BAB I Pendahuluan

Pada bab ini permasalahan dan latar belakang yang menjadi awal mula penelitian dijabarkan dalam subbab latar belakang dan rumusan masalah serta tujuan dari penelitian dijelaskan dalam subbab tujuan penelitian serta bab ini memuat tentang metodologi penelitian yang digunakan dalam penelitian.

1.1 Latar Belakang

Pada awalnya karya tulis hanya dituangkan dalam tulisan tangan manusia. Seiring berjalannya waktu, era digital semakin berkembang dan mengakibatkan karya tulis mulai direpresentasikan dalam bentuk digital dan dapat di sebarluaskan dengan mudah dengan adanya internet. Dipaparkan dalam [2] melalui hasil penelitian dari *IBM*, *Compac* dan *Alta Vista* terdapat lebih dari 600 juta laman unik yang memiliki variasi topik yang sangat luas. Hal itu memberikan fakta bahwa dalam pencarian suatu topik permasalahan dalam internet sebagian besar akan didapatkan hasil yang relevan dan sesuai dengan kebutuhan. Berdasarkan hal tersebut, akan timbul kecenderungan seorang penulis dapat dengan mudah melakukan tindakan plagiarisme. Seperti yang dijelaskan dalam [3] tindakan plagiarisme dapat dibedakan menjadi dua jenis yaitu *literal plagiarism* dan *intelligent plagiarism*. Teknik plagiarisme yang dapat dilakukan untuk *literal plagiarism* diantaranya adalah melakukan *copy-paste* secara langsung dan melakukan *copy-paste* dengan tambahan modifikasi didalamnya. Sedangkan teknik yang digunakan untuk *intelligent plagiarism* diantaranya adalah mengadopsi ide, mengartikan kedalam bahasa lain, melakukan teknik *paraphrasing* dan mengambil intisari dalam karya tulis tersebut. Tentu saja semua kegiatan tersebut dilakukan tanpa mencantumkan sumber referensi yang digunakan.

Pada dasarnya manusia dapat menganalisis apakah suatu dokumen tersebut terindikasi plagiat atau tidak, namun hal tersebut menjadi cukup sulit apabila terdapat banyak dokumen yang harus dianalisis dan setiap dokumen memiliki kata yang cukup banyak. Hal itu menyebabkan dibutuhkannya suatu sistem yang dapat mendeteksi plagiarisme dengan memberikan hasil analisis secara otomatis sehingga dapat mempermudah pihak pihak yang membutuhkan. Seperti yang telah dijelaskan sebelumnya, tindakan plagiarisme dapat dilakukan dalam berbagai cara, dan dengan cara yang berbeda, pendekatan untuk mendeteksi plagiarisme tersebut pun berbeda beda. Terdapat salah satu pendekatan yang dapat digunakan untuk mendeteksi plagiarisme, salah satunya adalah *text alignment*. Salah satu pengaplikasian dari *text alignment* adalah akan diberikan dua buah dokumen kemudian akan diidentifikasi bagian dokumen yang dipergunakan kembali [1]. Pada umumnya, sistem pendeteksian plagiat biasanya hanya dapat menentukan apakah kedua pasang dokumen yang dinputkan terindikasi plagiat atau tidak, namun disamping itu dibutuhkan juga sebuah sistem yang dapat memberikan informasi pasangan fragmen dari pasangan dokumen yang membuktikan bahwa pasangan tersebut terindikasi plagiat. Dengan

menggunakan pendekatan *text alignment*, setiap bagian dari dokumen dalam pasangan dokumen akan dilihat dan diidentifikasi untuk mendapatkan pasangan bagian dari dokumen sumber yang dipergunakan kembali oleh dokumen yang diduga plagiat. Berbagai cara untuk melakukan plagiarisme membuat beberapa kasus plagiarisme memiliki karakteristik yang berbeda sehingga dibutuhkan suatu sistem yang dapat menangani setiap karakteristik tersebut.

Dalam tugas akhir ini, sistem pendeteksi plagiarisme akan dikembangkan dalam 4 tahap yaitu *Preprocessing*, *Seeding*, *Extension* dan *Filtering*. Pada penelitian sebelumnya, setiap dokumen akan diubah menjadi sebuah unit unit kalimat, kemudian setiap kalimat akan diekstraksi fiturnya menggunakan TF-IDF *vector space model* dan dihitung nilai kesamaannya menggunakan *cosine similarity*, *dice coefficient* dan *jaccard coefficient* dengan tambahan parameter yang dapat menyesuaikan karakteristik tipe plagiarisme [4] [5]. Dalam penelitian [14], penggunaan metode *dice coefficient* dan *cosine similarity* dinilai lebih baik dibandingkan penggunaan *jaccard coefficient*. Selain itu, perhitungan nilai kesamaan dapat didapatkan melalui metode kesamaan semantik dengan mempertimbangkan arti dari setiap fitur dalam teks [6]. Sehingga pada penelitian ini akan menggunakan TF-IDF untuk mengekstraksi fitur dari tiap kalimat dan penggunaan metode *cosine similarity*, *dice coefficient*, dan kesamaan semantik untuk perhitungan nilai kesamaan antar kalimat dengan tujuan menganalisis pengaruh perubahan nilai untuk setiap parameter yang diacu dalam penelitian [4] dan pengaruh penambahan penggunaan kesamaan semantik dalam sistem.

1.2 Perumusan Masalah

Terkait permasalahan yang dipaparkan diatas, dapat dirumuskan masalah yang diangkat dalam Tugas Akhir ini, yaitu :

1. Bagaimana pengaruh perubahan nilai parameter *maxGap*, *threshold seeding*, *threshold extension* terhadap performansi sistem?
2. Bagaimana pengaruh penggunaan kesamaan semantik untuk menghitung kesamaan antar kalimat?
3. Bagaimana cara menguji performansi dari hasil yang didapatkan dari sistem pendeteksi plagiarisme?
4. Bagaimana cara menentukan apakah kedua pasang dokumen merupakan pasangan plagiat atau tidak?

1.3 Tujuan

Berdasarkan latar belakang dan rumusan masalah yang telah diuraikan diatas maka tujuan dari adanya penelitian ini adalah:

1. Dapat menganalisis pengaruh perubahan setiap nilai parameter yang digunakan terhadap performansi sistem yang dihasilkan.
2. Dapat menentukan suatu pasangan dokumen terindikasi plagiat atau tidak.

3. Dapat mengetahui pasangan fragmen yang saling ber-*align* antara dokumen sumber dan dokumen *suspicious*.

1.4 Batasan Masalah

Untuk menyelesaikan penelitian ini ada beberapa hal yang menjadi batasan agar penelitian ini lebih terarah dan tidak terlalu menyimpang dari inti permasalahan maka batasan penelitian ini adalah:

1. Data yang digunakan adalah data dokumen yang berbahasa Inggris yang diambil dari PAN Webis *Plagiarism Detection*[1] yang berisi *training corpus*, *testing corpus* 1 dan *testing corpus* 2.

2. Masukan untuk sistem merupakan pasangan dokumen *suspicious* dan dokumen sumber yang didapat dari *task source retrieval* [1] dengan tipe .txt.

3. Sistem hanya dapat mengeluarkan pasangan fragmen dari pasangan dokumen dan status pasangan tersebut plagiat atau tidak.

4. Setiap hasil keluaran sistem hanya diklasifikasikan kedalam 2 kelas yaitu plagiarisme atau tidak, bukan menklasifikasikan kedalam 4 tipe plagiarisme (*summary, none, random,* dan *translation*).

5. Kasus plagiarisme yang dideteksi dalam penelitian ini diasumsikan tidak memperdulikan adanya pengutipan dari sumber.

1.5 Metodologi Penelitian

Untuk menyelesaikan masalah dibutuhkan tahapan dalam pemecahan masalah tersebut. Metodologi yang dilakukan dalam menyelesaikan masalah adalah sebagai berikut:

1. Studi Literatur
   Dalam tahapan ini akan dilakukan pencarian informasi dan literatur terkait materi dan metode apa yang akan diimplementasikan dalam sistem sehingga dapat menghasilkan performansi sistem yang optimal dan mencapai tujuan penelitian.

2. Perumusan dan Analisis Permasalahan
   Dalam tahapan ini akan dirumuskan permasalahan apa yang akan dipecahkan dalam penelitian tugas akhir ini. Setelah masalah ditemukan kemudian penulis akan menganalisis untuk dapat menentukan rancangan sistem yang tepat untuk mengatasi permasalahan tersebut.

3. Perancangan Sistem
   Setelah menentukan permasalahan yang akan diselesaikan kemudian akan dibuat suatu rancangan sistem yang sesuai untuk menyelesaikan permasalahan yang ada pada tugas akhir ini.

[1] http://pan.webis.de/clef14/pan14-web/plagiarism-detection.html

4. Eksperimen
   Eksperimen bertujuan untuk mengetahui lebih dalam mengenai pengetahuan yang didapat dari proses sebelumnya yaitu studi literatur, dari eksperimen ini akan didapat permasalahan untuk dikaji lebih dalam. Di dalam tahap eksperimen dapat dilakukan percobaan beberapa pasang nilai parameter yang sesuai untuk setiap tipe plagiarisme.

5. Implementasi Hasil Eksperimen
   Implementasi dilakukan untuk merealisasikan hasil perancangan sistem tersebut terhadap dataset yang dipergunakan.

6. Analisis Hasil
   Dari hasil implementasi yang dilakukan akan dilakukan analisis parameter mana yang tepat sehingga menghasilkan performansi yang tinggi untuk setiap tipe plagiarisme yang terdapat dalam dataset yaitu *no plagiarism*, *no obfuscation*, *random obfuscation*, *translation obfuscation*, dan *summary obfuscation*.

7. Pembuatan Laporan
   Setelah melakukan implementasi dan analisis hasil yang didapatkan maka dilakukan kegiatan mendokumentasikan hasil dari penelitian yang dilakukan dan melampirkan dokumen pendukung yang berkaitan tentang penelitian ini.

1.6 Sistematika Penulisan

Penulisan tugas akhir ini akan dibagi menjadi lima bab yang masing masing berisi penjelasan sebagai berikut

1. Bab I Pendahuluan
   Pada bab ini permasalahan dan latar belakang yang menjadi awal mula penelitian dijabarkan dalam subbab latar belakang dan rumusan masalah serta tujuan dari penelitian dijelaskan dalam subbab tujuan penelitian serta bab ini memuat tentang metodologi penelitian yang digunakan dalam penelitian.

2. Bab II Tinjauan Pustaka
   Bab ini menjelaskan teori-teori yang digunakan dalam penelitian dan penjelasan mengenai metode yang digunakan akan dijabarkan dalam subbab dari bab 2.

3. Bab III Perancangan Sistem
   Pada bab ini akan dijelaskan bagaimana sistem dalam penelitian ini bekerja, termasuk detail setiap tahap yang dilakukan untuk membangun sistem dan mengindentifikasi kasus plagiarisme.

4. Bab IV Evaluasi dan Pengujian
   Bab ini akan membahas dokumentasi dan analisis dari hasil pengujian sistem terhadap parameter yang digunakan dalam sistem.

5. Bab V Kesimpulan dan Saran
   Dalam bab ini akan dijelaskan kesimpulan dari hasil penelitian yang telah dilaksanakan berserta saran untuk penelitian yang akan dilanjutkan setelahnya.