Abstrak

Salah satu tujuan dari *Natural Languange Processing* adalah mengidentifikasi parafrasa, yang berarti untuk mengajarkan kepada mesin apakah sebuah kalimat memiliki makna yang sama dengan kalimat lainnya. Seperti contoh, kalimat ”Disaat hujan turun” akan dikenali mesin sebagai kalimat yang serupa dengan ”Ketika hari hujan”. Parafrasa berarti pengungkapan kembali suatu tuturan dari sebuah tingkatan atau macam bahasa menjadi yang lain tanpa merubah pengertian. Dalam penelitian ini dilakukan klasifikasi untuk menentukan apakah dua kalimat Bahasa Indonesia termasuk kedalam parafrasa atau non-parafrasa. Tahapan yang dilakukan mencakup *preprocessing*, pembangunan *classifier*, dan evaluasi performansi.

Proses *preprocessing* terdiri dari *tokenization, normalization, stopword removal*, dan *stemming.* Hasil *preprocessing* yang ada kemudian dilakukan proses *feature extraction* untuk menemukan fitur-fitur untuk membangun *classifier*. Fituryang diekstraksi terdiri dari dua jenis fitur, yaitu *Levhenstein Distance* untuk fitur sintaktik dan *Wu and Palmer* untuk fitur semantik. Setelah dilakukan proses ekstraksi, dataset kemudian dibagi kedalam dua bagian yaitu Data *training* dan Data *testing.* Data *training* digunakan untuk melatih *classifier*, sedangkan data *testing* digunakan untuk menguji kinerja *classifier*. Setelah data dibagi, maka dilakukan proses pembangunan *classifier* dengan menggunakan *Naïve Bayes.* Performansi terbaik dari sistem menghasilkan akurasi 0.713, presisi 0.688, *recall* 0.798, dan *F1-Measure* 0.735.

Kata kunci :*Naive Bayes*, identifikasi parafrasa, *preprocessing.*