ABSTRAK

Penelitian ini membahas penerapan metode Hidden Markov Model (HMM) dengan algoritma Viterbi untuk melakukan *Part-of-Speech* (POS) *Tagging* pada teks berbahasa Jawa. Bahasa Jawa dipilih karena statusnya sebagai Bahasa daerah dengan jumlah penutur yang besar namun dengan representasi sumber daya digital yang rendah. *Dataset* yang digunakan berasal dari Universal Dependencies (UD) dengan lebih dari 14.000 kata beranotasi POS. Proses penelitian dimulai dengan tahap *preprocessing* berupa *case folding, tokenizing*, dan *data cleaning*. Kemudian dilakukan perhitungan probabilitas transisi dan emisi dalam model HMM, serta digunakan algoritma Viterbi untuk menentukan urutan *tag* yang paling mungkin. Evaluasi dilakukan menggunakan metode *K-Fold Cross Validation* dengan nilai k = 5 dan k = 10. Hasil menunjukkan bahwa penghapusan *tag* "SYM" dan "PUNCT" menyebabkan penurunan akurasi, dengan rata-rata akurasi turun dari sekitar 83–85% menjadi 67–70%. Hal ini menunjukkan bahwa meskipun tampak tidak signifikan, keberadaan simbol dan tanda baca membantu dalam menjaga konteks struktural kalimat. Penelitian ini menunjukkan bahwa model probabilistik HMM dengan algoritma Viterbi cukup efektif dalam melakukan POS *Tagging* pada Bahasa Jawa, dan kehadiran elemen linguistik seperti tanda baca berkontribusi terhadap peningkatan akurasi model.

Kata kunci: *Part-of-Speech Tagging*, Bahasa Jawa, Hidden Markov Model, Viterbi, *K-Fold Cross Validation*.

ABSTRACT

This research discusses the application of Hidden Markov Model (HMM) method with Viterbi algorithm to perform Part-of-Speech (POS) Tagging on Javanese text. Javanese was chosen because of its status as a regional language with a large number of speakers but with low representation of digital resources. The dataset used comes from from Universal Dependencies (UD) with more than 14,000 POS annotated words. The research process begins with a preprocessing stage in the form of case folding, tokenizing, and data cleaning. Then the calculation of transition and emission probabilities in the HMM model is carried out, and the Viterbi algorithm is used to determine the most likely tag order. Evaluation was performed using the K-Fold Cross Validation method with values of k = 5 and k = 10. Results show that the removal of the "SYM" and "PUNCT" tags led to a decrease in accuracy, with the average accuracy dropping from approximately 83 85% to 67-70%. This suggests that although it may not seem significant, the presence of symbols and punctuation helps in maintaining the structural context of the sentence. This study shows that the probabilistic HMM model with Viterbi algorithm is quite effective in performing POS Tagging on the Javanese language, and the presence of linguistic elements such as punctuation marks contributes to the improvement of the model's accuracy.

***Keywords**: Part-of-Speech Tagging, Javanese Language, Hidden Markov Model, Viterbi, K-Fold Cross Validation.*