

BAB I PENDAHULUAN

1.1 Latar Belakang

Information retrieval merupakan kegiatan memperoleh data informasi yang relevan, pencarian informasi dapat didasarkan pada teks lengkap atau pengindeksan berbasis konten lainnya [1]. *Information retrieval* terdiri dari tiga komponen utama, yaitu masukan (input), pemroses (processor), dan keluaran (output). Menurut Manning et al *Information Retrieval* (IR) adalah menemukan bahan (biasanya dokumen) yang bersifat tidak terstruktur (biasanya teks) yang memenuhi kebutuhan informasi dari dalam koleksi besar (biasanya disimpan di komputer) [2].

Teknologi pencarian berkembang dan semakin mudah untuk mendapatkan informasi di internet melalui situs web pencarian seperti Google memudahkan siapa pun untuk mendapatkan informasi yang baik dalam bentuk berita, sains, hasil penugasan dan lain-lain [3]. Plagiarisme mungkin merupakan pelanggaran akademik utama, namun sebagian dari institusi perguruan tinggi yang telah berkembang dengan baik sudah menerapkan kebijakan dan prosedur untuk menangani masalah tersebut. Selain itu, plagiarisme juga digambarkan sebagai tindakan pencurian karya orang lain dan membuat pekerjaan seolah-olah itu adalah pekerjaan sendiri [4].

Algoritma yang digunakan oleh Kusnawan dalam aplikasi pendeteksian plagiarisme di Indonesia hanya menggunakan cara sederhana pra-pemrosesan teks seperti menghilangkan spasi, dan karakter yang tidak dibutuhkan [5]. Metode semacam ini hanya bisa mendeteksi plagiarisme dengan cara sederhana tetapi tidak dapat mendeteksi bentuk yang lebih kompleks dengan tambahan proses *Stemming Enhanced Confix Stripping* dalam pra-pemrosesan. Pendekatan yang diusulkan menggunakan algoritma *Rabin-Karp* sebagai tahap membandingkan nilai hash antara *string* masukan dan *substring* dalam teks, jika nilai hash adalah sama, maka perbandingan akan dilakukan sekali lagi untuk karakter, tetapi jika nilai hash tidak sama, substring akan bergeser ke kanan [4].

Stemming adalah langkah dasar dalam pengolahan data teks sebelum melakukan pencarian informasi, pertambahan data, dan pengolahan bahasa alami. Tujuan utama *stemming* adalah untuk membakukan kata dengan mengurangi kata menjadi

kata dasar [6]. Metode *stemming* bahasa indonesia sudah banyak dikembangkan seperti Porter Stemmer, *Confix Stripping Stemmer*, Nazief & Andriani dan *Enhanced Confix Stripping Stemmer*.

Pada penelitian sebelumnya Dibrianto Sinaga menggunakan *Confix Stripping Stemmer* untuk mengetahui pengaruh *stemming* terhadap algoritma *Rabin-karp* dengan hasil waktu proses rata-rata 34,4881 detik dan tingkat kesamaan rata-rata 74,5% data yang digunakan merupakan potongan-potongan teks bahasa indonesia, dokumen yang diuji berjumlah 31 [4]. Pada penelitian tersebut masih memiliki masalah karena menggunakan algoritma *stemming* yang belum dikembangkan hal tersebut bisa mempengaruhi kecepatan proses, kemiripan dokumen maupun akurasi dari sistem yang dibangun.

Algoritma *Enhanced Confix Stripping Stemmer* merupakan perbaikan dari algoritma sebelumnya yaitu, algoritma *Confix Stripping Stemmer*. Setelah dilakukan beberapa percobaan dan analisis, ditemukan beberapa kata yang tidak dapat di-*stemming* menggunakan *Confix Stripping Stemmer* [7]. Pada 2013 Muhamad Alif melakukan penelitian metode ECS dengan hasil Uji coba 300 dokumen, *ECS* telah menghasilkan 81% dengan kecepatan 11,8311 detik [8].

Metode yang digunakan peneliti di atas menggunakan metode *Enhanced Confix Stripping Stemmer* dan *Rabin-Karp*, dimana metode *ECS* merupakan pengembangan algoritma terdahulunya dengan beberapa perbaikan kesalahan pada metode sebelumnya dan metode *Rabin-Karp* sebagai tahap membandingkan nilai hash antara string masukan dan substring dalam teks. Dengan selesainya penelitian ini diharapkan dapat menghasilkan sistem yang akurat untuk pencarian kemiripan dokumen

1.2 Rumusan Masalah

1.2.1 Permasalahan

Banyak algoritma *stemming* yang dikembangkan untuk bahasa Indonesia tetapi dari beberapa algoritma tersebut masih memiliki kekurangan diantaranya tingkat akurasi. Bagaimana perbedaan hasil yang didapatkan bila sistem pendeteksi kemiripan dokumen menggunakan proses *stemming* dengan algoritma *Enhanced*

Confix Stripping yang akan dipadukan *Rabin-Karp* dengan mengganti *Hasing* biasa menggunakan *Rolling Hash* yang dimiliki oleh algoritma *Winnowing*.

1.2.2 Usulan Solusi

Diusulkan gabungan algoritma untuk dilakukan pengujian yaitu *Stemming Enhanced Confix Stripping* dan *Rabin-Karp* untuk mendeteksi kemiripan dokumen teks bahasa indonesia.

1.2.3 Pertanyaan Penelitian

Seberapa besar kemiripan dokumen jika algoritma *Rabin-Karp* dikombinasikan dengan *Stemming Enhanced Confix Stripping*?

1.3 Tujuan Dan Manfaat

1.3.1 Tujuan

Penerapan algoritma *Stemming Enhanced Confix Stripping* yang dikombinasikan dengan *Rabin-Karp* bertujuan untuk memperoleh hasil yang lebih akurat ataupun lebih singkat dalam mendeteksi kemiripan dokumen teks bahasa indonesia.

1.3.2 Manfaat

Dihasilkan aplikasi yang mampu mendeteksi kemiripan dokumen yang lebih akurat.

1.4 Batasan Masalah

Batasan masalah pada penelitian ini adalah sebagai berikut :

1. Dokumen yang akan diuji merupakan dokumen teks.
2. Format dokumen yang digunakan dalam penelitian ini merupakan .pdf
3. Teks yang digunakan sebagai dokumen uji merupakan dokumen yang menggunakan bahasa Indonesia yang baku.
4. Sistem yang akan dibangun menggunakan bahasa pemrograman PHP.

1.5 Sistematika Penulisan

Sistematika penulisan Proposal ini terdiri dari beberapa bab, yaitu :

BAB I PENDAHULUAN

Pada bab ini membahas tentang latar belakang, rumusan masalah, tujuan, manfaat, batasan masalah, serta sistematika penulisan laporan.

BAB II KAJIAN PUSTAKA

Pada bab ini membahas tentang landasan teori dan penelitian sebelumnya.

BAB III METODE USULAN

Pada bab ini membahas tentang metode yang diusulkan (algoritma), arsitektur (Input, Proses, Ouput), Dataset, tahapan penelitian, serta rencana pengujian.

BAB IV HASIL DAN PEMBAHASAN

Pada bab ini berisi bagaimana sistem dibuat, hasil uji coba berdasarkan skenario uji coba dan analisa hasil uji coba.

BAB V PENUTUP

Pada bab ini berisi tentang kesimpulan dari analisa yang telah dilakukan dan saran untuk penelitian selanjutnya dalam kemiripan dokumen bahasa indonesia.