

## BAB 1

### PENDAHULUAN

#### 1.1 Latar Belakang

*Information Retrieval* adalah ilmu yang mempelajari tentang bagaimana suatu data dapat ditemukan dari berbagai sumber (*resources*) yang relevan atau koleksi sumber informasi yang dicari atau dibutuhkan. Dengan tindakan *index (indexing)*, panggilan (*searching*), pemanggilan data kembali (*recalling*), atau biasa disebut sebagai temu kembali informasi. Untuk pencarian data, beberapa jenis data dapat ditemukan diantaranya teks, tabel, gambar (*image*), video, audio. Dalam *Information Retrieval*, *preprocessing* merupakan tahapan yang sangatlah penting, di mana *preprocessing* adalah tahapan awal untuk mengubah data masukan menjadi data yang sesuai dan siap untuk diproses. Pada *preprocessing* terdapat *algoritma stemming* yang berfungsi untuk menghilangkan kata imbuhan menjadi kata dasar. Pada *algoritma stemming* Nazief & Adriani tidak semua kata yang memiliki imbuhan bisa diselesaikan, ada beberapa kata yang mengalami pemenggalan imbuhan yang lebih (*overstemming*) atau pemenggalan imbuhan yang terlalu sedikit (*understemming*).

Penelitian yang dilakukan oleh Erick Alfons Lisangan, Implementasi *N-Gram Technique* Dalam Deteksi Plagiarisme Pada Tugas Mahasiswa. Pada tahun 2013, menemukan bahwa hasil dari nilai *similarity Sorensen Dice Coefficient* berdasarkan *N-Gram Technique* sendiri akan lebih maksimal jika ada penambahan *algoritma stemming* sebelum proses *similarity*. *Algoritma stemming* yang akan ditambahkan adalah Nazief & Adriani, di mana proses yang akan dilakukan adalah dengan melakukan *stemming* terhadap kata yang akan dicari kata dasarnya dan kemudian akan dicari nilai *similarity* terbesar pada kata yang sesuai terhadap *N-Gram* untuk didefinisikan sebagai kata dasar dengan nilai yg di hasilkan.

Permasalahan yang dihadapi saat ini adalah bagaimana cara untuk mengimplementasikan metode *N-Gram* dan *Sorensen Dice Coefficient* pada pencarian kata dasar menggunakan metode Nazief&Adriani sehingga pencarian

kata dasar dapat dilakukan secara maksimal. Permasalahan utama pada kasus ini adalah pada *Understemming* (Kurangya pemenggalan imbuhan dalam sebuah metode *Stemming*) dan *Overstemming* (kelebihan pemenggalan imbuhan dalam sebuah metode *Stemming*) dalam pemenggalan imbuhan, misalnya pada kata membeku yang menghasilkan kata yakni “membeku” yang seharusnya kata dasar untuk kata “Membeku” adalah “Beku”.

Metode *Stemming* yakni Nazief & Adriani Algoritma ini mengacu pada aturan morfologi bahasa Indonesia yang mengelompokkan imbuhan, yaitu imbuhan yang diperbolehkan atau imbuhan yang tidak diperbolehkan. Pengelompokan ini termasuk imbuhan depan (awalan), imbuhan kata dibelakang (akhiran), imbuhan kata tengah (sisipan), dan kombinasi imbuhan pada awal dan akhir kata dasar misalnya, (membenarkan, terdapat imbuhan, -me dan -kan, pada kata dasar benar).

Penelitian yang akan dilakukan adalah stemming kata berimbuhan menggunakan metode Nazief & Adriani yang dikombinasikan dengan metode *N-Gram* dan *Sorensen-Dice Coefficient*. Dengan metode tersebut, setiap kata yang tidak ditemukan *RootWord* atau kata dasar akan dicari nilai *N-Gram* (*Uni-Gram, Ni-Gram, Tri-Gram*) pada tiap suku kata yang menjadi perbandingan dengan kamus Bahasa Indonesia. Setelah nilai *N-Gram* didapat, proses selanjutnya yakni menghitung kedekatan/jarak dengan menggunakan metode SDC (*Sorensen-Dice Coefficient*). Setelah nilai SDC diperoleh, langkah selanjutnya adalah mencari jarak terdekat mendekati nilai konstanta 0-1, kata yang mempunyai nilai terbesar diasumsikan menjadi kata dasar yang dicari dengan nilai relevansi yang ditampilkan.

## 1.2 Rumusan Masalah

Adapun permasalahan yang didapatkan pada penelitian ini adalah :

Bagaimana cara mengimplementasikan metode *N-Gram* dan *Sorensen Dice Coefficient* pada metode Nazief&Adriani sehingga dapat melakukan perbaikan kata dasar pada sebuah kata berimbuhan.

### 1.3 Tujuan Penelitian

Tujuan dari penulisan skripsi ini adalah menambahkan metode *N-Gram* dan *Sorensen Dice Coefficient* pada metode Nazief&Adriani

### 1.4 Manfaat Penelitian

Manfaat penelitian ini adalah sebagai bahan acuan dalam pegetikan kalimat berbahasa Indonesia yang baik dan benar, dan mengetahui kesalahan penulisan dari hasil term yang muncul.

### 1.5 Batasan Masalah

Agar masalah yang akan dibahas tidak meluas, maka batasan masalah yang dibahas pada penelitian ini hanya difokuskan pada :

1. Data yang dikumpulkan berasal dari korpus (*Library* kamus Bahasa Indonesia).
2. Nilai *N-Gram* yang digunakan adalah (*N3-Gram, N4-Gram, N5-Gram*).
3. Hasil yang ditampilkan adalah nilai pada ketiga *N-Gram* dari perhitungan *SDC (Sorensen-Dice Coefficient)*.

### 1.6 Metodologi Penelitian

Metode yang digunakan dalam perancangan sistem antara lain :

#### 1. Studi Literatur

Studi literatur ini dilakukan untuk mendapatkan informasi dengan cara mengumpulkan jurnal, skripsi, tesis, karya ilmiah, *website* dan lain sebagainya yang berhubungan dengan sistem temu kembali informasi dan analisis sentimen.

#### 2. Pengumpulan Data

Mengumpulkan data yakni kata berimbuhan. Data yang dikumpulkan akan dilakukan *preprocessing* dan pengujian sistem.

### 3. Analisis Sistem

Analisis pada pada sistem dilakukan berdasarkan hasil observasi dari pengumpulan data. Analisa kebutuhan sistem dilakukan untuk menentukan fitur-fitur apa saja yang nantinya akan diterapkan pada sistem.

### 4. Implementasi Sistem

Implementasi ini merupakan proses penerjemahan dari tahap perancangan ke dalam bentuk aplikasi dengan menggunakan bahasa pemrograman PHP dan *database* menggunakan MySQL.

### 5. Pengujian Sistem

Tahap selanjutnya adalah melakukan uji coba terhadap program yang dibangun untuk mengetahui sejauh mana kinerja sistem dan keakuratan metode yang diterapkan sehingga mampu menghasilkan informasi sesuai yang diharapkan.

### 6. Penyusunan Laporan

Melakukan penyusunan laporan dari penelitian berdasarkan sistematika penulisan.

## 1.7 Sistematika Penulisan

Dalam sistematika penulisan ini terdapat pembahasan yang tersusun dalam beberapa kelompok sehingga mempermudah dalam memahami maksud dan tujuan dalam penelitian ini.

## BAB I PENDAHULUAN

Bab ini membahas tentang latar belakang permasalahan, rumusan masalah yang akan diselesaikan, batasan masalah, tujuan penelitian, metode penelitian, sistematika penulisan, dan jadwal kegiatan yang direncanakan.

## BAB II LANDASAN TEORI

Bab ini membahas tentang teori-teori dasar yang berhubungan dengan permasalahan yang diambil, seperti penjelasan mengenai konsep Sistem Temu Kembali Informasi, metode Nazief&Adriani, *N-Gram*, dan metode *Sorensen Dice's Similarity*.

### **BAB III ANALISIS DAN PERANCANGAN SISTEM**

Bab ini menjelaskan tentang penganalisaan kebutuhan dan perancangan dari sistem. Meliputi analisis sistem, analisis kebutuhan sistem, perancangan arsitektur, perhitungan manual, basis data dan antarmuka sistem, serta langkah-langkah yang akan dilakukan untuk menyelesaikan permasalahan dan mencapai tujuan yang telah ditetapkan.

### **BAB IV IMPLEMENTASI DAN ANALISIS**

Bab ini menjelaskan tentang pengujian sistem secara umum maupun terperinci mengenai hasil penerapan sistem pada objek penelitian.

### **BAB V PENUTUP**

Bab ini terdiri dari simpulan dan saran, yang berisi tentang simpulan hasil penelitian dan saran-saran yang dibutuhkan guna pengembangan sistem lebih lanjut.

