Buletin Sistem Informasi dan Teknologi Islam
Vol 3, No 4, November 2022, pp. 268-273
ISSN: 2721-0901

Analisis Algoritma *Winnowing* pada Pendeteksian Plagiarisme Judul Tugas Akhir

Andi Yudra Bramantya Shadiqan Putra[a,1,*], Tasrif Hasanuddina[a,2], Fitriyani Umar[a,3]

[a] *Universitas Muslim Indonesia, Jalan Urip Sumoharjo, Makassar 90245, Indonesia*
[1] *ayudddos@gmail.com;* [2] *tasrif.hasanuddin@umi.ac.id;* [3] *fitriyani.umar@umi.ac.id*
**corresponding author*

INFORMASI ARTIKEL

Diterima : 06 – 10 – 2022
Direvisi : 22 – 11 – 2022
Diterbitkan : 30 – 11 – 2022

Kata Kunci:
judul tugas akhir
plagiarisme
similaritas
winnowing
jaccard coefficient

ABSTRAK

Dalam proses pengajuan judul atau topik tugas akhir yang dilakukan oleh mahasiswa pengecekan kesamaan judul tugas akhir dilakukan untuk mencegah kegiatan plagiarisme. Penelitian ini bertujuan untuk menganalisis algoritma *winnowing* pada pengecekan plagiarisme judul tugas akhir. Proses pengecekan plagiarism, dimulai dari pembentukan rangkaian *k-gram* dan *windows* hingga perhitungan tingkat kemiripan dengan menggunakan *jaccard coeficient*. Pada penelitian ini menggunakan 25 pasang nilai *k-gram* dan *windows*. Hasil dari penelitian ini adalah persentase tingkat kemiripan judul tugas akhir pada penelitian ini diperoleh variabel *k* lebih berpengaruh pada hasil similaritas dibandingkan dengan variabel *w*, selisih hasil kemiripan untuk perbedaan nilai k yaitu 5% - 9% sedangkan selisih hasil kemiripan untuk perbedaan nilai w yaitu 1% - 3%.

I. Pendahuluan

Skripsi merupakan tugas akhir yang wajib dikerjakan oleh Mahasiswa untuk menyelesaikan studi S1 yang bersifat mandiri. Proses pengerjaan tugas akhir ini dimulai dengan menentukan topik atau judul tugas akhir yang mahasiswa dapat peroleh dengan melakukan pencarian topik tugas akhir, membaca jurnal penelitian baik lokal, nasional maupun internasional mengikuti penelitian yang telah dilakukan dosen, atau tugas akhir yang telah dilakukan oleh mahasiswa sebelumnya. Dalam proses pengajuan judul atau topik tugas akhir yang dilakukan oleh mahasiswa, kemungkinan mahasiswa mengumpulkan judul atau topik yang sudah ada atau pernah dilakukan oleh mahasiswa sebelumnya. Skripsi merupakan tugas akhir yang wajib dikerjakan oleh Mahasiswa untuk menyelesaikan studi S1 yang bersifat mandiri. Proses pengerjaan tugas akhir ini dimulai dengan menentukan topik atau judul tugas akhir yang mahasiswa dapat peroleh dengan melakukan pencarian topik tugas akhir, membaca jurnal penelitian baik lokal, nasional maupun internasional mengikuti penelitian yang telah dilakukan dosen, atau tugas akhir yang telah dilakukan oleh mahasiswa sebelumnya. Dalam proses pengajuan judul atau topik tugas akhir yang dilakukan oleh mahasiswa, kemungkinan mahasiswa mengumpulkan judul atau topik yang sudah ada atau pernah dilakukan oleh mahasiswa sebelumnya. Upaya untuk mencegah proses pengusulan judul tersebut maka diperlukan sebuah sistem yang dapat digunakan oleh mahasiswa dalam proses pencarian judul untuk mengetahui judul yang diusulkan sudah pernah diusulkan atau tidak. sistem ini juga dapat digunakan oleh pihak program studi dalam memeriksa judul yang diusulkan oleh mahasiswa.

Membangun sebuah sistem yang dapat melakukan pengecekan kesamaan judul tugas akhir untuk mencegah kegiatan plagiarisme, dibutuhkan metode untuk proses pengecekan plagiarisme tersebut. Algoritma *winnowing* adalah salah satu algoritma yang dapat melakukan metode dokumen *fingerprint*ing untuk mendeteksi plagiarisme [1]. *Document fingerprinting* merupakan metode yang digunakan untuk mendeteksi keakuratan salinan antar dokumen atau hanya sebagian teks saja dengan prinsip kerja menggunakan teknik *hashing* [2]. Teknik *hashing* adalah sebuah fungsi yang mengkonversi setiap string menjadi bilangan [2]. Salah satu langkah dari algoritma *winnowing* yaitu mentransformasikan sebuah kata menjadi sebanyak ukuran *k-gram* ataupun *n-gram*, jika menggunakan *k-gram* didasarkan pada sekumpulan per karakter yang dipotong berdasarkan jumlah *k* namun terdapat pemotongan berdasarkan pemenggalan kata dengan metode memotong berdasarkan *unigram, bigram, trigram, four-gram* [3][4]. Setelah pemotongan berdasarkan jumlah *k-gram* ataupun *n-gram* selanjutnya pembuatan *hashing* pada setiap pemotongan tersebut menggunakan *rolling hash* yang hasilnya dikelompokkan menjadi beberapa *windows* dengan anggota *windows* yang ditentukan. Setelah dikelompokkan maka diidentifikasi nilai terkecil dari setiap *windows* yang menjadi nilai *fingerprint* [5]. Nilai

k-gram dan nilai *windows* ditentukan saat merancang dan membangun sebuah sistem plagiarisme menggunakan metode *winnowing*. Pada penelitian ini, dilakukan analisis judul tugas akhir dengan menggunakan nilai variable *k-gram* dan *windows* dan menghasilkan tingkat kemiripan judul.

II. Metode

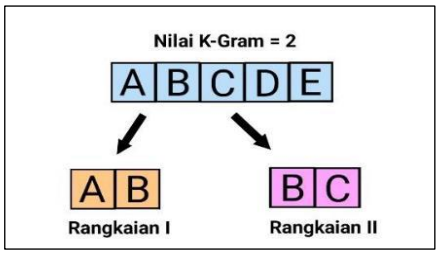
A. Metode Pendeteksi Plagiarisme

Metode Pendeteksi Plagiarisme dibagi menjadi tiga bagian yaitu metode perbandingan teks lengkap, metode dokumen *fingerprinting*, dan metode kesamaan kata kunci [6]. Dokumen *fingerprinting* merupakan metode yang digunakan untuk mendeteksi keakuratan salinan antar dokumen, baik semua teks yang terdapat di dalam dokumen atau hanya sebagian teks saja. Prinsip kerja dari metode dokumen *fingerprinting* ini adalah dengan menggunakan teknik *hashing* [6]. Teknik *hashing* adalah merubah data menjadi serangkaian bilangan bulat dengan jumlah data yang digunakan, dinyatakan dengan *k-gram* [7].

B. Algoritma Winnowing

Winowwing adalah algoritma yang digunakan untuk melakukan proses *document fingerprinting*. Prinsip kerja dari metode *document fingerprinting* ini adalah dengan menggunakan teknik *hashing*. Teknik *hashing* adalah sebuah fungsi yang mengonversi setiap *string* menjadi bilangan. Proses ini ditujukan agar dapat mengidentifikasi kemiripan, termasuk bagian-bagian kecil yang mirip dalam dokumen yang berjumlah banyak [2].

- *K-gram* adalah metode berbasis karakter. Langkah pertama dari *Winnowing* adalah pembentukan *k-gram*. Metode ini memiliki masukan *k*, dimana nilai ini diubah menjadi jumlah potongan atau *substring* dari teks atau kata-kata yang terbentuk. Pada proses ini membutuhkan *substring* karakter semua *k* struktur kalimat [8]. Metode *k-gram* adalah proses pemotongan sejumlah karakter pada sebuah *string* dengan jumlah tertentu (k) [4]. *K-gram* adalah nilai yang dibutuhkan untuk memecah setiap kalimat menjadi rangkaian kata dengan panjang kata sesuai dengan nilai *k*. Berikut contoh ilustrasi proses *k-gram* sebuah kalimat [1].



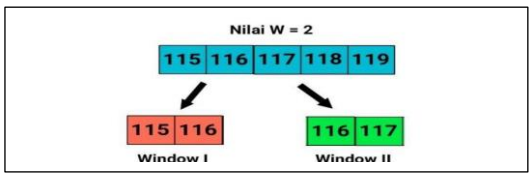
Gambar 1. Ilustrasi Proses *K-gram*

Proses dari *k-gram* dilakukan dengan memecah sebuah kalimat ke dalam beberapa rangkaian karakter dengan panjang yang sama. Berikut contoh penggunaan *k-gram*.

Kalimat : sayaanakteknikinformatika
Nilai *k-gram* : 4
Hasil proses *k-gram* (4-gram) adalah sebagai berikut:

saya ayaa yaan aana anak naki akin kinf info nfor form orma rmat mati atik tika

- *Window* adalah tahapan dimana pembagian *hash* yang telah diperoleh dari proses *Rolling Hash* ke beberapa kelompok berdasarkan nilai *window*. Pada proses ini nilai *hash* yang telah diperoleh dibagi ke dalam *window* yang berukuran w, berikut ilustrasi dari proses *window* [1].



Gambar 2. Ilustrasi Proses *Window*

Pada gambar 2 dapat dilihat contoh pembentukan *window* dengan nilai w = 2, dimana di setiap *window* terdapat dua nilai *hash*. *Window* 1 dimulai dari nilai *hash* yang pertama hingga w, *window* 2 dimulai dari nilai *hash* yang kedua hingga w dan seterusnya.

Langkah-langkah algoritma *winnowing* sebagai berikut [2]:

1) Pembuangan karakter yang tidak relevan
Pembuangan karakter yang tidak relevan memenuhi kebutuhan algoritma *winnowing* yaitu *whitespace insensitivity*.

2) Pembentukan Rangkaian K-gram
Pembentukan rangkaian *n-gram* dilakukan dengan cara membentuk rangkaian karakter sepanjang *n* dari hasil pembuangan karakter yang tidak relevan dengan ukuran *n.*

3) Perhitungan fungsi hash untuk setiap n-gram
Perhitungan nilai-nilai *hash* dari setiap *gram*, fungsi yang digunakan untuk menghasilkan nilai *hash* dari rangkaian *gram* dalam algoritma *winnowing* adalah *rolling hash*. Berikut persamaan untuk menghitung *rolling hash*:

$$H(c1 \ldots . ck) = \mathrm{c}_1 * b^{(k-1)} + \mathrm{c}_2 * b^{(k-2)} + \cdots + c_{(\mathrm{k}\text{-}1)} * b^{(k)} + c_{(\mathrm{k})} \; mod \; q \qquad (1)$$

Keterangan:
H: substring
c: nilai ascii per-karakter
b: konstan bilangan prima
k: banyak karakter
q: module bilangan prima

4) Pembentukan window dari nilai hash
Nilai *hash* yang dibentuk pada tahap sebelumnya dibagi ke dalam *window* (*w*). *Window* pertama berisi nilai hash pertama sampai nilai *hash* ke-*w*. *Window* kedua dibentuk dari nilai *hash* kedua sampai nilai *hash* ke-w+1 dan seterusnya sampai terbentuk *window* dari seluruh nilai *hash.*

5) Pemilihan fingerprint dari setiap window
Menentukan nilai *fingerprint* teks. Nilai *fingerprint* ditentukan dengan memilih nilai *hash* terkecil dari setiap *window.*

6) Persamaan Jaccard Coeficient
Jaccard Coeficient Similarity adalah mengukur kemiripan, ketidakmiripan, dan jarak dari dataset dengan mengukur antara dua dataset adalah hasil dari pembagian antara jumlah data yang sama dari kedua dataset dibagi dengan jumlah semua data pada dataset [9]. Berikut persamaan *Jaccard Coeficient* [10].

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

Keterangan:
Similarity : Nilai Similaritas / Nilai Kemiripan
$|X \cap Y|$: Jumlah *fingerprint* yang sama dari kedua dokumen.
$|X \cup Y|$: Total Jumlah *fingerprint* kedua dokumen.

III.Hasil dan Pembahasan

Analisis metode *winnowing* dilakukan dengan melihat hasil pengecekan kesamaan judul yang dihasilkan dengan berbagai nilai variabel *k* dan *w*. Dalam melakukan analisis pengaruh variabel *k* dan variabel *w*, digunakan dua judul tugas akhir yang dibandingkan hasil kesamaan antara kedua judul tersebut. Dari hasil tersebut dapat diketahui pengaruh variabel k dan w dalam menerapkan metode *Winnowing*. Berikut contoh judul yang digunakan dalam pengecekan kemiripan:

Tabel 1. Teks Uji dan Teks Asli Judul Tugas Akhir

Teks Uji	Rancang dan Bangun Aplikasi Bahasa Toraja Berbasis Mobile
Teks Asli	Perancangan Aplikasi Bahasa Wolio Berbasis Android

Hasil pembentukan *k-gram* dan *window* dari judul Tugas Akhir sebagai berikut :

Tabel 2. Nilai *K-Gram* dan *Window* Judul Tugas Akhir

No	Nilai	
	k	***w***
1	2	2
2	2	3
3	2	4
4	2	5
5	2	6
6	3	2
7	3	3
8	3	4
9	3	5
10	3	6
11	4	2
12	4	3
13	4	4

No	Nilai	
	k	***w***
14	4	5
15	4	6
16	5	2
17	5	3
18	5	4
19	5	5
20	5	6
21	6	2
22	6	3
22	6	3
23	6	4
24	6	5
25	6	6

Hasil pengecekan kemiripan kedua judul, didapatkan nilai similaritas yang berbeda-beda. Berikut hasil pengecekan kemiripan judul dengan berdasarkan nilai variabel *k* dan *w* masing-masing.

Tabel 3. Teks Uji dan Teks Asli Judul Tugas Akhir

No	Nilai *K-Gram*	Nilai *Window*	Hasil Kemiripan
1	2	2	68.52 %
2	2	3	81.63 %
3	2	4	77.55 %
4	2	5	80.85 %
5	2	6	84.44 %
6	3	2	36.92 %
7	3	3	40.32 %
8	3	4	39.34 %
9	3	5	43.10 %
10	3	6	44.64 %
11	4	2	29.85%
12	4	3	32.81%
13	4	4	36.07%
14	4	5	37.29%
15	4	6	38.60%
16	5	2	25.00 %
17	5	3	25.76 %
18	5	4	26.56 %
19	5	5	29.51 %
20	5	6	30.51 %
21	6	2	20.29 %

No	Nilai *K-Gram*	Nilai *Window*	Hasil Kemiripan
22	6	3	19.12 %
23	6	4	17.91 %
24	6	5	20.31 %
25	6	6	20.97 %

Berdasarkan hasil pengecekan kemiripan judul dari 25 pasang nilai *k* dan *w* dengan 5 nilai k yaitu 2, 3, 4, 5, dan 6 diperoleh rata-rata kemiripan yang berbeda-beda dimana untuk nilai k=2 mendapatkan nilai rata-rata kemiripan 78.60%, nilai k=3 mendapatkan nilai rata-rata kemiripan 40.86%, nilai k=4 mendapatkan nilai rata-rata kemiripan 34.92%, nilai k=5 mendapatkan nilai rata-rata kemiripan 27.47%, dan nilai k=6 mendapatkan nilai rata-rata kemiripan 30.97%. Perbedaan hasil rata-rata kemiripan untuk masing-masing nilai k memiliki selisih nilai kemiripan mulai dari 5% - 7% kecuali antara nilai k=2 dan nilai k=3 memiliki selisih jauh yaitu 37.73%.

Sementara hasil pengecekan kemiripan judul dengan memperhatikan nilai variabel *w* yang terdiri dari 2, 3, 4, 5, dan 6 diperoleh nilai kemiripan yang berbeda-beda setiap nilai *w*. Perbedaan hasil nilai *w* dapat dilihat dari hasil pengecekan kemiripan judul seperti pada kasus nomor 11 – 15 untuk nilai *k*=4 dan masing-masing nilai *w*. Dari percobaan tersebut diperoleh nilai kemiripan dari nilai *w*=2 yaitu 29.85%, nilai kemiripan untuk nilai *w*=3 yaitu 32.81%, nilai kemiripan untuk nilai *w*=4 yaitu 36.07%, nilai kemiripan dari nilai *w*=5 yaitu 37.29%, dan nilai kemiripan untuk nilai *w*=6 yaitu 38.60%. Perbedaan hasil nilai kemiripan untuk masing-masing nilai *w* memiliki selisih kemiripan mulai dari 1% - 3%.

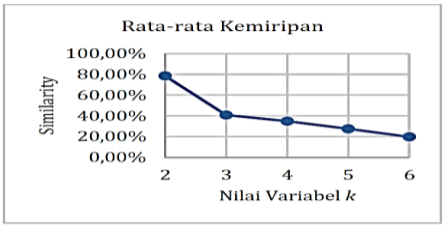
Variabel *k* dan variabel *w* merupakan variabel-variabel yang nilainya ditetapkan untuk melakukan beberapa proses. Variabel *k* untuk menentukan panjang *term* pada pembentukan rangkaian *k-gram*, dan variabel *w* untuk menentukan jumlah *term* yang dikelompokkan dalam pembentukan *window*. Dari hasil penelitian yang telah didapatkan dan dipaparkan pada tabel 3 menunjukan:

a) Semakin tinggi nilai dari variabel *k*, maka menghasilkan nilai similaritas yang rendah dan sebaliknya semakin rendah nilai dari variabel *k*, maka menghasilkan nilai similaritas yang tinggi. Rata-rata nilai kemiripan teks dengan nilai *k* sama dengan 2, 3, 4, 5, dan 6 dapat dilihat pada tabel berikut:

Tabel 4. Rata-rata Nilai Kemiripan

Nilai k	Kemiripan
2	78.60 %
3	40.86 %
4	34.92 %
5	27.47 %
6	19.72%

Persentase rata-rata tingkat kemiripan digambarkan dalam grafik sebagai berikut:



Gambar 3. Grafik Tingkat Kemiripan

b) Nilai similaritas yang dihasilkan oleh nilai k=2 atau nilai *k* yang terlalu rendah menghasilkan nilai paling tinggi. Pada pengujian teks judul, nilai k sama dengan 2-6 memiliki nilai selisih kemiripan 5% – 7%. Berbeda nilai selisih antara nilai *k*=2 dan nilai k=3 memiliki nilai kemiripan yang lebih tinggi yaitu 37.73%. Hal ini disebabkan karena peluang kemiripan *term* dengan panjang antara teks1 dengan teks2 lebih besar. Contoh:

- Rangkaian *k-gram* (*k*=2) dari kata rancang: *ra an nc ca an ng*
- Rangkaian *k-gram* (*k*=2) dari kata bangun: *ba an ng gu un*

Dengan nilai *k*=2, berpeluang menghitung *term* yang sama dari kata yang jauh berbeda seperti contoh di atas, didapatkan 2 *term* yang sama dari kata rancang dan bangun. Ini mempengaruhi nilai kemiripan.

Nilai *k*=2 menghasilkan nilai kemiripan yang lebih tinggi. Oleh karena itu nilai *k*=2 atau nilai *k* yang paling rendah tidak direkomendasikan untuk digunakan.

c) Hasil kemiripan yang didapatkan dengan nilai nilai *k* yang paling tinggi *k*=6 adalah hasil yang memiliki nilai kemiripan yang paling rendah. Fungsi nilai *k* terletak pada pembentukan *term* untuk rangkaian *k-gram*. Pembentukan *term* sepanjang 6 karakter mempengaruhi kata yang kurang dari 6 karakter, karena digabung dengan karakter lainnya pada pembentukan *term*. Pembentukan k-gram dari judul tugas akhir lain yaitu:

- Rangkaian k-gram (k=2) dari kalimat "Basis Web": basisw asiswe sisweb
- Rangkaian k-gram (k=2) dari kalimat "Basis Desktop": basisd asisde sisdes isdesk sdeskt deskto esktop

Hasil pembentukan *k-gram* di atas yaitu "Basis Web" dan "Basis Desktop" terdapat kemiripan yaitu kata "Basis", tetapi setelah pembentukan *term* dengan panjang k=6, berpengaruh terhadap kata "Basis" yang panjangnya hanya 5 karakter. Pada pembentukan *term* ditambahkan karakter lain menjadi "basisw" dan "basisd", sehingga kata basis dianggap tidak mirip. Dari simulasi tersebut diketahui bahwa nilai k untuk pembentukan *k-gram* yang terlalu tinggi berpengaruh kepada kata-kata yang panjang karakternya di bawah nilai *k* yang digunakan.

d) Semakin tinggi nilai dari variabel *w*, maka menghasilkan nilai similaritas yang tinggi dan sebaliknya semakin rendah nilai dari variabel w, maka menghasilkan nilai similaritas yang rendah. Variabel *w* digunakan pada proses pembentukan *window* yang berpengaruh pada pemilihan *fingerprint* dan perhitungan kemiripan. Variabel *w* berpengaruh terhadap banyaknya window yang terbentuk.

IV. Kesimpulan

Semakin tinggi nilai k maka semakin rendah hasil similaritas yang didapatkan dan semakin tinggi nilai w maka semakin tinggi nilai similaritas yang dihasilkan. Variabel *k* lebih berpengaruh pada hasil similaritas dibandingkan dengan variabel *w*, terbukti pada hasil penelitian selisih hasil kemiripan untuk perbedaan nilai k yaitu 5% - 9% sedangkan selisih hasil kemiripan untuk perbedaan nilai *w* yaitu 1% - 3%.

Daftar Pustaka

[1] M. S. Ramli, S. Cokrowibowo, and M. F. Rustan, "Uji Plagiarism pada Tugas Mahasiswa Menggunakan Algoritma Winnowing," *J. Appl. Comput. Sci. Technol.*, vol. 2, no. 2, pp. 108–112, 2021, doi: 10.52158/jacost.v2i2.177.

[2] N. Putra Bayu Pratama, Mustaqiem, and Minarni, "Pengembangan Sistem Informasi Pengelolaan Judul Skripsi dan Tugas Akhir dengan Fitur Deteksi Kemiripan Menggunakan Algoritma Winnowing," *Terap. Inform. Nusant.*, vol. 2, no. 5, pp. 271–278, 2021.

[3] W. A. Negoro, F. Amalia, and E. Santoso, "Pengembangan Aplikasi Resep Masakan dengan Rekomendasi berdasarkan Bahan-Bahan Makanan Berbasis Web," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 9, pp. 9212–9221, 2019, [Online]. Available: https://j-ptiik.ub.ac.id/

[4] H. Y. Pamungkas and Fitrianingsih, "Deteksi Similaritas Dokumen Ilmiah Menggunakan Algoritma Rabin-Karp," *J. Ilm. Inform. Komput.*, vol. 24, no. 3, pp. 209–219, 2019, doi: 10.35760/ik.2019.v24i3.2363.

[5] W. Hidayat, E. Utami, and A. D. Hartanto, "Pemilihan Parameter Terbaik pada Algoritma Winnowing dalam Mendeteksi Tingkat Kesamaan Dokumen Bahasa Indonesia Selection of the Best Parameters in the Winnowing Algorithm in Detecting the Level of Similarity in Indonesian Documents," *Citec J.*, vol. 7, no. 2, 2020.

[6] N. Alamsyah and M. Rasyidan, "Deteksi Plagiarisme Tingkat Kemiripan Judul Skripsi Pada Fakultas Teknologi Informasi Menggunakan Algoritma Winnowing," *Technol. J. Ilm.*, vol. 10, no. 4, p. 197, 2019, doi: 10.31602/tji.v10i4.2361.

[7] I. Widaningrum, D. Mustikasari, R. Arifin, and E. Dyah Cahyani, "Analisa Penggunaan K-Gram pada Karakter, Kata dan Kalimat untuk Mendeteksi Kesamaan Dokumen," *Pros. Semin. Nas. Teknoka*, vol. 5, no. 2502, pp. 59–64, 2020, doi: 10.22236/teknoka.v5i.333.

[8] A. Suroni and E. Setyati, "Klasifikasi Bank Soal Berdasarkan Kompetensi Dasar Menggunakan Hasil Similarity Tertinggi Algoritma Winnowing," *J. Ilm. Tek. Inform.*, vol. 15, no. 2, pp. 163–173, 2021.

[9] R. M. Fauzi and J. C. Wibawa, "Implementasi Algoritma Winnowing untuk Mendeteksi Kemiripan Teks pada Artikel," 2018.

[10] W. Desena and A. Solichin, "Pencarian Abstrak Tugas Akhir Mahasiswa Berdasarkan Tingkat Kemiripan Menggunakan Algoritma Winnowing dan Jaccard Similarity pada Universitas Budi Luhur," *Inform. J. Ilmu Komput.*, vol. 17, no. 2, p. 112, 2021, doi: 10.52958/iftk.v17i2.3628.