ABSTRAK

Penggunaan Bahasa Indonesia dalam media sosial dan platform digital sering kali tidak mengikuti kaidah kebakuan, baik dari segi ejaan maupun pemilihan kata. Fenomena ini menimbulkan tantangan dalam pengembangan sistem pemrosesan bahasa alami (Natural Language Processing/NLP), khususnya dalam deteksi kesalahan dan normalisasi teks. Penelitian ini bertujuan untuk mengembangkan sistem deteksi kata tidak baku dalam kalimat Bahasa Indonesia menggunakan model trigram karakter dengan penerapan teknik *Laplace Smoothing*. Pendekatan ini dipilih karena mampu memperkirakan probabilitas kemunculan kata berdasarkan konteks tiga karakter sebelumnya, sekaligus mengatasi masalah *data sparsity* pada model bahasa.

Data yang digunakan meliputi kamus KBBI sebagai acuan kata baku, *Spelling Error Correction in Indonesian Language (SPECIL)* sebagai sumber pasangan kata baku dan tidak baku, serta *Twitter Emotion Dataset* sebagai korpus pelatihan model trigram. Proses penelitian meliputi tahap pra-pemrosesan teks, pembentukan model trigram, penerapan teknik *Laplace Smoothing*, serta evaluasi sistem berdasarkan metrik *precision*, *recall*, dan *F1-score*. Sistem dirancang untuk mengklasifikasikan kata dalam kalimat menjadi tiga kategori, yaitu baku**,** tidak baku (dikenali SPECIL), dan tidak baku (probabilitas trigram rendah)**.**

Hasil pengujian menunjukkan bahwa sistem yang dikembangkan mampu mendeteksi kata tidak baku dengan tingkat ketepatan yang baik dan stabil, bahkan untuk kata dengan variasi penulisan yang jarang ditemukan dalam data pelatihan. Penerapan *Laplace Smoothing* terbukti efektif dalam mencegah probabilitas nol pada kombinasi trigram langka, sehingga meningkatkan kemampuan model dalam memprediksi bentuk kata yang lebih alami. Sistem ini dapat diimplementasikan dalam aplikasi berbasis web sebagai alat bantu pemeriksa ejaan atau normalisasi teks otomatis. Dengan demikian, penelitian ini memberikan kontribusi terhadap pengembangan teknologi NLP untuk Bahasa Indonesia, khususnya dalam peningkatan kualitas teks digital dan pelestarian kebakuan bahasa.

Kata kunci: deteksi kata tidak baku, trigram, Laplace smoothing, Bahasa Indonesia, NLP

ABSTRACT

The use of the Indonesian language on social media and digital platforms often deviates from standard linguistic rules, particularly in spelling and word choice. This phenomenon poses challenges for Natural Language Processing (NLP) applications, especially in error detection and text normalization. This study aims to develop a system for detecting non-standard words in Indonesian sentences using a character-based trigram model combined with the *Laplace Smoothing* technique. This approach was chosen because it estimates word occurrence probabilities based on three-character contexts while addressing the *data sparsity* problem in language models.

The research utilizes three main data sources: the *Kamus Besar Bahasa Indonesia (KBBI)* as the reference for standard words, the *Spelling Error Correction in Indonesian Language (SPECIL)* dataset as the source of standard–non-standard word pairs, and the *Twitter Emotion Dataset* as the corpus for trigram model training. The process involves text preprocessing, trigram model construction, application of Laplace Smoothing, and system evaluation using *precision*, *recall*, and *F1-score* metrics. The system classifies words in sentences into three categories: standard**,** non-standard (recognized by SPECIL), and non-standard (low trigram probability).

Experimental results indicate that the proposed system effectively detects non-standard words with stable and reliable accuracy, even for rare word variations not frequently observed in the training corpus. The implementation of *Laplace Smoothing* successfully prevents zero probabilities for unseen trigrams, enhancing the model's ability to predict more natural word forms. The developed system can be implemented as a web-based application for spell-checking or automatic text normalization. Overall, this research contributes to advancing NLP technology for the Indonesian language by improving digital text quality and supporting the preservation of linguistic standardization.

Keywords: *non-standard word detection, trigram, Laplace smoothing, Indonesian language, NLP.*