

DETEKSI PENGGUNAAN BAHASA JAKARTA SELATAN PADA MEDIA SOSIAL X MENGGUNAKAN MODEL INDOBERT BERBASIS BIDIRECTIONAL CONTEXT

Nama Mahasiswa : Niko Afandi Saputro
NIM : 11221039
Dosen Pembimbing Utama : Bima Prihasto, S.Si., M.Si., Ph.D
Dosen Pembimbing Pendamping : Boby Mugi Pratama, S.Si., M.Han.

ABSTRAK

Fenomena bahasa Jakarta Selatan yang mencampurkan kosakata bahasa Indonesia dengan bahasa Inggris melalui teknik *code-mixing* telah menjadi tren komunikasi yang masif di media sosial X. Namun, algoritma klasifikasi tradisional seperti *Support Vector Machine* dan *Random Forest* seringkali gagal menangkap makna kontekstual kalimat *code-mixing* yang kompleks karena ketergantungan pada fitur statistik *N-Gram* yang bersifat statis. Penelitian ini bertujuan untuk membangun sistem deteksi dan klasifikasi otomatis tipe *code-mixing* bahasa Jakarta Selatan menggunakan model IndoBERT berbasis *Bidirectional Context*, serta menganalisis efektivitas teknik augmentasi dan *balancing* data terhadap performa model. Metode penelitian menggunakan arsitektur klasifikasi dua tahap (*two-stage*). Stage 1 untuk deteksi biner (Bahasa Jaksel vs Bukan Bahasa Jaksel) dan Stage 2 untuk klasifikasi *multiclass* berdasarkan tipologi Muysken (*Insertion*, *Alternation*, dan *Congruent Lexicalization*). Dua eksperimen dilakukan dengan teknik augmentasi *rule-based replacement* yang identik, mengikuti alur data mentah ke kalimat baku kemudian diganti menggunakan aturan pemetaan kata, serta *balancing* rasio 50:50 untuk label biner dan 1:1:1 untuk jenis Muysken. Eksperimen utama menggunakan dataset IJELID (Hidayatullah, 2023) yang merepresentasikan teks media sosial X sebanyak 6.000 baris, sedangkan eksperimen pendukung menggunakan dataset transkrip YouTube sebanyak 8.000 baris sebagai pembanding. Model di *fine-tune* menggunakan IndoBERT-Base-P2 dengan *hyperparameter learning rate* 2e-5, *batch size* 16, dan 3 *epochs*. Hasil eksperimen utama menunjukkan bahwa model mencapai *F1 Macro* sebesar 0,92 pada Stage 1 dan 0,96 pada Stage 2, dengan kelas *Alternation* mencapai *F1-score* sempurna 1,00. Eksperimen pendukung memperoleh *F1 Macro* 0,94 pada Stage 1 dan 0,96 pada Stage 2. Berdasarkan hasil tersebut, dapat disimpulkan bahwa arsitektur *Two-Stage* IndoBERT dengan teknik augmentasi dan *balancing* 50:50 efektif dalam mendeteksi dan mengklasifikasikan fenomena bahasa Jakarta Selatan pada media sosial X.

Kata kunci: IndoBERT, Jakarta Selatan, *Code-Mixing*, *Bidirectional Context*, *Natural Language Processing*.