BAB I

PENDAHULUAN

1.1 Latar Belakang

Kemajuan teknologi dan kemudahan akses informasi yang dilakukan melalui internet ternyata tidak hanya memberikan dampak positif terhadap dunia pendidikan namun juga memberikan dampak negatif. Salah satu dampak negatifnya adalah merebaknya kasus plagiarisme karya tulis. Dalam Kamus Besar Bahasa Indonesia (2018) disebutkan: "Plagiat adalah pengambilan karangan (pendapat dan sebagainya) orang lain dan menjadikannya seolah-olah karangan (pendapat dan sebagainya) sendiri, misalnya menerbitkan karya tulis orang lain atas nama dirinya sendiri; jiplakan". Plagiarisme sendiri merupakan salah satu pelanggaran hukum. Undang-undang No. 20 Tahun 2003 mengatur sanksi bagi orang yang melakukan plagiat, khususnya yang terjadi di lingkungan akademik. Sanksi tersebut adalah sebagai berikut (Pasal 70): "Lulusan yang karya ilmiah yang digunakannya untuk mendapatkan gelar akademik, profesi, atau vokasi sebagaimana dimaksud dalam Pasal 25 Ayat (2) terbukti merupakan jiplakan dipidana dengan pidana penjara paling lama dua tahun dan/atau pidana denda paling banyak Rp 200.000.000,00 (dua ratus juta rupiah)". (Istiana & Purwoko, 2014)

Selain merugikan pihak yang diplagiasi, plagiarisme memberikan dampak negatif yang sangat besar terhadap pelakunya. Di samping itu, dengan melakukan plagiarisme, tentunya tingkat kreativitas dari pelaku akan menurun.

Dari besarnya dampak negatif plagiarisme, perlu dilakukan pencegahan khususnya dilakukan oleh badan/instansi pendidikan. Banyak upaya yang dilakukan untuk mencegah terjadinya kasus plagiarisme. Salah satunya adalah dengan meningkatkan peran pemeriksa tulisan dalam mencegah plagiarisme. Dari cara pencegahan ini, perlu dilakukan suatu proses validasi dan pendeteksian plagiarisme terhadap suatu tulisan sebelum dapat dipublikasikan. Plagiarisme sendiri dapat dideteksi dengan metode-metode yang telah dikembangkan oleh para ahli.

Levenshtein *Distance* merupakan salah satu metode yang bisa dipergunakan untuk mendeteksi plagiarisme. Levenshtein *Distance* adalah sebuah matriks *string* yang digunakan untuk mengukur perbedaan atau jarak *(distance)* antara dua *string*. Nilai *distance* antara dua *string* ini ditentukan oleh jumlah minimum dari operasi-operasi perubahan yang diperlukan untuk melakukan transformasi dari suatu *string* menjadi *string* lainnya. Operasi-operasi tersebut adalah penyisipan *(insertion)*, penghapusan *(deletion)*, atau penukaran *(subtitution)*. Algoritma Levenshtein *distance* dapat diimplementasikan untuk mendeteksi kemiripan suatu dokumen teks dengan dokumen teks lainnya dengan cara pembentukan suatu matriks *string* untuk mendapatkan nilai *distance*, kemudian melakukan perhitungan bobot *similarity* antar dokumen teks berdasarkan nilai *distance* tersebut. (Pratama & Pamungkas, 2016) Walaupun Levenshtein *Distance* efektif dipergunakan untuk melakukan pendeteksian plagiarisme, algoritma ini memerlukan waktu yang cukup lama pada pemrosesannya terlebih *string* yang akan dibandingkan berukuran besar. Pada penelitian (Balhaf, Shehab, & Al-Sarayrah, 2016) mengatakan bahwa algoritma

ini memiliki kompleksitas algoritma yang tergolong tinggi dan tidak disarankan untuk menerapkan pada kasus yang memiliki ukuran data yang besar. Kompleksitas waktu untuk algoritma ini adalah O(|s1|×|s2|) dimana ukuran s1 dan s2 merupakan ukuran dari *string-string* yang dibandingkan. (Mulyanto, 2010)

Metode lainnya yang biasa dipergunakan untuk melakukan pendeteksian plagiarisme adalah penggunaan metode n-gram. Teknik n-gram didasarkan pada pemisahan teks menjadi *string* dengan panjang *n* mulai dari posisi tertentu dalam suatu teks. Posisi n-gram berikutnya dihitung dari posisi yang sebenarnya bergeser sesuai dengan *offset* yang diberikan. Nilai *offset* bergantung pada pembagian yang digunakan dalam n-gram. Pembagian n-gram dapat bervariasi tergantung dari pendekatan dalam membagi teks menjadi bentuk n-gram. N-gram untuk setiap *string* dihitung dan kemudian dibandingkan satu per satu. N-gram dapat berupa unigram (n=1), bigram (n=2), trigram (n=3), dan seterusnya. (Sugianto, Liliana, & Rostianingsih, 2013)

Selain kedua metode yang telah disebutkan di atas, metode populer lainnya adalah Jaro-Winkler *Distance*. Jaro-Winkler *Distance* merupakan varian dari Jaro *Distance* metrik yaitu sebuah algoritma untuk mengukur kesamaan antara dua *string*, biasanya algoritma ini digunakan dalam pendeteksian duplikat. Semakin tinggi nilai Jaro-Winkler *Distance* untuk dua *string*, semakin mirip kedua *string* tersebut. Pada penelitian yang dilakukan oleh (Rochmawati & Kusumaningrum, 2016) mengatakan bahwa metode *Jaro-Winkler Distance* memiliki akurasi yang tinggi untuk melakukan pendeteksian kesamaan dua buah *string* yang ditunjukan dengan nilai *Mean Average Precision* (MAP) yang tinggi.

Pada bidang sistem informasi, beberapa sistem yang berhubungan langsung dengan bidang akademik kini telah menerapkan pendeteksian plagiarisme. Salah satu yang paling terkenal adalah Turnitin. Turnitin adalah aplikasi yang tidak hanya melakukan *scanning* untuk pengecekan similaritas, namun juga menyediakan layanan berupa manajemen aplikasi secara lebih terstruktur. Untuk dapat menggunakan aplikasi ini, pengguna diwajibkan berlangganan (berbayar). (Afdhal, Chalis, & Gani, 2014) Turnitin menggunakan sebuah *indeks* yang digunakan sebagai indikator *similarity* dari dokumen teks berbasis pada banyak kesamaan teks yang ditemukan. *Indeks* tersebut sering disebut Turnitin *Similarity Index*.

Selain Turnitin, pada Perguruan Tinggi Universitas Pendidikan Ganesha (Undiksha) telah mengembangkan dan menerapkan penggunaan suatu sistem *Plagiarism Detector* yang dapat diakses di laman *http://ejournal.undiksha.ac.id:8080/PlagiarismDetector* yang dikembangkan pada tahun 2013. Setiap karya ilmiah dosen yang akan dipublikasikan, akan melewati tahap pengecekan plagiarisme melalui sistem ini dengan demikian kualitas dari karya ilmiah yang akan dipublikasikan tetap terjaga. Namun, sistem informasi tersebut masih perlu untuk dikembangkan lagi terutama pada metode pengecekan plagiarisme.

Walaupun beberapa sistem informasi telah mengimplementasikan metode-metode yang dipergunakan untuk pengecekan suatu tulisan untuk pendeteksian plagiarisme, namun perlu dibuatnya suatu *service* yang berdiri sendiri dan dapat dengan mudah dipergunakan oleh suatu sistem lain dalam penentuan kesamaan dua *string* tanpa harus menanamkan metode-metode

pencocokan *string* pada sistem tersebut. Dengan dibuatkan *service* ini, sistem penggunanya juga dapat memanfaatkannya untuk melakukan pencarian dengan menggunakan kata kunci selain tujuan awal yaitu untuk mendeteksi plagiarisme.

Dari pemaparan tersebut di atas, peneliti ingin memberikan suatu solusi yang dapat dipergunakan oleh sistem-sistem yang membutuhkan proses pengecekan plagiarisme yaitu dengan membangun sebuah *rest web service* yang berguna untuk melakukan deteksi tingkat kesamaan antara dua *string*. Sistem yang akan dibangun akan mengimplementasikan 3 buah metode yang akan dibandingkan satu sama lain. Alasan penulis mengulas mengenai Levenshtein *Distance*, n-gram, dan Jaro-Winkler *Distance* adalah dalam kasus ini penulis ingin melakukan improvisasi terhadap metode Leveinstein *Distance* dan melakukan pengujian akurasi dengan metode Jaro-Winkler *Distance* sebagai pembanding. Telah disebutkan sebelumnya, bahwa metode Leveinstein *Distance* memiliki nilai kompleksitas algoritma yang tergolong tinggi dan membutuhkan waktu eksekusi yang besar yang dikarenakan ukuran ordo matrik yang diproses. Penulis ingin mencoba memadukan antara metode Leveinstein *Distance* dan n-gram. Penggunaan n-gram bertujuannya untuk mengurangi ukuran ordo matriks yang diproses pada metode Levenshtein *Distance* dengan harapan dapat mempercepat proses yang dilakukan. Jaro-Winkler *Distance* dalam hal ini akan dijadikan sebagai salah satu acuan pembanding akurasi *similarity* karena dikatakan bahwa Jaro-Winkler *Distance* memiliki tingkat akurasi yang tinggi.

Pengembangan *Rest Web Service* yang menerapkan metode Levenshtein *Distance*, Jaro-Winkler *Distance* serta Levenshtein *Distance* berbasis n-gram belum pernah dikembangkan sebelumnya. Sehingga dalam penelitian ini akan

dibangun sebuah sistem yang berupa *Rest Web Service* yang mengimplementasikan ketiga metode tersebut sebagai pengukur persentase plagiarisme pada *string*.

Meskipun tujuan utama pengembangan sistem ini adalah untuk memberikan kemudahan dalam melakukan deteksi plagiarisme dan improvisasi algoritma yang telah ada, tidak menutup kemungkinan metode-metode yang digunakan memiliki hasil yang kurang maksimal, baik dari akurasi pendeteksian maupun kecepatan prosesnya. Maka, dari hal itu dikembangkan tesis dengan judul "Penerapan Metode Levenshtein *Distance* Berbasis N-gram untuk Meningkatkan Akurasi pada Pengembangan *Rest Web Service* Deteksi Plagiarisme".

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, dapat dirumuskan beberapa masalah yang akan dibahas dalam penelitian ini.

1. Bagaimana implementasi algoritma Levenshtein *Distance* pada pengembangan *rest web service* deteksi plagiarisme?
2. Bagaimana implementasi algoritma Levenshtein *Distance* berbasis n-gram yang diharapkan untuk mempercepat proses yang dilakukan pada metode Levenshtein *Distance* pada pengembangan *rest web service* deteksi plagiarisme?
3. Bagaimana implementasi algoritma Jaro-Winkler *Distance* pada pengembangan *rest web service* deteksi plagiarisme?

4. Bagaimana perbandingan efektivitas Levenshtein *Distance*, Levenshtein *Distance* berbasis n-gram, serta Jaro-Winkler *Distance* pada pengembangan *rest web sevice* deteksi plagiarisme?

1.3 Ruang Lingkup dan Keterbatasan Penelitian

Berikut merupakan ruang lingkup dan keterbatasan penelitian yang akan diteliti oleh penulis.

1. Penelitian ini menggunakan data dokumen pada *e-journal* Universitas Pendidikan Ganesha bagian isi untuk studi kasus pengujian *string* ukuran panjang dan bagian abstrak untuk pengujian *string* ukuran sedang serta kata-kata yang bersumber dari kbbi untuk *string* pendek.
2. Penelitian ini menghasilkan nilai akurasi kesamaan *string* dan hasil pembandingan kecepatan serta akurasi algoritma Levenshtein *Distance* dan Levenshtein *Distance* berbasis n-gram dengan menggunakan hasil deteksi plagiarism dari Jaro-Winkler *Distance* dan *prepostseo* sebagai acuan.

1.4 Tujuan Penelitian

Tujuan penelitian ini adalah sebagai berikut :

1. Untuk mengetahui implementasi algoritma Levenshtein *Distance* pada pengembangan *rest web service* deteksi plagiarisme.
2. Untuk mengetahui implementasi algoritma Levenshtein *Distance* berbasis n-gram dengan harapan mempercepat proses yang dilakukan pada metode Levenshtein *Distance* pada pengembangan *rest web service* deteksi plagiarisme.

3. Untuk mengetahui implementasi algoritma Jaro-Winkler *Distance* pada pengembangan *rest web service* deteksi plagiarisme.
4. Untuk mengetahui perbandingan algoritma Levenshtein *Distance*, Levenshtein *Distance* berbasis n-gram, dan Jaro-Winkler *Distance* pada pengembangan *rest web service* deteksi plagiarisme.

BAB II

KAJIAN TEORI

2.1 Plagiat

Dalam Kamus Besar Bahasa Indonesia disebutkan: "Plagiat adalah pengambilan karangan (pendapat dan sebagainya) orang lain dan menjadikannya seolah-olah karangan (pendapat dan sebagainya) sendiri, misalnya menerbitkan karya tulis orang lain atas nama dirinya sendiri; jiplakan" (Arti kata plagiat, 2018). Plagiarisme merupakan salah satu bentuk tindak pidana. Undang-Undang No. 20 Tahun 2003 mengatur sanksi bagi orang yang melakukan plagiat, khususnya yang terjadi di lingkungan akademik. (Salmuasih & Sunyoto, 2013) Selain merugikan pihak yang diplagiasi, plagiarisme memberikan dampak negatif yang sangat besar terhadap pelakunya. Di samping itu, dengan melakukan plagiarisme, tentunya tingkat kreativitas dari pelaku akan menurun.

Banyak upaya yang dilakukan untuk mencegah terjadinya kasus plagiarisme. Salah satunya adalah dengan meningkatkan peran pemeriksa tulisan dalam mencegah plagiarisme dengan mendeteksi nilai plagiarisme dari suatu tulisan. Salah satu yang paling terkenal adalah Turnitin. Turnitin adalah aplikasi yang tidak hanya melakukan *scanning* untuk pengecekan similaritas, namun juga menyediakan layanan berupa manajemen aplikasi secara lebih terstruktur. Untuk dapat menggunakan aplikasi ini, pengguna diwajibkan berlangganan (berbayar).

(Afdhal, Chalis, & Gani, 2014) Turnitin menggunakan sebuah *indeks* yang digunakan sebagai indikator *similarity* dari dokumen teks berbasis pada banyak kesamaan teks yang ditemukan. Indeks tersebut sering disebut Turnitin *Similarity Index*.

Dari hal teknis, para ahli telah mengembangkan metode-metode untuk pendeteksian persentase *similarity* antar teks. Persentase *similarity* ini sering dipergunakan untuk mendeteksi plagiarisme antar tulisan. Konsep yang diterapkan pada metode pendeteksian plagiarisme adalah menghitung nilai kesamaan karakter atau kata pada 2 buah teks. Salah satu metode yang menerapkan konsep ini adalah metode Jaro-Winkler *Distance*. Selain itu, ada juga metode yang mencari nilai *distance* atau perbedaan karakter atau kata pada 2 buah teks untuk dipergunakan dalam mendeteksi nilai *similarity*. Salah satu contoh metode yang menerapkan hal ini adalah metode Levenshtein *Distance*.

2.2 Levenshtein *Distance*

Levenshtein *Distance* adalah sebuah matriks *string* yang digunakan untuk mengukur perbedaan atau jarak (*distance*) antara dua *string*. Nilai *distance* antara dua *string* ini ditentukan oleh jumlah minimum dari operasi-operasi perubahan yang diperlukan untuk melakukan transformasi dari suatu *string* menjadi *string* lainnya. Operasi-operasi tersebut adalah penyisipan (*insertion*), penghapusan (*deletion*), atau penukaran (*subtitution*). Levenshtein *Distance* merupakan salah satu algoritma yang dapat digunakan dalam mendeteksi kemiripan antara dua *string* yang berpotensi melakukan tindak plagiarisme.