Abstrak

Pengkategorian dokumen merupakan salah satu permasalahan dalam *text mining*. Salah satu cara agar suatu dokumen dapat dikategorikan adalah dengan menggunakan teknik klasifikasi. Sekumpulan dokumen selain memiliki *feature space* yang berdimensi tinggi dapat juga memiliki sifat data yang *imbalance*. Sifat *imbalance* tersebut akan mengakibatkan klasifikasi data yang akan dibentuk kurang akurat. Untuk meningkatkan efisiensi dan keakuratan dalam klasifikasi dokumen, salah satunya dengan menggunakan teknik *feature selection.*

Pada tugas akhir ini dilakukan analisis perbandingan metode *feature selection* antara lain *Odds Ratio* (OR), *GSS Coefficient*, *Information Gain* (IG), *improved OR* (iOR), dan *improved SIG* (iSIG). Metode-metode *feature selection* tersebut diterapkan secara *filter feature selection* sedangkan pada *wrapper feature selection* menerapkan *Odds Ratio* (OR). Penerapan dilakukan menggunakan teknik klasifikasi *multinomial naive bayes,* metode tersebut menggunakan algoritma *naive bayes* dengan memperhitungkan jumlah kemunculan kata dalam dokumen. Selain menggunakan *multinomial naive bayes*, pada penerapan *filter feature selection* dilakukan juga proses pengklasifikasian dokumen menggunakan *software* Weka 3.5. Dengan melakukan analisis perbandingan metode *feature selection* diketahui metode mana yang paling handal dalam menangani data *imbalance* dengan menguji tingkat akurasi data setelah dilakukan klasifikasi dengan *test set* yang diberikan. Data yang digunakan berasal dari Reuters 21578 dengan dokumen bersifat *multi-label.*

Kata Kunci: *text mining,* klasifikasi*, imbalance*, *feature selection, multinomial naive bayes.*