ABSTRAK

PENDEKATAN BARU *STEMMING* KATA DAN AUGMENTASI DATA TEKS PADA BAHASA LAMPUNG SEBAGAI *LOW-RESOURCE LANGUAGE*

Oleh

ZAENAL ABIDIN

Penelitian ini mengembangkan pendekatan *natural language processing* (NLP) untuk bahasa Lampung berfokus pada *stemming* kata dialek Tulang Bawang dan *Lampung Language Dialect Identification* (LLDI) pada dialek *Api* dan *Nyo*. Untuk *stemming* kata, lima modifikasi metode diuji secara komparatif: modifikasi Nazief-Adriani (MNA), modifikasi *Confix-Stripping* (MCS), MCS disertai *N-Gram Stemming*, *Morphological-based*, dan *N-Gram Stemming*. Eksperimen pada 500 kata data uji dan 200 kata independen menunjukkan MCS disertai *N-Gram Stemming* mencapai performa terbaik dengan nilai *Gold Standard Assessment* (GSA) 98,8%. Pendekatan terbaik dari *stemming* kata diimplementasikan pada aplikasi *Direct Machine Translation* (DMT) Tulang Bawang–Indonesia dan hasil pengujian penerjemahan menunjukan nilai *Bilingual Evaluation Understudy* (BLEU) mencapai 80,07%. Untuk identifikasi dialek, model LLDI dibangun menggunakan *data set* 3000 kalimat dialek *Api* dan 9078 kalimat dialek *Nyo*. *Text Data Augmentation* (TDA) melalui metode permutasi kalimat diterapkan untuk mengatasi ketidakseimbangan data, menghasilkan *data set* sintesis sebesar *n*! dari *n token* sebuah kalimat. Empat pendekatan klasifikasi—*Naive Bayes, Logistic Regression, Support Vector Machine* (SVM), dan *Random Forest*—dievaluasi menggunakan *5-fold cross validation*. Hasil eksperimen menunjukkan SVM *balanced class* mencapai performa tertinggi dengan akurasi 97,4%, diikuti *Random Forest balanced class* dengan akurasi 96,9%. Penyeimbangan kelas terbukti meningkatkan deteksi dialek minoritas (*Api*) tanpa mengorbankan performa dialek mayoritas (*Nyo*). Sebaliknya, kondisi *unbalanced* menghasilkan *precision* tinggi namun *recall* rendah untuk dialek *Api*. Penelitian ini memberikan kontribusi awal yang signifikan dalam pengembangan *Natural Language Understanding* (NLU) pada bahasa Lampung dan berpotensi dilanjutkan pada penelitian fonologi, sintaksis, semantik dan pragmatik bahasa Lampung secara komputasi.

Kata Kunci : *Stemming*, Nazief-Adriani, *Confix-Stripping*, *N-Gram Stemming*, *Text Data Augmentation*, Permutasi, *Lampung Language Dialect Identification*

ABSTRACT

A NEW APPROACH TO WORD STEMMING AND TEXT DATA AUGMENTATION IN LAMPUNG LANGUAGE AS A LOW-RESOURCE LANGUAGE

By

ZAENAL ABIDIN

This study develops a natural language processing (NLP) approach for Lampung language focusing on stemming words in Tulang Bawang dialect and Lampung language dialect identification (LLDI) in *Api* and *Nyo* dialects. For word stemming, five approaches were tested comparatively: modified Nazief-Adriani (MNA), modified Confix-Stripping (MCS), MCS with N-Gram stemming, morphological-based, and N-Gram stemming. Experiments on 500 test words and 200 independent words showed that MCS with N-Gram stemming achieved the best performance with a Gold Standard Assessment (GSA) value of 98,8%. The best approach of word stemming was implemented in the Tulang Bawang-Indonesia Direct Machine Translation (DMT) application and the translation test results showed a bilingual evaluation understudy (BLEU) value of 80,07%. For dialect identification, the LLDI model was built using a dataset of 3000 *Api* dialect sentences and 9076 *Nyo* dialect sentences. Text data augmentation (TDA) using a sentence permutation approach was applied to address data imbalance, generating a synthetic dataset of n! from n sentence tokens. Four classification approaches—Naive Bayes, Logistic Regression, Support Vector Machine (SVM), and Random Forest—were evaluated using 5-fold cross-validation. Experimental results showed that the balanced SVM class achieved the highest performance with 97,4% accuracy, followed by the balanced Random Forest class with 96,9% accuracy. Class balancing was shown to improve the detection of the minority dialect (*Api*) without sacrificing the performance of the majority dialect (*Nyo*). Conversely, the unbalanced condition resulted in high precision but low recall for the Api dialect. This research provides a significant initial contribution to the development of natural language understanding (NLU) in Lampung and has the potential to be continued in computational research on the phonology, syntax, semantics, and pragmatics of Lampung.

Key Words : Stemming, Nazief-Adriani, Confix-Stripping, N-Gram Stemming, Text Data Augmentation, Permutasi, Lampung Language Dialect Identification.