Bab I Pendahuluan

1.1 Latar Belakang

Perkembangan teknologi yang begitu pesat menuntut semakin besar ukuran informasi yang berbentuk teks digital. Dokumen teks dengan jumlah yang sedikit, tentunya mudah bagi manusia untuk melakukan kategorisasi secara manual, namun akan menjadi suatu hal yang sulit untuk dilakukan jika jumlah dokumennya mencapai ratusan hingga ribuan. Sehingga dibutuhkan sebuah sistem yang dapat mengelola dan mengkategorikan dokumen dengan jumlah besar tersebut secara otomatis. Salah satu solusi yang dapat dilakukan adalah dengan menerapkan *text categorization* agar dapat mengkategorisasi dokumen teks secara otomatis, selain itu juga dapat menghemat waktu dan biaya.

Text categorization adalah tugas khusus dari *text mining* yang bertujuan untuk mendefinisikan kategori dari dokumen yang belum diketahui kategorinya [1]. Pada kategorisasi teks, terlebih dahulu dilakukan proses pembelajaran dan pengklasifikasian dengan menggunakan data latih dan data uji sebelum dilakukan tahap klasifikasi. Metode klasifikasi seperti ini disebut *supervised learning* dimana salah satu algoritma dalam *supervised learning* adalah multinomial naïve bayes. Algoritma multinomial naïve bayes diketahui memiliki cara kerja yang sangat sederhana, efektif dan mempunyai performansi yang bagus [2].

Pada umumnya, sistem kategorisasi dokumen hanya sampai pada tahap kategorisasi saja. Pada penelitian tugas akhir ini, penulis akan menyimpan dokumen hasil kategorisasi tersebut ke dalam basisdata agar pengolahan data dapat dilakukan dengan mudah. Tren basis data yang berkembang saat ini yaitu NoSQL yang didesain khusus untuk memecahkan permasalahan *scalability* dan *reliability* [3]. Salah satu jenis dari NoSQL adalah *document oriented database* yang menyimpan data dalam format dokumen. Salah satu DBMS *document oriented* adalah MongoDB. MongoDB memiliki fitur GridFS yang dapat menyimpan data dalam bentuk file biner. Selain itu, MongoDB juga memiliki fitur *sharding* yang dapat mendistribusikan data ke dalam beberapa mesin. Dengan menempatkan data pada beberapa mesin memungkinkan untuk menyimpan lebih banyak data dan menangani beban lebih besar tanpa diperlukan adanya mesin dengan performansi tinggi [4].

Pada tugas akhir ini akan diimplementasikan sebuah sistem kategorisasi dokumen menggunakan algoritma multinomial naïve bayes. Kemudian file biner dari dataset akan disimpan menggunakan GridFS dengan *sharding* pada DBMS MongoDB.

1.2 Perumusan Masalah

Permasalahan yang diangkat dalam Tugas Akhir ini adalah:

1. Bagaimana cara menerapkan algoritma multinomial naïve bayes dalam kategorisasi dokumen?
2. Bagaimana performansi sistem kategorisasi dokumen menggunakan algoritma multinomial naïve bayes?

3. Bagaimana cara mengimplementasikan *GridFS* menggunakan *sharding* pada MongoDB?
4. Bagaimana performansi *GridFS* dengan *sharding* berdasarkan parameter uji *response time* dan *throughput*?

1.3 Tujuan

Tujuan yang ingin dicapai dari Tugas Akhir ini adalah:

1. Mampu melakukan kategorisasi dokumen menggunakan algoritma multinomial naïve bayes.
2. Menganalisis performansi sistem kategorisasi dokumen menggunakan algoritma multinomial naïve bayes.
3. Mengimplementasikan *GridFS* menggunakan *sharding* pada MongoDB
4. Menganalisis performansi *GridFS* dengan *sharding* berdasarkan parameter uji *response time* dan *throughput*.

1.4 Batasan Masalah

Adapun batasan masalah dalam Tugas Akhir ini adalah:

1. Dataset yang digunakan adalah jurnal-jurnal hasil dari tugas akhir mahasiswa jurusan S1 Teknik Informatika Universitas Telkom
2. Dokumen yang digunakan adalah file berekstensi *.pdf* dengan format penulisan jurnal
3. Teks yang diekstrak hanya abstraknya dan penulisan abstraknya menggunakan satu kolom
4. Implementasi system menggunakan bahasa pemrograman java
5. DBMS yang digunakan adalah MongoDB
6. Sistem tidak menangani masalah duplikasi dokumen pada basisdata dokumen

1.5 Metodologi Penyelesaian Masalah

Metodologi penyelesaian masalah terbagi dalam beberapa tahapan, yaitu:

1. Studi Literatur
   Mencari dan mempelajari sumber kajian yang berkaitan dengan Tugas Akhir. Referensi yang digunakan berasal dari buku dan jurnal ilmiah. Referensi yang terkait adalah *Text Mining*, *Text Categorization*, *Document Oriented Database* dan MongoDB.
2. Pengumpulan dan Pengolahan Data
   Mengambil data dari *Open Library* Telkom University yang akan digunakan dalam tugas akhir. Data yang diambil adalah jurnal-jurnal tugas akhir mahasiswa S1 Teknik Informatika Universitas Telkom.
3. Desain Sistem
   Tahapan ini merupakan tahapan yang dilakukan setelah pendefinisian masalah. Pada tahap ini, dilakukan perancangan sistem dan perumusan implementasi yang akan diterapkan.

4. Implementasi
   Mengimplementasikan kategorisasi dokumen dan *document oriented database* pada sistem sesuai dengan permasalahan yang telah didefinisikan dan sesuai dengan perancangan sistem.
5. Pengujian dan Analisis
   Melakukan pengujian terhadap sistem yang telah dibangun. Pengujian dilakukan dengan cara menghitung *precision, recall*, dan *accuracy* dari sistem kategorisasi dan menghitung *response time* dan *throughput* pada arsitektur sistem GridFS dengan *sharding*. Selanjutnya dilakukan analisis terhadap hasil pengujian.
6. Penyusunan Laporan Tugas Akhir
   Pada tahapan ini dilakukan penulisan dokumentasi semua tahap yang sudah dilakukan agar setiap kegiatan dapat dipertanggungjawabkan secara jelas.