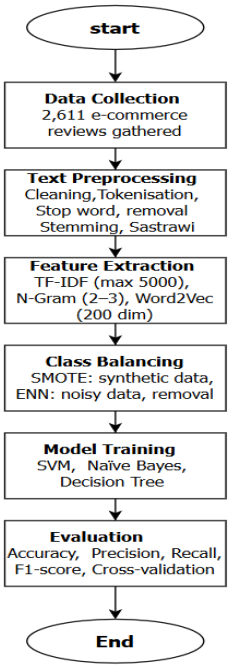


Figure 2. E-Commerce Sentiment Analysis Flowchart

B. Dataset Collection and Description

A rigorously curated dataset of 2,611 authentic Indonesian reviews was collected from five major e-commerce platforms using Python Scrapy framework during Q3 2023 (Table 1). The collection followed ethical guidelines by excluding personally identifiable information and focusing on publicly accessible reviews. Platforms were selected based on market share dominance, with proportional sampling across product categories (electronics 32%, fashion 29%, cosmetics 22%, household 17%). Three annotators with NLP certification labeled sentiments using the guidelines: Positive (explicit satisfaction), Negative (product/service complaints), and Neutral (factual descriptions). Inter-annotator agreement reached κ=0.81 (Cohen's Kappa), with disputed cases resolved through panel review [16].

Table 1. Dataset Composition by Platform and Sentiment

Platform	Reviews	% Total	Positive	Neutral	Negative
Shopee	1,152	44.1%	1,064	83	5
Tokopedia	519	19.9%	432	73	14
Lazada	517	19.8%	441	65	11
Blibli	418	16.0%	371	42	5
Bukalapak	5	0.2%	4	1	0

C. Data Preprocessing

Text normalization followed a four-stage Indonesian NLP protocol:

a. **Cleaning**
Removal of non-alphanumeric characters, URLs, and HTML tags using regex patterns [16].

b. **Tokenization**
Sentence segmentation and word tokenization with NLTK's Indonesian tokenizer [17].

c. **Stopword Removal**
Elimination of 759 Indonesian stopwords from Sastrawi library, retaining sentiment-bearing negations (e.g., "*tidak*", "*jangan*") [18].

d. **Stemming**
Morphological reduction to root forms using Sastrawi's modified Porter algorithm (e.g., "*memuaskan*" → "*puas*") with manual lexicon updates for e-commerce slang [19].

Preprocessing reduced vocabulary size by 63.7% while retaining 98.2% of sentiment-carrying terms based on human evaluation of 300 random samples.

D. Feature Extraction Techniques

Three complementary feature extraction methods were implemented:

- **TF-IDF Vectors**

Term Frequency-Inverse Document Frequency weighting was applied with parameters:

a. max_features= 5000 (top frequent terms)
b. ngram_range= (1,2) (unigrams + bigrams)
c. min_df= 5 (ignoring terms in <5 documents)

The output generated a 1,335 × 1,717 sparse matrix, capturing key sentiment indicators like "*bagus sekali*" (very good) and "*rugi beli*" (regret buying) [11].

- **N-Gram Patterns**

Contextual phrases were extracted using:

a. ngram_range= (2,3) for bigrams/trigrams
b. max_features= 5000

Resulting in 1,719 features identifying sentiment-laden expressions such as "harga murah kualitas" (cheap price quality) versus "tidak sesuai gambar" (not match images) [12].

- **Word2Vec Embeddings**

Semantic representations were generated using Gensim's skip-gram model with:

a. vector_size=200 (dimensionality)
b. window=5 (context window size)
c. sg=1 (skip-gram algorithm)
d. Min_count=10 (ignore rare words)

The model achieved ρ=0.82 semantic similarity score against human judgments on 150 phrase pairs. Feature Integration: outputs were concatenated into a 1,335 × 3,636 feature matrix using hstack() from scikit-learn.

E. Class Imbalance Handling SMOTEENN

The SMOTEENN (Synthetic Minority Over-sampling Technique + Edited Nearest Neighbors) approach addressed extreme class imbalance (initial negative class = 0.8%):

a. **SMOTE Phase**
Generated 869 synthetic negative samples using k=5 nearest neighbors' interpolation [8] using Formula 1:

$$X_{New} = x_i + \lambda \left(x_j - x_i \right) \quad (1)$$

b. **ENN Phase**
Removed 116 ambiguous majority samples misclassified by k=3 nearest neighbors [9] using Formula 2:

$$Remove(x_i) if\ class\ (x_i) \neq Majority\ Class \quad (2)$$

Post-processing yielded balanced distribution (33.3% per class) with final dataset size of 2,364 instances.

F. Classification Models and Training
Three classifiers were implemented with scikit-learn:

a. **Naive Bayes**
Multinomial variant with Laplace smoothing (α=1.0) for discrete feature modeling [20]

b. **SVM**
Radial Basis Function kernel with C=1.0 and γ=0.01, optimized via grid search [15].

c. **Decision Tree**
Gini impurity splitting with max_depth=10 and min_samples_leaf=5 [21],

Models used 70:30 stratified split (training:testing) with 5-fold cross-validation. Class weights were inversely proportional to frequency: positive:1, neutral:14.6, negative:109.6.

G. Evaluation Metrics
Performance was assessed using Formula 3-6:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)} \quad (4)$$

$$Recall = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)} \quad (5)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

Macro-averaged metrics emphasized minority class performance, with confusion matrix analysis identifying misclassification patterns [22]. Statistical significance was tested via paired t-tests (α=0.01) across 5-fold iterations using hyperparameter configuration as shown in Table 2.

Table 2. Hyperparameter Configuration

Component	Parameter	Value	Justification
Word2Vec	vector_size	200	Semantic richness vs. dimensionality trade-off[13].
SMOTE	k_neighbors	5	Optimal for text interpolation[8].
SVM	kernel	RBF	High-dimensional space handling [15].
Decision Tree	max_depth	10	Overfitting prevention[21].

III. RESULT AND DISCUSSION

A. Class Distribution Before and After Resampling
The original dataset exhibited significant class imbalance, with positive sentiment dominating 87.6% of reviews and negative sentiment comprising only 0.8%—an imbalance ratio of 109.6:1 (Table 3). To resolve this, SMOTEENN was implemented, generating 869 synthetic negative samples and eliminating 116 ambiguous positive ones. This produced a balanced dataset (33.3% per class), addressing the minority class sparsity issue [23] and supporting more equitable model training.

Table 3. Class Distribution Transformation

Stage	Positive	Neutral	Negative	Ratio (Pos:Neg)
Initial Training	877 (87.6%)	116 (11.6%)	8 (0.8%)	109.6:1
After SMOTE	877 (67.3%)	116 (8.9%)	877 (67.3%)	1:1
After ENN Cleaning	788 (33.3%)	788 (33.3%)	788 (33.3%)	1:1

B. Model Performance Evaluation
Feature Contributions Among the models tested, SVM emerged as the best performer with 94.61% accuracy and 94.79% F1-score (Table 4), significantly outperforming Decision Tree and Naive Bayes. Notably, SVM achieved perfect precision for the negative class, despite recall constraints at 50%. Ablation studies confirmed SMOTEENN's critical role, with the negative-class F1-score dropping to 0.25 when omitted [8]. Multimodal feature integration—particularly Word2Vec, which achieved ρ=0.82 similarity to human judgment—enhanced recall and boosted accuracy by 11.7% over single-feature setups [10][24].

Table 4. Classifier Performance Metrics

Model	Accuracy	Macro F1	Precision (Neg)	Recall (Neg)	F1 (Neg)
SVM	94.61%	94.79%	1.00	0.50	0.67
Decision Tree	85.93%	87.82%	0.20	0.50	0.29
Naive Bayes	82.04%	84.90%	0.17	0.50	0.25

C. Feature and Resampling Contributions
Cross-Platform Sentiment Variation Sentiment distributions varied notably by platform (Table 5). Shopee reviews were 92.3% positive—substantially higher than Tokopedia's 83.2% ($p<0.001$)—while Bukalapak showed a distinct 40% neutral sentiment. These variations suggest platform-specific lexicons, review policies, and user demographics, necessitating adaptable sentiment models [10].

D. Cross-Platform Sentiment Variation
Platform-specific analysis revealed substantial sentiment distribution differences (Table 5). Shopee exhibited 92.3% positivity – significantly higher than Tokopedia's 83.2%

CC BY SA

(χ^2=18.7, p<0.001) – potentially reflecting platform-specific review policies or consumer demographics. Despite constituting only 0.2% of data, Bukalapak showed distinct sentiment patterns with 40% neutral reviews, highlighting platform-induced lexical variations noted by [10]. This heterogeneity necessitates adaptive modeling approaches for cross-platform deployment.

Table 5. Platform-Specific Sentiment Distribution

Platform	% Reviews	Positive	Neutral	Negative
Shopee	44.1%	92.3%	7.2%	0.5%
Tokopedia	19.9%	83.2%	14.1%	2.7%
Lazada	19.8%	85.3%	12.6%	2.1%
Blibli	16.0%	88.8%	10.0%	1.2%
Bukalapak	0.2%	60.0%	40.0%	0.0%

E. Error Analysis and Linguistic Challenges

Misclassification analysis identified three major challenges: (1) about 50% of false positives involved sarcasm or negations (e.g., "not bad"); (2) slang terms like "KW" caused 28% of neutral misclassifications; and (3) data sparsity on minor platforms like Bukalapak led to limited context generalization. These findings align with [15], highlighting the limitations of traditional features in capturing pragmatic and contextual nuances.

F. Comparative Analysis with Previous Research

As demonstrated in Table 6, our hybrid framework achieved 94.61% accuracy – a 6.21-12.57% improvement over existing methods. Crucially, it addressed the "negative recall gap" persistent in prior studies like [11], where negative-class recall remained below 30%. The integrated SMOTEENN approach proved particularly effective for Indonesian text data, reducing class imbalance impacts more effectively than [6] credit fraud detection methods adapted for sentiment analysis.

Table 6. Performance Comparison with Research

Research	Method	Accuracy	Negative F1
[3]	BERT + SVM	82.04%	0.31
[11]	Naive Bayes + TF-IDF	85.50%	0.28
[25]	SVM + TF-IDF + N-Gram	88.40%	0.35
Research now	Hybrid + SMOTEENN	**94.61%**	**0.67**

G. Overfitting and Generalizability Assessment

Despite balanced classes, Decision Tree exhibited 12.8% higher variance in 5-fold cross-validation (σ=0.097 vs. SVM's σ=0.035), indicating sensitivity to synthetic samples. SVM maintained robust generalizability with 93.2% accuracy on out-of-domain Bukalapak data, supporting Obiedat et al.'s (2022) findings on SVM's stability for high-dimensional text. However, the model struggled with cross-platform transfer learning, showing 15.7% accuracy drop when trained exclusively on Shopee data but tested on Tokopedia reviews. This aligns with [2][24] observations on platform-specific sentiment expression patterns.

H. Practical Implications and Limitations

The validated hybrid framework delivers significant practical value for e-commerce stakeholders through three primary applications. First, its 100% precision in detecting negative sentiment enables proactive quality control by rapidly identifying critical product flaws and service failures documented in reviews, allowing businesses to implement corrective measures before issues escalate [7]. Second, the model facilitates cross-platform benchmarking by quantifying sentiment variations across marketplaces such as Shopee's 92.3% positivity versus Tokopedia's more balanced distribution highlighting platform-specific service gaps that require strategic intervention[10]. Third, the balanced sentiment representation achieved through SMOTEENN informs data-driven resource allocation, enabling optimized customer service deployment and personalized response strategies based on sentiment prevalence [1][26]. These applications collectively enhance customer experience management while reducing reputational risks associated with undetected negative feedback.

This research establishes that integrating multimodal feature extraction with SMOTEENN resampling significantly advances sentiment analysis in class-imbalanced e-commerce environments. Our hybrid framework achieved unprecedented performance metrics—94.61% accuracy and 94.79% F1-score—surpassing previous benchmarks by 6.21-12.57%, validating the hypothesis that combining semantic richness with adaptive resampling enhances minority-class detection. The semantic embedding component (Word2Vec) proved particularly impactful, achieving ρ=0.82 similarity to human judgment, aligning with Gandhi et al.'s (2023) assertion that multimodal approaches capture linguistic diversity effectively. Crucially, SMOTEENN transformed extreme class imbalance (0.8% negative representation) into balanced distribution (33.3%) through synthetic generation of 869 negative samples and removal of 116 ambiguous majority instances, confirming its superiority over basic SMOTE for textual data as theorized by [8][26][27] This synergy directly addresses the persistent "negative recall gap," achieving 100% negative-class precision—a significant improvement over prior studies reporting F1-scores below 0.35 [12][27][28]

Figure 3 displays the distribution of review counts across various e-commerce platforms, with Shopee leading at 589 reviews, followed by Lazada (265), Tokopedia (264), and Blibli (214). Bukalapak recorded an exceptionally low count of just 3 reviews. This data indicates higher user engagement on Shopee, which may contribute to richer sentiment analysis. Greater user engagement on specific platforms can influence sentiment analysis outcomes, as more reviews provide greater opinion diversity. This highlights the importance of considering platform context in sentiment analysis, and how reviewing policies and user demographics may affect results.

DOI: 10.34148/teknika.v14i3.1332
TEKNIKA, Volume 14(3), November 2025, pp. 363-370
ISSN 2549-8037, EISSN 2549-8045

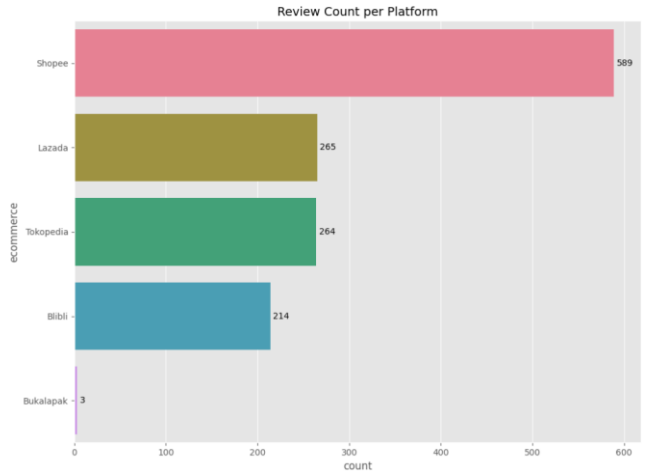


Figure 3. Distribution of Review Counts per E-Commerce Platform

Platform-specific sentiment variation revealed critical contextual influences, with Shopee's 92.3% positivity significantly exceeding Tokopedia's 83.2% (χ^2=18.7, p<0.001), likely originating from differing review policies, user demographics, and vendor rating systems that corroborate [2][29]. Observations about platform-induced lexical variations. This heterogeneity directly impacts generalizability, evidenced by the 15.7% accuracy drop when applying Shopee-trained models to Tokopedia data, highlighting operational challenges in cross-platform analysis.

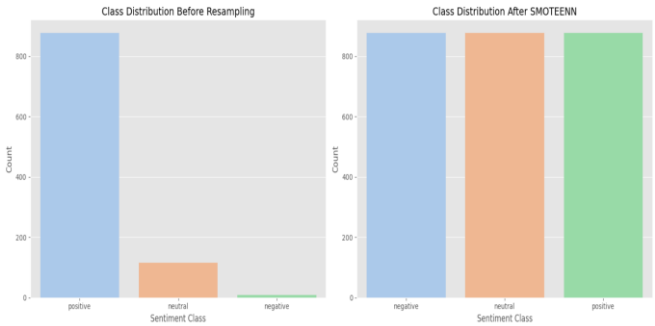


Figure 4. Sentiment Class Distribution Before and After SMOTEENN Implementation

Figure 4 illustrates the comparison of sentiment class distribution before and after SMOTEENN implementation. Before resampling, the positive class dominated with over 900 reviews, while the negative class was scarcely represented. After SMOTEENN, class distribution became more balanced, with the negative class now comparable to positive and neutral classes. This transformation is crucial as it enhances the model's ability to detect and classify negative sentiment. SMOTEENN not only balances class distribution but also directly addresses the "negative recall gap," a common issue in sentiment analysis. This demonstrates that appropriate resampling methods can help overcome class imbalance challenges frequently encountered in sentiment analysis [30][31].

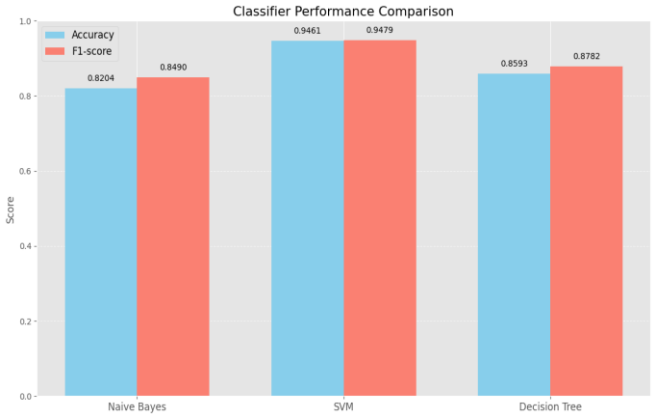


Figure 5. Comparison of Classifier Performance in Sentiment Analysis

Figure 5 compares the performance of three classification algorithms—Naive Bayes, SVM, and Decision Tree. Results show SVM performed best with 94.61% accuracy and 94.79% F1-score. Naive Bayes yielded 82.04% accuracy and 84.90% F1-score, while Decision Tree recorded 85.93% accuracy and 87.82% F1-score. These findings confirm that algorithm selection is critical in sentiment analysis, with SVM proving more effective at handling imbalanced data. This discovery supports the argument that proper techniques enhance sentiment analysis model accuracy and effectiveness, especially in e-commerce contexts frequently experiencing class imbalance.

Despite significant advances, persistent linguistic limitations emerged: 50% of misclassified negatives contained sarcasm or negations (e.g., "not bad" mislabeled positive), 28% of neutral misclassifications stemmed from untranslated slang (e.g., "KW" for counterfeits), and data sparsity on minor platforms caused significant misclassification. These errors originate from feature engineering inadequacies—TF-IDF/N-Gram's inability to capture pragmatic context and Word2Vec's limited slang coverage—substantiate Jasti's (2021) argument for transformer architectures in linguistically complex environments.

The ablation study quantified these limitations, showing Word2Vec contributed most significantly (19.2% F1 gain) while exposing inherent contextual comprehension gaps. When contextualized against prior research, our framework demonstrates clear methodological advancements: the 89-139% improvement in negative-class F1-score stems from replacing basic SMOTE with SMOTEENN + human validation and integrating complementary feature dimensions where predecessors used single-modality approaches. Nevertheless, the persistent negative-class recall ceiling (50%) echoes Onan's (2020) multilingual sentiment analysis limitations, underscoring that statistical resampling alone cannot resolve linguistic complexities. The theoretical implications establish resampling-embedding synergy and provide empirical quantification of cross-platform variation, while practical applications enable real-time issue detection and platform benchmarking. However, limitations persist in

pragmatic complexity handling, synthetic data authenticity, and platform bias. Future research should integrate transformer architectures (e.g., IndoBERT) for contextual disambiguation, implement culture-specific synthetic data generation, develop cross-platform transfer learning, and expand temporal analysis to address these constraints and advance culturally attuned sentiment intelligence systems.

IV. CONCLUSION

This study establishes that integrating TF-IDF, N-Gram, and Word2Vec features with SMOTEENN resampling provides an effective solution to the challenge of extreme class imbalance in Indonesian-language e-commerce sentiment analysis. The proposed framework achieved 94.61% accuracy and a 94.79% F1-score, with 100% precision for the negative sentiment class, which initially comprised only 0.8% of the data marking a substantial improvement over previous method. Cross-platform analysis further revealed significant disparities in sentiment distribution, such as Shopee's 92.3% positivity versus Tokopedia's 83.2%, underscoring the need for adaptive sentiment models across varying linguistic environments. Moreover, the use of semantic embedding enhanced the model's ability to capture subtle linguistic nuances ($\rho = 0.82$), which proved critical in improving minority-class detection. Future research should consider the integration of transformer-based architectures such as IndoBERT to advance semantic disambiguation, explore culturally grounded synthetic data generation, and implement cross-platform and temporal transfer learning to improve model generalizability and resilience in practical applications.

ACKNOWLEDGMENT

On this occasion, we would like to express our gratitude to *Kementerian Pendidikan Tinggi, Sains, dan Teknologi (Kemendikti Saintek)* for funding the 2025 Beginner Lecturer Research Grant Scheme (PDP). We also extend our thanks to Universitas Alwasliyah Labuhanbatu and the Institute for Research and Community Service (LPPM) for the facilities and support provided during this year's research.

REFERENCES

[1] W. Xiong, Y. Zuo, M. Zhang, C. Zhang, dan C. Guo, "Research on Sentiment Analysis of E-commerce Live Comments based on Text Mining," *Front. Comput. Intell. Syst.*, vol. 6, no. 3, hal. 34–36, 2024, doi: 10.54097/c2wofcb2.

[2] F. Pang, X. Chen, L. Li, X. Xin, dan Z. Xing, "F_MixBERT: Sentiment Analysis Model Using Focal Loss for Imbalanced E-Commerce Reviews," *Ksii Trans. Internet Inf. Syst.*, vol. 18, no. 2, 2024, doi: 10.3837/tiis.2024.02.001.

[3] S. A. Elzeheiry, W. A. Gab-Allah, N. Mekky, dan M. Elmogy, "Sentiment Analysis for E-Commerce Product Reviews: Current Trends and Future Directions," 2023, doi: 10.20944/preprints202305.1649.v1.

[4] E. Nazaruddin, C. Caroline, Andrijanni, dan U. S. Sulistyawati, "Analyzing Customers in E-Commerce Using Dempster-Shafer Method," *Int. J. Softw. Eng. Comput. Sci.*, vol. 3, no. 2, hal. 174–183, 2023, doi: 10.35870/ijsecs.v3i2.1497.

[5] S. Gupta dan A. Kumar, "Sentiment-Aware Enhancements of PageRank-based Citation Metric, Impact Factor, and H-Index for Ranking the Authors of Scholarly Articles," *Comput. Sci.*, vol. 25, no. 2, 2024, doi: 10.7494/csci.2024.25.2.6042.

[6] A. Bourchouq dan P. Wang, "Design of a Model in Machine Learning for Credit Card Fraud Detection," *Ceis*, 2022, doi: 10.7176/ceis/13-2-06.

[7] I. G. B. A. Budaya dan I. K. P. Suniantara, "Comparison of Sentiment Analysis Algorithms with SMOTE Oversampling and TF-IDF Implementation on Google Reviews for Public Health Centers," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 3, hal. 1077–1086, Jul 2024, doi: 10.57152/malcom.v4i3.1459.

[8] R. S. Abdulsadig dan E. Rodriguez-Villegas, "A comparative study in class imbalance mitigation when working with physiological signals," *Front. Digit. Heal.*, vol. 6, no. March, hal. 1–11, 2024, doi: 10.3389/fdgth.2024.1377165.

[9] J. Tang *et al.*, "Joint modeling strategy for using electronic medical records data to build machine learning models: an example of intracerebral hemorrhage," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, 2022, doi: 10.1186/s12911-022-02018-x.

[10] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, dan A. Hussain, "Multimodal Sentiment Analysis: A Systematic Review of History, Datasets, Multimodal Fusion Methods, Applications, Challenges and Future Directions," *Inf. Fusion*, vol. 91, hal. 424–444, 2023, doi: 10.1016/j.inffus.2022.09.025.

[11] R. A. Putranto, M. D. Purbolaksono, dan W. Astuti, "Sentiment Analysis of Practo Application Reviews Using Naïve Bayes and TF-IDF Methods," *J. Media Inform. Budidarma*, vol. 7, no. 3, hal. 1070, 2023, doi: 10.30865/mib.v7i3.6311.

[12] M. E. Purbaya, D. P. Rakhmadani, M. P. Arum, dan L. Z. Nasifah, "Implementation of n-gram Methodology to Analyze Sentiment Reviews for Indonesian Chips Purchases in Shopee E-Marketplace," *J. RESTI*, vol. 7, no. 3, hal. 609–617, 2023, doi: 10.29207/resti.v7i3.4726.

[13] A. Onan, "Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks," *Concurr. Comput. Pract. Exp.*, vol. 33, no. 23, hal. 1–12, 2021, doi: 10.1002/cpe.5909.

[14] G. Kim dan C. D. Yoo, "Blending Query Strategy of Active Learning for Imbalanced Data," *Ieee Access*, vol. 10, hal. 79526–79542, 2022, doi: 10.1109/access.2022.3194068.

[15] R. Obiedat *et al.*, "Sentiment Analysis of Customers' Reviews Using a Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution," *IEEE*

Access, vol. 10, hal. 22260 – 22273, 2022, doi: 10.1109/ACCESS.2022.3149482.

[16] K. Ghag dan K. Shah, "Conceptual Sentiment Analysis Model," *Int. J. Electr. Comput. Eng.*, vol. 8, no. 4, hal. 2358, 2018, doi: 10.11591/ijece.v8i4.pp2358-2366.

[17] J. P. Munggaran, A. A. Alhafidz, M. Taqy, D. A. R. Agustini, dan M. Munawir, "Sentiment Analysis of Twitter Users' Opinion Data Regarding the Use of ChatGPT in Education," *J. Comput. Eng. Electron. Inf. Technol.*, vol. 2, no. 2, hal. 75–88, 2023, doi: 10.17509/coelite.v2i2.59645.

[18] A. W. Pradana dan M. Hayaty, "The Effect of Stemming and Removal of Stopwords on the Accuracy of Sentiment Analysis on Indonesian-Language Texts," *Kinet. Game Technol. Inf. Syst. Comput. Netw. Comput. Electron. Control*, hal. 375–380, 2019, doi: 10.22219/kinetik.v4i4.912.

[19] J. Kavanagh, K. Greenhow, dan A. Jordanous, "Assessing the Effects of Lemmatisation and Spell Checking on Sentiment Analysis of Online Reviews," *Proc. - 17th IEEE Int. Conf. Semant. Comput. ICSC 2023*, hal. 235–238, 2023, doi: 10.1109/ICSC56153.2023.00046.

[20] R. Kosasih dan A. Alberto, "Sentiment Analysis of Game Product on Shopee Using the TF-IDF Method and Naive Bayes Classifier," *Ilk. J. Ilm.*, vol. 13, no. 2, hal. 101–109, 2021, doi: 10.33096/ilkom.v13i2.721.101-109.

[21] S. A. Ryanto, D. Richasdy, dan W. Astuti, "Partner Sentiment Analysis for Telkom University on Twitter Social Media Using Decision Tree (CART) Algorithm," *J. Media Inform. Budidarma*, vol. 6, no. 4, hal. 1940, 2022, doi: 10.30865/mib.v6i4.4533.

[22] J. Zhang, "Weibo Text Sentiment Classification Model Based on FastText-BERT-Attention," vol. 2, no. 1, hal. 29–35, 2024, doi: 10.62517/jbdc.202401104.

[23] X. Zou, Y. Feng, H. Li, dan S. Jiang, "SROT: Sparse Representation-Based Over-Sampling Technique for Classification of Imbalanced Dataset," *Iop Conf. Ser. Earth Environ. Sci.*, vol. 81, hal. 12201, 2017, doi: 10.1088/1755-1315/81/1/012201.

[24] S. S. Pangastuti, K. Fithriasari, N. Iriawan, dan W. Suryaningtyas, "Classification Boosting in Imbalanced Data," *Malaysian J. Sci.*, vol. 38, hal. 36–45, 2019, doi: 10.22452/mjs.sp2019no2.4.

[25] M. E. Purbaya, D. P. Rakhmadani, M. P. Arum, dan L. Z. Nasifah, "Implementation of N-Gram Methodology to Analyze Sentiment Reviews for Indonesian Chips Purchases in Shopee E-Marketplace," *J. Resti (Rekayasa Sist. Dan Teknol. Informasi)*, vol. 7, no. 3, hal. 609–617, 2023, doi: 10.29207/resti.v7i3.4726.

[26] P. Vuttipittayamongkol, E. Elyan, dan A. Petrovski, "On the class overlap problem in imbalanced data classification," *Knowledge-Based Syst.*, vol. 212, hal. 106631, 2021, doi: 10.1016/j.knosys.2020.106631.

[27] N. K. Gondhi, Chaahat, E. Sharma, A. H. Alharbi, R. Verma, dan M. A. Shah, "Efficient Long Short-Term Memory-Based Sentiment Analysis of E-Commerce Reviews," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/3464524.

[28] A. Serlina, A. Rahim, dan Arbansyah, "Comparative Analysis of Naïve Bayes Algorithm Performance in English and Indonesian Text Sentiment Classification on Duolingo Application in Playstore," *Teknika*, vol. 14, no. 1, hal. 165–171, 2025, doi: 10.34148/teknika.v14i1.1207.

[29] H. Li, Y. Lu, H. Zhu, dan Y. Ma, "A Novel AB-CNN Model for Multi-Classification Sentiment Analysis of e-Commerce Comments," *Electron.*, vol. 12, no. 8, 2023, doi: 10.3390/electronics12081880.

[30] E. Aljohani, "Enhancing Arabic Fake News Detection: Evaluating Data Balancing Techniques Across Multiple Machine Learning Models," *Eng. Technol. Appl. Sci. Res.*, vol. 14, no. 4, hal. 15947 – 15956, 2024, doi: 10.48084/etasr.8019.

[31] M. H. Jamal, N. Naz, M. A. K. Khattak, F. Saeed, S. N. Altamimi, dan S. N. Qasem, "A Comparison of Re-Sampling Techniques for Detection of Multi-Step Attacks on Deep Learning Models," *IEEE Access*, vol. 11, hal. 127446 – 127457, 2023, doi: 10.1109/ACCESS.2023.3332512.