Insearch
Information System Research Journal

E-ISSN: 2775-4669
Volume 6, Nomor 2, Agustus 2026

The Impact of Signal Deletion in Text Preprocessing on Timestamp Comment Detection under Severe Class Imbalance

Decha Danillo Novicahyanto[1], Albertus Dwivoga Widiantoro[2]
Information Systems Study Program, Faculty of Computer Science, Soegijapranata Catholic University
Jl. Pawiyatan Luhur IV/1, Bendan Duwur, Semarang 50234, Central Java, Indonesia
23n10007@student.unika.ac.id[1], yoga@unika.ac.id[2]

Received: 23 Juli 2026 | Revised: 30 Agustus 2026
Accepted: 31 Agustus 2026 | Published: 31 Agustus 2026

Abstrak

Praktik timestamping pada kolom komentar YouTube merupakan bentuk kurasi konten partisipatif yang khas. Penelitian ini menguji mengapa empat algoritma klasifikasi teks, yaitu Logistic Regression, Support Vector Machine, Random Forest, dan XGBoost, gagal sepenuhnya mendeteksi komentar bertimestamp pada korpus berbahasa Indonesia yang informal. Dari 14.149 komentar video Ruangguru Clash of Champions, rasio ketidakseimbangan mencapai 73,86:1. Seluruh model menghasilkan akurasi sekitar 98,7% dengan recall kelas minoritas nol, Balanced Accuracy 0,50, dan Cohen's Kappa nol atau negatif, yakni gejala klasik Accuracy Paradox. Kontribusi utama penelitian ini bersifat diagnostik: ketidakseimbangan kelas bukan satu-satunya penyebab. Penelusuran token pada alur praproses memperlihatkan bahwa penghapusan tanda baca yang diikuti penghapusan angka berdiri sendiri secara deterministik memusnahkan pola token yang mendefinisikan label positif, sementara pemangkasan term jarang pada ambang 0,99 membuang kosakata yang muncul pada kurang dari 1% dokumen, yaitu ambang yang berada di atas prevalensi kelas minoritas sebesar 1,34%. Pemulihan confusion matrix dari hasil yang dilaporkan memungkinkan penghitungan metrik tahan-ketidakseimbangan secara lengkap, yang menegaskan bahwa keempat model tidak dapat dibedakan dari pengklasifikasi konstan. Penelitian ini merancang ablasi praproses dan perbandingan strategi penanganan ketidakseimbangan sebagai agenda lanjutan, serta menegaskan bahwa sebelum kegagalan kelas minoritas diatribusikan pada ketidakseimbangan data, peneliti wajib memastikan praproses tidak menghapus sinyal yang hendak dipelajari.

Kata kunci: *klasifikasi teks, ketidakseimbangan kelas, paradoks akurasi*

Abstract

Timestamping in YouTube comment sections is a distinctive form of participatory content curation. This study examines why four text classification algorithms, namely Logistic Regression, Support Vector Machine, Random Forest, and XGBoost, fail completely at detecting timestamp-bearing comments in an informal Indonesian-language corpus. Across 14,149 comments from Ruangguru Clash of Champions videos, the imbalance ratio reaches 73.86:1. Every model returns an accuracy near 98.7% with a minority-class recall of zero, a Balanced Accuracy of 0.50, and Cohen's Kappa at or below zero, the textbook signature of the Accuracy Paradox. The contribution is diagnostic: class imbalance is not the sole cause. A token-level trace shows that punctuation removal followed by standalone-numeral removal deterministically destroys the pattern defining the positive label, while sparse-term pruning at a 0.99 threshold discards vocabulary occurring in fewer than 1% of documents, above the minority prevalence of 1.34%. Recovering the confusion matrices permits a complete imbalance-robust metric set to be computed, confirming that all four models are indistinguishable from a constant classifier that ignores its input. The study specifies the ablation required to separate these causes, and establishes that before minority-class failure is attributed to imbalance, researchers must verify that preprocessing has not deleted the signal being learned.

Keywords: *text classification, class imbalance, accuracy paradox*

I. INTRODUCTION

Video-sharing platforms are no longer mere content distribution channels. They have evolved into participatory ecosystems in which a single user may act as producer, curator, and consumer of information simultaneously [1]. YouTube exemplifies what Jenkins [2] termed a convergence culture, a setting in which the boundary between media production and consumption continues to erode. This shift is particularly visible in edutainment, where Ruangguru uses video content to bring academic competitions to a broad audience. Within this ecosystem, one digital social practice stands out: the culture of timestamping, in which viewers voluntarily annotate specific temporal markers in comment sections, typically formatted as MM:SS or embedded in URL parameters such as t=754. From a text-mining perspective, timestamp-bearing comments constitute a pragmatically distinct category, since they serve an instrumental purpose as navigational aids and content-indexing markers rather than as expressive reactions to the video as a whole.

Analyzing Indonesian-language social media content introduces well-known complications. Indonesian social media discourse is informal and structurally loose, marked by slang, acronyms, emoticons, and code-mixing among Indonesian, regional languages, and foreign vocabulary [3]. Timestamping also occurs far less frequently than ordinary commenting. Automated labeling via regular expression matching confirms this skew: Class 0 (non-timestamp comments) accounts for 13,960 instances, or 98.66%, while Class 1 (timestamp comments) contains 189 instances, or 1.34%, yielding an imbalance ratio of 73.86:1. Following the taxonomy of Haixiang et al. [4], who reserve the label extreme for ratios exceeding 100:1, a ratio of 73.86:1 is more accurately described as severe or high imbalance. That terminology is used consistently below, and the reader should note that the failure documented here occurs below the conventional extreme threshold, which makes it more, not less, consequential for applied work.

The Accuracy Paradox, in which a model achieves high accuracy on skewed data while failing entirely at its actual task, is well established [5], [6]. A study that merely reproduces it adds little. The contribution of the present work lies elsewhere. When we re-audited our own pipeline in response to an anomalous result, we found that the failure of all four classifiers had a second and more elementary cause, one that class-imbalance remedies alone would not have fixed. The preprocessing sequence removes punctuation before removing standalone numerals. A comment containing 12:34 therefore becomes 1234 at step five, and 1234 is then deleted in its entirety by the standalone-numeral pattern at step six. The token that defines the positive label is removed by construction, in every document, before feature extraction begins. Sparse-term pruning compounds this: removeSparseTerms() at a 0.99 threshold retains only terms appearing in at least 1% of documents, while the entire minority class constitutes 1.34% of the corpus.

This distinction matters because it changes what the experiment can be said to demonstrate. A classifier that fails on a feature space stripped of the discriminative signal has not demonstrated the limits of learning under imbalance; it has demonstrated the consequences of an unaudited preprocessing pipeline. Accordingly, this study poses four questions. RQ1: how do conventional text classification algorithms perform at identifying timestamp-bearing comments under severe class imbalance? RQ2: to what extent is the observed failure attributable to class imbalance, and to what extent to preprocessing-induced deletion of the label-defining signal? RQ3: do standard imbalance-handling strategies recover minority-class performance, and does their effect depend on whether the preprocessing signal has been preserved? RQ4: which evaluation metrics remain informative at this degree of skew? The paper makes three contributions: it documents a failure mode we term preprocessing-induced signal deletion together with evidence that its symptoms are indistinguishable from imbalance-driven failure; it provides a factorial evaluation crossing two preprocessing variants with four imbalance-handling strategies; and it contributes a corrected evaluation protocol for severely imbalanced Indonesian text classification.

II. LITERATURE REVIEW

A. Learning from Imbalanced Data

Class imbalance is a long-standing problem in machine learning. Japkowicz and Stephen [7] showed that minority-class degradation scales with both the imbalance ratio and concept complexity, while He and Garcia [8] organised solutions into three levels: data-level resampling, algorithm-level cost-sensitive learning, and ensemble approaches. SMOTE [9] remains the reference method at the data level, with ADASYN [10] as its adaptive extension, while Ting [11] addresses cost weighting at the

algorithm level. Haixiang et al. [4] provide the severity taxonomy adopted in this study.

B. The Accuracy Paradox and Metric Selection

Accuracy becomes uninformative when the class distribution is severely skewed, because True Negatives dominate both numerator and denominator. Hand [6] and Fernandez et al. [5] argue that under such conditions accuracy is not merely uninformative but actively misleading. Luque et al. [12] analyse systematically how imbalance biases metrics derived from the binary confusion matrix, while Chicco and Jurman [13] argue for the Matthews Correlation Coefficient over F1 and accuracy. Davis and Goadrich [14] show that Precision-Recall curves are more discriminative than ROC curves when the positive class is rare, which motivates the choice of PR-AUC as the primary metric here. Fawcett [15] provides the interpretive basis for ROC analysis.

C. Processing Informal Indonesian Text

Indonesian user-generated text departs substantially from the standard register, which has prompted dedicated lexical resources for Indonesian microblogs [16] and normalisation pipelines for code-mixed Indonesian-English social media data [3]. Pre-trained models such as IndoBERT [17] and the IndoNLU benchmark [18] demonstrate that transformer architectures capture contextual dependencies beyond the reach of bag-of-words representations. TF-IDF, although efficient and interpretable, discards word order and contextual dependency [19], a limitation of particular importance for a phenomenon defined by a positional numeric pattern.

D. Research Position

To the authors' knowledge, the existing literature treats minority-class failure as a consequence of imbalance but has not examined the possibility that a routine cleaning step may delete the label-defining feature and thereby produce an identical metric signature. This study addresses that gap.

III. RESEARCH METHOD

A. Data Source

The dataset comprises 14,149 YouTube comment records from videos of the Ruangguru Clash of Champions academic tournament, stored in CSV format. Each record carries nine attributes: authorChannelId, authorDisplayName, id, isPublic, likeCount, publishedAt, textDisplay, totalReplyCount, and videoId. The field textDisplay, holding the textual content of each comment, is the sole analytical input. The corpus was assembled as part of a broader study by the first author [20].

B. Automatic Labeling and Label Validation

Target-class labeling was performed automatically through regular expression matching on the raw, unprocessed comment text. Two patterns were applied: t=\d+ for YouTube URL temporal parameters, and \b\d{1,2}:\d{2}\b for standard temporal expressions such as 12:34. A comment matching either pattern was labeled Class 1, producing 13,960 instances (98.66%) in Class 0 and 189 instances (1.34%) in Class 1.

Because every downstream result depends on the correctness of these labels, two limitations of the patterns are acknowledged explicitly rather than left implicit. The second pattern admits invalid times such as 5:67, since it does not constrain the minute field, and it also matches non-temporal uses of the colon-digit form, for example score reports or Qur'anic verse citations such as 2:255, which are plausible in an Indonesian educational-content comment section. Conversely, the patterns miss timestamps written in other conventions common in Indonesian comments, such as menit 12, 12.34, or 1 jam 5 menit. No manual annotation was carried out against these regex labels in the present study, so the resulting label noise is unquantified. This bounds every result reported below and is treated as a limitation in Section IV-F rather than as a solved problem; independent human annotation of a stratified sample is the first item of the follow-up agenda in Section IV-C.

C. Preprocessing and the Destruction of the Label-Defining Token

The original pipeline proceeded through seven steps: case folding; HTML tag removal; URL elimination; non-ASCII character removal; punctuation elimination; standalone numeral removal via \b\d+\b; and whitespace normalization. Records reduced to empty text were then discarded. The interaction between steps three, five, and six is destructive for this task, and this is the pivot of the present study. Table I traces a representative Class 1 comment through the pipeline.

TABLE I. TOKEN-LEVEL TRACE OF A CLASS 1 COMMENT

Preprocessing step	Resulting text
Raw input	Bagian terbaiknya di 12:34 wajib nonton
(1) Case folding	bagian terbaiknya di 12:34 wajib nonton
(4) Non-ASCII removal	bagian terbaiknya di 12:34 wajib nonton
(5) Punctuation removal	bagian terbaiknya di 1234 wajib nonton
(6) Standalone numeral removal	bagian terbaiknya di wajib nonton
(7) Whitespace normalization	bagian terbaiknya di wajib nonton

After step six, the surviving text is lexically indistinguishable from an ordinary comment. The same holds for the URL-parameter form: t=754 is removed outright by step three when it appears inside a hyperlink, and otherwise reduces to the singleton token t754, whose document frequency falls far below the retention threshold applied later. This trace is deductive rather than experimental: given the stated order of operations, the deletion follows necessarily for every document containing a timestamp, and no run of the pipeline is required to establish it. What a trace cannot establish is how much of the observed failure this deletion accounts for relative to class imbalance, which requires the ablation specified in Section IV-C.

D. Feature Extraction

A corpus was constructed with the tm library in R, followed by a Document-Term Matrix weighted by TF-IDF. Dimensionality was reduced with removeSparseTerms() at a sparsity threshold of 0.99, retaining terms appearing in at least 1% of corpus documents. This threshold is analytically problematic for a minority class of 1.34% prevalence: a term appearing in every Class 1 document but no others would sit at 1.34% document frequency, barely clearing the cutoff, while a term appearing in three-quarters of Class 1 documents would fall below it and be discarded. Sparse-term pruning calibrated on the majority class is therefore a second, independent mechanism of minority-signal loss, and relaxing this threshold is one of the conditions specified for the ablation in Section IV-C.

Two reporting issues in the original analysis are corrected here. First, TF-IDF weights were computed over the complete corpus before the train-test split, allowing document frequencies from the test set to influence the training representation; although the practical effect is small, this is a form of information leakage, and vocabulary and IDF should be fitted on the training partition only. Second, the original text reports a modeling matrix of 13,971 rows while also reporting a split of 11,320 and 2,829, which sums to 14,149. These are mutually inconsistent. The reported accuracies resolve the conflict: a test set of 2,829 with an all-majority accuracy of 0.9869 implies exactly 2,792 majority and 37 minority instances, and no such clean decomposition exists for a test set of 2,794. The split was therefore performed on the full 14,149 records, and the figure of 13,971 does not describe the matrix that was actually modeled.

E. Data Splitting and Validation

Stratified partitioning used createDataPartition() from the caret library with an 80:20 split and a fixed random seed of 2026, yielding 11,320 training and 2,829 testing records. Evaluation rests on this single hold-out partition; no cross-validation, repeated sampling, or significance testing was performed. This is a material limitation rather than a design choice. The test partition contains only 37 minority instances, so every minority-class metric carries a wide confidence interval, and the differences among models reported in Table III, on the order of 0.0007 in accuracy, fall far below the resolution this design can support. No claim of superiority among the four classifiers is made anywhere in this paper, and none would be defensible from these data.

F. Models and Hyperparameter Selection

Four classifiers were evaluated. Logistic Regression used glm() with family = binomial and predicted probabilities thresholded at 0.5. The SVM used svm() from the e1071 library with a linear kernel, producing class labels directly. Random Forest used randomForest(), and XGBoost used xgb.train() with objective = 'binary:logistic', again thresholded at 0.5. The hyperparameter values are given in Table II. These were selected manually and held fixed; no grid search, random search, or nested tuning was performed, and the values were not optimized against any validation criterion. A fair comparison among classifiers would require tuning each model under an identical protocol, optimizing a metric appropriate to the skew such as PR-AUC rather than accuracy, since an accuracy-optimizing search at a ratio of 73.86:1 returns the trivial majority-class model for every configuration in any grid.

TABLE II. HYPERPARAMETER VALUES USED (MANUALLY SELECTED, NOT TUNED)

Model	Function	Settings
Logistic Reg.	glm()	family = binomial; threshold = 0.5
SVM	svm() [e1071]	linear kernel; direct class output
Random Forest	randomForest()	ntree = 50; direct class output
XGBoost	xgb.train()	objective = binary:logistic; max_depth = 6; eta = 0.3; nrounds = 30; threshold = 0.5

G. Absence of Imbalance Handling

No imbalance-handling mechanism was applied at any point. Training used the original distribution of 98.66% against 1.34%, with no class weighting, no resampling, and no adjustment of the 0.5 decision threshold. Every result in Section IV therefore describes the behaviour of unmodified classifiers on unmodified data. This is stated plainly because it determines what the results can and cannot show: they document how conventional pipelines fail, not whether that failure is remediable. The strategies that would test remediability are specified in Section IV-C.

H. Evaluation Metrics and Positive-Class Declaration

The positive class in this study is Class 1, the minority class of timestamp-bearing comments. This declaration is stated explicitly because its absence was the source of a substantive misinterpretation in the original analysis. Reported metrics are Accuracy; Precision, Recall, and F1 for Class 1; Specificity; Macro-F1; Balanced Accuracy; Cohen's Kappa [21]; MCC; ROC-AUC; and PR-AUC. Given a positive-class prevalence of 1.34%, PR-AUC is treated as the primary metric, since ROC-AUC remains optimistic under severe skew [14], and Accuracy is reported only to document the paradox, never as evidence of performance.

IV. RESULTS AND DISCUSSION

A. Correcting the Confusion-Matrix Interpretation

Table III reproduces the original results exactly as reported. These figures are internally consistent, but only under a convention the original manuscript did not state: caret::confusionMatrix() assigns the positive class to the first factor level, which for a factor(c("0","1")) label is Class 0, the majority class.

TABLE III. ORIGINAL REPORTED RESULTS (POSITIVE CLASS = CLASS 0, MAJORITY)

Model	Acc.	Kappa	Sens.	Spec.
Logistic Regression	0.9869	0.0000	1.0000	0.0000
SVM	0.9869	0.0000	1.0000	0.0000
Random Forest	0.9862	−0.0013	0.9993	0.0000
XGBoost	0.9866	−0.0007	0.9996	0.0000

Read under that convention the table is coherent: Sensitivity of 1.0000 means every majority-class instance was captured, and Specificity of 0.0000 means no minority-class instance was. The original manuscript, however, described Specificity as the proportion of actual positive-class instances correctly identified and attributed its zero value to True Positives never rising above zero, statements that hold only if the positive class is Class 1, contradicting the convention that produced the numbers. The fault is one of undeclared convention rather than miscalculation.

Because the reported values fully determine the underlying counts, the confusion matrices can be recovered exactly. With a test set of 2,829 instances, an all-majority accuracy of 0.9869 implies 2,792 majority and 37 minority instances, not the approximately 38 stated in the original text. Table IV presents the recovered matrices under the correct convention.

TABLE IV. RECOVERED CONFUSION MATRICES (POSITIVE CLASS = CLASS 1; n = 2,829)

Model	TP	FP	FN	TN
Logistic Regression	0	0	37	2,792
SVM	0	0	37	2,792
Random Forest	0	2	37	2,790
XGBoost	0	1	37	2,791

Recomputing Cohen's Kappa from these matrices returns −0.0013 for Random Forest and −0.0007 for XGBoost, reproducing the reported values to four decimal places and confirming the reconstruction. The slight negativity is not a substantive finding: two Random Forest predictions and one XGBoost prediction fell on the positive class and both were wrong, pushing observed agreement marginally below chance expectation. With 37 positive instances, this is sampling noise.

B. The Complete Metric Set Derived from the Confusion Matrices

The confusion matrices in Table IV determine every imbalance-robust metric exactly, so these follow by arithmetic rather than by a further experiment. For Logistic Regression and SVM,

recall on Class 1 is 0.0000, F1 is 0.0000, specificity is 1.0000, Macro-F1 is 0.4967, Balanced Accuracy is 0.5000, and Cohen's Kappa is 0.0000. Precision and MCC are undefined for both, because neither model issued a single positive prediction and the denominators are therefore zero; reporting them as 0.0000 would present a structural absence as a measurement. Random Forest and XGBoost differ only marginally: recall and F1 remain 0.0000, specificity is 0.9993 and 0.9996, Macro-F1 is 0.4965 and 0.4966, Balanced Accuracy is 0.4996 and 0.4998, MCC is −0.0031 and −0.0022, and Kappa is −0.0013 and −0.0007.

These figures acquire their meaning against a single benchmark. A constant classifier that ignores its input entirely and returns Class 0 for every comment achieves, on this partition, an accuracy of 0.9869, a Balanced Accuracy of 0.5000, a Macro-F1 of 0.4967, and a Kappa of 0.0000. Logistic Regression and SVM reproduce those values exactly, to four decimal places, on every metric. Random Forest and XGBoost fall marginally below them, because the two and one positive predictions they respectively emitted were all incorrect, which is why their Kappa and MCC are slightly negative rather than zero. Four learning algorithms, given 11,320 training examples and 125 features, extracted nothing that a model ignoring the input did not already achieve. ROC-AUC and PR-AUC are not reported, since both require predicted probability scores rather than the hard class predictions preserved in the original analysis; for reference, a random classifier would achieve a PR-AUC of 0.0134 here. The gap between an accuracy of 98.7% and a Balanced Accuracy of 0.5000, describing the same models, is the Accuracy Paradox in its clearest form, and the second number is the honest one.

C. The Ablation This Study Does Not Report, and Why It Is Required

Sections III-C and III-D establish two mechanisms by which the label-defining signal is removed before any model sees it. What they do not establish is the share of the observed failure attributable to each mechanism as against class imbalance itself, because all three are present simultaneously in the single pipeline that was run, and their metric signatures are identical: zero minority recall, near-perfect accuracy, and Kappa at zero. No metric computed on one pipeline can separate them. Reporting the diagnosis without this separation would overstate what the present data support, so the ablation is specified here as the required next step rather than presented as a result.

The design is a two-factor comparison. The first factor is preprocessing, contrasting the original pipeline against a variant in which timestamp patterns are replaced by a reserved sentinel token before punctuation removal, with the sparse-term threshold relaxed from 0.99 to 0.995 as a nested condition. The second factor is imbalance handling, contrasting the unmodified training set against class weighting, SMOTE applied strictly within training folds, and threshold optimization on minority-class F1. Evaluation would use repeated stratified five-fold cross-validation reporting PR-AUC and minority-class F1, with McNemar tests under Holm correction for pairwise comparisons and a Friedman test across the four classifiers. The prediction is falsifiable: if class imbalance is the sole cause, the original-pipeline rows improve substantially under every imbalance strategy; if signal deletion dominates, those rows stay near zero under all of them, because no reweighting scheme can recover a feature absent from the design matrix.

One outcome of that design must be interpreted with care rather than counted as success. A sentinel token derived from the same regular expression that produced the labels is near-perfectly correlated with them by construction, so a classifier scoring highly under the preservation condition has learned the labeling rule, not the language. This circularity is inherent to regex-derived labels. Its value is diagnostic only, in establishing whether the original feature space contained usable signal at all; genuine linguistic generalization would have to be demonstrated against independently annotated labels, which returns to the annotation gap noted in Section III-B.

D. Why Conventional Loss Functions Converge on the Trivial Solution

Under a class prior of 98.66% against 1.34%, any learner minimizing aggregate error, whether cross-entropy for Logistic Regression and XGBoost, hinge loss for SVM, or Gini impurity for Random Forest, finds that predicting the majority class unconditionally yields 98.66% correct classifications [7], [8]. This is a global optimum of the stated objective and a complete failure of the intended task. That four structurally distinct paradigms converge on it is not a coincidence but the expected consequence of optimizing an unweighted objective under skew.

What this study adds is the observation that the well-understood mechanism above was, in this case,

not the binding constraint. When the discriminative feature has already been deleted, the trivial solution is not merely the easiest available strategy but the only one the feature space permits. Distinguishing these two situations requires an ablation, since no metric computed on a single pipeline can separate them: their signatures are identical, namely zero recall, near-perfect accuracy, and Kappa at zero. Studies reporting minority-class collapse are therefore encouraged to include a preprocessing trace of the kind shown in Table I before attributing the collapse to imbalance.

E. Limitations

Five limitations bound these conclusions. First, the ground truth derives from regular expressions with no independent human annotation, so the label noise discussed in Section III-B is unquantified and every metric below inherits it. Second, no imbalance-handling strategy was evaluated, so this study documents that conventional pipelines fail but not whether that failure is remediable. Third, evaluation rests on a single hold-out partition containing 37 minority instances, without cross-validation or significance testing, which is why no claim of superiority among the four classifiers is made. Fourth, hyperparameters were fixed manually rather than tuned under a common protocol. Fifth, the corpus is drawn from a single content series in one genre, and comment conventions may differ across Indonesian YouTube more broadly. The first four are addressed directly by the design specified in Section IV-C.

V. CONCLUSION

This study compared four conventional text classification algorithms on the task of identifying timestamp-bearing comments in an informal Indonesian-language YouTube corpus with a severe class imbalance of 73.86:1. All four returned an accuracy near 98.7% together with a minority-class recall of exactly zero, a Balanced Accuracy of approximately 0.50, and Cohen’s Kappa at or below zero. Read through accuracy alone, the models appear excellent; read through any imbalance-robust metric, they are indistinguishable from a constant classifier.

The more consequential finding is diagnostic. The failure is not attributable to class imbalance alone. A token-level trace established that punctuation removal followed by standalone-numeral removal deterministically destroys the temporal token pattern defining the positive label, and that sparse-term pruning at a 0.99 threshold sits above the minority class prevalence of 1.34%. The classifiers were asked to learn a distinction whose evidence had already been erased. Because all three mechanisms operated at once in the single pipeline that was run, their relative contributions remain unseparated, and the ablation specified in Section IV-C is the necessary next step rather than an optional extension.

Three recommendations follow. For practitioners, preprocessing pipelines should be audited against the definition of the target label before any modeling result is interpreted; a trace of a single positive instance through each step, as in Table I, is sufficient to catch this class of error and takes minutes. For researchers reporting minority-class collapse, an ablation separating signal availability from class imbalance should accompany the claim, since the two produce identical metric signatures and only one of them is remediable by resampling. For evaluation practice under severe skew, accuracy should be reported only to document the paradox, with PR-AUC, Macro-F1, MCC, and Balanced Accuracy carrying the analytic weight, and the positive class declared explicitly wherever a confusion matrix appears.

REFERENCES

[1] M. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," ACM Computing Surveys, vol. 52, no. 4, pp. 1–36, 2019.

[2] H. Jenkins, Convergence Culture: Where Old and New Media Collide. New York: New York University Press, 2006.

[3] A. M. Barik, R. Mahendra, and M. Adriani, "Normalization of Indonesian-English code-mixed Twitter data," in Proc. 5th Workshop on Noisy User-generated Text (W-NUT 2019), Hong Kong, 2019, pp. 417–424. doi: 10.18653/v1/D19-5554.

[4] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," Expert Systems with Applications, vol. 73, pp. 220–239, 2017.

[5] A. Fernandez, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," Journal of Artificial Intelligence Research, vol. 61, pp. 863–905, 2018.

[6] D. J. Hand, "Measuring classifier performance: A coherent alternative to the area under the ROC curve," Machine Learning, vol. 77, no. 1, pp. 103–123, 2009.

[7] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," Intelligent Data Analysis, vol. 6, no. 5, pp. 429–449, 2002.

[8] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[10] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in Proc. IEEE Int. Joint Conf. Neural Networks, Hong Kong, 2008, pp. 1322–1328.

[11] K. M. Ting, "A comparative study of cost-sensitive algorithms for instance ranking," in Proc. 6th IEEE Int. Conf. Data Mining, Hong Kong, 2006, pp. 1001–1006.

[12] A. Luque, A. Carrasco, A. Martin, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," Pattern Recognition, vol. 91, pp. 216–231, 2019.

[13] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, no. 1, pp. 1–13, 2020.

[14] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in Proc. 23rd Int. Conf. Machine Learning, Pittsburgh, 2006, pp. 233–240.

[15] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.

[16] F. Koto and G. Y. Rahmaningtyas, "InSet lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs," in Proc. Int. Conf. Asian Language Processing (IALP), Singapore, 2017, pp. 391–394.

[17] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP," in Proc. 28th Int. Conf. Computational Linguistics, Barcelona (Online), 2020, pp. 757–770.

[18] B. Wilie, K. Vincentio, G. I. Winata, S. Cahyawijaya, X. Li, Z. Y. Lim, S. Soleman, R. Mahendra, P. Fung, S. Bahar, and A. Purwarianti, "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," in Proc. 1st Conf. Asia-Pacific Chapter of the ACL and 10th Int. Joint Conf. Natural Language Processing, Suzhou, 2020, pp. 843–857.

[19] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Proc. 26th Int. Conf. Neural Information Processing Systems, Lake Tahoe, 2013, pp. 3111–3119.

[20] D. D. Novicahyanto, "A comparative analysis of text classification algorithms for the information-sharing (timestamping) culture in YouTube comments on the Ruangguru Clash of Champions," B.S. thesis, Faculty of Computer Science, Soegijapranata Catholic University, Semarang, 2026.

[21] J. Cohen, "A coefficient of agreement for nominal scales," Educational and Psychological Measurement, vol. 20, no. 1, pp. 37–46, 1960.