Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 6, No. 5, October 2025, Page. 3769-3786
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5177

Enhancing Fake News Detection on Imbalanced Data Using Resampling Techniques and Classical Machine Learning Models

Dodo Zaenal Abidin *[1], Agus Siswanto[2], Chindra Saputra[3], Bhetantio[4], Afrizal Nehemia Toscany[5]

[1, 4] Magister of Information System, Faculty of Computer Science, Universitas Dinamika Bangsa, Jambi, Indonesia
[2,3, 5] Informatics Engineering, Faculty of Computer Science, Universitas Dinamika Bangsa, Jambi, Indonesia

Email: [1]dodozaenalabidin@gmail.com

Received : Jul 26, 2025; Revised : Aug 12, 2025; Accepted : Aug 13, 2025; Published : Oct 22, 2025

Abstract

Class imbalance remains a critical challenge in fake news detection, particularly in domains such as entertainment media where class distributions are highly skewed. This study evaluates seven resampling techniques—Random Oversampling, SMOTE, ADASYN, Random Undersampling, Tomek Links, NearMiss, and No Resampling—applied to three classical machine learning models: Logistic Regression, Support Vector Machine (SVM), and Random Forest. Using the imbalanced GossipCop dataset comprising 24,102 news headlines, the proposed pipeline integrates TF-IDF vectorization, stratified 3-fold cross-validation, and five evaluation metrics: F1-score, precision, recall, ROC AUC, and PR AUC. Experimental results show that oversampling methods, particularly SMOTE and Random Oversampling, substantially improve minority class (fake news) detection. Among all model–resampling combinations, SVM with SMOTE achieved the highest performance (F1-score = 0.67, PR AUC = 0.74), demonstrating its robustness in handling imbalanced short-text classification. Conversely, undersampling methods frequently reduced recall, especially with ensemble models like Random Forest. This approach enhances model robustness in fake news detection on skewed datasets and contributes a reproducible, domain-specific framework for developing more reliable misinformation classifiers.

Keywords : *class imbalance, fake news classification, logistic regression, resampling techniques, random forest, support vector machine.*

1. INTRODUCTION

The proliferation of digital media has facilitated the rapid spread of fake news, which can significantly distort public opinion and undermine trust in reliable sources [1], [2].]. Headlines, as short-text representations of news content, are particularly vulnerable to manipulation and misinterpretation on social platforms [3]. Automated fake news detection has therefore emerged as a critical area in natural language processing (NLP), requiring accurate and generalizable models to distinguish between true and false information [4], [5]. While numerous studies have addressed this problem, relatively few have systematically investigated the combined impact of multiple resampling techniques with different classical machine learning models, leaving a methodological gap that this study aims to address.

A major challenge in fake news classification is the class imbalance, where real news instances significantly outnumber fake news. This imbalance often biases models toward the majority class, resulting in poor performance on the minority class—typically the more critical one to detect [6], [7]. Classical models such as logistic regression, support vector machines (SVM), and ensemble methods like random forest are commonly applied, but their effectiveness can deteriorate under highly imbalanced distributions [8].

Vol. 6, No. 5, October 2025, Page. 3769-3786
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5177

To mitigate this issue, numerous resampling techniques have been introduced. Oversampling methods (e.g., SMOTE, ADASYN) synthesize minority class samples, while undersampling techniques (e.g., Tomek Links, NearMiss) reduce overrepresented majority samples [9], [10], [11]. Despite their usefulness, many existing studies evaluate only one technique or rely on single train-test splits, without addressing imbalance systematically—limiting reproducibility and raising concerns of biased performance estimates.

For instance, studies like Khanal et al. [12] conducted a benchmark study of various machine learning and deep learning models for online fake news detection on diverse datasets; however, their work did not explicitly focus on or systematically address the critical challenge of class imbalance. Similarly, Elsaeed et al. [13] proposed a voting classifier for fake news detection, achieving high accuracy, but did not indicate specific strategies for handling imbalanced data, which can lead to misleading results for the minority class. While Hossain et al. [9] explored imbalance handling (using SMOTE) and model stacking for fake news detection in Bangla, their scope of resampling techniques was limited. Furthermore, methodologically relevant studies such as Yao et al. [10] and Budhi et al. [11] conducted comparative analyses of resampling techniques, but in the distinct domain of fake online review detection, which limits direct applicability to the fake news domain. These findings underscore the pressing need for a more systematic and robust comparison across various resampling strategies and classifier types specifically within the context of fake news classification.

Beyond that, it is quite clear that very few studies genuinely integrate diverse resampling techniques and classifier models into a single, cohesive experimental pipeline, complete with thorough cross-validation, careful metric aggregation, and insightful visualization [14]. Consequently, understanding which resampling method and classifier pairings consistently perform well across each data fold remains a significant challenge [15]. Ironically, despite their immense value, ROC and PR curve visualizations are still rarely utilized in fake news classification research. Yet, these plots offer crucial insights into how a classifier truly behaves at different thresholds, especially when dealing with imbalanced datasets [16].

By offering a thorough comparison of seven resampling methods—including SMOTE, ADASYN, TomekLinks, and others—applied to three popular classifiers—logistic regression, support vector machines, and random forest—this study fills that gap [17]. To verify reliability, the models are assessed using performance metrics such as F1-score, precision, recall, ROC AUC, and PR AUC under a stratified 3-fold cross-validation scheme [18]. To aid in interpretability and reproducibility, visual comparisons and aggregated classification reports are also included [19].

This study makes three main contributions. First, it offers a methodical benchmarking of oversampling and undersampling methodologies to overcome the class imbalance in the categorization of fake news [20], [17]. Second, it uses reliable, cross-validated metrics including F1-score, precision, recall, ROC AUC, and PR AUC to assess different combinations of resampling techniques and classification models [21]. Third, it employs a pipeline-based, organized experimentation methodology that incorporates metric aggregation and thorough visualizations to assist interpretability, ensuring reproducibility.

The remainder of this article is organized as follows: Section 2 presents a review of related work on class imbalance and fake news detection, emphasizing prior limitations in resampling evaluation. Section 3 outlines the research methodology, including dataset preprocessing, resampling strategies, and model configurations used in the experimental pipeline. Section 4 reports the experimental results along with visual and metric-based analyses, highlighting key findings from 21 model–resampling combinations. Finally, Section 5 concludes the study and discusses directions for future research, including hybrid resampling, integration with deep learning models, and domain-adaptive modeling for broader applicability.

Vol. 6, No. 5, October 2025, Page. 3769-3786
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5177

2. METHOD

This study implements a systematic and reproducible workflow to evaluate the effectiveness of various resampling techniques for handling class imbalance in fake news classification. The experimental pipeline focuses on benchmarking three classical machine learning models—Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF)—using multiple resampling methods across stratified cross-validation folds [11]. The study is based on the GossipCop dataset, a widely used benchmark in fake news detection research [22]. To ensure robust evaluation, all experiments were repeated over three folds with aggregated metrics for consistent comparison.

The methodological framework is organized into eight sequential stages: (1) data collection, (2) preprocessing and TF-IDF transformation, (3) application of resampling techniques, (4) model training across defined combinations, (5) stratified 3-fold cross-validation, (6) metric aggregation, (7) result visualization, and (8) best combination selection. The pipeline also incorporates confusion matrix analysis, per-class performance reports, and graphical evaluation (ROC and PR curves). Each stage is illustrated in Figure 1 and elaborated in the following subsections.

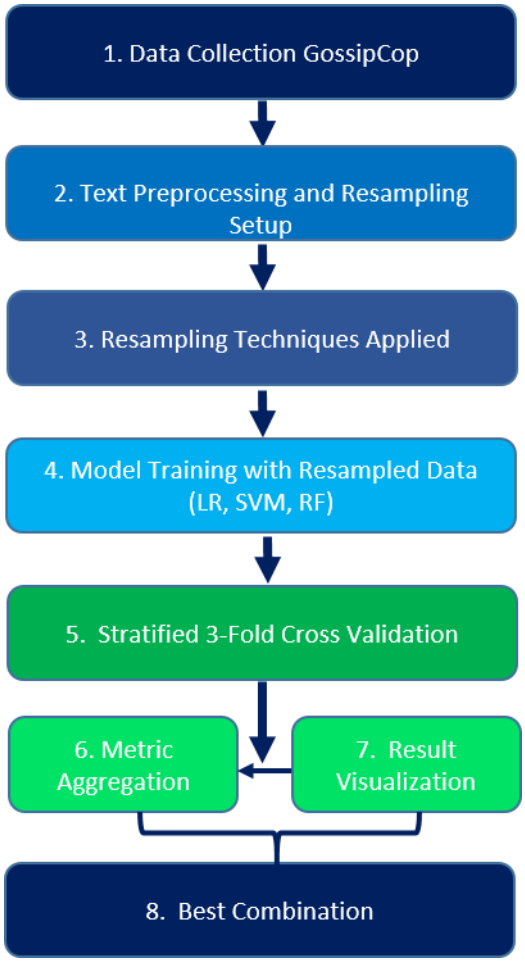

Figure 1. Workflow of the Proposed Resampling-Based Fake News Classification Pipeline.

Figure 1 presents the proposed workflow for evaluating resampling methods in the context of fake news classification through cross-validation. The process begins with collecting data from the GossipCop dataset, consisting of news headlines labeled as either real or fake [23], [24]. During preprocessing, the raw text is normalized and transformed into numerical representations using the TF-IDF method [25]. To address class imbalance, seven resampling techniques are applied: No Resampling

Vol. 6, No. 5, October 2025, Page. 3769-3786
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5177

(baseline), Random Oversampling (ROS), SMOTE, ADASYN, Random Undersampling (RUS), Tomek Links, and NearMiss [10]. Each balanced dataset is then classified using one of three algorithms—Logistic Regression, Support Vector Machine (SVM), or Random Forest. Model performance is assessed under a stratified 3-fold cross-validation scheme, with metrics including F1-score, precision, recall, ROC AUC, and PR AUC, along with confusion matrix components. The outcomes are aggregated and visualized to identify the most effective resampling–model combinations [25]. To enhance clarity and reproducibility, the key stages of this experimental pipeline are summarized in Table 1, which complements Figure 1 by providing a concise, step-by-step description of each process from data collection to result visualization, fully aligned with both the implemented code and the methodological workflow.

Table 1. Summary of the Experimental Pipeline for Fake News Detection

Step No	Process Stage	Description
1	Data Collection	Load the GossipCop dataset containing 24,102 news headlines labeled as real (0) or fake (1) from the FakeNewsNet repository. Handle missing values by removing rows with null entries.
2	Text Preprocessing	Convert all text to lowercase and remove missing entries. No stemming or stopword removal is performed to preserve semantic context, while TF-IDF internally handles basic tokenization and punctuation removal.
3	Feature Extraction (TF-IDF)	Convert text into numerical vectors using Term Frequency–Inverse Document Frequency (TF-IDF). Fitting is performed only on the training folds to prevent data leakage.
4	Resampling	Apply one of seven techniques: No Resampling (baseline), Random Oversampling, SMOTE, ADASYN, Random Undersampling, Tomek Links, or NearMiss. Implemented using the imbalanced-learn library.
5	Model Selection	Choose one of three classifiers: Logistic Regression (LR), Support Vector Machine (SVM), or Random Forest (RF). All hyperparameters are set to default values unless otherwise specified.
6	Model Training	Integrate TF-IDF, resampling, and model training into a unified pipeline (Pipeline or ImbPipeline). Each model–resampling combination is trained separately.
7	Cross-Validation	Perform Stratified 3-Fold Cross-Validation to maintain the same class distribution in each fold. Use random_state=42 for reproducibility.
8	Evaluation Metrics	$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$, $Precision = \frac{TP}{TP+FP}$ $Recall = \frac{TP}{TP+FN}$, $F1-Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$
9	Result Aggregation	Average metric scores across the three folds to obtain stable performance estimates for each model–resampling configuration.

2.1. Dataset

This study utilizes the GossipCop dataset, a widely adopted benchmark for fake news detection in the entertainment domain [22], [23]. The dataset comprises short-text news headlines, each assigned a binary label: 0 for real news and 1 for fake news. These labels originate from the FakeNewsNet repository, which aggregates data from fact-checking websites and credible news sources. For this

Vol. 6, No. 5, October 2025, Page. 3769-3786
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5177

research, only the headline text is retained to align with short-text classification objectives, excluding article body content to ensure uniform feature length and minimize noise.

The curated version used in this work contains 24,102 entries after duplicate removal, handling of missing values, and basic text normalization (lowercasing, punctuation removal). Each instance consists of a headline as the sole predictive feature and its corresponding binary label as the classification target.

A notable characteristic of the dataset is its class imbalance, a critical issue in supervised learning [10], [26]. As shown in Figure 2, the majority class (real news, label 0) comprises 16,817 instances, while the minority class (fake news, label 1) contains only 5,323 instances, resulting in an approximate 3:1 imbalance ratio. This imbalance tends to bias classifiers toward the majority class, reducing recall and F1-score for fake news detection. Therefore, multiple resampling techniques are explored in this study to rebalance the training data, aiming to improve the minority-class detection capability and overall classification reliability [27].

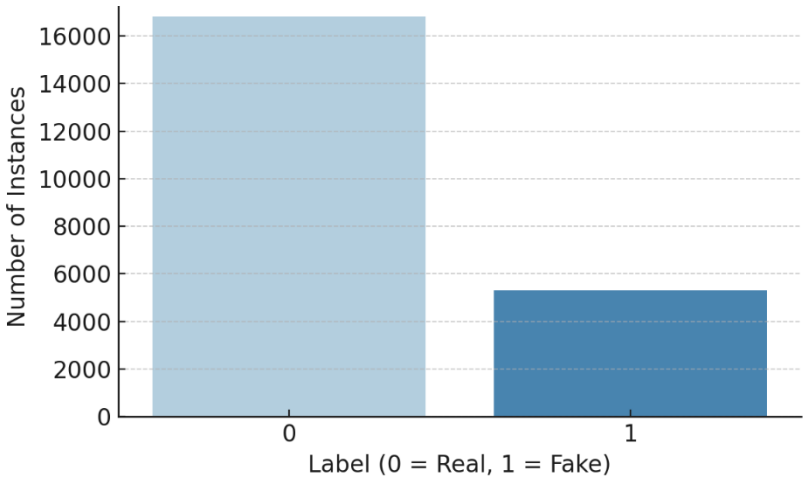

Figure 2. Class distribution of the GossipCop dataset used in this study

Given this imbalance, the GossipCop dataset offers a robust benchmark for assessing the impact of various resampling strategies on fake news classification performance. In this study, oversampling, undersampling, and hybrid techniques are systematically applied under identical experimental conditions to evaluate their effects on minority-class recall, precision, and overall model robustness. This controlled framework ensures a fair comparison across all model–resampling configurations, enabling the identification of optimal combinations for imbalanced short-text classification. The findings are expected to provide practical guidance for building more reliable and generalizable hoax detection systems in domains characterized by highly skewed class distributions and concise textual content.

2.2. Text Preprocessing and Resampling Setup

Prior to model training, textual preprocessing was applied to prepare raw news headlines extracted from the GossipCop dataset, which consists of short textual claims annotated as either real or fake [28]. Initial preprocessing included standard normalization steps such as lowercasing and the removal of missing entries. Considering the concise nature of the headlines, advanced linguistic operations like stemming or stopword removal were deliberately omitted to preserve contextual semantics and avoid discarding potentially informative tokens.

Subsequently, the cleaned textual data were transformed into numerical feature representations using the Term Frequency–Inverse Document Frequency (TF-IDF) approach [29]. This method

Vol. 6, No. 5, October 2025, Page. 3769-3786
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5177

emphasizes discriminative terms by scaling their frequency inversely with their overall occurrence across the corpus, thus reducing the impact of overly common words. To ensure methodological rigor, TF-IDF fitting was exclusively conducted on training folds within each cross-validation split to prevent information leakage from test data.

To address the class imbalance present in the dataset, where fake news samples are substantially outnumbered by real news instances, a diverse set of resampling strategies was introduced [30]. Seven techniques were evaluated: No Resampling (as baseline), Random Undersampling (RUS), Tomek Links, NearMiss, Random Oversampling (ROS), Synthetic Minority Oversampling Technique (SMOTE), and Adaptive Synthetic Sampling (ADASYN). These methods reflect both undersampling and oversampling paradigms widely applied in imbalanced classification scenarios. RUS performs random deletion of majority class samples; Tomek Links eliminates ambiguous boundary instances; NearMiss selects majority instances based on nearest-neighbor proximity to the minority class. In contrast, ROS duplicates existing minority samples, while SMOTE and ADASYN synthetically generate new minority examples via interpolation and adaptive neighborhood distributions, respectively [31],[32].

All resampling techniques were incorporated into modular pipelines alongside TF-IDF vectorization and classifier instantiation. This unified design, implemented using the imbalanced-learn library, ensures consistency, reproducibility, and fair comparative evaluation across multiple classifier–resampling configurations. The adoption of a modular pipeline design offers several advantages for this study. By encapsulating preprocessing, resampling, and classification within a single reproducible workflow, the approach minimizes human error, facilitates transparent experiment replication, and ensures identical data transformations across all folds and configurations. Moreover, this design enables a controlled comparison of resampling methods by keeping all other processing stages constant, thereby isolating the effect of each technique on model performance. Such methodological rigor aligns with best practices in imbalanced classification research and supports the reproducibility standards expected in high-quality academic publications [30], [32].

2.3. Applied Resampling Techniques

To address the class imbalance present in the GossipCop dataset, this study implemented seven well-established resampling techniques. These methods span both undersampling and oversampling paradigms, facilitating a comparative evaluation of their influence on binary fake news classification [34]. All techniques were integrated within a unified pipeline using the imbalanced-learn library to ensure consistency and reproducibility across models and folds [35].

Random Undersampling (RUS) reduces the majority class by randomly eliminating samples. While computationally efficient, this method risks discarding potentially informative data, which may degrade model generalizability. Tomek Links refine decision boundaries by identifying overlapping sample pairs from different classes—termed Tomek links—and removing the majority class sample. This results in cleaner class separation and reduced boundary noise [33].

NearMiss, particularly the NearMiss-1 variant used here, selects majority class samples closest to minority instances. Though it enhances class proximity, excessive data removal may impact overall learning capacity. Random Oversampling (ROS) replicates minority class instances to balance the class distribution. Despite its simplicity, ROS can lead to overfitting due to redundant information [34]].

Synthetic Minority Oversampling Technique (SMOTE) improves upon ROS by synthetically generating new minority samples through linear interpolation between nearest neighbors, thereby expanding the decision region and improving diversity. Adaptive Synthetic Sampling (ADASYN) further enhances SMOTE by generating more synthetic data for harder-to-learn instances. This adaptive focus helps improve decision boundaries for imbalanced data [35].

Vol. 6, No. 5, October 2025, Page. 3769-3786
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5177

The integration of these resampling methods within the classification pipeline—comprising TF-IDF vectorization and classical classifiers (Logistic Regression, Support Vector Machine, and Random Forest)—supports robust, fair benchmarking. The selected methods are widely cited and have shown effectiveness in handling class imbalance in natural language processing and fake news detection contexts [36]. The selection of these seven techniques reflects a deliberate methodological choice to cover complementary imbalance-handling strategies, ensuring that both data reduction and data generation approaches are represented. This comprehensive inclusion enables the study to capture performance variations arising from fundamentally different resampling philosophies, thereby providing a more complete and unbiased benchmarking of resampling–classifier interactions in the context of fake news detection.

2.4. Model Training with Resampled Data

This research employed three widely recognized traditional machine learning algorithms—Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF)—to evaluate the impact of different resampling strategies on binary fake news classification [37]. These models were selected due to their proven performance in text classification, interpretability, and ability to handle high-dimensional sparse features typical of TF-IDF representations.

To ensure methodological consistency and reproducibility, all classifiers were implemented within a unified modular pipeline comprising three core components: (1) TF-IDF vectorization, (2) resampling, and (3) model training. First, TF-IDF converted raw news headlines into numerical feature vectors, with fitting performed exclusively on the training partition within each cross-validation fold to prevent information leakage [38]. Second, one of the seven resampling techniques described in Section 2.3 was applied; for the baseline scenario, this step was omitted. Finally, the chosen classifier was trained on the resampled TF-IDF features [39].

The pipelines were constructed using the Pipeline and ImbPipeline utilities from the scikit-learn and imbalanced-learn libraries, respectively. Model evaluation followed a Stratified 3-Fold Cross-Validation (CV) protocol to preserve the original class distribution within each fold, ensuring balanced and unbiased testing. For each fold, the pipeline was trained on resampled data and evaluated on unseen validation data, producing performance metrics including F1-score, precision, recall, ROC AUC, and PR AUC, alongside confusion matrix analysis to assess class-specific behavior.

By integrating vectorization, resampling, and classification into a single modular framework, this study ensured that every model–resampling combination was evaluated under identical preprocessing and validation conditions. This design not only enhanced experimental fairness but also enabled robust, direct comparisons across classifiers and resampling strategies, strengthening the validity of the conclusions.

2.5. Cross-Validation Strategy

To ensure fair and reliable evaluation in imbalanced binary classification, this study adopted a Stratified 3-Fold Cross-Validation (CV) scheme [40]. Unlike standard k-fold partitioning, stratified CV preserves the original class distribution within each fold, which is critical for imbalanced datasets to avoid skewed representation of the majority class and to strengthen evaluation robustness [41].

In this setup, the dataset was randomly shuffled using a fixed random_state=42 for reproducibility before being divided into three stratified subsets. In each iteration, two folds were used for training and one fold for validation, ensuring that all folds served as the validation set exactly once. The choice of three folds balances computational efficiency and reliability, as increasing the number of folds would improve stability but significantly raise processing time when evaluating multiple models and resampling strategies.

Vol. 6, No. 5, October 2025, Page. 3769-3786
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5177

All preprocessing stages—TF-IDF fitting, resampling, and model training—were performed exclusively on the training folds to prevent any information leakage into the validation data. The held-out validation fold remained untouched until the evaluation stage, adhering to best practices in text classification.

For each fold, evaluation metrics including precision, recall, F1-score, ROC AUC, and PR AUC were computed and subsequently averaged to yield stable overall performance estimates. This averaging process reduces metric variance and offers a more realistic measure of generalization capability in real-world scenarios. Integrating Stratified 3-Fold CV into a unified modular pipeline ensured methodological rigor, reproducibility, and equitable benchmarking across all classifier–resampling combinations [42].

2.6. Evaluation Metrics

This study adopted a comprehensive set of performance metrics to evaluate the effectiveness of different resampling–classifier combinations under imbalanced binary classification conditions [42]. For each fold in the Stratified 3-Fold Cross-Validation, five primary metrics were computed—Precision, Recall, F1-score, ROC AUC, and PR AUC—and then averaged to produce stable overall estimates. All computations were performed using the sklearn.metrics module within the evaluation stage of the unified pipeline.

The F1-score was emphasized as the primary metric, as it provides a harmonic balance between precision and recall, making it particularly suitable for imbalanced tasks such as fake news detection. Precision, defined as the ratio of true positives to all predicted positives, reflects the model's ability to minimize false alarms. Recall, calculated as the proportion of true positives among actual positives, measures the model's sensitivity in identifying minority-class instances [43].

For probabilistic evaluations, ROC AUC and PR AUC were included to capture model performance across varying classification thresholds. ROC AUC offers a threshold-independent measure of class separability [44], while PR AUC focuses specifically on minority-class performance, making it more informative in imbalanced settings [45]. These metrics were computed using roc_auc_score and average_precision_score when probabilistic outputs (predict_proba) were available.

Additionally, confusion matrices were generated for each model–resampling configuration to provide an interpretable view of misclassification patterns, highlighting Type I errors (false positives) and Type II errors (false negatives). The matrices were aggregated across folds and stored for further analysis, supporting error diagnosis and comparative interpretation.

By calculating and averaging these metrics within the cross-validation framework, the evaluation process ensured a fair, rigorous, and reproducible comparison of all 21 tested configurations, thereby offering a robust foundation for result interpretation.

3. RESULT AND DISCUSSIONS

This chapter reports the main findings from our evaluation of fake news classification on the imbalanced GossipCop dataset, following the methodological setup in Figure 1 and Table 1. Three classical classifiers—Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF)—were tested with seven resampling strategies (Random Undersampling, Tomek Links, NearMiss, Random Oversampling, SMOTE, ADASYN) plus a baseline without resampling. Each model–resampling pair was evaluated within a unified TF-IDF-based pipeline using stratified 3-fold cross-validation to ensure balanced class distribution, fairness, and reproducibility. Sections 3.1–3.4 outline the evaluation protocol, present aggregated metrics and visual comparisons, and discuss the findings in relation to prior studies, highlighting the most effective configurations for imbalanced fake news detection.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 5, October 2025, Page. 3769-3786
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5177

3.1. Evaluation Protocol and Metric Design

To assess model performance on the imbalanced GossipCop dataset, we employed stratified 3-fold cross-validation, preserving class distribution across folds while ensuring computational feasibility. This approach enables a fair comparison among model–resampling configurations and reduces sampling bias.

Each fold followed a unified pipeline comprising TF-IDF vectorization, optional resampling, and model training, all applied exclusively to the training data to avoid information leakage. The test set in each fold remained unseen until final prediction. Hyperparameters and random seeds were fixed for consistency, and the pipeline was implemented using Pipeline and StratifiedKFold from scikit-learn and imbalanced-learn.

Five evaluation metrics were used: F1-score, precision, recall, ROC AUC, and PR AUC. F1-score served as the primary metric due to its suitability for imbalanced classification. Precision measures the ability to avoid false alarms, while recall reflects sensitivity to hoax detection. ROC AUC and PR AUC offer threshold-independent evaluation, with PR AUC providing better insight into minority-class detection.

Confusion matrices were generated for each fold and aggregated to reveal dominant error patterns (false positives and false negatives), ensuring robustness, reproducibility, and interpretability across all 21 model–resampling configurations..

3.2. Model and Resampling Performance Overview

This section compares 21 model–resampling combinations using macro F1-score, ROC AUC, and PR AUC, supported by aggregated confusion matrices. Figure 3 shows that Random Oversampling and SMOTE consistently deliver the highest F1-scores across all models, with Logistic Regression and SVM exceeding 0.66. In contrast, NearMiss and No Resampling yield the lowest scores, highlighting their limited effectiveness in addressing class imbalance.

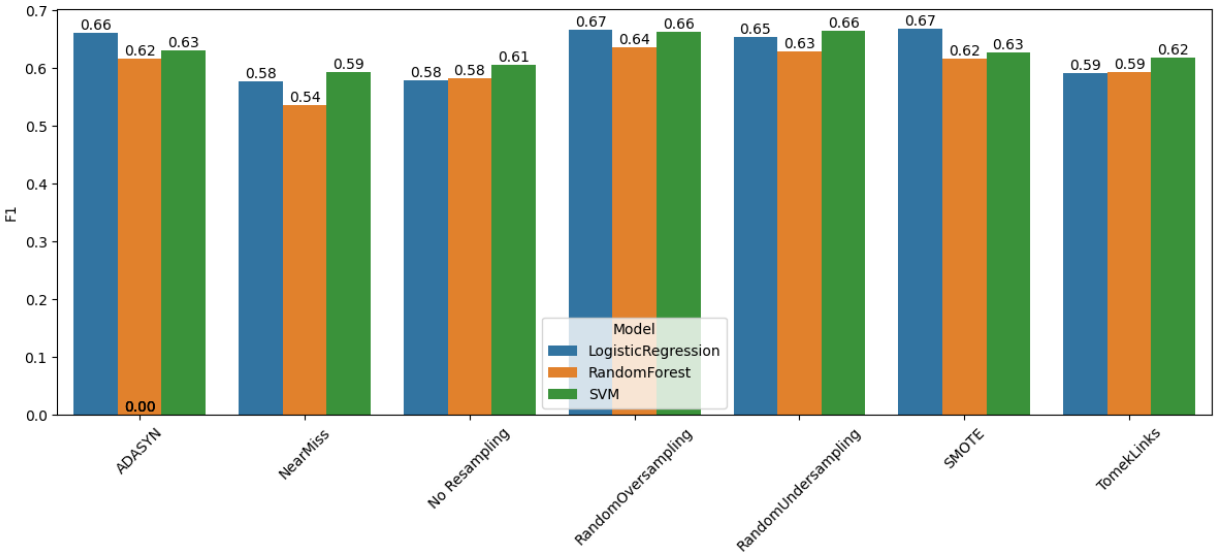

Figure 3. F1-score comparison across resampling techniques and models

Figure 3. F1-score comparison across resampling techniques and models. The chart shows macro-averaged F1-scores for Logistic Regression, Random Forest, and SVM across seven resampling strategies. Random Oversampling and SMOTE consistently outperform other methods, with Logistic Regression and SVM achieving F1-scores up to 0.67. In contrast, NearMiss and No Resampling yield the weakest results, particularly for Random Forest, underscoring the limitations of aggressive

Vol. 6, No. 5, October 2025, Page. 3769-3786
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5177

undersampling or the absence of class balancing in capturing minority patterns. ADASYN performs competitively for Logistic Regression and SVM but fails entirely with Random Forest, likely due to poor synthetic sample generation for tree-based models in this setting.

Figure 4 extends the F1-score analysis by comparing PR AUC values, which are more indicative in imbalanced settings. Results reaffirm the strength of synthetic oversampling—particularly SMOTE, Tomek Links, and Random Oversampling—when paired with margin-based or linear models. SVM with SMOTE or Tomek Links achieved the highest PR AUC, reflecting strong precision–recall trade-offs. In contrast, ADASYN with Random Forest produced the lowest scores (F1 = 0.00, PR AUC = 0.00), likely due to noisy synthetic samples causing overfitting. These results underscore the need to match resampling methods with model characteristics to maximize learning and generalization.

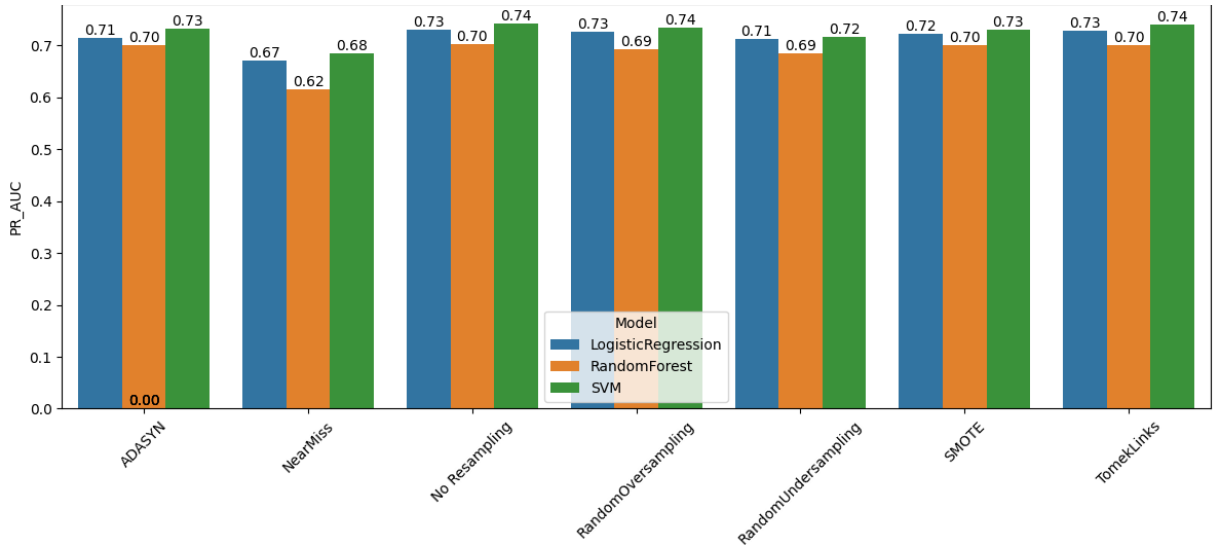

Figure 4. PR AUC Comparison across Resampling Techniques and Models

Figure 4. PR AUC comparison across resampling techniques and models. SVM, when paired with SMOTE, TomekLinks, or RandomOversampling, consistently achieves the highest PR AUC values (0.74), reflecting a strong balance between precision and recall under imbalanced conditions. Logistic Regression delivers comparable results, particularly with RandomOversampling and TomekLinks, also reaching PR AUC = 0.74. In contrast, Random Forest shows greater variability across resampling strategies. The poorest performance is observed with Random Forest + ADASYN, where PR AUC drops to 0.00—indicating a complete inability to identify minority class instances, likely due to noisy or unrepresentative synthetic samples generated by ADASYN, which severely hinder the model's learning capability.

Similarly, Random Forest combined with NearMiss records a PR AUC of only 0.62, with Table 2 revealing an excessively high false positive (FP) count of 1,660. Such aggressive undersampling may remove valuable majority-class information, degrading precision and limiting recall. By contrast, oversampling techniques such as RandomOversampling and SMOTE provide the most consistent precision–recall trade-off, particularly for margin-based and linear models, and maintain balanced performance across folds.

To further illustrate predictive behavior, Table 2 presents aggregated confusion matrix results. SVM paired with RandomOversampling achieved a TP of 1,063.67 with an FP of just 372, while SVM with SMOTE and TomekLinks recorded even lower FP counts (223.33 and 197.0, respectively) albeit with slightly lower TP values. These patterns confirm that the top PR AUC–yielding combinations are

Vol. 6, No. 5, October 2025, Page. 3769-3786
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5177

also those with low FP rates, underscoring their superior capability to generalize and accurately identify fake news in imbalanced datasets.

Table 2. Averaged Confusion Matrix per Model–Resampling Combination (3-Fold CV)

Resampling	Model	TP	FP	FN	TN
ADASYN	LogisticRegression	1298.0	857.0	476.33	4748.67
	RandomForest	958.33	376.0	816.0	5229.67
	SVM	920.0	226.33	854.33	5379.33
NearMiss	LogisticRegression	1265.33	1345.33	509.0	4260.33
	RandomForest	1257.67	1660.0	516.67	3945.67
	SVM	1250.67	1197.33	523.67	4408.33
No Resampling	LogisticRegression	799.67	186.67	974.67	5419.0
	RandomForest	834.67	253.0	939.67	5352.67
	SVM	845.67	172.0	928.67	5433.67
RandomOversampling	LogisticRegression	1277.0	782.33	497.33	4823.33
	RandomForest	1051.0	477.67	723.33	5128.0
	SVM	1063.67	372.0	710.67	5233.67
RandomUndersampling	LogisticRegression	1322.33	943.0	452.0	4662.67
	RandomForest	1311.67	1086.0	462.67	4519.67
	SVM	1317.67	871.67	456.67	4734.0
SMOTE	LogisticRegression	1234.67	686.67	539.67	4919.0
	RandomForest	944.0	341.33	830.33	5264.33
	SVM	913.67	223.33	860.67	5382.33
TomekLinks	LogisticRegression	832.0	208.0	942.33	5397.67
	RandomForest	869.33	288.33	905.0	5317.33
	SVM	881.0	197.0	893.33	5408.67

To further analyze predictive behavior, Table 2 presents the aggregated confusion matrix results across all resampling techniques. Models trained with RandomOversampling and SMOTE consistently achieved higher true positive (TP) counts while maintaining relatively low false positive (FP) rates, especially when paired with SVM. For instance, SVM with RandomOversampling recorded a TP of 1,063.67 and an FP of 372, indicating strong sensitivity and balanced generalization in detecting fake news.

Conversely, NearMiss produced markedly higher FP rates across all models—exceeding 1,600 in the Random Forest configuration—suggesting that aggressive undersampling removed critical majority-class information and degraded decision boundaries. Similarly, while ADASYN yielded competitive results for Logistic Regression and SVM, it performed poorly with Random Forest, producing a TP of

Vol. 6, No. 5, October 2025, Page. 3769-3786
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5177

only 958.33, likely due to the generation of noisy or unrepresentative synthetic instances that failed to enhance minority-class learning.

Overall, the confusion matrix analysis aligns with scalar metric trends, reinforcing that effective class imbalance handling not only improves F1-score and AUC metrics but also yields more favorable TP–FP distributions. These patterns provide further evidence supporting the importance of model–resampling compatibility, setting the stage for a deeper per-class performance analysis in the following section..

3.3. Per-Class Performance and Classification Reports

To gain a deeper understanding of model behavior, it is necessary to complement aggregate metrics with class-level evaluation. In hoax detection, the minority class (class 1) typically suffers from reduced recall and F1-score due to imbalance, even when overall accuracy is high. Per-class precision, recall, and F1-score reveal how effectively models differentiate legitimate news (class 0) from hoaxes (class 1). This is particularly important for imbalanced datasets like GossipCop, where the majority class is easier to classify, while minority detection remains more challenging. Accordingly, Table 3 presents per-class F1-scores for each model–resampling configuration, offering insight into sensitivity and generalization toward minority class predictions.

Table 3. Per-Class F1-Scores by Model and Resampling Method
(Averaged over 3-Fold Cross-Validation)

Model	Resampling	F1 (Class 0)	F1 (Class 1)
LogisticRegression	No Resampling	0.9032	0.5793
LogisticRegression	RandomUndersampling	0.8699	0.6547
LogisticRegression	TomekLinks	0.9037	0.5913
LogisticRegression	NearMiss	0.8213	0.5771
LogisticRegression	RandomOversampling	0.8829	0.6662
LogisticRegression	SMOTE	0.8892	0.6682
LogisticRegression	ADASYN	0.8769	0.6607
SVM	No Resampling	0.9080	0.6058
SVM	RandomUndersampling	0.8770	0.6649
SVM	TomekLinks	0.9084	0.6177
SVM	NearMiss	0.8367	0.5924
SVM	RandomOversampling	0.9063	0.6627
SVM	SMOTE	0.9085	0.6277
SVM	ADASYN	0.9087	0.6300
RandomForest	No Resampling	0.8998	0.5833
RandomForest	RandomUndersampling	0.8537	0.6288
RandomForest	TomekLinks	0.8991	0.5930
RandomForest	NearMiss	0.7838	0.5361
RandomForest	RandomOversampling	0.8952	0.6364
RandomForest	SMOTE	0.8999	0.6171
RandomForest	ADASYN	0.8977	0.6166

As shown in Table 3, all models consistently achieved higher F1-scores for class 0 than for class 1, reaffirming the tendency of imbalanced datasets to favor majority-class accuracy. For Logistic

Vol. 6, No. 5, October 2025, Page. 3769-3786
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5177

Regression, RandomOversampling and SMOTE produced the most balanced results, elevating the minority class F1-score to 0.6662 and 0.6682, respectively, with only minor reductions in class 0 performance. SVM demonstrated stable improvements across most oversampling methods, particularly RandomOversampling (F1=0.6627) and ADASYN (F1=0.6300), while maintaining class 0 scores above 0.90. Random Forest showed a similar trend, with SMOTE (F1=0.6171) and ADASYN (F1=0.6166) improving class 1 recognition without severely degrading class 0 accuracy.

These results suggest that synthetic oversampling methods—especially SMOTE and ADASYN—are generally more effective than undersampling in mitigating class imbalance, particularly when paired with linear models like Logistic Regression or margin-based models like SVM.

To visually reinforce these patterns, Figure 5 compares class 1 (hoax) F1-scores across all model–resampling combinations. The chart clearly highlights the consistent advantage of SMOTE and RandomOversampling, especially with SVM and Logistic Regression, both achieving class 1 F1-scores above 0.66.

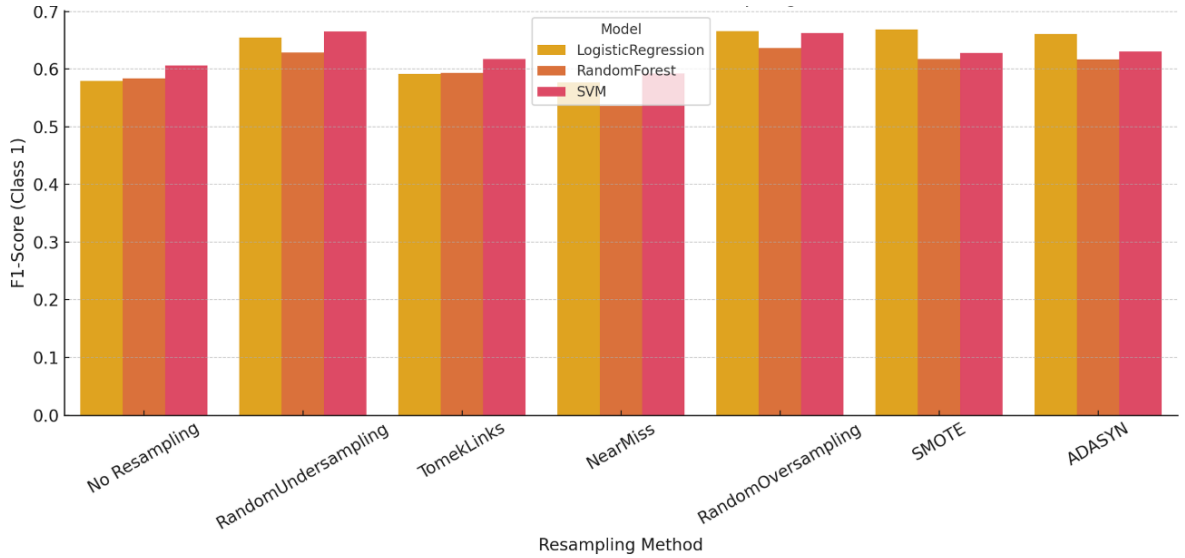

Figure 5. Bar Chart of F1-Score for Class 1 Across Model–Resampling Combinations

Based on the data presented in Figure 5, the highest minority-class (Class 1) F1-scores are achieved by Logistic Regression + SMOTE (0.6682), Logistic Regression + RandomOversampling (0.6662), and SVM + RandomOversampling (0.6627), confirming that oversampling approaches are particularly effective in enhancing hoax detection. These methods consistently improve Class 1 performance while maintaining high Class 0 scores, thereby achieving an optimal precision–recall trade-off.

Logistic Regression demonstrates substantial gains under SMOTE and RandomOversampling, whereas SVM exhibits greater stability across resampling methods, with most Class 1 F1-scores exceeding 0.62, even under less favorable configurations such as ADASYN or TomekLinks. In contrast, Random Forest displays greater variability, and combinations with aggressive undersampling techniques such as NearMiss yield the lowest Class 1 F1-score (0.5361).

Overall, the per-class evaluation underscores the pivotal role of resampling strategies in addressing class imbalance for hoax classification. Oversampling methods—particularly SMOTE and RandomOversampling—stand out as the most consistent performers, especially when paired with linear or margin-based classifiers. These findings highlight that selecting an appropriate model–resampling combination is critical for achieving reliable minority-class recognition without compromising majority-class accuracy in imbalanced fake news detection scenarios.

Vol. 6, No. 5, October 2025, Page. 3769-3786
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5177

To provide a concise yet insightful synthesis of these findings, Table 4 presents the Top-3 model–resampling combinations ranked by their Class 1 (hoax) F1-scores. This targeted summary distills the most effective configurations from the broader experimental results, enabling a clearer understanding of which approaches achieve the optimal balance between minority-class detection and overall classification robustness. By highlighting the highest-performing pairings, the table offers practical guidance for both researchers and practitioners in selecting strategies that not only enhance minority-class recognition but also preserve high predictive reliability across all classes in imbalanced fake news detection scenarios.

Table 4. Top-3 Model–Resampling Combinations Ranked by Class 1 F1-Score

Rank	Model – Resampling Combination	F1-Score (Class 1)	F1-Score (Class 0)
1	Logistic Regression + SMOTE	0.6682	0.8892
2	Logistic Regression + RandomOversampling	0.6662	0.8829
3	SVM + RandomOversampling	0.6627	0.9063

As shown in Table 4, all three top-performing combinations share two defining characteristics: the use of oversampling strategies (SMOTE or RandomOversampling) and the application of linear or margin-based classifiers (Logistic Regression and SVM). These methods are inherently well-suited for imbalanced text classification tasks, as they can form stable decision boundaries even when the minority class is artificially expanded through synthetic instances.

Logistic Regression + SMOTE secures the highest minority-class F1-score (0.6682), benefiting from SMOTE's ability to generate representative synthetic samples that expand the decision space for the hoax class, while maintaining a strong majority-class F1-score (0.8892). Logistic Regression + RandomOversampling follows closely, demonstrating that even a simpler oversampling method can yield competitive performance when paired with a robust linear classifier. SVM + RandomOversampling ranks third, combining the enriched minority-class representation with SVM's high discriminative power, and achieving the highest majority-class F1-score (0.9063) among the top combinations—indicating superior generalization to legitimate news instances.

From a practical perspective, these results emphasize that combining oversampling techniques with linear or margin-based classifiers provides a reliable and computationally efficient approach for hoax detection in highly imbalanced datasets. Such configurations not only elevate minority-class recognition but also preserve robust majority-class accuracy, offering a well-balanced and interpretable classification pipeline. Furthermore, their stability across cross-validation folds suggests consistent performance in real-world scenarios where data distributions may shift—making them technically sound and operationally viable for deployment.

3.4. Discussion

Building upon the research gap outlined in the Introduction, this section situates the present findings within the context of relevant prior studies, highlighting both methodological parallels and key advancements. Unlike earlier works by Hossain et al. [9], Yao et al. [10], Budhi et al. [11], Khan et al. [12], and Elsaeed et al. [13], which either applied resampling in different domains or did not systematically address class imbalance in hoax detection, this study provides a domain-specific, cross-validated evaluation of seven resampling strategies across three classical models, yielding 21 unique model–resampling combinations. The integration of TF-IDF vectorization, stratified cross-validation,

Vol. 6, No. 5, October 2025, Page. 3769-3786
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5177

and comprehensive metric aggregation represents a methodological contribution aimed at enhancing reproducibility and comparability in future research.

The experimental results consistently highlight the critical role of resampling in mitigating class imbalance. SMOTE and RandomOversampling emerged as the most effective techniques, improving minority-class (Class 1) F1-scores without substantially compromising majority-class (Class 0) performance. Logistic Regression and SVM, when paired with these oversampling methods, delivered the most stable and competitive outcomes—an observation aligned with the findings of Hossain et al. [9], who reported improved recall in Bangla fake news detection using SMOTE, although their work employed model stacking. Similarly, Yao et al. [10] demonstrated that balanced datasets enhance ensemble-based detection of fake online reviews, while our results confirm this advantage in the fake news domain using simpler, interpretable models.

In contrast, undersampling methods such as NearMiss and RandomUndersampling produced the lowest Class 1 F1-scores, particularly when combined with ensemble models like Random Forest, reinforcing prior concerns from Budhi et al. [11] about information loss in aggressive undersampling. Khan et al. [12] benchmarked various classifiers for fake news detection but did not employ imbalance-handling strategies, which may explain why their minority-class metrics lag behind our oversampling-based configurations. Elsaeed et al. [13] proposed a high-accuracy voting classifier but similarly omitted explicit resampling, raising concerns of inflated majority-class performance—an issue our approach directly addresses.

From a practical standpoint, the strong performance of Logistic Regression + SMOTE, Logistic Regression + RandomOversampling, and SVM + RandomOversampling suggests that combining oversampling techniques with linear or margin-based classifiers offers a reliable and computationally efficient solution for real-world hoax detection systems. Such configurations are not only robust across cross-validation folds but also maintain interpretability, making them well-suited for deployment in fact-checking tools or early-warning misinformation platforms.

Nevertheless, this study has limitations. The GossipCop dataset focuses exclusively on entertainment news, potentially limiting generalizability to other domains such as political, financial, or health-related misinformation. Additionally, the evaluation is confined to classical machine learning models. Future work should investigate hybrid pipelines incorporating transformer-based architectures (e.g., BERT, RoBERTa) with resampling strategies, assess robustness against adversarial inputs, and explore real-time deployment scenarios.

In conclusion, by systematically comparing 21 model–resampling configurations within a unified pipeline, this research extends prior literature by delivering empirical evidence that synthetic oversampling—particularly SMOTE and RandomOversampling—can substantially improve minority-class detection while maintaining strong overall accuracy. These findings reinforce the value of data-centric preprocessing in imbalanced text classification and provide a tested framework for advancing fake news detection in diverse application domains.

4. CONCLUSION

SMOTE combined with Logistic Regression achieves the most balanced and consistently high performance for minority-class detection in hoax classification, demonstrating its effectiveness as a practical and interpretable solution for real-world fact-checking systems. The study contributes a reproducible evaluation framework that integrates TF-IDF vectorization, seven resampling strategies, and three classical classifiers, enabling systematic assessment of class imbalance handling in textual classification tasks. These results underscore the critical role of tailored oversampling in improving recall without sacrificing majority-class accuracy, making it well-suited for deployment in constrained environments such as misinformation early-warning systems. While the analysis is limited to classical

Vol. 6, No. 5, October 2025, Page. 3769-3786
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5177

machine learning models and an entertainment-focused dataset, extending this approach to transformer-based architectures and diverse misinformation domains (e.g., political or health-related) could enhance generalizability, semantic representation, and scalability.

ACKNOWLEDGEMENT

We want to express our sincere gratitude to the Indonesian Ministry of Research, Technology, and Higher Education for its generous financial support and the research facilities provided. This assistance has been invaluable in the completion of this study. We also greatly appreciate the faculty members and research staff for their invaluable technical guidance and constructive input, which have significantly contributed to the improvement of the quality of this research.

REFERENCES

[1] D. H. Lan and T. M. Tung, "Exploring fake news awareness and trust in the age of social media among university student tiktok users," *Cogent Soc. Sci.*, vol. 10, no. 1, Dec. 2024, doi: 10.1080/23311886.2024.2302216.

[2] M. A. Alonso, D. Vilares, C. Gómez-Rodríguez, and J. Vilares, "Sentiment analysis for fake news detection," *Electronics*, vol. 10, no. 11, p. 1348, Jun. 2021, doi: 10.3390/electronics10111348.

[3] B. Collins, D. T. Hoang, N. T. Nguyen, and D. Hwang, "Trends in combating fake news on social media – a survey," *J. Inf. Telecommun.*, vol. 5, no. 2, pp. 247–266, Apr. 2021, doi: 10.1080/24751839.2020.1847379.

[4] S. Mishra, P. Shukla, and R. Agarwal, "Analyzing machine learning enabled fake news detection techniques for diversified datasets," *Wirel. Commun. Mob. Comput.*, vol. 2022, pp. 1–18, Mar. 2022, doi: 10.1155/2022/1575365.

[5] M. F. Mridha, A. J. Keya, Md. A. Hamid, M. M. Monowar, and Md. S. Rahman, "A comprehensive review on fake news detection with deep learning," *IEEE Access*, vol. 9, pp. 156151–156170, 2021, doi: 10.1109/access.2021.3129329.

[6] S. K. Hamed, M. J. Ab Aziz, and M. R. Yaakub, "A review of fake news detection approaches: a critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion," *Heliyon*, vol. 9, no. 10, p. e20382, Oct. 2023, doi: 10.1016/j.heliyon.2023.e20382.

[7] F. Gulzar Hussain, M. Wasim, S. Hameed, A. Rehman, M. Nabeel Asim, and A. Dengel, "Fake news detection landscape: datasets, data modalities, ai approaches, their challenges, and future perspectives," *IEEE Access*, vol. 13, pp. 54757–54778, 2025, doi: 10.1109/access.2025.3553909.

[8] Q. Li, C. Zhao, X. He, K. Chen, and R. Wang, "The impact of partial balance of imbalanced dataset on classification performance," *Electronics*, vol. 11, no. 9, p. 1322, Apr. 2022, doi: 10.3390/electronics11091322.

[9] M. M. Hossain, Z. Awosaf, M. S. H. Prottoy, A. S. M. Alvy, and M. K. Morol, "Approaches for improving the performance of fake news detection in bangla: imbalance handling and model stacking," Mar. 22, 2022, *arXiv*: arXiv:2203.11486. doi: 10.48550/arXiv.2203.11486.

[10] J. Yao, Y. Zheng, and H. Jiang, "An ensemble model for fake online review detection based on data resampling, feature pruning, and parameter optimization," *IEEE Access*, vol. 9, pp. 16914–16927, 2021, doi: 10.1109/access.2021.3051174.

[11] G. S. Budhi, R. Chiong, and Z. Wang, "Resampling imbalanced data to detect fake reviews using machine learning classifiers and textual-based features," *Multimed. Tools Appl.*, vol. 80, no. 9, pp. 13079–13097, Apr. 2021, doi: 10.1007/s11042-020-10299-5.

[12] J. Y. Khan, Md. T. I. Khondaker, S. Afroz, G. Uddin, and A. Iqbal, "A benchmark study of machine learning models for online fake news detection," *Mach. Learn. Appl.*, vol. 4, p. 100032, Jun. 2021, doi: 10.1016/j.mlwa.2021.100032.

Vol. 6, No. 5, October 2025, Page. 3769-3786
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5177

[13] E. Elsaeed, O. Ouda, M. M. Elmogy, A. Atwan, and E. El-Daydamony, "Detecting fake news in social media using voting classifier," *IEEE Access*, vol. 9, pp. 161909–161925, 2021, doi: 10.1109/access.2021.3132022.

[14] M. S. Kraiem, F. Sánchez-Hernández, and M. N. Moreno-García, "Selecting the suitable resampling strategy for imbalanced data classification regarding dataset properties. an approach based on association models," *Appl. Sci.*, vol. 11, no. 18, p. 8546, Sep. 2021, doi: 10.3390/app11188546.

[15] D. Z. Abidin, M. Rosario, and A. Sadikin, "Improving term deposit customer prediction using support vector machine with smote and hyperparameter tuning in bank marketing campaigns," vol. 6, no. 3, 2025, doi: doi.org/10.52436/1.jutif.2025.6.3.4585.

[16] M. Khushi *et al.*, "A comparative performance analysis of data resampling methods on imbalance medical data," *IEEE Access*, vol. 9, pp. 109960–109975, 2021, doi: 10.1109/access.2021.3102399.

[17] G. S. Budhi, R. Chiong, and Z. Wang, "Resampling imbalanced data to detect fake reviews using machine learning classifiers and textual-based features," *Multimed. Tools Appl.*, vol. 80, no. 9, pp. 13079–13097, Apr. 2021, doi: 10.1007/s11042-020-10299-5.

[18] E. Richardson, R. Trevizani, J. A. Greenbaum, H. Carter, M. Nielsen, and B. Peters, "The receiver operating characteristic curve accurately assesses imbalanced datasets," *Patterns*, vol. 5, no. 6, p. 100994, Jun. 2024, doi: 10.1016/j.patter.2024.100994.

[19] X. Chao, G. Kou, Y. Peng, and A. Fernández, "An efficiency curve for evaluating imbalanced classifiers considering intrinsic data characteristics: experimental analysis," *Inf. Sci.*, vol. 608, pp. 1131–1156, Aug. 2022, doi: 10.1016/j.ins.2022.06.045.

[20] C.-M. Lai, M.-H. Chen, E. Kristiani, V. K. Verma, and C.-T. Yang, "Fake news classification based on content level features," *Appl. Sci.*, vol. 12, no. 3, p. 1116, Jan. 2022, doi: 10.3390/app12031116.

[21] S. Farhadpour, T. A. Warner, and A. E. Maxwell, "Selecting and interpreting multiclass loss and accuracy assessment metrics for classifications with class imbalance: guidance and best practices", doi: doi.org/ 10.3390/rs16030533.

[22] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: a data repository with news content, social context and spatialtemporal information for studying fake news on social media," Mar. 27, 2019, *arXiv*: arXiv:1809.01286. doi: 10.48550/arXiv.1809.01286.

[23] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: a data mining perspective," Sep. 03, 2017, *arXiv*: arXiv:1708.01967. doi: 10.48550/arXiv.1708.01967.

[24] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, "Review of classification methods on unbalanced data sets," *IEEE Access*, vol. 9, pp. 64606–64628, 2021, doi: 10.1109/access.2021.3074243.

[25] S. Rawat, A. Rawat, D. Kumar, and A. S. Sabitha, "Application of machine learning and data visualization techniques for decision support in the insurance sector," *Int. J. Inf. Manag. Data Insights*, vol. 1, no. 2, p. 100012, Nov. 2021, doi: 10.1016/j.jjimei.2021.100012.

[26] F. Olan, U. Jayawickrama, E. O. Arakpogun, J. Suklan, and S. Liu, "Fake news on social media: the impact on society," *Inf. Syst. Front.*, vol. 26, no. 2, pp. 443–458, Apr. 2024, doi: 10.1007/s10796-022-10242-z.

[27] S. Hakak, M. Alazab, S. Khan, T. R. Gadekallu, P. K. R. Maddikunta, and W. Z. Khan, "An ensemble machine learning approach through effective feature extraction to classify fake news," *Future Gener. Comput. Syst.*, vol. 117, pp. 47–58, Apr. 2021, doi: 10.1016/j.future.2020.11.022.

[28] A. M. Elmogy, U. Tariq, A. Ibrahim, and A. Mohammed, "Fake reviews detection using supervised machine learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 1, 2021.

[29] M. Z. Naeem, F. Rustam, A. Mehmood, Mui-zzud-din, I. Ashraf, and G. S. Choi, "Classification of movie reviews using term frequency-inverse document frequency and optimized machine learning algorithms," *PeerJ Comput. Sci.*, vol. 8, p. e914, Mar. 2022, doi: 10.7717/peerj-cs.914.

[30] R. M. Pereira, Y. M. G. Costa, and C. N. Silla Jr., "Toward hierarchical classification of imbalanced data using random resampling algorithms," *Inf. Sci.*, vol. 578, pp. 344–363, Nov. 2021, doi: 10.1016/j.ins.2021.07.033.

Vol. 6, No. 5, October 2025, Page. 3769-3786
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5177

[31] M. Imani, A. Beikmohammadi, and H. R. Arabnia, "Comprehensive analysis of random forest and xgboost performance with smote, adasyn, and gnus under varying imbalance levels," *Technologies*, vol. 13, no. 3, p. 88, Feb. 2025, doi: 10.3390/technologies13030088.

[32] M. Altalhan, A. Algarni, and M. Turki-Hadj Alouane, "Imbalanced data problem in machine learning: a review," *IEEE Access*, vol. 13, pp. 13686–13699, 2025, doi: 10.1109/access.2025.3531662.

[33] M. S. Ebrahimi Shahabadi, H. Tabrizchi, M. Kuchaki Rafsanjani, B. B. Gupta, and F. Palmieri, "A combination of clustering-based under-sampling with ensemble methods for solving imbalanced class problem in intelligent systems," *Technol. Forecast. Soc. Change*, vol. 169, p. 120796, Aug. 2021, doi: 10.1016/j.techfore.2021.120796.

[34] A. Mahabub, "A robust technique of fake news detection using ensemble voting classifier and comparison with other classifiers," *SN Appl. Sci.*, vol. 2, no. 4, Apr. 2020, doi: 10.1007/s42452-020-2326-y.

[35] M. Thanh Vo, A. H. Vo, T. Nguyen, R. Sharma, and T. Le, "Dealing with the class imbalance problem in the detection of fake job descriptions," *Comput. Mater. Contin.*, vol. 68, no. 1, pp. 521–535, 2021, doi: 10.32604/cmc.2021.015645.

[36] M. Carvalho, A. J. Pinho, and S. Brás, "Resampling approaches to handle class imbalance: a review from a data perspective," *J. Big Data*, vol. 12, no. 1, Mar. 2025, doi: 10.1186/s40537-025-01119-4.

[37] I. Ahmad, M. Yousaf, S. Yousaf, and M. O. Ahmad, "Fake news detection using machine learning ensemble methods," *Complexity*, vol. 2020, pp. 1–11, Oct. 2020, doi: 10.1155/2020/8885861.

[38] S. Kaur, P. Kumar, and P. Kumaraguru, "Automating fake news detection system using multi-level voting model," *Soft Comput.*, vol. 24, no. 12, pp. 9049–9069, Jun. 2020, doi: 10.1007/s00500-019-04436-y.

[39] T. Jiang, J. P. Li, A. U. Haq, A. Saboor, and A. Ali, "A novel stacking approach for accurate detection of fake news," *IEEE Access*, vol. 9, pp. 22626–22639, 2021, doi: 10.1109/access.2021.3056079.

[40] C. A. Ramezan, T. A. Warner, and A. E. Maxwell, "Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification," *Remote Sens.*, vol. 11, no. 2, p. 185, Jan. 2019, doi: 10.3390/rs11020185.

[41] W. H. Bangyal *et al.*, "Detection of fake news text classification on covid-19 using deep learning approaches," *Comput. Math. Methods Med.*, vol. 2021, pp. 1–14, Nov. 2021, doi: 10.1155/2021/5514220.

[42] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, "Evaluating the quality of machine learning explanations: a survey on methods and metrics," *Electronics*, vol. 10, no. 5, p. 593, Mar. 2021, doi: 10.3390/electronics10050593.

[43] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, "Evaluating trust prediction and confusion matrix measures for web services ranking," *IEEE Access*, vol. 8, pp. 90847–90861, 2020, doi: 10.1109/access.2020.2994222.

[44] M. N. Razali, S. A. Manaf, R. B. Hanapi, M. R. Salji, L. W. Chiat, and K. Nisar, "Enhancing minority sentiment classification in gastronomy tourism: a hybrid sentiment analysis framework with data augmentation, feature engineering and business intelligence," *IEEE Access*, vol. 12, pp. 49387–49407, 2024, doi: 10.1109/access.2024.3362730.

[45] Q. Li *et al.*, "A survey on text classification: from traditional to deep learning," *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 2, pp. 1–41, Apr. 2022, doi: 10.1145/3495162.