ABSTRAK

Deny Supanji, I Komang (2026), *Evaluasi Multi-Dimensi Model Transformer untuk Analisis Sentimen Bahasa Bali: Performa, Tokenisasi, dan Interpretabilitas*. Tesis, Ilmu Komputer, Program Pascasarjana, Universitas Pendidikan Ganesha.

Tesis ini sudah disetujui dan diperiksa oleh Pembimbing I : Dr. I Gede Aris Gunadi, S.Si., M.Kom. dan Pembimbing II: Dr. Ir. I Made Gede Sunarya, S.Kom., M.Cs.

Kata-kata kunci: analisis sentimen, bahasa Bali, model *pre-trained Transformer,* audit *tokenizer*, SHAP

Bahasa Bali tergolong sebagai *low-resource language* yang menghadapi keterbatasan sumber daya pemrosesan bahasa alami. Penelitian ini mengevaluasi tiga model *pre-trained* berbasis Transformer, yaitu IndoBERT, mBERT, dan XLM-RoBERTa, untuk analisis sentimen teks berbahasa Bali pada *dataset* NusaX melalui kerangka evaluasi multi-dimensi yang mengintegrasikan performa klasifikasi, audit kualitas tokenisasi, dan interpretabilitas SHAP. Audit tokenisasi dilakukan terhadap leksikon bahasa Bali yang telah melalui proses pembersihan dari 1.063 entri mentah menjadi 830 kata unik. Hasil menunjukkan IndoBERT mencapai performa terbaik dengan akurasi 78,00% dan *F1-Macro* 77,23%, mengungguli XLM-RoBERTa (65,25%, 64,39%) dan mBERT (64,75%, 63,59%) secara signifikan berdasarkan uji McNemar. Audit *tokenizer* menunjukkan IndoBERT memiliki *fertility rate* terendah (1,9373) dan *perfect match rate* tertinggi (30,36%) yang berkorelasi dengan performa terbaik. Analisis *OOV rate* dihitung melalui pendekatan *vocabulary lookup* yang konsisten untuk WordPiece dan SentencePiece. Analisis SHAP menunjukkan IndoBERT menghasilkan atribusi token yang lebih terfokus secara semantik, sementara model *multilingual* cenderung mengandalkan tanda baca dan fragmen *subword* tidak bermakna akibat fragmentasi tokenisasi. Penelitian ini menyimpulkan bahwa kesamaan kosakata antara data pelatihan dan dataset tampak lebih berpengaruh terhadap performa dibandingkan ukuran model, dan menghasilkan kerangka evaluasi multi-dimensi yang dapat direplikasi sebagai prosedur evaluasi pada penelitian sejenis, dengan penyesuaian terhadap karakteristik kosakata masing-masing dataset. Perlu ditekankan bahwa keunggulan IndoBERT bersumber dari tumpang tindih leksikal dengan bahasa Indonesia, mengingat dataset NusaX bahasa Bali merupakan hasil terjemahan yang memuat kata serapan dan potensi *code-mixing*, sehingga sebagian besar kosakata Bali murni tetap terfragmentasi pada seluruh model. Keberlakuan temuan pada bahasa daerah lain karenanya masih memerlukan pengujian lebih lanjut.

ABSTRACT

Deny Supanji, I Komang (2026), Multi-Dimensional Evaluation of Transformer Models for Balinese Sentiment Analysis: Performance, Tokenization, and Interpretability. Thesis, Computer Science, Postgraduate Program, Universitas Pendidikan Ganesha.

This thesis has been approved and examined by: Supervisor I: Dr. I Gede Aris Gunadi, S.Si. M.Kom. Supervisor II: Dr. Ir. I Made Gede Sunarya, S.Kom., M.Cs.

Keywords: sentiment analysis, Balinese language, pre-trained Transformer models, tokenizer audit, SHAP

Balinese is classified as a low-resource language that faces limited natural language processing resources. This study evaluates three Transformer-based pre-trained models, namely IndoBERT, mBERT, and XLM-RoBERTa, for sentiment analysis of Balinese text on the NusaX dataset through a multi-dimensional evaluation framework that integrates classification performance, tokenization quality auditing, and SHAP interpretability. The tokenization audit was conducted on a Balinese lexicon that was cleaned from 1,063 raw entries into 830 unique words. The results show that IndoBERT achieves the best performance with an accuracy of 78.00% and a macro F1-score of 77.23%, significantly outperforming XLM-RoBERTa (65.25%, 64.39%) and mBERT (64.75%, 63.59%) based on McNemar's test. The tokenizer audit shows that IndoBERT has the lowest fertility rate (1.9373) and the highest perfect match rate (30.36%), which correlates with its superior performance. The OOV rate analysis was computed through a vocabulary lookup approach applied consistently to both WordPiece and SentencePiece. The SHAP analysis indicates that IndoBERT produces more semantically focused token attributions, whereas the multilingual models tend to rely on punctuation and meaningless subword fragments resulting from tokenization fragmentation. This study concludes that the vocabulary overlap between the training data and the dataset appears to be more influential on performance than model scale, and it produces a multi-dimensional evaluation framework that can be replicated as an evaluation procedure for similar studies, with adjustments to the vocabulary characteristics of each dataset. It should be emphasized that IndoBERT's advantage stems from its lexical overlap with Indonesian, given that the Balinese NusaX dataset is a translation containing loanwords and potential code-mixing, so that most genuine Balinese vocabulary remains fragmented across all models. The applicability of these findings to other regional languages therefore requires further testing.