

BAB I PENDAHULUAN

1.1 Latar Belakang

Natural Language Processing(NLP) dianggap masih sangat sulit untuk dilakukan. Namun seiring dengan perkembangan teknologi, NLP semakin mudah dan perlu untuk dikembangkan. Pengidentifikasian bahasa (*language identification*) menjadi kebutuhan agar sistem NLP dapat berjalan [AHM04], karena dengan mengetahui bahasa yang digunakan, proses terhadap suatu *natural language* akan lebih mudah.

Identifikasi bahasa menggunakan metode yang sederhana dapat dilakukan dengan menyimpan leksikon tiap bahasa. Setelah itu, dilakukan pencocokkan setiap kata yang terdapat di dalam dokumen tersebut terhadap seluruh leksikon bahasa yang dimiliki. Leksikon yang memiliki kelompok kecocokan kata dengan jumlah terbanyak akan dianggap sebagai bahasa yang digunakan oleh dokumen.

Untuk membuat leksikon yang cukup representatif bukanlah hal yang mudah. Terlebih lagi jika bahasa tersebut menggunakan banyak bentuk untuk setiap kata yang ada. Hal ini dapat mengakibatkan ukuran leksikon menjadi sangat besar [CAV94].

Metode lain yang dapat digunakan untuk melakukan pengidentifikasian suatu bahasa adalah dengan menggunakan analisis *n-gram*. Ide awalnya adalah menggunakan *n-gram* yang frekuensi kemunculannya dapat menunjukkan apakah teks tersebut menggunakan bahasa tertentu [CAV94].

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, terdapat beberapa permasalahan yang dapat dirumuskan, antara lain:

1. Bagaimana melakukan pengidentifikasian bahasa menggunakan metode *n-gram*.

2. Bagaimana metode perbandingan *n-gram frequency rank* antar dokumen sehingga dapat memberikan hasil yang baik.
3. Bagaimana membuat sampel bahasa yang baik dan bisa digunakan sebagai dasar identifikasi bahasa.

1.3 Tujuan

Tujuan dari Tugas Akhir ini adalah:

1. Memahami cara melakukan identifikasi bahasa dengan menggunakan metode *n-gram*.
2. Membangun sistem identifikasi bahasa yang menggunakan metode *n-gram*.
3. Melakukan pengujian terhadap performansi sistem identifikasi bahasa yang telah dibuat.
4. Menganalisis hasil pengujian kinerja sistem identifikasi bahasa yang menggunakan metode *n-gram*.

1.4 Batasan Masalah

Batasan masalah dalam Tugas Akhir ini adalah:

1. Sampel bahasa yang akan digunakan akan dibuat secara manual bukan berdasarkan penelitian statistik.

1.5 Metodologi

Dalam pengerjaan Tugas Akhir ini, metodologi yang digunakan dapat dijabarkan dalam langkah-langkah sebagai berikut:

1. Studi Literatur

Studi literatur dilakukan dengan mempelajari literatur-literatur, jurnal ilmiah, dan halaman situs web yang berkaitan dengan teknik identifikasi bahasa.

2. Analisis Masalah

Pada tahapan ini dilakukan analisis terhadap masalah-masalah yang muncul pada proses identifikasi bahasa serta analisis terhadap alternatif penyelesaiannya.

3. Perancangan

Pada tahapan ini dilakukan perancangan terhadap perangkat lunak yang akan dibangun.

4. **Implementasi**

Pada tahap ini dilakukan implementasi perangkat lunak berdasarkan hasil perancangan yang telah dilakukan.

5. **Pengujian**

Pengujian dilakukan untuk mengukur kinerja perangkat lunak yang akan dibangun.

6. **Analisis Hasil dan Penarikan Kesimpulan**

Tahapan yang terakhir adalah menganalisis hasil pengujian dan menarik kesimpulan dari hasil tersebut.

1.6 Sistematika Pembahasan

Sistematika penulisan laporan Tugas Akhir ini adalah sebagai berikut:

1. BAB I PENDAHULUAN, berisi tentang latar belakang, rumusan masalah, tujuan, ruang lingkup, batasan, dan metodologi dalam pelaksanaan Tugas Akhir serta sistematika pembahasan.
2. BAB II DASAR TEORI, berisi dasar-dasar teori yang digunakan dalam analisis, perancangan, dan implementasi.
3. BAB III ANALISIS, berisi analisis mengenai pengenalan bahasa alami menggunakan metode *N-gram*.
4. BAB IV ANALISIS PERANGKAT LUNAK DAN PERANCANGAN, berisi analisis dan perancangan perangkat lunak yang akan dibuat.
5. BAB V IMPLEMENTASI PERANGKAT LUNAK, berisi implementasi perangkat lunak yang dibuat.
6. BAB VI PENGUJIAN, berisi penjabaran umum hasil implementasi dari perancangan dan pengujian kinerja hasil implementasi yang dilakukan.