

BAB 1 PENDAHULUAN

1.1 Latar Belakang

Di zaman yang sedang mengalami pertumbuhan teknologi yang sangat pesat, teknologi informasi menjadi salah satu bidang yang terdampak berkembangnya teknologi. Dampak berkembangnya teknologi informasi memberikan kemudahan seseorang untuk mendapatkan informasi yang dibutuhkan, dimana informasi yang dulunya hanya dapat diakses lewat media cetak dan televisi kini dapat diakses lewat telepon genggam (*smartphone*), ataupun laptop dengan menghubungkan ke internet. Beberapa website yang menyajikan berita seperti kompas.com, tribun.com dan masih banyak lainnya. Artikel beritanya yang telah didapatkan seringkali terdapat persamaan yang mempengaruhi pembaca mengenai kemiripan artikel berita satu dan berita lainnya, oleh karena itu diperlukan suatu sistem untuk mendeteksi kemiripan teks atau kalimat untuk membandingkan antar artikel berita sehingga diketahui tingkat kemiripan pada artikel berita. Untuk mengetahui tingkat kemiripan antar artikel perlu dilakukan pengecekan. Pengecekan dapat dilakukan dengan dua cara. pertama dilakukan secara manual dengan melakukan perbandingan antar artikel berita satu – persatu. Kedua dilakukan dengan menggunakan sistem pengukuran kemiripan menggunakan algoritma dari suatu metode yang ada[1]. Beberapa permasalahan yang ada pada saat melakukan deteksi kemiripan dokumen terdapat kata yang sama namun tidak terdeteksi oleh sistem, beberapa kata yang berbeda namun memiliki makna yang sama (sinonim) sulit untuk di deteksi. Maka dari itu harus ada sebuah sistem yang dapat melakukan deteksi kemiripan sebuah dokumen dengan tingkat akurasi yang bagus serta waktu perhitungan yang cepat, dan yang dapat mendeteksi kata yang bersinonim[2]. Pada penelitian 2019 yang dilakukan oleh Shella Batrisya Lupitasari menggunakan Metode *Levenshtein Distance* disebutkan saat dilakukan pengujian pada dokumen, kata yang memiliki makna sama tidak dapat dideteksi oleh metode *Levenshtein Distance*[3].

Beberapa proses menghitung kemiripan dokumen sendiri diantara pendekatan *Rabin Karp*, *Winnowing*, *Jaccard Coefficient*, dan *Levenshtein Distance*. Timothy Bolle dkk mengatakan bahwa dibandingkan dengan metode *Jaccard*, *Jaro Winkler* dan *Cosine Similarity*, pada metode *Levenshtein Distance* diperoleh nilai rasio terbaik pada rasio true positif dan terdapat beberapa *false negatif*[4]. Ade Riyanidkk menerapkan metode *Cosine Similarity* untuk mendeteksi kemiripan pada abstrak skripsi. Pada uji coba yang dilakukan menggunakan 1 dokumen uji yang akan dibandingkan dengan 5 abstrak. 1 Abstrak yang dibandingkan dengan 5 abstrak yang ada diperoleh dari Institut Teknologi Telkom

Purwokerto. Dari pengujian yang dilakukan, beberapa kata yang memiliki kesamaan makna namun beda kata (sinonim) tidak dapat dideteksi oleh metode *Cosine Similarity*[5]. Prima Putra dkk menggunakan metode *Rabin – Karp* dan pendekatan *Synonym Recognition* untuk mengantisipasi plagiarisme. Dimana dari penelitian yang dilakukan disimpulkan bahwa dengan menggunakan *Synonym Recognition* hasil yang diperoleh lebih baik dibandingkan tanpa menggunakan *Synonym Recognition*[6].

Penelitian ini menggunakan metode *Levenshtein Distance* yang dikombinasikan dengan *Synonym Recognition*. Dimana *Synonym Recognition* akan melakukan pengecekan dengan membandingkan kata antar dokumen untuk menemukan kata yang memiliki makna yang sama untuk dirubah ke kata yang terdapat pada database atau kamus *thesaurus*[6]. Dimana pada *Levenshtein Distance* digunakan sebuah matrik untuk membandingkan dan mengukur perbedaan jumlah *string* atau karakter pada metode *Levenshtein* ini[1]. Pada metode *Levenshtein Distance* ini tingkat kecepatan waktu proses mengikuti jumlah kata (*count word*). Semakin banyak kata pada sebuah dokumen maka waktu yang diperlukan untuk melakukan perbandingan semakin lama. Muhammad Hendra Febiawan dkk menerapkan metode *Levenshtein* untuk melakukan deteksi dini plagiarisme pada dokumen dengan data uji sebuah dokumen dengan jumlah kata 4405 dibandingkan dengan 3 dokumen dengan jumlah kata 13465 dengan kecepatan perhitungan sebesar 3,67 detik[7]. Deteksi kemiripan yang telah ada pada jurnal-jurnal sebelumnya telah ditentukan keakuratan dengan kecepatan waktu dalam proses lebih cepat dan hasil yang lebih baik dibandingkan tanpa menggunakan metode *Levenshtein Distance*[8]. Metode ini dipilih untuk mengetahui berapa akurasi yang akan didapat serta untuk menilai seberapa layaknnya metode ini digunakan untuk melakukan deteksi kemiripan pada artikel berita, agar menjadi salah satu acuan untuk para peneliti di masa mendatang.

1.2 Rumusan Masalah

1.2.1 Permasalahan

Pada penelitian sebelumnya disebutkan Metode *Levenshtein Distance* tidak dapat mendeteksi kata yang bersinonim. Oleh karena itu diperlukan penggunaan metode *Synonym Recognition* yang dikombinasikan dengan *Levenshtein Distance* untuk memperbaiki hasil pengujian dan waktu proses serta dapat mendeteksi kata yang saling bersinonim antar dokumen.

1.2.2 Usulan Solusi

1. Menggunakan metode *Levenshtein Distance* agar nilai keakuratan kemiripan dokumen dan waktu proses pada pengujian menghasilkan nilai yang baik.
2. Menggunakan *Synonym Recognition* untuk mendeteksi kata yang bersinonim antar dokumen.

1.2.3 Pertanyaan Penelitian

Berapa tingkat akurasi metode *Levenshtein Distance* yang dikombinasikan dengan *Synonym Recognition* dalam proses mengecek dokumen satu dengan dokumen lainnya yang akan dibandingkan kemiripannya?

1.3 Tujuan dan Manfaat

1.3.1 Tujuan Penelitian

Tujuan pada penelitian ini untuk mengetahui hasil akurasi kemiripan dokumen dengan pendekatan *Synonym Recognition* yang dikombinasikan dengan *Levenshtein Distance*.

1.3.2 Manfaat

1. Penelitian ini dapat menjadi sebuah saran ataupun pertimbangan untuk meningkatkan efisiensi pengukuran deteksi kemiripan antar dokumen pada penelitian selanjutnya.
2. Penggunaan *Synonym Recognition* pada metode ini dapat menjadi sarana ataupun pertimbangan untuk mengembangkan penggunaan *Synonym Recognition* dengan mengkombinasikan dengan metode – metode yang ada.

1.4 Batasan

1. Data yang akan diuji hanya dapat mendeteksi teks saja dan tidak dapat mendeteksi gambar, tabel dan angka seperti rumus.
2. Kesalahan penulisan atau ejaan pada dokumen tidak akan diperhatikan.
3. Data yang akan digunakan yaitu berita pariwisata.
4. Data dengan topik yang sama.
5. Kamus thesaurus yang digunakan terbatas
6. Data yang digunakan hanya berbahasa Indonesia.

1.5 Metode Penelitian

a. Studi Pustaka

Pada tahap studi pustaka pada metode penelitian ini dilakukan pengumpulan informasi serta mempelajari informasi yang terdapat di jurnal – jurnal dan dari media digital yang telah ada sebelumnya guna untuk membantu menyelesaikan penelitian ini.

b. Analisa dan Perancangan Sistem

Setelah data telah diperoleh, maka tahap selanjutnya membuat rancangan sistem, yang dimana pada rancangan sistem akan dibangun beberapa rancangan arsitektur yang akan digunakan pada tahap pengujian. Selain itu nantinya akan dilakukan suatu perancangan data, dan perancangan tampilan.

c. Implementasi Sistem

Pada tahap implementasi akan dilakukan text preprocessing, lalu melakukan pengecekan kata yang memiliki arti yang sama (sinonim). Tahap selanjutnya menghitung nilai distance dari dokumen yang dibandingkan, langkah terakhir menghitung bobot similaritas.

d. Uji Coba Sistem

Uji coba pada penelitian ini dilakukan dengan mengujikan data yang telah disiapkan, kemudian dilakukan analisa dan evaluasi untuk diketahui akurasi dari sistem yang telah dibuat.

e. Evaluasi

Setelah melakukan tahap uji coba maka masuk pada tahap evaluasi. Dimana evaluasi dilakukan untuk mengetahui keakuratan dari hasil uji coba yang telah dilakukan oleh sistem.

f. Dokumentasi

Dokumentasi akan dilakukan pada tahap studi pustaka hingga tahapan evaluasi. Tahap ini dilakukan untuk penyusunan laporan terhadap implementasi penelitian yang telah dilakukan serta mengambil kesimpulan dari tahap pengujian.

1.6 Sistematika Laporan Skripsi

BAB I Pendahuluan

Membahas latar belakang dari proposal, rumusan masalah yang ada, usulan solusi yang dikemukakan, tujuan dan manfaat, batasan masalah dan sistematika penulisan.

BAB II Tinjauan Pustaka

Membahas dasar tinjauan pustaka, landasan teori dan penelitian terkait.

BAB III Metode Penelitian

Membahas metode yang digunakan (*Synonym Recognition dan Levenshtein Distance*) serta tahapan – tahapan yang ada seperti dataset, proses arsitektur sistem, tahapan penelitian dan proses pengujian.

BAB IV Hasil Dan Pembahasan

Bab ini membahas implementasi dari metode yang digunakan, sistem yang telah dibuat, proses uji coba, serta analisis hasil dan evaluasi dari sistem yang telah dibuat.

BAB V Penutup

Bab ini membahas kesimpulan dari penelitian yang telah dilakukan dengan menyesuaikan dengan pertanyaan penelitian dan saran terkait pengembangan sistem selanjutnya.