

IMPLEMENTASI *ITERATIVE MASKING* PADA INDOBERTTWEET-BILSTM UNTUK DETEKSI UJARAN KEBENCIAN BERBAHASA INDONESIA DI PLATFORM X

Nama Mahasiswa : Fakhrizal Naufal
NIM : 11221057
Dosen Pembimbing Utama : Bima Prihasto, S.Si, M.Si., Ph.D.
Dosen Pembimbing Pendamping : Ramadhan Paninggalih S.Si., M.Si., M.Sc.

ABSTRAK

Perkembangan teknologi komunikasi pada platform media sosial X telah memicu penyebaran ujaran kebencian (*hate speech*) yang masif di Indonesia. Karakteristik teks pada platform tersebut yang cenderung menggunakan bahasa informal, kata tidak baku (*slang*), serta ejaan yang tidak konsisten menjadi tantangan signifikan dalam proses deteksi otomatis. Penelitian terdahulu telah menggunakan model IndoBERTTweet-BiLSTM untuk menangkap representasi kontekstual, namun kinerja model tersebut masih belum optimal dalam membedakan ujaran kebencian murni dengan keluhan atau sentimen negatif lainnya yang menggunakan bahasa non-formal, sehingga membatasi tingkat akurasi klasifikasi. Tujuan dari penelitian ini adalah meningkatkan akurasi deteksi ujaran kebencian melalui pengembangan arsitektur *Dual-Branch Feature Fusion* yang mengintegrasikan fitur kontekstual mendalam dengan fitur leksikal dan statistik secara paralel. Metode penelitian yang diusulkan terdiri dari dua jalur ekstraksi fitur utama. Jalur pertama menggunakan IndoBERTTweet sebagai *encoder* yang dipadukan dengan lapisan *Bidirectional Long Short-Term Memory* (BiLSTM) untuk mengekstraksi informasi sekuensial dari dua arah. Jalur kedua menerapkan metode *Iterative Masking* berbasis *FastText* untuk mendapatkan fitur statistik berupa penurunan probabilitas (*probability drop*), serta fitur leksikal menggunakan kamus kata kasar (*abusive dictionary*) untuk menangkap keberadaan elemen bahasa ofensif secara eksplisit. Hasil ekstraksi dari kedua jalur tersebut kemudian diintegrasikan menggunakan teknik *Feature Fusion* melalui tahap konkatenasi sebelum diproses oleh lapisan klasifikasi akhir. Penelitian ini diharapkan dapat memberikan solusi yang lebih tangguh dalam mengenali pola ujaran kebencian pada teks media sosial yang dinamis dan tidak standar.

Kata kunci: *Hate Speech*, IndoBERTTweet, BiLSTM, *Iterative Masking*, *Feature Fusion*