p-ISSN 2301–4156 | e-ISSN 2460–5719

Pengaruh *Synonym Recognition* dalam Deteksi Kemiripan Teks Menggunakan *Winnowing* dan *Cosine Similarity*

Santi Purwaningrum[1], Agus Susanto[2], Ari Kristiningsih[3]

[1,2] Jurusan Teknik Informatika dan Bisnis Politeknik Negeri Cilacap, Cilacap, 53212 INDONESIA (tel.: 0282-533329; fax: 0274-4321982, email: [1]santi.purwaningrum@pnc.ac.id, [2]agussusanto@pnc.ac.id)
[3] Jurusan Pengembangan Produk Agroindustri Politeknik Negeri Cilacap, Cilacap, 53212 INDONESIA (tel.: 0282-533329; fax: 0274-4321982, email: [3]ari.kristiningsih@pnc.ac.id)

[Diterima: 19 Januari 2023, Revisi: 4 Juli 2023]
Corresponding Author: Santi Purwaningrum

INTISARI — Plagiarisme adalah tindakan meniru dan mengutip bahkan menyalin atau mengakui hasil karya orang lain sebagai hasil karya diri sendiri. Tugas akhir merupakan salah satu syarat wajib mahasiswa untuk menyelesaikan pembelajaran pada perguruan tinggi. Tugas akhir harus disusun mahasiswa berdasarkan ide sendiri. Akan tetapi, banyak terjadi plagiarisme karena mudahnya melakukan kegiatan tersebut, yaitu hanya dengan menyalin teks gagasan orang lain kemudian ditempelkan dalam lembar kerja dan diakui bahwa gagasan tersebut adalah ide sendiri. Selain itu, mengganti beberapa kata dalam kalimat gagasan orang lain dengan gaya bahasa sendiri tanpa menuliskan sumber aslinya juga termasuk plagiarisme. Pengecekan tugas akhir yang masih manual juga menjadi masalah bagi koordinator tugas akhir, yang membutuhkan ketelitian tinggi dan waktu yang cukup banyak untuk mengecek plagiarime pada dokumen tugas akhir. Maka, deteksi plagiarisme sangat dibutuhkan untuk mencegah tindakan plagiarisme makin berkembang. Menyikapi hal tersebut, penelitian ini bermaksud mengembangkan sistem yang dapat mendeteksi persamaan antardokumen teks yang berfokus pada kata yang mengandung sinonim pada suatu kalimat. Salah satu algoritma yang digunakan adalah *synonym recognition*, yang berfungsi untuk mendeteksi kata yang mengandung sinonim, dengan proses membandingkan setiap kata dengan kata yang terdapat pada kamus. *Synonym recognition* dikombinasikan dengan metode *winnowing*, yang berfungsi untuk pembobotan teks berbasis *fingerprint*. Setelah diperoleh bobot dari masing-masing dokumen, tingkat kemiripan antardokumen dihitung dengan algoritma *cosine similarity*. Hasil rata-rata nilai kemiripan untuk deteksi judul dan abstrak dengan menambahkan *synonym recognition* meningkat sebesar 3,11% daripada tanpa menggunakan *synonym recognition* yang dikombinasikan dengan metode pembobotan *winnowing*. Hasil pengujian menunjukkan bahwa algoritma-algoritma yang digunakan akurat dengan pengujian akurasi dan *root mean squared error* (RMSE).

KATA KUNCI — *Synonym Recognition*, *Winnowing*, *Cosine Similarity*, Plagiarisme.

I. PENDAHULUAN

Tugas akhir adalah salah satu syarat wajib bagi mahasiswa untuk dapat menyelesaikan perkuliahan dalam jenjang diploma untuk mendapatkan gelar ahli madya dan sarjana terapan. Tugas akhir ditempuh setelah menyelesaikan magang industri, sehingga diharapkan mahasiswa mendapatkan gagasan yang dapat diangkat menjadi topik tugas akhir dari studi kasus di tempat magang tersebut. Studi kasus tidak diharuskan dilakukan di tempat magang industri. Mahasiswa diperbolehkan mencari tempat studi kasus yang lain dengan topik yang sesuai dengan kebutuhan tempat tersebut.

Bagi mahasiswa vokasi, implementasi sistem bukanlah masalah besar. Namun, terdapat kesulitan dalam penyusunan laporan tugas akhir, terutama dalam penulisan ide atau gagasan, agar mahasiswa tidak dianggap melakukan plagiarisme. Plagiarisme dapat terjadi dengan kesengajaan maupun tanpa disengaja. Karena kurangnya membaca atau kesulitan dalam mencari referensi, gagasan atau ide yang akan dijadikan topik tugas akhir bisa saja sama atau pernah dibuat sebelumnya [1], [2].

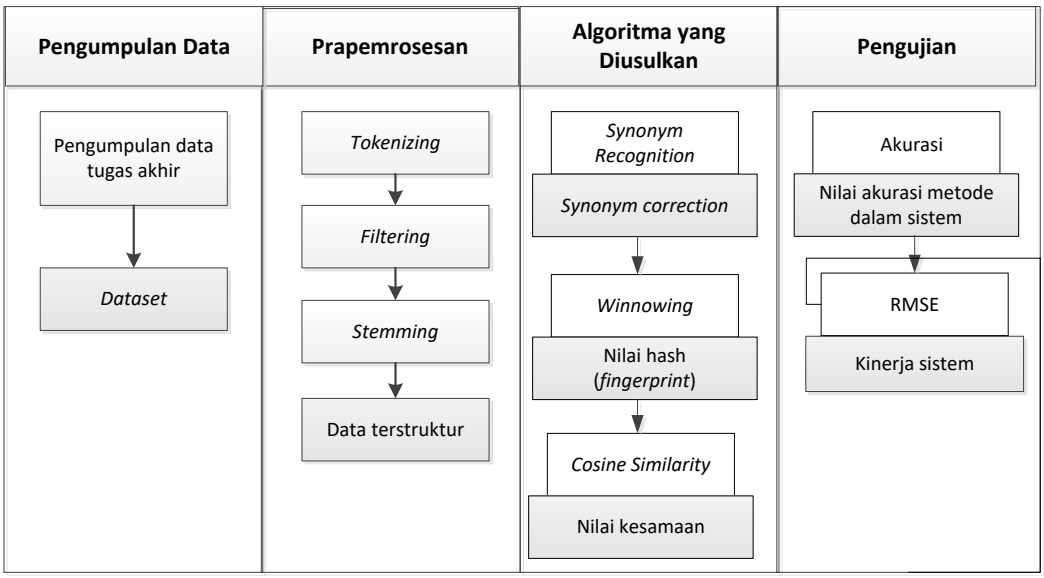
Proses mengunggah judul, abstrak, proposal, jurnal, poster, hingga laporan tugas akhir ke sistem dilakukan oleh koordinator tugas akhir. Namun, koordinator tugas akhir hanya mengunggah tanpa memperhatikan dengan pasti setiap judul, abstrak, hingga laporan tugas akhir mahasiswa, sehingga banyak laporan tugas akhir yang memiliki kemiripan antarangkatan, bahkan seangkatan, tidak terdeteksi. Di sisi lain, proses pemeriksaan oleh koordinator tugas akhir akan membutuhkan waktu yang cukup lama dan konsentrasi yang tinggi karena harus melakukan pemeriksaan satu per satu.

Selain permasalahan tersebut, budaya salin-tempel juga sudah berkembang, dengan hanya mengganti kata yang mengandung sinonim tanpa mengutip sumbernya. Deteksi dini kemiripan judul dan abstrak karya ilmiah sangat dibutuhkan untuk mengurangi tindakan plagiarisme yang berawal dari kurangnya kreativitas mahasiswa.

Banyak peneliti tertarik melakukan penelitian tentang kasus plagiarisme. Metode yang digunakan untuk pembobotan teks deteksi plagiarisme antara lain algoritma *winnowing*, algoritma Boyer-Moore, dan algoritma Rabin-Karp (atau disebut juga algoritma Karp-Rabin) [3], [4]. Salah satu penelitian menggunakan algoritma *winnowing* untuk mendeteksi tingkat kemiripan judul skripsi yang diajukan dengan judul skripsi yang telah ada sebelumnya [5]. Penelitian ini menyimpulkan bahwa dengan nilai *n-gram* = 3, *window* = 3, dan bilangan prima = 2, dihasilkan nilai kemiripan sebesar 73,86%, yang menunjukkan bahwa judul yang dibandingkan memiliki tingkat kemiripan yang tinggi. Dengan masukan *n-gram* = 7, *window* = 9, dan bilangan prima = 2, dihasilkan nilai kemiripan 19,82%, yang menunjukan tingkat plagiarisme ringan [5].

Sebuah penelitian menjelaskan gambaran tentang definisi plagiarisme, peralatan untuk deteksi plagiarisme, matriks perbandingan, metode *obfuscation*, kumpulan data yang digunakan untuk perbandingan, dan jenis algoritma [6]. Disimpulkan bahwa tiga algoritma yang sering digunakan dalam deteksi plagiarisme adalah *running Karp-Rabin greedy*

p-ISSN 2301–4156 | e-ISSN 2460–5719



Gambar 1. Tahapan penelitian.

string tiling (RKR-GST), algoritma *winnowing*, dan proses implementasi dalam tokenisasi [6].

Penelitian lainnya menggunakan algoritma Rabin-Karp pada proses pembobotan, yang dikombinasikan dengan pendekatan *synonym recognition*, untuk menemukan kata-kata yang sudah diubah ke bentuk kata yang mempunyai makna sama. Pada proses pembobotan menggunakan algoritma Rabin-Karb, terdapat langkah *rolling hash* yang harus memasukkan suatu nilai basis. Pada algoritma ini, tidak semua angka dapat digunakan dalam nilai basis. Penyebabnya adalah pada beberapa kasus, nilai basis yang salah akan mengakibatkan nilai *hash* yang dihasilkan sama dengan nilai *hash* lain yang mempunyai kata berbeda [7].

Sebuah penelitian yang bertujuan mengurangi bias pada sistem penilaian jawaban uraian menggunakan metode *cosine similarity* dengan pembobotan *term frequency-inverse document frequency* (TF-IDF) dan pencocokan kata dengan menambahkan regresi linier telah dilakukan [8]. Sistem penilaian lebih sering memberikan keluaran dengan nilai lebih tinggi daripada penilaian secara manual oleh dosen, yang artinya nilai tersebut mengalami bias. Oleh karena itu, digunakan regresi linier untuk mengurangi bias agar nilai jawaban uraian yang dihasilkan sistem tidak cenderung lebih maupun kurang dari nilai manual yang diberikan dosen. Pada proses regresi linier, nilai sistem mengacu pada kunci jawaban yang telah dibuat, yang disesuaikan dengan penilaian dosen, sehingga diharapkan dosen memberikan penilaian secara objektif [8].

Telah dilakukan juga penelitian yang menerapkan metode *synonym recognition* untuk mendeteksi kata yang mengandung sinonim dan *cosine similarity* untuk mencari nilai kemiripan. Proses pembobotan teks setiap dokumen yang dibandingkan menggunakan *term frequency* (TF) dan pengujian akurasi menggunakan *root mean squared error* (RMSE). Terdapat beberapa tahapan dalam pengujian RMSE, yaitu 20, 50, 100, dan 116 data. Nilai RMSE tertinggi yang diperoleh pada jawaban nomor 1 adalah pada jumlah siswa 20, yaitu 2,07, dan yang terendah pada jumlah siswa 50, yaitu 6,16. Pada jawaban nomor 2, RMSE tertinggi diperoleh pada jumlah mahasiswa 50, yaitu 8,94, dan terendah pada jumlah siswa 20, yaitu 8,00. Hasil pengujian akurasi mempunyai nilai *error* yang cukup tinggi karena kata pada jawaban siswa termasuk dalam kamus *stopword*, sehingga kata tersebut hilang pada proses *filtering*. Maka, dalam penelitian selanjutnya, diharapkan nilai kemiripan antara dua buah teks dapat ditingkatkan [9].

Penelitian-penelitian sebelumnya sudah banyak yang membahas tentang deteksi kemiripan teks dengan banyak metode yang digunakan untuk pembobotan suatu teks, seperti Rabin-Karp, *winnowing*, Smith-Waterman dan Manber, serta TF [10]-[12]. Kemudian, algoritma yang digunakan untuk mencari nilai kemiripan, antara lain Jaccard, *dice similarity*, *neural network*, dan *cosine similarity*, dinilai memiliki tingkat kemiripan lebih baik [13]. Referensi [14] mengkaji penggunaan kolaborasi metode *cosine similarity* dengan algoritma pembobotan yang lain, seperti *artificial neural network* (ANN) dan *support vector machine* (SVM), tetapi belum dengan kombinasi metode *winnowing*.

Tujuan utama dari penelitian ini adalah mendeteksi dini kemiripan judul dan abstrak pada tugas akhir mahasiswa dengan berfokus pada kata yang mengandung sinonim. Metode yang digunakan untuk menghitung tingkat kemiripan antardokumen adalah *cosine similarity*. Dalam metode *similarity*, biasanya proses pembobotan dilakukan menggunakan TF-IDF untuk mengetahui nilai bobot dari setiap dokumen. Akan tetapi, dalam penelitian ini pembobotan nilai dari setiap dokumen menggunakan algoritma *winnowing*, yaitu pembobotan berbasis *fingerprint*. Algoritma ini dikombinasikan dengan metode *synonym recognition* yang digunakan untuk mendeteksi kata-kata pada dokumen yang dibandingkan dengan kamus sinonim. Dengan kombinasi *sinonym recognition*, diharapkan tugas akhir dengan parafrasa yang berbasis sinonim dapat dideteksi. Kinerja setiap metode yang dikombinasikan diukur menggunakan parameter akurasi, sedangkan RMSE digunakan untuk mengetahui nilai *error* dari nilai yang direkomendasikan oleh sistem terhadap nilai yang sebenarnya [15].

II. METODOLOGI

Pada tahap ini, dijelaskan pengaruh algoritma *synonym recognition* yang dikombinasikan dengan algoritma *winnowing* dan *cosine similarity* untuk mendeteksi kemiripan judul dan abstrak tugas akhir yang berfokus pada kata yang mengandung sinonim.

p-ISSN 2301–4156 | e-ISSN 2460–5719

TABEL I
CONTOH DATA MASUKAN

Judul	Abstrak
Aplikasi Media Pembelajaran Pengenalan Aksara Jawa Menggunakan *Augmented Reality* di Smartphone Android (Studi Kasus SDN 07 Adipala)	Teknologi mengalami kemajuan yang pesat dan mempengaruhi dunia pendidikan juga berbagai inovasi telah dilakukan untuk menunjang kegiatan belajar mengajar. Proses belajar mengajar bahasa Jawa yang dilakukan guru saat ini masih menggunakan cara ceramah juga membosankan inovasi teknologi yang dapat digunakan dan dapat membantu siswa untuk belajar bahasa Jawa salah satunya adalah *Augmented Reality* (*AR*). Penelitian ini memiliki rumusan masalah yaitu bagaimana merancang dan membuat aplikasi *Augmented Reality* (*AR*). Pembelajaran pengenalan aksara Jawa di *smartphone Android* Penulis menggunakan metode literature dan observasi lapangan sebagai metode pengumpulan data dan sebagai metode pengembangan sistem penulis menggunakan *MDLC* (*Multimedia Development Life Cycle*). Kata kunci: *Augmented Reality*, *MDLC*, *Smartphone Android*, Aksara Jawa, Teknologi, Pendidikan, Belajar Mengajar
Aplikasi Pembelajaran Pengenalan Rangka Manusia Menggunakan *Augmented Reality* (*AR*). Berbasis *Smartphone Android* (Studi Kasus: SD Negeri Jepara Kulon 01, Binangun)	Pemahaman belajar siswa dapat meningkat dengan tersedianya media belajar yang menarik. Media belajar yang menarik dapat memudahkan pengajar dalam menyampaikan materi. Penelitian ini diambil dari studi lapangan dengan guru dan juga siswa sekolah dasar di SDN Jepara Kulon 01, Binangun. Kurangnya alat peraga rangka manusia dan metode pembelajaran di SDN Jepara Kulon 01, Kecamatan Binangun yang kurang menarik karena hanya tersedianya gambar rangka manusia 2D membuat siswa merasa bosan, sehingga dapat menjadi salah satu penyebab kurangnya pemahaman siswa terhadap materi yang diajarkan. Penelitian ini dilakukan dengan menerapkan teknologi *Augmented Reality (AR)* yang diimplementasikan dalam bentuk aplikasi berbasis android. Teknologi *Augmented Reality* (AR) merupakan perpaduan antara 2D, 3D, dan dunia nyata yang digabung dalam satu objek dengan satu teknologi yang dapat digunakan sebagai media pembelajaran di bidang multimedia. Aplikasi ini dikembangkan menggunakan metode *Multimedia Development Life Cycle* (*MDLC*). Hasil kuisioner menunjukan aplikasi ini dapat membantu guru dan siswa dalam mengenalkan serta memahami bagian - bagian rangka manusia dan juga menarik minat belajar dibanding menggunakan buku dengan persentase yang didapat 28% menyatakan setuju dan 72% menyatakan sangat setuju dengan aplikasi ini. Kata Kunci: android, *Augmented Reality (AR)*, media pembelajaran, rangka manusia

Gambar 1 mengilustrasikan tahapan-tahapan dalam penelitian, dari proses awal, yaitu pengumpulan data, pengolahan data awal, implementasi algoritma, yaitu *synonym recognition* untuk deteksi teks yang mengandung sinonim dan algoritma *winnowing* untuk pembobotan suatu teks pada dokumen (*fingerprint*), hingga proses akhir yang memperoleh persentase hasil nilai kemiripan dari dokumen yang dibandingkan dan nilai akurasi sistem yang menggunakan metode *winnowing*, *synonym recognition*, dan *cosine similarity*. Untuk mengetahui pengaruh *synonym recognition* pada sistem dengan kombinasi pembobotan *winnowing* dan *cosine similarity*, dilakukan pengujian akurasi dan RMSE.

A. PENGUMPULAN DATA

Tahap ini adalah langkah awal untuk memulai penelitian, yaitu cara dan tempat data penelitian diperoleh. Data diperoleh dengan pengumpulan data primer, yaitu data yang diperoleh secara langsung pada tempat studi kasus, yaitu pada jurusan Teknik Informatika Politeknik Negeri Cilacap.

B. PENGOLAHAN DATA

Dataset berupa judul dan abstrak bahasa Indonesia tugas akhir mahasiswa Politeknik Negeri Cilacap jurusan Teknik Informatika angkatan 2021 dan 2022. Terdapat 182 laporan tugas akhir yang kemudian dibagi menjadi tiga bagian berdasarkan tema tugas akhir, yaitu *augmented reality* (AR) sebanyak 42, sistem pendukung keputusan sebanyak 63, dan sistem informasi berbasis *e-commerce* sebanyak 77. Rata-rata panjang judul antara 12 sampai 15 kata per judul, sedangkan panjang abstrak rata-rata 350 sampai 500 kata per abstrak.

Pada proses ini data yang harus diunggah dipersiapkan, yaitu dokumen teks judul dan abstrak tugas akhir, agar dapat diproses pada metode pengukuran kemiripan teks. Data diperoleh dari sistem informasi tugas akhir jurusan Teknik Informatika Politeknik Negeri Cilacap. Data yang diperoleh masih dalam laporan tugas akhir penuh, yang kemudian dipisah setiap judul dan abstraknya, lalu disimpan dalam dokumen yang berbeda. Contoh data masukan berupa data laporan tugas akhir yang dipisahkan antara judul dan abstrak yang disimpan berupa *file* PDF.

Tabel I memperlihatkan contoh masing-masing dokumen dengan judul dan abstrak yang dijadikan *dataset* untuk pengujian model. Dokumen dari masing-masing judul dan abstrak tersebut akan dibandingkan dan dicari kata yang mengandung sinonim, kemudian dibuat pembobotan teks yang akan menghasilkan *fingerprint*. Kemudian, *fingerprint* dari setiap dokumen tersebut dihitung nilai kemiripannya.

C. PRAPEMROSESAN

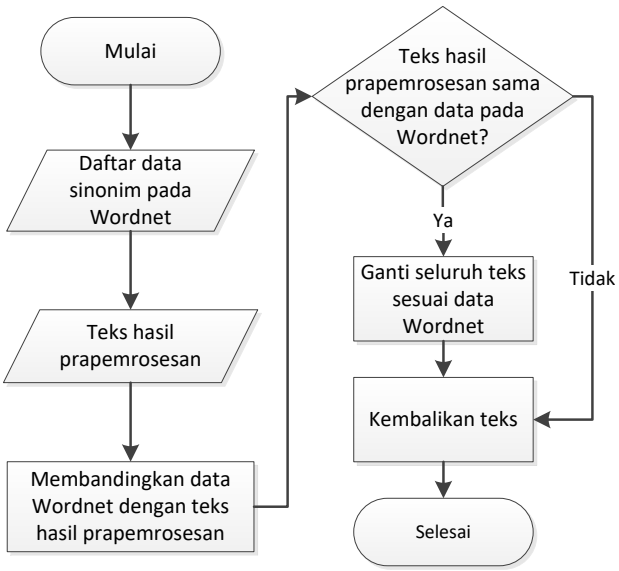
Prapemrosesan adalah tahap awal proses penambangan teks (*text mining*). Prapemrosesan berfungsi mengubah data teks yang belum terstruktur menjadi data teks terstruktur [16], [17]. Pada prapemrosesan ini, dilakukan langkah-langkah untuk menghapus bagian-bagian teks yang tidak diperlukan pada sebuah dokumen karena akan menjadi *noise* pada proses selanjutnya. Prapemrosesan dibagi menjadi tiga tahap, yaitu *tokenizing*, *filtering*, dan *stemming* [18], [19].

Tokenizing adalah proses untuk memisahkan setiap kata pada suatu dokumen yang disebuat sebagai token [20]. Proses ini juga mengubah semua huruf besar menjadi huruf kecil. Selain itu, *tokenizing* juga menghapus semua tanda baca, angka, dan simbol-simbol karena tidak memiliki nilai unik sebab tidak berkaitan dengan *string* yang akan diproses.

Fungsi *filtering* adalah untuk membuang kata-kata yang tidak mempunyai makna. Kata-kata yang tidak mempunyai makna biasa disebut *stopword*. Beberapa contoh kata *stopword* adalah "juga," "dan," "untuk," serta "adalah" [21]. Proses penghapusan *stopword* ini sangat diperlukan karena jika kata-kata penghubung pada suatu kalimat sering muncul, persentase kemiripan teks akan sangat tinggi dan akan mengganggu akurasi metode kemiripan teks [22].

Stemming berfungsi menghilangkan imbuhan pada suatu kata pada dokumen teks, sehingga kata yang diambil adalah kata dasar. Hal ini dilakukan untuk mempermudah proses selanjutnya. Beberapa contoh kata imbuhan adalah "mem-," "-kan," "ber-," "-pun," dan "me-an" [23]. Kata dasar yang

p-ISSN 2301–4156 | e-ISSN 2460–5719



Gambar 2. Proses *synonym recognition*.

diperoleh dijadikan sebagai token-token pada setiap dokumen teks agar lebih cepat dan tepat dalam pencocokan sintaksis. Sebagai contoh, pada dokumen 1 terdapat kata "belajar" dan pada dokumen 2 terdapat kata "mengajar". Setelah dilakukan proses *stemming*, kata "belajar dan mengajar" berubah menjadi "ajar" karena "ajar" adalah kata dasar dari "belajar" dan "mengajar".

D. SYNONYM RECOGNITION

Synonym recognition merupakan salah satu metode pendeteksian kegiatan plagiat teks dengan pendekatan sinonim [7]. *Wordnet* merupakan suatu basis data leksikal untuk digunakan di bawah kendali suatu program [24]. Basis data tersebut berisi berbagai jenis kosakata yang diatur sesuai dengan sinonimnya menurut sebuah konsep leksikal. Satu kata mempunyai dua bahkan lebih makna kata yang sama yang disebut sinonim. Pada proses *synonym recognition*, dibutuhkan daftar kata yang mempunyai makna sinonim. Daftar kata ini diperoleh dari *website* Kamus Besar Bahasa Indonesia (KBBI) daring dan kamusbesar.com [15].

Gambar 2 menjelaskan tahapan-tahapan pada metode *synonym recognition*, dari masukan hingga keluaran. Teks hasil prapemrosesan masuk ke tahap *synonym recognition* dengan tujuan untuk mendeteksi suatu kegiatan plagiat pada suatu dokumen teks. Tahap ini membandingkan dokumen satu dengan dokumen yang lain dengan cara mendeteksi kata-kata yang mengandung sinonim. Proses *synonym recognition* membandingkan kata yang terdapat dalam dokumen dengan tesaurus yang terdapat dalam basis data.

Algoritma *winnowing* saja tidak dapat mendeteksi plagiat suatu dokumen jika kata dalam dokumen tersebut mengalami perubahan kata, tetapi memiliki arti yang sama. Maka, dibutuhkan algoritma *synonym recognition* untuk mengatasi masalah perubahan kata tersebut. Proses *synonym recognition* mengubah semua kata yang didiagnosis sebagai sinonim dari suatu kata yang terdapat dalam kamus sinonim yang dianggap sebagai kata utama [25].

E. WINNOWING

Pada proses *winnowing*, data diolah dari bentuk *string* menjadi data numerik. Pada proses tersebut, pengguna harus memasukkan nilai parameter, yaitu *k-gram*, *hash*, dan *window*. Untuk menentukan nilai *k-gram*, *hash*, dan *window*, harus dilakukan pengujian parameter dengan data sampel tersebut agar diperoleh nilai kemiripan yang baik.

Masukan dari algoritma *winnowing* adalah teks kata dasar dari hasil prapemrosesan data. Kemudian, dihasilkan keluaran berupa kumpulan nilai *hash*. Nilai *hash* adalah suatu nilai numerik yang dibentuk dari perhitungan tabel ASCII dari setiap karakter. Nilai *hash* juga dapat disebut sebagai *fingerprint*, yang digunakan sebagai indikator untuk membandingkan kemiripan antardokumen teks [26].

Sebelum memulai proses *winnowing*, dilakukan proses *synonym recognition*, yaitu pengolahan kata dasar hasil prapemrosesan yang dibandingkan dengan kamus sinonim. Selanjutnya, data masuk pada proses *k-gram*, yang digunakan untuk mendapatkan kumpulan *string* yang baru dari kumpulan *string* yang lama, dengan nilai *k-gram* yang ditentukan oleh pengguna. Kumpulan *string* dikelompokkan menjadi kumpulan *string* yang baru, yang merupakan penggabungan *string* awal, dengan panjang *string* yang digabungkan adalah *k*.

Kemudian, dilakukan proses *rolling hash*, yang berfungsi untuk menghasilkan nilai *hash* dari setiap *gram* yang telah dibentuk. Pengubahan serangkaian karakter menjadi suatu nilai atau kode yang kemudian menjadi tanda dari rangkaian karakter tersebut disebut *hashing* dan nilai yang dihasilkan dapat disebut sebagai nilai *hash*. Nilai *hash* yang diperoleh sudah berupa data numerik. Proses ini dapat dilakukan menggunakan rumus sebagai berikut.

$$H(c_1..c_l) = c_1.b^{(l-1)} + c_2.b^{(l-2)} + .. + c_{(l-1)}.b + c_1 \quad (1)$$

dengan c merupakan nilai ASCII pada setiap karakter, l merupakan panjang *string*, dan b adalah nilai basis *hash* yang ditentukan oleh pengguna. Dari proses tersebut, akan didapatkan nilai *hash* dari setiap *gram*.

Langkah selanjutnya adalah membentuk *window* pada hasil nilai *hash* dari setiap *gram* sebelumnya. Besar *window* juga ditentukan oleh pengguna. Proses *window* dilakukan dengan menentukan nilai *hash* dari setiap *window* untuk dijadikan *fingerprint* dokumen dan jika terdapat nilai *hash* yang sama, akan dipilih nilai *hash* yang paling kanan. Kemudian, proses terakhir dari algoritma *winnowing* adalah menentukan nilai *hash* terkecil dari setiap *window* untuk dijadikan *fingerprint* pada suatu dokumen.

F. COSINE SIMILARITY

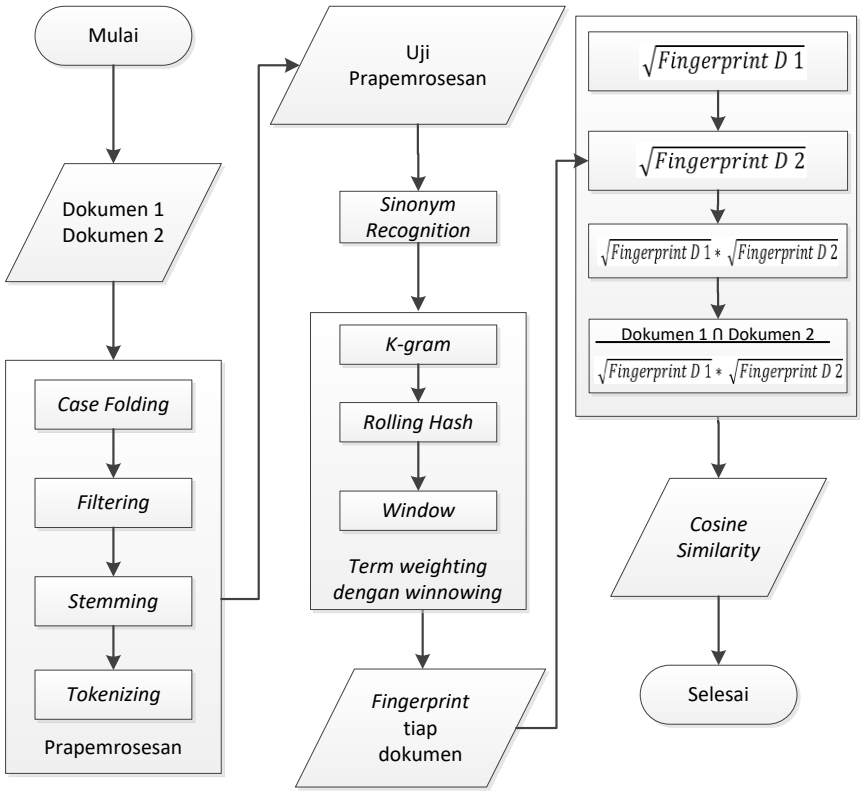
Cosine similarity merupakan salah satu metode yang dapat digunakan untuk menghitung tingkat kemiripan teks pada suatu kalimat atau dokumen. *Cosine similarity* mempunyai nilai akurasi yang tinggi untuk menentukan tingkat kemiripan karena tidak berpengaruh pada panjangnya suatu kata atau kalimat pada suatu dokumen yang dibandingkan [27], [28]. *Cosine similarity* dihitung dengan rumus sebagai berikut.

$$similarity(X,Y) = \frac{(X \cap Y)}{|X|^{0,5}.|Y|^{0,5}} \quad (2)$$

dengan $X \cap Y$ merupakan jumlah kata yang terdapat pada dokumen X dan yang terdapat pada dokumen Y, $|X|$ adalah jumlah kata yang terdapat pada dokumen X, dan $|Y|$ adalah jumlah kata yang terdapat pada dokumen Y.

G. AKURASI

Akurasi digunakan untuk memberikan penilaian hasil prediksi yang sama dengan data aktual. Makin tinggi nilai akurasi, makin akurat atau bagus kinerja metode yang digunakan. Fungsi akurasi adalah untuk mengetahui tingkat



Gambar 3. Diagram alir sistem deteksi kemiripan judul dan abstrak.

ketepatan rekomendasi metode yang digunakan pada suatu sistem. Rumus akurasi dituliskan sebagai berikut [29]-[31].

$$Akurasi = \frac{TP+TN}{TP+FP+TN+FN} \quad (3)$$

dengan *TN* adalah jumlah data negatif yang terdeteksi dengan benar (*true negative*), *FP* adalah data negatif yang terdeteksi sebagai data positif (*false positive*), *TP* adalah data positif yang terdeteksi benar (*true positive*), dan *FN* adalah data positif yang terdeteksi sebagai data negatif (*false negative*).

H. ROOT MEAN SQUARE ERROR (RMSE)

RMSE digunakan untuk mengetahui kinerja sistem yang dibuat. RMSE adalah akar kuadrat dari rata-rata perbedaan kuadrat antara prediksi dan observasi aktual [32]. Makin banyak nilai kemiripan antara sistem dan perhitungan manual, makin tepat metode yang digunakan dalam sistem.

$$RMSE = \sqrt{\frac{1}{n}} \sum_{j=1}^{n} (y_j - \hat{y}_j)^2. \quad (4)$$

III. HASIL DAN PEMBAHASAN

Gambar 3 menjelaskan proses algoritma-algoritma yang digunakan dalam deteksi kemiripan teks dalam judul dan abstrak pada suatu tugas akhir yang berfokus pada teks yang mengandung sinonim.

Untuk memperoleh hasil yang maksimal dalam deteksi kemiripan pada judul dan abstrak dalam proses pembobotan menggunakan metode *winnowing*, perlu dilakukan pengujian nilai parameter *k-gram*, *rolling hash*, dan *window*. Pengujian parameter dilakukan menggunakan contoh dua judul dan abstrak sistem pendukung keputusan dengan nilai-nilai parameter seperti pada Tabel II.

Tabel II menyajikan hasil pengujian parameter *winnowing*. Masing-masing parameter diuji dengan nilai yang berbeda menggunakan *dataset* dua judul dan abstrak tugas akhir dengan

TABEL II
PENGUJIAN PARAMETER

K-gram	*Hash*	*Window*	Kemiripan Judul (%)	Kemiripan Abstrak (%)
5	7	2	24,50	26,21
7	2	5	18,81	19,37
2	5	7	55,84	59,42
2	7	5	51,74	57,33
7	5	2	19,44	20,50
5	2	7	25,55	27,67

tema sistem pendukung keputusan. Hasil parameter terbaik dalam pembobotan teks diperoleh dengan nilai *k-gram* = 2, *hash* = 5, dan *window* = 7, dengan rata-rata nilai kemiripan judul sebesar 55,84% dan kemiripan abstrak sebesar 59,42%. Maka, untuk menganalisis pengaruh algoritma *synonym recognition* terhadap judul dan abstrak tugas akhir dalam pembobotan teks pada algoritma *winnowing*, pada penelitian ini digunakan parameter tersebut, yang akan dikombinasikan dengan metode *cosine similarity* untuk menentukan kemiripan antardokumen.

Proses awal dilakukan dengan membandingkan dua dokumen teks, yang setiap dokumen akan masuk dalam prapemrosesan. Contoh hasil prapemrosesan pada judul 1 adalah "aplikasi media belajar kenal aksara Jawa *Augmented Reality smartphone android* studi kasus SDN Adipala", dan hasil prapemrosesan judul 2 adalah "Aplikasi belajar kenal rangka manusia *Augmented Reality* basis *smartphone android* studi kasus SD Negeri Jepara Kulon Binangun".

Setelah prapemrosesan selesai, langkah selanjutnya adalah proses deteksi kata yang mengandung sinonim. Metode *synonym recognition* ini dapat menambah, menghapus, dan mengubah kata-kata sinonim yang terdapat pada basis data kamus sinonim. *String* pada judul atau abstrak dibandingkan

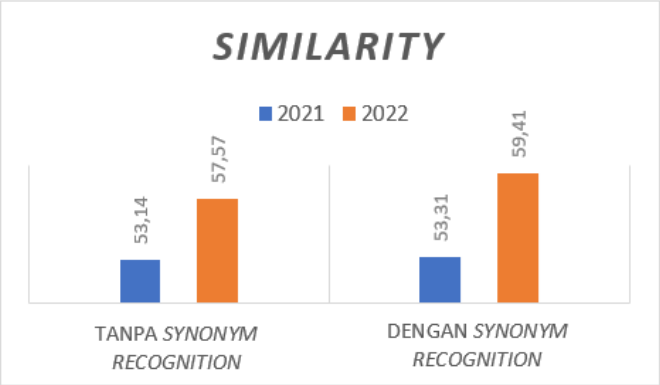
p-ISSN 2301–4156 | e-ISSN 2460–5719

atau dicocokkan dengan kamus sinonim. Kamus sinonim merupakan kumpulan kata berbasis sinonim dalam bahasa Indonesia. Kamus sinonim berfokus pada tiga tema besar, yaitu tentang AR, sistem informasi berbasis *e-commerce*, dan sistem pendukung keputusan. Jika terdapat kata judul dan abstrak yang sama dengan kamus sinonim, *string* yang terdapat pada judul atau abstrak akan diganti dengan *string* seperti yang ada pada kamus sinonim. Sebagai contoh, judul 1 setelah melalui prapemrosesan dan *synonym recognition* menjadi "Aplikasi media belajar kenal aksara *Jawa Augmented Reality smartphone android* studi kasus SDN Adipala" dan judul 2 menjadi "aplikasi tatar kenal rangka manusia *Augmented Reality basis smartphone android* studi kasus SD Negeri Jepara Kulon Binangun". Terdapat satu *string* judul dokumen 1 dan 2 yang teridentifikasi memiliki kata yang mengandung sinonim yang terdapat pada kamus basis data sinonim, yaitu "belajar" menjadi "tatar".

Setelah *synonym recognition*, proses selanjutnnya adalah pembobotan teks dari setiap dokumen yang dibandingkan. Proses pembobotan menggunakan metode *winnowing* dilakukan dengan mencari nilai *fingerprint* setiap dokumen. Proses awal pembobotan adalah memasukkan nilai parameter *k-gram*, *rolling hash*, dan *window*.

Contoh hasil proses *winnowing* dengan memasukkan nilai *k-gram* 2 pada judul 1 adalah "ap, pl, li, ik, ka, as, si, it, ta, at, ta, ar, rk, ke, en, na, al, lr, ra, an, ng, gk, ka, am, ma, an, nu, us, si, ia, aa, au, ug, gm, me, en, nt, te, ed, dr, re, ea, al, li, it, ty, yb, ba, as, si, is, ss, sm, ma, ar, rt, tp, ph , ho, on, ne, ea, an, nd, dr, ro, oi, id, ds, st, tu, ud, di, ik, ka, as, su, us, ss, sd, dn, ne, eg, ge, er, rj, je, ep, pa, ar, ra, ak, ku , ul, lo, on, nb, bi, in, na, an, ng, gu, un". Proses kedua dalam *winnowing* adalah *rolling hash*. Sebagai contoh, nilai *hash* yang dimasukkan adalah 5, maka hasil dari judul 1 pada *string* pertama "ap", *string* "a" dalam tabel ASCII memiliki nilai 97, dan "p" memiliki nilai 112, sehingga perhitungan dalam rumus *rolling hash* adalah $(97\times5^{(2-1)})+ (112\times5^{(1-1)}) = 485+112 = 597$. Maka, hasil *rolling hash* keseluruhan pada judul 1 adalah [ap : 597, pl : 668, li : 645, ik : 632, ka : 632, as : 600, si : 680, it : 641, ta : 677, at : 601, ta : 677, ar : 599, rk : 677, ke : 636, en : 615, na : 647, al : 593, lr : 654, ra : 667, an : 595, ng : 653, gk : 622, ka : 632, am : 594, ma : 642, an : 595, nu : 667, us : 700, si : 680, ia : 622, aa : 582, au : 602, ug : 688, gm : 624, me : 646, en : 615, nt : 666, te : 681, ed : 605, dr : 614, re : 671, ea : 602, al : 593, li : 645, it : 641, ty : 701, yb : 703, ba : 587, as : 600, si : 680, is : 640, ss : 690, sm : 684, ma : 642, ar : 599, rt : 686, tp : 692, ph : 664, ho : 631, on : 665, ne : 651, ea : 602, an : 595, nd : 650, dr : 614, ro : 681, oi : 660, id : 625, ds : 615, st : 691, tu : 697, ud : 685, di : 605, ik : 632, ka : 632, as : 600, su : 692, us : 700, ss : 690, sd : 675, dn : 610, ne : 651, eg : 608, ge : 616, er : 619, rj : 676, je : 631, ep : 617, pa : 657, ar : 599, ra : 667, ak : 592, ku : 652, ul : 693, lo : 651, on : 665, nb : 648, bi : 595, in : 635, na : 647, an : 595, ng : 653, gu : 632, un : 695].

Langkah ketiga dalam *winnowing* adalah *window*. Proses *window* sama seperti *k-gram*, tetapi data yang diolah sudah dalam bentuk angka dari proses *rolling hash* sebelumnya. Sebagai contoh, nilai masukan window adalah 5, sehingga hasil dari judul 1 yaitu {597, 668, 645, 632, 632, 600, 680}, {668, 645, 632, 632, 600, 680, 641}, {645, 632, 632, 600, 680, 641, 677}, {632, 632, 600, 680, 641, 677, 601}, {632, 600, 680, 641, 677, 601, 677}, {600, 680, 641, 677, 601, 677, 599}, {680, 641, 677, 601, 677, 599, 677}, {641, 677, 601, 677, 599, 677, 636}, {677, 601, 677, 599, 677, 636, 615} … {595, 635, 647, 595, 653, 632, 695}.



Gambar 4. Pengaruh *synonym recognition* terhadap nilai kemiripan (*similarity*).

Langkah terakhir dari proses *winnowing* adalah mencari nilai minimal dari setiap *window*. Jika terdapat nilai minimal yang sama dari setiap *window*, hanya satu nilai saja yang diambil, yaitu nilai yang paling kanan. Hal ini dilakukan untuk menghindari kelebihan (*redundancy*), sehingga memengaruhi akurasi dan mengurangi waktu proses. Nilai minimal dari setiap *window* disebut *fingerprint*. Hasil nilai *fingerprint* pada judul 1 setelah proses *winnowing* selesai adalah [597, 600, 599, 593, 594, 582, 602, 605, 587, 595, 614, 615, 608, 592].

Setelah diperoleh nilai *fingerprint* dari setiap dokumen, proses selanjutnya adalah mencari nilai kemiripan dokumen teks yang dibandingkan menggunakan metode *cosine similarity*. Proses ini dilakukan dengan membagi jumlah *fingerprint* yang sama dari dokumen yang dibandingkan dengan hasil penjumlahan akar kuadrat *fingerprint* dari setiap dokumen yang dibandingkan.

Proses menggunakan *synonym recognition* dan tanpa *synonym recognition* yang dikombinasikan dengan metode pembobotan menggunakan *winnowing* serta metode pencarian kemiripan menggunakan *cosine similarity* memberikan hasil sebagai berikut. Pada tema sistem pendukung keputusan, nilai rata-rata selisih kemiripan judul pada tahun 2021 dan 2022 adalah 6,8%, pada tema sistem informasi berbasis *e-commerce* sebesar 1,11%, dan pada tema AR sebesar 6,07%. Sementara itu, nilai rata-rata selisih kemiripan abstrak pada tema sistem pedukung keputusan sebesar 1,41%, pada tema sistem informasi berbasis *e-commerce* sebesar 2,75%, dan pada tema AR sebesar 0,52%. Nilai rata-rata kemiripan dari semua tema pada judul dan abstrak yang dibandingkan dengan menggunakan *synonym recognition* dikombinasikan dengan metode pembobotan menggunakan *winnowing* dan metode pencarian kemiripan menggunakan *cosine similarity* adalah 59,32% dan 55,19%, sedangkan nilai rata-rata semua tema pada judul dan abstrak tanpa *synonym recognition* yang dikombinasikan dengan metode pembobotan menggunakan *winnowing* dan metode pencarian kemiripan menggunakan *cosine similarity* sebesar 54,67% dan 53,63%.

Gambar 4 menunjukkan nilai rata-rata kemiripan judul dan abstrak akibat pengaruh penggunaan *synonym recognition* dan tanpa *synonym recognition* pada tahun 2021 dan 2022. Dari gambar tersebut, terlihat bahwa pada tahun 2021 dan 2022 rata-rata nilai kemiripan meningkat sebesar 2,73%.

Sistem deteksi kemiripan judul dan abstrak menampilkan hasil persentase kemiripan dari setiap judul dan abstrak yang dibandingkan. Untuk mengetahui tingkat keberhasilan algoritma-algoritma pada sistem yang digunakan, dicari nilai akurasi dan RMSE. Pengaruh penggunaan *synonym recognition* yang dikombinasikan dengan metode pembobotan

menggunakan *winnowing* dan metode pencarian kemiripan menggunakan *cosine similarity* terhadap deteksi kemiripan judul dan abstrak dibandingkan dengan perhitungan manual dan diperoleh rata-rata akurasi sebesar 81,05%. Nilai akurasi diperoleh dari nilai *TP* sebesar 147,57% dibagi dengan jumlah data. Sementara itu, nilai rata-rata RMSE adalah 4,38%. Nilai ini diperoleh dari data aktual sebesar 63,7% dikurangi nilai hasil peramalan sebesar 2,01%, kemudian dari hasil tersebut dihitung akar pangkatnya, lalu dibagi dengan jumlah data.

IV. KESIMPULAN

Berdasarkan hasil pengujian yang telah dilakukan, disimpulkan bahwa nilai parameter *k-gram* yang makin kecil cukup memengaruhi nilai tingkat kemiripan antardokumen, sedangkan nilai masukan untuk *hash* dan *window* tidak terlalu berpengaruh terhadap nilai tingkat kemiripan antardokumen. Masukan nilai parameter paling tinggi dalam pembobotan teks nilai kemiripan adalah *k-gram* = 2, *hash* = 5 dan *window* = 7. Pengaruh *synonym recognition* terhadap pembobotan teks mengggunakan algoritma *winnowing* dan penghitungan nilai kemiripan menggunakan *cosine similarity* menunjukkan hasil rata-rata kenaikan nilai kemiripan judul dan abstrak sebesar 3,11% dibandingkan dengan *synonym recognition* tanpa algoritma *winnowing*. Kenaikan nilai kemiripan judul dan abstrak diperoleh dari nilai rata-rata judul dan abstrak menggunakan pembobotan *winnowing* yang dikombinasikan dengan *synonym recognition* dan menghitung nilai kemiripan menggunakan *cosine similarity*, sebesar 57,26%, dikurangi nilai rata-rata tanpa *synonym recognition*, sebesar 54,15%. Penambahan metode *synonym recognition* yang dikombinasikan dengan pembobotan *winnowing* dan penghitungan nilai kemiripan mengunakan *cosine similarity* dapat meningkatkan nilai kemiripan, terutama pada teks berbasis sinonim. Pengujian terhadap pengaruh metode *synonym recognition* yang dikombinasikan dengan algoritma *winnowing* dan *cosine similarity* menghasilkan nilai akurasi sebesar 80,97% dan RMSE 26,14%. Hasil ini diperoleh dari *dataset* tugas akhir mahasiswa Politeknik Negeri Cilacap untuk pembuatan model karena jika digunakan *dataset* yang lain, ada kemungkinan didapatkan hasil yang berbeda.

KONFLIK KEPENTINGAN

Penulis menyatakan bahwa tidak terdapat konflik kepentingan dalam penelitian dan penyusunan makalah ini.

KONTRIBUSI PENULIS

Konseptualisasi, Santi Purwaningrum dan Agus Susanto; metodologi, Santi Purwaningrum dan Agus Susanto; perangkat lunak, Santi Purwaningrum; validasi, Santi Purwaningrum dan Agus Susanto; analisis formal, Santi Purwaningrum dan Agus Susanto; analisis formal, Santi Purwaningrum; investigasi, Santi Purwaningrum; analisis formal, Santi Purwaningrum; sumber daya, Santi Purwaningrum; kurasi data, Santi Purwaningrum; penulisan—penyusunan draf asli, Santi Purwaningrum; penulisan—peninjauan dan penyuntingan, Ari Kristiningsih; visualisasi, Santi Purwaningrum; pengawasan, Santi Purwaningrum.

UCAPAN TERIMA KASIH

Terima kasih disampaikan kepada Politeknik Negeri Cilacap yang telah memberikan dukungan dengan hibah dana untuk penelitian ini serta kepada pihak-pihak yang telah membantu penyelesaian penelitian ini.

REFERENSI

[1] M.H.P. Swari, C.A. Putra, dan I.P.S. Handika, "Plagiarism Checker pada Sistem Manajemen Data Tugas Akhir," *J. Sains, Inform.*, Vol. 7, No. 2, hal. 192–201, Nov. 2021, doi: 10.34128/jsi.v7i2.338.

[2] M.H.P. Swari dan C.A. Putra, "Sistem Manajemen Data Skripsi (Studi Kasus: Perpustakaan Fakultas Ilmu Komputer UPN "Veteran" Jawa Timur)," *J. Pendidik. Teknol., Kejuru.*, Vol. 17, No. 2, hal. 198–209, Jul. 2020, doi: 10.23887/jptk-undiksha.v17i2.25436.

[3] F.E. Kurniawati dan W.M. Pradnya, "Implementasi Algoritma Winnowing pada Sistem Penilaian Otomatis Jawaban Esai pada Ujian Online Berbasis Web," *J. Tek. Komput. AMIK BSI*, Vol. 6, No. 2, hal. 169–175, Jul. 2020, doi: 10.31294/jtk.v6i2.7838.

[4] I. Ahmad, R.I. Borman, G.G. Caksana, dan J. Fakhrurozi, "Implementasi String Matching dengan Algoritma Boyer-Moore untuk Menentukan Tingkat Kemiripan pada Pengajuan Judul Skripsi/TA Mahasiswa (Studi Kasus: Universitas XYZ)," *SINTECH (Sci., Inf. Technol. J.)*, Vol. 4, No. 1, hal. 53–58, Apr. 2021, doi: 10.31598/sintechjournal.v4i1.699.

[5] N. Alamsyah dan M. Rasyidan, "Deteksi Plagiarisme Tingkat Kemiripan Judul Skripsi pada Fakultas Teknologi Informasi Menggunakan Algoritma Winnowing," *Technologia*, Vol. 10, No. 4, hal. 197–201, Okt.-Des. 2019, doi: 10.31602/tji.v10i4.2361.

[6] M. Novak, M. Joy, dan D. Kermek, "Source-Code Similarity Detection and Detection Tools Used in Academia: A Systematic Review," *ACM Trans. Comput. Educ.*, Vol. 19, No. 3, hal. 1-37, Mei 2019, doi: 10.1145/3313290.

[7] N.P. Putra dan Sularno, "Penerapan Algoritma Rabin-Karp dengan Pendekatan Synonym Recognition Sebagai Antisipasi Plagiarisme pada Penulisan Skripsi," *J. Teknol., Sist. Inf. Bisnis*, Vol. 1, No. 2, hal. 130–140, Jul. 2019, doi: 10.47233/jteksis.v1i2.52.

[8] S. Fauziati dkk., "Regresi Linear untuk Mengurangi Bias Sistem Penilaian Uraian Singkat," *J. Nas. Tek. Elekt., Teknol. Inf.*, Vol. 10, No. 3, hal. 221–228, Agu. 2021, doi: 10.22146/jnteti.v10i3.1983.

[9] I. Mufiid, S. Lestanti, dan N. Kholila, "Aplikasi Penilaian Jawaban Esai Otomatis Menggunakan Metode Synonym Recognition dan Cosine Similarity Berbasis Web," *J. Mnemonic: J. Tek. Inform.*, Vol. 4, No. 2, hal. 31–37, Sep. 2021, doi: 10.36040/mnemonic.v4i2.4067.

[10] B. Sari dan Y. Sibaroni, "Deteksi Kemiripan Dokumen Bahasa Indonesia Menggunakan Algoritma Smith-Waterman dan Algoritma Nazief & Andriani," *Ind. J. Comput.*, Vol. 4, No. 3, hal. 87–98, Des. 2019, doi: 10.21108/indojc.2019.4.3.365.

[11] M.R. Parvez, W. Hu, dan T. Chen, "Comparison of the Smith-Waterman and Needleman-Wunsch Algorithms for Online Similarity Analysis of Industrial Alarm Floods," *2020 IEEE Elect. Power, Energy Conf. (EPEC)*, 2020, hal. 1-6, doi: 10.1109/EPEC48502.2020.9320080.

[12] V. Kumar, C. Bhatt, dan V. Namdeo, "A Framework for Document Plagiarism Detection Using Rabin Karp Method," *Int. J. Innov. Res. Technol., Manage.*, Vol. 5, No. 4, hal. 17–30, Agu. 2021.

[13] T. Wahyuningsih, Henderi, dan Winarno, "Text Mining an Automatic Short Answer Grading (ASAG), Comparison of Three Methods of Cosine Similarity, Jaccard Similarity and Dice's Coefficient," *J. Appl. Data Sci.*, Vol. 2, No. 2, hal. 45–54, Mei 2021, doi: 10.47738/jads.v2i2.31.

[14] L. Meilina, I.N.S. Kumara, dan N. Setiawan, "Literature Review Klasifikasi Data Menggunakan Metode Cosine Similarity dan Artificial Neural Network," *Maj. Ilm. Teknol. Elekt.*, Vol. 20, No. 2, hal. 307-314, Jul.-Des. 2021, doi: 10.24843/mite.2021.v20i02.p15.

[15] M.N. Cholis, E. Yudaningtyas, dan M. Aswin, "Pengaruh Penggunaan Synonym Recognition dan Spelling Correction pada Hasil Aplikasi Penilaian Esai dengan Metode Longest Common Subsequence dan Cosine Similarity," *InfoTekJar (J. Nas. Inform., Teknol. Jar.)*, Vol. 3, No. 2, hal. 242–246, 2019, doi: 10.30743/infotekjar.v3i2.1061.

[16] Sunardi, A. Yudhana, dan I.A. Mukaromah, "Indonesia Words Detection Using Fingerprint Winnowing Algorithm," *J. Inform.*, Vol. 13, No. 1, hal. 7-15, Jan. 2019, doi: 10.26555/jifo.v13i1.a8452.

[17] M.R. Faisal, D. Kartini, A.R. Arrahimi, dan T.H. Saragih, *Belajar Data Science: Text Mining Untuk Pemula I*, Banjarbaru, Indonesia: Scripta Cendekia, 2023.

[18] H.A. Rouf, A. Wijayanto, dan A. Aziz, "Deteksi Plagiarisme Skripsi Mahasiswa dengan Metode Single-link Clustering dan Jaro-Winkler Distance," *J. Pilar Teknol.*, Vol. 5, No. 1, hal. 26–31, Mar. 2020, doi: 10.33319/piltek.v5i1.50.

[19] C.D. Manning, P. Raghavan, dan H. Schütze, *Introduction to Information Retrieval*, Cambridge, Inggris: Cambridge University Press, 2008.

[20] S.P. Gunawan, L. Dwika, dan A.R. Chrismanto, "Analisis Fitur Stilometri dan Strategi Segmentasi pada Sistem Deteksi Plagiasi Intrinsik Teks," *J. RESTI (Rekayasa Sist., Teknol. Inf.)*, Vol. 4, No. 5, hal. 988-997, Okt. 2020, doi: 10.29207/resti.v4i5.2486.

[21] N.C. Haryanto, L.D. Krisnawati, dan A.R. Chrismanto, "Temu Kembali Dokumen Sumber Rujukan dalam Sistem Daur Ulang Teks," *J. Teknol.*,

Sist. Komput., Vol. 8, No. 2. hal. 140–149, Apr. 2020. doi: 10.14710/jtsiskom.8.2.2020.140-149.

[22] I.M.S. Putra, P. Jhonarendra, dan N.K.D. Rusjayanthi, "Deteksi Kesamaan Teks Jawaban pada Sistem Test Essay Online dengan Pendekatan Neural Network," *J. RESTI (Rekayasa Sist., Teknol. Inf.)*, Vol. 5, No. 6, hal. 1070–1082, Des. 2021, doi: 10.29207/resti.v5i6.3544.

[23] N.L.W.S.R. Ginantra dan N.W. Wardani, "Implementasi Metoda Naïve Bayes dan Vector Space Model dalam Deteksi Kesamaan Artikel Jurnal Berbahasa Indonesia," *J. Infomedia*, Vol. 4, No. 2, hal. 94–100, Des. 2019, doi: 10.30811/jim.v4i2.1530.

[24] R.P. Nuristiqomah dan Y. Anistyasari, "Pengembangan Kamus Istilah Basis Data Berbasis Website Menggunakan Algoritma Cosine Similarity untuk Meningkatkan Hasil Belajar Siswa," *J. IT-EDU*, Vol. 5, No. 2, hal. 621–630, 2021.

[25] R. Nishiyama, "Adaptive Use of Semantic Representations and Phonological Representations in Verbal Memory Maintenance," *J. Mem. Lang.*, Vol. 111, hal. 1-11, Apr. 2020, doi: 10.1016/j.jml.2019.104084.

[26] S. Inturi dan S. Dusa, "Assessment of Descriptive Answers in Moodle-Based E-Learning Using Winnowing Algorithm," *J. Contemp. Issues Bus. Gov.*, Vol. 27, No. 3, hal. 2759–2769, 2021, doi: 10.47750/cibg.2021.27.03.331.

[27] E. Siswanto dan Y.C. Giap, "Implementasi Algoritma Rabin-Karp dan Cosine Similarity untuk Pendeteksi Plagiarisme Pada Dokumen," *J. ALGOR*, Vol. 1, No. 2, hal. 16–22, Mei 2020.

[28] Y. Nurdiansyah, A. Andrianto, dan L. Kamshal, "New Book Classification Based on Dewey Decimal Classification (DDC) Law Using TF-IDF and Cosine Similarity Method," *J. Phys. Conf. Ser.*, Vol. 1211, hal. 1-9, 2019, doi: 10.1088/1742-6596/1211/1/012044.

[29] R.N. Harahap dan K. Muslim, "Peningkatan Akurasi pada Prediksi Kepribadian Mbti Pengguna Twitter Menggunakan Augmentasi Data," *J. Teknol. Inf., Ilmu Komput.*, Vol. 7, No. 4, hal. 815-822, Agu. 2020, doi: 10.25126/jtiik.2020743622.

[30] J. Xu, Y. Zhang, dan D. Miao, "Three-Way Confusion Matrix for Classification: A Measure Driven View," *Inf. Sci. (Ny).*, Vol. 507, hal. 772–794, Jan. 2020, doi: 10.1016/j.ins.2019.06.064.

[31] Y. Zhang dan J.T. Yao, "Gini Objective Functions for Three-Way Classifications," *Int. J. Approx. Reason.*, Vol. 81, hal. 103–114, Feb. 2017, doi: 10.1016/j.ijar.2016.11.005.

[32] E. Sutoyo dan A. Almaarif, "Educational Data Mining untuk Prediksi Kelulusan Mahasiswa Menggunakan Algoritme Naïve Bayes Classifier," *J. RESTI (Rekayasa Sist., Teknol. Inf.)*, Vol. 4, No. 1, hal. 95–101, Feb. 2020, doi: 10.29207/RESTI.V4I1.1502.