STUDI KOMPARASI METODE SUPPORT VECTOR MACHINES

DAN RANDOM FOREST UNTUK KLASIFIKASI KATA KERJA

BAHASA JEPANG

Oleh

I Nyoman Sweta (2129101005)

Program Studi Ilmu Komputer

ABSTRAK

Bahasa Jepang memiliki sistem konjugasi kata kerja yang kompleks, di mana akhiran -ru dapat muncul pada semua kelompok konjugasi (K1, K2, K3), sehingga menimbulkan ambiguitas, khususnya pada kata kerja baru hasil serapan atau pembentukan kreatif. Proses pelabelan kelompok kata kerja ini umumnya dilakukan oleh anotator, namun tingkat kesepakatan antar anotator (inter-annotator agreement) dilaporkan relatif rendah ($\kappa \approx 0{,}68$) dan berpotensi menghasilkan label yang tidak konsisten. Penelitian ini bertujuan mengembangkan model klasifikasi kata kerja bahasa Jepang berbasis machine learning yang dapat dimanfaatkan sebagai alat rekomendasi label untuk membantu anotator meningkatkan konsistensi dan efisiensi pelabelan. Data yang digunakan berupa kata kerja dalam bentuk dasar (dictionary form) yang dipetakan ke tiga kelompok konjugasi utama (K1, K2, K3). Fitur yang digunakan adalah ciri morfologis sederhana, yaitu keberadaan akhiran -ru (ContainRU), satu huruf sebelum -ru (OcbRU), dan dua huruf sebelum -ru (TcbRU). Dua algoritma dibandingkan, yaitu Support Vector Machines (SVM) dan Random Forest (RF), dengan evaluasi menggunakan akurasi, precision, recall, F1-score, ROC-AUC, Log Loss, serta waktu pelatihan dan prediksi. Hasil penelitian menunjukkan bahwa kedua algoritma mencapai akurasi, precision, recall, dan F1-score sebesar 0,95. RF sedikit unggul pada ROC-AUC (0,9809 vs 0,9758), sedangkan SVM unggul pada Log Loss (0,2098 vs 0,3658), waktu pelatihan (0,1273 detik vs 0,147 detik), dan waktu prediksi (0,0068 detik vs 0,0077 detik). Berdasarkan kombinasi akurasi, reliabilitas probabilitas, dan efisiensi komputasi, SVM direkomendasikan sebagai algoritma yang lebih sesuai untuk mendasari alat rekomendasi label bagi anotator.

.

Kata Kunci: Bahasa Jepang, konjugasi kata kerja, *machine learning*, *Support Vector Machines*, *Random Forest*, rekomendasi label, anotator.

A COMPARATIVE STUDY OF SUPPORT VECTOR MACHINES

AND RANDOM FOREST METHODS FOR CLASSIFYING

JAPANESE VERBS

By

I Nyoman Sweta (2129101005)

Master's Program in Computer Science

ABSTRACT

Japanese verb conjugation is complex, as the -ru ending can appear in all conjugation groups (K1, K2, K3), creating ambiguity, especially in newly coined verbs derived from loanwords or creative word formation. The classification of verb groups is typically performed by human annotators; however, the inter-annotator agreement has been reported to be relatively low ($\kappa \approx 0.68$), leading to potential inconsistencies in labeling. This study aims to develop a machine learning-based Japanese verb classification model that can serve as a label recommendation tool to assist annotators in improving the consistency and efficiency of verb labeling.

The dataset consists of Japanese verbs in their dictionary form, mapped into three main conjugation groups (K1, K2, K3). The features used are simple morphological cues: the presence of the -ru ending (ContainRU), one character before -ru (OcbRU), and two characters before -ru (TcbRU). Two algorithms, Support Vector Machines (SVM) and Random Forest (RF), were compared and evaluated using accuracy, precision, recall, F1-score, ROC-AUC, Log Loss, as well as training and prediction time.

The results show that both algorithms achieved the same accuracy, precision, recall, and F1-score of 0.95. RF slightly outperformed SVM in ROC-AUC (0.9809 vs 0.9758), while SVM outperformed RF in Log Loss (0.2098 vs 0.3658), training time (0.1273 s vs 0.147 s), and prediction time (0.0068 s vs 0.0077 s). Considering the combination of accuracy, probability reliability, and computational efficiency, SVM is recommended as the most suitable algorithm for developing a label recommendation tool for annotators.

.

Keywords: Japanese language, verb conjugation, machine learning, SVM, Random Forest, label recommendation, annotator.