Abstrak

Pada *natural language processing* identifikasi parafrasa merupakan proses yang penting oleh karena itu diperlukan mesin untuk membedakan secara otomatis frasa-frasa yang berbeda bentuk namun memiliki makna yang sama. Misalnya pada kalimat “penyebab kebakaran hutan”, seharusnya komputer akan mengenali bahwa kalimat tersebut serupa dengan kalimat “sumber kebakaran hutan”. Parafrasa sendiri merupakan pengungkapan kembali suatu tuturan dari sebuah tingkatan atau macam Bahasa menjadi yang lain tanpa mengubah pengertian; Parafrasa dapat diartikan juga sebagai penguraian kembali suatu teks dalam bentuk yang lain, dengan maksud untuk dapat menjelaskan makna yang tersembunyi. Pada penelitian ini dilakukan klasifikasi dua kalimat bahasa Indonesia apakah termasuk parafrasa atau bukan parafrasa. Tahapan yang dilakukan ada tiga yaitu proses *preprocessing,* pembangunan *classifier* dan evaluasi performansi.

Proses *preprocessing* terdiri dari tiga tahap yaitu *tokenization, non-alphanumerical removal,* dan *stemming*. Data hasil *preprocessing* tersebut lalu dilakukan proses *feature extraction* yang bertujuan untuk membangun fitur-fitur baru dari data set tersebut. Fitur yang pertama adalah fitur sintaktik yang merupakan hasil dari perhitungan jarak antara dua kalimat, perhitungan jarak tersebut menggunakan metode *Normalized Levhenstein Distance.* Fitur yang kedua adalah fitur semantik, fitur ini menghitung kemiripan pasngan kalimat berdasarkan pohon semantik, perhitungan jarak semantik dilakukan dengan menggunakan metode Wu *and* Palmer. Setelah dilakukan ekstraksi fitur, dataset tersebut dibagi menjadi dua bagian yaitu data *training* dan data *testing.* Data *training* digunakan untuk melatih *classifier,* sedangkan data *testing* digunakan untuk menguji performansi *classifier*. Setelah data selesai dibagi, maka dilakukan diskritisasi nilai fitur dengan *clustering* menggunakan metode K-Means. Metode yang digunakan untuk melatih *classifier* adalah *Bayesian Networks*. Perhitungan parameter yang digunakan *classifier* ini adalah MAP(*Maximum A Posteriori*) dan *Multinomial Distribution Probability*. Hasil dari pengujian data *testing* terhadap *classifier* yang didapatkan nilai *Precision* 61.2%, *Recall* 84.8%, Akurasi 66.2%, and *F1-Measure* 71.5%.

Kata kunci: identifikasi parafrasa, *preprocessing*, *bayesian networks*, MAP