1. Pendahuluan

1.1 Latar belakang

Text mining secara luas dapat diartikan sebagai proses penemuan informasi yang sebelumnya tidak terungkap dengan memproses dan menganalisa data dalam jumlah besar. Tujuan dari *text mining* adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen. Jadi, dokumen yang diproses pada *text mining* biasanya berupa *unstructured text*. *Text mining* mengadopsi dan mengembangkan banyak teknik dan solusi dari bidang lain, seperti *Data Mining, Machine Learning, Natural Language Processing (NLP), Information Retrieval (IR),* dan *Knowledge Management*. [1,3,9]

Biasanya, algoritma klasifikasi menggunakan Vector Space Model (VSM) untuk mengkodekan dokumen. VSM mengaitkan *term* dengan dokumen, dan karena *term* yang berbeda memiliki kepentingan yang berbeda dalam suatu dokumen, sebuah *term weights* berhubungan dengan setiap *term* [13]. *Term weights* ini berasal dari frekuensi dari sebuah *term* dalam dokumen atau kumpulan dokumen.

Banyak skema pembobotan *term* yang telah diajukan [10,13,17]. Sebagian besar metode yang ada tersebut bekerja berdasarkan asumsi bahwa seluruh kumpulan data tersedia dan statis. Sebagai contoh, untuk menggunakan pendekatan Term Frequency - Inverse Document Frequency (TF-IDF) dan variannya, orang perlu mengetahui jumlah dokumen di mana sebuah *term* muncul minimal sekali (dokumen frekuensi). Hal ini memerlukan pengetahuan apriori data, dan bahwa kumpulan data tidak berubah selama perhitungan *term weights*. Kebutuhan pengetahuan dari seluruh kumpulan data secara signifikan membatasi penggunaan skema ini dalam aplikasi dimana aliran data kontinu harus dianalisa secara real-time. Untuk setiap dokumen baru, pembatasan ini mengarah pada pembaruan dokumen frekuensi dari banyak *term* dan karena itu, semua *term weights* yang dihasilkan sebelumnya membutuhkan kalibrasi ulang. Dengan menggunakan metode pembobotan TF-ICF, proses kalibrasi ulang ini tidak perlu dilakukan lagi sehingga motode TF-ICF secara signifikan lebih cepat dibandingkan dengan metode tradisional lainnya.

Dalam rangka untuk mengatasi masalah menemukan dan mengelola informasi dari aliran dokumen dinamis, maka pada tugas akhir ini akan dilakukan analisis dan implementasi sebuat metode pembobotan *term* yang disebut Term Frequency - Inverse Corpus Frequency (TF-ICF). Metode pembobotan ini tidak memerlukan informasi frekuensi *term* dari dokumen-dokumen lain dan dengan demikian dapat memproses aliran dokumen dalam waktu linier. [12]

1.2 Perumusan masalah

Dalam Tugas Akhir ini terdapat beberapa permasalahan yang akan dijadikan sebagai objek penelitian, yaitu:

1. Bagaimana mengimplementasikan Term Frequency - Inverse Corpus Frequency (TF-ICF) pada klasifikasi Teks Mining dengan aliran data dinamis.
2. Bagaimana menganalisa dan mengukur performansi hasil klasifikasi dengan metode pembobotan Term Frequency - Inverse Corpus Frequency (TF-ICF) dan membandingkannya dengan metode pembobotan yang lain, seperti Term Frequency - Inverse Document Frequency (TF-IDF) dan ATC.

Adapun batasan masalah dalam tugas akhir ini adalah:

1. Data set bersih yang digunakan sudah mengalami pre-prosesing data.
2. Algoritma klasifikasi yang digunakan adalah C4.5.
3. Parameter yang digunakan adalah *F-measure*, ROC dan waktu pembobotan.

1.3 Tujuan

Tujuan dari Tugas Akhir ini adalah :

1. Membangun aplikasi yang dapat mengimplementasikan skema pembobotan Term Frequency - Inverse Corpus Frequency (TF-ICF).
2. Mengevaluasi perbandingan performansi yang berupa nilai *F-measure*, ROC serta waktu pembobotan dari metode pembobotan TF-ICF, TF-IDF dan ATC.

1.4 Metodologi penyelesaian masalah

Metodologi yang digunakan untuk menyelesaikan masalah dalam Tugas Akhir ini adalah :

1. Studi literatur.
   Mencari referensi dan sumber-sumber lain yang berhubungan dengan Term Frequency - Inverse Corpus Frequency (TF-ICF). Selain itu juga mengumpulkan dokumen yang akan digunakan pada tahap analisis.
2. Pencarian dan pengumpulan data.
   Mengumpulkan data berupa dokumen berita berbahasa Indonesia yang dibutuhkan untuk keperluan proses implementasi dan pengujian metode yang digunakan.
3. Analisis kebutuhan dan perancangan aplikasi yang akan dibangun.
   Tahapan ini dilakukan dengan menganalisa kebutuhan perangkat lunak dan merancang perangkat lunak menggunakan konsep analisis dan disain yang berorientasikan object. Perancangan akan menggunakan UML (Unified Modelling Language) sebagai pemodelan.
4. Implementasi dan Pengujian
   Mengimplementasikan hasil analisis dan perancangan perangkat lunak serta melakukan pengujian terhadap data set yang akan diklasifikasikan untuk mengukur performansi.
5. Analisa hasil pengujian dan penarikan kesimpulan.

Perangkat lunak yang dihasilkan dievaluasi berdasarkan data yang diperoleh dari hasil pengujian.

6. Penyusunan laporan tugas akhir.
   Pembuatan laporan tugas akhir yang mendokumentasikan tahap-tahap kegiatan dan hasil dalam tugas akhir ini.