ABSTRAK

Indonesia memiliki berbagai keragaman salah satunya memiliki 718 bahasa daerah termasuk Bahasa Jawa yang digunakan oleh banyak masyarakat. Namun dalam era modern, penggunaan Bahasa Jawa mengalami tantangan identitas akibat pengaruh budaya asing dan perubahan sosial. Salah satu upaya pelestarian bahasa daerah adalah melalui penelitian kebahasaan menggunakan *Natural Language Processing* (NLP), khususnya *Part of Speech tagging* yang memberi label kelas kata untuk memahami struktur dan makna teks. Salah satu metode yang digunakan dalam *POS tagging* adalah Hidden Markov Model (HMM) yang dapat ditingkatkan menggunakan algoritma Baum-Welch. Data yang digunakan berasal dari korpus UD Javanese CSUI yang terdiri dari 17 *tag* serta 1.000 kalimat dengan total 14.000 kata dalam Bahasa Jawa Ngoko dan sebagian dalam Bahasa Jawa Krama. Data ini akan melalui tahap *pre-processing* dan dibagi ke dalam beberapa subset menggunakan K-Fold Cross-Validation dengan nilai k = 5 dan k = 10. Kemudian, model HMM dilatih menggunakan algoritma Baum-Welch dan akurasi diukur berdasarkan lima nilai *threshold* yang berbeda yaitu 10^{-2}, 10^{-4}, 10^{-10}, 10^{-20}, dan 10^{-30}. Hasil penelitian menunjukkan bahwa akurasi terbaik adalah sebesar 74,92% yang diperoleh pada skenario k = 5 serta pada nilai *threshold* 10^{-2} dan 10^{-4}, dengan menggunakan 17 *tag*.

Kata Kunci : Bahasa Jawa, *Part of Speech Tagging*, Hidden Markov Model (HMM), algoritma Baum-Welch.

ABSTRACT

Indonesia has a variety of diversity, one of which has 718 regional languages including Javanese which is used by many people. However, in the modern era, the use of Javanese faces identity challenges due to foreign cultural influences and social changes. One of the efforts to preserve the Javanese language is through linguistic research using *Natural Language Processing* (NLP), especially Part of Speech tagging which labels word classes to understand the structure and meaning of text. The data used comes from the UD Javanese CSUI corpus which consists of 17 tags and 1,000 sentences with a total of 14,000 words in Javanese Ngoko, with some in Javanese Krama. The data will go through text preprocessing and divided into subsets using K-Fold Cross-Validation with k = 5 and k = 10. Then, the HMM model is trained using the Baum-Welch algorithm, and accuracy is measured based on five different *threshold* values of 10^{-2}, 10^{-4}, 10^{-10}, 10^{-20}, dan 10^{-30}. The results showed that the best accuracy was 74.92% obtained in the k = 5 scenario and at *threshold* values of 10^{-2} and 10^{-4}, using 17 tags.

Keywords : Javanese, Part of Speech Tagging, Hidden Markov Model (HMM), Baum-Welch algorithm.