ABSTRAK

Penelitian ini mengkaji penerapan teknik *part of speech tagging* (POS *tagging*) untuk Bahasa Jawa menggunakan metode Support Vector Machine (SVM). Bahasa Jawa, yang digunakan oleh mayoritas penduduk Pulau Jawa, memiliki potensi besar dalam bidang *Natural Language Processing* (NLP). Dalam penelitian ini, data Bahasa Jawa yang digunakan berasal dari Universal Dependencies (UD) Javanese CSUI, yang mencakup 1000 kalimat dan 14.000 kata dengan anotasi manual. Data ini akan melalui tahap *preprocessing* untuk membersihkan dan menyiapkan data sebelum digunakan dalam model. Fitur yang digunakan meliputi TF-IDF, gabungan TF-IDF dan atribut lainnya, serta *pre-trained embeddings* seperti FastText dan BERT. Pembagian data dilakukan menggunakan *K-Fold Cross Validation* dengan nilai k = 5. Model SVM kemudian dilatih menggunakan empat jenis *kernel* yang berbeda, yaitu linear, RBF, *polynomial*, dan *sigmoid*. Hasil penelitian menunjukkan bahwa kombinasi fitur gabungan TF-IDF dan atribut lainnya dengan *kernel* linear menghasilkan akurasi tertinggi sebesar 87,42%, dengan *precision* 87,40%, *recall* 87,42%, dan *F1-score* 87,28%. Hasil penelitian ini diharapkan dapat berkontribusi dalam penelitian dan pengembangan di bidang NLP untuk Bahasa Jawa.

Kata Kunci: *Part of Speech Tagging* (POS *Tagging*), Bahasa Jawa, Support Vector Machine (SVM), *Natural Language Processing* (NLP)

ABSTRACT

This study examines the application of part-of-speech tagging (POS tagging) techniques for Javanese using the Support Vector Machine (SVM) method. Javanese, which is spoken by the majority of the population of Java Island, has great potential in the field of Natural Language Processing (NLP). In this study, the Javanese language data used comes from the Universal Dependencies (UD) Javanese CSUI corpus, which includes 1,000 sentences and 14,000 words with manual annotations. This data will undergo a preprocessing stage to clean and prepare the data before it is used in the model. The features used include TF-IDF, a combination of TF-IDF and other attributes, as well as pre-trained embeddings such as FastText and BERT. Data division was performed using K-Fold Cross Validation with a value of k = 5. The SVM model was then trained using four different kernels, namely linear, RBF, polynomial, and sigmoid. The results show that the combination of TF-IDF and other attributes with a linear kernel produces the highest accuracy of 87.42%, with precision of 87.40%, recall of 87.42%, and F1-score of 87.28%. The results of this study are expected to contribute to research and development in the field of NLP for Javanese.

Keywords : Part of Speech Tagging (POS Tagging), Javanese, Support Vector Machine (SVM), Natural Language Processing (NLP)