Journal of Artificial Intelligence and Legal Technology

SAH PUBLISHER

e-ISSN 3123-786X
Vol. 2, No. 2, August 2026

Explainable XGBoost for Indonesian Hoax Detection under the Electronic Transactions Law

Rizka Nadialif[1]*, Landung Sudarmana[2], Selvi Dwi Hartiyani[3], Sapriani Gustina[4], Ardy Wicaksono[5], Agatha Pricillia Sekar Tamtomo[6]

[1,2,3,4] Information Technology, Universitas Proklamasi 45, Sleman, Indonesia
[5] Computer Science, Universitas Sugeng Hartono, Sukoharjo, Indonesia
[6] Business Digital, Universitas Sugeng Hartono, Sukoharjo, Indonesia
Email: ardywicaksono166@gmail.com

Abstract. The rapid circulation of misleading information in digital spaces creates a need for screening tools that are accurate, transparent, and suitable for human review. This study develops a reproducible Indonesian hoax-detection pipeline and examines whether model explanations can support cautious legal review. The experiment uses a political-hoax text corpus with fixed training, validation, and test splits. The primary classifier combines word- and character-level TF-IDF features with XGBoost. A frozen-encoder IndoBERT-Lite implementation is included solely as a small CPU-based feasibility pilot to verify the data, tokenizer, and inference workflow; it is not a competitive baseline and its score is not used for model ranking. TreeSHAP summarizes global and local feature contributions, and LIME is used to inspect borderline predictions. On the held-out test set, XGBoost achieved 0.9725 accuracy, 0.9724 macro-F1, 0.9957 ROC-AUC, and a 0.0235 Brier score. The CPU pilot reached a validation macro-F1 of 0.3343 under its deliberately restricted setting. The most influential features included source and article-genre markers such as “baca juga,” “referensi,” “Kompas,” “Facebook,” and “foto hoaks,” indicating that the model may learn publisher or writing-style shortcuts in addition to claim-related signals. The audit also identified normalized duplicate overlap across the training-validation and training-test splits, while the released label-to-class mapping could not be verified. These issues may inflate apparent generalization and prevent class-specific outputs from being interpreted as verified truth-status findings. The resulting system should therefore support triage, explanation, and documentation by trained reviewers, not serve as a standalone basis for determining the truth of a claim or establishing an Electronic Information and Transactions Law violation.

Keywords: Indonesian hoax detection, IndoBERT, XGBoost, Electronic Information and Transactions Law, source-bias audit
Received Jun 2026 / **Revised** Aug 2026 / **Accepted** Aug 2026

This work is licensed under a Creative Commons Attribution 4.0 International License.

INTRODUCTION

The rapid circulation of misleading information has made digital content assessment a practical problem for journalists, fact-checkers, public institutions, and citizens. Indonesian political narratives are particularly difficult to screen because a single post can combine a true event, an old photograph, a partial quotation, and an unsupported interpretation. The same text may also be copied across websites and social-media accounts, making apparent volume a poor indicator of independent evidence. A useful detection system must therefore distinguish repeated wording from genuinely new claims and must remain cautious when the available label represents a dataset convention rather than a legal finding. The legal context increases the cost of an unexamined prediction. Law No. 1 of 2024 amends the Electronic Information and Transactions Law and addresses misleading information in electronic transactions. However, a statutory provision contains elements concerning conduct, context, consequence, intent, and evidence that cannot be reduced to a binary text label. A high classifier score may help a reviewer locate material for examination, but it cannot establish that a person acted knowingly, that a particular harm occurred, or that the legal threshold for an offense has been met. The Constitutional Court's interpretation of public unrest further illustrates why context matters: a model that recognizes alarming words cannot decide whether a legal element is satisfied. For this reason, the present study treats machine learning as a review-support instrument and makes its limitations explicit.

Recent Indonesian studies show that transformer representations and tree-based classifiers can perform well on benchmark hoax data. IndoBERT-based systems benefit from contextual token representations and have

been examined for imbalance handling and Indonesian political news classification [1], [2], [3], [4], [5], [6]. XGBoost systems using TF-IDF features are attractive when computing resources are limited because they train quickly, are easy to serialize, and can expose feature contributions [4], [7]. IndoBERTweet extends the language-model perspective to Indonesian Twitter language, which is relevant when a future system must process short posts and repost conventions [8]. These results establish a useful predictive baseline, but they do not by themselves show whether the model is learning claim content or publisher-specific style. Explainability research addresses part of this problem. SHAP assigns additive feature contributions relative to a model baseline, while LIME constructs a local approximation around one prediction. Surveys and empirical studies recommend using both global and local views, testing explanation stability, and reporting the limits of attribution methods [9], [10], [6], [11]. An explanation can reveal that the model relies on words such as "referensi," a URL marker, a publisher name, or a phrase such as "baca juga." Such a finding is valuable precisely because it may expose a shortcut rather than confirm a fact. In a legal-support setting, the explanation must be readable, reproducible, and clearly separated from the evidence reviewed by a human examiner. The literature therefore points to a combined requirement: strong predictive evaluation must be accompanied by leakage audits, source-bias analysis, and an operational account of how explanations are reviewed.

The unresolved issue is the absence of an end-to-end, auditable workflow that connects Indonesian hoax classification, explanation, and legal caution. Many reports emphasize a single accuracy or F1 value, although near-duplicate claims across train and test splits can inflate that value. Other reports present feature importance without checking whether the features describe the truth of a claim, the identity of its source, or the genre of an article [7]. The problem is compounded when the label dictionary is not machine-readable: a model may be highly consistent with the annotation convention while the direction of the "hoax" label remains undocumented. These weaknesses limit the use of published scores in real review environments. This study asks three practical questions. First, how well does a reproducible TF-IDF and XGBoost pipeline classify the supplied Indonesian political-hoax benchmark when the repository split is preserved? Second, what lexical and source-style signals drive global and local explanations, and are those signals plausible indicators of claim status or possible shortcuts? Third, how should the resulting score and explanation be documented so that they support, rather than replace, human assessment under the Electronic Information and Transactions Law? The contribution is consequently methodological and operational rather than purely competitive. The study reports a fixed experimental protocol, a CPU-oriented IndoBERT-Lite CPU pilot, class-wise and aggregate metrics, a normalized overlap audit, SHAP/LIME reason codes, and a governance workflow that records model versions, text hashes, reviewer decisions, and override rationales. By separating statistical prediction from legal judgment, the paper defines a safer boundary for explainable machine learning in hoax-content screening. Related work on political hate-speech detection, real-time fake-news screening, Indonesian disinformation, and evidence-aware explanations is also relevant to the proposed workflow [12], [13], [11].

METHODS

The study follows a leakage-aware experimental sequence. First, the three repository CSV files are loaded without changing their original split assignment. Second, text is normalized using whitespace cleanup and deterministic replacement of URLs and usernames. Third, the training split is transformed into sparse word and character TF-IDF representations; the vectorizers are fitted only on training text and then applied unchanged to validation and test text. Fourth, XGBoost, the sole model evaluated as the primary classifier, is trained with a fixed random seed and early stopping on the validation split. Fifth, the held-out test split is evaluated with threshold-based and ranking-based metrics. Finally, global TreeSHAP contributions and local LIME explanations are exported together with model probabilities and text hashes. Separately, a frozen IndoBERT-Lite encoder with a linear head is run on 500 stratified training examples for one epoch on CPU. This deliberately constrained run is an implementation feasibility pilot only: it verifies the tokenizer, data path, checkpoint compatibility, and explanation interface. It is not a competitive baseline, is not included in the held-out benchmark comparison, and cannot support claims about the relative performance of XGBoost and a fully fine-tuned transformer. No test instance is used for feature fitting, hyperparameter selection, or explanation calibration. The reported configuration records the dataset version, checkpoint name, seed, feature limits, and evaluation outputs.

TOPSIS Method Stages

TOPSIS is applied as a transparent multi-criteria ranking stage for comparing candidate model configurations after predictive evaluation. The alternatives are the evaluated XGBoost and IndoBERT

configurations, while the criteria are macro-F1, ROC-AUC, explanation stability, and computational cost. Macro-F1 and ROC-AUC are benefit criteria; cost is treated as a cost criterion. The ranking is not a legal decision and is used only to document why one configuration is preferable for a review-support prototype.

$$r_{ij} = x_{ij} / \sqrt{(\Sigma\, x_{ij}^{2})} \quad (1)$$

Equation (1). Vector normalization of criterion j across m candidate alternatives.

In (1), X_{ij} is the observed value of criterion *j* for alternative *i*, r_{ij} is its normalized value, *m* is the number of alternatives, and the denominator is the Euclidean norm of criterion *j*. Normalization removes differences in measurement units before criteria are combined.

$$v_{ij} = w_j\, r_{ij}, \;\; with \;\; \Sigma\, w_j = 1 \quad (2)$$

Equation (2). Weighted normalized decision matrix.

In (2), v_{ij} is the weighted normalized value, w_j is the pre-specified weight of criterion *j*, and *n* is the number of criteria. The weights are fixed before inspecting the test result; in this study predictive quality receives the largest share, while stability and cost prevent a high-scoring but opaque or impractical model from dominating the ranking.

$$A_j^{+} = max_i(v_{ij})\;[benefit], min_i(v_{ij})\;[cost]; \;\; A_j^{-} = min_i(v_{ij})\;[benefit], max_i(v_{ij})\;[cost] \quad (3)$$

Equation (3). Positive and negative ideal solutions for benefit and cost criteria.

The positive ideal solution *A+* contains the most desirable value for each criterion, whereas the negative ideal solution *A-* contains the least desirable value. For predictive metrics, larger values are preferred; for computational cost, a smaller value is preferred.

$$D_i^{+} = \sqrt{\Sigma(v_{ij} - A_j^{+})^2}, \;\; D_i^{-} = \sqrt{\Sigma(v_{ij} - A_j^{-})^2} \quad (4)$$

Equation (4). Euclidean distance of alternative i from the two ideal solutions.

In (4), D_i^{+} and D_i^{-} measure how far alternative i is from the positive and negative ideals, respectively. *A* strong alternative should be close to A^{+} and far from A^{-}.

$$C_i = D_i^{-}/(D_i^{+} + D_i^{-}), \;\; 0 \le C_i \le 1 \quad (5)$$

Equation (5). TOPSIS closeness coefficient used to rank candidate configurations.

In (5), C_i is the relative closeness of alternative *i* to the positive ideal; a larger value indicates a more balanced configuration under the selected criteria. The resulting rank is reported with the underlying metrics and explanation audit so that reviewers can reproduce the decision rather than treating the coefficient as an unexplained recommendation.

Research Stages

The study proceeds through a fixed sequence: dataset acquisition and overlap audit, deterministic preprocessing, model training, held-out evaluation, explanation generation, and human/legal-context review. Each stage produces a versioned artifact so that a reviewer can trace a reported result back to the source split, model configuration, and explanation record. The flowchart in Figure 1 summarizes this sequence; it is a methodological roadmap rather than a claim that automated output replaces expert judgment.

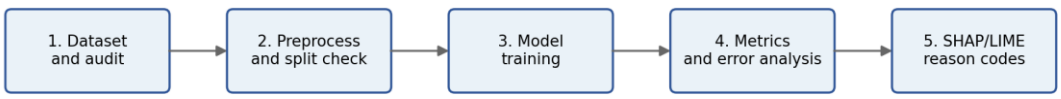


Figure 1. Research stages from dataset audit to human/legal-context review.

Dataset and label caveat

The Indonesia Political Hoax Dataset v1 is the primary corpus because it contains an explicit train-validation-test organization and a large Indonesian political-news sample [14]. The training set contains 15,187 instances, the validation set contains 1,561 instances, and the test set contains 1,562 instances. However, the release does not provide a machine-readable dictionary that verifies which numeric class corresponds to a particular truth status. Consequently, the present study can measure consistency with the released numeric annotations but cannot safely name class 0 or class 1 as "true" or "hoax." This is a measurement-validity problem, not merely a reporting inconvenience: precision, recall, false-positive, and false-negative labels cannot be given substantive truth-status meanings until the mapping is confirmed. In addition, normalized duplicate overlap across the supplied splits means that some evaluation items may reproduce claims seen during training. The reported metrics therefore describe performance on the repository split and may overestimate performance on genuinely novel claims. The corpus is treated as a text-classification benchmark, not as an adjudicated evidence repository. The proposed Hugging Face resource was not merged because it is an Indonesian hate-speech superset rather than a hoax dataset and its access is gated. Any future combination would require a new annotation protocol, documentation of annotator agreement, and an ethics review addressing privacy and potential harm to named individuals or groups.

RESULTS AND DISCUSSION

The Results and Discussion section first reports the primary XGBoost findings, then examines errors, explanations, legal interpretation, and threats to validity. Detailed model settings are provided in Methods and are not repeated here. The IndoBERT-Lite run is reported only as evidence that the CPU implementation path executed successfully; because it used a frozen encoder, one epoch, and only 500 stratified training examples, it is not a competitive baseline and is excluded from comparative ranking.

The primary evaluation uses the untouched test split once after validation-based early stopping. Accuracy and balanced accuracy summarize thresholded performance; class-wise precision, recall, and F1 expose asymmetric errors; ROC-AUC assesses ranking from continuous probabilities; and the Brier score assesses probability quality. The fixed classifier threshold is an operating point for triage, not a legal risk boundary. Metrics and explanation outputs are stored separately so that the reported tables and figures can be regenerated without manual transcription.

Table 1 reports the principal XGBoost test performance. Accuracy is 0.9725, balanced accuracy is 0.9722, macro-F1 is 0.9724, ROC-AUC is 0.9957, and the Brier score is 0.0235. The small gap between accuracy and balanced accuracy indicates similar aggregate performance across the numeric classes, while the ROC-AUC and Brier score respectively indicate strong ranking and probability quality on this supplied split. These values are benchmark-positioning results for the TF-IDF-XGBoost pipeline only. They do not demonstrate that the model understands factual truth, generalizes to unseen claims or publishers, or satisfies any legal element. The separate CPU IndoBERT-Lite feasibility pilot obtained a validation macro-F1 of 0.3343 under its restricted setup. That number is reported for implementation transparency, not as evidence that XGBoost outperforms IndoBERT. A competitive transformer comparison would require full-data fine-tuning, an equivalent test protocol, multiple seeds, confidence intervals, and matched preprocessing and selection conditions.

Table 1. Predictive performance and data-quality audit

Model	Accuracy	Macro-F1	ROC-AUC	Audit
TF-IDF + XGBoost	0.9725	0.9724	0.9957	T-V 101; T-Test 85
IndoBERT pilot	-	val 0.3343	-	CPU pilot
Split audit	-	-	-	V-Test 3

The aggregate score must be interpreted together with the unresolved label mapping and duplicate audit. Normalized overlap includes 101 training-validation matches and 85 training-test matches. Test items that duplicate training claims are not independent evidence of generalization, so their correct classification can inflate accuracy, macro-F1, ROC-AUC, and apparent calibration. Because a deduplicated re-estimate is not yet available, the magnitude of inflation cannot be quantified and the reported 0.9725 accuracy should be treated as an upper-bound description of performance on the supplied split, not an estimate for new real-world claims. The unresolved numeric-label mapping creates a separate interpretive constraint: the two classes can be evaluated statistically, but false-positive and false-negative counts cannot safely be translated into errors about "hoax" versus "true" content. Qualitative error analysis nevertheless identifies plausible mechanisms. Legitimate reporting may be misclassified when it quotes a false claim for debunking, uses

sensational wording, or reproduces a familiar headline; misleading content may be missed when it uses neutral language, novel spelling, or an unseen source. Future work should report claim-level deduplicated metrics and error slices by source, publication year, claim length, and URL presence.

Figure 2 shows the global TreeSHAP profile. It reveals what the model treats as predictive, not which facts a legal reviewer should accept as evidence. Highly ranked terms such as kata, baca juga, referensi, URL, politik, foto, Facebook, partai, Kompas, and foto hoaks often describe source, article format, or editorial convention. Their prominence is therefore a warning of possible shortcut learning rather than confirmation that a claim is false. A legal reviewer should use the display to formulate verification questions: Is the model reacting to the alleged claim itself, to quoted material, to a publisher name, or to fact-checking language? The reviewer must then return to the complete text, provenance, publication date, corroborating sources, and any admissible digital evidence. Local LIME explanations can help inspect borderline cases, but they do not establish intent, knowledge, causation, harm, authorship, or the satisfaction of a statutory element. Explanations should be stress-tested through deletion, paraphrase, and controlled-perturbation checks before they are relied upon in an operational workflow.

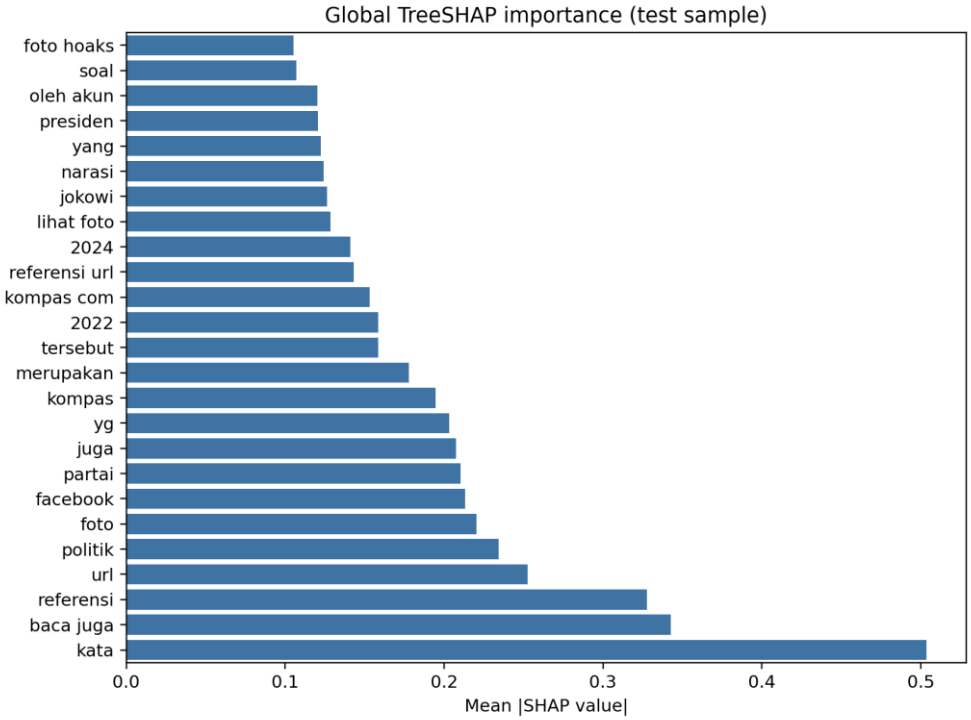


Figure 2. Global TreeSHAP importance for the XGBoost test sample

SHAP and LIME are diagnostic aids, not legal reasons or independent evidence. A positive SHAP value means that a feature moves the model output toward a numeric class relative to the model baseline; it does not show that the feature caused the underlying event, that the statement is factually false, or that a person acted with legally relevant intent. LIME approximates one prediction in a perturbed neighborhood, so its token weights may change with the sampling seed, tokenization, Indonesian morphology, code switching, or social-media syntax. Legal reviewers should therefore read both positive and negative contributions, inspect the exact source passage, and record whether a feature reflects claim content, quotation, source identity, genre, or preprocessing artifacts. Every explanation record should retain the model and data version, probability, method and parameters, original text span, reviewer assessment, and any override rationale. If an explanation is unstable, dominated by source markers, or inconsistent with the full text, it should be marked inconclusive and must not be converted into a reason for adverse action. Even a stable explanation describes model behavior only; factual verification and legal analysis remain separate human tasks.

The proposed legal-support workflow has three operational stages. During triage, the classifier prioritizes items for human attention and presents a calibrated probability with a warning that the output is a statistical score. During review, the examiner checks the full text, source history, publication time, corroborating evidence, and the model's top positive and negative features. During documentation, the system stores the text hash, model and data versions, probability, explanation output, reviewer decision, and an override reason. This audit trail supports accountability without converting an explanation into a legal conclusion. In particular, a high hoax score does not establish intent, knowledge, causation, consumer loss, or the public-order context required by a legal provision. A lawful decision must remain grounded in applicable legislation, judicial interpretation, evidence standards, and due process. The design also requires access controls, retention limits, and a mechanism for correcting false positives. These safeguards are essential because automated suspicion can amplify existing source or political biases.

The tool should assist trained reviewers, and every operational deployment should be preceded by a domain-specific validation with legal and fact-checking experts. Governance should be built into the system rather than added after deployment. Access to raw text and user identifiers should be separated from access to aggregate model diagnostics. Retention rules should specify when text, probabilities, and explanation records are deleted or anonymized. Reviewers should be able to override a model recommendation and record the reason without changing the original prediction. Periodic audits should check whether error rates or top features differ across publishers, political topics, time periods, or language varieties. A change in the explanation distribution can be an early warning of domain shift even when a small labeled sample is unavailable. These controls make the model's limitations visible to the organization using it. Recent legal and AI-governance studies emphasize human oversight, digital-evidence safeguards, and caution against treating automated scores as legal conclusions [15], [16], [13], [5], [6].

Table 2 reports the class-wise confusion pattern so that the overall score can be read alongside the types of errors made by the classifier. The thresholded test predictions contain 28 false positives and 15 false negatives; the class-wise values should be interpreted together with the unresolved label mapping.

Table 2. Class-wise performance and confusion-matrix summary

Class	Precision	Recall	F1	Count
Class 0	0.9798	0.9630	0.9713	757
Class 1	0.9658	0.9814	0.9735	805
Total errors	-	-	-	43 (28 FP; 15 FN)

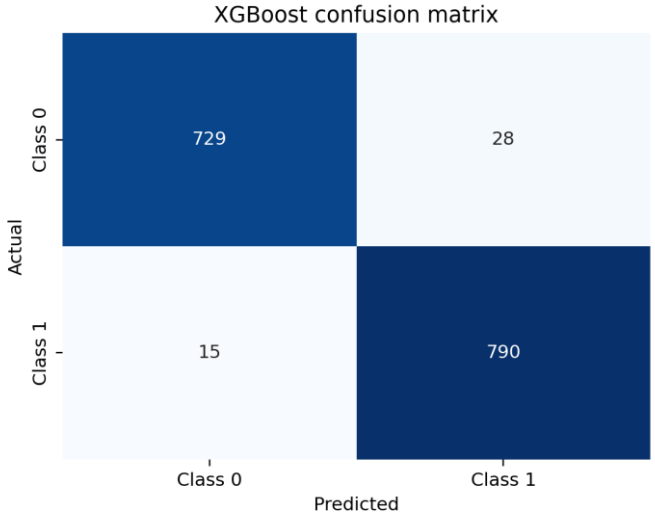


Figure 2. XGBoost confusion matrix on the held-out test set

Several threats to validity materially constrain the findings. First, because the numeric label dictionary is unresolved, the study cannot claim semantic accuracy for the categories “true” and “hoax”; class-wise outputs remain statistical descriptions of numeric labels. Second, normalized duplicate overlap includes 101 training-validation and 85 training-test matches. Exposure to repeated claims can inflate every headline metric and weaken the independence of the held-out test, especially when distinctive wording is preserved.

Without rerunning the experiment on a claim-level deduplicated split, the degree of optimism is unknown. Third, source- and genre-related SHAP features suggest that the model may exploit publisher distribution or article conventions. Performance may therefore fall on X/Twitter posts, government releases, independent fact-checks, new publishers, or later political events. Fourth, the IndoBERT-Lite result is only a constrained CPU implementation pilot and provides no competitive benchmark evidence. Fifth, probability calibration and explanation stability have not been validated on deduplicated external data or assessed by independent annotators. Finally, classification metrics do not capture legal and social harms such as chilling effects, selective enforcement, privacy exposure, reputational injury, and automation bias. Collectively, these limitations mean that the present results support a reproducible prototype and audit framework, not deployment readiness or legal decision-making.

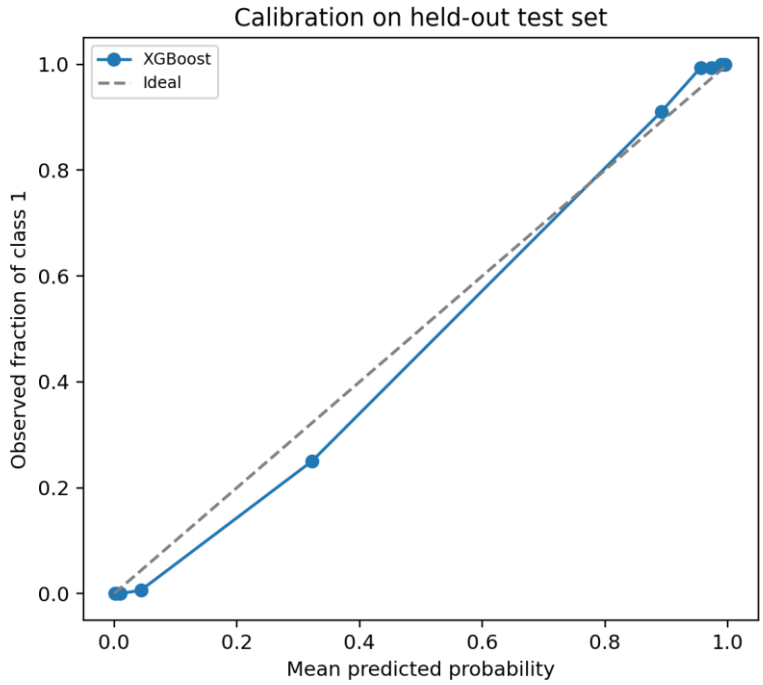


Figure 3. XGBoost probability calibration on the held-out test set

A stronger follow-up study should first verify and publish the label dictionary, document annotation provenance and agreement, and construct claim-level deduplicated train, validation, and test splits. Reporting performance before and after deduplication would quantify the inflation caused by repeated claims. Temporal and cross-source tests are also required because the current article-style corpus may not represent short, multimodal, or conversational misinformation. Indonesian social-media language includes abbreviations, code switching, emojis, repost markers, and spelling variation that are not fully captured by conventional TF-IDF features. The model may therefore be less reliable on X/Twitter data than on the benchmark corpus. Moreover, a fact-checking label does not map automatically to the elements of a criminal offense. These measurement and domain differences should appear in deployment documentation and be explained to reviewers before any model score is displayed.

Global TreeSHAP summarizes aggregate feature contributions, whereas local LIME is used to inspect borderline cases. Before operational deployment, explanations should be tested for stability across seeds, paraphrases, token deletion, and controlled perturbations, and their usefulness should be assessed independently by legal and fact-checking reviewers. Reproducibility requires archiving dataset checksums, split files, preprocessing configuration, vectorizer vocabulary, model parameters, random seeds, and the IndoBERT checkpoint version. Explanation records should link to a stable text hash rather than an undisclosed personal identifier. A data card should describe collection boundaries, label provenance, duplicate policy, source imbalance, and foreseeable misuse. The planned validation program should combine temporal testing, cross-source testing, claim-level deduplication, and independent ratings of explanation sufficiency, consistency, and usefulness. X/Twitter data should be added only under a lawful collection and anonymization protocol. Any future competitive IndoBERT evaluation must use the same

cleaned splits as XGBoost, full fine-tuning, comparable model-selection rules, and confidence intervals across multiple seeds.

These steps convert the current pilot into an auditable research program rather than a one-off demonstration. The final manuscript should provide a compact algorithmic description rather than source-code listings. In prose, the algorithm is: load the fixed splits; fit preprocessing on training text; transform validation and test text; train XGBoost with early stopping; compute held-out metrics; generate global TreeSHAP and local LIME outputs; attach hashes and versions; and export the audit files. This description is sufficient for independent implementation while avoiding the inclusion of program listings prohibited by the journal instructions. A flow diagram can show the same sequence from data intake to human review and legal documentation.

CONCLUSION

The experiment answers the central question cautiously. On the supplied held-out split, the TF-IDF-XGBoost pipeline achieved 97.25% accuracy, 97.24% macro-F1, 99.57% ROC-AUC, and a Brier score of 0.0235, with 28 false positives and 15 false negatives in numeric-label terms. TreeSHAP and LIME expose lexical and source-style cues that drive model output, but they explain model behavior rather than provide factual or legal evidence. IndoBERT-Lite was included solely as a constrained CPU implementation feasibility pilot; it is not a competitive baseline, and no conclusion about relative model superiority is drawn from its validation score. The unresolved label mapping prevents semantic interpretation of the numeric classes, while duplicate overlap across splits may inflate the reported performance and weakens claims of generalization. The evidence therefore does not show that the model determines the truth of a claim, performs reliably on novel real-world sources, or can establish an Electronic Information and Transactions Law violation. Its defensible role is limited to prioritizing material for trained human review and documenting model behavior. Before operational use, the labels must be verified, claim-level duplicates removed, performance re-estimated on temporal and cross-source data, explanation stability independently assessed, and privacy, appeal, and legal-oversight safeguards implemented. Subject to these conditions, explainable machine learning may support evidence organization and review efficiency without replacing factual and legal judgment.

REFERENCES

[1] M. Fathin, Y. Sibaroni, and S. Prasetyowati, “Handling Imbalance Dataset on Hoax Indonesian Political News Classification using IndoBERT and Random Sampling,” Jurnal Media Informatika Budidarma, 8(1), 2024. doi: https://doi.org/10.30865/mib.v8i1.7099.

[2] L. A. Pekandi, R. G. Widjaja, A. Ananta, J. Harefa, and K. Jingga, “Evaluating IndoBERT for Indonesian Hoax News Detection: A Comparative Study with Ensemble and CNN-LSTM Models,” Procedia Computer Science, vol. 269, pp. 1625–1633, 2025. doi: https://doi.org/10.1016/j.procs.2025.09.105.

[3] R. Kozik, M. M. Ficco, A. Pawlicka, M. Pawlicki, F. Palmieri, and M. Choraś, “When Explainability Turns into a Threat—Using xAI to Fool a Fake News Detection Method,” Computers & Security, vol. 137, 103599, 2024. doi: https://doi.org/10.1016/j.cose.2023.103599.

[4] V. Prisscilya and A. S. Girsang, “Classification of Indonesia False News Detection Using BERTopic and IndoBERT,” Jurnal Indonesia Sosial Teknologi, vol. 5, no. 8, pp. 3913–3931, 2024. doi: https://doi.org/10.59141/jist.v5i8.1310.

[5] S. M. T. Situmeang, D. Pudjiastuti, and Sutarjo, “Adequacy of Indonesian Cybercriminal Law Against Generative AI-Based Social Engineering Attacks,” Res Nullius Law Journal, vol. 6, no. 1, pp. 82–97, 2024. doi: https://doi.org/10.34010/rnlj.v6i1.19927.

[6] N. P. R. Adiati, D. F. Priambodo, Girinoto, S. Indarjani, A. Rizal, A. Prayoga, and Y. Beatrix, “Comparative Study of Predictive Models for Hoax and Disinformation Detection in Indonesian News,” International Journal of Advances in Intelligent Informatics, vol. 10, no. 3, pp. 504–516, 2024. doi: https://doi.org/10.26555/ijain.v10i3.878.

[7] J. Chen, “Identification and Analysis of Real and Fake News by XGBoost Algorithm of Machine Learning,” Applied and Computational Engineering, vol. 40, no. 1, pp. 255–262, 2024. doi: https://doi.org/10.54254/2755-2721/40/20230661.

[8] P. A. Bhagaskara S., S. S. Prasetiyowati, and Y. Sibaroni, “Hoax Detection of Indonesian News Media on Twitter Using IndoBERT with Word Embedding Word2Vec,” Jurnal Media Informatika Budidarma, vol. 7, no. 3, 2023. doi: https://doi.org/10.30865/mib.v7i3.6367.

[9] C. J. L. Tobing, I. G. N. Lanang Wijayakusuma, and L. P. I. Harini, “Perbandingan Kinerja IndoBERT dan MBERT untuk Deteksi Berita Hoaks Politik dalam Bahasa Indonesia,” Jurnal Sains dan Teknologi, vol. 14, no. 1, pp. 114–123, 2025. doi: https://doi.org/10.23887/jstundiksha.v14i1.92126.

[10] M. Ahmed et al., “Explainable Text Classification Model for COVID-19 Fake News Detection,” JISIS, 12(2), 2022. doi: https://doi.org/10.22667/JISIS.2022.05.31.051.

[11] S.-Y. Chien, C.-J. Yang, and F. Yu, “XFlag: Explainable Fake News Detection Model on Social Media,” International Journal of Human–Computer Interaction, vol. 38, no. 18–20, pp. 1808–1827, 2022. doi: https://doi.org/10.1080/10447318.2022.2062113.

[12] C. Zhang et al., “A computational approach for real-time detection of fake news,” Expert Systems with Applications, 221, 119656, 2023. doi: https://doi.org/10.1016/j.eswa.2023.119656.

[13] A. S. Dwiandari and R. Arifin, “Criminal Law Enforcement on Digital Identity Misuse in AI Era for Commercial Interests in Indonesia,” Indonesian Journal of International Clinical Legal Education, vol. 7, no. 1, 2025. doi: https://doi.org/10.15294/iccle.v7i1.25525.

[14] N. N. Hidayati, A. Santosa, and S. Shaleha, “Indonesia Political Hoax Dataset,” Mendeley Data, v1, 2025. doi: https://doi.org/10.17632/mzx5rd397v.1.

[15] E. Mutmainnah, “The Phenomenon of Artificial Intelligence Hallucination,” Jurnal Rechtsvinding, 13(2), 2024. doi: https://doi.org/10.33331/rechtsvinding.v13i2.1781.

[16] S. N. Dachlan, D. E. S. Karauwan, and N. Lahangatubun, “The Role of Artificial Intelligence in Law Enforcement: Towards a More Accurate and Efficient Justice System,” Sinergi International Journal of Law, vol. 2, no. 3, pp. 198–207, 2024. doi: https://doi.org/10.61194/law.v2i3.157.