1. Pendahuluan

1.1 Latar belakang masalah

Saat ini informasi artikel berbahasa Indonesia berbasiskan web semakin banyak jumlahnya. Hal ini menyebabkan diperlukannya suatu kategorisasi terhadap artikel tersebut untuk memudahkan pembaca dalam mencari topik berita yang mereka inginkan. Salah satu cara yang dapat dilakukan sebagai solusi untuk masalah ini adalah dengan menggunakan proses klasifikasi teks dalam *text mining.*

Informasi yang akan digali pada *text mining* memiliki struktur sembarang. Oleh karena itu, diperlukan proses pengubahan bentuk menjadi data yang terstruktur sesuai kebutuhannya untuk pengolahan lebih lanjut. Proses ini sering disebut *text preprocessing* yang terdiri dari tahap *case folding*, *tokenization*, dan *filtering*. Setelah data menjadi data terstruktur, lalu dilakukan proses *term weighting* untuk memberikan bobot pada setiap *term* yang ditemukan pada sekumpulan dokumen teks. Bobot ini menyatakan kepentingan/kontribusi *term* terhadap suatu dokumen.

Term frequency telah lama digunakan sebagai metode pembobotan *term* pada dokumen teks [5]. Dalam metode ini, tiap *term* diasumsikan memiliki nilai kepentingan yang sebanding dengan jumlah kemunculan *term* tersebut pada dokumen teks. Hal inilah yang menjadi kelemahan utama dari *term frequency* sehingga mengakibatkan *term* menjadi independen dan mengabaikan ketergantungan yang mungkin ada antar *term* dalam dokumen. Hal ini mungkin efektif untuk mengambil *term-term* yang relevan dalam konteks lokal, tapi tidak dalam konteks global dimana keberadaan suatu *term* berpengaruh pada teks secara keseluruhan.

Kelemahan metode *term frequency* diatas dapat diatasi dengan mengimplementasikan metode pembobotan *random-walk* [11]. Metode ini direpresentasikan dengan algoritma perankingan berbasis graf yang diterapkan dalam graf tekstual yang dapat mengintegrasikan keterhubungan antar *term* dan konteks sekitarnya. Pada perankingan graf tekstual, teks direpresentasikan menjadi sebuah graf. *Vertex/node* pada graf tekstual adalah unit teks yang akan diranking, yaitu berupa *term-term* dalam dokumen. *Edge/link* dalam graf menunjukkan keterhubungan yang bermakna antar *vertex/node*.

Penelitian akan dilakukan dengan menerapkan metode *random walk* dan *term frequency* pada beberapa dataset artikel berita berbahasa Indonesia. Ada dua skema yang akan digunakan, yaitu skema *tf - rw* sebagai skema pertama dan *tf.idf - rw.idf* sebagai skema kedua. Setelah itu dilakukan proses klasifikasi dokumen dengan menggunakan klasifier pada *tools* Weka. Analisa difokuskan pada pengaruh *random walk* dan *term frequency* terhadap performansi klasifier berdasarkan nilai *akurasi* dan *macro-average F-measure.*

1.2 Perumusan masalah

Objek penelitian pada Tugas Akhir ini adalah implementasi metode *random-walk* dan *term frequency* untuk menghitung bobot term dalam suatu dokumen. Akan tetapi, tugas akhir ini diutamakan kepada analisis karakteristik *random-walk* dan *term frequency* terhadap performansi klasifier berdasarkan nilai *akurasi* dan *macro-average F-measure*.

Adapun batasan masalah yang dipakai pada Tugas Akhir ini adalah :

1. Dataset yang digunakan adalah artikel berita berbahasa Indonesia yang diperoleh dari web, bersifat *offline* dan disimpan dalam *file* berekstensi .txt.
2. Tidak melakukan *stemming* pada tahap *text preprocessing* dengan pertimbangan bahwa penggunaan *stemming* tidak terlalu berpengaruh pada proses pembobotan.
3. Proses klasifikasi yang akan dilakukan pada dokumen dengan menggunakan klasifier pada *tools data mining* yaitu Weka.

1.3 Tujuan

Tujuan dari Tugas Akhir ini adalah :

1. Mengimplementasikan *random-walk* dan *term frequency* untuk menghitung bobot *term* dalam suatu dokumen.
2. Melakukan klasifikasi dokumen terhadap dataset yang telah dilakukan pembobotan *random-walk* dan *term frequency*.
3. Menganalisis pengaruh *random-walk* dan *term frequency* terhadap performansi klasifier berdasarkan nilai *akurasi* dan *macro-average F-measure*. Analisis tersebut dilakukan pada skema *tf - rw* dan skema *tf.idf - rw.idf*.

1.4 Metodologi penyelesaian masalah

Metodologi penyelesaian masalah dalam Tugas Akhir ini adalah :

1. Studi literatur
   Mencari referensi dan sumber-sumber yang berhubungan dengan permasalahan yang ada seperti *text mining*, *text classification*, *term weighting*, dll.
2. Pencarian dan pengumpulan data
   Data yang akan digunakan berupa artikel berita berbahasa Indonesia yang diambil dari web.
3. Analisis kebutuhan dan implementasi sistem
   Analisis kebutuhan dilakukan dengan pembuatan sistem kebutuhan perangkat lunak. Melakukan implementasi sistem dengan membangun perangkat lunak sesuai dengan perancangan yang telah dilakukan.
4. Pengujian sistem dan analisa hasil
   Pengujian dilakukan terhadap beberapa metode pembobotan *term* berdasarkan parameter yang telah didefinisikan di awal.
5. Pengambilan keputusan dan penyusunan Tugas Akhir
   Melaporkan semua yang telah dilakukan selama implementasi dan pengujian dalam penulisan Tugas Akhir ini.