

ABSTRAK

Stemming adalah sebuah proses untuk menghilangkan imbuhan-imbuhan yang terdapat pada sebuah kata biasanya diaplikasikan pada kasus pencarian teks, penerjemahan bahasa, ringkasan dokumen dan pada klasifikasi teks. Teknik *stemming* berbeda untuk masing-masing bahasa. Seperti *stemming* untuk teks berbahasa Inggris berbeda dengan teks berbahasa Indonesia. Pada bahasa Indonesia, *stemming* sangat penting sekali; Imbuhan pada kata-kata bahasa Indonesia terdiri dari awalan, akhiran, sisipan dan gabungan dari awalan, akhiran yang menyebabkan mencocokkan kata-kata tersebut sulit.

Pada tugas akhir saya ini diimplementasikan ketiga algoritma untuk *stemming* bahasa Indonesia yaitu algoritma Nazief&Adriani, algoritma Arifin&Setiono, algoritma Vega serta akan dibandingkan performa dari ketiga algoritma *stemming* tersebut. Parameter performa *stemming* yang digunakan adalah banyaknya jumlah term yang di-*stemming* dengan benar dibagi dengan jumlah seluruh term. Hasil dari pengujian menunjukkan bahwa *stemming* yang menggunakan algoritma Nazief&Adriani yang dikembangkan memiliki nilai performa sebesar 96.5447%. Oleh karena itu dapat disimpulkan bahwa *stemming* untuk teks berbahasa Indonesia sebaiknya dilakukan menggunakan pendekatan algoritma Nazief and Adriani yang telah dikembangkan.

Pada tugas akhir ini juga akan dianalisis pengaruh proses *stemming* terhadap kategorisasi. *Stemming* dilakukan dengan algoritma Nazief&Adriani, algoritma Arifin&Setiono, algoritma Vega dan *stemming* murni yang dilakukan secara manual. Parameter performa kategorisasi yang digunakan adalah akurasi dan *f-measure*. Kemudian masing-masing akurasi dari keempat kasus diatas akan dibandingkan dengan akurasi pada dataset yang tidak di-*stemming*. Kategorisasi dilakukan dengan metode Multinomial Naïve Bayes. Hasil percobaan menunjukkan bahwa dataset yang di-*stemming* murni secara manual memiliki nilai akurasi kategorisasi yang paling baik yaitu sebesar 97.5% dibandingkan 4 kasus lainnya (*stemming* dengan Algoritma Nazief&Adriani, *stemming* dengan algoritma Arifin&Setiono, *stemming* dengan algoritma Vega dan dataset yang tidak di-*stemming*). Tetapi untuk ketiga algoritma *stemming* yang diimplementasikan, algoritma Nazief&Adriani memiliki akurasi kategorisasi yang paling baik yaitu 95%. Berdasarkan percobaan yang telah dilakukan maka dapat disimpulkan bahwa proses *stemming* sangat berpengaruh pada proses kategorisasi karena proses *stemming* dapat meningkatkan akurasi kategorisasi.

Kata Kunci: *Stemming, Kategorisasi Teks, Akurasi, f-measure*