BAB II
TINJAUAN PUSTAKA

2.1. Algoritma Winnowing

Algoritma winnowing merupakan salah satu algoritma yang bertujuan untuk membuat dokumen *fingerprinting*. Algoritma ini merupakan pengembangan dari Rabin-Karp dengan penambahan metode window (Schleimer, Wilkerson, & Aiken, 2003). *Fingerprint* ini dihasilkan setelah melalui rangkaian proses sanitize, pembentukan n-gram, pembentukan hash dari n-gram, pembentukan window dan selanjutnya menghasilkan *fingerprint*. Teknik hash yang digunakan ada bermacam-macam. Dalam penelitian ini menggunakan teknik rolling hash untuk membuat hash dari rangkaian n-gram. Penjelasan proses lengkapnya adalah sebagai berikut.

a. Sanitize, yaitu pembersihan/penghapusan teks dari karakter yang bukan merupakan huruf dan angka seperti tanda baca dan karakter simbol. Kemudian proses case-folding untuk mengubah huruf kapital menjadi huruf kecil. Sebagai contoh, sebuah teks "Info: Penggunaan masker bermanfaat untuk mencegah penularan virus berbahaya" akan menjadi "info penggunaan masker bermanfaat untuk mencegah penularan virus berbahaya".
b. Pembentukan rangkaian n-gram yang merupakan pengelompokan kata yang berdekatan dengan **n** panjang tiap gramnya. Rangkaian yang digunakan dalam penelitian ini adalah rangkain kata sebanyak **n**. Pada contoh ini diberikan perumpamaan nilai n-gram = 5. Dari teksi diatas akan menjadi rangkaian : *[info penggunaan masker bermanfaat untuk] [penggunaan masker bermanfaat untuk mencegah] [masker bermanfaat untuk mencegah penularan] [bermanfaat untuk mencegah penularan virus] [untuk mencegah penularan virus berbahaya]*.
c. Penghitungan Hash dari setiap gramnya. Hashing adalah pengubahan teks menjadi kunci yang panjangnya tetap sama meskipun teks yang dihasing memiliki panjang yang berbeda, kunci ini mewakili teks aslinya. Proses pengubahan menjadi nilai hash menggunakan fungsi rolling hash.

d. Pembagian rangkaian hash kedalam window dengan panjang **n**. Teknik ini mirip dengan n-gram, namun objeknya adalah rangkain hash yang menjadi beberapa window dengan panjang tiap window sejumlah n.
e. Pemilihan beberapa nilai hash untuk membentuk *fingerprint* dokumen. Proses ini adalah pemilihan nilai paling kecil di tiap window. Jika ada dua nilai terkecil dalam window maka dipilih nilai paling kanan (urutan terbakhir). Hasil pemilihan tersebut disimpan dalam rangkaian sehingga menghasilkan *fingerprint* dokumen.

2.2. N-Gram

Model dokumen N-gram memungkinkan kesamaan terjadi diukur berdasarkan urutan kata-kata terdekat yang tumpang tindih (frasa) daripada kata-kata individual. Ngram menangkap beberapa bentuk kesamaan sintaksis dan kontek kalimat antara dokumen dan menghindari kekurangan asumsi independensi kata yang membatasi model *bag of word* (Johnson & Zhang, 2015).

N-gram pada dasarnya digunakan dalam penelitian ini untuk membedakan dan mengkategorikan dokumen berdasarkan ukuran n-gram yang serupa. Misalnya mayoritas dokumen yang sangat mirip dapat dideteksi menggunakan jumlah urutan n-gram lebih tinggi (n-gram lebih panjang) dari dokumen yang ditinjau dengan ringan. Karenanya menggunakan n-gram panjang (ukuran) tertentu bisa membantu membedakan satu kelas dokumen yang serupa dari yang lain. Namun ukuran model n-gram harus dipilih dengan cermat untuk menghindari memintas dokumen berpotensi menjiplak atau mendeteksi dokumen dari kategori terdekat menghasilkan false positif dan penurunan kinerja. Untuk mendapatkan n-gram terbaik untuk kategori tertentu perlu menguji n-gram dengan panjang berbeda secara berurutan (Thompson, Panchev, & Oakes, 2015).

2.3. Rolling Hash

Rolling hash adalah metode hashing yang digunakan untuk mencari nilai hash dari rangkaian grams yang telah terbentuk dan memberikan kemampuan untuk menghitung nilai tanpa mengulangi seluruh string (Sunardi, Yudhana, & Mukaromah, 2018). Rolling hash (juga dikenal sebagai rekursif hashing atau rolling checksum) adalah fungsi hash di mana input di-hash dalam window yang bergerak

melalui input. Beberapa fungsi hash memungkinkan rolling hash untuk dihitung dengan sangat cepat. nilai hash baru dihitung dengan cepat mengingat hanya nilai hash lama, nilai lama dihapus dari *window*, dan nilai baru ditambahkan ke *window*. ini mirip dengan cara fungsi rata-rata bergerak dapat dihitung jauh lebih cepat daripada filter low-pass lainnya.

2.4. Cosine Similarity

Metode Cosine Similarity merupakan metode ukuran kesamaan yang menghitung sudut antara dua vector (Sugiyamta, 2015). Vector disini adalah *fingerprint* dokumen dari dua dokumen. Jika vektor adalah satuan panjang, cosinus dari sudut antara mereka hanyalah dot product dari vektor, persamaannya sebagai berikut.

$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

A : vektor A

B : vektor B

|A| : besar dari vektor A

|B| : besar dari vektor B

2.5. Jaccard Similarity

Metode Jaccard Similarity adalah salah satu metode yang dipakai untuk menghitung similarity antara dua obyek. Ditemukan oleh Paul Jaccard yang merupakan metode ukuran kesamaan yang digunakan untuk membandingkan kesamaan dan keragaman 2 set sampel (Sunardi et al., 2018). Persamaannya sebagai berikut.

$$Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

J(A,B) : nilai similaritas antara set A dan set B

|A n B| : banyaknya elemen irisan antara set A dan set B.

|A u B| : banyak gabungan elemen yang ada pada set A dan set B

|A| : banyak elemen yang terdapat pada set A

|B| : banyak elemen yang terdapat pada set B

2.6. Dice Similarity

Metode Dice Similarity 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 (Fikri, 2019). Rumus ini memiliki kesamaan dengan rumus Jaccard. Namun, perbedaannya terletak pada adanya pencocokan dua kali pada rumus Dice's coefficient. Berikut adalah rumus Dice's coefficient.

$$dice(A,B) = \frac{2\,|A \cap B|}{|A| + |B|}$$

A : vektor A

B : vektor B

|A| : besar dari vektor A

|B| : besar dari vektor B

2.7. Euclidean Distance

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 (Nishom, 2019).

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

Dimana D adalah jarak kedua titik, b adalah koordinat titik akhir dan a adalah koordinat titik awal

2.8. Manhattan Distance

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 (Nishom, 2019). Prosedur ini disebut blok absolute atau lebih dikenal dengan city block distance.

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

2.9. Minkowski Distance

Minkowski Distance merupakan generalisasi dari Euclidean dan Manhattan Distance (Nishom, 2019). Dalam pengukuran jarak objek menggunakan minkowski distance biasanya digunakan nilai r adalah 1 atau 2. Berikut rumus yang digunakan menghitung jarak dalam metode ini.

$$D(a,b) = \left(\sum_{j=1}^{d} | x_j - y_j |^r \right)^{\frac{1}{2}}, r \geq 1$$

Nilai r disebut sebagai order dari Minkowski distance. Jika r = 2 dan 1, maka formulasi jarak tersebut masing-masing adalah Euclidean Distance dan Manhattan Distance.

2.10. Mahalanobis Distance

Mahalanobis Distance dapat mengurangi distorsi (penyimpangan) jarak yang disebabkan oleh kombinasi linier dari atribut (Zhang, Zhou, & Feng, 2015). Berikut ini adalah rumus persamaan metode ini.

$$D_{mah}(a,b) = \sqrt{(a-b)\sum^{-1}(a-b)^T}$$

(.)T : Transpose dari sebuah matriks

(.)-1 : Inverse dari sebuah matriks

Σ : Variance covariance matrix

2.11. Weighted Distance

Pada Weighted Distance tiap variabel dapat diberi bobot sesuai tingkat kepentingannya. Metode ini merupakan bentuk modifikasi dari metode Euclidean dengan penambahan bobot pada tiap variable. Contoh weighted pada Euclidean Distancenya dapat dihitung sebagai berikut (Shirkhorshidi, Aghabozorgi, & Ying Wah, 2015).

$$D(a,b) = \sqrt{w_1 \mid a_1 - b_1 \mid^2 + w_2 \mid a_2 - b_2 \mid^2 + \ldots + w_n \mid a_n - b_n \mid^2}$$

w = bobot untuk tiap vector

a = vektor a

b = vektor b

|a - b| = nilai mutlak dari pengurangan vektor a dan b

2.12. Bilingual

Bilingual menurut KBBI yaitu bersangkutan dengan atau mengandung dua bahasa. 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.

2.13. Google Translate API

Application Programming Interface (API) atau Antarmuka Pemrograman Aplikasi adalah sekumpulan perintah, fungsi, dan protokol yang dapat digunakan oleh programmer saat membangun perangkat lunak untuk sistem operasi tertentu (Ichwan & Hakiky, 2011). API adalah sebuah teknologi untuk memfasilitasi pertukaran informasi atau data antara dua atau lebih aplikasi perangkat lunak. API adalah antarmuka virtual antara dua fungsi perangkat lunak yang saling bekerja sama. Sebuah API mendefinisikan suatu cara bagaimana seorang pengembang apikasi memanfaatkan suatu fitur tertentu dari sebuah komputer tanpa perlu mengetahui atau mengakses kode sumber maupun semua fungsi aplikasi tersebut.

Google Translate adalah layanan dari Google sebagai mesin terjemahan mesin yang mendukung banyak bahasa untuk menerjemahkan teks. Google Translate menawarkan antarmuka web, aplikasi seluler untuk Ponsel pintar secara gratis, dan beberapa kondisi untuk layanan API berbayar yang membantu pengembang membangun ekstensi peramban dan aplikasi perangkat lunak. Google translate API merupakan API dari mesin penerjemahan Google yang bisa digunakan oleh para akademisi dalam menerjemahkan literatur-literatur yang ditulis dalam bahasa asing, misalnya dalam bahasa Inggris dan juga digunakan para pengembang aplikasi untuk menjadikan fleksibiltas bahasa aplikasi yang tampilkan untuk pengguna. Google translate dapat menerjemahkan mulai dari kata, kalimat dan bahkan semua isi dokumen. Mesin penerjemahan ini sangat membantu untuk menerjemahkan bahasa asing ke dalam bahasa Indonesia (Pujiati 2017). Metode yang dipakai di Google Translate API yaitu:

Translate : Menterjemahkan suatu sumber teks ke bahasa yang dimaksudkan.
Detect : Mendeteksi bahasa yang ada pada sumber teks.
Languages : Daftar bahasa yang didukung untuk penterjemahan.

Fitur Translate and Detect mengharuskan pengembang membayar sejumlah karakter yang telah ditentukan. Namun anda bisa memakai method yang ketiga yaitu Languages. Akan tetapi kita perlu menentukan terlebih dahulu bahasa apa yang diinginkan secara manual. Untuk biaya pemakain API ini perbulannya adalah sebagai berikut.

Translation API version	Feature	Up to 500,000 characters	500,000 –1 billion characters
Translation API v2, v3	Language detection	$20 per million characters*	$20 per million characters*
Translation API v2	Text translation	$20 per million characters*	$20 per million characters*
Translation API v3	Text translation (PBMT general models)	$20 per million characters*	$20 per million characters*
Translation API v3	Text translation (NMT general models)	Free	$20 per million characters*
Translation API v3	Text translation (AutoML models)	Free	$80 per million characters*

Gambar 2.13.1 Google API Translate

Harga dihitung per karakter yang dikirim ke API untuk diproses, termasuk karakter spasi. Kueri kosong juga dikenai biaya untuk satu karakter. Google membebankan biaya berdasarkan karakter, bahkan karakternya adalah beberapa byte. Untuk tingkat gratis hanya tersedia untuk API v3. Biaya akan dikenakan ketika karakter yang diproses mencapai 500.000 – 1 juta karakter dengan biaya 20 – 80 dolar.

2.14. Python

Python adalah salah satu bahasa pemrograman tingkat tinggi yang bersifat interpreter, interaktif, object-oriented dan dapat beroperasi di hampir sistem operasi. sebagai bahasa tingkat tinggi, Python menjadi bahasa pemrograman yang mudah untuk dipelajari karena sintaks yang didesain bersifat *readable* untuk manusia (Rosmala & L, 2012). Kemudian pada proses *runtime* program lebih cepat dikarenakan bersifat interpreted yang mana kode sumber tidak perlu melewati compiler namun langsung pada interpreternya. Python juga memiliki banyak modul-modul yang mendukung berbagai operasi khususnya pada bidang saintis dan matematika seperti Scipy, Numpy dan banyak lainnya. Oleh karena itu Bahasa Python penulis pilih untuk digunakan dalam melakukan penelitian ini.

2.15. Penelitian Terdahulu

Ilham dkk, (2017) melakukan penelitian tentang "Penerapan Algoritma Winnowing Untuk Mendeteksi Kemiripan Pada Karya Tulis Mahasiswa". Dalam penelitian tersebut para peneliti melakukan perancangan aplikasi pendeteksi plagiasi guna menghindari kesamaan pada karya tulis masing-masing mahasiswa. Proses pendetekisan pada algoritma Winnowing yang dilakukan untuk menemukan hasil presentasi akhir yaitu dengan menghilangkan tanda baca, menghapus spasi dan mengubah teks menjadi kata tanpa huruf kapital. Selanjutnya menentukan kgram dan mencari nilai hash dan proses windowing untuk membagi tiap hashnya. Kemudian hasil dari pemrosesan adalah *fingerprinting* dokumen yang akan dibandingkan dengan *fingerprint* dokumen pembanding. Namun disini tidak dijelaskan secara eksplisit metode apa yang digunakan dalam membandingkan antar *fingerprint*. Dari hasil tersebut ditemukan bahwa penerapan algoritma

Winnowing dapat menghitung dan menentukan hasil persentasi kemiripan antar karya ilmiah.

Reynald Karisma Wibowo dkk. (2016) melakukan penelitian tentang "Penerapan Algoritma Winnowing Untuk Mendeteksi Kemiripan Teks Pada Tugas Akhir Mahasiswa". Dalam penelitian ini para peneliti melakukan perancangan aplikasi pendeteksi plagiasi guna menghindari kesamaan judul dan naskah pada teks tugas akhir mahasiswa. Algoritma Winnowing dilakukan untuk menemukan hasil presentasi akhir berupa *fingerprinting* dokumen. Kemudian *fingerprinting* dokumen dibandingan dengan metode Jaccard Coefficient. Dari hasil penelitian tersebut ditemukan bahwa penerapan algoritma Winnowing dengan rumus Jaccard Coefficient dapat menghitung dan menentukan hasil persentasi kemiripan antar teks tugas akhir mahasiswa.

Suwanto Sanjaya dkk. (2015) melakukan penelitian tentang "Pengelompokan Dokumen Menggunakan Winnowing *Fingerprint* dengan Metode K-Nearest Neighbour". Dalam penelitian ini para peneliti menggunakan algoritma Winnowing untuk mengkelompokkan dokumen dengan metode K-NN. Pengukuran jarak kemiripan menggunakan *Euclidean distance*. Hasil pengujian akurasi terhadap 10 dokumen, persentase akurasi yang didapat adalah 80%. Hal ini disebabkan ada kelompok yang tidak revelan. Kelompok yang tidak relevan dipengaruhi oleh beberapa faktor seperti: nilai k-gram, nilai k tetangga terdekat, dan panjang dokumen.

M. Nishom (2019) melakukan penelitian tentang "Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square". Pada penelitian ini, beliau membahas perbandingan akurasi Euclidean, Minkowski dan Manhattan apabila diterapkan pada algoritma K-Means Clustering. Hasil penelitan ini menyimpulkan bahwa perbandingan akurasi metode pengukuran jarak (euclidean, manhattan, dan minkowski) untuk pelabelan klaster status disparitas kebutuhan Guru telah dilakukan dan memberikan nilai atau tingkat akurasi yang tinggi, yaitu 84.47% (untuk metode euclidean distance), 83.85% (untuk metode manhattan distance), dan

83.85% (untuk metode minkowski). Metode euclidean merupakan metode terbaik untuk diterapkan dalam algoritma KMeans Clustering.

Ali Shirkhorshidi (2015) melakukan penelitian tentang "A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data". Pada penelitian ini membahas tentang perbandingan pengukuran kesamaan dan ketidaksamaan pada Clustering data berkelanjutan. Pengukuran yang dibandingkan antara lain Euclidean Distance, Weighted Euclidead, Mahalanobis, Cosine, Manhattan, Pearson Correlation dan lainnya. Evaluasi hasil pengukurannya menggunakan Rand Index (RI). Indeks Rand sering digunakan dalam mengukur kualitas pengelompokan. Ini adalah ukuran kemiripan antara dua set objek: pertama adalah set yang dihasilkan oleh proses clustering dan yang lainnya ditentukan oleh kriteria eksternal. Tujuan dari penelitian ini adalah untuk mengklarifikasi ukuran kemiripan mana yang lebih tepat untuk dimensi rendah dan mana yang lebih baik untuk dataset dimensi tinggi. Berdasarkan hasil dalam penelitian ini, secara umum, korelasi Pearson tidak direkomendasikan untuk dataset dimensi rendah. Itu juga tidak kompatibel dengan algoritma berbasis centroid. Namun, ukuran ini sebagian besar direkomendasikan untuk dataset dimensi tinggi dan oleh menggunakan pendekatan.

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 Winnowung dengan membandingkan delapan metode yaitu Cosine, Jaccard, Dice, Euclidean, Manhattan, Minkowski, Mahalobis dan Weighted 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 diterjemahkan begitu saja. Dengan dilakukan penelitian ini penulis berharap kedepannya dapat meminimalisir tindakan plasiarisme dan tingkat keaslian karya akan meningkat juga dapat dipertanggungjawabkan.