BAB II

TINJAUAN PUSTAKA

II.1 Sistem Deteksi Plagiarisme Berbasis Teks

Beberapa penelitian telah mengusulkan metode berbeda untuk menangani plagiarisme. Diantaranya yaitu:

(Widjaja & Gunawan, 2020) Deteksi plagiarisme pada kode bahasa pemrograman Java menggunakan model pembelajaran mesin yang berbeda di XGBoost. Penelitian ini mengimplementasikan fungsi biner menggunakan tiga algoritma utama yaitu Levenshtein distance, Greedy string tiling dan bigram. Dataset yang digunakan berasal dari file penyerahan kode bahasa Java, yang selanjutnya diolah menjadi pasangan plagiat/non plagiat. Materi dibagi menjadi training set dan test set dengan komposisi 80% : 20% (toleransi ± 2%). Berbagi data memastikan pengidentifikasi memiliki panjang yang berbeda untuk memberikan informasi yang lebih umum dan mengurangi bias. Model XGBoost dengan kombinasi fitur yang diusulkan pada penelitian ini memiliki hasil pengujian yang lebih baik dibandingkan penelitian sebelumnya yaitu presisi 99%, akurasi 98%, recall 100% dan skor f1 99%.

(Siswanto & Ceng Giap, 2020) Penelitian dengan judul "Implementasi Algoritma Rabin-Karp Dan Cosine Similarity Untuk Pendeteksi Plagiarisme Pada Dokumen" penelitian ini menggunakan algoritma Rabin-Karp untuk mendeteksi kemiripan antara dokumen-dokumen dan melakukan pencocokan substring antara

teks dokumen yang dibandingkan. Dengan menggunakan algoritma ini, aplikasi dapat mengidentifikasi bagian-bagian yang serupa antara dokumen-dokumen tersebut. Kemudian mengunakan metode Cosine Similarity untuk menghasilkan nilai similarity antara dokumen-dokumen. Metode ini menghitung kemiripan vektor antara dokumen-dokumen berdasarkan kata-kata yang digunakan. Semakin tinggi nilai similarity, semakin mirip dokumen-dokumen tersebut.

(Utomo et al., 2020) Melakukan Deteksi Plagiat Tugas Akhir dengan Metode Jaccard Similarity. Metode yang digunakan untuk menghitung kesamaan adalah metode Jaccard Similarity atau Jaccard Coefficien. Jaccard Similarity adalah salah satu metode yang dipakai untuk menghitung Similarity antara dua objek (items), pada penelitian ini bertujuan mendesain sistem rekomendasi dengan metode Jaccard’s Coefficient, dan menguji kinerja temu kembali menggunakan recall dan precision, penerapan metode jaccard's Coefficient dalam penelitian tersebut ialah setiap dokumen harus dikorelasikan dengan subyek dengan relasi many to many, artinya satu subyek bisa memiliki beberapa dokumen, sebaliknya satu dokumen bisa juga memiliki beberapa subyek.

Penelitian selanjutnya yang dilakukan (Sipayung, 2021) dengan judul Identifikasi Tingkat Kemiripan Dokumen Teks Menggunakan Fungsi Hash Pada Algoritma Winnowing dan Pattern Recognation Pada Algoritma Rafcliff/Obershelp. Metode yang digunakan dalam pendeteksian plagiarisme pada penelitian yang dilakukan dibagi menjadi tiga yaitu perbangingan teks lengkap, dokument fingerprint dan kesamaan kata kunci. Pada algoritma Winnowing termasuk pada metode fingerprint dan sedangkan pada algoritma

Ratcliff/Obershelp termasuk keadalam metode perbandingan teks lengkap dan kesamaan kata kunci. data yang digunakan 10 file dokumen dengan ekstensi docx. kesimpulan hasil yang di peroleh pada penelitian yang dilakukan menggunakan kedua algoritma winnowing dan Ratcliff/Obershelp berjalan dengan baik, dari 10 dokumen yang dilakukan pengujian diperolah tingkat kemiripan pada dokumen 1 dan 2 sebesar 65.68%, pada dokumen 3 dan 4 sebesar 37.67%. pada dokumen 5 dan 6 sebesar 36.76, pada dokumen 7 dan 8 sebesar 36.67, pada dokumen 9 dan 10 sebesar 29.41%.

(Kambey & Dkk, 2020) Melakukan Penerapan Clustering pada Aplikasi Pendeteksi Kemiripan Dokumen Teks Bahasa Indonesia. Tujuan penelitian yang dilakukan mendeteksi kemiripan isi dokumen teks dengan memisahkan tiap kalimatnya menjadi data tersendiri dan mengelompokan tiap data kalimat yang ada pada isi dokumen teks untuk mempercepat waktu komputasi. Penerapan clustering pada pendeteksi tingkat kemiripan dokumen teks menggunakan algoritma *Density Based Spatial Clustering of Application with Noise* (DBSCAN). Hasil penelitian yang dilakukan dengan perbandingan akurasi tanpa penerapan clustering & penerapan clustering dengan nilai tertinggi mencapai 62%. Sedangkan saat tidak menerapkan clustering data yang dibandingkan secara keseluruhan sehingga mendapat nilai presentase akurasi tinggi, yaitu mencapai 100%. Kesimpulan dari penelitian pendeteksi tingkat kemiripan mendapat hasil yang tidak tepat atau kurang akurasi ketika menerapkan metode clustering dibandingkan saat tidak menerapkan metode tersebut.

(Efriyanto & Hayaty, 2022) Penelitian dengan judul: *Jaro Winkler Algorithm For Measuring Similarity Online News*. Tujuan dari penelitian ini adalah untuk mendeteksi plagiarisme dalam media berita online menggunakan algoritma Jaro Winkler. Penelitian ini mengukur tingkat kesamaan antara data berita online dengan data yang telah dikumpulkan oleh peneliti lain menggunakan metode yang berbeda. Penggunaan algoritma Jaro Winkler diusulkan karena penelitian sebelumnya menunjukkan bahwa algoritma ini menghasilkan tingkat kesamaan yang lebih tinggi dari pada metode lain seperti *Latent Semantic Analysis*. Hasil kesimpulan penelitian menunjukkan rata-rata tingkat kesamaan berita online adalah 74,49% dengan menggunakan 55 data sebagai sampel. Dari 55 data tersebut, 43 data memiliki tingkat plagiarisme tinggi dan 12 data memiliki tingkat plagiarisme sedang.

(Agustian & Sucipto, 2020) Penelitian dengan judul: Pengambilan sumber dalam deteksi plagiarisme berdasarkan sidik jari dua kata menggunakan model ruang vektor. Tujuan dari penelitian ini adalah untuk membantu mengidentifikasi apakah suatu dokumen mengandung plagiarisme dengan mencari beberapa sumber dokumen yang disusun dalam suatu korpus yang sistematis. Ketika calon sumber ditemukan, carilah potongan teks yang dijiplak. Penelitian ini menggunakan sidik jari dan pencarian dokumen asli dari dokumen plagiat. Hasilnya sistem bekerja dengan baik dalam pengumpulan dokumen ketika teks tidak diturunkan, menggunakan kata 3 gram dan memilih 5 karakter paling sering dari dokumen yang dicurigai sebagai pertanyaan, terlihat bahwa proporsi teks yang terdeteksi sebagai plagiarisme cukup besar. tinggi. tinggi yaitu 36,46%. Berdasarkan hasil yang

diperoleh pada percobaan penelitian dapat disimpulkan bahwa penggunaan sidik jari pada dokumen panjang dapat membantu mendeteksi plagiarisme dan mengurangi proses pencocokan kalimat demi kalimat dan kata demi kata dengan hasil yang memuaskan.

(Ariyanti et al., 2021)Melakukan Perancangan aplikasi pendeteksi plagiarisme teks menggunakan metode Rational Unified Process pada aplikasi pembelajaran online. Plagiarisme dapat dideteksi pada dokumen teks dengan menggunakan sistem deteksi komputer. Permasalahan yang diangkat pada penelitian ini adalah bagaimana merancang sebuah aplikasi e-learning yang mampu mendeteksi plagiarisme pada suatu teks dokumen. Metode yang digunakan adalah Rational Unified Process (RUP), tahapan metode ini adalah: start,develop,build,transition. Data yang digunakan adalah data sekunder yang berasal dari data buku, internet dan data primer yang dikumpulkan langsung dari pengguna dan pemangku kepentingan pengguna. Algoritma yang digunakan untuk membangun model analisis ini adalah algoritma n-gram yang menggambarkan secara jelas alur pengolahan kata dokumen.

(Sunardi et al., 2018) Makalah penelitian berjudul: Penerapan Deteksi Plagiarisme Menggunakan Metode Kemiripan N-Gram dan Jaccard pada Algoritma Menampi. Permasalahan dalam penelitian ini adalah mahasiswa atau peneliti menggunakan karya orang lain tanpa memberikan atribusi yang semestinya. Tujuan dari penelitian ini adalah menggunakan metode kemiripan n-gram dan Jaccard dengan algoritma winnowing untuk mendeteksi plagiarisme pada dokumen akademik. Dengan menggabungkan metode n-gram, kesamaan Jaccard dan

algoritma winnowing, plagiarisme pada dokumen akademik dapat dideteksi dan diatasi dengan membandingkan karakter alfanumerik dan menghitung kemiripan antar dokumen yang diteliti. Hasilnya, pendeteksi plagiarisme mampu menghasilkan 100% persen dokumen atau sampel dengan nilai 3 n-gram dan 5 w-gram menggunakan kesamaan Jaccard. Dibandingkan dengan metode k-gram yang digunakan, diperoleh kemiripan sebesar 83%. Dari hasil penelitian disimpulkan bahwa pendeteksi plagiarisme dengan metode n-gram, kesamaan Jaccard dan algoritma winnowing cukup baik untuk membandingkan kesamaan dua dokumen dan meminimalisir tindakan plagiarisme.

(AL-Jibory, Farah K, 2021) Penelitian dengan judul: Hybrid System for Plagiarism Detection on A Scientific Paper. Tujuan penelitian adalah mendeteksi kejadian plagiarisme menggunakan sistem deteksi plagiarisme. Sistem ini menggunakan teknik pemrosesan bahasa alami (*natural language processing*/NLP) dan pembelajaran mesin (*machine learning*/ML), serta strategi deteksi plagiarisme eksternal berbasis penambangan teks dan analisis kesamaan. Metode yang diusulkan menggunakan kombinasi dari kesamaan Jaccard dan kosinus untuk mendeteksi plagiarisme. Kesimpulan dari penelitian tingkat akurasi sebesar 0,96, recall sebesar 0,86, F-measure sebesar 0,86, dan skor PlagDet sebesar 0,865. Hasil ini menunjukkan bahwa sistem yang diusulkan efektif dalam mendeteksi plagiarisme dalam publikasi ilmiah.

(Nugroho et al., 2021) Penelitian dengan judul: Penerapan Algoritma Cosine Similarity untuk Deteksi Kesamaan Konten pada Sistem Informasi Penelitian dan Pengabdian Kepada Masyarakat. permasalahan pada penelitian yang dilakukan

tidak adanya sistem yang digunakan untuk mengelola kegiatan penelitian dan pengabdian kepada masyarakat serta data terkait catatan kegiatan penelitian dan pengabdian kepada masyarakat yang telah dilakukan oleh 2.613 dosen. Tujuan dari penelitian ini adalah mengembangkan sistem informasi yang dapat mengelola kegiatan penelitian dan pengabdian kepada masyarakat serta mendeteksi kesamaan konten menggunakan algoritma Cosine Similarity. Data yang digunakan dalam penelitian ini adalah catatan kegiatan penelitian dan pengabdian kepada masyarakat. Metode pengembangan sistem yang digunakan adalah metode waterfall, yang merupakan metode pengembangan perangkat lunak yang berurutan dan berdasarkan fase. Hasil pengembangan yang dilakukan, sistem yang diusulkan terbukti mampu mengelola kegiatan penelitian, namun pada informasi yang diberikan dikesimpulan tidak memberikan tingkat akurasi keberhasilan sistem dalam mendeteksi kesamaan konten menggunakan algoritma Cosine Similarity.

(Yuslena et al., 2021) Melakukan Deteksi Plagiarisme menggunakan Algoritma Levenshtein Distance. Permasalahan pada penelitian ini plagiarisme sering terjadi pada dokumen di lingkungan akademik, yang bertujuan mengembangkan sistem deteksi plagiarisme untuk mendukung penjagaan orisinalitas akademik. Data yang digunakan adalah dokumen-dokumen tesis mahasiswa. Hasil kesimpulan dari penelitian yang dilakukan, rata-rata waktu proses sistem tanpa preprocessing adalah 6,283ms. sedangkan rata-rata waktu proses sistem dengan preprocessing adalah 4,920ms. Dengan menggunakan algoritma Levenshtein Distance dan preprocessing, sistem deteksi plagiarisme dapat berjalan lebih cepat dan efektif dalam mendeteksi plagiarisme pada dokumen.

II.2 Pengertian Text

Text ialah ungkapan bahasa yang menurut isi, sintaksis, dan pragmatik merupakan satu kesatuan. Istilah teks sebenarnya berasal dari kata text yang berarti ‘tenunan’. Teks dalam filologi diartikan sebagai ‘tenunan kata-kata’, yakni serangkaian kata-kata yang berinteraksi membentuk satu kesatuan makna yang utuh. Baried mengungkapkan bahwa teks adalah kandungan atau muatan naskah, sesuatu yang abstrak hanya dapat dibayangkan saja. Teks terdiri atas isi, yaitu ide-ide atau amanat yang hendak disampaikan pengarang kepada pembaca. Dan bentuk, yaitu cerita dalam teks yang dapat dibaca dan dipelajari menurut berbagai pendekatan melalui alur, perwatakan, gaya bahasa, dan sebagainya(Purba & Situmorang, 2017).

Dari pendapat diatas dapat disimpulkan bahwa text adalah simbol atau karakter yang membentuk bahasa tertulis dan dapat digunakan untuk berbagai tujuan komunikasi, text juga dapat merujuk pada tipe data string yang digunakan untuk menyimpan dan memanipulasi text dalam konteks komputasi.

(Pratiwi et al., 2022) Klasifikasi teks adalah teknik dalam teks mining yang bertujuan untuk menempatkan teks ke dalam kategori yang sesuai dengan karakteristiknya, menggunakan aturan-aturan tertentu. Terdapat dua metode dasar dalam klasifikasi teks yaitu Unsupervised Text Classification (klasifikasi teks tanpa aturan sebelumnya) dan Supervised Text Classification (klasifikasi teks berdasarkan aturan yang sudah ditentukan melalui proses pembelajaran). Dapat disimpulkan pengertian teks dari kalimat tersebut adalah bahwa teks merujuk pada data berupa teks yang akan dikelompokkan atau diklasifikasikan menggunakan

aturan atau pola dalam teks mining, dengan metode klasifikasi teks yang dapat berupa Unsupervised Text Classification atau Supervised Text Classification.

II.3 Plagiarisme

(Ginting et al., 2022) Plagiarisme atau pelanggaran hak cipta adalah tindakan yang dilakukan dengan sengaja untuk mendapatkan manfaat pribadi, seperti mendapatkan nilai atau karya ilmiah, tanpa mencantumkan sumber asal dari karya yang digunakan. Beberapa tindakan yang dapat dianggap sebagai plagiarisme meliputi:

1. Salin dan tempel secara langsung tanpa melakukan modifikasi atau perubahan apapun pada kata-kata yang diambil.
2. Pelanggaran hak cipta yang melibatkan tindakan menyembunyikan bagian yang direplikasi. Hal ini dapat dilakukan melalui beberapa strategi seperti shake and paste (menyisipkan secara asal), pencurian sastra luas (mengambil sebagian besar konten dari sumber lain), pemalsuan kontraktif (memalsukan sumber referensi), dan plagiat mosaic (menggabungkan beberapa sumber menjadi satu).
3. Penyamaran teknis yang dilakukan untuk menyembunyikan konten yang dicuri dari alat pendeteksi otomatis. Misalnya, dengan memperbarui alfabet pada simbol bahasa asing agar tidak terdeteksi.
4. Parafrase yang tidak tepat, yaitu melakukan parafrase atau perubahan kata-kata dari ide-ide orang lain tanpa menyebutkan sumbernya dengan tepat, dengan tujuan menyembunyikan asal aslinya.

5. Plagiarisme yang diterjemahkan, yaitu mengonversi karya dari satu bahasa ke bahasa asing lain tanpa mencantumkan sumber asal.
6. Ide plagiarisme, yaitu memanfaatkan gagasan atau pikiran orang lain tanpa mencantumkan sumbernya.

Dengan demikian, kesimpulan pengertian plagiarisme adalah tindakan yang disengaja untuk menggunakan karya orang lain tanpa mencantumkan sumbernya, baik dengan cara menyalin secara langsung maupun dengan menggunakan strategi lainnya yang mencoba menyembunyikan asal aslinya.

Plagiarisme merupakan pelanggaran berat terhadap principles etika ilmiah yang dengannya sebuah artikel mewakili kontrak implisit antara penulis karya itu dan pembacanya. Dengan demikian, pembaca berasumsi bahwa penulis adalah satu-satunya pencetus karya tulis dan bahwa setiap materi, teks, data atau ide yang dipinjam dari orang lain secara jelas diidentifikasi seperti itu oleh pertemuan ilmiah yang mapan, seperti catatan kaki, teks blok-indentasi, dan tanda kutip 'yang mengungkapkan asal materi melalui kutipan langsung, parafrase atau meringkas(*Elsevier Enhanced Reader,* n.d.,2021). Ambil ide atau kata-kata orang lain dan sebut saja itu milik Anda. Plagiarisme adalah salah satu bentuk pencurian intelektual. Plagiarisme dapat terjadi dalam berbagai bentuk, mulai dari penipuan yang disengaja hingga penyalinan sumber secara sengaja tanpa pengakuan (Apridiansyah et al., 2022).

Ada tiga metode untuk mendeteksi plagiarisme (Apridiansyah et al., 2022) yaitu:

1. Perbandingan teks lengkap

   Metode ini digunakan ketika membandingkan keseluruhan isi suatu dokumen. Namun pendekatan ini memakan waktu lama namun cukup efektif

2. Dokumen sidik jari

   Sidik jari dokumen merupakan suatu metode untuk menentukan keakuratan kesamaan dokumen. Prinsip kerja metode dokumen sidik jari ini adalah penggunaan teknik hashing. Teknik hashing adalah fungsi yang mengubah string apa pun menjadi angka. Algoritma yang digunakan dalam metode ini adalah algoritma Winnowing, algoritma Manber dan algoritma Rabin-Krap.

3. Kesamaan kata kunci

   Prinsip dari metode kata kunci adalah mencari kata kunci pada dokumen kemudian membandingkannya dengan kata kunci pada dokumen lain

II.4 *Text Processing*

Pada penelitian (Liken Anggoro et al., 2023) Preprocessing teks adalah proses yang dilakukan sebelum analisis teks atau penambangan teks. Tujuan dari text preprocessing adalah untuk mengolah teks agar dapat dijadikan masukan untuk analisis lebih lanjut. Proses Preprocess Text melibatkan beberapa tahapan, antara lain:

1. Tokenisasi: Teks dipisahkan menjadi unit yang lebih kecil yang disebut token. Token bisa berupa kata, frasa, atau karakter tertentu.
2. Transformasi: Melibatkan konversi teks ke dalam format yang lebih sesuai untuk analisis. Misalnya, mengubah huruf menjadi huruf kecil semua, menghapus tanda baca atau karakter khusus, dan lain sebagainya.
3. Normalisasi: Melakukan normalisasi pada teks, seperti menggantikan kata-kata yang memiliki variasi penulisan yang sama (misalnya, "computer" dan "cpt") menjadi bentuk yang seragam.
4. Filtering: Melakukan filtering untuk menghapus kata-kata atau token yang tidak relevan atau tidak diperlukan dalam analisis. Misalnya, kata-kata umum seperti "dan", "atau", atau stop words.

(Dimas Diandra Audiansyah, Dian Eka Ratnawati, 2022)Pengertian text preprocessing adalah proses normalisasi atau mengubah teks yang tidak terstruktur menjadi data terstruktur, proses ini dilakukan bertujuan untuk mendapatkan data latih dengan kualitas yang baik dan mengekstrak data sesuai yang dibutuhkan, sehingga mampu menyederhanakan pengolahan data. Dalam prosesnya dilakukan analisis semantik (makna yang benar) dan sintaksis (susunan yang benar) terhadap teks. Tujuan dari pengolah kata adalah untuk mempersiapkan teks untuk informasi yang akan diproses di masa depan.

Proses pengolahan kata meliputi case folding, tokenisasi, pemfilteran, dan stemming (Normah et al., 2022).

1. *Case Folding* adalah proses mengubah seluruh huruf dalam suatu dokumen menjadi huruf kecil. Digunakan untuk memudahkan pencarian.

Dalam Case Folding huruf yang diterima adalah dari “a” sampai “z”, karakter selain huruf-huruf ini dihilangkan dan diperlakukan sebagai pembatas.

2. Tokenisasi berfungsi untuk mengekstrak kata-kata pada data masukan.
3. Filtrasi Pemfilteran adalah proses mengekstraksi kata-kata penting dari sebuah string. Penyaringan dilakukan dengan menghilangkan kata-kata yang terdaftar pada stop word/stop list. Kata-kata berhenti adalah kata-kata yang sering muncul dalam sebuah teks dan dianggap tidak mempunyai arti penting.
4. Stemming adalah proses pencarian kata dasar hasil filter. Mencari root atau kata kunci dapat melemahkan hasil indeks tanpa kehilangan makna.

Penelitian(Saeed & Taqa, 2022) bahwa Pre-process the text merujuk pada serangkaian teknik yang diterapkan pada dokumen sumber dan dokumen yang mencurigakan sebelum diproses lebih lanjut. Teknik-teknik yang digunakan dalam pre-process teks meliputi tokenisasi, mengubah huruf menjadi huruf kecil (lowercasing), penghapusan tanda baca dan stop word, segmentasi kalimat, serta perbaikan lainnya menggunakan teknik NLP (Natural Language Processing) yang cerdas. Tujuan dari pre-process the text adalah untuk mempersiapkan korpus atau data teks sebelum menerapkan pendekatan word embedding. Hal ini disebabkan oleh karakteristik data yang tidak sesuai untuk diproses langsung. Dalam tahap pre-process beberapa langkah penting yang perlu dilakukan termasuk membersihkan data dengan menghilangkan tanda baca, stop word, dan melakukan stemming (penghapusan infleksi kata menjadi kata dasarnya).

Dapat disimpulkan Pre-process the text mengacu pada serangkaian teknik yang digunakan untuk membersihkan dan mempersiapkan data teks sebelum menerapkan pendekatan word embedding atau pendekatan lainnya dalam analisis atau pemrosesan teks.

II.5 *Hashing*

(Alifah Akbar Pertiwi1, 2023)Hash adalah kode alfanumerik (terdiri dari huruf dan angka) dengan panjang tetap yang dipakai untuk mewakili suatu data, kata, atau pesan. Hashing dilakukan dengan cara mengubah variable sized input menggunakan rumus matematika (hash function) hingga menghasilkan fixed sized output. Hash dapat digunakan untuk beberapa fungsi salah satunya untuk mengubah tulisan asli (seperti password) untuk diubah dengan metode hashing agar password asli berubah menjadi kode acak.

Hashing adalah proses mengubah data kata yang memiliki variabel *string* menjadi nilai sesuai dengan ASCII (*american standard code for information interchange*) dalam bentuk *array*. Proses *hashing* merupakan proses untuk melakukan konversi karakter dengan variabel string menjadi sebuah nilai dengan panjang tetap serta dapat mewakili karakter dari *string* asli (Tsaqif, 2021). Adapun rumus yang digunakan untuk melakukan proses *hashing* yaitu:

$$H\ (C2 \ldots Cn) = C1 * b^{(n-1)} + C2 * b^{(n-2)} + \ldots + C(n-1) * b + Cn$$

C : nilai ASCII karakter

B : basis bilangan prima (ditentukan oleh *user*)

N : banyaknya karakter (ke – n)

Contoh perhitungan nilai hash pada kata “saya” menggunakan basis bilangan prima (2):

$$H(saya) = ascii(s) * 2^{(4-1)} + ascii(a) * 2^{(4-2)} + ascii(y)\ 2^{(4-3)} + ascii(a) * 2^{0}$$

$$= 115 * 2^{3} + 97 * 2^{2} + 121 * 2^{1} + 97 * 2^{0}$$

$$= 1647$$

(Rahim et al., 2022)Hashing merupakan metode yang digunakan untuk mengubah sebuah string menjadi karakter-karakter yang merepresentasikan string tersebut dengan menggunakan operasi aritmatika. Dalam konteks bahasa, hashing berasal dari kata "hash" yang berarti memenggal dan menggabungkan. Metode hashing digunakan untuk penyimpanan data dalam sebuah larik (array) agar operasi penyimpanan, pencarian, penambahan, dan penghapusan data dapat dilakukan dengan cepat. Dasar dari proses hashing adalah dengan menghitung posisi record yang dicari dalam array berdasarkan nilai hashnya.

II.6 Metode N-gram

N-gram adalah metode mengambil serangkaian substring dari string n (urutan karakter dengan panjang n). Dalam deteksi plagiarisme, n-gram digunakan untuk mengekstraksi karakter atau string sepanjang n dari suatu kata atau dokumen secara terus menerus (kontinuitas) hingga berpindah ke posisi halaman tertentu atau akhir kata atau dokumen. N-gram dalam pendeteksian plagiarisme sangat mempengaruhi keakuratan atau tingkat kemiripan dokumen yang dibandingkan, sehingga teknik n-gram dipadukan dengan pendekatan statistik untuk mencapai kemiripan dokumen, seperti kesamaan sederhana, kesamaan kosinus, kemiripan

jaccard, dan dice coeficient(Sunardi et al., 2018). Contoh penggunaan dari N-grams, fourgram dari Mengimplementasi yaitu “Meng, engi, ngim, gimp, impl, mple, plem, leme, emen, ment, enta, ntas, tasi”.

II.7 Algoritma Winnowing

Algoritma Winnowing Algoritma Winnowing merupakan algoritma yang dapat digunakan untuk mengukur tingkat kemiripan kata (fingerprinting suatu dokumen) untuk mendeteksi plagiarisme. Rolling hash adalah algoritma yang digunakan untuk menemukan hash selama hashing. Nilai hash adalah nilai numerik yang terdiri dari perhitungan ASCII untuk setiap karakter (Hizkia et al., 2022). Nilai hash ditentukan untuk setiap k-gram dengan algoritma Winnowing, yang dalam hal ini menggunakan fungsi rolling hash. Hash yang dihasilkan digunakan untuk membuat jendela. Di setiap interval, nilai hash terkecil dipilih. Nilai hash paling kanan dipilih jika beberapa nilai hash memiliki nilai terkecil yang sama. Hash yang dipilih kemudian disimpan sebagai sidik jari dokumen (Saputra et al., 2022).

(Hidayat et al., 2020) Langkah preprocessing teks pada algoritma Winnowing adalah sebagai berikut:

1. Memotong substring

   Metode untuk kliping teks ada yaitu k-gram dengan teknik kliping berbasis karakter dan n-gram dengan implementasi kliping berbasis kata.

2. Rooling Hash

Berdasarkan pemotongan substring yang dibuat, semua substring yang terbentuk diproses selama hashing dengan teknik Rolling Hash. Berikutnya adalah Persamaan 1 (Rolling Hash).

$$H(cc1\ldots cccc) = cc1 * bb(cc-1) + cc2 * bc2 * bc2 * bb)\ bbcc + cccc \quad (1)$$

Keterangan:

c: nilai ascii karakter

b: basis (bilangan prima)

k: banyak karakter

3. Window

Setelah membuat hash untuk setiap rantai parsial, kelompokkan hash ke dalam jendela sehingga jumlah anggota di setiap jendela ditentukan oleh nilai jendela.

4. Fingerprint

Selanjutnya mencari nilai sidik jari yaitu berdasarkan nilai hash terkecil dari setiap jendela. Jika nilai terkecilnya sama, ambil nilai paling kanan.

II.8 Cosine Similarity

(Adnan et al., 2022) Kesamaan kosinus secara sederhana diartikan sebagai cara membandingkan persamaan antara dua dokumen atau teks. Kesamaan kosinus adalah metrik yang digunakan untuk mengukur kesamaan antara dua vektor. Secara khusus, kesamaan kosinus mengukur kesamaan arah atau orientasi vektor, dengan

mengabaikan perbedaan ukuran atau skalanya. Ide di balik metode ini adalah untuk menormalkan panjang vektor data dengan membandingkan N-gram paralel dari dua perbandingan. Kesamaan kosinus digunakan dalam ruang positif yang hasilnya dibatasi antara 0 dan 1. Jika skornya 0 maka dokumen dikatakan berbeda, dan sebaliknya jika skor kesamaan kosinus 1 maka kedua dokumen dikatakan berbeda. serupa. Layanan informasi mengasumsikan bahwa setiap kata/istilah memiliki dimensi yang berbeda dan dokumen diwakili oleh vektor, dimana nilai setiap dimensi sesuai dengan berapa banyak istilah yang muncul dalam dokumen. Berikut rumus cosine-similarity :

$$\cos\theta = sim(x, y) = \frac{x.y}{||x||||y||}, \quad (1)$$

Keterangan:

A : Vector

B : Vector

Ai = bobot term I dalam blok Ai

Bi = bobot term I dalam blok Bi

I = jumlah term dalam kalimat

II.9 Jaccard Similarity

Jaccard similarity atau jaccard Coefficient merupakan algoritma yang fungsinya untuk membandingkan dua sampel yaitu dokumen yang satu dengan dokumen yang lainnya berdasarkan kata yang dimilikinya. Jaccard similarity

biasanya digunakan untuk membandingkan dokumen dan menghitung nilai kemiripan (similarity) dari dua buah dokumen atau objek(Sunardi et al., 2018). Jaccard similarity dapat dirumuskan sebagai berikut:

$$similarity = \left(\frac{X \cap Y}{X \cup Y}\right) \text{ x } 100\ \% \quad (2)$$

II.10 Penelitian Sebelumnya

Tabel II.1 mencantumkan beberapa penelitian atau jurnal yang menjadi referensi dalam penelitian ini.

Tabel II.1 Ringkasan penelitian deteksi plagiarisme

No.	Penelitian/ Tahun	Judul Penelitian	Persamaan	Perbedaan
1.	(Rahmatulloh et al., 2019)	Perbandingan antara Efek Porter Stemmer dan Nazief-Adriani pada Kinerja Algoritma Winnowing untuk Mengukur Plagiarisme	Sama-sama menggunakan algoritma winnowing dalam mengukur plagiarisme	Tidak menggunakan Stemmer Porter
2.	(V.A.R.Barao et al., 2022)	Deteksi Plagiarisme Teks Dan Gambar Berbasis Model Kecerdasan Buatan Berbasis Algoritma K-Mean Clustering	Deteksi Plagiarisme berbasis Teks	Tidak menggunakan teknik SimHash dan N-gram
3.	(Faisal et al., 2020)	Deteksi Plagiarisme Menggunakan Algoritma Manber dan Winnowing	Sama-sama menggunakan algoritma Winnowing	Tidak menggunakan Algoritma Manber
4.	(Duan et al., 2017)	Algoritma Deteksi Plagiarisme berdasarkan Extended Winnowing	Sama-sama menggunakan algoritma winnowing	Tidak menggunakan teknik SimHash dan N-gram
5.	(n.d.,2019)	Sistem deteksi plagiarisme untuk dokumen berbasis teks Malayalam dengan	Sistem deteksi plagiarisme untuk dokumen berbasis teks.	Tidak menggunakan teknik SCAM dan

		salinan lengkap dan sebagian		PPCHECKER untuk mendeteksi kalimat yang disalin sebagian atau seluruhnya.
6.	(*Reader*, n.d.,2022)	Deteksi Duplikat pada Dokument dengan menggunakan algoritma sidik jari Simhash	Sama – sama menggunakan teknik Simhash	Tidak menggunakn teknik k-gram untuk membandingkan beberapa dokumen
7.	(*Elsevier Enhanced Reader* n.d.,2022)	Deteksi Plagiarisme menggunakan kesamaan dokumen berdasarkan representasi terdistribusi	Tidak menggunakan algoritma winnowing dengan teknik n-gram	Tidak menggunakan metode deteksi dokument plagiarisme lcs, llcs dan wllc
8.	(Kuruvila et al., 2017)	Sistem Deteksi Plagiarsisme Pemrosesan Gambar Diagram Alir	Sama-sama membandingkan bentuk dari model akurasi data uji	Tidak menggunakan sistem deteksi berbasis diagram alir
9.	(Zouaoui & Rezeg, 2022)	Deteksi Plagiarisme Multi-Agents Indexing System	Tidak menggunakan pengindeksan semantik dan multi-agen.	Tidak menggunakan algoritma winnowing
10.	(Tresnawati et al., 2012)	Desain Sistem deteksi plagiarisme berbasis moodle	Tidak menggunakan software deteksi seperti SID, MOS Jplag, YAP	Sama-sama membandingkan representasi strukutur program string parse teks
11.	(Hart et al., 2023)	Pencegahan Plagiarisme di CS1 Melalui Data Keystroke	Tidak menggunakan dokumen berbasis teks	Tidak menggunakan plugin IDE software dan Keystroke
12.	(*Project* n.d.,2021)	Implementasi Algoritma Winnowing dalam mendeteksi Plagiarisme Judul dan Abstrak Tugas Akhir Mahasiswa	Sama-sama menggunakan Algoritma Winnowing	Tidak menggunakan Simhash

13.	(Widaningrum et al., 2020)	Evaluasi akurasi algoritma winnowing, rabin karp dan knuth morris pratt dalam aplikasi pendeteksi plagiarisme	Sama-sama menggunakan algoritma winnowing dan morris pratt	Tidak menggunakan rabin karp
14.	(Ahnaf et al., 2023)	Pendekatan deteksi plagiarisme monolingual ekstrinsik yang ditingkatkan dari teks Bengali	Sama-sama menggunakan algoritma pengukuran kesamaan teks yang berbeda yaitu, kesamaan kosinus dan kesamaan Jaccard	Tidak menggunaan algoritma algoritma Cosine similarity measure
15.	(Kurniati et al., 2019)	Web Scraping dan Winnowing Algoritma untuk Deteksi Plagiarisme Final Judul Proyek	Sama-sama menggunakan algoritma Winnowing	Tidak menggunakan web Scraping

II.11 Kontribusi Penelitian

Adapun kontribusi dalam penelitian ini adalah membandingkan sebagaimana tingkat similarity antara algoritma Cosine Similarity dan Jaccard Similarity untuk deteksi kesamaan text terkait dengan variasi nilai N-gram 3, 5 dan 7.