# ABSTRAK

Penyebaran berita hoaks di era digital menjadi ancaman bagi integritas informasi, sehingga diperlukan sistem deteksi otomatis yang akurat dan transparan. Pendekatan klasifikasi teks menggunakan model IndoBERT terbukti memiliki performa tinggi, namun sifatnya yang *black-box* membuat alasan di balik pengambilan keputusannya sulit dipahami. Penelitian ini bertujuan untuk menerapkan *Explainable Artificial Intelligence* (XAI) menggunakan metode *SHapley Additive exPlanations* (SHAP) pada model IndoBERT untuk mengidentifikasi dan mengukur kontribusi fitur linguistik yang mempengaruhi keputusan model dalam mendeteksi berita hoaks berbahasa Indonesia. Penelitian ini menggunakan kerangka kerja OSEMN (*Obtain, Scrub, Explore, Model, iNterpret*) dengan memanfaatkan 24.740 data hasil *web scraping* dari portal berita seperti CNN, Kompas, dan Tempo untuk teks fakta serta Turnbackhoax untuk teks hoaks. Hasil evaluasi pemodelan menunjukkan bahwa hasil yang terbaik adalah pada rasio pembagian data 70:20:10 dengan F1-Score yaitu 99,46%. Meskipun matrik evaluasi menunjukkan hasil yang sangat tinggi, interpretasi melalui SHAP mengungkap bahwa model cenderung bertindak sebagai pengklasifikasi gaya bahasa dibandingkan pengklasifikasi kebenaran faktual, selain itu moel terindikasi *overfitting* ringan. Model sangat bergantung pada kosa kata formal untuk memprediksi kelas fakta, serta bahasa hiperbola dan provokatif untuk mengenali kelas hoaks. Untuk mengatasi keterbatasan sifat *black-box* tersebut, hasil interpretasi SHAP digabungkan dengan penjelasan naratif menggunakan pendekatan *Rule-Based Natural Language Generation* (NLG) yang diimplementasikan ke dalam antarmuka berbasis web.



**Kata Kunci:** Deteksi Hoaks, IndoBERT, *Explainable Artificial Intelligence*, SHAP, *Natural Language Generation*.

## *ABSTRACT*

*The spread of misinformation in the digital era poses a significant threat to information integrity, thereby necessitating accurate and transparent automated detection systems. Text classification approaches using the IndoBERT model have demonstrated high performance; however, their black-box nature makes the reasoning behind their decisions difficult to interpret. This study aims to implement Explainable Artificial Intelligence (XAI) using the SHapley Additive exPlanations (SHAP) method on the IndoBERT model to identify and quantify the contribution of linguistic features influencing the model's decisions in detecting Indonesian-language misinformation. This research adopts the OSEMN (Obtain, Scrub, Explore, Model, iNterpret) framework, utilizing 24,740 data points collected through web scraping from news portals such as CNN, Kompas, and Tempo for factual texts, as well as Turnbackhoax for misinformation texts. The modeling evaluation results indicate that the best performance is achieved with a 70:20:10 data split ratio, yielding an F1-Score of 99.46%. Despite the exceptionally high evaluation metrics, SHAP-based interpretation reveals that the model tends to function more as a stylistic classifier rather than a factual truth classifier,* furthermore the model exhibits indications of mild overfitting. *The model heavily relies on formal vocabulary to predict factual content, while hyperbolic and provocative language is strongly associated with misinformation classification. To address the limitations of the black-box nature, SHAP interpretation results are integrated with narrative explanations using a Rule-Based Natural Language Generation (NLG) approach, which is implemented within a web-based interface.*



***Keywords:*** *Hoax Detection, IndoBERT, Explainable Artificial Intelligence, SHAP, Natural Language Generation.*