

ABSTRAK

Stopword merupakan kata yang tidak memiliki arti penting pada sebuah dokumen. Penghapusan *stopword* dinilai mampu untuk meringkas dokumen. *Stopword* Bahasa Jawa belum tersedia hingga saat ini. Menurut (Tsz-Wai Lo et al., 2005), Term Based Random Sampling merupakan suatu metode pembuatan *stopword* otomatis yang menghitung seberapa informatif sebuah istilah, sehingga dapat menghasilkan beberapa kumpulan *stopword* yang berbeda.

Penelitian ini bertujuan untuk membuat *stopword* bahasa Jawa menggunakan metode Term Based Random Sampling dengan rumus perhitungan Kullback Leibler. Untuk menilai performa yang lebih baik, perhitungan Kullback Leibler menggunakan TF dan TF-IDF. Selain itu digunakan juga parameter Term Based Random Sampling, yaitu parameter Y untuk menentukan jumlah iterasi dan parameter L untuk menentukan banyak *stopword* yang digunakan. Parameter Y yang digunakan adalah 50. Dokumen referensi dalam penelitian ini juga dibagi menjadi dua, yaitu dokumen referensi yang diambil selain dari dokumen spesifik dan dokumen referensi yang diambil dari Djaka Lodhang. Dalam pengujian analisis sentimen NusaX tanpa *stopword* didapatkan hasil macro F1 score 69,4%. Dari percobaan yang telah dilakukan, hasil terbaik penggunaan *stopword* Bahasa Jawa untuk analisis sentimen adalah dengan perhitungan TF-IDF, parameter L=10%, dokumen referensi yang diambil selain dari dokumen spesifik, dan hasilnya adalah 71,29%.

Kata kunci: *Stopword*, Bahasa Jawa, Term Based Random Sampling, Kullback Leibler, Naive Bayes, analisis sentimen

ABSTRACT

Stopwords are words that do not carry important meaning in a document. Removing stopwords is considered capable of summarizing documents. Javanese stopwords are not yet available to date. According to Tsz-Wai Lo et al. (2005), Term Based Random Sampling is a method for automatically generating stopwords by calculating how informative a term is, thereby producing several different sets of stopwords.

This study aims to generate Javanese stopwords using the Term Based Random Sampling method with the Kullback–Leibler divergence formula. To evaluate better performance, the Kullback–Leibler calculation is conducted using TF and TF-IDF. In addition, the Term Based Random Sampling parameters are used, namely parameter Y to determine the number of iterations and parameter L to determine the number of stopwords used. The value of parameter Y used is 50. The reference documents in this study are also divided into two types: reference documents taken from sources other than the specific documents, and reference documents taken from Djaka Lodhang. In sentiment analysis experiments on the NusaX dataset without using stopwords, a macro F1 score of 69.4% was obtained. Based on the experiments conducted, the best result for using Javanese stopwords in sentiment analysis was achieved using TF-IDF, parameter L = 10%, reference documents taken from sources other than the specific documents, yielding a macro F1 score of 71.29%.

Keyword: Stopword, Javanese language, Term Based Random Sampling, Kullback Leibler, Naive Bayes, sentiment analysis