Abstrak

Saat ini data berupa teks sangat banyak dijumpai. Teks bahasa manusia tersebut merupakan data yang tidak terstruktur atau semi terstrukur sehingga diperlukan proses konversi data teks yang tidak terstruktur tersebut menjadi representasi teks yang terstruktur. Pada proses konversi tersebut, pembobotan *term* merupakan salah satu bagian yang biasanya dilakukan.

Belum ada studi perbandingan yang komprehensif pembobotan *term* untuk korpus-korpus berbahasa Indonesia. Padahal bahasa Indonesia adalah bahasa nasional bangsa Indonesia, dan merupakan salah satu bahasa yang digunakan oleh banyak orang.

Untuk itu, dalam Tugas Akhir ini dilakukan sebuah studi perbandingan yang komprehensif terhadap metoda-metoda pembobotan *term* untuk korpus-korpus bahasa Indonesia. Studi dibatasi pada penerapan untuk klasifikasi, sebuah pekerjaan data mining yang penting. Melakukan penelitian pengaruh pemilihan pembobotan *term* berpengaruh terhadap efektivitas klasifikasi. Ada dua sisi mengapa kami menyatakan sebagai sebuah studi yang komprehensif. Pertama, metoda yang akan diteliti adalah seluruh metode yang kami jumpai yaitu *TF*, *IDF*, *ITF*, *TF-IDF*, *RF*, *TF-RF*, *OR*, *TF-OR*, *IG*, *TF-IG*, *NGL*, *GSS*, *QF*, *IQF*, *ICF*, *VRF*, *QF-IQF-ICF*, *CHI-SQUARE*, *TF-X^2* , *BINARY*, dan *MI*. Kedua, korpus bahasa Indonesia yang akan dipakai beragam. Dari korpus yang menggunakan penulisan formal seperti artikel berita, non formal seperti email, yang relatif panjang seperti artikel berita maupun yang pendek seperti tweet.

Pada penelitian ini menunjukkan bahwa secara keseluruhan metode *ngl* memiliki performansi dengan *nilai f-measure* lebih baik dibandingkan metode yang lainya untuk korpus bahasa Indonesia, khususnya untuk proses klasifikasi, dan diketahui bahwa nilai performansi yang didapatkan dari setiap metode pembobotan bergantung pada data dan *classifier* yang dipakai.

Kata Kunci : *text preprocessing*, *term*, *term weighting*.