1. Pendahuluan

1.1 Latar belakang

Data Mining merupakan salah satu bidang yang berkembang pesat karena besarnya kebutuhan akan nilai tambah dari database skala besar sebagai tuntutan dari pertumbuhan teknologi informasi. *Data Mining* digunakan untuk menemukan informasi atau pola yang bermanfaat atau bernilai ekonomis yang tinggi melalui metode-metode dalam *Data Mining itu*. *Data mining* secara otomatis mengekstraksi informasi dari sekumpulan data yang berukuran besar. Karena itulah, Data Mining merupakan bagian dari *Knowledge Discovery in Database* atau disingkat sebagai KDD.

Dalam *data mining* diperkenalkan salah satu teknik pengelompokan yaitu *clustering*. Idenya adalah dengan mengelompokkan beberapa dokumen yang belum berlabel ke dalam kelompok-kelompok atau *clusters* berdasarkan kemiripan antar berita. Teknik ini bertujuan menemukan distribusi data serta mengelompokkannya berdasarkan kriteria homogenitas atau sejenis sehingga data dapat di *cluster* berdasarkan ditemukannya kesamaan antar data tersebut serta dapat menunjukkan perbedaan data antara satu *cluster* dengan data yang ada di cluster yang lainnya. Klaster yang baik adalah klaster yang memiliki persamaan (*similarity*) intraklaster yang tinggi dan perbedaan (*dissimilarity)* antarklaster yang tinggi.

Sekarang ini jumlah berita yang beredar yang bersifat *unsupervised* sangatlah tidak sedikit. Tenaga manusia kurang efisien untuk menanganinya. Untuk itu diperlukan perangkat yang dapat membantu dalam pengolahan berita secata otomatis, khususnya dalam pengelompokan berita yang belum berlabel (*unsupervised*). Perangkat dibuat dengan mengaplikasikan berbagai bidang ilmu dari mulai matematika, statistika, kecerdasan buatan *data mining*, *information retrieval*, dsb. Secara teoritis dapat membantu dalam menyelesaikan masalah pengelompokan data.

Perkembangan teknik pengelompokan dokumen teks telah menarik perhatian para penelitian untuk mengembangkan dan menemukan suatu teknik atau algoritma yang dapat memperbaiki teknik-teknik *clustering* yang sudah ada. *Suffix Tree Clustering* (STC) adalah algoritma pertama yang menggunakan frasa (multi-word terms) sehingga prosesnya lebih sederhana dibandingkan dengan algoritma yang lain. STC adalah algoritma incremental, kompleksitas waktu perhitungannya linear O(*n*) dan memenuhi *document clustering*. STC tidak memperlakukan dokumen sebagai kumpulan dari kata tetapi lebih dari sebuah string dan didasarkan pada *suffix tree* dalam mengenali kumpulan dokumen serta menggunakan informasi tersebut untuk membangun *cluster*. Oleh karena menggunakan frasa, maka STC mampu mengurangi dimensionalitas himpunan dokumen, hal ini akan mempengaruhi kecepatan dan efisiensi dari STC. Dengan mengembangkan algoritma STC untuk *document clustering*, akan diperoleh manfaat proses *clustering* memberikan

performasi yang baik, dengan adanya beberapa keunggulan yang dimiliki seperti yang telah disebutkan diatas. Algoritma lain yang sudah banyak diaplikasikan dalam Data Mining adalah *K-means*. *K-means* mengelompokkan data berdasarkan kedekatan *data point* dengan *centroid.* Jumlah *centroid* yang ditentukan adalah sebanyak jumlah *cluster* yang ingin dihasilkan. Proses pada tugas akhir ini adalah membandingkan performansi dari algoritma *K-means* dengan algoritma *Suffix Tree Clustering* dalam pengelompokan berita dalam bahasa Indonesia. *K-means* adalah algoritma yang *clustering* yang sederhana dan mudah untuk diimplementasikan. Dengan pembuatan sistem untuk mengelompokkan berita berbahasa Indonesia yang menggunakan algoritma tersebut diatas kita dapat mengetahui algoritma terbaik dalam pengelompokan berita. Dan menghasilkan berita yang telah terkelompokkan dengan *noise* yang sangat minimum lebih dapat digunakan untuk keperluan selanjutnya.

1.2 Perumusan masalah

Clustering dalam *data mining* digunakan untuk mengelompokkan data berupa berita yang bersifat *unsupervised.* Data-data yang ada belum memiliki kelasnya masing-masing. Berita akan dikelompokkan sesuai dengan kemiripan-kemiripan yang ada antara berita yang satu dengan berita yang lain sehingga membentuk kelompok-kelompok berita.
Permasalahannya adalah:

- Bagaimana mengelompokan berita dengan tepat dan memiliki akurasi yang tinggi.
- Membandingkan performansi dari kedua algoritma yang dipakai dalam pengelompokan berita dalam bahasa Indonesia.

Permasalahan dalam tugas akhir ini memiliki batasan sebagai berikut :

1. Tidak membahas sistem *data mining* secara keseluruhan, hanya salah satu
   fungsionalitas data mining yaitu *clustering*.
2. Berita yang digunakan adalah berita berbahasa Indonesia.
3. Berita yang digunakan tidak diambil langsung dari web, tetapi dari database.
4. Tiap berita hanya akan terkelompokkan ke dalam satu kelompok berita.
5. Tidak menerapkan proses stemming dalam dalam proses *pre-processing*
   data.

1.3 Tujuan

Tujuan dari pembuatan tugas akhir ini adalah sebagai berikut :

1. Merancang dan membangun perangkat lunak untuk mengelompokkan berita berbahasa Indonesia dengan algoritma *Suffix Tree Clustering* dan algoritma *K-means*.
2. Menganalisis dan mengevaluasi performansi system dalam hal ini yaitu akurasi *cluster* yang dihasilkan sistem dalam melakukan *clustering*.
3. Membandingkan performansi (tingkat akurasi) dari kedua algoritma dalam pengelompokan berita berbahasa Indonesia.

1.4 Metodologi penyelesaian masalah

Metodologi yang akan digunakan untuk menyelesaikan tugas akhir ini adalah:

1. Studi Literatur
   Pada tahap ini akan dilakukan pendalaman materi, identifikasi masalah, dan metodologi yang akan digunakan dalam pemecahan masalah dengan mencari informasi dan referensi dari berbagai sumber seperti artikel, informasi dari buku maupun internet.
2. Perancangan
   Melakukan analisa penerapan metode yang digunakan dan perancangan akan sistem yang akan diimplementasikan.
3. Implementasi
   Mengimplementasikan desain perangkat lunak yang telah dinyatakan fixed kedalam bahasa pemograman untuk menghasilkan suatu program yang dapat menganalisis berdasarkan perumusan masalah yang telah diuraikan diatas.
4. Pengujian
   Melakukan pengujian dari sistem yang telah dibangun pada tahap implementasi, serta melakukan perbaikan terhadap *bug* dan *error* yang ditemukan pada perangkat lunak yang dibangun.
5. Analisa dan pembuatan laporan
   Membuat analisis dari hasil implementasi yang telah dibuat sesuai dengan parameter yang telah ditentukan sebelumnya dan membuat laporan hasil analisa.