SYNTIA – Systems and Technologies in Artificial Intelligence Applications.
Volume 1 Issue 2 (June 2026). Page: 55-66.

Random Forest on Frozen IndoBERT Embeddings for Indonesian Political Hoax Detection

Daniel Irwansyah[1], Anita Ahmad Kasim[2*], Hajra Rasmita Ngemba[3]

[1] *Informatics Engineering Study Program, Faculty of Engineering, Universitas Tadulako, Palu, Indonesia.*
[2] *Department of Information Technology, Faculty of Engineering, Universitas Tadulako, Palu, Indonesia.*
[3] *Doctoral Programme, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Malaysia.*

**Corresponding author: anita.ahmad@untad.ac.id*

Abstract: Political hoaxes shape public opinion, yet Indonesian-language detectors are often reported with a single accuracy figure. This study classifies political news as hoax or non-hoax with frozen IndoBERT embeddings (768 dimensions, mean pooling) and a Random Forest. A public collection of 2,262 articles was split 80:20, leaving 453 test articles (215 hoax, 238 non-hoax). A default forest reached 91.61% accuracy (F1-score 91.40%, AUC 0.9746) and a forest tuned by five-fold grid search reached 92.72% (F1-score 92.52%, AUC 0.9753). The 95% Wilson intervals of the two accuracies (88.7–93.8% and 90.0–94.8%) overlap almost entirely, and a net difference of five articles cannot reach $p < 0.05$ in an exact McNemar test whatever the overlap of errors (best case $p = 0.0625$), so tuning is not shown to help. The tuned forest flagged 22 of 238 genuine articles as hoax and missed 11 of 215 hoaxes; 12–13% of test articles received a forest vote share between 0.4 and 0.6. Hoax articles come from a fact-checking site and genuine articles from three news portals, so the classifier may separate sources instead of truth, and no experiment here excludes this. We specify a six-check audit, including keyword ablation, a cross-source test, and a lexical baseline, and conclude that the results show separability within this collection, not hoax detection in general.

Keywords : hoax detection; IndoBERT; political news; Random Forest; text classification.

Received: February 23, 2026 | Revised: April 6, 2026 | Accepted: May 4, 2026 | Published: June 30, 2026

1. Introduction

Political hoaxes spread quickly and shape public opinion. A large study of Twitter found that false news diffused farther and faster than true news, and that the difference came from people sharing it, not from automated accounts [1]. Automatic detection of fake news is therefore an active area, and a data-mining survey groups the approaches into those that use news content and those that use social context, including propagation [2]. Content-based methods, which use only the text of an article, are attractive because they need no social-network data and can be applied before an article circulates. For Indonesian, they have become practical with pretrained language models such as IndoBERT, released together with benchmark suites for Indonesian language understanding [3], [4], and the multilingual BERT family on which they build [5].

Indonesian hoax detection has been studied with several designs (Table 1). Fine-tuned IndoBERT reached 94.1% accuracy on a balanced Kaggle sample of political news [6], an IndoBERT–LSTM hybrid reached 93.2% accuracy on articles scraped from a fact-checking site and a news portal [7], and a Random Forest tuned by grid search on TF-IDF features improved from 96% to 97% on another collection [8]. These reports use their own collections, which limits comparison, and in at least one of them [7] the hoax and genuine articles come from different outlets, as in the present data. A classifier can then reach a high score by recognizing the outlet, its layout, or its vocabulary, which is a case of shortcut learning, where a model succeeds on a benchmark by a route that does not transfer [9]. The same phenomenon is documented in natural language inference, where annotation artifacts alone allow above-chance prediction [10].

SYNTIA – Systems and Technologies in Artificial Intelligence Applications is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

Table 1. Selected Indonesian hoax-detection studies.

Study	Features / model	Reported result
Jocelynne et al. [6]	Fine-tuned IndoBERT; 1,302 articles per class	94.1% accuracy, AUC 0.991
Yefferson et al. [7]	IndoBERT + LSTM; 5,876 articles	93.2% accuracy, F1 90.8%
Riyandari et al. [8]	TF-IDF + Random Forest, grid search	96% → 97% (tuned)
Present study	Frozen IndoBERT + Random Forest; 2,262 articles	91.61% (default), 92.72% (tuned)

Results are as reported by each study; collections and protocols differ, so rows are not a ranking.

A cheap alternative to fine-tuning is to keep the language model frozen, extract a fixed vector for each text, and train a classical classifier on it. This needs a single forward pass per article, no GPU for training, and a model whose behavior is easy to inspect. It is a natural baseline whose strength should be measured before heavier models are adopted. This study evaluates that baseline for Indonesian political hoaxes: 768-dimensional frozen IndoBERT embeddings feed a Random Forest [11], trained with default settings and with a five-fold grid search. Beyond the headline scores, we ask how large the uncertainty on the 453-article test set is, whether tuning produced a measurable gain, which error types remain, how the forest's probabilities are distributed, and what the construction of the collection allows us to conclude.

The choice between a frozen and a fine-tuned encoder is a trade-off between cost and accuracy. Fine-tuning adapts all layers to the task and usually gives the best scores, but it needs labeled data, a GPU, and careful regularization on small collections; a frozen encoder keeps the general-purpose representation learned in pretraining, and the downstream model can be changed, inspected, or retrained cheaply [5], [12]. For a collection of about two thousand articles, the frozen route is plausible, but its accuracy relative to fine-tuning or to a lexical model on the same data is rarely reported together, and the present study cannot supply that comparison either; it supplies a measured baseline and the uncertainty around it.

2. Method

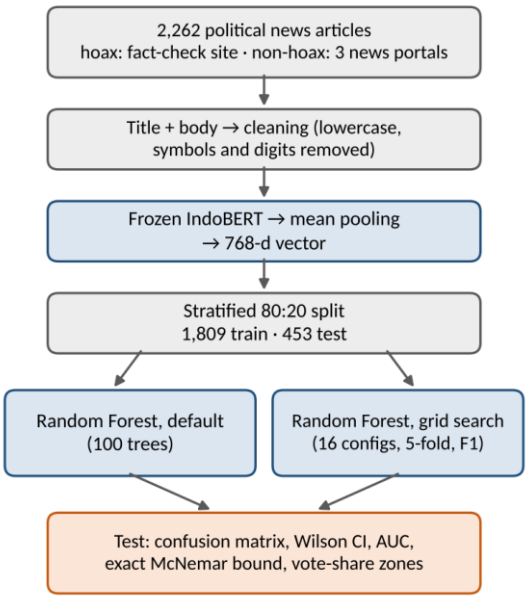


Figure 1. Research workflow.

The contributions are: (i) a reproducible baseline with a frozen encoder and a classical classifier for Indonesian political hoax text; (ii) interval estimates for every proportion and an exact analysis of what a paired test could show about the effect of tuning; (iii) an analysis of the forest's probability output and of the cost of false alarms at realistic hoax prevalence; and (iv) an explicit statement of the source confound in the collection, with a six-check audit protocol that can be run on the same data. Section 2 describes data and methods, Section 3 reports and discusses results, and Section 4 concludes.

2.1. Data

The data are a public collection of Indonesian political news taken from the Kaggle platform. Hoax articles come from TurnBackHoax.id, the fact-checking site run by MAFINDO, and non-hoax articles come from three news portals: CNN Indonesia, Kompas, and Detik. Each record has a title, timestamp, full text, tags, author, and a source URL. Only the title and full text were used; they were concatenated into one text field. Labels are 1 for hoax and 0 for non-hoax. The collection has 2,262 articles. A stratified 80:20 split gave 1,809 training and 453 test articles (Table 2), and the test set holds 238 non-hoax and 215 hoax articles (47.5% hoax).

Table 2. Data partition.

Set	Non-hoax	Hoax	Total
Training	n/a	n/a	1,809
Test	238	215	453
Total	n/a	n/a	2,262

n/a: class counts of the training set are not given in the source; stratification implies a similar proportion to the test set.

The class definitions deserve attention. The hoax class is built from items that a fact-checking organization has already examined, and the non-hoax class from articles that mainstream outlets published. The task that the classifier solves is therefore to tell fact-checking-site items from news-portal items, which coincides with hoax versus genuine only to the degree that outlet and style carry no information beyond veracity. Fact-check entries also commonly quote the false claim before rebutting it, so a text labeled hoax may describe the hoax instead of being it. Both the title and the full article text were used for every record, hoax and non-hoax alike (Section 2.1), so this confound applies uniformly rather than following from a field choice specific to one class. These points are examined in Section 3.5.

2.2. Preprocessing and embedding

Text was lower-cased, and symbols, punctuation, hashtags, and digits were removed. Stop words were kept and no stemming was applied, because the encoder works on sub-word tokens of running text. Each text was tokenized with the WordPiece vocabulary of the encoder, with [CLS] and [SEP] markers and padding [5]. We used the encoder without fine-tuning, loading the IndoLEM/IndoBERT base-uncased checkpoint [3] with its matching WordPiece tokenizer and truncating or padding every sequence to 256 tokens. For a text of N tokens with final-layer vectors h_i and attention mask m_i, the embedding is the masked mean of the token vectors, a common way to obtain a fixed-size sentence representation [12].

$$\mathbf{e} = \frac{\sum_{i=1}^{N} m_i \mathbf{h}_i}{\sum_{i=1}^{N} m_i} \in \mathbb{R}^{768} \quad (1)$$

The embeddings of the 1,809 training and 453 test articles formed the feature matrices. Because the encoder is frozen and the embeddings are computed per article, they carry no information from the test set into training.

2.3. Classifiers and tuning

Random Forest [11] as implemented in scikit-learn [13] was used. The default model has 100 trees, no depth limit, a minimum of 2 samples to split a node, a minimum of 1 sample per leaf, the square root of the feature count tried at each split, and a fixed seed of 42. The tuned model came from an exhaustive search over 16 configurations (Table 3) with five-fold stratified cross-validation on the training set and the F1-score as selection criterion. The best configuration had 200 trees, maximum depth 20, minimum split size 5, and minimum leaf size 1, with a mean cross-validated F1-score of 0.9252.

Table 3. Hyperparameter grid of the tuned Random Forest.

Parameter	Default	Grid	Selected
Number of trees	100	100, 200	200
Maximum depth	None	10, 20	20
Min. samples to split	2	2, 5	5
Min. samples per leaf	1	1, 5	1

Cross-validated F1-score of the selected configuration: 0.9252.

Selecting hyperparameters and reporting performance on the same data inflates the estimate, which is why the search used only the training data and the test set was used once for reporting [14]. The grid is small and contains the default values only partially: the default has no depth limit, whereas the grid limits the depth to 10 or 20.

A Random Forest grows T decision trees on bootstrap samples of the training set, choosing at each split the best among a random subset of features by an impurity criterion, here the Gini impurity of a node with class proportions p_k, $G = 1 - \Sigma\, p_k^2$. The forest's probability for the hoax class is the average of the tree outputs [11],

$$\hat{P}(\text{hoax} \mid \mathbf{e}) = \frac{1}{T}\sum_{t=1}^{T} P_t\,(\text{hoax} \mid \mathbf{e}) \tag{2}$$

and the predicted class is hoax when this value exceeds 0.5. Random subsets of the 768 embedding dimensions decorrelate the trees. Because each tree sees a different bootstrap sample, about one third of the training articles are left out of each tree, which gives an out-of-bag estimate that could be used for monitoring without a separate validation set, although it was not used here. Grid search with five folds fits $16 \times 5 = 80$ forests and then refits the best configuration on all 1,809 training articles, so the tuned model has seen the same training articles as the default model.

2.4. Evaluation and statistics

Accuracy, precision, recall of the hoax class, specificity, F1-score, and the area under the ROC curve (AUC) were computed on the 453 test articles. For every proportion we report a 95% Wilson score interval [15],

$$\mathrm{CI}_{95\%} = \frac{\hat{p} + \frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}} \tag{3}$$

where $\hat{p}$ is the observed proportion, n the number of articles, and $z = 1.96$. F1-score and AUC are shown without intervals because their intervals require per-article predictions, which are not available. The two forests were evaluated on the same articles, so their comparison is paired, and the McNemar test is appropriate [16], [17]. With b articles correct for the default forest only and c for the tuned forest only, the exact two-sided p-value is

$$p = \min\left\{1,\; 2\sum_{i=0}^{\min(b,c)} \binom{b+c}{i} \left(\frac{1}{2}\right)^{b+c}\right\} \tag{4}$$

Per-article predictions were not stored, so b and c are unknown. The default forest made 38 errors and the tuned forest 33, so $b - c = 5$. We therefore computed the p-value over all possible values of $b + c$, which gives the most and least favorable cases for significance without assuming anything about overlap.

The forest's probability for the hoax class is the mean over trees of the fraction of hoax training articles in the reached leaf. Its distribution over the test set was summarized in five bins of width 0.2, and the central bin (0.4–0.6) was treated as an uncertainty zone. Probabilities of forests are known to be pulled away from 0 and 1, so they require calibration before being read as confidences [18], [19].

2.5. Reproducibility and ethics

Everything reported in the Results was either read from the experiment outputs of the original study or derived from its confusion matrices: Wilson intervals, the McNemar analysis, chance-corrected metrics, the prevalence table, and the sample-size calculation. The random seed of the default forest was 42, and the split, the grid, and the selected configuration are given in Tables 2 and 3. The code, the embeddings, and the list of test articles should be released together with the data citation. The data are published articles and fact-check entries; no personal data were collected, but the texts name public figures, which matters for the discussion of bias in Section 3.7.

3. Results and Discussion

3.1. Classification performance

Figure 2 shows the confusion matrices. The default forest classified 415 of 453 test articles correctly and the tuned forest 420 (Table 4). Their accuracies are 91.61% and 92.72%, with Wilson intervals of 88.7–93.8% and 90.0–94.8%. Precision for the hoax class was 88.99% and 90.27%, recall 93.95% and 94.88%, and specificity 89.50% and 90.76%. F1-scores were 91.40% and 92.52%, and the AUC 0.9746 and 0.9753 (Fig. 3 shows the ROC curve of the tuned forest). All intervals overlap almost completely between the two models. The tuned forest is nominally better on every metric, by about one point, but each interval is about five points wide.

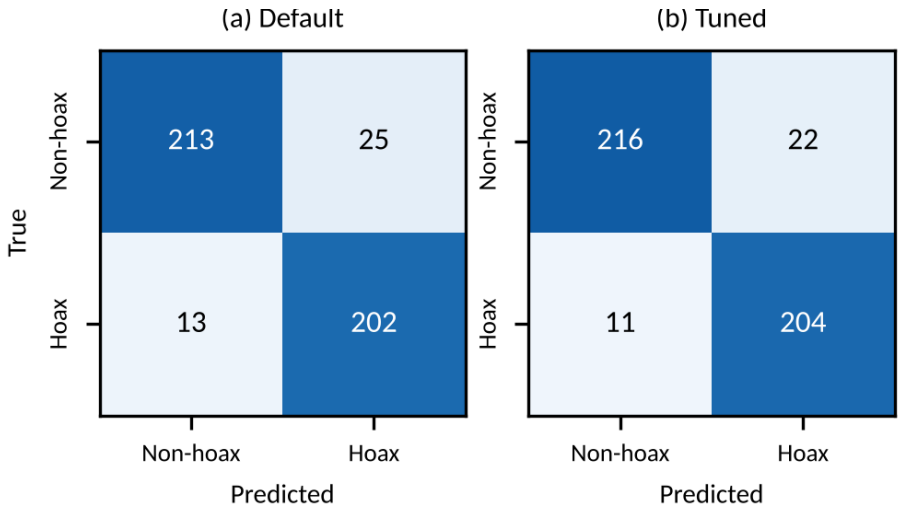


Figure 2. Confusion matrices of the default (a) and tuned (b) forests on the 453 test articles.

Table 4. Test-set performance (n = 453) with 95% Wilson intervals for proportions.

Metric	Default	Tuned
Accuracy (%)	91.61 (88.70–93.83)	92.72 (89.95–94.77)
Precision, hoax (%)	88.99 (84.25–92.43)	90.27 (85.70–93.48)
Recall, hoax (%)	93.95 (89.93–96.43)	94.88 (91.07–97.12)
Specificity (%)	89.50 (84.95–92.78)	90.76 (86.40–93.82)
F1-score (%)	91.40	92.52
AUC	0.9746	0.9753
Errors (FP + FN)	38 (25 + 13)	33 (22 + 11)

FP: non-hoax predicted as hoax; FN: hoax predicted as non-hoax.

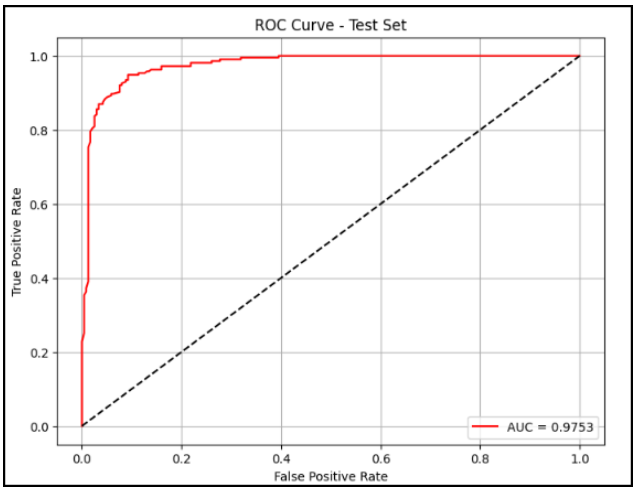


Figure 3. ROC curve of the tuned forest on the test set (AUC = 0.9753); the dashed line is chance.

The AUC values are high, which means that the hoax probability ranks hoax articles above genuine ones in about 97.5% of random pairs. This is a stronger statement about separability than accuracy at one threshold, and it holds for both forests, so the choice of threshold, not the forest, is the main lever for trading false alarms against misses. The F1-score of the tuned forest on the test set (92.52%) equals its mean cross-validated F1-score on the training set (0.9252). Such a coincidence to four digits is possible, but it is unusual enough that the logs of the grid search and the test evaluation should be checked before the two numbers are presented as independent.

Because accuracy depends on class balance and a single threshold, Table 5 adds per-class and chance-corrected summaries computed from the confusion matrices. Precision for the non-hoax class (94.25% and 95.15%) is higher than for the hoax class, as is expected when false alarms outnumber misses. Macro-averaged F1-score, balanced accuracy, Cohen's kappa, and the Matthews correlation coefficient all tell the same story as accuracy: about 0.83–0.86 for kappa and MCC, which is conventionally read as strong agreement, and a gain of two points on those scales for the tuned forest that is as uncertain as the accuracy gain. The positive likelihood ratio, 8.9 and 10.3, says that a hoax flag raises the odds that an article is a hoax by roughly nine to ten times relative to the prior odds, while the negative likelihood ratio, 0.068 and 0.056, says that an unflagged article has about one-seventeenth to one-eighteenth of its prior odds of being a hoax.

Table 5. Per-class and chance-corrected metrics from the confusion matrices.

Metric	Default	Tuned
Precision / recall / F1, non-hoax (%)	94.25 / 89.50 / 91.81	95.15 / 90.76 / 92.90
Precision / recall / F1, hoax (%)	88.99 / 93.95 / 91.40	90.27 / 94.88 / 92.52
Macro F1-score (%)	91.61	92.71
Balanced accuracy (%)	91.72	92.82
Cohen's kappa	0.832	0.854
Matthews correlation	0.833	0.855
Positive / negative likelihood ratio	8.94 / 0.068	10.26 / 0.056

Computed from the confusion matrices in Fig. 2; hoax is the positive class.

3.2. Effect of tuning

Tuning reduced errors from 38 to 33, a net gain of five articles out of 453 (1.1 points). Such a gain is within sampling noise, and the exact McNemar analysis shows more than that. If the two forests shared all but five of their errors, the five articles would all favor the tuned forest and the exact two-sided p-value would be 0.0625. This is the best case, and it already exceeds 0.05. In any other configuration of shared errors the p-value is larger, up to 0.64 when the forests share no errors at all (Fig. 4b and Table 6). A net

difference of five articles on a test set of this size therefore cannot be statistically significant under any pattern of overlap, and the statement that tuning improved the model significantly is not supported.

Table 6. Exact McNemar p-value for a net difference of five articles (38 vs 33 errors).

Discordant articles (b + c)	b / c	Exact p
5 (best case)	5 / 0	0.063
9	7 / 2	0.180
15	10 / 5	0.302
31	18 / 13	0.473
71 (worst case)	38 / 33	0.635

b: articles correct only for the default forest; c: only for the tuned forest.

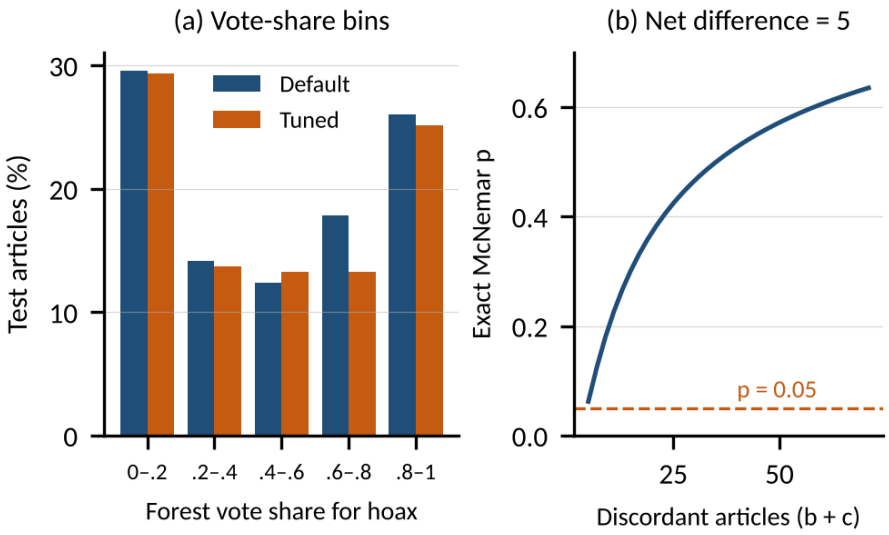


Figure 4. (a) Share of the 453 test articles in each bin of forest probability for hoax. The tuned bins sum to 94.7% in the original results, so about 5% of articles are unaccounted for. (b) Exact McNemar p-value against the number of discordant articles when the net difference is five.

How large would a test set need to be? With a net gain of 1.1 points and 80% power, the normal approximation for the paired design needs about 3,200 articles if the models disagree on 5% of them and 9,700 if they disagree on 15% [20]; the disagreement cannot exceed 15.7%, because the forests made 8.4% and 7.3% errors. With 453 articles the power is only 9–18%, so this test set can detect differences of roughly three points or more, and the absence of a significant difference is not evidence that tuning has no effect.

Table 7. Test-set size needed to detect a net gain of 1.1 points with 80% power (two-sided $\alpha = 0.05$).

Disagreement rate q	Articles needed	Power at n = 453 (%)
5%	3,241	18
10%	6,485	11
15%	9,728	9

Normal approximation for the paired design; q is unknown because per-article predictions were not stored.

3.3. Error structure and the cost of false alarms

The tuned forest flagged 22 of 238 genuine articles as hoax (9.2%) and missed 11 of 215 hoaxes (5.1%); Table 8 gives intervals for these rates. The default forest flagged 25 (10.5%) and missed 13 (6.0%). False alarms outnumber misses in both models, and in a moderation setting they are also the costlier error, because a genuine report is labeled false. Precision therefore matters for avoiding the wrongful labeling of true news. One practical response is to raise the decision threshold above 0.5; the shape of the ROC curve suggests that a few points of recall might be traded for a larger reduction in false alarms, but the exact trade cannot be computed without per-article scores.

Table 8. False-alarm and miss rates on the test set with 95% Wilson intervals.

Error rate	Default	Tuned
False alarms: 25 and 22 of 238	10.5% (7.2–15.0)	9.2% (6.2–13.6)
Misses: 13 and 11 of 215	6.0% (3.6–10.1)	5.1% (2.9–8.9)
Articles in 0.4–0.6 zone	12.4% (9.6–15.7)	13.2% (10.4–16.7)

The intervals of the two models overlap almost entirely for every row.

The test set is nearly balanced (47.5% hoax), whereas hoaxes are a small share of real news traffic. Precision depends on prevalence, so the 90.3% precision of the tuned forest cannot be carried over to a live feed. Holding the measured sensitivity (94.88%) and false-positive rate (9.24%) fixed, Table 9 gives the share of flagged articles that would be genuine hoaxes at lower prevalence. At 5% hoax prevalence only about one flagged article in three would be a true hoax, and at 1% about one in eleven. These numbers assume that the error rates remain the same under the shift, which is itself doubtful, but they show that a classifier with 92.7% accuracy on a balanced test set is far from ready for unsupervised flagging.

Table 9. Share of flagged articles that are hoaxes (positive predictive value) at different prevalence, for the tuned forest.

Hoax prevalence	PPV (%)	NPV (%)
47.5% (test set)	90.3	95.1
20%	72.0	98.6
10%	53.3	99.4
5%	35.1	99.7
1%	9.4	99.9

Derived from sensitivity 204/215 and false-positive rate 22/238; assumes the rates do not change with prevalence.

3.4. Probability distribution and uncertainty zone

For the default forest, 29.58% and 26.05% of test articles had probabilities in the outer bins (0–0.2 and 0.8–1), 12.36% in the central bin, and 14.13% and 17.88% in the two intermediate bins (Fig. 4a). For the tuned forest the outer bins held 29.36% and 25.17% and the central bin 13.25%. The original analysis counts 55.6% and 54.5% of predictions as extreme and reads this as high confidence. The bins of the tuned forest sum to 94.7%, not 100%, so the remaining 5.3% (about 24 articles) are not accounted for, and the intermediate-zone percentage given in the original analysis (32.2%) does not follow from its own table; we therefore do not interpret the tuned distribution beyond the central bin.

Two cautions apply to reading these bins as confidence. First, a forest probability is an average over trees and is not calibrated; for tree ensembles the predicted probabilities are typically pulled away from 0 and 1 [18], so a probability of 0.8 does not mean that 80% of such articles are hoaxes. A reliability diagram or the expected calibration error [19] on held-out data would measure this directly. Second, the share of articles in the central bin (12–13%, or 56 and 60 articles) is of the same order as the number of errors (38 and 33), but without per-article results we cannot tell how many errors lie in that zone. If most do, routing the 0.4–0.6 zone to a human fact-checker would remove most errors at the cost of reviewing about one article in eight; if few do, the forest is confidently wrong, which is the worse case. This is a cheap analysis for future work.

3.5. What the collection can and cannot show

The high scores indicate that, within this collection, the two classes are separable by the embedding of the text. They do not indicate that the classifier detects false claims. Three features of the collection allow separation without any notion of truth. First, outlet and style: the hoax class comes from one site with a fixed template and recurring vocabulary (for example words used to introduce a claim or a rebuttal), and the genuine class from three portals with their own boilerplate, such as related-article prompts or agency bylines. Second, time: if the fact-check items and the news items were collected over different periods, topical vocabulary (names, events) differs by class. Third, content type: fact-check entries discuss a claim,

whereas news reports describe an event. A frozen encoder with a Random Forest can pick up each of these cues, and none of them transfers to a new outlet.

These are hypotheses about the collection and not findings, because the experiments needed to test them were not run. The point is that the reported accuracy cannot discriminate between the hypothesis that the classifier reads veracity and the hypothesis that it reads source, and the second is the more likely for any dataset assembled this way [9], [10]. The conclusion that the combination of IndoBERT and Random Forest is effective and reliable for detecting political hoaxes is therefore too broad. What the data support is that the combination is effective at separating these fact-check items from these news items.

3.6. Audit protocol

Table 10 lists six checks that can be run on the same collection with modest effort and that would turn the present baseline into evidence about hoax detection. Each needs a pass criterion fixed in advance and justified by test-set size; for example, an accuracy fall of more than five points after ablation would show that source cues carry much of the signal, and cross-source accuracy near 50–60% would show that the model recognizes outlets, not hoaxes.

Table 10. Audit checks and the evidence each would give.

Check	Procedure	What it shows
Keyword and boilerplate ablation	Remove outlet names, bylines, template phrases, and claim-introducing words; re-run	Accuracy drop measures reliance on source cues
Cross-source test	Train on two news portals, test on the third; add a second fact-check site	Transfer to unseen outlets
Lexical baseline	TF-IDF with logistic regression or Random Forest on the same split	Whether IndoBERT adds anything
Fine-tuned comparison	Fine-tune IndoBERT [3] with the same split and tuning budget	Gap between frozen and fine-tuned encoders
Repeated splits	Repeat split and tuning 5–10 times; corrected t-test	Whether tuning helps beyond noise
Near-duplicate check	Hash and embedding similarity between train and test	Leakage between splits

The lexical baseline is the most informative single check: if TF-IDF with a linear model reaches a similar accuracy, the encoder adds little beyond surface vocabulary. The cross-source test also needs hoax examples from outside TurnBackHoax.id, for instance from another fact-checking organization; otherwise a model that recognizes the template of the fact-check site would still score well on the hoax class. Each held-out outlet needs about a hundred articles per class for the Wilson interval to stay within ten points. Previous Indonesian studies [6]–[8] each report only one family of features, so a joint comparison on one split would also help to place their results.

A last limit concerns content-only detection: a text classifier sees how an article is written, not whether its claims are true. Systems that retrieve evidence or use social context address this directly, at a higher cost [2]; a text-only baseline such as this one is best seen as a first filter whose output goes to a human with the evidence displayed.

3.7. Bias, preprocessing, and what the results can support

Political text raises a bias risk beyond source style: if fact-checkers examine claims about certain figures more often, a classifier can learn that mentions of those figures predict the hoax class and then flag genuine reports about them more often, a harm as well as an accuracy problem. The check is to mask named entities or compute the false-alarm rate per frequently named actor; with 238 genuine test articles such rates are imprecise, so a larger sample is needed. Preprocessing also matters: removing digits and symbols discards dates, vote counts, and hashtags that may differ between classes, and mean pooling of a frozen encoder dilutes rare but decisive tokens [12]. A lexical baseline and a cleaning ablation would show whether

these choices help or hurt. Table 11 states the claims that a result of this kind is often taken to support and what the evidence in this study allows.

Table 11. Claims and the strength of the evidence in this study.

Claim	Evidence in this study
The models separate these hoax and non-hoax articles	Supported: accuracy 91.6–92.7%, AUC about 0.975
Tuning improved the model	Not shown: net 5 of 453, $p \geq 0.0625$ in the best case
The classifier detects falsehood	Not tested: classes differ by source
IndoBERT contributes strongly	Not tested: no lexical baseline
Predictions are confident	Not shown: probabilities uncalibrated
Results generalize to other outlets	Not tested: single source mix, single split

3.8. Position relative to earlier studies

The accuracy of the present forests (91.6–92.7%) is below the 94.1% of fine-tuned IndoBERT [6] and the 93.2% of the IndoBERT–LSTM hybrid [7], and below the 96–97% of Random Forest on TF-IDF features [8]. The comparison is only indicative because the collections differ in size, source mix, and balance, but its direction is consistent with the expectation that fine-tuning an encoder adds accuracy over freezing it. The TF-IDF result is the most important for interpretation. If a lexical model reaches 96–97% on another collection of fact-check and news items, high scores are easy to obtain from surface cues, which supports the caution of Section 3.5. The present study cannot decide the issue without a lexical baseline on its own data.

Frozen embeddings have a practical appeal. The encoder needs one forward pass per article, a Random Forest on 768 features trains quickly, and no GPU is needed for training. For a newsroom or a fact-checking unit that needs a first-pass filter, this matters more than a point of accuracy. The finding that grid search adds little also has a positive reading: the default forest is already near the plateau, so practitioners can skip tuning and spend effort on data quality and on calibration instead.

3.9. Limitations and future work

Several limitations qualify the results. The results come from a single 80:20 split with 453 test articles, so every estimate has an interval of about five points. The origin of the data is a compiled collection whose period, topic filter, and duplicates are unknown to us. The encoder checkpoint, maximum sequence length, and the truncation policy were not documented in the source material, and long articles may have been truncated, with the consequence that only the start of each article (which often differs in style between sources) was embedded. No lexical, fine-tuned, or alternative-classifier baseline was run, so the contribution of the encoder is not isolated.

Future work should run the audit of Table 10, report calibration, and add per-article predictions to allow paired tests and error analysis. Fine-tuning IndoBERT with the same split, comparing with other classifiers such as gradient boosting or a linear model on the same embeddings, and testing on hoaxes from outlets outside the collection are the natural next experiments. A final extension is to define the task more closely: classifying the original circulating text of a hoax, not the fact-checker's article about it, would address the problem that is of practical interest.

3.10. Test-set size and recommendations

The width of the intervals follows directly from the size of the test set. At 92% accuracy, the half-width of the 95% Wilson interval is about 2.5 points for 453 articles, 5.4 for 100, 1.7 for 1,000, and 0.8 for 4,000. Halving the interval needs about four times as many articles, which is why differences of one point between models are rarely resolvable on test sets of a few hundred items.

From the analysis above we draw five recommendations for studies of Indonesian hoax detection with pretrained encoders. They are practical and each was motivated by a specific weakness found here (Table 12).

Table 12. Recommendations motivated by the analysis.

Recommendation	Weakness addressed
Report the outlet composition of each class and the period covered	Source and time confounds
Add a lexical baseline and, where feasible, a fine-tuned model	Contribution of the encoder unknown
Evaluate on outlets not seen in training	Shortcut learning on source
Give an interval for every proportion and store per-article predictions	Over-reading of small gaps; no paired tests
Report calibration and PPV at realistic prevalence	Confidence and false-alarm cost

4. Conclusion

Frozen IndoBERT embeddings with a Random Forest separated hoax from non-hoax political news in a collection of 2,262 Indonesian articles with 91.61% accuracy (default) and 92.72% (tuned) on 453 test articles, with AUC near 0.975. The accuracy intervals are about five points wide and overlap almost completely. A net gain of five articles from tuning cannot be significant in an exact McNemar test whatever the overlap of errors, so the benefit of tuning is not established. False alarms outnumber misses, and at realistic hoax prevalence they would dominate the flagged articles. Because hoax and genuine articles come from different outlets, the scores show separability within this collection and not detection of falsehood. The six-check audit (keyword ablation, cross-source testing, a lexical baseline, a fine-tuned comparison, repeated splits, and a near-duplicate check) is the route from this baseline to a conclusion about hoax detection.

AI Disclosure Statement

The authors used Generative AI and AI-Assisted Technologies throughout the preparation of this manuscript: to draft and revise portions of the text, to write and run the scripts behind the experiments and statistical analyses, to generate the figures included in the revision, and to format the document from the authors' own research data and source code. The authors verified the scripts and their outputs, reviewed the resulting text, cross-checked every reference against its original source, and accept full responsibility for the content of this publication.

Author Contributions Statement

Daniel Irwansyah: Conceptualization, Methodology, Software, Investigation, Formal Analysis, Data Curation, Visualization, Writing – Original Draft.
Anita Ahmad Kasim: Supervision, Validation, Writing – Review and Editing.
Hajra Rasmita Ngemba: Validation, Writing – Review and Editing.

References

[1] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, Mar. 2018, doi: 10.1126/science.aap9559.

[2] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explor. Newsl.*, vol. 19, no. 1, pp. 22–36, Sep. 2017, doi: 10.1145/3137597.3137600.

[3] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP," in *Proc. 28th Int. Conf. Comput. Linguist. (COLING)*, 2020, pp. 757–770, doi: 10.18653/v1/2020.coling-main.66.

[4] B. Wilie et al., "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," in *Proc. 1st Conf. Asia-Pacific Chapter Assoc. Comput. Linguist. and 10th Int. Joint Conf. Natural Lang. Process.*, 2020, pp. 843–857, doi: 10.18653/v1/2020.aacl-main.85.

[5] 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 Amer. Chapter Assoc. Comput. Linguist.: Human Lang. Technol. (NAACL-HLT)*, vol. 1, Minneapolis, MN, USA, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[6] C. Jocelynne, I. G. N. L. Wijayakusuma, and L. P. I. Harini, "Detection of political hoax news using fine-tuning IndoBERT," *J. Appl. Informatics Comput.*, vol. 9, no. 2, pp. 354–360, Mar. 2025, doi: 10.30871/jaic.v9i2.8989.

[7] D. Y. Yefferson, V. Lawijaya, and A. S. Girsang, "Hybrid model: IndoBERT and long short-term memory for detecting Indonesian hoax news," *IAES Int. J. Artif. Intell.*, vol. 13, no. 2, pp. 1913–1924, Jun. 2024, doi: 10.11591/ijai.v13.i2.pp1913-1924.

[8] L. Riyandari, S. S. Andiko, S. Delimasari, and S. Briandoko, "Optimization of random forest model via GridSearchCV for hoax news detection," in *Proc. Nat. Conf. Elect. Eng., Informatics, Ind. Technol., Creative Media (CENTIVE)*, 2025, pp. 208–213, doi: 10.20895/centive.v2025i1.545.

[9] R. Geirhos et al., "Shortcut learning in deep neural networks," *Nat. Mach. Intell.*, vol. 2, no. 11, pp. 665–673, Nov. 2020, doi: 10.1038/s42256-020-00257-z.

[10] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. Bowman, and N. A. Smith, "Annotation artifacts in natural language inference data," in *Proc. 2018 Conf. North Amer. Chapter Assoc. Comput. Linguist.: Human Lang. Technol.*, vol. 2, 2018, pp. 107–112, doi: 10.18653/v1/N18-2017.

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

[12] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. 2019 Conf. Empirical Methods Natural Lang. Process. and 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 3980–3990, doi: 10.18653/v1/D19-1410.

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

[14] G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *J. Mach. Learn. Res.*, vol. 11, pp. 2079–2107, 2010.

[15] L. D. Brown, T. T. Cai, and A. DasGupta, "Interval estimation for a binomial proportion," *Stat. Sci.*, vol. 16, no. 2, pp. 101–133, May 2001, doi: 10.1214/ss/1009213286.

[16] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, Jun. 1947, doi: 10.1007/BF02295996.

[17] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998, doi: 10.1162/089976698300017197.

[18] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proc. 22nd Int. Conf. Mach. Learn. (ICML)*, 2005, pp. 625–632, doi: 10.1145/1102351.1102430.

[19] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, 2017, arXiv:1706.04599.

[20] R. J. Connor, "Sample size for testing differences in proportions for the paired-sample design," *Biometrics*, vol. 43, no. 1, p. 207, Mar. 1987, doi: 10.2307/2531961.