ABSTRAK

Adyanata, I Komang (2026). Analisis Komparatif mBERT Dan IndoBERT untuk Klasifikasi Sentimen *Tweet* Berbahasa Indonesia Informal dan *Code-Mixing*. Tesis, Ilmu Komputer, Program Pascasarjana, Universitas Pendidikan Ganesha.

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

Kata Kunci: analisis sentimen, IndoBERT, mBERT, *hyperparameter tuning, class weighting*

Pertumbuhan pesat penggunaan media sosial Twitter di Indonesia memunculkan kebutuhan akan analisis sentimen yang mampu menangani karakteristik bahasa informal, *slang*, singkatan, serta pencampuran bahasa Indonesia dengan bahasa lainnya (*code-mixing*). Penelitian ini membandingkan dua model berbasis *transformer*, yakni *Multilingual* BERT (mBERT) *base-uncased* yang dilatih pada 104 bahasa dan IndoBERT *base-p1* yang khusus bahasa Indonesia, pada klasifikasi sentimen tiga kelas (negatif, netral, positif). *Dataset* yang digunakan adalah *Indonesian General Sentiment Analysis Dataset* sebanyak 10.806 *tweet* dengan distribusi label tidak seimbang (netral 49,3%, negatif 26,7%, positif 24,0%) dan terdapat pencampuran bahasa dalam *dataset*. Pra-pemrosesan dilakukan secara terbatas, mencakup *case folding* dan pembersihan URL, *mention*, *hashtag*, serta karakter spesial, dengan mempertahankan *slang*, singkatan, dan karakter berulang sebagai ciri khas data. Ketidakseimbangan kelas ditangani dengan *balanced class weighting* pada fungsi *loss CrossEntropy*, dan *hyperparameter tuning* dilakukan dalam dua tahap. Eksplorasi data menunjukkan tingginya bahasa nonbaku: 37,8% *tweet* memuat *slang* dan 31,7% memuat singkatan. Analisis tokenisasi memperlihatkan mBERT memecah kata secara berlebihan (1,529 *subword* per kata), sedangkan IndoBERT lebih hemat (1,192 *subword* per kata, sekitar 22%). Pada *test set* 2.162 *tweet*, IndoBERT mencatat *accuracy* 69,84%, *F1-macro* 69,07%, dan *F1-weighted* 69,93%, sementara mBERT mencapai *accuracy* 63,00%, *F1-macro* 61,43%, dan *F1-weighted* 63,00%, sehingga IndoBERT unggul +6,85 poin *accuracy* dan +7,64 poin *F1-macro*. Dibandingkan *baseline* KNN dan *SVM* (62,9%), IndoBERT meningkat +6,94 poin. Keunggulan IndoBERT terbukti signifikan secara statistik melalui uji *McNemar* ($p<0,0001$) dengan selang kepercayaan *bootstrap* 95% yang tidak tumpang-tindih, sementara mBERT tidak berbeda signifikan dari *baseline* klasik *TF-IDF + SVM* ($p=1,000$). Analisis bertingkat menunjukkan selisih performa terbesar pada *tweet* berkarakter berulang (*F1-macro* 12,51 poin) dan *tweet* pendek di bawah 10 kata (10,32 poin), mengindikasikan ketepatan tokenisasi sebagai penentu utama keberhasilan klasifikasi pada konteks bahasa nonbaku.

ABSTRACT

***Adyanata, I Komang (2026)**. Comparative Analysis of mBERT and IndoBERT for Sentiment Classification of Informal and Code-Mixing Indonesian Tweets. Thesis, Computer Science, Postgraduate Program, Ganesha University of Education.*

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

Keywords: sentiment analysis, IndoBERT, mBERT, hyperparameter tuning, class weighting

The rapid growth of Twitter usage in Indonesia has created a need for sentiment analysis capable of handling informal language characteristics, slang, abbreviations, and the mixing of Indonesian with other languages (code-mixing). This study compares two transformer-based models, namely Multilingual BERT (mBERT) base-uncased trained on 104 languages and IndoBERT base-p1 dedicated to the Indonesian language, on a three-class sentiment classification task (negative, neutral, positive). The dataset used is the Indonesian General Sentiment Analysis Dataset, comprising 10,806 tweets with an imbalanced label distribution (neutral 49.3%, negative 26.7%, positive 24.0%) and the presence of language mixing within the data. Preprocessing was performed in a limited manner, covering case folding and the removal of URLs, mentions, hashtags, and special characters, while preserving slang, abbreviations, and repeated characters as distinctive features of the data. Class imbalance was addressed using balanced class weighting on the CrossEntropy loss function, and hyperparameter tuning was conducted in two stages. Exploratory data analysis revealed a high prevalence of non-standard language: 37.8% of tweets contained slang and 31.7% contained abbreviations. Tokenization analysis showed that mBERT over-fragmented words (1.529 subwords per word), whereas IndoBERT was more efficient (1.192 subwords per word, approximately 22% fewer). On the test set of 2,162 tweets, IndoBERT achieved an accuracy of 69.84%, an F1-macro of 69.07%, and an F1-weighted of 69.93%, while mBERT obtained an accuracy of 63.00%, an F1-macro of 61.43%, and an F1-weighted of 63.00%, so that IndoBERT outperformed mBERT by +6.85 points in accuracy and +7.64 points in F1-macro. Compared with the KNN and SVM baselines (62.9%), IndoBERT improved by +6.94 points. The superiority of IndoBERT was statistically significant based on the McNemar test ($p<0.0001$) with non-overlapping 95% bootstrap confidence intervals, whereas mBERT did not differ significantly from the classical TF-IDF + SVM baseline ($p=1.000$). Stratified analysis showed that the largest performance gaps occurred on tweets with repeated characters (F1-macro 12.51 points) and short tweets of fewer than 10 words (10.32 points), indicating that tokenization accuracy is the primary determinant of classification success in non-standard language contexts.