Vol. 8, No. 4, August 2026

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Comparative Analysis of Four Machine Learning Classifiers for Indonesian Hoax News Detection

Dedi Irawan[1], Sudarmaji[2]

[1,2]Faculty of Computer Science, Universitas Muhammadiyah Metro, Metro, Lampung, Indonesia

Received:
February 28, 2026
Revised:
June 18, 2026
Accepted:
July 22, 2026
Published:
August 22, 2026

Corresponding Author:
Dedi Irawan*:
Dedi Irawan
Email*:
dedimti@ummetro.ac.id

DOI:
10.63158/journalisi.v8i4.1710

Abstract. Across Indonesian online platforms, fabricated news spreads faster than fact-checkers can confirm. Because much of the literature relies on resource-intensive deep models, one applied question stays unsettled: which lighter, more transparent classifier best detects Indonesian hoaxes? We assessed four algorithms, Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB), and Extreme Gradient Boosting (XGBoost), on 1,116 Indonesian articles from the MAFINDO/TurnBackHoax repository, using one shared preprocessing pipeline and two feature schemes, Bag-of-Words (BoW) and TF-IDF, under an 80:20 stratified split. On the held-out test set, Random Forest with BoW performed best at 98.66% accuracy and 98.59% macro F1-score, missing three of 224 cases, with XGBoost next at 98.21%. Under repeated five-fold cross-validation, however, XGBoost with BoW attained a significantly higher mean (98.36% versus 97.44% macro F1), so the two ensembles are best regarded as closely competitive rather than decisively ranked. BoW also beat TF-IDF for every classifier, most sharply for Naïve Bayes (84.70% versus 70.19% F1). Because the study uses lexical features alone on a single dataset, the findings indicate that classical ensembles can rival reported deep-learning figures at far lower cost, rather than forming a universal conclusion; broader validation across other sources and periods is still needed.

Keywords: hoax detection; machine learning; Indonesian language; Random Forest; Bag-of-Words

1. INTRODUCTION

With one of the largest and most online populations on earth, Indonesia enjoys broad digital connectivity that, in turn, has created fertile ground for a continual flow of hoaxes and false information. Spreading across news sites and social platforms, invented stories are capable of dividing public opinion, triggering widespread alarm, and gradually undermining social order [1]. To push back, the Indonesian Anti-Hoax Society (MAFINDO) runs the fact-checking service TurnBackHoax.id, on which submitted hoaxes are checked and recorded by hand. Such manual review, though, cannot match the daily torrent of new material a gap that turns automated detection from a convenience into something genuinely required [2], [3]. The present study scopes this problem deliberately: it addresses repository-based hoax detection over curated, article-length fact-check entries, rather than the broader and harder problem of detecting every form of misinformation as it circulates live across social media.

A scalable answer to this bottleneck comes from machine learning, and a solid body of work on Indonesian hoax detection has built up over time. Early studies demonstrated that a plain Naïve Bayes model could identify Indonesian hoax pieces from lexical signals by themselves [4], a result that larger hoax collections subsequently confirmed [5]. The field has more recently pivoted to deep learning: one transformer experiment using material harvested from TurnBackHoax.id achieved as much as 90% accuracy with BERT, surpassing convolutional and recurrent comparators including CNN, BiLSTM, and combined CNN–BiLSTM architectures [6], while tuned IndoBERT variants have been carried over to hoax and sentiment problems on a range of Indonesian media and social networks [7], [8]. Read collectively, this research underscores two points detection can be automated, and which model is chosen carries substantial weight [9]. The same family of text-classification techniques has also proven adaptable to adjacent Indonesian tasks such as sentiment analysis of social-media and election discourse [10], [11], [12], which share the short, noisy, lexically driven character of hoax text.

Even so, two shortcomings remain. The first concerns focus: much of the latest research favors deep learning and transformer architectures which, however accurate, require intensive computing, sizeable labeled datasets, and yield limited interpretability conditions that sit uneasily with the constrained environments where many Indonesian

organizations work [9], [13]. For fact-checking teams, community newsrooms, and educational institutions that operate without dedicated GPUs, lightweight classical models that train in seconds on commodity hardware remain not merely a fallback but often the only deployable option. Interpretability matters here for a further reason: in hoax detection an automated flag must be defensible to a human reviewer, and models whose decisions trace back to explicit lexical cues are far easier to audit and trust than opaque neural representations. The second concerns comparison. Conventional classifiers have largely been examined in isolation, leaving structured side-by-side assessments of multiple lightweight, interpretable models on one uniformly labeled Indonesian hoax-news corpus, sharing the same preprocessing and spanning more than a single feature representation, in short supply [14], [15], [16]. As a result, practitioners lack solid, data-backed direction on which traditional classifier and which feature scheme together achieve the most favorable accuracy–efficiency trade-off for this language and problem.

The present work tackles this gap head-on, benchmarking four commonly applied classifiers Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB), and Extreme Gradient Boosting (XGBoost) on the two-class task of separating Indonesian hoax from non-hoax news. The intended beneficiaries are concrete: fact-checking teams such as MAFINDO, online news platforms moderating user-submitted content, and educational institutions building media-literacy tools, all of whom need a fast, transparent first-pass screen rather than a heavyweight research model. Using an openly available set of hoax and genuine articles from the MAFINDO/TurnBackHoax repository, each model is run through one identical preprocessing pipeline and evaluated under two rival feature schemes, Bag-of-Words (BoW) and Term Frequency Inverse Document Frequency (TF-IDF), keeping the contest even. Performance is judged jointly by accuracy, precision, recall, and F1-score.

There are three contributions. First, the study supplies a controlled and repeatable benchmark of four classical classifiers on an Indonesian hoax-news collection under a single shared pipeline and two feature representations. Second, it reveals an unanticipated empirical pattern in this corpus the plainer Bag-of-Words scheme reliably outdoes TF-IDF, overturning the customary assumption and bearing directly on how features should be chosen. Third, it translates these observations into practical guidance

on selecting a classifier for first-pass screening of Indonesian hoax articles, rather than for final fact-checking judgment, while remaining frank about the constraints that come with strictly lexical features and with a single curated repository. The rest of the article is organized thus: Section 2 describes the dataset, preprocessing, feature extraction, classification models, and evaluation metrics; Section 3 reports and interprets the outcomes; and Section 4 concludes with the study's findings, implications, limitations, and future research avenues.

2. METHODS

A quantitative, experiment-driven design underpins this work, set up to weigh the performance of four machine learning classifiers for detecting Indonesian hoax news. The complete procedure unfolds in five ordered phases: data collection, text preprocessing, feature extraction, model training and classification, and performance evaluation, as illustrated in Figure 1.

Figure 1. Research Framework of the Proposed Comparative Study

2.1. Data Collection

The data analyzed here comprise Indonesian hoax and genuine (non-hoax) news entries sourced from the MAFINDO/TurnBackHoax repository, an open collection kept by the Indonesian Anti-Hoax Society [17]. Every entry holds a headline and the body text of an article alongside a two-value label, in which 0 marks a hoax (an invented claim) and 1 marks a legitimate item (an official clarification or rebuttal of a circulating claim). Before modeling, the labeling direction was confirmed by manually reviewing sample records from both classes. In total the corpus holds 1,116 entries, made up of 683 hoaxes (61.2%) and 433 legitimate items (38.8%), pointing to a moderately skewed class balance that the evaluation takes into account. Table 1 summarizes this distribution. It is worth stressing at the outset that the legitimate class is not made up of ordinary neutral news reporting but of official clarification and rebuttal texts that explicitly debunk a circulating claim; this characteristic shapes the vocabulary of the class and is revisited when the results

are interpreted. Prior to splitting, the corpus was screened for exact duplicate documents, of which only four were found. Because exact matching alone would miss articles that are reworded or partially overlapping, a similarity-based near-duplicate check was also carried out: pairwise cosine similarity was computed over TF-IDF representations of all 1,116 processed documents, yielding five pairs above a similarity of 0.95 and eleven above 0.80. With so few overlapping items in a corpus of this size, no documents were removed, and the residual risk of near-duplicate leakage across the split is negligible. Applying an 80:20 stratified split then yielded a training set of 892 entries (546 hoax, 346 legitimate) and a held-out test set of 224 entries (137 hoax, 87 legitimate), preserving the original class proportions in both partitions.

Table 1. Class Distribution of the Dataset

Class	Number of Items	Proportion (%)
Hoax	683	61.2
Non-Hoax (Valid)	433	38.8
Total	**1,116**	**100.0**

2.2. Text Preprocessing

Because unprocessed Indonesian articles carry noise that can hurt classification, a chain of preprocessing operations was run to regularize the text ahead of feature extraction. These operations ran in this order: (1) case folding, lowering every character; (2) cleaning, stripping out URLs, mentions, hashtags, digits, punctuation, and non-letter symbols; (3) tokenization, breaking the text into word tokens; (4) stopword removal, dropping frequent Indonesian function words via the Sastrawi stopword list and removing tokens of three characters or fewer; and (5) stemming, collapsing each token to its base form with the Sastrawi Indonesian stemmer. To bound the cost of dictionary-based Sastrawi stemming, which scales with document length, the truncation was applied to the article body only: each body was cut to its opening 200 words and the full, untruncated headline was then prepended to it, so that the two fields were concatenated into a single text field in which no headline text is ever discarded. This choice is supported by the structure of news writing, where the lead paragraphs typically carry the core claim and its framing; because the average processed document still contained 158.8 tokens, well below the cap, the truncation left most documents untouched and is not expected to materially affect

classification, though its effect on the minority of longer articles is acknowledged among the study limitations. These token streams became the input to the feature-extraction step.

2.3. Feature Extraction

Turning the preprocessed text into numeric feature vectors involved building and contrasting two encoding schemes: Bag-of-Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF). Under Bag-of-Words, a document becomes a vector of plain term counts that ignores word order and treats every term as equally important; representations of this vector-space kind are a cornerstone of text mining [18]. TF-IDF builds on it by scaling each term by both how often it appears in a document and how uncommon it is over the whole corpus, so that common yet uninformative terms are discounted. Equation 1 gives the TF-IDF weight of term t in document d:

$$w(t,d) = tf(t,d) \times idf(t) \quad (1)$$

where tf(t,d) is the frequency of term t in document d, and idf(t) is defined in Equation 2:

$$idf(t) = log\ (\ N\ /\ df(t)\ ) \quad (2)$$

in which N stands for the overall document count and df(t) is the count of documents that contain term t. To keep the comparison fair, both encodings were produced under the same capped vocabulary of 5,000 features. This ceiling was set deliberately: retaining the 5,000 most frequent terms covers the great majority of discriminative vocabulary in a corpus of this size while trimming the long tail of rare tokens that mostly add sparsity and noise, and it keeps the feature matrices compact enough to train every classifier quickly on commodity hardware, consistent with the lightweight, resource-aware aim of the study.

2.4. Classification Models

Four supervised classifiers were trained and tested under the same settings. Random Forest (RF) combines decision trees grown on bootstrap samples and merges their outputs by majority vote, which curbs variance and overfitting [19]. Support Vector Machine (SVM) looks for the best separating hyperplane that widens the margin between

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

classes and copes well with sparse, high-dimensional text features [20]. In its multinomial variant, Naïve Bayes (NB) uses Bayes' theorem together with a conditional independence assumption among features and acts as a quick baseline. Extreme Gradient Boosting (XGBoost) is a regularized boosting ensemble that grows trees one after another to fix earlier mistakes [21] and stays robust under class imbalance [22]. An 80:20 stratified split divided the data into training and test portions while keeping the class proportions intact. To prevent any information leakage, both the BoW and TF-IDF vectorizers were fitted exclusively on the training partition and then used to transform the held-out test set, so that no test-set vocabulary or document-frequency statistics informed feature construction. Table 2 lists the principal hyperparameters. Every experiment was coded in Python 3.12 with scikit-learn 1.8 [23], XGBoost 3.3 [21], and Sastrawi 1.0.1 for Indonesian stemming, all under fixed random seeds and matching conditions so that the results can be reproduced exactly.

Table 2. Hyperparameter Configuration for Each Classifier

Classifier	Key Hyperparameters
Random Forest	n_estimators = 200; random_state = 42
SVM	linear kernel; C = 1.0
Naïve Bayes	multinomial; alpha = 1.0
XGBoost	n_estimators = 200; learning_rate = 0.1; max_depth = 6

2.5. Evaluation Metrics

Four conventional metrics taken from the confusion matrix accuracy, precision, recall, and F1-score, all in common use for assessing machine learning classifiers [24] gauged model quality and are defined in Equations 3 to 6. Given the moderate class imbalance, precision, recall, and F1-score were obtained through macro-averaging so that each class counted equally.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (3)$$

$$Precision = TP / (TP + FP) \quad (4)$$

$$Recall = TP / (TP + FN) \quad (5)$$

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

here TP, TN, FP, and FN stand respectively for true positives, true negatives, false positives, and false negatives. The F1-score received special emphasis because it reconciles precision with recall and is a steadier gauge than accuracy by itself when classes are unevenly distributed.

3. RESULTS AND DISCUSSION

3.1. Comparative Classifier Performance

Every one of the four classifiers was fitted and assessed with both the Bag-of-Words and TF-IDF representations under the same 80:20 stratified split, giving eight model–feature pairings in all. Table 3 sets out the accuracy, precision, recall, and macro F1-score for each pairing.

Table 3. Performance Comparison of the Four Classifiers under BoW and TF-IDF

Classifier	Features	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	**BoW**	**98.66**	**98.69**	**98.49**	**98.59**
Random Forest	TF-IDF	98.66	98.93	98.28	98.58
XGBoost	BoW	98.21	98.12	98.12	98.12
XGBoost	TF-IDF	97.77	97.75	97.55	97.65
SVM	BoW	93.30	94.29	91.80	92.77
SVM	TF-IDF	90.18	90.17	89.03	89.52
Naïve Bayes	BoW	85.27	84.36	85.23	84.70
Naïve Bayes	TF-IDF	75.89	82.37	69.59	70.19

Table 3 reports the ranking on the held-out test set. The best performer on this split is Random Forest with Bag-of-Words at 98.66% accuracy and a macro F1-score of 98.59%, narrowly ahead of both its TF-IDF counterpart and XGBoost (98.21% accuracy); as Section 3.3 shows, this narrow margin does not persist once performance is averaged over repeated folds. At the bottom is Naïve Bayes paired with TF-IDF, lagging the others at

75.89% accuracy and 70.19% F1. The contrast is easier to grasp in Figure 2, which lines up the F1-scores of all classifier–feature combinations together.

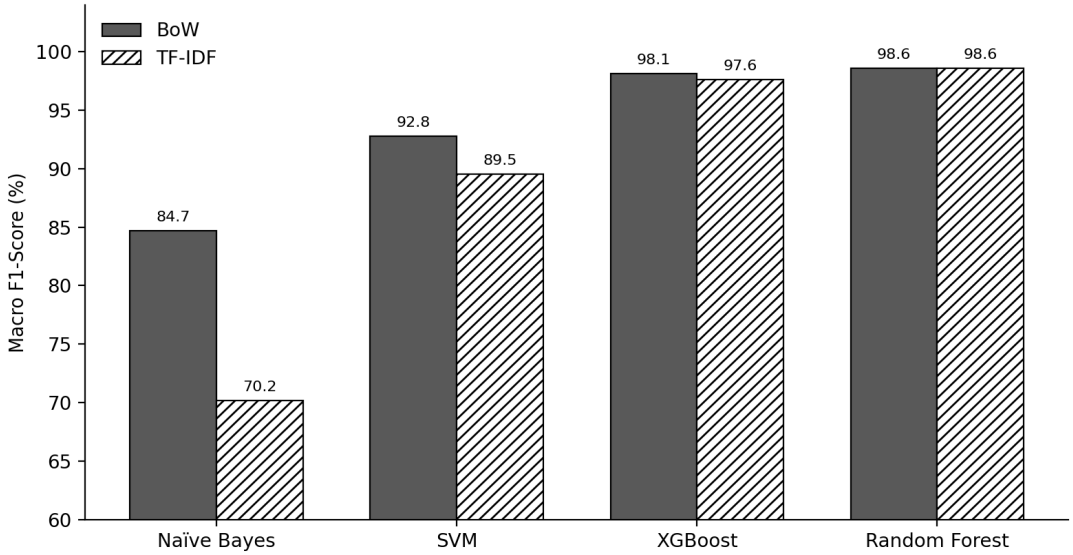

Figure 2. F1-Score Comparison across Classifiers and Feature Representations

3.2. Confusion Matrix of the Best Model

Figure 3 shows exactly where the front-runner, Random Forest with Bag-of-Words, gets things right and wrong. Out of the 224 test cases it tagged 136 of 137 hoaxes and 85 of 87 legitimate articles correctly, producing only three mistakes a single false positive and two false negatives. Its class-wise F1-scores, 0.989 for hoaxes and 0.983 for genuine news, are nearly the same a meaningful detail, since it shows the model performing on the minority class as strongly as on the majority class in spite of the moderate imbalance present.

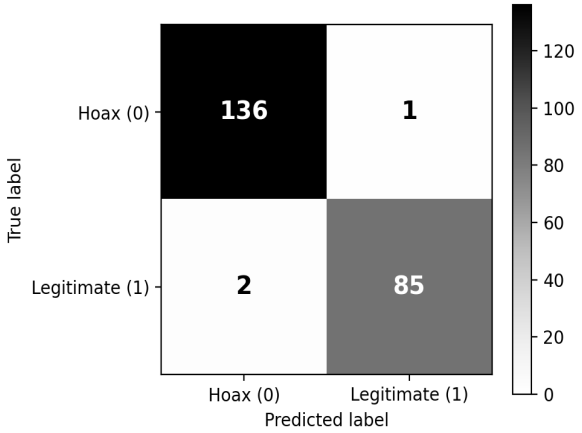

Figure 3. Confusion Matrix of the Random Forest (BoW) Classifier

3.3. Stability under Cross-Validation

Because a single 80:20 split can flatter or penalize a model by chance, the experiment was repeated under stratified five-fold cross-validation run three times with different shuffles, giving fifteen train-test folds per configuration. Table 4 reports the mean and standard deviation of accuracy and macro F1-score over these folds. The standard deviations are small, confirming that the leading scores are stable rather than an artifact of one fortunate partition, but the ordering of the two leading models does not carry over unchanged from the held-out split. Under repeated cross-validation XGBoost with BoW attained the highest mean macro F1 at 98.36% (SD 1.20), ahead of Random Forest with BoW at 97.44% (SD 1.25); the same reversal appears in accuracy (98.45% against 97.58%). Because both models were evaluated on identical folds, the two can be compared directly: XGBoost outscored Random Forest in twelve of the fifteen folds, with a mean macro F1 advantage of 0.92 percentage points (95% confidence interval 0.31 to 1.54), and the difference is statistically significant under both a paired t-test ($t = 3.23$, $p = 0.006$) and a Wilcoxon signed-rank test ($p = 0.006$), with a large effect size (Cohen's $d = 0.83$). The evidence therefore does not support treating either ensemble as the single best model across all evaluation settings: Random Forest with BoW is the best performer on the held-out test set, where it misclassified three cases against XGBoost's four, whereas XGBoost with BoW is the stronger performer on average across repeated folds. A one-case difference on a 224-document test set is well within the fold-to-fold variation reported here, which suggests the held-out ordering reflects the particular partition rather than a durable advantage. The two tree ensembles are best described as closely competitive, and both clearly outperform SVM and Naïve Bayes, the latter remaining the weakest and most variable, especially under TF-IDF (70.91% F1, SD 3.78).

Table 4. Repeated Stratified Five-Fold Cross-Validation Results (Mean ± SD)

Classifier	Features	Accuracy (%)	Macro F1 (%)
Random Forest	**BoW**	**97.58 ± 1.18**	**97.44 ± 1.25**
Random Forest	TF-IDF	97.10 ± 1.45	96.90 ± 1.58
XGBoost	BoW	98.45 ± 1.13	98.36 ± 1.20
XGBoost	TF-IDF	98.15 ± 1.31	98.04 ± 1.38
SVM	BoW	93.43 ± 1.04	92.98 ± 1.12
SVM	TF-IDF	90.11 ± 1.94	89.41 ± 2.10
Naïve Bayes	BoW	85.69 ± 2.81	84.90 ± 2.98
Naïve Bayes	TF-IDF	76.73 ± 2.37	70.91 ± 3.78

3.4. Discriminating Performance and Computational Cost

Because accuracy and F1-score summarize behaviour at a single decision threshold, the threshold-independent ranking ability of each model was also examined through the receiver operating characteristic (ROC) and precision-recall curves, treating hoax as the positive class. Figure 4 plots the ROC curves under the BoW representation, and the corresponding areas under the curve confirm the headline results: Random Forest reached an ROC-AUC of 0.9995 and a PR-AUC of 0.9996, XGBoost followed at 0.9992 and 0.9995, SVM at 0.9856 and 0.9899, and Naïve Bayes trailed at 0.9096 and 0.9090. The near-perfect AUC values for the two ensembles show that their separation of hoax from legitimate text is not confined to one operating point but holds across the full range of thresholds, which matters for screening tools whose alert sensitivity may be tuned in practice.

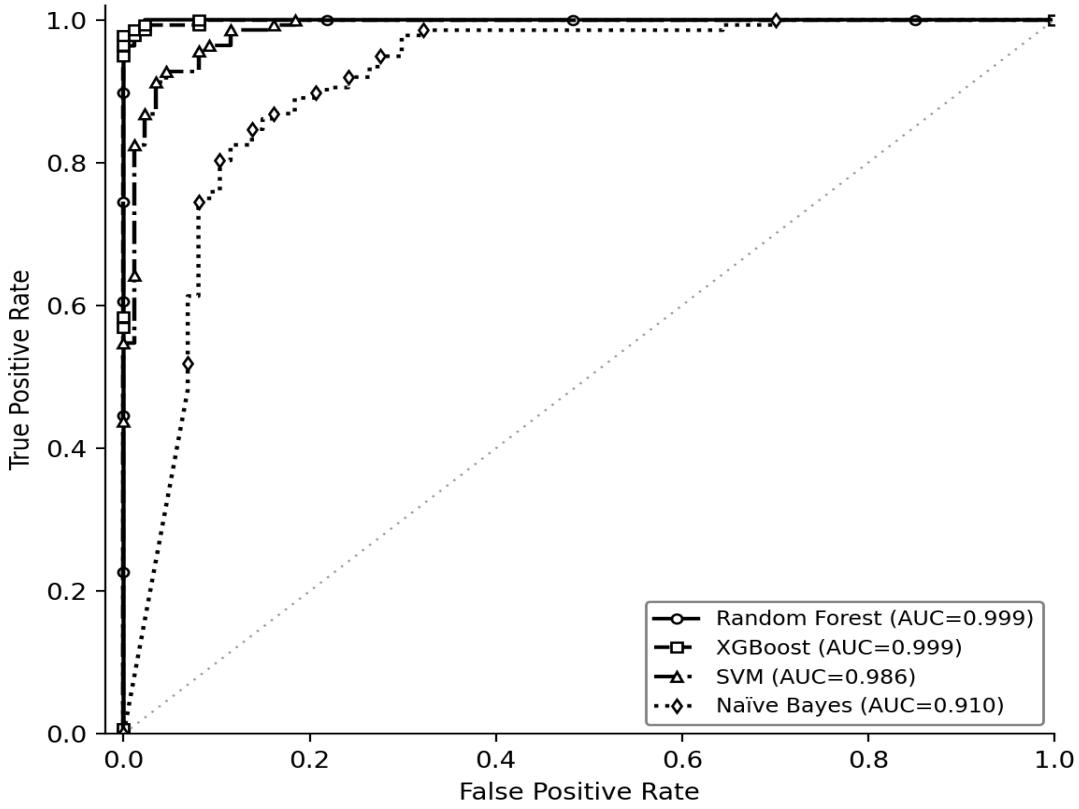

Figure 4. ROC Curves of the Four Classifiers under Bag-of-Words

The efficiency argument that motivates lightweight models was checked directly by timing training and inference on the BoW features, as reported in Table 5. All timings were obtained on a single-core Intel Xeon CPU at 2.80 GHz with 3 GB of RAM running

Ubuntu 24.04, with no GPU of any kind, a deliberately modest configuration that approximates the commodity hardware available to the resource-limited settings this study targets. Every model trained in under one second on that machine: Naïve Bayes and SVM were effectively instantaneous (about 0.003 s and 0.024 s), while the two ensembles, despite fitting 200 trees each, still trained in well under a second (Random Forest 0.860 s, XGBoost 0.640 s) and classified the entire 224-document test set in a few milliseconds. These times confirm that the accuracy of the ensembles carries no prohibitive computational penalty at this corpus scale, reinforcing their suitability as a first-pass screen on modest hardware.

Table 5. Training and Inference Time on Bag-of-Words Features

Classifier	Training Time (s)	Inference Time (s)
Random Forest	0.860	0.0324
XGBoost	0.640	0.0033
SVM	0.024	0.0004
Naïve Bayes	0.003	0.0004

To open the otherwise opaque ensemble decisions to inspection, the terms most responsible for separating the two classes were extracted, addressing the interpretability concern raised earlier. Ranking the Bag-of-Words vocabulary by Random Forest impurity-based importance surfaced function-level cues tied to the fact-checking genre; the fifteen most influential terms are listed in Table 6. Two tokens dominate by a wide margin, “kategori” (category) at 0.090 and “klarifikasi” (clarification) at 0.073, together accounting for far more importance than any topical word, and they are followed by further genre markers such as “daring” (online), “hoaks”, “fakta” (fact), and “referensi” (reference). A complementary view from the Naïve Bayes log-probability difference, which leans each term toward one class, is given in Table 7. The hoax-leaning terms are dominated by concrete topical and named-entity tokens (for example “pki”, “ahok”, “rohingya”, “klik”), whereas the legitimate-leaning terms include survey, electoral, and institutional vocabulary (for example “survei”, “capres”, “caleg”, “sanksi”). This pattern is revealing, and the magnitudes make the point sharper still: the two single most important features for Random Forest are “kategori” and “klarifikasi”, editorial labels of the fact-check format itself rather than words that carry any evidence about whether a claim is

true. A meaningful share of the model's discriminative power therefore rests on the editorial markers of debunking texts rather than on intrinsic signals of falsehood, which suggests the reported accuracy partly measures how well the classifier recognises the fact-checking genre. This is taken up directly in the discussion of dataset bias.

Table 6. Top 15 Random Forest Features by Impurity-Based Importance (Bag-of-Words)

Rank	Term (Indonesia)	English Gloss	Importance
1	kategori	category	0.09006
2	klarifikasi	clarification	0.07250
3	lengkap	complete	0.02636
4	isi	content	0.01507
5	kata	word	0.01445
6	daring	online	0.01399
7	kait	related	0.01150
8	ujar	said	0.01092
9	pihak	party	0.01045
10	hoax	hoax	0.01018
11	beri	give	0.00860
12	hoaks	hoax	0.00786
13	tanggap	respond	0.00765
14	referensi	reference	0.00760
15	fakta	fact	0.00663

Table 7. Top Class-Leaning Terms (Naïve Bayes Log-Probability Difference)

Hoax-Leaning Terms (Indonesia)	Legitimate-Leaning Terms (Indonesia)
pki, ahok, rohingya, myanmar, klik, rusuh, copy, salman, roti, toyota	survei, capres, cawapres, caleg, pks, sanksi, mahfud, lhkpn, bpn, lombok

3.5. Error Analysis of the Best Model

Inspecting the three test documents that Random Forest with BoW misclassified is informative, because all three share a single linguistic pattern. The two false negatives, legitimate items wrongly flagged as hoax, are both clarification texts that quote at length the false claim they set out to debunk, so their surface vocabulary is saturated with the

very topical and sensational terms that mark hoaxes (for instance a rebuttal concerning a public figure that repeats the disputed accusation throughout). The single false positive is the mirror image: a hoax post whose text incorporates the words “klarifikasi” and “hoaks” while itself spreading a fabricated narrative. In every error the model is misled by lexical overlap between debunking language and the claims being debunked, which is consistent with the feature-importance findings above and underlines that the classifier reasons over word occurrence rather than meaning. The errors are therefore not random but cluster around texts where the two classes are lexically entangled, a limitation intrinsic to surface-level features.

3.6. Effect of Feature Representation

A single trend ran through every result: Bag-of-Words bettered TF-IDF for each classifier without exception. With the tree ensembles the difference was trivial Random Forest shifted from 98.58% to 98.59% F1 and XGBoost from 97.65% to 98.12% yet for the linear and probabilistic models the difference was marked, SVM picking up about 3.3 F1 points and Naïve Bayes roughly 14.5 when using BoW. This ranking runs counter to received wisdom, given that TF-IDF weighting is normally expected to sharpen class separation. The most plausible reason concerns how hoax and legitimate Indonesian texts diverge: the separating cue rests in a few high-frequency words, and TF-IDF’s tendency to discount common words ends up muting precisely those signals. Naïve Bayes is hit hardest since its multinomial likelihood depends straight on raw term counts.

3.7. Discussion

The wider takeaway is that compact classical classifiers are anything but outdated for Indonesian hoax detection. Reaching 98.66% accuracy, Random Forest with BoW sits well above the early Naïve Bayes baselines on this task [4], [5]. It is tempting to set this figure against the up-to-90% accuracy reported for transformer-based BERT models [25] on TurnBackHoax data [6], but such a direct comparison would be misleading and is deliberately avoided here: the studies differ in dataset version, sample size, class composition, and train-test split, any of which can move accuracy by several points, so the numbers are not measured on common ground. The defensible claim is narrower, namely that on this corpus and under this protocol classical ensembles remain highly competitive while giving up some semantic modeling power in return for substantially lower computing demand, quicker training, and considerably more interpretability a trade

that is particularly appealing for the resource-limited institutions carrying out much of this work in Indonesia. Table 8 places the current findings next to representative earlier studies on Indonesian hoax and fake-news detection.

Table 8. Comparison with Representative Prior Studies on Indonesian Hoax Detection

Study	Method	Dataset / Source	Best Accuracy
Pratiwi et al. [4]	Naïve Bayes	250 Indonesian news articles	~83%
Darmawan et al. [26]	SVM, Random Forest	Twitter (IKN relocation hoaxes)	~88%
Awalina et al. [6]	CNN, BiLSTM, BERT	TurnBackHoax.id	~90% (BERT)
Rachmawati & Darmawan [8]	RF, LSTM, IndoBERT	Indonesian political hoax	RF highest
This study	**RF, SVM, NB, XGBoost (BoW/TF-IDF)**	**MAFINDO/TurnBackHoax (1,116)**	**98.66% (RF, BoW)**

The reason the ensembles come out on top is fairly plain. Random Forest and XGBoost are able to model non-linear interplay among lexical features and absorb noise and skewed classes, while Naïve Bayes is constrained by its premise that features act independently a premise this corpus breaks, because terms that co-occur jointly carry discriminative value. That weakness lines up with the inconsistent Naïve Bayes outcomes seen in other Indonesian hoax-detection studies [27], [28]. For anyone balancing accuracy against efficiency, two practical lessons follow: lean toward an ensemble where resources permit, and resist treating TF-IDF as the automatic choice returning to plain Bag-of-Words may well reward you.

A number of qualifications soften these results. The first is the dependence on lexical features: BoW and TF-IDF capture which words occur, not their meaning, so the models may falter on hoaxes worded much like real news, and the top-line accuracy might partly reflect corpus-specific word patterns instead of true language comprehension. The second, and most consequential, is a form of dataset bias that the feature-importance

and error analyses above make concrete: because the legitimate class consists of official clarification and rebuttal texts rather than neutral reporting, the most discriminative terms the models learn are editorial markers of the debunking genre ("klarifikasi", "kategori", "hoaks") rather than intrinsic signals of truth or falsehood. The classifier may therefore be distinguishing a fact-check from a hoax post as much as it is distinguishing true from false, which inflates separability on this corpus; the three observed errors, all caused by lexical overlap between the two genres, are the visible symptom of this.

Concretely, the legitimate class is made up mostly of official clarification and rebuttal texts rather than plain neutral reporting, and the repeated cues in such texts (terms along the lines of "clarifies" or "denies") probably broaden the separation between the classes; the reported numbers are thus most safely treated as a ceiling for this specific corpus rather than an estimate of real-world accuracy, a generalization gap repeatedly documented for token-based fake-news detectors [29]. A direct consequence is that the superiority of Bag-of-Words observed here may itself be corpus-specific: where the decisive signal lies in a handful of high-frequency genre markers, count-based features preserve them while TF-IDF down-weights them, but on a corpus without such markers the advantage could shrink or reverse. Equally, because the models key on a fixed learned vocabulary, their performance on genuinely new hoax topics that introduce unfamiliar terms, or on informal social-media text whose register differs from curated articles, is likely to degrade and remains untested. Third, every record originates from one repository (MAFINDO/TurnBackHoax) within a limited time window, which restricts how well the conclusions generalize to other sources, platforms, or fresher topics, and external testing on other Indonesian news and social-media datasets is needed before the model is trusted in operation.

A particularly informative test would be to evaluate the trained models on fresh, previously unseen articles drawn from time periods later than the collection window, since this would reveal how quickly performance decays as hoax topics and vocabulary drift. Fourth, capping article bodies at 200 words, while it made stemming feasible, unavoidably drops content that might matter in longer articles. Together these constraints motivate the future directions described next.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

4. CONCLUSION

Pitting Random Forest, SVM, Naïve Bayes, and XGBoost against one another on 1,116 labeled Indonesian articles, under a single preprocessing pipeline and two feature representations, identified the best performer on the held-out test set within the evaluated dataset and split: Random Forest with Bag-of-Words, scoring 98.66% accuracy and a 98.59% macro F1-score. XGBoost with Bag-of-Words proved highly competitive and was in fact slightly stronger under repeated cross-validation, where its mean macro F1 of 98.36% exceeded Random Forest's 97.44% by a statistically significant margin, so the two ensembles should be read as closely matched rather than definitively ranked. Across both evaluation settings BoW outdid TF-IDF for every classifier examined. Its practical weight is that interpretable, inexpensive classical ensembles can rival the accuracies typically tied to deep learning at a fraction of the computational cost, offering Indonesian fact-checking bodies a workable means of lightweight first-pass screening of suspected hoaxes rather than a substitute for final fact-checking judgment; human verification remains necessary after any automated flag. These results carry candid qualifications a dependence on lexical features, a single-source dataset whose legitimate class is dominated by debunking texts, and shortened article bodies all limit the study's semantic and cross-domain scope, so broader validation is needed before operational deployment. The logical next moves stem from those limits: incorporate contextual embeddings like IndoBERT, broaden the corpus over additional sources, platforms, and time periods, and carry out explicit cross-domain and temporal testing to learn whether the edge found here carries past this single dataset.

DATA AND CODE AVAILABILITY

The data examined here can be obtained freely from the MAFINDO/TurnBackHoax fact-checking collection [17] (accessed 15 January 2026). To support reproducibility, every step reported in Section 2 is executable from the public repository: the preprocessing pipeline (case folding, cleaning, tokenisation, stopword removal, and Sastrawi stemming, with the 200-word body truncation), the feature-extraction and classification scripts for both Bag-of-Words and TF-IDF, the cross-validation and statistical-comparison scripts, and the processed data and result files backing these findings are all openly accessible at https://github.com/dedimti/hoax-detection-ml. All experiments use fixed random seeds,

so the reported figures can be regenerated exactly. The code is released under the MIT License, and any additional information is available from the corresponding author upon reasonable request.

ACKNOWLEDGMENT

The authors are grateful to the Indonesian Anti-Hoax Society (MAFINDO) for keeping up the openly available TurnBackHoax dataset drawn on in this work, and recognize the open-source contributors who assembled and released it.

REFERENCES

[1] D. K. S. Sekarhati, "Combating Hoax and Misinformation in Indonesia Using Machine Learning: What Is Missing and Future Directions," *Eng. Math. Comput. Sci. J. (EMACS)*, vol. 6, no. 2, pp. 143–150, 2024. doi: 10.21512/emacsjournal.v6i2.11556

[2] D. Rahmawan, I. Garnesia, and R. Hartanto, "Checking the Fact-Checkers: Analyzing the Content of Fact-Checking Organizations as Initiatives for Hoax Eradication in Indonesia," *J. ASPIKOM*, vol. 8, no. 2, pp. 241–256, 2023. doi: 10.24329/aspikom.v8i2.1267

[3] J. Fawaid et al., "TurnBackHoax Dataset: Indonesian News Dataset from turnbackhoax.id," 2021. [Online]. Available: https://github.com/JibranFawaid/turnbackhoax-dataset.

[4] I. Y. R. Pratiwi, R. A. Asmara, and F. Rahutomo, "Study of hoax news detection using naïve Bayes classifier in Indonesian language," *in Proc. 2017 11th Int. Conf. Inf. Commun. Technol. Syst. (ICTS)*, Surabaya, Indonesia, 2017, pp. 73–78. doi: 10.1109/ICTS.2017.8265649

[5] F. Rahutomo, I. Y. R. Pratiwi, and D. M. Ramadhani, "Naïve Bayes's experiment on hoax news detection in Indonesian language," *J. Penelit. Komun. dan Opini Publik*, vol. 23, no. 1, pp. 1–15, 2019.

[6] A. Awalina, J. Fawaid, R. Y. Krisnabayu, and N. Yudistira, "Indonesia's Fake News Detection using Transformer Network," *in Proc. 6th Int. Conf. Sustainable Inf. Eng. Technol. (SIET)*, Malang, Indonesia: ACM, 2021, pp. 247–251, doi: 10.1145/3479645.3479666.

[7] D. Wijaya, G. M. Sasmitha, and W. O. Vihikan, "Sentiment Analysis of Indonesian Citizens on Electric Vehicle Using FastText and BERT Method," *J. Inf. Syst. Informatics*, vol. 6, no. 3, pp. 1360–1372, 2024. doi: 10.51519/journalisi.v6i3.784

[8] O. C. R. Rachmawati and Z. M. E. Darmawan, "The Comparison of Deep Learning Models for Indonesian Political Hoax News Detection," *CommIT J.*, vol. 18, no. 2, pp. 123–135, 2024.

[9] M. Granik and V. Mesyura, "Fake news detection using naïve Bayes classifier," *in Proc. 2017 IEEE 1st Ukraine Conf. Electr. Comput. Eng. (UKRCON)*, 2017, pp. 900–903. doi: 10.1109/UKRCON.2017.8100379

[10] F. Panjaitan, W. Ce, H. Oktafiandi, G. Kanugrahan, Y. Ramdhani, and V. H. C. Putra, "Evaluation of Machine Learning Models for Sentiment Analysis in the South Sumatra Governor Election Using Data Balancing Techniques," *J. Inf. Syst. Informatics*, vol. 7, no. 1, pp. 461–478, 2025. doi: 10.51519/journalisi.v7i1.1019

[11] P. R. A. Savitri, I M. A. D. Suarjaya, and W. O. Vihikan, "Sentiment Analysis of X (Twitter) Comments on The Influence of South Korean Culture in Indonesia," *J. Inf. Syst. Informatics,* vol. 6, no. 2, 2024. doi: 10.51519/journalisi.v6i2.738

[12] A. N. Maulana, W. M. Y. bin Wan Yaacob, and Felawati, "Analyzing Public Sentiment on the Proposal to Return Regional Head Elections to DPRD on Platform X Using the C4.5 Algorithm," *J. Inf. Syst. Informatics*, vol. 8, no. 2, pp. 2529–2549, 2026. doi: 10.63158/journalisi.v8i2.1521

[13] F. Sebastiani, "Machine learning in automated text categorization," *ACM Comput. Surv.*, vol. 34, no. 1, pp. 1–47, 2002. doi: 10.1145/505282.505283

[14] R. N. Devita, H. W. Herwanto, and A. P. Wibawa, "Perbandingan kinerja metode naïve Bayes dan K-nearest neighbor untuk klasifikasi artikel berbahasa Indonesia," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 4, pp. 427–434, 2018. doi: 10.25126/jtiik.201854773

[15] I. Y. R. Pratiwi and A. F. Nugraha, "Hoax news identification using machine learning model from online media in Bahasa Indonesia," *Matrix: J. Manaj. Teknol. dan Inform.,* vol. 12, no. 2, pp. 58–67, 2022. doi: 10.31940/matrix.v12i2.58-67

[16] I. F. Putra and A. Purwarianti, "Improving Indonesian text classification using multilingual language model," *in Proc. 2020 7th Int. Conf. Adv. Informatics: Concept, Theory Appl. (ICAICTA)*, Tokoname, Japan: IEEE, 2020, pp. 1–5, doi: 10.1109/ICAICTA51733.2020.9429038.

[17] Y. Setiawan et al., "FakeNews-Mafindo: An Indonesian Fact-Checking Dataset," 2023. [Online]. Available: https://huggingface.co/datasets/nlp-brin-id/fakenews-mafindo. [Accessed: Jan. 15, 2026].

[18] C. C. Aggarwal and C. Zhai, Mining Text Data. New York: *Springer*, 2012. doi: 10.1007/978-1-4614-3223-4

[19] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324

[20] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995. doi: 10.1007/BF00994018

[21] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 785–794. doi: 10.1145/2939672.2939785

[22] P. Zhang, Y. Jia, and Y. Shang, "Research and application of XGBoost in imbalanced data," *Int. J. Distrib. Sens. Netw.*, vol. 18, no. 6, 2022. doi: 10.1177/15501329221106935

[23] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[24] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci. Rep.*, vol. 14, no. 1, p. 6086, 2024. doi: 10.1038/s41598-024-56706-x

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *in Proc. 2019 Conf. North Am. Chapter Assoc. Comput. Linguistics (NAACL-HLT)*, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423

[26] A. K. Darmawan, M. W. Al Wajieh, M. B. Setyawan, T. Yandi, and H. Hoiriyah, "Hoax News Analysis for the Indonesian National Capital Relocation Public Policy with the Support Vector Machine and Random Forest Algorithms," *J. Inf. Syst. Informatics*, vol. 5, no. 1, pp. 150–173, 2023. doi: 10.51519/journalisi.v5i1.438

[27] D. A. N. Krisna and U. Salamah, "Perbandingan algoritma naïve Bayes dan K-nearest neighbor untuk klasifikasi berita hoax kesehatan di media sosial Twitter," J. *Tek. Inform. Kaputama (JTIK)*, vol. 6, no. 2, 2022.

[28] E. Rasywir and A. Purwarianti, "Eksperimen pada sistem klasifikasi berita hoax berbahasa Indonesia berbasis pembelajaran mesin," *J. Cybermatika*, vol. 3, no. 2, 2016.

[29] N. Hoy and T. Koulouri, "An exploration of features to improve the generalisability of fake news detection models," *Expert Syst. Appl.*, vol. 275, art. 126949, 2025. doi: 10.1016/j.eswa.2025.126949.