ABSTRAK

Bahasa Jawa merupakan salah satu bahasa daerah dengan jumlah penutur terbanyak di Indonesia, terutama di Pulau Jawa, dan memiliki potensi besar dalam pengembangan sistem pemrosesan bahasa alami. Namun, pengolahan teks Bahasa Jawa masih menghadapi beberapa tantangan seperti kesalahan pelabelan dan keterbatasan cakupan kata. Salah satu tahapan penting dalam pemrosesan bahasa alami adalah Part-of-Speech (POS) *tagging*, yaitu proses pelabelan kelas kata dalam suatu kalimat. Penelitian ini menerapkan metode Conditional Random Fields (CRF) untuk melakukan POS *tagging* pada Bahasa Jawa, karena CRF dinilai lebih akurat dibanding metode lain seperti Hidden Markov Model (HMM). Data yang digunakan pada penelitian ini berasal dari UD Javanese-CSUI, yang mencakup kalimat dari buku referensi tata bahasa, korpus WikiMatrix, dan berita daring dari Solopos. Hasil evaluasi menunjukkan bahwa pendekatan manual memiliki akurasi rendah (12–21%), sedangkan penggunaan *library* sklearn-crfsuite secara signifikan meningkatkan akurasi hingga 88,58%, baik dengan maupun tanpa penghapusan tag PUNCT dan SYM. Penelitian ini membuktikan bahwa metode CRF sangat efektif dalam melakukan POS *tagging* Bahasa Jawa, khususnya dengan dukungan fitur yang tepat dan pemanfaatan *library* yang optimal.

Kata kunci: Part-of-Speech Tagging, Bahasa Jawa, Conditional Random Fields.

ABSTRACT

Javanese is one of the regional languages with the largest number of speakers in Indonesia, particularly on the island of Java, and holds significant potential in the development of natural language processing (NLP) systems. However, processing Javanese text still faces several challenges, such as tagging errors and limited vocabulary coverage. One of the crucial steps in NLP is Part-of-Speech (POS) tagging, which involves labeling each word in a sentence according to its grammatical category. This study applies the Conditional Random Fields (CRF) method for POS tagging in the Javanese language, as CRF is considered more accurate than other methods such as the Hidden Markov Model (HMM). The dataset used in this research is UD Javanese-CSUI, which includes sentences from Javanese grammar reference books, the WikiMatrix corpus, and online news articles from Solopos. Evaluation results show that manual approaches yield low accuracy (12–21%), while the implementation using the sklearn-crfsuite library significantly improves accuracy up to 88.58%, with or without the removal of PUNCT and SYM tags. This research demonstrates that the CRF method is highly effective for POS tagging in the Javanese language, particularly when supported by appropriate feature engineering and optimal library implementation.

Keywords: *Part-of-Speech Tagging, Javanese, Conditional Random Fields.*