1. Pendahuluan

1.1. Latar belakang

Kategorisasi teks merupakan sebuah permasalahan pemilihan kategori untuk sebuah teks yang memiliki karakteristik atau tanda-tanda umum yang dimiliki oleh teks, artikel atau dokumen tersebut [13]. Suatu teks, artikel, atau dokumen berita dapat dikategorisasikan oleh beberapa, tepat satu, atau tidak termasuk dalam kategori manapun [8].

Permasalahan yang signifikan dalam kategorisasi teks adalah dimensionalitas data yang sangat tinggi. Dimensi data yang sangat tinggi tersebut menyebabkan waktu pemrosesan menjadi lebih lama. Salah satu cara untuk mengatasi hal tersebut adalah dengan melakukan *feature selection* terhadap data. *Feature selection* dilakukan untuk memilih fitur-fitur penting dan relevan terhadap data dan membuang fitur-fitur yang tidak berpengaruh terhadap proses kategorisasi teks. Dengan demikian, dimensionalitas data dapat dikurangi.

Dalam tugas akhir ini metode kategorisasi teks yang digunakan adalah N-Gram. Pemilihan metode tersebut didasarkan pada penelitian-penelitian sebelumnya yang menyebutkan bahwa metode tersebut memiliki performansi yang baik dalam permasalahan kategorisasi teks. Selain itu, pada N-Gram, jumlah fitur sangat mempengaruhi proses komputasinya, yang berpengaruh pula pada waktu yang dibutuhkan untuk kategorisasi itu sendiri.

Metode N-Gram menurut Trenkle & Cavnar (1994) adalah munculnya makna atau kata baru dari seperangkat karakter hasil pemotongan pada sebuah kata. Dari penelitian yang telah dilakukan, kategorisasi *Tri*-gram menunjukkan permformansi yang paling baik di antara yang lain, dengan persentase tertinggi 81,25% dan terendah 5,357% [16]. N-gram merupakan urutan karakter dengan panjang *n* yang di dapatkan dari ekstraksi dokumen atau teks. Dengan menggunakan N-gram sistem dapat mencapai *Language Independence* dimana kebanyakan *Information Rietrieval System* memiliki ketergantungan terhadap bahasa yang digunakan[1,10]. Selain itu, N-Gram dapat diterapkan untuk teks yang memiliki kesalahan penulisan, penggunaan karakter, dan tanda asing.

Dalam tugas akhir ini, penulis menganalisis pengaruh dari *feature selection* menggunakan *Information Gain* dan *Chi-Square* terhadap performansi kategorisasi teks yang dilakukan dengan metode Ngram. Analisis dilakukan terhadap waktu kategorisasi dan akurasi kategorisasi yang diukur dengan nilai *F-Measure* dari kategorisasi menggunakan N-Gram.

1.2. Perumusan masalah

Dalam tugas akhir ini permasalahan yang akan di bahas diantaranya adalah:

a. Bagaimana pengaruh *feature selection* terhadap performansi kategorisasi menggunakan klasifier N-gram?
b. Bagaimana pengaruh perubahan jumlah dokumen terhadap performansi *feature selection* pada kategorisasi teks?
c. Bagaimana pengaruh perubahan jumlah kategori terhadap performansi *feature selection* pada kategorisasi teks?

Hipotesis :

Waktu yang dibutuhkan untuk kategorisasi teks dengan *feature selection* lebih kecil dibandingkan kategorisasi teks tanpa *feature selection*. Nilai *F-Measure* yang dihasilkan dari kategorisasi teks dengan *feature selection* lebih besar dibandingkan kategorisasi teks tanpa *feature selection.*

1.3. Tujuan

Tujuan yang ingin dicapai dari Tugas Akhir ini adalah :

a. Mengimplementasikan dan menganalisis *feature selection* dengan *Information Gain* dan *Chi-Square* pada kategorisasi teks dengan metode N-Gram.
b. Menganalisis pengaruh *feature selection* terhadap performansi hasil kategorisasi berdasarkan *F-Measure* dan waktu kategorisasi.
c. Menganalisis pengaruh jumlah dokumen training dan variansi data terhadap nilai *F-Measure* dan waktu yang dibutuhkan untuk kategorisasi teks dengan *feature selection*.

1.4. Batasan Masalah

Permasalahan dalam Tugas Akhir ini dibatasi untuk hal-hal sebagai berikut.

a. Dokumen atau teks yang digunakan sebagai data latih dan data uji merupakan dokumen atau teks berita berbahasa Indonesia yang diambil dari situs berita **www.okezone.com** dan disimpan dalam file berekstensi *.txt.
b. Dokumen atau teks yang digunakan sebagai data latih dan data uji merupakan dokumen dengan kapasitas maksimal 4 kb.

1.5. Metodologi Penyelesaian Masalah

Metodologi yang digunakan adalah :

a. Studi literatur

Pada tahap ini, penulis mencari dan mempelajari konsep dan teori yang berhubungan dengan topik yaitu *feature selection*, *Information Gain*, *Chi-Square*, *Text Categorization* (TC), dan N-Gram.

b. Pengumpulan data

Pada tahap ini dilakukan pencarian data, data yang akan digunakan pada tugas akhir ini berupa dokumen atau teks-teks berbahasa Indonesia yang disimpan dalam file berekstensi *.txt dengan ukuran maksimal 4 kb.

c. Analisis dan perancangan sistem

Melakukan analisis dan perancangan terhadap sistem yang dibangun, menganalisis metode yang akan digunakan untuk menyelesaikan permasalahan, termasuk menentukan bahasa pemrograman yang digunakan, arsitektur, fungsionalitas, dan antarmuka sistem. Input dari sistem ini adalah dokumen atau teks berbahasa Indonesia yang akan ditentukan kategorinya. Sedangkan output dari sistem adalah kategori dari teks atau dokumen yang telah di input.

Sistem dalam tugas akhir ini terdiri dari *feature selection*, pembentukan model, dan analisis performansi untuk masing-masing metode.

1) *Feature selection*
   Pada proses ini dilakukan pemilihan fitur yang akan digunakan dalam proses kategorisasi. Fitur ini berupa *term* (token) yang diambil dari dokumen.
2) Proses *Learning*
   Dalam proses ini, akan dilakukan *learning* dengan metode yang telah ditentukan terhadap dokumen atau teks-teks yang telah dikumpulkan sebagai data latih.
3) Analisis performansi
   Analisis performansi dilihat untuk masing-masing metode dengan parameter *F-Measure* dan waktu kategorisasi.

d. Implementasi dan pembangunan sistem
   Sistem ini diimplementasikan dengan bahasa pemrograman *java*.
e. Pengujian dan analisis
   Pengujian dan analisis dilakukan dengan cara:
   1) Mengkategorisasikan dokumen atau teks yang digunakan untuk pengujian (data uji) terhadap sistem.
   2) Mengkalkulasikan akurasi kategorisasi.
   3) Mengkalkulasikan waktu yang dibutuhkan untuk melakukan kategorisasi.
f. Penyusunan laporan Tugas Akhir.
   Dalam tahap ini akan disusun buku laporan Tugas Akhir yang berisi segala implementasi dari sistem yang telah dikerjakan.

1.6. Sistematika Penulisan

Tugas akhir ini disusun berdasarkan sistematika berikut :

BAB 1 : Pendahuluan
Yang dipaparkan dalam bab ini adalah latar belakang penelitian, perumusan masalah, tujuan penelitian, batasan masalah, metodologi penelitian, dan sistematika penulisan tugas akhir.

BAB 2 : Landasan Teori
Pada bab ini dijelaskan mengenai teori-teori yang berhubungan dengan tugas akhir ini.

BAB 3 : Analisis dan Perancangan Sistem
Pada bab ini akan dijelaskan proses perancangan dan implementasi sistem yang digunakan dalam tugas akhir ini.

BAB 4 : Implementasi dan Evaluasi Hasil
Pada bab ini, dipaparkan mengenai pengimplementasian sistem dan evaluasi hasil pengujian sistem.

BAB 5 : Kesimpulan dan Saran
Pada bab ini dipaparkan beberapa kesimpulan mengenai permasalahan yang dibahas. Selain itu, pada bab ini, juga akan diberikan saran untuk pengembangan selanjutnya.