JSCE (Journal of System and Computer Engineering)
Vol.7, No.3, July 2026, pp. 229-240
ISSN (online): 2723-1240
DOI : https://doi.org/10.61628/jsce.v7i3.2688

Comparison of SMOTE, Class Weighting, and Classical Machine Learning Models on the ID-SMSA Indonesian Stock Market Dataset

I Komang Adyanata [1,a,*]; I Gede Aris Gunadi[2,b]; I Made Gede Sunarya[3,c]
[1,2,3] *Program Studi Ilmu Komputer, Pascasarjana, Universitas Pendidikan Ganesha, Jl. Udayana No.11, Singaraja 81116, Indonesia*
[a] *ikomangadyanata@gmail.com;* [b] *igedearisgunadi@undiksha.ac.id;* [c] *sunarya@undiksha.ac.id*
** Corresponding author*

Abstract

Sentiment classification of social-media text related to the Indonesian stock market is a growing research area. The ID-SMSA dataset is the publicly available labelled corpus for this domain, yet class-imbalance handling strategies on this dataset have not been systematically compared across multiple classifiers. This paper evaluates Multinomial Naive Bayes, linear Support Vector Machine (SVM), and Random Forest under three imbalance-handling conditions: no handling, class weighting, and SMOTE. All experiments use 3,287 tweets (1 tweet was removed from the original 3,288 during preprocessing after producing an empty data) with an 80:20 stratified split and report macro F1 as the primary metric. SMOTE consistently improves macro F1 across all classifiers. The largest gain is on Naive Bayes (+0.137, from 0.589 to 0.726). The best configuration is SVM with SMOTE, achieving macro F1 of 0.752 and accuracy of 0.784. Class weighting benefits Random Forest (+0.011) but slightly reduces SVM, confirming that linear SVM on TF-IDF is robust to moderate imbalance at IR = 2.41. Per-issuer evaluation reveals macro F1 variation from 0.647 on TPIA to 0.881 on BBNI, shaped by vocabulary consistency, class dominance, and domain specificity. These results provide a transparent and reproducible classical baseline that situates transformer-based and deep-learning approaches on ID-SMSA within a well-defined reference frame.

Keywords: *Imbalanced Data, Sentiment Analysis, TF-IDF, Class Weighting, SMOTE.*

1. Introduction

The Indonesia Stock Exchange (IDX) has experienced a surge in retail investor participation over the past five years, driven largely by younger demographic cohorts who rely heavily on social-media platforms for market information and peer discussion. Platform X (formerly Twitter) occupies a central role in this ecosystem: opinions about specific issuers spread rapidly, and the aggregate sentiment of a discussion thread can inform, or misinform, trading decisions within hours. Automating the interpretation of such sentiment streams is therefore an increasingly important task for market analysts, risk managers, and regulators alike.

Indonesian-language sentiment analysis has been an active research area, with a growing body of work spanning multiple application domains. Classical machine-learning models, namely Naive Bayes, SVM, and Random Forest, have been applied to public-service feedback (Saputri et al., 2024), healthcare reviews (M. H. Setiawan et al., 2023), public opinion on COVID-19 vaccination (Raharjo et al., 2022), community responses to the pandemic itself (Ariyani et al., 2025), and public policy discourse on social media (A. Setiawan & Suryono,

2024). Pairwise comparisons of classical models are also common: Random Forest and SVM have been compared on student feedback (Sidik et al., 2025), and SVM has been coupled with part-of-speech tagging for aspect-based analysis (Sanjaya et al., 2024). The representation of text features has been studied as well; TF-IDF has been found to outperform bag-of-words for Indonesian text (Maharani et al., 2026). On the deep-learning side, transformer models based on IndoBERT have been employed for tourism-news classification (Dewi et al., 2024; Wijaya et al., 2025), and government-programme sentiment (Sari et al., 2026), among other applications.

In the financial domain, studies on Indonesian stock-market sentiment have been limited by the absence of a publicly labelled corpus. Ensemble machine-learning methods have been applied to financial-news articles from IDX-listed companies (Munthe et al., 2025), while the Indonesian Stock Market Sentiment Analysis (ID-SMSA) dataset, released in early 2025, provided a large, human-labelled collection of 3,288 tweets in Bahasa Indonesia covering the ten largest issuers by market capitalization (Hartanto, Liundi, et al., 2025a). The dataset opened a new benchmark space for Indonesian financial NLP. However, like many real-world social-media corpora, ID-SMSA exhibits class imbalance, a challenge that has received broad attention in the machine-learning literature.

Class imbalance is a pervasive challenge across machine learning tasks, with established mitigation strategies categorized into data re-balancing, feature representation, training strategy, and ensemble approaches (Gao et al., 2026). The ID-SMSA class distribution is inherently imbalanced: 53.8% of tweets are labelled Positive, 23.9% Negative, and 22.3% Neutral, yielding an imbalance ratio of 2.41. Under such conditions, models trained without mitigation tend to over-predict the majority class, inflating overall accuracy while producing poor minority-class recall, the very recall that carries the greatest decision-making value for investors monitoring negative signals. SVM was compared with IndoBERT-base and IndoBERT-large under an out-of-vocabulary evaluation framework with ID-SMSA dataset. However, that study used accuracy as the sole metric and did not apply imbalance handling (Simanihuruk et al., 2025). Four IndoBERT variants were trained with back-translation augmentation and class weighting (Hartanto, Sutoyo, et al., 2025). Yet, neither of those studies systematically compares multiple class-imbalance mitigation strategies across different classifier families, and none evaluate classical machine-learning models.

Among the studies reviewed, classical machine-learning models have not been benchmarked on ID-SMSA, leaving a gap in the baseline literature that limits comparability across studies. Direct precedent exists in the literature. SVM with SMOTE has improved classification on imbalanced health-domain data (Aziz et al., 2025), Random Forest with SMOTE and random under-sampling has benefited imbalanced customer-review sentiment (Istiqamah & Rijal, 2024), and SMOTE has been shown to improve both SVM and Naive Bayes on imbalanced Indonesian tweet classification (A. Setiawan & Suryono, 2024), all finding meaningful improvements over unmitigated baselines.

Motivated by this gap, the present study evaluates SMOTE and class weighting against a no-handling baseline across three TF-IDF-based classical classifiers on the complete ID-SMSA dataset, with three objectives: to quantify the effect of each technique on macro F1 and minority-class recall, to determine whether one technique is consistently superior across classifiers, and to assess per-issuer performance variation and its structural determinants. The contributions of this work are as follows:

- This study systematically compares three imbalance-handling strategies, namely no handling, class weighting, and SMOTE, across three classical machine-learning classifiers on the complete ID-SMSA dataset. Prior studies on this dataset have focused exclusively on transformer architectures without evaluating classical models and comparing multiple imbalance-handling strategies.
- It establishes a transparent and reproducible classical baseline on ID-SMSA, enabling future deep-learning and transformer studies to directly measure their performance gains against a well-defined reference point.

- It provides a per-issuer performance analysis that reveals how issuer-level corpus characteristics, such as vocabulary consistency, class distribution, and domain specificity, shape classifier behaviour beyond what aggregate metrics can capture.

Together, these contributions address a gap in the ID-SMSA literature by providing a classical machine-learning perspective that complements the transformer-based work of prior studies and serves as a reference point for future research.

2. Method

The overall research workflow is illustrated in Figure 1 and comprises five sequential stages: data acquisition and exploratory analysis, text preprocessing, feature representation with stratified data splitting, classifier training under three imbalance handling scenarios, and performance evaluation at both the corpus and per-issuer levels.

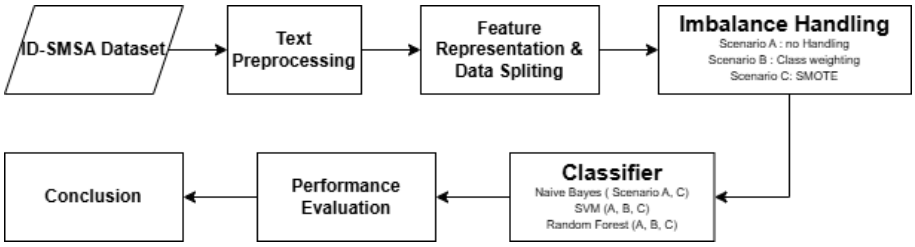

Figure 1. Research workflow diagram.

2.1 Dataset

The study uses *ID-SMSA* version 3, which is publicly available on Mendeley Data under a CC-BY 4.0 licence (Hartanto, Liundi, et al., 2025b). The dataset was constructed by crawling the public Twitter API for tweets mentioning the ten IDX issuers with the highest market capitalization as of March 2023: BBCA, BBRI, BMRI, BBNI, ASII, TLKM, UNVR, HMSP, TPIA, and BYAN. Collection covers the period from 14 December 2020 to 1 March 2024. Each tweet was independently labelled by two trained annotators as *Positive*, *Neutral*, or *Negative*; the resulting inter-annotator agreement, measured by Cohen's kappa, is 0.779(Hartanto, Liundi, et al., 2025a). The full corpus comprises 3,288 labelled tweets distributed as 1,769 Positive (53.8%), 786 Negative (23.9%), and 733 Neutral (22.3%), giving an imbalance ratio IR = 1769/733 = 2.41. After text preprocessing, 1 tweet produces an empty string and is discarded, reducing the corpus to 3,287 data used for modelling. Figure 2 visualizes the overall class distribution of the dataset.

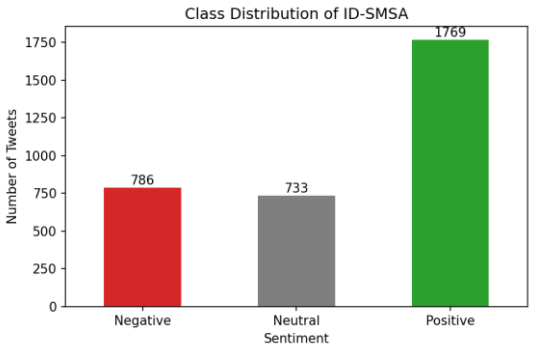

Figure 2. Overall class distribution of the ID-SMSA dataset.

2.2 Text Preprocessing

Text preprocessing transforms raw social-media content into a clean, normalized form suitable for feature extraction; the specific steps required depend heavily on the language and platform characteristics of the corpus (Hidayah et al., 2025). In this study, each tweet passes through a six-step preprocessing pipeline. (i) *Case folding* converts all characters to lowercase. (ii) *Placeholder removal* eliminates the tokens [USERNAME], [URL], and [HASHTAG] that the dataset creators inserted during annotation. (iii) *Ticker normalization* converts cashtag forms (e.g., $BBCA) to plain lowercase tokens (bbca), preserving issuer identity for downstream feature extraction. (iv) *Noise removal* strips residual URLs, user mentions, hashtags, punctuation, and digit strings. (v) *Stopword removal* applies the Sastrawi Indonesian stopword list, extended with a small hand-crafted block list for common domain-neutral financial terms such as "saham" (stock). (vi) *Stemming* applies via Sastrawi library, reducing inflected and derived forms to their base roots. After this pipeline, 1 tweet produces an empty string and is discarded, as noted in Section 2.1.

2.3 Feature Representation and Data Splitting

Text is represented as term frequency–inverse document frequency (TF-IDF) vectors with unigrams and bigrams and a vocabulary ceiling of 5,000 features. This choice is supported by empirical evidence that TF-IDF outperforms bag-of-words for Indonesian-language text classification (Maharani et al., 2026). The corpus is partitioned into a training set (80%, 2,629 documents) and a held-out test set (20%, 658 documents) using stratified sampling to preserve the original class proportions in both partitions. A fixed random seed (value = 42) is applied to all stochastic operations throughout the experiment to ensure full reproducibility. The TF-IDF vectoriser is fitted exclusively on the training partition so that no test-side lexical information influences feature selection or weighting.

2.4 Classifiers

Three classical text classifiers are employed. *Multinomial Naive Bayes* (NB) is a generative, probabilistic model whose parameters are estimated directly from word-count frequencies, and has been widely applied to Indonesian short-text sentiment (Ariyani et al., 2025; Raharjo et al., 2022; Saputri et al., 2024). *Linear Support Vector Machine* (SVM) learns a maximum-margin hyperplane in the high-dimensional TF-IDF feature space and is a consistently strong text-classification baseline, including for multi-class Indonesian-language text with unigram and bigram features (Indrawan et al., 2022; Munthe et al., 2025; Sanjaya et al., 2024; Simanihuruk et al., 2025). *Random Forest* (RF) constructs 200 decision trees on bootstrap samples and aggregates predictions by majority vote, and has shown meaningful improvements on imbalanced Indonesian text when combined with SMOTE (Istiqamah & Rijal, 2024; Sidik et al., 2025). Classical models are chosen because they have not been systematically evaluated on ID-SMSA with varied imbalance-handling strategies, they provide a transparent and reproducible reference frame for future deep-learning and transformer research on this dataset, and they remain computationally practical compared with pre-trained transformer models (Maharani et al., 2026).

2.5 Imbalance Handling Scenarios

Each classifier is evaluated under three scenarios, yielding eight classifier–scenario combinations in total. In the no-handling scenario, the raw training distribution is used without modification, serving as the performance baseline. In the class-weighting scenario, scikit-learn's balanced class-weight mode assigns sample weights inversely proportional to class frequencies, penalising misclassifications of minority-class instances more heavily without altering the dataset itself. In the SMOTE scenario, the *Synthetic Minority Oversampling Technique* synthesises new minority-class instances by linear interpolation between existing samples and

their k = 3 nearest neighbours in TF-IDF space; oversampling is applied exclusively to the training partition to prevent information leakage into the test set, following best practices established in comparable imbalanced sentiment classification studies (Nassr et al., 2025).

An important caveat applies to Naive Bayes: because its likelihood estimates are derived directly from raw word counts, the standard scikit-learn implementation does not support class-weight arguments for fitting. Consequently, only the no-handling and SMOTE scenarios are evaluated for NB.

2.6 Evaluation

Performance on the test set is reported through five metrics: accuracy, macro-averaged precision, macro-averaged recall, macro-averaged F1-score (macro F1), and weighted-average F1-score (F1-weighted). Macro F1 is designated as the primary comparison metric because it averages each class's F1 score without weighting by class size, thereby penalizing models that achieve high aggregate accuracy at the expense of minority-class recognition. Formal justification for this preference is provided for multi-class settings with unequal class frequencies, showing that weighted and micro averages can overstate model quality when the majority class is large (Hinojosa Lee et al., 2024). Macro F1 is particularly appropriate in settings with more than two classes, as demonstrated in multi-class Indonesian text classification studies where per-class performance varies substantially across label categories (Sillviari et al., 2025). Confusion matrices are also examined for the most informative configurations to characterise error distributions qualitatively. Following the overall evaluation, the best model configuration is applied to the test partition and its predictions are further stratified by issuer ticker, enabling a per-issuer assessment that surfaces differences in model performance across the ten IDX issuers covered by ID-SMSA.

3. Results And Discussion

3.1 Positioning Relative to Prior Work

Table 1 summarizes the three studies that directly used ID-SMSA alongside the present work. A notable feature of the prior literature is its exclusive reliance on transformer architectures: all three studies evaluate IndoBERT variants, whereas classical models have not yet been benchmarked on this dataset. Moreover, one prior study applied imbalance mitigation through class weighting combined with back-translation (Hartanto, Sutoyo, et al., 2025), one applied none (Simanihuruk et al., 2025). The Present study fills these gaps by providing a systematic multi-technique comparison on classical models while using macro F1 as the primary criterion.

Table 1. Prior studies that directly used ID-SMSA and the position of the present work.

Study	Models Used	Imbalance Handling
Simanihuruk et al. (2025)	SVM, IndoBERT-base, and IndoBERT-large.	None reported
Hartanto et al. (2025)	4 IndoBERT variants	Class weighting + back-translation
This study	Naive Bayes, SVM, and Random Forest.	No handling, Class-weighting, and SMOTE.

3.2 Per-issuer Sentiment Composition

Before examining classifier performance, Table 2 and Figure 3 describe the sentiment landscape within ID-SMSA at the issuer level. Among the four banking tickers, namely BMRI (73.5%), BBRI (70.6%), BBNI (66.3%), and BBCA (61.3%), positive sentiment dominates

overwhelmingly, broadly consistent with the strong aggregate performance of the Indonesian banking sector during 2021–2024. UNVR, by contrast, stands as the only issuer where Negative tweets constitute a majority (54.5%). Closer examination reveals that the negative discourse around UNVR is anchored by three recurring terms: "boikot" (boycott, 30 occurrences), "turun" (decline, 67 occurrences), and "produk" (product, 27 occurrences); lexical markers directly traceable to the consumer-boycott movement against Unilever products that intensified in late 2023 and had a demonstrable impact on the share price. HMSP and ASII exhibit more balanced distributions, with HMSP negative discourse centred on "cukai" (excise, 23 occurrences) and "rokok" (cigarette, 36 occurrences) alongside references to the competing tobacco issuer GGRM, reflecting the ongoing policy debates surrounding tobacco excise taxation. These observations confirm that ID-SMSA is not merely a generic opinion corpus but a domain-specific resource that encodes real macroeconomic and socio-political events.

Table 2. Per-issuer composition of ID-SMSA.

Issuer	Total	Positive	Neutral	Negative	Dominant Class
ASII	416	170 (40.9%)	101 (24.3%)	145 (34.9%)	Balanced
BBCA	442	271 (61.3%)	103 (23.3%)	68 (15.4%)	Positive
BBNI	371	246 (66.3%)	87 (23.5%)	38 (10.2%)	Positive
BBRI	551	389 (70.6%)	106 (19.2%)	56 (10.2%)	Positive
BMRI	426	313 (73.5%)	67 (15.7%)	46 (10.8%)	Positive
BYAN	386	213 (55.2%)	97 (25.1%)	76 (19.7%)	Positive
HMSP	386	147 (38.1%)	94 (24.4%)	145 (37.6%)	Balanced
TLKM	331	183 (55.3%)	71 (21.5%)	77 (23.3%)	Positive
TPIA	335	177 (52.8%)	81 (24.2%)	77 (23.0%)	Positive
UNVR	418	113 (27.0%)	77 (18.4%)	228 (54.5%)	Negative

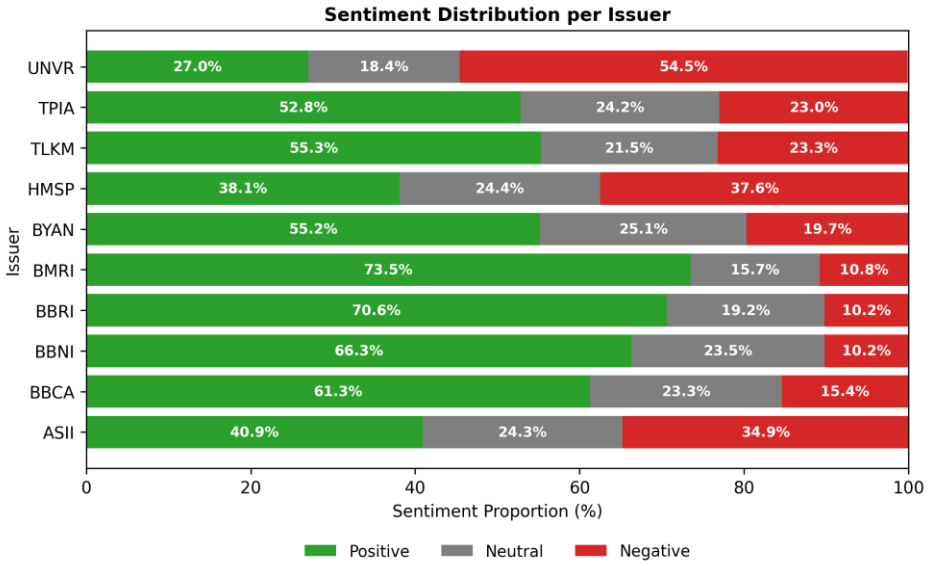

Figure 3. Sentiment distribution per issuer as proportions of each issuer's total tweets.

3.3 Overall Classification Performance

Table 3 presents all eight classifier–scenario combinations. Across the no-handling scenario, linear SVM achieves the highest macro F1 of 0.750, followed by Random Forest (0.671) and Naive Bayes (0.589), an ordering consistent with the relative ranking of SVM over Random Forest reported for short Indonesian text (Sidik et al., 2025).

Table 3. Full classification results

Classifier	Scenario	Acc.	P-mac	R-mac	F1-mac	F1-w
Naive Bayes	No handling	0.690	0.778	0.566	0.589	0.649
Naive Bayes	SMOTE	0.748	0.722	0.741	0.726	0.751
SVM (linear)	No handling	0.787	0.766	0.738	0.750	0.783
SVM (linear)	Class weighting	0.780	0.750	0.741	0.745	0.778
SVM (linear)	SMOTE	0.784	0.753	0.750	0.752	0.784
Random Forest	No handling	0.736	0.782	0.644	0.671	0.714
Random Forest	Class weighting	0.743	0.774	0.657	0.682	0.724
Random Forest	SMOTE	0.742	0.748	0.672	0.690	0.728

In Table 3, Acc. denotes accuracy; P-mac, R-mac, and F1-mac denote macro-averaged precision, recall, and F1-score, respectively; and F1-w denotes the weighted F1-score. Figure 4 visualises the macro F1 scores across models and imbalance-handling scenarios, making the direction and magnitude of each technique's effect immediately apparent.

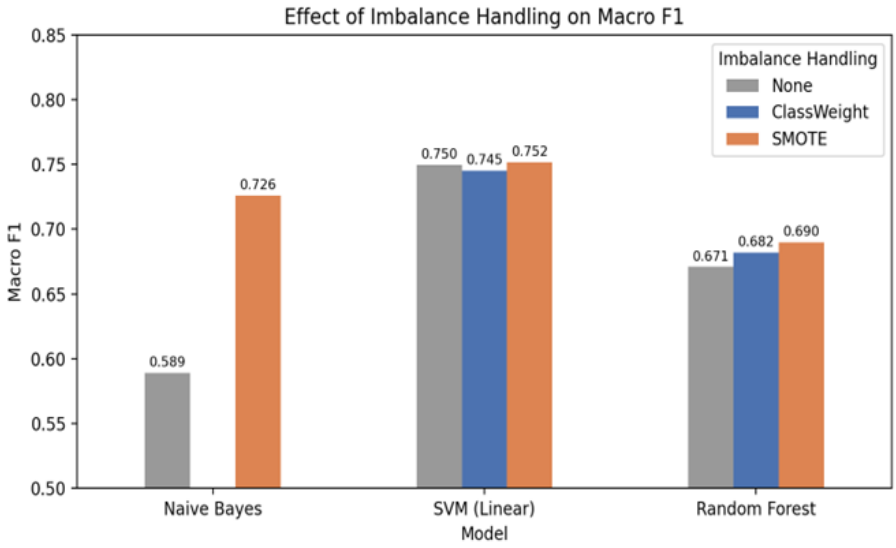

Figure 4. Macro F1 by classifier and imbalance-handling scenario.

SMOTE improves macro F1 for every classifier. The gain is most pronounced for Naive Bayes (+0.137, from 0.589 to 0.726) because NB is directly sensitive to marginal class frequencies: without oversampling, the model internalizes a strong prior toward the Positive class, and SMOTE counteracts this by equalizing apparent class frequencies in the training data. For Random Forest the improvement is moderate (+0.019), consistent with similar gains observed when applying SMOTE to Random Forest on imbalanced Indonesian text (Istiqamah & Rijal, 2024). For SVM the gain is marginal (+0.002), which is expected given that linear SVM with hinge loss on high-dimensional TF-IDF vectors is structurally robust to moderate imbalance. This pattern is also consistent with findings that SVM+SMOTE remains effective on imbalanced Indonesian classification tasks (Aziz et al., 2025).

Class weighting produces smaller and less consistent effects. It raises Random Forest macro F1 by 0.011, leaves Naive Bayes inapplicable, and slightly reduces SVM by 0.005. The SVM reduction is not statistically alarming at this scale, but suggests that at IR = 2.41 the linear-SVM decision boundary is already close to the globally optimal position, and inverse-frequency reweighting perturbs it marginally. Taking all results together, SVM with SMOTE emerges as the best overall configuration with macro F1 of 0.752 and accuracy of 0.784.

3.4 Confusion Matrix Analysis

Figure 5 displays confusion matrices for Naive Bayes without handling, Naive Bayes with SMOTE, and SVM with SMOTE. The unmitigated Naive Bayes matrix epitomizes the majority-class bias problem: recall for Positive reaches 0.96 while recall for Negative falls to 0.50 and, most strikingly, recall for Neutral collapses to 0.24. In practice, nearly three-quarters of Neutral tweets are mislabelled as Positive, rendering the model essentially a binary Positive/non-Positive discriminator rather than a three-class classifier.

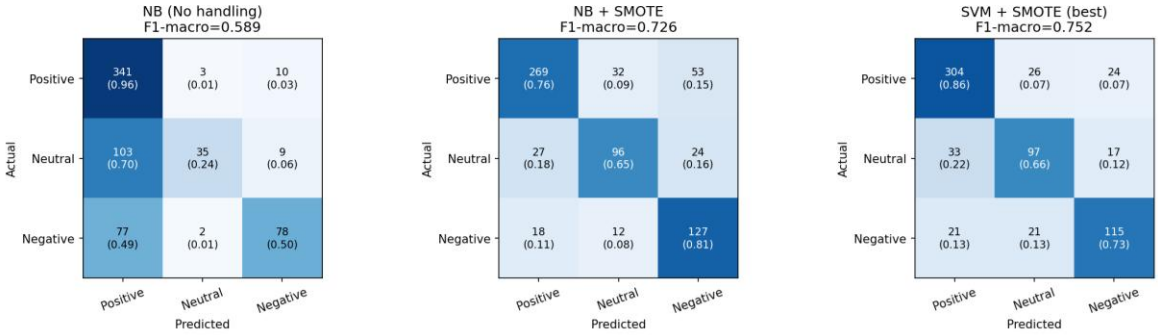

Figure 5. Confusion matrices for NB without handling (left), NB with SMOTE (center), and SVM with SMOTE (right).

SMOTE transforms this picture substantially. Negative recall increases to 0.81 (+0.31) and Neutral recall to 0.65 (+0.41), while Positive recall decreases to 0.76, a deliberate redistribution of predictive effort. SVM with SMOTE refines this balance further, attaining per-class recalls of 0.86, 0.66, and 0.73 for Positive, Neutral, and Negative, respectively. From a financial-analysis perspective, the improvement in Negative recall is particularly consequential: negative-sentiment signals constitute the most actionable early-warning indicators for risk management, and a classifier that misses half of them provides materially less value than one that captures nearly three-quarters.

3.5 Per-issuer Performance of The Best Model

Table 4 and Figure 6 decompose SVM+SMOTE predictions by issuer on the 658-tweet test set. Macro F1 ranges from 0.881 on BBNI to 0.647 on TPIA, a spread of more than 0.23 points, which considerably exceeds the variation attributable to random sampling and points to genuine structural differences among the issuer corpora.

Table 4. Per-issuer accuracy and macro F1 for the best model (SVM + SMOTE).

Ticker	N (test)	Accuracy	Macro F1	Interpretation
ASII	90	0.778	0.758	Balanced class distribution
BBCA	91	0.802	0.752	Banking; class profile similar to BMRI
BBNI	70	0.929	0.881	Highest; consistent banking vocabulary
BBRI	113	0.805	0.663	Most discussed; many multi-issuer tweets causing ambiguity
BMRI	80	0.863	0.773	Banking; strong positive dominance (73.5%)
BYAN	74	0.770	0.719	Coal mining; high proportion of technical content
HMSP	85	0.718	0.697	Tobacco; narrow domain vocabulary (excise, cigarette)
TLKM	56	0.857	0.844	Telecommunications; structured discussion patterns

TPIA	61	0.754	0.647	Lowest; petrochemical domain, sparse specialized vocabulary
UNVR	83	0.819	0.778	Negative-dominated (54.5%); boycott-driven vocabulary

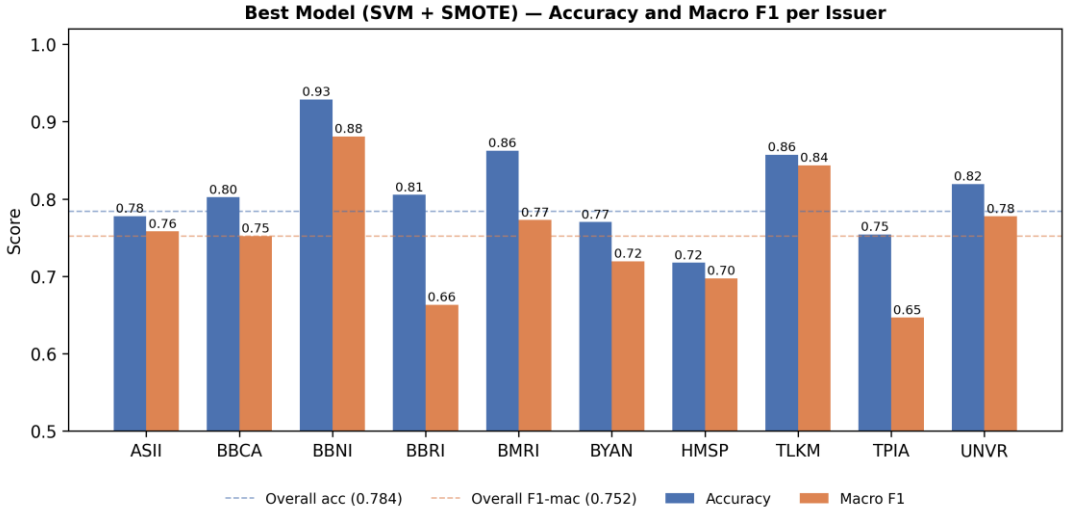

Figure 6. Per-issuer accuracy and macro F1 for the best model (SVM + SMOTE).

Three structural factors explain the bulk of this variation. Vocabulary consistency is the principal driver of high performance: the banking issuers (BBNI, TLKM, BMRI, BBCA) attract a comparatively stable lexicon centred on dividends, earnings upgrades and downgrades, and quarterly results, providing TF-IDF distributions that the linear classifier can separate reliably. Distinctive class vocabulary matters independently of consistency: UNVR, despite hosting a Negative majority, achieves macro F1 of 0.778 because its negative discourse is anchored by highly distinctive terms (boikot, turun, produk) that rarely appear in Positive or Neutral tweets, making the boundaries between classes cleaner than simple imbalance ratios would suggest. Domain-vocabulary sparsity degrades performance for TPIA and HMSP: petrochemical discussions involve domain-specific terms such as "nafta" (naphtha) and "crackers" that appear infrequently in the training vocabulary, while tobacco discussions around "cukai" are similarly narrow. Both effects reduce effective feature coverage and force the model toward noisier statistical proxies.

BBRI is a noteworthy outlier. Although it contributes the largest test sub-corpus (n = 113), its macro F1 of 0.663 is the second lowest. Qualitative review of misclassified instances reveals a recurring pattern: many BBRI tweets reference multiple issuers within a single post; for example, "BBRI, BBCA, BMRI naik bareng" (BBRI, BBCA, BMRI rising together), so that the single gold label attached to the tweet may accurately describe only one of the named issuers. This multi-issuer ambiguity is an inherent limitation of the current annotation scheme, not a weakness of the classifier per se, and it suggests that future dataset extensions might benefit from sentence-level or mention-level annotation that can resolve polarity at the issuer mention rather than the document level.

3.6 Research Limitations

This study provides a systematic and reproducible evaluation of class-imbalance handling strategies on the ID-SMSA dataset; nevertheless, acknowledging the following limitations will help contextualize the findings and guide more comprehensive investigations in future work.

- All experiments use a single 80:20 stratified train-test split with a fixed random seed, which ensures full reproducibility and allows direct comparison across configurations.

Future work employing k-fold cross-validation would further strengthen the reliability of these findings across different data partitions.

- The study intentionally focuses on TF-IDF-based classical machine-learning models to establish a transparent and reproducible baseline. This provides a well-defined reference point against which future studies using contextual representations such as IndoBERT, which captures subword-level and contextual information, can directly measure their performance gains.
- The ID-SMSA dataset exhibits moderate class imbalance at IR = 2.41, which reflects a realistic distribution found in many real-world social-media corpora. Studies on datasets with more severe imbalance ratios would offer an opportunity to further examine the conditions under which SMOTE and class weighting provide larger performance gains.
- Multi-issuer tweets, where a single post mentions several issuers simultaneously, present an interesting annotation challenge that affects per-issuer model performance, as reflected in the BBRI results. Addressing this through sentence-level or mention-level annotation in future dataset extensions would provide a richer and more precise evaluation framework for any classifier applied to this corpus.

4. Conclusions

This paper evaluates two class-imbalance mitigation strategies, SMOTE and class weighting, against a no-handling baseline on the complete ID-SMSA dataset, using three TF-IDF-based classical classifiers. SMOTE consistently improves macro F1 across all compatible classifiers, with the largest absolute gain on Naive Bayes (+0.137, from 0.589 to 0.726). The best overall configuration is linear SVM combined with SMOTE, achieving macro F1 of 0.752 and accuracy of 0.784. Class weighting provides smaller and less consistent improvements, modestly aiding Random Forest (+0.011) while marginally reducing SVM, confirming that linear SVM on TF-IDF is already robust to moderate imbalance at IR = 2.41.

Per-issuer evaluation reveals that macro F1 varies substantially across the ten IDX issuers, from 0.881 on BBNI to 0.647 on TPIA. Three structural factors explain this spread: the lexical consistency of issuer-specific discourse, the distinctiveness of minority-class vocabulary, and domain-vocabulary sparsity for narrow-sector issuers such as petrochemicals and tobacco. The relatively poor result on BBRI, the most frequently discussed issuer, is driven by multi-issuer tweet ambiguity rather than model deficiency, a structural annotation characteristic that future work could address. Linguistically, the negative-class language reflects concrete socio-economic events: the boycott movement in UNVR tweets and excise-policy debates in HMSP tweets are clearly legible in the top-frequency negative terms, confirming the domain-sensitivity of ID-SMSA.

These results provide a transparent, reproducible classical baseline on ID-SMSA. Prior studies on this dataset have focused exclusively on transformer architectures; the present benchmark complements that work by establishing reference points that future deep-learning and transformer studies can measure against, and by empirically quantifying the impact of three imbalance-handling strategies across three structurally different classifiers.

References

Ariyani, P. W., Sunarya, I. M. G., & Gunadi, I. G. A. (2025). ANALISIS SENTIMEN MASYARAKAT TERHADAP VIRUS CORONA BERDASARKAN OPINI DARI TWITTER MENGGUNAKAN METODE NAÏVE BAYES DAN K-NEAREST NEIGHBOR. *Jurnal Pendidikan Teknologi Dan Kejuruan*, *22*(2), 128–138. https://doi.org/10.23887/jptk-undiksha.v22i2.103233

Aziz, F., Jeffry, J., Ayu Asrhi, N., & La Wungo, S. (2025). Classification of Chocolate Consumption Using Support Vector Machine Algorithm. *Journal of System and Computer Engineering (JSCE)*, *6*(2), 180–187. https://doi.org/10.61628/jsce.v6i2.1860

Dewi, N. L. P. R., Wijaya, I. N. S. W., Purnamawan, I. K., & Marti, N. W. (2024). Model Classifer Judul Berita Pariwisata Indonesia Berdasarkan Sentimen. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *11*(1), 117–124. https://doi.org/10.25126/jtiik.20241117617

Gao, X., Xie, D., Zhang, Y., Wang, Z., Chen, C., He, C., Yin, H., & Zhang, W. (2026). A comprehensive survey on imbalanced data learning. *Frontiers of Computer Science*, *20*(11), 2011622. https://doi.org/10.1007/s11704-025-50274-7

Hartanto, J., Liundi, T., Sutoyo, R., & Andangsari, E. W. (2025a). ID-SMSA: Indonesian stock market dataset for sentiment analysis. *Data in Brief*, *60*, 111571. https://doi.org/10.1016/j.dib.2025.111571

Hartanto, J., Liundi, T., Sutoyo, R., & Andangsari, E. W. (2025b). ID-SMSA: Indonesian Stock Market Dataset for Sentiment Analysis (Version 3) [Data set]. *Mendeley Data*. https://doi.org/10.17632/tn4vzs8tdw.3

Hartanto, J., Sutoyo, R., Liundi, T., & Andangsari, E. W. (2025). A Deep Learning Approach to Sentiment Analysis of Indonesian Stock Market Using IndoBERT. *2025 5th International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS)*, 30–35. https://doi.org/10.1109/ICE3IS66769.2025.11281119

Hidayah, A. K., Sunardi, D., & Sahputra, E. (2025). *Analisis Sentimen di Era Digital: Konsep, Algoritma, dan Studi Kasus Media Sosial* (I). Litnus.

Hinojosa Lee, M. C., Braet, J., & Springael, J. (2024). Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores. *Applied Sciences*, *14*(21), 9863. https://doi.org/10.3390/app14219863

Indrawan, G., Setiawan, H., & Gunadi, A. (2022). Multi-class SVM Classification Comparison for Health Service Satisfaction Survey Data in Bahasa. *HighTech and Innovation Journal*, *3*(4), 425–442. https://doi.org/10.28991/HIJ-2022-03-04-05

Istiqamah, N., & Rijal, M. (2024). Klasifikasi Ulasan Konsumen Menggunakan Random Forest dan SMOTE. *Journal of System and Computer Engineering (JSCE)*, *5*(1), 66–77. https://doi.org/10.61628/jsce.v5i1.1061

Maharani, M. D., Indradewi, I. G. A. A. D., & Wijaya, I. N. S. W. (2026). PERBANDINGAN PERFORMA TF-IDF DAN BOW PADA ANALISIS SENTIMEN BPJS KESEHATAN MENGGUNAKAN XGBOOST. *Jurnal Pendidikan Teknologi Dan Kejuruan*, *23*(1), 13–24. https://doi.org/10.23887/jptk-undiksha.v23i1.109604

Munthe, I. R., Sumijan, & Tajuddin, M. (2025). Enhancing Machine Learning Algorithms for Analyzing Financial Condition News of Companies Listed on the Indonesia Stock Exchange. *2025 3rd International Conference on Computer System, Information Technology, and Electrical Engineering (COSITE)*, 328–333. https://doi.org/10.1109/COSITE68330.2025.11414342

Nassr, Z., Benabbou, F., Sael, N., & Hamim, T. (2025). Improving Sentiment Analysis Performance on Imbalanced Moroccan Dialect Datasets Using Resample and Feature Extraction Techniques. *Information*, *16*(1), 39. https://doi.org/10.3390/info16010039

Raharjo, R. A., Sunarya, I. M. G., & Divayana, D. G. H. (2022). Perbandingan Metode Naïve Bayes Classifier Dan Support Vector Machine Pada Kasus Analisis Sentimen Terhadap Data Vaksin Covid-19 Di Twitter. *Elkom : Jurnal Elektronika Dan Komputer*, *15*(2), 456–464. https://doi.org/10.51903/elkom.v15i2.918

Sanjaya, I. G. A. S., Candiasa, I. M., & Dewi, L. J. E. (2024). Analisis Sentimen Berbasis Aspek Kinerja Polri Menggunakan SVM dengan Pendekatan POS Tagging. *SINTECH (Science and Information Technology) Journal*, *7*(2), 112–124. https://doi.org/10.31598/sintechjournal.v7i2.1568

Saputri, N. K. T. A., Gunadi, I. G. A., & Sunarya, I. M. G. (2024). Analisis Sentimen Pelayanan Daring di Fakultas Teknik dan Kejuruan Universitas Pendidikan Ganesha Menggunakan Algoritma Naïve Bayes dan LSTM. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(3), 1120–1129. https://doi.org/10.57152/malcom.v4i3.1336

Sari, A. M., Sunarya, I. M. G., & Maysanjaya, I. M. D. (2026). Perancangan Aplikasi Analisis Sentimen Mengenai Program Petani Milenial Berbasis IndoBERT. *Kumpulan Artikel Mahasiswa Pendidikan Teknik Informatika (KARMAPATI)*, *15*(1). https://doi.org/10.23887/karmapati.v15i1.108665

Setiawan, A., & Suryono, R. R. (2024). Analisis Sentimen Ibu Kota Nusantara menggunakan Algoritma Support Vector Machine dan Naïve Bayes. *Edumatic: Jurnal Pendidikan Informatika*, *8*(1), 183–192. https://doi.org/10.29408/edumatic.v8i1.25667

Setiawan, M. H., Gunadi, I. G. A., & Indrawan, G. (2023). Klasifikasi Pelayanan Kesehatan Berdasarkan Data Sentimen Pelayanan Kesehatan menggunakan Multiclass Support Vector Machine. *Jurnal Sistem Dan Informatika (JSI)*, *17*(1), 47–54. https://doi.org/10.30864/jsi.v17i1.512

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

Sillviari, N. P. D., Candiasa, I Made, & Indrawan, G. (2025). Classification of Anxiety Levels in Vocational Students Through Life Story Analysis Using Multi-class SVM. *Tekno - Pedagogi : Jurnal Teknologi Pendidikan*, *15*(2), 1–21. https://doi.org/https://doi.org/10.22437/teknopedagogi.v15i2.46957

Simanihuruk, R., Yulianti, E., Azizah, K., & Jatmiko, W. (2025). Sentiment Analysis on Indonesian Stock Market Texts: A Comparative Study of Support Vector Machine (SVM) and IndoBERT. *2025 IEEE International Conference on Data and Software Engineering (ICoDSE)*, 455–460. https://doi.org/10.1109/ICoDSE68111.2025.11351619

Wijaya, W., Seputra, K. A., & Dewi, N. P. N. P. (2025). FINE TUNNING MODEL INDOBERT UNTUK ANALISIS SENTIMEN BERITA PARIWISATA INDONESIA. *Jurnal Pendidikan Teknologi Dan Kejuruan*, *22*(2), 195–204. https://doi.org/10.23887/jptk-undiksha.v22i2.104056