Abstrak

Salah satu permasalahan yang signifikan dalam kategorisasi teks adalah dimensionalitas data yang sangat tinggi yang menyebabkan waktu pemrosesan menjadi lebih lama. Salah satu cara untuk mengatasi hal tersebut adalah dengan melakukan *feature selection*. *Feature selection* dilakukan untuk memilih fitur-fitur penting dan relevan terhadap data dan membuang fitur-fitur yang tidak berpengaruh. Dengan demikian, dimensionalitas data dapat dikurangi.

Dalam tugas akhir ini, permasalahan yang diangkat berkaitan dengan *feature selection* menggunakan *Information Gain* dan *Chi-Square*, pada kategorisasi teks dengan *classifier* N-gram. Kategorisasi dilakukan dengan menghitung jarak profil kategori ke profil dokumen, di mana profil dibentuk dari fitur-fitur yang ada. Sehingga jumlah fitur sangat mempengaruhi waktu yang dibutuhkan dalam proses kategorisasi. Hasil *F-measure* yang didapatkan pada kategorisasi teks dengan N-gram tanpa *feature selection* adalah 0.89, di mana gram yang digunakan adalah 2-gram. Dan ketika mengalami *feature selection* dengan menggunakan *Information Gain* sebanyak 80 %, *F-Measure* meningkat menjadi 0.935, serta ketika mengalami *feature selection* sebanyak 20 % dengan *Chi-Square*, *F-Measure* meningkat menjadi 0.94.

Proses pemilihan fitur dengan menggunakan *Information Gain* lebih cepat dibandingkan dengan *Chi-Square*. Akan tetapi, secara keseluruhan performansi yang dihasilkan oleh fitur-fitur hasil pemilihan *Chi-Square* memberikan hasil yang lebih baik.

Kata kunci: *feature selection*, 2-gram, *Information Gain*, *Chi-Square*, *F-Measure*.