BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Pengukuran kesamaaan suatu dokumen sangat bermanfaat untuk mencegah tindakan plagiarisme. Plagiarisme atau sering disebut plagiat adalah penjiplakan atau pengambilan karangan, pendapat, dan sebagainya dari orang lain dan menjadikannya seolah karangan dan pendapat sendiri [1]. Namun jika pengambilan karangan tersebut disertai dengan mencantumkan asalnya yaitu nama pengarang serta judul karangan yang diambil, maka tindakan pengambilan karangan tersebut bukan merupakan plagiat [1]. Oleh karena itu, perlu dilakukan pemeriksaan kemiripan antar dokumen, dalam hal ini adalah dokumen teks, sebagai langkah validasi keterkaitan dan hubungan antar dokumen tersebut. Kemiripan suatu dokumen atau plagiarisme dibagi menjadi 2 berdasarkan tingkat kesulitan dalam pendeteksiannya, yang pertama yaitu plagiarisme secara verbatim atau biasa dikenal dengan nama *exact copy* [2]. Tindakan plagiarisme ini cukup mudah untuk di deteksi karena hanya mengambil teks yang ada secara langsung tanpa melakukan pengubahan pada isi dokumen. Kedua yaitu plagiarisme *obfuscation* (pengaburan), plagiarisme jenis ini sangat komplek dan sangat sulit untuk di deteksi, yaitu dapat berupa mengurangi, merubah struktur kalimat, mengubah istilah atau bahasa yang digunakan.

Saat ini aplikasi mendeteksi kesamaan dokumen sudah banyak dikembangkan baik yang berbayar maupun gratis. Jurnal tentang mendeteksi kesamaan dokumen juga sudah banyak dikeluarkan dengan berbagai macam metode. Beberapa penelitian yang berkaitan dengan mendeteksi kesamaan dokumen diantaranya yang dilakukan oleh David Erwinson dengan metode *Vector Space Model,* dokumen uji dan latih berupa abstrak paper jurnal berbahasa Indonesia. Hasilnya menunjukkan rata-rata persentase *similarity* sebesar 9,575%, sistem belum bisa menunjukan detail kesamaan dan akurasi yang dihasilkan belum diketahui [3]. Selanjutnya penelitian yang dilakukan M Isa dengan dengan metode yang sama yaitu *Vector Space Model,* dokumen yang diujikan berupa text

bahasa Indonesia. Hasilnya yaitu berupa persentase kesamaan dan akurasi yang dihasilkan tidak diketahui [4]. Begitu pula penelitian yang dilakukan Tudesman, metode yang digunakan sama yaitu *Vector Space Model*, dokumen yang diujikan berupa text bahasa Indonesia. Hasilnya yaitu hanya berupa persentase kesamaan dan akurasi yang dihasilkan tidak diketahui [5].

Sedangkan penelitian yang dilakukan Miguel A Sanchez-Perez, dkk, terdapat 4 tahapan yang dilakukan yaitu pertama proses *preprocessing* yang terdiri dari *special character removal, case folding, stemming* [6]. Kedua proses *seeding* dimana tahap ini untuk menghitung kesamaan antar 2 paragraf menggunakan *tf-idf Vector Space Model* yang dibandingkan dengan *cosine similarity* dan *dice coefficient*. Ketiga *extension* yang terdiri dari 2 langkah yaitu *clustering* dan *validation.* Keempat *filtering* untuk meningkatkan presisi. Hasilnya *recall 89,57%, precission* 91,25%, *PlagDet* 90% [7]. Penelitian Sanchez-Perez menggunakan data uji dan latih berbahasa inggris PAN2014 [7]. Untuk pengujian dokumen bahasa Indonesia belum pernah diuji sebelumnya.

Oleh karena itu penelitian ini bermaksud untuk membangun sistem untuk mendeteksi kesamaan suatu dekumen yang berbahasa Indonesia dengan metode yang digunakan oleh Sanchez-Perez yaitu *Vector Space Model* dan *Clustering* karena nilai akurasi keseluruhan merupakan yang tertinggi dalam kompetisi PAN2015.

1.2 Identifikasi Masalah

Berdasarkan latar belakang yang telah dijelaskan di atas, maka dapat ditarik identifikasi masalah yang timbul yaitu sebagai berikut.

1. Pendeteksian kesamaan dokumen dalam bahasa Indonesia dengan menggunakan metode *Vector Space Model* hanya menghasilkan persentase kesamaan, sistem belum bisa menentukan kalimat mana yang sama tersebut.
2. Belum diketahuinya akurasi yang dihasilkan dalam deteksi kesamaan dokumen dalam bahasa Indonesia dengan menggunakan metode *Vector Space Model*.

1.3 Maksud dan Tujuan

Berdasarkan permasalahan yang ada, maksud dari penelitian ini adalah untuk membangun sistem deteksi kesamaan dokumen berbahasa Indonesia menggunakan metode *Vector Space Model* dan *Clustering*.

Sedangkan tujuan dari penelitian ini adalah untuk mengukur akurasi kemiripan suatu dokumen teks berbahasa Indonesia dan mengetahui banyaknya kalimat yang sama dengan menggunakan metode *Vector Space Model* dan *Clustering*.

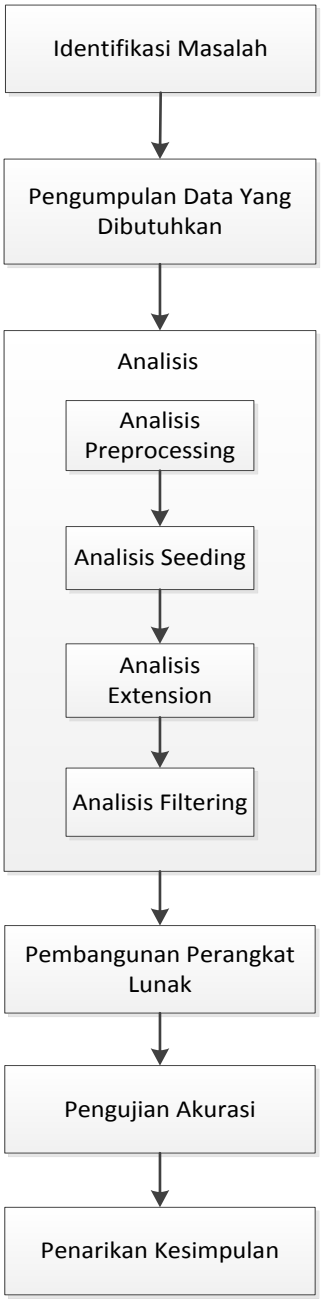
1.4 Batasan Masalah

Batasan masalah dari penelitian ini yaitu :

1. Dokumen uji dan latih berupa abstrak jurnal berbahasa Indonesia
2. Format file dokumen uji dan latih adalah .txt
3. Metode yang digunakan yaitu *Vector Space Model* dan *Clustering*
4. Sistem tidak melihat acuan atau referensi yang digunakan dokumen

1.5 Metode Penelitian

Metode penelitian yang digunakan dalam penelitian ini adalah metode deskriptif. Penelitian ini menggambarkan secara sistematis fakta dan karakteristik objek dan subjek yang diteliti secara tepat. Alur penelitian dapat dilihat pada gambar berikut.



Gambar 1.1 Metode Penelitian

1.5.1 Identifikasi Masalah

Berdasarkan latar belakang, identifikasi masalah yang timbul yaitu pendeteksian kesamaan dokumen dalam bahasa Indonesia dengan metode *Vector Space Model* pada penelitian sebelumnya hanya menghasilkan persentase kesamaan saja, sistem belum bisa menentukan kalimat mana yang sama tersebut dan akurasi yang dihasilkan belum diketahui.

1.5.2 Pengumpulan Data

Tahapan ini melakukan studi kepustakaan dari hasil penelitian yang telah dilakukan sebelumnya oleh orang lain, artikel–artikel yang terkait dengan pembuatan deteksi kesamaan dokumen, membaca paper dan jurnal mengenai plagiarisme, serta mempelajari teknik dan algoritma yang tepat untuk dapat meningkatkan kemampuan sistem deteksi kesamaan dokumen yang sudah dilakukan pada penelitian sebelumnya.

1.5.3 Analisis

Pada tahap ini dijelaskan tahapan-tahapan dalam mendeteksi kesamaan dokumen dengan menggunakan metode *Vector Space Model* dan *Clustering*. Adapun tahapannya sebagai berikut :

1. *Preprocessing*

   Merupakan tahapan awal dalam mengolah data input sebelum memasuki proses tahapan utama. Dalam penelitian ini tahap preprocessing terdiri dari: tokenisasi kalimat, *case folding*, tokenisasi kata, filter kata, *stemming*, filter kalimat, TF-IDF.

2. *Seeding*

   Pada tahap ini setiap kalimat pada dokumen *suspicious* dan *source* dibandingkan dengan menggunakan *Cosine Similarity* dan *Dice Coefficient*.

3. *Extension*

   Mengkluster nilai-nilai dari proses *seeding* yang berupa pasangan kalimat sama berdasarkan nilai *max_gap* sehingga menciptakan nilai baru yang

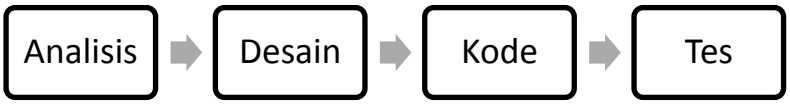
berupa kumpulan kalimat yang sama. Nilai hasil kluster kemudian di validasi menggunakan *Cosine Similarity*.

4. *Filtering*

   Menyeleksi nilai hasil dari *extension*.

1.5.4 Pembangunan Perangkat Lunak

Metode pembangunan perangkat lunak yang digunakan adalah dengan menggunakan metode *sequntial linier model* [8] yang dapat dilihat pada Gambar 1.2.



Gambar 1.2 Metode Pembangunan Perangkat Lunak [8]

1. Analisis Kebutuhan

   Analisis kebutuhan meliputi penentuan perangkat lunak, penentuan perangkat keras dan analisis data masukan yaitu berupa abstrak jurnal berbahasa Indonesia.

2. Desain Sistem dan Perangkat Lunak

   Merancang tahapan dari proses deteksi kesamaan dokumen dan spesifikasi aplikasi yang akan dibuat.

3. Implementasi

   Tahap ini merupakan tahap pembuatan aplikasi sesuai dengan desain yang telah dibuat pada tahap perancangan.

4. Pengujian

   Pengujian yang dilakukan menggunakan metode ROUGE [9] untuk mengetahui akurasi dari metode *Vector Space Model* dan *Clustering* untuk dokumen berbahasa Indonesia.

1.6 Sistematika Penulisan

Sistematika penulisan penelitian ini adalah sebagai berikut.

BAB 1 PENDAHULUAN

Bab ini akan membahas kerangka penelitian, meliputi latar belakang masalah, perumusan masalah, tujuan, batasan masalah, metode penyelesaian masalah, dan sistematika penulisan dari penelitian mendeteksi kesamaan kalimat dalam tulisan bahasa Indonesia dengan menggunakan metode *Vector Space Model* dan *Clustering*.

BAB 2 TINJAUAN PUSTAKA

Bab ini memuat berbagai dasar teori yang mendukung dan mendasari penulisan penelitian ini, di antaranya mengenai konsep plagiarisme, metode *VSM*, dan *Clustering* beserta pemodelan, bahasa pemrograman, dan perangkat lunak yang digunakan.

BAB 3 ANALISIS DAN PERANCANGAN SISTEM

Bab ini berisi tentang tahapan analisis masalah dan solusi, analisis data masukan, analisis proses, analisis data keluaran, dan analisis kebutuhan fungsional dan nonfungsional, serta perancangan sistem yang mencakup perancangan antar muka, dan struktur menu.

BAB 4 IMPLEMENTASI DAN PENGUJIAN

Bab ini membahas tentang penerapan metode yang telah dianalisis. Dalam bab ini juga dilakukan pengujian untuk mengukur tingkat similaritas antar dua buah teks berbahasa Indonesia.

BAB 5 KESIMPULAN DAN SARAN

Pada bab ini membahas tentang kesimpulan dari penelitian ini dan saran yang dapat dijadikan masukan untuk pengembangan penelitian deteksi kesamaan dokumen selanjutnya.