Abstrak

Kata-kata dalam suatu dokumen yang sering muncul tapi kurang berarti dalam proses kategorisasi disebut sebagai *stopword.* Untuk kata-kata yang dikategorisasikan ke dalam *stopwords* dianggap tidak memiliki kontribusi dalam proses kategorisasi, seharusnya dihapus sewaktu pengindeksan sebelum proses kategorisasi dilakukan. Bagaimanapun, penggunaan satu daftar *stopword* untuk koleksi dokumen yang berbeda-beda bisa mengurangi performansi dari peng-kategorisasian.

Pada tugas akhir ini digunakan pendekatan *Term-Based Random Sampling* menghasilkan daftar *stopword* secara otomatis untuk dokumen yang diberikan. Pendekatan ini, menentukan seberapa besar informasi yang dimiliki suatu kata (*term*). Dengan ini akan bisa ditentukan suatu daftar *stopword* secara otomatis. Dalam tugas akhir ini digunakan koleksi dokumen Reuter. Untuk daftar *stopword* yang dihasilkan akan dievaluasi dengan melakukan kategorisasi pada dokumen yang menggunakan daftar *stopword* yang dihasilkan.

Pendekatan ini juga nanti akan dievaluasi dengan membandingkan hasil performansi kategorisasi yang dihasilkan dengan pre-proses pembuangan daftar *stopword* menggunakan daftar *stopword* yang dihasilkan dengan menggunakan pendekatan ini, hasil performansi kategorisasi menggunakan daftar *stopword* Salton and Buckley I dan Salton and Buckley II, Google *stopword*, *defalutl English Stopword* dan hasil performansi kategorisasi tanpa menggunakan pre-proses pembuangan *stopword.*

Dari hasil evaluasi yang dilakukan, daftar *stopword* yang lebih efektif bisa diperoleh dengan menggunakan pendekatan Term-Based Random Sampling. Dengan akurasi pengkategorisasi sebesar 88.24%.

Keyword
Kategorisasi, *term, stopword*