1. PENDAHULUAN

1.1 Latar Belakang Masalah

Perangkat pencarian informasi dapat dikatakan sedang dalam keadaan diproses adalah ketika *user* memasukkan sebuah *query* ke dalam si*stem*. *Query* tersebut dapat dimisalkan pada sebuah *string* pencarian *web* pada mesin pencari. Ada beberapa *string* yang dapat langsung cocok namun belum tentu relevan. Saat ini kebutuhan pengguna akan perangkat pencarian informasi semakin meningkat dan jumlah dokumen teks yang dapat diakses juga semakin banyak, hal ini dapat mengakibatkan user semakin sulit menemukan dokumen yang relevan dengan query yang diinputkan. Si*stem Information Retrieval* yang ideal adalah si*stem* dimana dapat menemukan informasi yang relevan sesuai permintaan pengguna. Indikator yang lazim dipakai untuk menilai relevansi hasil pencarian suatu dokumen adalah kesesuaian antara *query* yang diberikan dan dokumen yang diperoleh. Akan tetapi, *term-term* yang terdapat di dokumen dan di *query* sering memiliki banyak varian morfologik, sehingga pasangan *term* seperti "memakan" ,"dimakan" dan "makan" tidak akan dianggap ekivalen oleh si*stem* tanpa suatu bentuk *Natural Language Processing* (NLP). [10]

Pada beberapa kasus, varian morfologik dari *term-term* memiliki interpretasi semantik yang sama dan dapat dianggap ekivalen oleh si*stem*. Jika dicari suatu dokumen dengan judul "baca buku" dengan menggunakan *query* "membaca", dokumen yang dimaksud tidak akan pernah terdapat dalam hasil pencarian. Dengan *stemming*, *term* seperti "membaca" dan "dibaca" akan dianggap memiliki interpretasi yang sama yaitu menjadi *term* "baca" sehingga antara *term* pada dokumen *index* dengan *query* bisa cocok. Dengan begini pencarian dokumen akan berhasil.

Stemming merupakan suatu proses untuk menemukan kata dasar dari sebuah kata dengan menghilangkan semua imbuhan (*affixes*) baik yang terdiri dari awalan (*prefixes*), sisipan (*infixes*), akhiran (*suffixes*) dan *confixes* (kombinasi dari awalan dan akhiran) pada kata turunan [4]. Dalam *Information Retrieval*, algoritma *stemming* digunakan untuk mengurangi perbedaan bentuk dari suatu kata dengan mengembalikannya ke dalam bentuk kata dasar. Hal ini bertujuan untuk meningkatkan kemampuan sistem dalam menemukan dokumen relevan sesuai *query* yang ada sehingga proses temu kembali menjadi lebih efisien. Algoritma *Stemming* untuk bahasa satu dengan bahasa lainnya berbeda. Untuk *stemming* Bahasa Indonesia, ada beberapa teknik pendekatan seperti Jelita Asian (2005), Arifin & Setiono (2002), Nazief & Adriani (1996), Ahmad Yusoff Sembok (1996), Vega (2001), dan Idris (2001). Teknik *stemming* dikembangkan untuk alasan mereduksi *term* menjadi bentuk dasarnya. *Term* yang ada pada dokumen dan *query* memiliki banyak varian morfologik maka akan sulit *term-term* tersebut dianggap ekivalen. Namun dalam beberapa kasus tertentu varian morfologik *term-term* memiliki interpretasi semantik yang sama dan dapat dikategorikan ekivalen.

Dalam tugas akhir ini, penulis menganalisis 2 algoritma stemming untuk bahasa indonesia yaitu Algoritma Nazief & Adriani dan algoritma Ahmad Yusoff Sembok. Perbedaan kedua algoritma ini ada pada proses penghilangan imbuhan

pada kata berbahasa Indonesia. Untuk Algoritma Nazief & Adriani, proses penghilangan imbuhan dilakukan pada akhiran (*suffixes*) terlebih dahulu. Tetapi pada algoritma Ahmad Yusoff Sembok penghilangan imbuhan dilakukan di awalan (*prefixes*). Persamaan dari kedua algoritma tersebut adalah keduanya sama-sama menggunakan kamus untuk pengecekan kata dasar. Dengan persamaan dan perbedaan tersebut itulah mengapa penulis memilih kedua algoritma ini sebagai topik tugas akhir.

Algoritma Ahmad Yusoff Sembok dikembangkan sebagai pendekatan baru dalam *stemming*. *Stemmer* ini tidak sebaik *stemmer* lain namun ada beberapa kata yang tidak dapat diproses pada *stemmer* lain seperti Nazief&Adriani namun dapat diproses oleh Ahmad Yusoff Sembok. Maka dari itu saya memilih algoritma Ahmad Yusoff Sembok sebagai pembanding algoritma Nazief&Adriani yang sudah lebih dulu dikenal.[1]

Stemming yang diimplementasikan pada tugas akhir ini digunakan pada *Information Retrieval*. Kemudian dianalisis pengaruh proses *stemming* tersebut terhadap proses *Information Retrieval* sehingga dapat disimpulkan teknik *stemming* yang terbaik atau yang paling cocok untuk digunakan pada proses *Information Retrieval* untuk teks berbahasa Indonesia. Selain itu, penulis membandingkan performa dan tingkat keakuratan untuk masing-masing algoritma. Mengacu pada 2 konsep algoritma yang diimplementasikan pada proses *stemming* untuk tugas akhir ini maka *stemming* untuk teks berbahasa Indonesia ini diharapkan dapat mendukung proses *Information Retrieval*.

1.2 Perumusan masalah

Berdasarkan latar belakang masalah yang dikemukakan diatas, penulis merumuskan bahwa masalah-masalah yang akan diselesaikan yaitu:

1. Bagaimana mengimplementasikan Algoritma Nazief & Adriani dan Algoritma Ahmad Yusoff Sembok dalam proses *stemming* pada teks berbahasa Indonesia?
2. Bagaimana perbandingan antara algoritma Nazief&Adriani dan algoritma Ahmad Yusoff Sembok jika ditinjau dari pengaruhnya terhadap *recall, precision, non-interpolated average precision, factor kompresi index (icf),* rata-rata jumlah *term* dalam suatu *conflation class (wc),* serta tingkat keakuratan hasil *term* yang telah di-*stemming*?
3. Bagaimana pengaruh proses *stemming* yang telah diimplementasikan terhadap proses *Information Retrieval* untuk teks berbahasa Indonesia?

Adapun batasan masalah pada Tugas Akhir ini adalah :

1. Teks yang digunakan untuk *document collection* merupakan berkas berita teks berbahasa Indonesia dengan *query* dan *relevance judgments* yang telah ditentukan sebelumnya yang didapat dari hasil riset *research group* Laboratorium Data Mining Centre (DMC).
2. Pengujian dilakukan secara *offline.*
3. Dokumen merupakan dokumen *free text (unstructured text).*
4. Parameter tingkat keakuratan algoritma *stemming* berdasarkan pada nilai *stem* yang di *stemming* dengan benar.

5. Parameter tingkat kekuatan *stemmer* (*stemmer strength)* dalam mereduksi index *term* berdasarkan pada analisis *icf* (*Index Compression Factor)* dan *wc (Number Of Word Per Conflation Class).*
6. Parameter tingkat performansi terhadap *Information Retrieval* berdasarkan *precision* & *recall* dan *non-interpolated average precision.*

1.3 Tujuan

Tujuan yang ingin dicapai dalam pengerjaan Tugas Akhir ini adalah sebagai berikut :

1. Melakukan implementasi dari algoritma *stemming* yang dipilih yaitu Algoritma Nazief & Adriani dan Algoritma Ahmad Yusoff Sembok.
2 Melakukan perbandingan antara algoritma Nazief&Adriani dan algoritma Ahmad Yusoff Sembok yang ditinjau dari segi *recall, precision, non-interpolated average precision, factor kompresi index (icf),* rata-rata jumlah *term* dalam suatu *conflation class (wc),* serta tingkat keakuratan hasil *term* yang telah di-*stemming*.
3 Menganalisis pengaruh dari masing-masing algoritma *stemming* yang diimplementasikan terhadap *Information retrieval*.

1.4 Metodologi penyelesaian masalah

Metodologi penyelesaian masalah yang akan dilakukan dalam Tugas Akhir ini adalah :

1. Studi Literatur
   Tahap ini akan melakukan pencarian referensi-referensi dan materi yang ada di internet serta memahami dan mempelajarinya sehingga dapat digunakan untuk menyelesaikan permasalahan dalam tugas akhir ini. Pencarian referensi berkaitan dengan pembangunan aplikasi search engine berbasis Web, *Information Retrieval* dan lebih mendalam mengenai algoritma stemming Nazief & Adriani dan Ahmad Yusoff Sembok.
2. Pengumpulan Data
   Tahap ini akan melakukan pengumpulan data berupa dokumen berita Bahasa Indonesia. Penulis juga membuat sekumpulan dokumen uji untuk membantu dalam skenario pengujian.
3. Analisis Kebutuhan Sistem
   Tahap ini akan melakukan analisis sistem dan menentukan kebutuhan dari sistem yang akan dibangun seperti kebutuhan fungsional sistem, spesifikasi perangkat lunak dan perangkat keras yang digunakan, serta pemodelan sistem yang akan dibangun.
4. Perancangan Sistem
   Tahap ini akan melakukan perancangan sistem dan perangkat lunak serta menerapkan Algoritma Nazief & Adriani dan Ahmad Yusoff Sembok untuk stemmingnya.
5. Implementasi dan Pengujian
   Tahap ini akan melakukan implementasi hasil perancangan dan pengujian terhadap performansi sistem yang telah dibangun. Pada proses ini dilakukan 3 skenario pengujian yakni menguji pengaruh penerapan kedua algoritma terhadap performansi *Information Retrieval*, menguji *stemmer strength* dari

algoritma *stemming dan* menganalisis performansi dari tiap-tiap algoritma yang diimplementasikan.

7. Analisis
   Tahap ini akan melakukan analisa hasil pengujian dan pengukuran performansi berdasarkan data yang diuji serta mengambil kesimpulan dari hasil yang telah dianalisa. Pengujian metode dilakukan dengan menggunakan sistem yang sebelumnya telah diimplementasikan pada tahap implementasi.
8. Pembuatan Laporan
   Tahap ini akan melakukan dokumentasi tahap-tahap kegiatan dan hasil yang di dapat ke dalam laporan Tugas Akhir. Di tahap ini akan dijelaskan pula mengenai langkah-langkah secara detail dalam menganalisis kebutuhan dari awal, perancangan sistem, implementasi, pengujian, serta analisisnya.