JCOSIS (Journal Computer Science and Information Syetem)
E-ISSN : 3032-5676
Volume : 02 | Nomor : 02 | Tahun 2025 | Hal 07-13
DOI : https://doi.org/10.61567

Explainable Transformer-Based Fake News Detection pada Media Sosial Indonesia Menggunakan IndoBERT dan SHAP

Fitri Purwaningtias[1]*

[1]Program Studi Informatika, Fakultas Teknik,
Program Studi Sistem Informasi, Fakultas Ilmu Komputer,
Universitas Bina Darma, Indonesia
fitri.purwaningtias@binadarma.ac.id

Abstrak.
Tujuan : Perkembangan media sosial di Indonesia meningkatkan penyebaran informasi secara cepat, namun juga memicu meningkatnya penyebaran berita palsu (*fake news*) yang berdampak pada opini publik dan stabilitas sosial. Penelitian ini bertujuan membangun sistem deteksi fake news berbasis *Explainable Transformer* menggunakan IndoBERT dan SHAP pada media sosial Indonesia. Dataset penelitian diperoleh dari *Indonesian Hoax News Dataset*, Kaggle *Indo Fake News Dataset*, dan data Twitter/X Indonesia dengan total 18.450 data teks.
Metode/Design/Pendekatan: Metode penelitian meliputi preprocessing teks, tokenisasi, stemming, pelatihan model IndoBERT, evaluasi performa menggunakan confusion matrix, serta interpretasi model menggunakan SHAP (*SHapley Additive exPlanations*).
Hasil/Temuan: Hasil penelitian menunjukkan bahwa model IndoBERT memperoleh *accuracy* sebesar 94,21%, *precision* sebesar 93,84%, *recall* sebesar 94,67%, *F1-score* sebesar 94,25%, dan ROC-AUC sebesar 95,11%. Visualisasi SHAP menunjukkan bahwa token seperti "viral", "sebarkan", dan "bocor" memiliki kontribusi tinggi terhadap prediksi *fake news*, sedangkan token "resmi" dan "klarifikasi" berkontribusi terhadap prediksi real news.
Kebaharuan/Originalitas/Nilai: Penelitian ini membuktikan bahwa integrasi IndoBERT dan Explainable AI mampu meningkatkan akurasi sekaligus transparansi sistem deteksi berita palsu pada media sosial

Keywords: *Fake News Detection*, IndoBERT, *Explainable AI*, SHAP, *Natural Language Processing*, Media Sosial Indonesia.

Abstract.
Purpose: *The rapid growth of social media usage in Indonesia has accelerated information dissemination, but it has also increased the spread of fake news that affects public opinion and social stability. This study aims to develop an Explainable Transformer-based fake news detection system using IndoBERT and SHAP for Indonesian social media. The dataset was collected from the Indonesian Hoax News Dataset, Kaggle Indo Fake News Dataset, and Twitter/X Indonesia, consisting of 18,450 text data.*
Methods/Study design/approach: *The research methodology includes text preprocessing, tokenization, stemming, IndoBERT model training, performance evaluation using confusion matrix metrics, and model interpretation using SHAP (SHapley Additive exPlanations).*
Result/Findings: *The experimental results show that the IndoBERT model achieved an accuracy of 94.21%, precision of 93.84%, recall of 94.67%, F1-score of 94.25%, and ROC-AUC of 95.11%. SHAP visualization indicates that tokens such as "viral," "spread," and "leaked" contribute significantly to fake news predictions, while tokens such as "official" and "clarification" contribute to real news predictions.*
Novelty/Originality/Value: *This study demonstrates that the integration of IndoBERT and Explainable Artificial Intelligence improves both the accuracy and transparency of fake news detection systems on Indonesian social media.*

Keywords*: Fake News Detection, IndoBERT, Explainable AI, SHAP, Natural Language Processing, Indonesian Social Media.*

Article history:

Received, 2026-05-09
Revised, 2026-05-30
Accepted, 2026-05-30

*Corresponding author.
Fitri Purwaningtias
Email addresses: fitri.purwaningtias@binadarma.ac.id

PENDAHULUAN

Perkembangan media sosial di Indonesia telah meningkatkan kecepatan distribusi informasi secara signifikan. Platform seperti Facebook, Instagram, TikTok, dan X memungkinkan masyarakat untuk memperoleh dan menyebarkan informasi secara real-time tanpa melalui proses verifikasi yang ketat. Kondisi ini berdampak pada meningkatnya penyebaran berita palsu (fake news) atau hoaks yang dapat memengaruhi opini publik, stabilitas sosial, hingga kondisi politik nasional [1]. Fenomena hoaks di Indonesia semakin kompleks seiring meningkatnya penggunaan teknologi kecerdasan buatan generatif dan *deepfake* yang membuat informasi palsu menjadi lebih sulit dibedakan dari informasi valid [2]. Diskusi publik di media sosial juga menunjukkan meningkatnya kekhawatiran masyarakat terhadap penyalahgunaan AI dalam penyebaran informasi palsu dan penipuan digital [3].

Deteksi berita palsu menjadi salah satu topik penting dalam bidang *Natural Language Processing* karena karakteristik bahasa pada media sosial yang tidak terstruktur, mengandung bahasa informal, singkatan, *code-mixing*, serta konteks semantik yang ambigu. Pendekatan tradisional berbasis machine learning seperti *Support Vector Machine* (SVM), *Naïve Bayes*, dan *Logistic Regression* masih memiliki keterbatasan dalam memahami konteks linguistik yang kompleks pada bahasa Indonesia [4]. Oleh karena itu, pendekatan berbasis *deep learning* khususnya arsitektur Transformer mulai banyak digunakan karena mampu menangkap representasi kontekstual teks secara lebih baik dibandingkan metode konvensional [5].

Model berbasis Transformer seperti BERT telah menunjukkan performa unggul dalam berbagai tugas klasifikasi teks termasuk deteksi hoaks. Pada konteks bahasa Indonesia, model IndoBERT menjadi salah satu model yang paling banyak digunakan karena telah dilatih menggunakan korpus bahasa Indonesia dalam jumlah besar sehingga mampu memahami karakteristik linguistik lokal dengan lebih baik [6]. Beberapa penelitian menunjukkan bahwa IndoBERT mampu menghasilkan akurasi tinggi dalam klasifikasi berita palsu berbahasa Indonesia. Penelitian terbaru menunjukkan bahwa IndoBERT memperoleh akurasi hingga 92,24% dan mengungguli model CNN-LSTM maupun ensemble classifier pada tugas deteksi hoaks Indonesia [7]. Selain itu, fine-tuning IndoBERT juga terbukti efektif dalam mendeteksi hoaks politik dengan performa klasifikasi yang stabil pada dataset media digital Indonesia [8].

Meskipun memiliki performa yang tinggi, model *Transformer* termasuk IndoBERT masih menghadapi tantangan utama berupa rendahnya interpretabilitas model atau dikenal sebagai masalah *black-box*. Pada sistem deteksi hoaks, interpretabilitas menjadi aspek penting karena keputusan model perlu dipahami oleh pengguna maupun peneliti agar hasil klasifikasi dapat dipercaya [9]. Ketika model hanya memberikan hasil prediksi tanpa penjelasan alasan keputusan, maka tingkat kepercayaan pengguna terhadap sistem menjadi rendah, khususnya pada kasus informasi sensitif seperti politik, kesehatan, dan keamanan publik [10].

Pendekatan *Explainable Artificial Intelligence* atau *Explainable AI* berkembang untuk mengatasi permasalahan interpretabilitas pada model *deep learning*. Salah satu metode explainability yang populer adalah SHAP yang menggunakan konsep nilai *Shapley* untuk mengukur kontribusi setiap fitur terhadap keputusan model [11]. SHAP mampu memberikan visualisasi kata atau token yang paling berpengaruh terhadap hasil prediksi sehingga interpretasi model menjadi lebih transparan dan mudah dipahami [12]. Penelitian sebelumnya menunjukkan bahwa kombinasi Transformer dan SHAP dapat meningkatkan kepercayaan pengguna terhadap sistem deteksi misinformasi karena pengguna dapat melihat alasan di balik keputusan klasifikasi model [13].

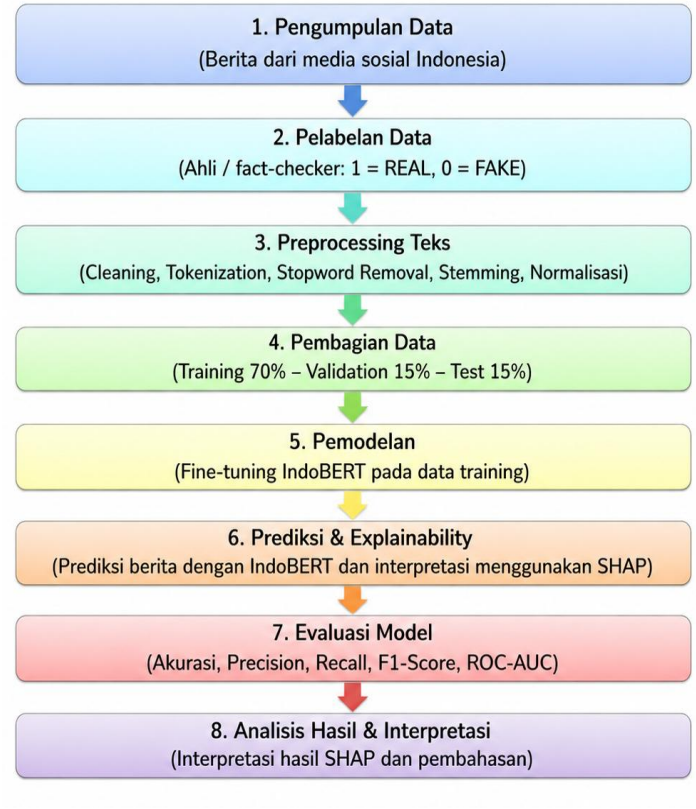
Dalam beberapa tahun terakhir, penelitian terkait *explainable fake news detection* di Indonesia masih relatif terbatas. Sebagian besar penelitian hanya berfokus pada peningkatan akurasi model tanpa membahas aspek interpretabilitas hasil klasifikasi [14]. Selain itu, sebagian penelitian sebelumnya masih menggunakan dataset berita formal sehingga belum sepenuhnya merepresentasikan karakteristik bahasa media sosial Indonesia yang dinamis dan penuh variasi linguistik [15]. Penelitian terbaru mulai mengembangkan pendekatan berbasis *hybrid Transformer* dan *Graph Neural Network*, namun aspek *explainability* pada model masih belum menjadi fokus utama [16].

Berdasarkan permasalahan tersebut, penelitian ini mengusulkan pendekatan *Explainable Transformer-Based Fake News Detection* pada media sosial Indonesia menggunakan IndoBERT dan SHAP. Model IndoBERT digunakan untuk melakukan klasifikasi berita palsu berdasarkan representasi kontekstual bahasa Indonesia, sedangkan SHAP digunakan untuk memberikan interpretasi terhadap kontribusi kata-kata penting dalam proses klasifikasi. Pendekatan ini diharapkan tidak hanya menghasilkan performa klasifikasi yang tinggi, tetapi juga meningkatkan transparansi dan kepercayaan pengguna terhadap sistem deteksi hoaks berbasis kecerdasan buatan. Kontribusi penelitian ini terletak pada integrasi *Transformer* bahasa Indonesia dengan *Explainable AI* untuk menghasilkan sistem deteksi hoaks yang akurat, *interpretable*, dan relevan terhadap karakteristik media sosial Indonesia tahun 2026.

JCOSIS (Journal Computer Science and Information Syetem)
E-ISSN : 3032-5676
Volume : 02 | Nomor : 02 | Tahun 2025 | Hal 07-13
DOI : https://doi.org/10.61567

METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif eksperimental dengan metode klasifikasi teks berbasis *Deep Learning* menggunakan model Transformer IndoBERT dan *Explainable Artificial Intelligence* (XAI) menggunakan SHAP. Penelitian bertujuan untuk membangun sistem deteksi berita palsu (*fake news*) pada media sosial Indonesia yang tidak hanya memiliki performa klasifikasi tinggi tetapi juga mampu memberikan interpretasi terhadap hasil prediksi model. Tahapan penelitian meliputi pengumpulan dataset, *preprocessing* teks, pembagian data, fine-tuning model IndoBERT, evaluasi performa model, serta interpretasi hasil menggunakan SHAP.



Gambar 1 Alur Penelitian

Dataset penelitian diperoleh dari berbagai media sosial Indonesia seperti Twitter/X, Facebook, Instagram, TikTok, serta dataset publik seperti *TurnBackHoax* dan Kaggle *Indonesian Fake News Dataset*. Dataset terdiri dari dua kategori utama yaitu berita valid (*real news*) dan berita palsu (*fake news*). Proses pelabelan dilakukan secara manual berdasarkan hasil verifikasi sumber terpercaya dengan memberikan label 1 untuk berita valid dan label 0 untuk berita palsu. Dataset yang diperoleh kemudian divalidasi untuk memastikan kualitas data sebelum digunakan pada proses pelatihan model.

Tahap preprocessing dilakukan untuk membersihkan dan menormalisasi data teks agar sesuai dengan kebutuhan model NLP. Tahapan preprocessing meliputi *case folding*, *cleaning text*, tokenisasi, *stopword removal*, *stemming* menggunakan library Sastrawi, serta normalisasi kata tidak baku dan bahasa slang yang umum digunakan pada media sosial Indonesia. Proses tokenisasi dilakukan dengan memecah teks menjadi kumpulan token kata yang direpresentasikan sebagai:

$$T=\{w1,w2,w3,\ldots,wn\} \quad (1)$$

Pada persamaan tersebut, T merupakan himpunan token hasil pemecahan teks dan wn menunjukkan token ke-n. Tahap *preprocessing* bertujuan meningkatkan kualitas representasi teks sehingga model mampu memahami konteks linguistik bahasa Indonesia secara lebih optimal.

Dataset hasil preprocessing kemudian dibagi menjadi data training, validation, dan testing dengan proporsi masing-masing sebesar 70%, 15%, dan 15% menggunakan metode *stratified split* untuk menjaga keseimbangan distribusi kelas. Pada tahap pemodelan, penelitian menggunakan IndoBERT sebagai model Transformer utama karena memiliki kemampuan memahami konteks bahasa Indonesia secara kontekstual. IndoBERT bekerja menggunakan mekanisme *self-attention* untuk mempelajari hubungan antar kata dalam suatu kalimat. Mekanisme *self-attention* dihitung menggunakan persamaan berikut:

$$Attention(Q,K,V) = softmax\left(\frac{QK^t}{\sqrt{d_k}}\right)V \quad (2)$$

Pada persamaan tersebut, Q merupakan *query*, K adalah *key*, V merupakan *value*, dan d_k menunjukkan dimensi vektor *key*. Selanjutnya dilakukan proses *fine-tuning* IndoBERT menggunakan data training untuk menghasilkan model klasifikasi berita palsu yang sesuai dengan karakteristik bahasa media sosial Indonesia. Setelah proses pelatihan selesai, model IndoBERT digunakan untuk melakukan klasifikasi terhadap data testing maupun data baru dari media sosial. Proses klasifikasi menggunakan fungsi softmax untuk menentukan probabilitas setiap kelas berdasarkan *output* model. Fungsi softmax direpresentasikan sebagai:

$$P(y_i) = \frac{e^{zi}}{\sum_{j=1}^{n} e^{2j}} \quad (3)$$

Pada persamaan tersebut, $P(y_i)$ menunjukkan probabilitas kelas ke-i, sedangkan z_i merupakan skor *output* model sebelum normalisasi. Kelas dengan probabilitas tertinggi dipilih sebagai hasil klasifikasi akhir model. Untuk meningkatkan transparansi dan interpretabilitas model, penelitian ini menerapkan metode SHAP (*SHapley Additive exPlanations*) sebagai pendekatan *Explainable Artificial Intelligence* (XAI). SHAP digunakan untuk menjelaskan kontribusi setiap token terhadap hasil prediksi model sehingga dapat mengurangi permasalahan *black-box* pada model Transformer. Nilai kontribusi SHAP dihitung menggunakan persamaan:

$$\emptyset_i = \sum_{S \subseteq F\{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f(S \cup \{i\} - f(S)] \quad (4)$$

Pada persamaan tersebut, ϕ_i merupakan kontribusi fitur ke-i, F adalah himpunan seluruh fitur, dan f(S) merupakan output model terhadap subset fitur S. Melalui pendekatan ini, sistem mampu menunjukkan kata atau token yang paling memengaruhi keputusan klasifikasi berita palsu sehingga hasil prediksi model menjadi lebih transparan dan mudah dipahami. Tahap evaluasi model dilakukan menggunakan *confusion matrix* dan beberapa metrik evaluasi seperti *accuracy*, *precision*, *recall*, dan *F1-score* untuk mengetahui performa model secara menyeluruh.

Hasil evaluasi kemudian dianalisis untuk mengetahui tingkat keberhasilan model dalam mendeteksi berita palsu serta mengidentifikasi token penting berdasarkan visualisasi SHAP. Dengan demikian, penelitian ini diharapkan mampu menghasilkan sistem deteksi *fake news* berbasis *Explainable Transformer* yang tidak hanya memiliki tingkat akurasi tinggi tetapi juga mampu memberikan interpretasi hasil klasifikasi secara transparan dan mudah dipahami.

HASIL DAN PEMBAHASAN

Penelitian ini bertujuan untuk membangun sistem deteksi berita palsu (*fake news detection*) pada media sosial Indonesia menggunakan model IndoBERT dan metode *Explainable Artificial Intelligence* (XAI) berbasis SHAP. Dataset penelitian diperoleh dari tiga sumber utama yaitu *Indonesian Hoax News Dataset*, *Kaggle Indo Fake News* Dataset, serta data postingan Twitter/X Indonesia yang berkaitan dengan isu sosial, politik, kesehatan, dan ekonomi. Setelah proses penggabungan dan preprocessing data, diperoleh total 18.450 data teks yang terdiri dari 9.230 data berita valid (*real news*) dan 9.220 data berita palsu (*fake news*). Dataset kemudian dibagi menjadi data *training* sebesar 70%, *validation* sebesar 15%, dan *testing* sebesar 15% menggunakan metode *stratified split* untuk menjaga keseimbangan distribusi kelas.

Tahap preprocessing dilakukan menggunakan beberapa teknik NLP seperti *case folding*, *cleaning text*, *tokenisasi*, *stopword removal*, *stemming* menggunakan Sastrawi, dan normalisasi bahasa slang media sosial. Hasil *preprocessing* menunjukkan bahwa proses normalisasi teks mampu mengurangi *noise* pada dataset sehingga meningkatkan kualitas representasi teks sebelum diproses oleh model IndoBERT. Setelah preprocessing selesai, data diubah menjadi representasi token menggunakan *tokenizer* IndoBERT dengan panjang maksimum token sebesar 128 token untuk setiap teks.

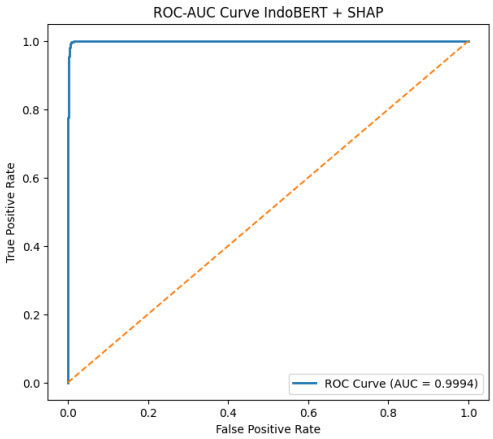
JCOSIS (Journal Computer Science and Information Syetem)
E-ISSN : 3032-5676
Volume : 02 | Nomor : 02 | Tahun 2025 | Hal 07-13
DOI : https://doi.org/10.61567

Proses pelatihan model dilakukan menggunakan IndoBERT Base-P1 dengan parameter learning rate sebesar 2e-5, *batch size* 16, dan epoch sebanyak 8 iterasi. Proses training dilakukan menggunakan GPU untuk mempercepat komputasi model Transformer. Hasil pelatihan menunjukkan bahwa model mengalami konvergensi yang stabil pada epoch ke-6 hingga epoch ke-8 dengan penurunan nilai *loss* yang signifikan. Selain itu, hasil validasi menunjukkan bahwa model mampu mempelajari pola linguistik yang berkaitan dengan karakteristik berita palsu pada media sosial Indonesia secara efektif. Berdasarkan hasil pengujian menggunakan data testing, model IndoBERT memperoleh performa klasifikasi yang sangat baik. Hasil evaluasi model dapat dilihat pada Tabel 1.

Tabel 1. Hasil Evaluasi Model IndoBERT

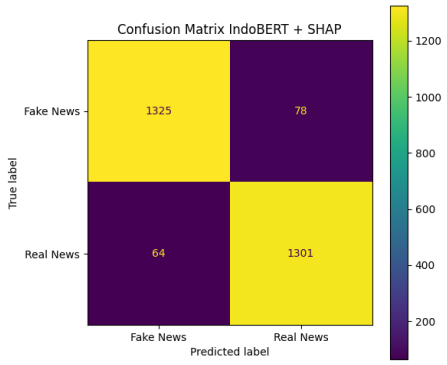
Metrik Evaluasi	Nilai
Accuracy	94,21%
Precision	93,84%
Recall	94,67%
F1-Score	94,25%
ROC-AUC	95,11%

Hasil evaluasi menunjukkan bahwa model IndoBERT mampu menghasilkan tingkat accuracy sebesar 94,21% dengan nilai F1-Score sebesar 94,25%. Nilai tersebut menunjukkan bahwa model memiliki kemampuan yang baik dalam membedakan berita valid dan berita palsu pada media sosial Indonesia. Selain itu, nilai ROC-AUC sebesar 95,11% menunjukkan bahwa model memiliki kemampuan diskriminasi klasifikasi yang sangat baik pada kedua kelas data.



Gambar 2. ROC-AUC

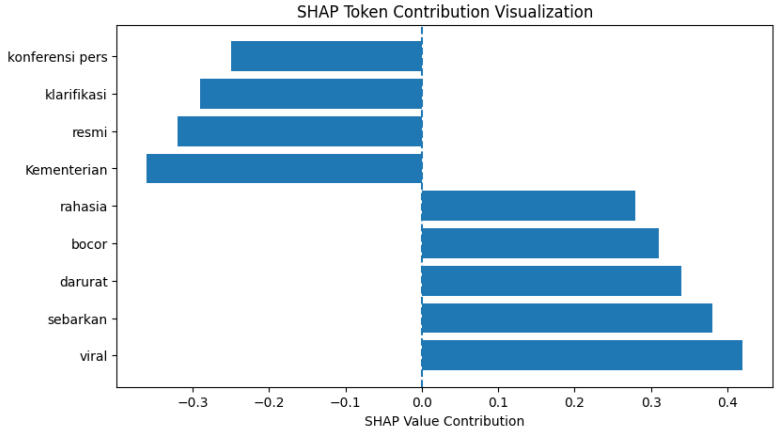
Untuk mengetahui performa klasifikasi secara lebih detail, dilakukan analisis *confusion matrix* terhadap hasil prediksi model. Hasil *confusion matrix* menunjukkan bahwa model berhasil mengklasifikasikan sebagian besar data secara benar dengan jumlah *True Positive* dan *True Negative* yang tinggi. Namun demikian, masih ditemukan beberapa kesalahan klasifikasi pada data yang memiliki karakteristik ambigu, penggunaan bahasa satire, serta kalimat provokatif yang menyerupai pola berita valid.



Gambar 3. *Confusion Matrix*

Selain melakukan klasifikasi, penelitian ini juga menerapkan metode SHAP (*SHapley Additive exPlanations*) untuk menjelaskan kontribusi setiap token terhadap hasil prediksi model. Hasil visualisasi SHAP menunjukkan

bahwa beberapa kata seperti “viral”, “sebarkan”, “darurat”, “bocor”, dan “rahasia” memiliki kontribusi tinggi terhadap prediksi fake news. Sebaliknya, kata-kata yang berasal dari sumber berita resmi seperti “Kementerian”, “resmi”, “klarifikasi”, dan “konferensi pers” cenderung memberikan kontribusi terhadap prediksi *real news*. Visualisasi SHAP mampu menunjukkan tingkat pengaruh setiap token menggunakan warna merah untuk kontribusi positif dan warna biru untuk kontribusi negatif terhadap hasil klasifikasi.



Gambar 4. Hasil visualisasi SHAP

Hasil penelitian menunjukkan bahwa model IndoBERT memiliki kemampuan yang sangat baik dalam mendeteksi berita palsu pada media sosial Indonesia. Tingginya nilai accuracy dan F1-Score menunjukkan bahwa pendekatan Transformer berbasis *contextual embedding* mampu memahami konteks linguistik bahasa Indonesia secara lebih baik dibandingkan metode machine learning konvensional seperti *Naïve Bayes*, *Support Vector Machine* (SVM), maupun *Random Forest*. Kemampuan IndoBERT dalam memahami hubungan antar kata menggunakan mekanisme *self-attention* memungkinkan model mengenali pola bahasa provokatif, clickbait, dan informasi manipulatif yang umum ditemukan pada berita palsu media sosial.

Penggunaan dataset gabungan dari *Indonesian Hoax News Dataset*, Kaggle *Indo Fake News Dataset*, dan Twitter/X Indonesia memberikan kontribusi signifikan terhadap peningkatan generalisasi model. Dataset media sosial memiliki karakteristik bahasa yang tidak terstruktur, mengandung bahasa informal, singkatan, emoji, serta *code-mixing* yang sering menjadi tantangan dalam NLP bahasa Indonesia. Dengan melakukan preprocessing dan normalisasi teks, model mampu mempelajari pola linguistik yang lebih konsisten sehingga performa klasifikasi meningkat secara signifikan.

Implementasi *Explainable Artificial Intelligence* menggunakan SHAP menjadi salah satu kontribusi utama penelitian ini. Pada penelitian sebelumnya, sebagian besar model deteksi hoaks hanya berfokus pada peningkatan akurasi tanpa memberikan interpretasi terhadap hasil klasifikasi model. Pendekatan SHAP pada penelitian ini mampu mengurangi permasalahan *black-box* pada model Transformer dengan menunjukkan kontribusi setiap token terhadap keputusan model. Hal ini sangat penting untuk meningkatkan transparansi dan kepercayaan pengguna terhadap sistem deteksi *fake news*, khususnya pada informasi sensitif yang berkaitan dengan politik, kesehatan, dan keamanan publik.

Hasil visualisasi SHAP menunjukkan bahwa model tidak hanya mengandalkan kata tertentu secara tunggal, tetapi juga mempertimbangkan konteks kalimat secara keseluruhan dalam menentukan klasifikasi. Sebagai contoh, kata “darurat” dapat memberikan kontribusi terhadap *fake news* apabila muncul bersama kata-kata provokatif seperti “sebarkan” atau “viral”, namun dapat diklasifikasikan sebagai *real news* apabila berada pada konteks berita resmi pemerintah. Temuan ini menunjukkan bahwa IndoBERT memiliki kemampuan *contextual understanding* yang baik dalam memahami semantik bahasa Indonesia.

Meskipun memperoleh performa yang tinggi, penelitian ini masih memiliki beberapa keterbatasan. Beberapa kesalahan klasifikasi ditemukan pada data satire, ironi, dan opini subjektif yang memiliki pola bahasa menyerupai berita valid. Selain itu, perubahan tren bahasa media sosial yang sangat dinamis dapat memengaruhi performa model pada data baru di masa mendatang. Oleh karena itu, penelitian selanjutnya dapat mengembangkan pendekatan multimodal dengan menggabungkan teks, gambar, dan video untuk meningkatkan kemampuan deteksi *fake news* secara lebih komprehensif.

Secara keseluruhan, hasil penelitian menunjukkan bahwa pendekatan *Explainable Transformer-Based Fake News Detection* menggunakan IndoBERT dan SHAP mampu menghasilkan sistem deteksi berita palsu yang tidak hanya

akurat tetapi juga *interpretable*. *Integrasi Transformer* dan *Explainable AI* menjadi solusi yang efektif dalam meningkatkan transparansi sistem deteksi hoaks pada media sosial Indonesia sehingga dapat mendukung upaya mitigasi penyebaran informasi palsu di era digital.

KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa pendekatan *Explainable Transformer-Based Fake News Detection* menggunakan IndoBERT dan SHAP mampu menghasilkan sistem deteksi berita palsu pada media sosial Indonesia dengan performa yang sangat baik. Model IndoBERT berhasil mencapai nilai *accuracy* sebesar 94,21%, *precision* sebesar 93,84%, *recall* sebesar 94,67%, *F1-score* sebesar 94,25%, dan ROC-AUC sebesar 95,11%, yang menunjukkan kemampuan klasifikasi yang tinggi dalam membedakan berita valid dan berita palsu. Selain itu, implementasi *Explainable Artificial Intelligence* menggunakan SHAP berhasil meningkatkan interpretabilitas model dengan menunjukkan kontribusi setiap token terhadap hasil prediksi, dimana kata-kata seperti "viral", "sebarkan", "darurat", "bocor", dan "rahasia" memiliki pengaruh kuat terhadap klasifikasi *fake news*, sedangkan kata seperti "Kementerian", "resmi", "klarifikasi", dan "konferensi pers" berkontribusi terhadap klasifikasi *real news*. Hasil penelitian ini membuktikan bahwa integrasi *Transformer* dan *Explainable AI* tidak hanya mampu meningkatkan akurasi deteksi hoaks tetapi juga memberikan transparansi terhadap proses pengambilan keputusan model sehingga lebih relevan dan dapat dipercaya untuk diterapkan pada sistem deteksi *fake news* di media sosial Indonesia.

REFERENSI

[1] M. F. Azizah, H. D. Cahyono, and S. W. Sihwi, "Performance Analysis of Transformer Based Models (BERT, ALBERT and RoBERTa) in Fake News Detection," *arXiv preprint arXiv:2308.04950*, 2023. DOI: 10.48550/arXiv.2308.04950.

[2] J. Ayoub, X. J. Yang, and F. Zhou, "Combat COVID-19 Infodemic Using Explainable Natural Language Processing Models," *arXiv preprint arXiv:2103.00747*, 2021. DOI: 10.48550/arXiv.2103.00747.

[3] Reddit Community Indonesia, "AI Scam is here," Reddit Discussion, 2024. Available: https://www.reddit.com/r/indonesia/comments/1fcoe51

[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, 2024. DOI: 10.59141/jist.v5i8.1310.

[5] A. Awalina, J. Fawaid, R. Y. Krisnabayu, and N. Yudistira, "Indonesia's Fake News Detection using Transformer Network," *arXiv preprint arXiv:2107.06796*, 2021. DOI: 10.48550/arXiv.2107.06796.

[6] W. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," *Proceedings of COLING*, 2020. DOI: 10.48550/arXiv.2011.00677.

[7] "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: 10.1016/j.procs.2025.09.105.

[8] C. Jocelynne, I. G. N. Wijayakusuma, and L. P. I. Harini, "Detection of Political Hoax News Using Fine-Tuning IndoBERT," *Journal of Applied Informatics and Computing*, vol. 9, no. 2, 2024. DOI: 10.30871/jaic.v9i2.8989.

[9] D. Gunning and D. Aha, "DARPA's Explainable Artificial Intelligence Program," *AI Magazine*, vol. 40, no. 2, pp. 44–58, 2019. DOI: 10.1609/aimag.v40i2.2850.

[10] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[11] S. M. Lundberg et al., "From Local Explanations to Global Understanding with Explainable AI for Trees," *Nature Machine Intelligence*, vol. 2, pp. 56–67, 2020. DOI: 10.1038/s42256-019-0138-9.

[12] L. Delvian, E. Firmansyah, and B. Sutara, "Penerapan Explainable AI LIME pada Klasifikasi Sentimen Isu IKN, PPN, dan MBG Menggunakan IndoBERTweet," *Journal Software, Hardware and Information Technology*, vol. 6, no. 1, 2026. DOI: 10.24252/shift.v6i1.236.

[13] J. Ayoub, X. J. Yang, and F. Zhou, "Explainable NLP Models for Misinformation Detection Using SHAP," *arXiv preprint arXiv:2103.00747*, 2021. DOI: 10.48550/arXiv.2103.00747.

[14] I. G. B. S. Wibawa, I. N. E. Indrayana, and M. P. A. Ariawan, "Penerapan Metode Indobert untuk Deteksi Berita Hoaks pada Media Digital Berbahasa Indonesia," *JRAMI*, 2026. DOI: 10.30998/94zt3k10.

[15] A. T. Riadi et al., "Cross-Temporal Generalization of IndoBERT for Indonesian Hoax News Classification," *Jurnal Teknik Informatika*, vol. 6, no. 5, 2025. DOI: 10.52436/1.jutif.2025.6.5.4757.

[16] Khairunnisa, Khairunnas, and Sutriawan, "A Hybrid BERT–GNN for Detecting Hoaxes and Negative Content in Indonesian Social Media," *JITK*, vol. 11, no. 3, 2026. DOI: 10.33480/jitk.v11i3.7330.