Jurnal Vocational Teknik Elektronika dan Informatika
http://ejournal.unp.ac.id/index.php/voteknika/index
**Vol. 12, No. 4, Desember 2024** **P- ISSN: 2302-3295, E-ISSN : 2716-3989**

# Perbandingan Metode Pembobotan Teks dari Algoritma Winnowing dan TF-IDF dikombinasikan Algoritma Cosine Similarity

**Santi Purwaningrum[1*], Oman Somantri [2], Nur Wachid Adi Prasetya[3]**
[1,2,3] Politeknik Negeri Cilacap, Indonesia
Jl. Dr. Soetomo No.1, Sidakaya, Cilacap, Jawa Tengah, Indonesia
**Corresponding author e-mail* : santi.purwaningrum@pnc.ac.id

## ABSTRAK

Tugas akhir di perguruan tinggi adalah syarat kelulusan untuk mendapatkan gelar sarjana atau ahli madya. Tingginya keinginan mahasiswa untuk segera lulus terkadang membuat mahasiswa melakukan tindakan plagiarisme. Plagiarisme adalah tindakan meniru dan mengutip bahkan menyalin atau mengakui hasil karya orang lain sebagai hasil karya dirinya sendiri. Penelitian ini bertujuan untuk mengembangkan sistem yang mendeteksi kesamaan antar dokumen teks berbahasa Indonesia dengan membandingkan dua metode pembobotan teks. Algoritma *Winnowing* dan TF-IDF adalah metode pembobotan teks yang dikombinasikan dengan metode *Cosine Similarity*. *Cosine Similarity* merupakan algoritma yang berfungsi untuk mencari nilai kesamaan antar dokumen dari hasil pembobotan algoritma *winnowing* dan TF-IDF. Hasil penelitian menunjukkan bahwa algoritma *Winnowing* memiliki nilai kesamaan rata-rata 66%, lebih tinggi dibandingkan TF-IDF yang hanya memiliki rata-rata 57%. Performa algoritma diukur menggunakan akurasi dan RMSE. Nilai akurasi pada algoritma *Winnowing* adalah 90.47% dan algoritma TF-IDF 81.84%. Nilai RMSE pada algoritma *Winnowing* sebesar 5,44 dan TF-IDF sebesar 5,34.

**Kata kunci** : *Winnowing*, TF-IDF, *Cosine Similarity*.

## *ABSTRACT*

*The final project at a higher education institution is a graduation requirement to obtain a bachelor's or associate degree. The strong desire of students to graduate quickly sometimes leads them to commit plagiarism. Plagiarism is the act of imitating, quoting, or even copying or acknowledging someone else's work as their own. This research aims to develop a system that detects similarities between Indonesian text documents by comparing two text weighting methods. The Winnowing and TF-IDF algorithms are text weighting methods combined with the cosine similarity method. Cosine similarity is an algorithm used to find the similarity value between documents based on the weighting results of the Winnowing and TF-IDF algorithms. The results of the study showed that the Winnowing algorithm had an average similarity value of 66%, higher than TF-IDF which only had an average of 57%. The performance of the algorithm uses measurements and RMSE. The algorithm's performance was measured using accuracy and RMSE. The accuracy value of the winnowing algorithm is 90.47% and the TF-IDF algorithm is 81.84%. The RMSE value of the Winnowing algorithm is 5.44 and TF-IDF is 5.34.*

***Keywords:*** *Winnowing, TF-IDF, Cosine Similarity.*

## I. PENDAHULUAN

Perkembangan sistem informasi yang pesat di dunia maya secara perlahan mengubah gaya hidup manusia menuju era digital. Salah satu bidang yang berdampak pada pesatnya perkembangan sistem informasi adalah pada bidang pendidikan. Sistem administrasi hingga pengajaran pada bidang pendidikan sudah menggunakan sistem informasi. Pengajaran pada perguruan tinggi kini dapat dilaksanakan secara luring maupun daring.

Tugas akhir atau skripsi merupakan salah satu matakuliah yang wajib diikuti mahasiswa tingkat akhir untuk dapat lulus dan mendapat gelar pada perguruan tinggi. Sistem penulisan tugas akhir juga mengalami perkembangan mulai dengan ditulis tangan, menggunakan mesin ketik hingga sekarang sudah berkembang dengan diketik pada mesin komputer.

Penggunaan komputer dalam penulisan tugas akhir mempermudah revisi dan pencarian referensi, namun juga meningkatkan risiko plagiarisme. Selain

lamanya waktu mahasiswa revisi, keinginan tinggi mahasiswa yang ingin segera lulus merupakan salah satu alasan mahasiswa melakukan plagiarisme.

Plagiarisme yaitu kegiatan adanya kesamaan antar dokumen dengan dokumen yang menjadi sumber. Plagiarisme dapat dilakukan dengan menyalin teks dari satu dokumen ke dokumen lain tanpa menyebutkan sumber asli [1].

Tindakan plagiarisme diatur dalam UU nomor 19 tahun 2002 tentang hak cipta. Macam-macam plagiarisme dapat berupa kata, ide, sumber dan kepengarangan. Salah satu foktor pendorong seseorang melakukan plagiarisme adalah kurangnya kreativitas dan kurangnya pengetahuan atau pengalaman dalam melakukan sitasi [2]. Kegiatan plagiarisme banyak ditemui dalam berbagai lembaga, salah satunya adalah lembaga pendidikan. Pengaruh prokrastinasi akademik terhadap tingkat plagiarisme di kalangan mahasiswa mencapai 71%. Analisis menunjukkan bahwa peningkatan plagiarisme berkaitan dengan ketersediaan informasi di internet dan tekanan untuk publikasi di lingkungan akademis. Faktor lain yang berkontribusi termasuk kurangnya kepercayaan diri dan keterampilan menulis ilmiah, ketidakpahaman penulis tentang konsep plagiarisme, serta kebiasaan menunda dalam menyelesaikan tugas [3]. Pencegahan plagiarisme dapat dilakukan melalui sosialisasi dan deteksi dini dengan mengembangkan sistem text mining untuk mendeteksi kesamaan teks antar dokumen.

Beberapa metode yang digunakan dalam sistem informasi deteksi plagiarisme, dalam setiap metode tersebut terdapat beberapa algoritma pembobotan dan algoritma perhitungan tingkat kesamaan pada setiap dokumen. Algoritma pembobotan teks antara lain *Winnowing*, *Term Frequency dan inverse document frequenc* (TF-IDF), *Levenshtein Distance*, dll[4]. Sedangkan algoritma untuk menghitung tingkat kesamaan antar dokumen antara lain *Dice Similarity*, *Cosine Similarity, Vector Space Model* (VSM), *Jaccard Distance* dll[5]. [5]. Dalam penelitian ini berfokus pada pembandingan dua metode pembobotan teks yaitu algoritma TF-IDF dan *Winnowing* yang dikombinasikan dengan algoritma *Cosine Similarity*.

Terdapat banyak peneliti yang tertarik untuk meneliti tentang teks mining. Pada penelitian yang dilakukan oleh yulena sari dkk dengan judul perbandingan metode pembobotan TF-RF dan TF-IDF dengan dikombinasikan dengan *Weighted Tree Similarity* untuk sistem rekomendasi buku [6]. Pada penelitian ini sistem informasi dapat membantu memberikan rekomendasi buku yang sesuai dengan kata kunci yang dicari pada unit pusat terpadu universitas lambung mangkurat. Motode yang digunakan untuk menghitung nilai kesamaan adalah *Weighted Tree*, nilai parameter metode *Weighted Tree Similarity* mengacu pada penelitian [7] yaitu 0,4 untuk bobot judul, 0,2 untuk bobot pengarang, 0,4 untuk bobot sinopsis. Metode TF-IDF dinilai mempunyai akurasi lebih tinggi dibandingkan dengan metode TF-RF dan mempunyai nilai presisi 98%. Implementasi *Natural Language Processing* (NLP) dan algoritma *Cosine Similarity* dalam penilaian ujian esai otomatis merupakan judul penelitian dari Daniel Oktodeli Sihombing Program [8].

Pada penelitian tersebut menggunakan metode NLP mulai dari proses *preprocessing*, pembobotan teks menggunakan algoritma *term frequency* (TF) hingga penilaian jawaban secara otomastis. Metode NLP dikombinasikan dengan metode *cosine similarity* yang bertujuan untuk menghitung nilai kesamaan dari setiap jawaban esai dengan jawaban yang benar. Hasil dari penelitian ini adalah dengan menggunakan penilaian multisoal, nilai kesamaan dari setiap soal menjadi dasar dari perhitungan nilai akhir dari setiap soal dan selanjutnya ditambahkan semuanya untuk medapatkan nilai keseluruhan hasil ujiain esai dari setiap mahasiswa. Algoritma *Cosine Similirity* memperoleh nilai rata-rata kesamaan dari soal-soal esai tersebut sebanyak 84.99%. Algoritma *Winnowing* juga pernah menjadi salah satu metode yang digunakan dalam penelitian Lilik Sugiarto dengan judul analisa algoritma *String Matching* dan *Winnowing* untuk deteksi kemiripan judul tugas akhir perguruan tinggi. Algoritma *String Matching* yang digunakan dalam penelitian ini adalah *Knout Morris Pratt.* Algoritma *String Matching* dinilai lebih efisien akan tetapi tidak lebih efektif dari pada *Winnowing*, untuk menghitung nilai kemiripan menggunakan *Jaccard Coeficient* [9].

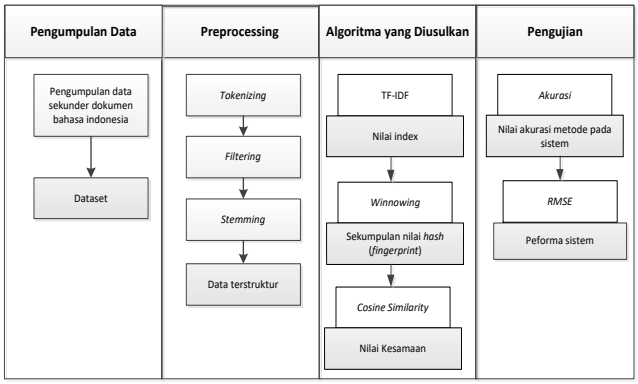
Berdasarkan beberapa penelitian yang terkait dengan metode pembobotan dan perhitungan nilai kesamaan pada suatu teks maka dalam penelitian ini bertujuan untuk membandingkan metode pembobotan teks yang menggunakan algoritma TF-IDF dan *Winnowing* yang dikombinasikan dengan metode perhitungan nilai kesamaan yaitu *Cosine Similarity*.

## II. METODE

Pada penelitian yang dilakukan berdasarkan maraknya plagiarisme di era digitalisasi yang semakin berkembang pesat saat ini, baik dalam karya tulis kalangan akademis ataupun pada karya tulis yang lain. Sehingga banyak penelitian tentang deteksi kesamaan teks akan tetapi hasil yang dicapai belum mendapat hasil yang maksimal. Perkembangan teknologi yang berkembang saat ini diharapkan dapat memberi solusi dari permasalahan kegiatan plagiarisme dan menjadi pertimbangan bagi peneliti yang lain. Penelitian ini membandingkan dua metode pembobotan teks, yaitu algoritma *Winnowing* dan TF-IDF, yang dikombinasikan dengan metode *Cosine Similarity* untuk menghitung kesamaan dokumen.


P-ISSN: 2302-3295

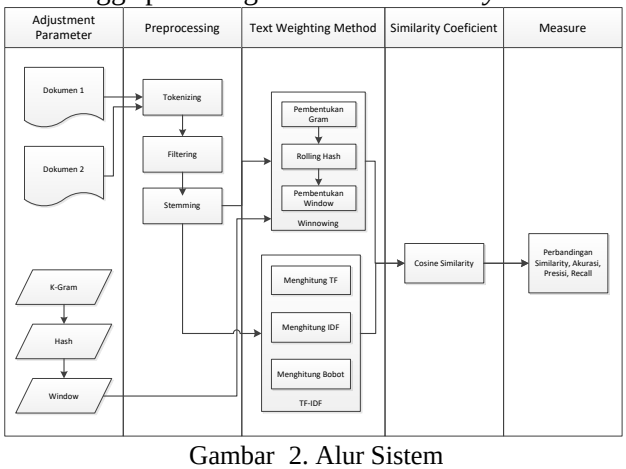
Algoritma *Winnowing* digunakan untuk menghasilkan nilai *hash* dari dokumen teks dan menentukan kemiripan, sementara TF-IDF digunakan untuk menghitung frekuensi kemunculan kata sebagai indeks bobot teks[9]. Kedua algoritma tersebut dikombinasikan dengan algoritma *Cosine Similarity* yang berfungsi untuk menghitung nilai kesamaan dokumen. Berikut kerangka pemikiran dari penelitian ini yang dapat dilihat pada gambar 1.




Gambar 1. Metode penelitian

Gambar diatas menjelaskan tentang penelitian dari proses awal yaitu pengumpulan data. Proses selanjutnya adalah *preprocessing* yang digunakan untuk memperoleh kata terstruktur pada suatu dokumen menggunakan proses *tokenizing*, *filtering*, dan *stemming*. Algoritma yang diusulkan yaitu membandingkan metode pembobotan teks antaralain algoritma TF-IDF untuk menghitung nilai index dari setiap teks dan algoritma *Winnowing* untuk menghitung nilai *fingerprint*. Setelah diperoleh nilai *index* dan *fingerprint* proses selanjutnya yaitu menghitung persentase hasil nilai kemiripan dari setiap teks pada dokumen yang dibandingkan menggunakan algoritma *cosine similarity*. Pengaruh dari metode pembobotan teks, yaitu Algoritma *Winnowing* dan TF-IDF yang dikombinasikan dengan *Cosine Similarity*, diukur menggunakan akurasi dan *Root Mean Square Error* (RMSE).

Alur proses sistem dijelaskan melalui Gambar 2, yang menunjukkan langkah-langkah dari pengumpulan data hingga perhitungan *Cosine Similarity*.




Gambar 2. Alur Sistem

Alur sistem menggambarkan alur logika pemrograman untuk membandingkan metode pembobotan teks menggunakan algoritma *Winnowing* dan TF-IDF yang dikombinasikan dengan algoritma *Cosine Similarity*. Awal proses sistem dengan menginput dokumen yang akan dibandingkan tingkat kesamaannya, selanjutnya pada proses *preprocessing* yang terdiri dari *tokenizing*, *filtering* dan *stemming* yang berfungsi untuk membuat teks menjadi terstruktur dari setiap dokumen. Setelah teks menjadi terstruktur masuk dalam proses pembobotan teks. Inti dari penelitian ini adalah membandingkan metode pembobotan teks. Pembobotan teks yang pertama menggunakan algoritma *Winnowing* yang menggunakan nilai *k-gram*, *window* dan *hash* untuk memperoleh nilai *fingerprint*. *K-gram* berfungsi untuk membentuk *substring* pada teks. Inti dari algoritma *Winnowing* adalah pembobotan teks dengan cara mengubah teks menjadi angka dengan berpatokan tabel ASCII pada proses *hash* [10]. Tabel ASCII adalah representasi numerik yang menghubungkan karakter-karakter teks dengan angka. Terdapat 255 karakter standar yang akan digunakan untuk mewakili setiap karakter, termasuk huruf, angka, tanda baca, serta karakter kontrol [11]. Proses *hash* menggunakan rumus berikut :

$$c_1 * b^{k-1} + c_2 * b^{k-2} + \cdots + c_{k-1} * b + c_{k+1} \quad (1)$$

c adalah nilai ASCII untuk setiap karakter, b merupakan basis nilai *hash* yang diinput, k merupakan nilai k-gram yang di inputkan[10].

Pada proses *window*, nilai *hash* dari k-gram dikelompokkan menggunakan *window* dengan ukuran yang ditentukan oleh parameter tertentu, biasanya lebih besar dari k. *Window* adalah subset dari nilai *hash* yang berisi sejumlah k-*grams* berturut-turut. Proses terekahir pada algoritma *Winnowing* adalah *fingerprint*, proses *fingerprint* yaitu dari setiap *window*, *hash* terkecil (atau terbesar, tergantung pada implementasi) dipilih sebagai "*fingerprint*" dari jendela tersebut. *Fingerprint* tersebut adalah perwakilan dari substruktur dalam teks yang digunakan untuk identifikasi atau pencocokan.

Pembobotan yang selanjutnya menggunakan algoritma TF-IDF yang dimulai dari proses menghitung term *Frequency*, *Inverse Document Frequency* hingga menghitung TF-TDF, rumus sebagai berikut:

$$Wdt = tfdt\ x\ idft \quad (2)$$

Wdf merupakan nilai bobot dokumen ke-d terhadap kata ke-t, tfdt merupakan banyaknya jumlah kata yang dicari dalam suatu dokumen. Idft merupakan *Infers Document Frequency* dengan rumus (log(N/df)). N merupakan jumlah dari

keseluruhan dokumen dan df merupakan banyaknya jumlah dokumen yang mengandung kata yang dicari [12]. Pada metode *Similarity Coefficient* yang akan dikombinasikan dengan kedua algoritma pembobotan tersebut adalah dengan menggunakan algoritma *Cosine Similarity. Cosine Similarity* merupakan salah satu algoritma yang digunakan untuk menghitung tingkat kesamaan suatu teks pada kalimat atau dokumen. *Cosine Similarity* mempunyai nilai akurasi yang tinggi untuk menentukan tingkat kesamaan karena tidak berpengaruh pada panjangnya suatu kata atau kalimat pada suatu dokumen yang dibandingkan [13]. Rumus *Cosine Similarity* adalah sebagai berikut:

$$Cos\ \alpha = \frac{(A\ .\ B)}{|A|.|B|} = \frac{\sum_{i=1}^{n} Ai\ x\ xBi}{\sqrt{\sum_{i=1}^{n}(Ai)^2}\ X\ \sqrt{\sum_{i=1}^{n}(Bi)^2}} \quad (3)$$

A dan B merupakan vektor dari A dan B yang dibentuk dari masing-masing dokumen A dan B. |A| dan |B| merupakan panjang vektor A dan B.

Dalam pengujian perbandingan metode pembobotan menggunakan akurasi dan *Root Mean Square Error* (RMSE). Akurasi digunakakan untuk memberikan penilaian hasil prediksi yang sama dengan data aktual, semakin tinggi nilai akurasi maka performa dari metode yang digunakan semakin akurat atau bagus [14]. Fungsi akurasi yaitu untuk mengetahui seberapa tepatnya rekomendasi metode yang digunakan pada suatu sistem. RMSE digunakan untuk mengetahui performa sistem yang dibuat pada penelitian ini. RMSE adalah akar kuadrat dari rata-rata perbedaan kuadrat antara prediksi dan observasi aktual [15]. Semakin banyak nilai kesamaan antara sistem dan perhitungan manual maka semakin tepat metode yang digunakan dalam sistem.

## III. HASIL DAN PEMBAHASAN

Data yang digunakan dalam penelitian ini adalah abstrak dari tugas akhir mahasiswa pada jurusan Teknik Informatika di Politeknik Negeri Cilacap. Data berupa *softcopy* sebanyak 336 dokumen abstrak yang dibedakan menjadi 3 tema yaitu sistem informasi, *e-commers* dan multimedia. Jumlah data pada setiap tema digambarkan pada tabel 1.

Tabel 1. Dataset

| Tema | Jumlah |
|---|---|
| Sistem Informasi | 178 |
| *e-commers* | 101 |
| Multimedia | 57 |

### *Winnnowing*

Awal proses sistem *text mining* adalah pada tahap *preprocessing* dengan mengubah huruf kapital ke huruf kecil, menghapus karakter dan angka yang dianggap sebagai *noise* biasa disebut sebagai *tokenizing* [16], meghapus kata sambung atau kata yang tidak mempunyai arti disebut *filtering* [17]. Kata sambung antara lain dan, dengan, atau dll. *Steming* merupakan proses dengan mencari kata dasar yang menghilangkan kata hubung seperti ke-, di-, me-, -kan dll.

Hasil *Preprocessing* sebagai berikut :

Dokumen 1:
Pahammaterisiswasiswitingkatsediamediaajartarikmediaajartarikmudahajarsampaimateritelitiambilstudilapanggurusiswasiswisekolahdasarlingkupsdnbojongsarimediabelajarsdnbojongsaridikagamaislampaikelasivbabshalatfardhugunametodetekstualsulitsiswasiswiingathafalmateritelititerapteknologiaugmentedrealityimplementasiaplikasibasisandroidteknologiaugmentedrealityarpaduduniayatagabungobjekteknologibidangmultimediametodegunamdlcmultimediadevelopmentlifecyclekuisioneraplikasiingathafalmateribelajarmudahgunatarikminatajarbandinggunabukupersentasedapatkunciandroidaugmentedrealitymdlcmediabelajarshalatfardhu

Dokumen 2:
Salinpanjangakibatibusalindehidrasilelahakibatuteruslemahkuattindakansiapcegahjadisalinpanjangsenamibuhamilsenambawahpandubidanjadwalbidanpadatpandubutuhaplikasimediapanduibuhamilmandiribangunaplikasiterapteknologiaugmentedrealitybasisandroidpanduibuhamilgunametodekembangsystemmdlcmultimediadevelopmentlifecycleaplikasikembangmarkerfiturtampilaplikasivideodinamismateripreeklamsiafitur isigeraksenamfiturvideoisigerasenamdapatiramairingsenamaplikasirupaaplikasimobileandroidinformasipreeklamsiakualitasaplikasirespondenkuncisenamibuhamilaugmentedrealityandroidmdlc

Kata dasar yang diperoleh dari proses *preprocessing* dijadikan sebagai token-token pada setiap dokumen teks agar lebih cepat dan tepat dicocokan secara syntax [18]. Kemudian token tersebut akan diproses menjadi *fingerprint* dan index yang akan dihitung menggunakan metode pembobotan teks.

Algoritma pertama yang diproses pada metode pembobotan teks adalah *Winnowing*. Pada proses algoritma *Winnowing* diawal proses harus memasukkan parameter k-*gram*, *hash* dan *window*. Tujuan akhir dari ketiga proses *Winnowing* ini adalah mendapatkan nilai dari setiap kata pada suatu dokumen. *K-gram* adalah urutan k karakter berturut-turut dalam teks. Misalnya, jika k = 5, maka *k-gram* pertama adalah lima huruf pertama dalam teks, *k-gram* kedua adalah huruf ke-2 sampai ke-6, dan seterusnya. Ini membantu dalam mengidentifikasi bagian-bagian kecil dari teks yang akan dijadikan sebagai kandidat untuk *fingerprint*. Contoh terdapat dokumen dengan teks “salin panjang akibat ibu salin dehidrasi” dengan k-gram=5, maka hasil dari k-gram adalah salin, alinp, linpa, inpan, npanj, panja, anjan, dst.


P-ISSN: 2302-3295

Proses *Winowing* kedua adalah *hash* yang berfungsi untuk mengkonversi setiap *k-gram* menjadi nilai integer yang unik. Ini memungkinkan representasi yang lebih efisien dan memungkinkan pengidentifikasi k-gram yang sama dengan nilai *hash* yang sama. *Hash* value ini kemudian digunakan dalam proses selanjutnya. Proses *hash* dapat menggunakan fungsi sederhana yang hanya mengubah setiap karakter menjadi kode ASCII-nya. Contohnya:

salin = 115x45-1 + 97x45-2 + 108x45-3 + 105x45-4 + 110x45-5

= 19440 + 6208 + 1728 + 420 + 110 = 37906

slinp = 97x45-1 + 108x45-2 + 105x45-3 + 110x45-4 + 112x45-5

= 24831 + 6912 + 1680 + 440 + 112 = 33976

linpa = 108x45-1 + 105x45-2 + 110x45-3 + 112x45-2 + 975-5

= 27648 + 6720 + 1760 + 448 + 97 = 36673

inpan = 105x45-1 + 110x45-2 + 112x45-3 + 975-4 + 1105-5

= 26880 + 7040 + 1792 + 288 +110 =36210

Npanj = 110x45-1 + 112x45-2 + 975-3 + 1105-4 + 1065-5

= 28160 + 7168 + 1552 + 440 + 106 = 37426

Hasil hash pada dokumen 1 = [paham : 85319, ahamm : 76704, hamma : 80492, ammat : 77576, mmate : 84856, mater : 83769, ateri : 78325, teris : 88615, erisi : 80680, risis : 87890, isisw : 83319, siswa : 88567, iswas : 83575, swasi : 89855, wasis : 90015, asisw : 78319, siswi : 88575, iswis : 83615, swise : 90051, wised : 90980, isedi : 83130, sedia : 87622, ediam : 78844, diame : 78696, iamed : 81080, amedi : 77380, media : 83872, ediaa : 78832, diaaj : 78641, iaaja : 80802, aajar : 75999, ajart : 76986, jarta : 81902 … tymdl : 90958, ymdlc : 92389, mdlcm : 83929, dlcme : 79121, lcmed : 83205, cmedi : 78630, media : 83872, ediaa : 78832, diaaj : 78641, iaaja : 80802, aajar : 75999, ajars : 76985, jarsh : 81904, arsha : 78367, rshal : 88818, shala : 87937, halat : 80426, alatf : 77232, latfa : 83132, atfar : 78274, tfard : 88345, fardh : 79329, ardhu : 78012]

Hasil hash pada dokumen 2 : [salin : 87335, alinp : 77412, linpa : 84032, inpan : 82770, npanj : 85831, panja : 85502, anjan : 77620 … uhami : 89200, hamil : 80483, amila : 77512, milau : 84552, ilaug : 82238, laugm : 83174, augme : 78471, ugmen : 89340, gment : 81191, mente : 84181, ented : 80380, ntedr : 86389, tedre : 88296, edrea : 79077, dreal : 79868, reali : 86945, ealit : 78591, ality : 77451, litya : 84227, ityan : 83745, tyand : 90700, yandr : 91114, andro : 77556, ndroi : 84760, droid : 80150, roidm : 88359, oidmd : 85645, idmdl : 81458, dmdlc : 79264]

Setelah diperoleh nilai *hash* langkah selanjutnya pada proses algoritma *Winnowing* adalah *window* dengan mengelompokkan nilai *hash* sesuai dengan *window* yang telah ditepkan. Seandainya *window* diinput dengan 5 maka hasilnya adalah sebagai berikut :

Hasil *window* pada dokumen 1 : {85319, 76704, 80492, 77576, 84856} , {76704, 80492, 77576, 84856, 83769} , {80492, 77576, 84856, 83769, 78325}, {77576, 84856, 83769, 78325, 88615}, {84856, 83769, 78325, 88615, 80680}, {83769, 78325, 88615, 80680, 87890}, {78325, 88615, 80680, 87890, 83319}, {88615, 80680, 87890, 83319, 88567}, {80680, 87890, 83319, 88567, 83575} … {80802, 75999, 76985, 81904, 78367}, {75999, 76985, 81904, 78367, 88818}, {76985, 81904, 78367, 88818, 87937}, {81904, 78367, 88818, 87937, 80426}, {78367, 88818, 87937, 80426, 77232}, {88818, 87937, 80426, 77232, 83132},{87937, 80426, 77232, 83132, 78274}, {80426, 77232, 83132, 78274, 88345}, {77232, 83132, 78274, 88345, 79329}, {83132, 78274, 88345, 79329, 78012}

Hasil *window* pada dokumen 2 : {87335, 77412, 84032, 82770, 85831}, {77412, 84032, 82770, 85831, 85502}, {84032, 82770, 85831, 85502, 77620}, {82770, 85831, 85502, 77620, 85078}, {85831, 85502, 77620, 85078, 81745}, {85502, 77620, 85078, 81745, 77573}, {77620, 85078, 81745, 77573, 84857}, {85078, 81745, 77573, 84857, 80650}, {81745, 77573, 84857, 80650, 81472} … {78591, 77451, 84227, 83745, 90700}, {77451, 84227, 83745, 90700, 91114}, {84227, 83745, 90700, 91114, 77556}, {83745, 90700, 91114, 77556, 84760}, {90700, 91114, 77556, 84760, 80150}, {91114, 77556, 84760, 80150, 88359}, {77556, 84760, 80150, 88359, 85645}, {84760, 80150, 88359, 85645, 81458}, {80150, 88359, 85645, 81458, 79264}

Setelah diperoleh hasil *window* kemudian diproses menjadi *fingerprint* dengan mencari nilai paling rendah pada setiap *window*. Nilai *fingerprint* tersebut merupakan hasil dari proses algoritma *Winnowing* yang akan dihitung pada *cosine similarity*.

*Fingerprint* dokumen 1 : [76704, 77576, 78325, 80680, 83319, 78319, 83130, 78844, 78696, 77380, 75999, 76986, 78144, 78832, 78641, 82410, 77816, 76677, 76979, 78201, 80681, 79935, 77333, 77766, 82697, 78922, 77703, 77582, 80862, 82550, 83575, 83171, 79912, 76725, 78099, 78103, 78210, 81350, 79361, 78440, 81850, 78146, 76985, 78360, 78105, 78757, 76567, 77262, 77640, 77038, 78313, 76504, 76367, 77232, 78012, 79071, 81001, 78637, 80072, 77693, 83640, 82112, 80017, 78312, 76449, 77451, 83317, 80476, 78137, 79913, 84051, 80613, 78471, 80380, 79077, 78591, 82691, 79941, 78222, 77957, 78212, 76890, 77556, 80150, 80170, 78070, 80497, 78001, 76532, 76453, 79251, 77676, 80575, 77470, 77612, 81581, 82380, 77466, 78458, 79081, 78835, 78651, 76341, 78688, 80069, 80420, 79554,

79128, 78552, 80065, 82869, 80468, 78260, 78285, 77902, 77124, 82360, 78177, 76968, 76730, 77510, 78988, 79247, 80865, 80375, 80074, 78075, 78174, 83795, 79121, 78630, 78367]

*Fingerprint* dokumen 2 : [77412, 77620, 77573, 79343, 77400, 78350, 79364, 78266, 79687, 77231, 80983, 79802, 76982, 78660, 77895, 77090, 77872, 77708, 77664, 77093, 77626, 79950, 77482, 79069, 77530, 79957, 77645, 77558, 77470, 77610, 76693, 77200, 77760, 76242, 79464, 77957, 78271, 78707, 77700, 77565, 77524, 77514, 79098, 76757, 78306, 80476, 78137, 79913, 84051, 80613, 78471, 80380, 79077, 78591, 77451, 76890, 77556, 80150, 79575, 81001, 78414, 76755, 77646, 82466, 80083, 79137, 79081, 82380, 78835, 78651, 76341, 78688, 80069, 80420, 79554, 79128, 78497, 78261, 79795, 76749, 77592, 78169, 80474, 78320, 78505, 78969, 77559, 78325, 80674, 79062, 77747, 76697, 83241, 81692, 80442, 77490, 77441, 83738, 81310, 78540, 80305, 77542, 77314, 77235, 81922, 79942, 77343, 78312, 83457, 76195, 78281, 77677, 78996, 81117, 78299, 77518, 77427, 78093, 78296, 80965, 78527, 80245, 78490, 77512, 79264]

*Fingerprint* dokumen yang sama : [78325, 81001, 78312, 77451, 80476, 78137, 79913, 84051, 80613, 78471, 80380, 79077, 78591, 77957, 76890, 77556, 80150, 77470, 82380, 79081, 78835, 78651, 76341, 78688, 80069, 80420, 79554, 79128]

## Term Frequency Dan Inverse Document Frequency (TF-IDF)

Langkah pertama dalam proses TF-IDF adalah menentukan kata kunci dari dokumen-dokumen yang dibandingkan. Hasil dari proses menentukan kata kunci pada TF-IDF adalah hasil dari proses *preprocessing* setelah proses *stemming* [19]. Sehingga pada proses TF-IDF setiap dokumen sudah menjadi kata kunci, kemudian kata kunci dari setiap dokumen dapat langsung diproses.

Tabel 2. Perhitungan TF-IDF

| TF | | | jumlah (DF) | IDF=log (n/df)+1 | TF-IDF | |
|---|---|---|---|---|---|---|
| Kata | Q/ D1 | D2 | | | Q /D1 | D2 |
| agama | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| ajar | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| ambil | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| android | 1 | 1 | 2 | 1 | 1.30 | 1.30 |
| aplikasi | 1 | 1 | 2 | 1 | 1.30 | 1.30 |
| ar | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| augmented | 1 | 1 | 2 | 1 | 1.30 | 1.30 |
| bab | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| banding | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| basis | 1 | 1 | 2 | 1 | 1.30 | 1.30 |
| belajar | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| bidang | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| bojongsari | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| buku | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| cycle | 1 | 1 | 2 | 1 | 1.30 | 1.30 |
| dapat | 1 | 1 | 2 | 1 | 1.30 | 1.30 |
| dasar | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| development | 1 | 1 | 2 | 1 | 1.30 | 1.30 |
| didik | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| dunia | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| fardhu | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| gabung | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| guna | 1 | 1 | 2 | 1 | 1.30 | 1.30 |
| guru | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| hafal | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| implementasi | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| ingat | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| islam | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| iv | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| kelas | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| kuisioner | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| kunci | 1 | 1 | 2 | 1 | 1.30 | 1.30 |
| lapang | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| life | 1 | 1 | 2 | 1 | 1.30 | 1.30 |
| lingkup | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| materi | 1 | 1 | 2 | 1 | 1.30 | 1.30 |
| mdlc | 1 | 1 | 2 | 1 | 1.30 | 1.30 |
| media | 1 | 1 | 2 | 1 | 1.30 | 1.30 |
| metode | 1 | 1 | 2 | 1 | 1.30 | 1.30 |
| minat | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| mudah | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| multimedia | 1 | 1 | 2 | 1 | 1.30 | 1.30 |
| nyata | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| objek | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| padu | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| paham | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| pai | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| persentase | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| reality | 1 | 1 | 2 | 1 | 1.30 | 1.30 |
| sampai | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| sdn | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| sedia | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| sekolah | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| shalat | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| siswa | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| siswi | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| studi | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| sulit | 1 | 0 | 1 | 1.30 | 1.30 | 0 |


P-ISSN: 2302-3295

| TF | | | jumlah (DF) | IDF=log (n/df)+1 | TF-IDF | |
|---|---|---|---|---|---|---|
| Kata | Q/ D1 | D2 | | | Q /D1 | D2 |
| tarik | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| teknologi | 1 | 1 | 2 | 1 | 1.30 | 1.30 |
| tekstual | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| teliti | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| terap | 1 | 1 | 2 | 1 | 1.30 | 1.30 |
| tingkat | 1 | 0 | 1 | 1.30 | 1.30 | 0 |
| akibat | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| bangun | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| bawah | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| bidan | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| butuh | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| cegah | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| dehidrasi | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| dinamis | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| fitur | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| gerak | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| hamil | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| ibu | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| informasi | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| irama | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| iring | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| isi | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| jadi | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| jadwal | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| kembang | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| kualitas | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| kuat | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| lelah | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| mandiri | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| marker | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| mdl | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| padat | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| pandu | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| panjang | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| preeklamsia | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| responden | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| rupa | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| Salin | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| senam | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| siap | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| system | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| tampil | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| tindakan | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| uterus | 0 | 1 | 1 | 1.30 | 0 | 1.30 |
| video | 0 | 1 | 1 | 1.30 | 0 | 1.30 |

### *Cosine Similarity*

Setalah masing-masing teks pada setiap dokumen mendapatkan bobot dari metode *winnowing* dan TF-IDF langkah selanjutnya adalah proses algoritma *Cosine Similarity*. Algoritma *Cosine Similarity* digunakan untuk menghitung tingkat kesamaan suatu teks pada kalimat atau dokumen. Pada proses *Similarity coefficient* menggunakan algoritma *cosine similarity*. Sesuai persamaan 2 nilai kesamaan diperoleh dengan menghitung dot product dari vector D1 dan D2 yang diperoleh dari jumlah hasil kali antara komponen *vector*. Dengan kata lain, komponen pertama *vector* D1 dikalikan dengan komponen kedua D2, begitu seterusnya sampai komponen terakhir. Perhitungan *Coefisien* pada algoritma *Winnowing* adalah dengan menghitung jumlah dari perkalian saklar antara *fingerprint* Q dengan masing-masing *fingerprint* dokumen. *Fingerprint* Q adalah gabungan antara *fingerprint* antar dokumen. Langkah selanjutnya hasil dari perkalian saklar tersebut dipangkat dua kemudian ditotal dan diakar pangkat dua. Langkah terakhir dari *Cosine Similarity* adalah mencari hasil *Cosine Similarity* dari setiap dokumen dengan cara dari total dokumen yang dikalikan Q dibagi dengan total akar kuadrat dari kuadrat setiap dokumen dikali dengan total akar kuadrat dari kuadrat Q.

Hasil *Cosine Similarity* pada dokumen 1 dengan pembobotan algoritma *Winnowing*:

$$sim = \frac{132}{\sqrt{229}\,x\,\sqrt{132}} = 0.759$$

Hasil *Cosine Similarity* pada dokumen 2 dengan pembobotan algoritma *Winnowing*:

$$sim = \frac{125}{\sqrt{229}\,x\,\sqrt{229}} = 0.54$$

Dari perhitungan *Cosine Similarity* menggunakan pembobotan *Winnowing* mendapatkan nilai rata-rata kesamaan dari dokumen 1 dan dokumen 2 sebesar 65%.

Proses *Cosine Similarity* pada hasil pembobotan pada algoritma TF-IDF dengan menggunakan kedua vektor untuk menghitung nilai kesamaan dokumen D1 dan D2. Perhitungan *Cosine Similarity* pada algoritma TF-IDF adalah dengan menggunakan nilai index disetiap term. Langkah pertama dengan menghitung perkalian saklar antara TF-IDF dari masing-masing dokumen dengan TF-IDF, kemudian dihitung jumlahnya. Selanjutnya menghitung kuadrat dari TF-IDF dari masing-masing dokumen kemudian dicari totalnya dan dari total tersebut diakar kuadratkan. Perhitungan terakhir *Cosine Similarity* terhadap TF-IDF dengan mencari hasil *Cosine Similarity* dari setiap dokumen, dengan menghitung dari total dokumen dikalikan dengan Q pada setiap dokumen kemudian dibagi dengan

perkalian dari total akar kuadrat dari kuadrat TF-IDF pada masing-masing dokumen dan dikalikan dengan total akar kuadrat dari kuadrat TF-IDF dari Q, dengan perhitungan:

Hasil *Cosine Similarity* pada dokumen 1 dengan pembobotan algoritma TF-IDF:

$$sim = \frac{101.28}{\sqrt{161.87}\ x\ \sqrt{162.27}} = 0.62$$

Hasil *Cosine Similarity* pada dokumen 2 dengan pembobotan algoritma TF-IDF:

$$sim = \frac{89.43}{\sqrt{161.87}\ x\ \sqrt{142.21}} = 0.59$$

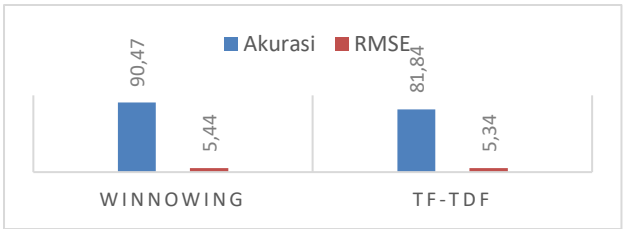
Berdasarkan hasil perhitungan *Cosine Similarity* menggunakan pembobotan TF-IDF mendapatkan nilai rata-rata kesamaan dari dokumen 1 dan dokumen 2 adalah 60%.

Berdasarkan hasil perhitungan, didapat bahwa nilai kesamaan antara dokumen 1 dan dokumen 2 adalah sebesar 14%. Hasil keseluruhan tema sistem informasi, *e-commers* dan multimedia sebagai berikut:

Tabel 3. Persentase Perbedaan Algoritma *Winnowing* dan TF-IDF

| | Winnowing | TF IDF |
|---|---|---|
| Sistem Informasi | 77% | 67% |
| E-commers | 67% | 63% |
| Multimedia | 53% | 40% |

Dari hasil tabel diatas nilai persentase nilai pembobotan teks menggunakan algoritma *Winnowing* yang dikombinasikan dengan *Cosine Similarity* mempunyai selisih sebesar 7% dan lebih tinggi tingkat kesamaannya daripada algoritma TF-IDF. Adapun grafik akurasi dan RMSE dari algoritma *Winnwoing* dan TF-IDF yang di kombinasikan dengan *Cosine Similarity* adalah sebagai berikut :




Gambar 3. Performa algoritma *Winnwoing* dan TF-IDF

Nilai akurasi agoritma *Winnwoing* diperoleh dari nilai true positif sebesar 304 dibagi dengan total jumlah data, sedangkan pada TF-IDF nilai *true positif* sebesar 275 dibagi dengan total jumlah data. Nilai akurasi pada algoritma *Winnwoing* dan TF-IDF cukup tinggi yaitu lebih dari 80%, yang menunjukkan bahwa dua algoritma tersebut mampu mendeteksi lebih banyak pasangan dokumen yang benar-benar serupa.

Nilai rata-rata RMSE pada algoritma *Winnowing* yang diperoleh dari data aktual sebesar 70.42% dikurangkan dengan nilai hasil peramalan sebesar 3.82 % kemudian dihitung akar pangkat dari selisih nilai data tersebut yang kemudian dibagi dengan total jumlah data. Nilai rata-rata RMSE pada algoritma TF-IDF yang diperoleh dari data aktual sebesar 67.7% dikurangkan dengan nilai hasil peramalan sebesar 3.38% kemudian dihitung akar pangkat dari selisih nilai data tersebut yang kemudian dibagi dengan total jumlah data. Hasil RMSE pada algoritma *Winnowing* dan TF-IDF menunjukkan bahwa nilainya rendah yaitu dibawah 5.5%. Nilai RMSE yang rendah menunjukkan bahwa nilai kesamaan yang diperkirakan algoritma Winnowing dan TF-IDF mendekati nilai kesamaan yang sesungguhnya.

## IV. KESIMPULAN

Hasil penelitian menunjukkan bahwa metode pembobotan teks menggunakan algoritma *Winnowing* yang dikombinasikan dengan *Cosine Similarity* memberikan nilai kesamaan yang lebih tinggi dibandingkan TF-IDF. Algoritma Winnowing memiliki nilai kesamaan rata-rata 7% lebih tinggi dibandingkan TF-IDF. Algoritma *Winnowing* terbukti lebih akurat dibandingkan TF-IDF dengan selisih 8,63% dan selisih nilai RMSE sebesar 0.1%, sehingga lebih efektif dalam mendeteksi kesamaan dokumen teks. Temuan ini menunjukkan bahwa algoritma *Winnowing* lebih unggul dalam mendeteksi kesamaan teks, yang dapat digunakan untuk meningkatkan akurasi sistem deteksi plagiarisme.

Penelitian ini memberikan kontribusi dalam pemilihan algoritma yang lebih tepat untuk sistem deteksi plagiarisme, khususnya dalam kasus teks berbahasa Indonesia. Namun, penelitian ini masih terbatas pada data teks bahasa Indonesia. Penelitian selanjutnya dapat memperluas penggunaan metode ini untuk bahasa lain atau jenis teks yang berbeda.

## DAFTAR PUSTAKA

[1] M. A. Shadiqi, “Memahami dan Mencegah Perilaku Plagiarisme dalam Menulis Karya Ilmiah,” *Bul. Psikol.*, vol. 27, no. 1, p. 30, 2019, doi: 10.22146/buletinpsikologi.43058.

[2] M. S. Wahyuni, D. Setiawan, and T. Syahputra, “Sistem Temu Kembali Informasi Dengan Latent Semantic Analisys Pada Kesamaan Tugas Akhir Mahasiswa,” *J. Tek.*, vol. 1, no. 1, 2021.

[3] “View of PENGARUH PROKRASTINASI AKADEMIK TERHADAP TINGKAT PLAGIARISME MAHASISWA PSIKOLOGI UNIVERSITAS YUDHARTA PASURUAN.pdf.”


P-ISSN: 2302-3295

[4] S. Susanti, M. Azmi, E. Ali, R. Rahmaddeni, and Y. Saputra Wijaya, "Perbandingan Boolean Model Dan Vector Space Model Dalam Pencarian Dokumen Teks," *Digit. Zo. J. Teknol. Inf. dan Komun.*, vol. 11, no. 2, pp. 268–277, 2020, doi: 10.31849/digitalzone.v11i2.4168.

[5] A. Meitaningsih, A. S. Aribowo, and N. H. Cahyana, "Text Mining Untuk Mendeteksi Plagiasi Dokumen Dengan Penerapan Stemming Nazief-Adriani Dan Algoritma Smith-Waterman," *Telematika*, vol. 17, no. 2, p. 99, 2020, doi: 10.31315/telematika.v1i1.3377.

[6] Y. Sari, A. R. Baskara, P. B. Prakoso, and N. Royani, "Perbandingan Metode Pembobotan Tf-Rf Dan Tf-Idf Dikombinasikan Dengan Weighted Tree Similarity Untuk Sistem Rekomendasi Buku," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 9, no. 6, pp. 1323–1332, 2022, doi: 10.25126/jtiik.2022935709.

[7] R. Sarno and F. Rahutomo, "Penerapan Algoritma Weighted Tree Similarity Untuk Pencarian Semantik," *JUTI J. Ilm. Teknol. Inf.*, vol. 7, no. 1, p. 39, 2008, doi: 10.12962/j24068535.v7i1.a60.

[8] D. O. Sihombing, "Implementasi Natural Language Processing ( NLP ) dan Algoritma Cosine Similarity dalam Penilaian Ujian Esai Otomatis," vol. 4, pp. 396–406, 2022, doi: 10.30865/json.v4i2.5374.

[9] L. Sugiarto, C. Mulyadi, and S. Rihastuti, "Analisa Algoritma String Matching Dan Winnowing Untuk Deteksi Kemiripan Judul Tugas Akhir Perguruan Tinggi," *J. Teknol. Inf.*, vol. 6, no. 2, pp. 97–106, 2021, doi: 10.52643/jti.v6i2.1141.

[10] R. A. Putra, F. P. Utama, and A. Erlansari, "Penerapan Algoritma Winnowing Pada Sistem Pengelolaan Kerja Praktik Dengan Pendekatan Human-Centered Design (Studi Kasus : Program Studi S-1 Informatika Universitas Bengkulu)," *Pseudocode*, vol. 10, no. 1, pp. 30–44, 2023, doi: 10.33369/pseudocode.10.1.30-44.

[11] Y. W. Hasibuan, R. B. Veronica, J. Matematika, U. N. Semarang, K. S. Gunungpati, and I. Artikel, "How to Cite," vol. 11, no. 1, pp. 54–68, 2022.

[12] A. W. Nila Andriani, "Implementasi Text Mining Klasifikasi Topik Tugas Akhir Mahasiswa Teknik Informatika Menggunakan Pembobotan TF-IDF dan Metode Cosine Similarity Berbasis Web," *Senamika*, no. September, pp. 130–137, 2021, [Online]. Available: https://conference.upnvj.ac.id/index.php/senamika/article/view/1807%0Ahttps://conference.upnvj.ac.id/index.php/senamika/article/download/1807/1350

[13] E. Siswanto and Y. Ceng Giap, "Implementasi Algoritma Rabin-Karp dan Cosine Similarity untuk Pendeteksi Plagiarisme Pada Dokumen," *J. Algor*, vol. 1, no. 2, pp. 16–22, 2020.

[14] M. Azmi, "Analisis Tingkat Plagiasi Dokumen Skripsi Dengan Metode Cosine Similarity Dan Pembobotan Tf-Idf," *Tek. Teknol. Inf. dan Multimed.*, vol. 2, no. 2, pp. 90–95, 2022, doi: 10.46764/teknimedia.v2i2.51.

[15] E. Sutoyo and A. Almaarif, "Educational Data Mining untuk Prediksi Kelulusan Mahasiswa Menggunakan Algoritme Naïve Bayes Classifier," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 1, pp. 95–101, 2020, doi: 10.29207/RESTI.V4I1.1502.

[16] S. P. Gunawan *et al.*, "Analisis Fitur Stilometri dan Strategi Segmentasi pada Sistem Deteksi Plagiasi Intrinsik Teks," *RESTI (Rekayasa Sist. dan Teknol. Inf.*, vol. 4, no. 5, 2021.

[17] N. C. Haryanto, L. D. Krisnawati, and A. R. Chrismanto, "Temu Kembali Dokumen Sumber Rujukan dalam Sistem Daur Ulang Teks," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 2. pp. 140–149, 2020. doi: 10.14710/jtsiskom..8.2.2020.140.140-149.

[18] S. Purwaningrum, A. Susanto, and ..., "Comparation of Dice Similarity and Jaccard Coefficience Against Winnowing Algorithm For Similarity Detection of Indonesian Text Documents," *J. Appl. …*, vol. 6, no. 1, pp. 10–22, 2021, [Online]. Available: http://publikasi.dinus.ac.id/index.php/jais/article/view/4453

[19] R. R. Anugrah, J. Rekayasa, and S. Komputer, "Penerapan Cosine Similarity Dan Pembobotan Tf-Idf Untuk Klasifikasi Pengaduan Masyarakat Berbasis Web ( Studi Kasus : Bagwassidik Ditreskrimum Polda Kalbar)," vol. 11, no. 01, 2023.