BAB II

TINJAUAN PUSTAKA

1.1. Penelitian Terdahulu

Ledisma Juliana Purba, dkk (2018) melakukan penelitian berjudul "Perancangan Aplikasi untuk Menghitung Persentase Kemiripan Proposal dan Isi Skripsi dengan Algoritma Rabin-Karp". Dalam penelitian tersebut para peneliti melakukan perancangan aplikasi pendeteksi plagiasi guna menghindari kesamaan judul dan naskah pada proposal dan juga isi dari skripsi masing-masing mahasiswa. Langkah atau proses pada algoritma Rabin-Karp yang dilakukan untuk menemukan hasil presentasi akhir yaitu dengan menghilangkan tanda baca, menghapus spasi dan mengubah teks menjadi kata tanpa huruf kapital. Selanjutnya menentukan k-gram dan mencari nilai hash. Dari hasil penelitian tersebut ditemukan bahwa penerapan algoritma Rabin-Karp dengan rumus *Jaccard Coefficient* dapat menghitung dan menentukan hasil persentasi kemiripan antar proposal ataupun skripsi. Waktu yang digunakan dalam pengecekan dokumen bergantung pada banyaknya jumlah kata (Purba and Sitorus 2018).

Hirroe Kesuma, dkk (2016) dalam jurnal Teknik Elektro Universitas Negeri Semarang melakukan penelitian berjudul "Penerapan Cosine Similarity dalam Aplikasi Kitab Undang-Undang Hukum Dagang (Wetboek Van Koophandle Voor Indonesia)". Penelitian tersebut berjalan berdasarkan latar belakang permasalahan yaitu dalam aplikasi kitab Undang-undang hukum dagang terdapat banyak bab dan ratusan pasal yang akan menyulitkan jika mencari pasal

dengan cara manual ditiap lembar halaman. Maka dibutukan aplikasi yang dapat mencari indeks kata berdasarkan input pencarian dari user. Salah satu algoritma yang dapat digunakan untuk mencocokkan data dari undang-undang kitab dagang dengan kata yang dicari oleh pengguna yaitu *cosine similarity* untuk melihat kedekatan antar keduanya. Hasil dari penelitian ini yaitu metode tersebut berhasil melakukan fungsi dalam pencarian kata dengan tingkat keberhasilan berdasarkan nilai uji *performance measure* sebesar 55,04% (Wijaya, Kesuma, and Pribadi 2016).

Sugiyamto, dkk (2014) dalam jurnal Masyarakat Informatika Universitas Diponegoro melakukan penelitian berjudul "Analisa Performa Metode Cosine dan Jacard pada Pengujian Kesamaan Dokumen". Penelitian tersebut dilatarbelakangi oleh maraknya kegiatan dalam melakukan plagiat dalam karya ilmiah baik sengaja ataupun tidak sengaja. Analisa peforma dilakukan pada metode Cosine dan Jaccard untuk menentukan tingkat akurasi dalam kemiripan abstrak karya ilmiah dan menggunakan metode *single pass clustering* untuk mengklasifikasi dokumen. Hasil dari penelitian tersebut yaitu menunjukan bahwa tingkat akurasi metode Cosine jauh lebih tinggi jika dibandingkan dengan Jaccard namun hanya selisih 0,000731(Sugiyamto, Surarso, and Sugiharto 2014).

Berdasarkan Penelitian terdahulu yang dijadikan referensi, terdapat kesamaan tujuan yaitu untuk mengetahui tingkat akurasi dan meminimalisir tindakan memplagiat hasil karya orang lain. Perbedaan dari penelitian terdahulu yaitu belum ada yang melakukan proses evaluasi pengukuran pada algoritma Rabin Karp dengan membandingkan metode Cosine dan Jaccard pada

pendeteksian dokumen berdasarkan acuan dari dokumen berbahasa inggris (bilingual) sehingga nantinya dapat diketahui apakah dokumen yang dibuat merupakan hasil plagiasi dari karya ilmiah berbahasa inggris yang dengan sengaja atau tidak hanya di-translate begitu saja. Dengan dilakukan penelitian ini penulis berharap akan meminimalisir tindakan plasiarisme dan tingkat orisinalitas karya akan meningkat juga dapat dipertanggungjawabkan.

1.2. Plagiarisme

Pengertian plagiarisme menurut KBBI V merupakan tindakan penjiplakan yang melanggar hak cipta. Sedangkan plagiat berarti pengambilan atau penggunaan karangan milik orang lain dalam berbagai bentuk yang nantinya dijadikan seolah-olah milik sendiri.

Ruang lingkup plagiarisme menurut (Sukaesih 2018) yaitu tindakan menggunakan kalimat atau kata-kata, gagasan, teori, data dan informasi dari orang lain tanpa memberikan tanda kutip dan menyebutkan identitas sumber. Tindakan berikutnya ialah menyerahkan karya hasil penjiplakan dengan berbagai kepentingan kepada pihak lain seolah-olah karya tersebut milik sendiri.

1.3. Algoritma Rabin Karp

Algoritma Rabin Karp merupakan suatu algoritma yang bertujuan untuk melakukan pencarian string namun bukan untuk menemukan string yang cocok dengan masukan melainkan menentukan *pattern* atau pola yang dianggap sesuai dengan sting masukan. Algoritma Rabin Karp menggunakan teknik *hashing* untuk menemukan substring dalam sebuah teks. Algoritma ini diciptakan oleh Michael

O. Rabin dan Richard M. Karp ditahun 1987 (Herriyance, Handrizal, and Fadila 2017)

Cara kerja dari algoritma Rabin Karp yaitu dengan melakukan *hashing* pada *string* yang dicari (m) dengan *substring* pada teks (n). kemudian setelah dilakukan *hasing*, apabila nilai keduanya sama *(hash value)* maka akan dilakukan perbandingan sekali lagi pada karakter-karakternya. Namun apabila hasil keduanya tidak sama maka substring akan melakukan pergeseran ke kanan sebanyak (n-m) kali (Putra, Sujaini, and Pratiwi 2015).

1.4. Cosine

Cosine merupakan salah satu metode yang dapat digunakan untuk menghitung tingkat kesamaan antara 2 buah objek. Peerhitungan tersebut didasarkan pada *vector space similarity measure* yaitu dalam perhitungan tingkat kesamaan, objek akan dinyatakan dalam bentuk vector dengan menggunakan kata kunci (*keyword)* sebagai ukuran (Sugiyamta 2015). Perhitungan *Cosine Similarity* dirumuskan dengan:

$$Similarity\ x, y = 1 - \frac{|X \cap Y|}{|X^{\frac{1}{2}}| . |Y^{\frac{1}{2}}|}$$

Dimana

a. $|X \cap Y|$ adalah jumlah term yang ada pada dokumen X dan yang ada pada dokumen Y X
b. |X| adalahjumlah term yang ada pada dokumen X

c. |Y| adalah jumlah term yang ada pada dokumen Y

1.5. Jaccard

Jaccard Coefficient merupakan metode yang berfungsi untuk membandingkan 2 buah sampel teks atau dokumen berdasarkan kesamaan kata yang dimilikinya (Sunardi, Yudhana, and Mukaromah 2018). *Jaccard Similarity* dirumuskan dengan:

$$Similarity\ (X,) = |X \cap Y|\ |X \cup Y|$$

Dimana:

X = Dokumen 1

Y = Dokumen 2

1.6. Koefisien Dice

Yang dikenal juga dengan sebutan *Sørensen–Dice index* merupakan metode yang digunakan untuk membandingkan tingkat similaritas dari dua objek. Metode ini dipublikasikan oleh Sørensen dan Lee Raymond Dice pada 1948 dan 1945 secara berturut-turut. Rumus ini memiliki kesamaan dengan rumus Jaccard. Namun, perbedaannya terletak pada adanya pencocokan dua kali pada rumus Dice's coefficient (Sunyoto 2013). Berikut adalah rumus *Dice's coefficient*:

$$\text{dice}(D,Q) = \frac{2|D \cap Q|}{|D|+|Q|}$$

Dimana:

D = himpunan set string dokumen 1

Q = himpunan set string dokumen 2

1.7. Euclidean

Euclidean distance adalah perhitungan jarak dari 2 buah titik dalam euclidean space. Euclidean berkaitan dengan teorema phytagoras, dan dapat diterapkan biasanya pada ruang 1, 2 dan 3 dimensi, Euclidean juga tetap sederhana jika diterapkan pada dimensi yang lebih tinggi (Rizaldi, Kurniawati, and Angkoso 2018). *Euclidean* dirumuskan dengan:

$$D(a,b) = \sqrt{\sum_{i=1}^{n}(b_i - a_i)^2}$$

Dimana:

D = Jarak kedua titik

b = koordinat titik akhir

a = koordinat titik awal

1.8. Manhattan

Manhattan distance adalah salah satu pengukuran yang paling banyak digunakan meliputi penggantian perbedaan kuadrat dengan menjumlahkan perbedaan absolute antar variable, dimana **a** dan **b** adalah koordinat titik atau vektor kedua dokumen. Prosedur ini disebut blok absolute atau lebih dikenal dengan city block distance (Arifin 2014). *Manhattan* dirumuskan dengan:

$$D = \sum_{i=1}^{n} |b_i - a_i|$$

1.9. Minkowski

Minkowski distance merupakan sebuah metrik dalam ruang vektor di mana suatu norma didefinisikan (normed vector space) sekaligus dianggap sebagai generalisasi dari Euclidean distance dan Manhattan distance. Dalam pengukuran jarak objek menggunakan minkowski distance biasanya digunakan nilai p adalah 1 atau 2. Berikut rumus yang digunakan menghitung jarak dalam metode ini (Hartono and Lusiana 2017). *Minkowski distance* dirumuskan dengan:

$$D(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

Dimana:
d = jarak antara x dan y
x = data pusat klaster
y = data pada atribut
i = setiap data
n = jumlah data,
xi = data pada pusat klaster ke i
yi = data pada setiap data ke i
p = power

1.10. Mahalanobis

Mahalanobis menggunakan variabel dengan sampel matriks varian-kovarian, karena matriks kovarian juga menggunakan rata-rata korelasi diantara variabel (Jannah 2010). *Mahalanobis* dirumuskan dengan:

$$D(x,y) = \sqrt{(a-b)\sum^{-1}(a-b)^T}$$

dimana:

T = Transpose dari sebuah matriks

-1 = Inverse dari sebuah matriks

Σ = Variance covariance matrix

1.11. **Weighted Distance**

Pada **Weighted Distance** tiap variabel dapat diberi bobot sesuai tingkat kepentingannya (Rao and Singh 2012). Contoh weighted pada Euclidean Distancenya dapat dihitung sebagai berikut:

$$D(x,y) = \sqrt{w_1|a_1 - b_1|^2 + w_2|a_2 - b_2|^2 + \cdots + w_n|a_n - b_n|^2}$$

1.12. Bilingual

Bilingual menurut KBBI V yaitu penggunaan dua bahasa atau lebih oleh penutur bahasa atau suatu kelompok masyarakat. Menurut (Sugianto 2014) penggunaan bilingual dalam kehidupan sehari-hari dikarenakan bahasa pada perbedaan bahasa ditiap tempat, keterbatasan sumber informasi membuat penggunaan bilingual menjadi hal yang wajar dan telah diterapkan pada beberapa instansi pendidikan. Di Indonesia sering diartikan dengan penggunaan bahasa Indonesia dan Inggris.

1.13. Python

Dalam penelitiannya (Rosmala and Dwipa 2012) menjelaskan, Python adalah salah satu bahasa pemrograman tingkat tinggi yang bersifat interpreter, interaktif, object-oriented dan dapat beroprasi di hamper semua platform. Sebagai bahasa tingkat tinggi, Python termasuk salah satu bahasa pemrograman yang

mudah untuk dipelajari karena sintaks yang jelas dan elegan, dikombinasikan dengan penggunaan module-module siap pakai dan struktur data tingkat tinggi yang efisien. Python merupakan bahasa pemrograman dinamis yang mendukung pemrograman berbasis objek. Python didistribusikan dengan beberapa lisensi yang berbeda dari beberapa versi. Namun pada prinsipnya Python dapat diperoleh dan dipergunakan secara bebas, bahkan untuk kepentingan komersial.