Abstrak

Kategorisasi teks (atau juga dikenal dengan klasifikasi teks) adalah suatu task yang mengurutkan kumpulan dokumen D kedalam kategori yang telah ditentukan secara otomatis ke dalam kategori C. Dalam kategorisasi teks salah satu prosesnya adalah teks preprocessing yang termasuk didalamnya meliputi feature selection dan term weighting.

Salah satu metode pembobotan yang dikenal adalah TF-IDF dimana dalam metode ini setiap term/kata dalam sebuah dokumen dihitung frekuensinya dalam sebuah dokumen (*term frequency*) yang kemudian hasilnya dikombinasikan dengan frekuensi kemunculan term pada suatu kumpulan dokumen (*inverse document frequency*). Term yang sering muncul pada dokumen tapi jarang muncul pada kumpulan dokumen memberikan nilai bobot yang tinggi. TF-IDF akan meningkat dengan jumlah kemunculan term pada sebuah dokumen dan berkurang dengan jumlah term yang muncul pada kumpulan dokumen.

Namun mengingat kategorisasi teks bersifat terawasi dimana menggunakan dataset yang dibagi menjadi dataset training dan dataset testing, maka diperlukan suatu metode yang memenuhi syarat diatas. Dalam konteks standar Information Retrieval, asumsi IDF cukup beralasan karena dapat menginterpretasikan term dengan baik karena term yang sering muncul dalam banyak dokumen adalah diskriminator yang tidak baik. Tapi ketika data training untuk query tersedia, cara yang lebih baik harus digunakan yang dapat membedakan term yang terdistribusi ke dalam kumpulan data training baik kategori positif maupun negative. Data training tidak tersedia dalam query di konsep standar IR, namun lebih sering tersedia untuk kategori dalam konteks TC, dimana gagasan "relevansi dengan query" digantikan dengan "keanggotaan dalam kategori". Maka dari itu digunakanlah *Category-based Function* yang ada pada *Term Evalution Function* seperti *Chi-square, Information Gain (IG),* dan *Gain Ratio (GR)* sebagai pengganti fungsi IDF pada TF-IDF. Metode ini disebut *Supervised Term Weighting*. Dan metode inilah yang digunakan dalam Skripsi ini.

Pada metode ini dikombinasikan antara nilai *term evaluation function* dari setiap term yang terpilih dengan nilai *term frequency* di setiap dokumen kemudian dilakukan klasifikasi dengan metode SVM dan dievaluasi dengan parameter precison, recall, f-measure dan akurasinya. Hasil penelitian menunjukkan bahwa metode *Supervised Term Weighting* memberikan performansi yang lebih baik dibandingkan TF-IDF khususnya pada threshold local policy.

Kata kunci : kategorisasi teks, *term frequency, inverse document frequency, term weighting, supervised term weighting*.