BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Stemming merupakan suatu proses untuk menemukan kata dasar dari sebuah kata dengan cara menghilangkan semua imbuhan (afiks) baik itu yang terdiri dari awalan (prefiks), sisipan (infiks), akhiran (suffiks) dan kombinasi awalan-akhiran (konfiks) pada kata turunan. *Stemming* adalah inti dari teknik pemrosesan *natural languange* untuk mendapatkan informasi kembali (*Information Retrieval*)[5]. *Information Retrieval* adalah suatu sistem dimana *user* akan memasukan suatu *query* tertentu dan sistem tersebut akan mengembalikan suatu informasi, dokumen maupun data yang mempunyai relevansi terhadap *query* yang telah diberikan *user*. Saat ini kebutuhan pengguna akan pencarian informasi semakin meningkat dan jumlah dokumen teks yang dapat diakses juga semakin banyak, hal ini dapat mengakibatkan pengguna semakin sulit menemukan dokumen yang relevan dengan *query* yang dimasukkan[1]. Indikator yang lazim digunakan untuk menilai relevansi hasil pencarian suatu dokumen adalah kesesuaian antara *query* yang diberikan dengan dokumen yang dicari. Metode *stemming* digunakan untuk mengatasi masalah pencarian informasi yang tersimpan didalam dokumen secara efektif dan efisien dan juga digunakan untuk mengurangi perbedaan bentuk dari suatu kata dengan mengembalikannya ke dalam bentuk kata dasar. Misal pada kasus ini dengan *stemming* bahasa Indonesia, dicari suatu dokumen dengan judul "baca buku" dengan menggunakan *query* "membaca", dokumen yang dimaksud tidak akan pernah terdapat dalam hasil pencarian. Namun, dengan menggunakan *stemming*, *query* seperti "membaca" dan "dibaca" akan dianggap memiliki interpretasi yang sama yaitu menjadi "baca" sehingga antara kata pada dokumen dengan *query* bisa cocok. Dengan begini pencarian dokumen akan berhasil[3].

S*temming* Bahasa Indonesia memiliki beberapa teknik pendekatan seperti Jelita Asian tahun 2005, Arifin & Setiono tahun 2002, Nazief & Adriani tahun 1996, Ahmad Yusoff Sembok tahun 1996, Vega tahun 2001, Idris tahun 2001 dan *ECS Stemmer* tahun 2008. Teknik *stemming* dikembangkan untuk alasan mereduksi *term*

menjadi bentuk dasarnya. Dalam penelitian ini akan dilakukan analisis perbandingan dari dua buah algoritma yang termasuk dalam metode *Stemming* yaitu Algoritma Idris dan Algoritma ECS *(Enhanced Confix Stripping) Stemmer* yang dapat diterapkan untuk pencarian dokumen teks bahasa Indonesia. Analisis performansi ini dilakukan dengan tujuan mendapatkan informasi dari kedua algoritma yaitu berupa informasi kecepatan dan akurasi serta jumlah langkah dari Algoritma Idris dan Algoritma ECS dalam penerapannya pada sebuah simulator. Adapun penelitian sebelumnya yang dilakukan oleh Eka yang berjudul "Analisis dan Implementasi *Stemming* Menggunakan Algoritma Idris Pada Dokumen Teks Berbahasa Indonesia"[4] dijelaskan bahwa Algoritma Idris cocok dan efektif untuk mereduksikan kata – kata yang digunakan pada dokumen teks bahasa Indonesia.

Pada penelitian yang dilakukan oleh Noverdy yang berjudul "Implementasi Modifikasi Algoritma *Enhanced Confix Stripping (ECS) Stemmer* Pada Teks Bahasa Indonesia"[5] dijelaskan bahwa Algoritma ECS *Stemmer* yang merupakan perbaikan dari algoritma *Confix Stripping Stemmer*, memiliki akurasi yang tinggi dalam men*stemming* dokumen bahasa Indonesia serta mampu menutupi kesalahan pada *Confix Stripping*. Tetapi tidak diketahui apakah hasil akan sama jika jumlah data serta parameter yang diteliti sama, begitu juga dengan performansinya serta tingkat kecocokan untuk *stemming* dokumen teks bahasa Indonesia dari kedua algoritma tersebut apakah Algoritma Idris akan tetap lebih akurat dan cepat serta memiliki performasi yang tinggi dibanding dengan Algoritma ECS *Stemmer*, begitu juga sebaliknya.

1.2 Identifikasi Masalah

Dari latar belakang masalah yang sudah dijelaskan maka identifikasi masalah adalah apakah performasi Algortima Idris akan lebih baik dari pada Algortima *Enhanced Confix Stripping (ECS) Stemmer* begitu juga sebaliknya dalam pencarian kata dalam bahasa Indonesia.

1.3 Maksud dan Tujuan

1.3.1 Maksud

Maksud dari penelitian ini adalah untuk menganalisis Algoritma Idris dan Algoritma *Enhanced Confix Stripping (ECS) Stemmer* dalam *stemming* kata pada bahasa Indonesia.

1.3.2 Tujuan

Sedangkan tujuan yang akan dicapai adalah mengetahui algoritma mana yang memiliki performasi yang paling baik dalam men*stemming* dokumen pada bahasa Indonesia.

1.4 Batasan Masalah

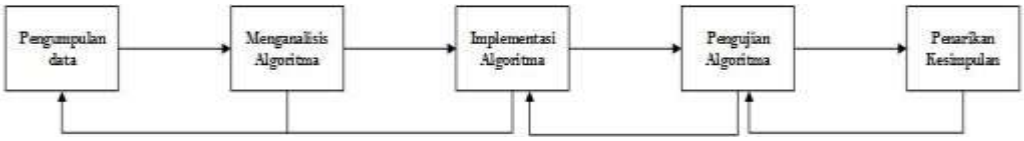
Adapun batasan masalah dalam penelitian adalah sebagai berikut:

1. Dokumen yang digunakan untuk proses *stemming* ini merupakan dokumen teks dengan *query* yang telah ditentukan sebelumnya.
2. Parameter tingkat keakuratan algoritma *stemming* berdasarkan nilai *stem* yang di *stemming* dengan benar.
3. Parameter *output* berdasarkan kata dasar dari Kamus Besar Bahasa Indonesia (KBBI).
4. Parameter tingkat kekuatan *stemmer (stemmer strength)* dalam mereduksi indeks *term* berdasarkan pada analisis *icf (Index Compression Factor)* serta persentase kekuatan *stemmer*.
5. Dokumen yang akan digunakan adalah berekstensi .txt atau .doc.
6. Dokumen yang akan digunakan merupakan novel atau kalimat yang menggunakan kata baku.
7. Sistem ini berbasis *dekstop* dan menggunakan bahasa pemograman C#.
8. *Database* yang akan digunakan adalah MySQL.

1.5 Metodologi Penelitian

Metodologi penelitian yang dimaksud adalah segala hal yang berhubungan dengan metode-metode yang digunakan dalam melakukan penelitian ini dengan cara melakukan pendekatan terhadap metode-metode yang telah ada. Dalam penelitian menggunakan beberapa tahap, yaitu tahap pengumpulan data,

menganalisis algortima, implementasi algortima, pengujian, penarikan kesimpulan yang dapat dilihat pada Gambar 1.1 sebagai berikut :



Gambar 1.1 Tahap Penelitian

Tahapan – tahapan metodologi penelitian pada Gambar 1.1 yang digunakan dalam penelitian tugas akhir ini adalah sebagai berikut :

1. Pengumpulan data

Pengumpulan data adalah cara atau prosedur yang sistematis dan standar untuk memperoleh data yang diperlukan. Adapun metode pengumpulan data pada penelitian ini yaitu studi pustaka dengan cara mengkaji buku, media, atau hasil penelitian orang lain yang memiliki keterkaitan dengan penelitian ini.

2. Menganalisis Algoritma

Pada tahap ini akan dilakukan analisis dan kebutuhan dari algortima Idris dan algortima *Enhanced Confix Stripping* (ECS) *Stemmer*, analisis ini dilakukan guna mempersiapkan hal – hal apa saja yang akan digunakan untuk dapat mengimplementasikan algoritma yang akan dibangun.

3. Implementasi Algoritma

Pada tahap ini analisis algoritma yang telah dilakukan akan diimplemetasikan kedalam sebuah simulator yang berguna untuk membantu simulasi perbandingan algoritma Idris dan algortima *Enhanced Confix Stripping (ECS) Stemmer*. Simulator yang dibangun menggunakan bahasa C# dengan *database* MySQL sesuai perencanaan atau analisis yang sudah dibuat.

4. Pengujian

Setelah simulator dibangun maka tahapan selanjutnya adalah pengujian hasil implemntasi algortima Idris dan algortima *Enhanced Confix Stripping (ECS) Stemmer* pada pencarian dokumen teks bahasa Indonesia. Pada tahap ini akan ditinjau mengenai bagaiman keberhasilan algortima Idris dan algortima *Enhanced Confix Stripping (ECS) Stemmer* dapat digunakan untuk pencarian kata dasar bahasa Indonesia. Pengujian ini dilakukan dengan memfokuskan pengujian pada

penguian akurasi data yang dihasilkan pada saat proses *stemming*. Pengujian pada tahap ini akan menggunakan dua skenario yaitu skenario pertama untuk melihat kekuatan *stemmer* dari kedua algoritma yang diujikan yaitu algoritma Idris dan algoritma ECS *Stemmer*. Kekuatan *stemmer* ini mengacu pada segi kekuatan berdasarkan nilai icf dan persentase serta melihat tingkat kecepatan juga dari kedua algoritma tersebut dalam menstemmer dokumen. Skenario kedua dilakukan untuk menguji keakuratan dari kedua algoritma yaitu algoritma Idris dan algoritma ECS *Stemmer* terhadap *stem* yang dihasilkan, skenario yang dilakukan adalah dengan menguji kedua algoritma menggunakan dokumen uji yang telah ditentukan dan akan menganalisis isi dokumen uji berdasarkan jumlah kata yang tidak berhasil distemming.

5. Penarikan Kesimpulan

Berdasarkan hasil pengujian langkah selanjutnya adalah penarikan kesimpulan berdasarkan parameter yang telah ditentukan yaitu dari nilai icf, persentase serta waktu pemrosesan dari algortima Idris dan algortima *Enhanced Confix Stripping (ECS) Stemmer.*

1.6 Sistematika Penulisan

Penyusunan skripsi ini dibagi ke dalam beberapa bab secara sistematis sesuai dengan pokok – pokok permasalahan yang dibahas. Sistematika penulisan secara umum adalah sebagai berikut :

BAB 1 PENDAHULUAN

Bab ini menjelaskan secara singkat mengenai latar belakang masalah, identifikasi masalah, maksud dan tujuan, batasan masalah, metodologi penelitian, serta sistematika penulisan yang dimaksudkan agar dapat memberikan gambaran tentang urutan pemahaman dalam menyajikan laporan ini.

BAB 2 LANDASAN TEORI

Pada bab ini berisi mengenai landasan terori yang digunakan untuk menganalisis masalah dan teori yang dipakai dalam mengelolah data penelitian yaitu mengenai teori sistem temu kembali, *stemming* bahasa Indonesia, algorima Idris, algoritma

Enhanced Confix Stripping (ECS) Stemmer dan teori mengenai *software* pembangun simulator.

BAB 3 ANALISIS DAN KEBUTUHAN ALGORTIMA

Bab ini menerangkan analisis yang dilakukan terhadap cara kerja aplikasi yang dibuat yaitu menerangkan tentang analisis perbandingan algoritma Idris dengan algoritma *Enhanced Confix Stripping (ECS) Stemmer* pada dokumen teks bahasa Indonesia.

BAB 4 IMPLEMENTASI DAN PENGUJIAN

Pada bab ini berisi tentang implementasi perbandingan algortima Idris dengan algoritma *Enhanced Confix Stripping (ECS) Stemmer* dalam menentukan perbandingan performansi pada dokumen teks bahasa Indonesia dari kedua algortima tersebut sehingga dapat terlihat hasilnya.

BAB 5 KESIMPULAN DAN SARAN

Bab ini berisi hal-hal yang bisa disimpulkan dari hasil analisis perbandingan algoritma Idris dengan algortima *Enhanced Confix Stripping (ECS) Stemmer*, serta saran untuk pengembangan aplikasi ini ke depannya.