

KOMPARASI ALGORITMA WINNOWING DAN ALGORITMA MANBER DALAM MENDETEKSI KEMIRIPAN TUGAS MAHASISWA

Ida Bagus Ketut Surya Arnawa
ITB STIKOM Bali
Jl. Raya Puputan No 86 Renon, Denpasar
E-mail : arnawa@stikom-bali.ac.id

ABSTRAK

Seiring dengan perkembangan teknologi yang begitu pesat, proses belajar mengajar baik yang dilaksanakan disekolah maupun diperguruan tinggi sudah banyak memanfaatkan teknologi informasi untuk menunjang proses belajar mengajar. Institut Teknologi dan Bisnis STIKOM Bali merupakan salah satu perguruan tinggi yang sebagian besar proses belajar mengajar menggunakan teknologi informasi. Dalam melaksanakan perkuliahan online beberapa dosen mengalami permasalahan terutama dalam mengkoreksi tugas yang dikumpulkan oleh mahasiswa. Dalam beberapa kasus terdapat tugas mahasiswa terindikasi hampir sama persis dengan tugas mahasiswa yang lainnya serta beberapa tugas mahasiswa tidak mencantumkan sumber yang digunakan. Untuk mengatasi permasalahan tindakan plagiarisme, perlu mengimplementasikan teknik pencocokan dokumen tugas guna untuk mengetahui tingkat kemiripan dokumen tugas mahasiswa. Banyak algoritma text mining dapat digunakan untuk melakukan pencocokan string. Penelitian sebelumnya penulis telah melakukan perbandingan unjuk kerja algoritma Rabin-Karp, Winnowing, Jaro Winkler dengan hasil algoritma Winnowing memiliki unjuk kerja yang lebih baik diantara algoritma Rabin-Karp dan Jaro Winkler. Dalam penelitian ini penulis membandingkan algoritma Winnowing dengan algoritma Manber dalam melakukan pencocokan dokumen. Hasil dari penelitian ini yaitu algoritma winnowing memiliki unjuk kerja yang lebih baik dari pada algoritma manber.

Kata kunci : Text Mining, Winnowing, Manber

ABSTRACTS

Along with the rapid development of technology, the teaching and learning process, both carried out in schools and universities, has used information technology a lot to support the teaching and learning process. STIKOM Bali Institute of Technology and Business is one of the universities where most of the teaching and learning process uses information technology. In carrying out online lectures, several lecturers experienced problems, especially in correcting assignments collected by students. In some cases, student assignments are indicated to be almost exactly the same as other student assignments and some student assignments do not include the sources used. To overcome the problem of plagiarism, it is necessary to implement the matching technique of assignment documents in order to determine the level of similarity of student assignment documents. Many text mining algorithms can be used to perform string matching. In previous research, the author has compared the performance of the Rabin-Karp, Winnowing, Jaro Winkler algorithms with the results that the Winnowing algorithm has better performance between the Rabin-Karp and Jaro Winkler algorithms. In this study, the authors compare the Winnowing algorithm with the Manber algorithm in document matching. The result of this research is that the winnowing algorithm has better performance than the manber algorithm.

Keywords: Text Mining, Winnowing, Manber

1. PENDAHULUAN

Manusia merupakan satu satunya makhluk ciptaan Tuhan yang memiliki akal dan pikiran. Hal yang membedakan manusia dengan ciptaan Tuhan yang lainnya yaitu manusia diberikan pikiran, akal, perasaan untuk meningkatkan kualitas hidup. Dalam meningkatkan kualitas hidup manusia tidak akan pernah berhenti untuk belajar, baik secara formal maupun informal. Dengan belajar memberikan beberapa manfaat diantaranya otak akan tetap sehat, menambah pengetahuan dan pengalaman serta dapat meningkatkan kualitas hidup. Indonesia merupakan salah satu negara yang mewajibkan warga negaranya untuk belajar selama 9 tahun di sekolah. Hal ini tertuang dalam Peraturan Pemerintah Republik Indonesia Nomor 47 Tahun 2008 Tentang Wajib Belajar. Seiring dengan perkembangan teknologi yang begitu pesat, proses belajar mengajar baik yang dilaksanakan disekolah maupun diperguruan tinggi sudah banyak memanfaatkan teknologi informasi untuk menunjang proses belajar mengajar.

Institut Teknologi dan Bisnis STIKOM Bali merupakan salah satu perguruan tinggi yang sebagian besar proses belajar mengajar menggunakan teknologi informasi. Saat ini Institut Teknologi dan Bisnis STIKOM Bali memiliki beberapa sistem yaitu SION (Sistem informasi online) yang diperuntukan bagi mahasiswa untuk melihat berbagai informasi mengenai perkuliahan seperti jadwal kuliah, hasil studi, indeks prestasi semester, indeks prestasi kumulatif serta informasi-informasi lainnya yang terkait dengan perkuliahan. Selain SION Institut Teknologi dan Bisnis juga memiliki sistem E-learning dimana sistem ini diperuntukan bagi mahasiswa dan dosen. Melalui sistem ini perkuliahan secara online dilaksanakan, dosen dapat memberikan materi perkuliahan, forum diskusi, pengumuman serta memberikan tugas kuliah. Mahasiswa dapat mengunduh materi perkuliahan yang diberikan oleh dosen, ikut

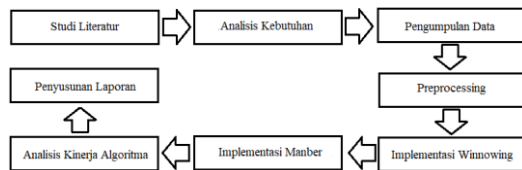
serta dalam forum diskusi serta mahasiswa dapat mengumpulkan tugas yang diberikan oleh dosen.

Dalam melaksanakan perkuliahan online beberapa dosen mengalami permasalahan terutama dalam mengoreksi tugas yang dikumpulkan oleh mahasiswa. Dalam beberapa kasus terdapat tugas mahasiswa terindikasi hampir sama persis dengan tugas mahasiswa yang lainnya serta beberapa tugas mahasiswa tidak mencantumkan sumber yang digunakan. Hal ini sudah termasuk tindakan plagiarisme. Plagiarisme merupakan tindakan yang tidak terpuji dimana seseorang dengan sengaja atau tidak mengambil ide/gagasan/karya orang lain tanpa seijin yang memiliki ide/gagasan/karya. Ada beberapa faktor yang mendorong mahasiswa melakukan tindakan plagiarisme diantaranya kurangnya ide dalam membuat tugas, kurangnya rasa percaya diri yang dimiliki, serta kurangnya pemahaman tentang plagiarisme dan undang-undang hak cipta[1].

Untuk mengatasi permasalahan tindakan plagiarisme, perlu mengimplementasikan teknik pencocokan dokumen tugas guna untuk mengetahui tingkat kemiripan dokumen tugas mahasiswa. Banyak algoritma text mining dapat digunakan untuk melakukan pencocokan string. Penelitian sebelumnya penulis telah melakukan perbandingan unjuk kerja algoritma Rabin-Karp, Winnowing, Jaro Winkler dengan hasil algoritma Winnowing memiliki unjuk kerja yang lebih baik diantara algoritma Rabin-Karp dan Jaro Winkler. Merujuk dari uraian diatas penulis tertarik untuk mencari algoritma yang memiliki unjuk kerja yang lebih baik dengan melakukan perbandingan algoritma Winnowing dimana algoritma ini dalam penelitian sebelumnya memiliki unjuk kerja yang paling baik. Dalam penelitian ini penulis akan membandingkan algoritma Winnowing dengan algoritma Manber dalam melakukan pencocokan dokumen.

2.METODE PENELITIAN

Adapun alur penelitian yang digunakan dalam penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Alur Penelitian

Gambar 1. merupakan alur penelitian yang digunakan dalam penelitian ini. Tahap pertama diawali dengan melakukan studi literatur terkait bidang penelitian yang dibuat yaitu text mining. Tahap kedua melakukan analisis kebutuhan yang diperlukan dalam melakukan penelitian dan dilanjutkan dengan pengumpulan data, dimana data yang digunakan bersumber dari tugas mahasiswa ITB STIKOM Bali. Tugas mahasiswa tersebut akan melalui proses preprocessing seperti case folding / lowercase, tokenizing / pemotongan kata, filtering / membuang kata kurang penting. Setelah proses preprocessing dilakukan maka tahap selanjutnya dilakukan implementasi algoritma winnowing dan algoritma manber dalam mengukur tingkat kemiripan tugas mahasiswa. Hasil dari masing-masing algoritma akan dianalisis dan dibandingkan untuk mengetahui algoritma yang memiliki kinerja yang paling baik.

2.1. Algoritma Winnowing

Algoritma winnowing merupakan salah satu algoritma yang dapat digunakan untuk mengukur tingkat kesamaan kata (document fingerprinting) untuk mendeteksi plagiarisme. Rolling hash merupakan algoritma yang digunakan mencari nilai hash dalam winnowing. Nilai hash merupakan nilai numerik yang terbentuk dari perhitungan ASCII setiap karakter. Langkah-langkah dalam penerapan Algoritma Winnowing adalah sebagai berikut [2]:

1. Langkah Pertama

Penghapusan karakter yang tidak relevan yaitu berupa spasi, tanda baca serta spesial karakter seperti !, @, #, \$, %, ^, &, *, (,), _ , - ?

Contoh : Belajar Pemrograman C# belajarpemrogramanc

Akan dirubah menjadi

3.Langkah Kedua

Pembentukan Rangkaian n-gram yaitu dengan cara membentuk rangkaian karakter sepanjang n dari hasil pembuangan karakter yang tidak relevan pada langkah pertama. Dari text diatas telah dibersihkan dengan ukuran k=5

belaj elaja lajar ajarp jarpe arpem
rpemr pemro emrog mrogr rogra
ogram grama raman amanc

4.Langkah Ketiga

Menghitung fungsi hash untuk setiap n-gram yaitu dengan cara menghitung nilai - nilai hash dari setiap gram, algoritma holling hash digunakan untuk menghitung nilai hash dalam algoritma winnowing. Rolling hash merupakan suatu teknik untuk mentransformasikan sebuah string menjadi nilai yang unik dengan panjang tertentu yang berfungsi sebagai penanda string tersebut. Fungsi tersebut menghasilkan nilai ini disebut fungsi hash, sedangkan nilai yang dihasilkan disebut nilai hash. Fungsi hash $H(c1...ck)$ didefinisikan sebagai berikut :

$$H(ck) = c1 * b^{(k-1)} + c2 * b^{(k-2)} + ... + ck * b^{(k-k)}$$

Keterangan :

c = nilai ascii karakter

b = basis (bilangan prima)

k = banyak karakter

hasil rolling hash dari kalimat diatas yaitu

1583490 1635489 1724342 1574366
1696180 1586683 1831680 1788777
1638938 1762281 1830629 1777214
1672790 1812547 1578302

5.Langkah Keempat

Pembentukan window dari nilai hash dari window dengan ukuran 3 yaitu sebagai berikut :

{1583490 1635489 1724342}
 {1635489 1724342 1574366}
 {1724342 1574366 1696180}
 {1574366 1696180 1586683}
 {1696180 1586683 1831680}
 {1586683 1831680 1788777}
 {1831680 1788777 1638938}
 {1788777 1638938 1762281}
 {1638938 1762281 1830629}
 {1762281 1830629 1777214}
 {1830629 1777214 1672790}
 {1777214 1672790 1812547}
 { 1672790 1812547 1578302 }

6. Langkah Kelima

Langkah terakhir yaitu memilih nilai terkecil dari setiap window untuk dijadikan fingerprint, hasil dari nilai fingerprintnya sebagai berikut:

[1583490,0] [1574366,3] [1586683,5]
 [1638938,8] [1762281,9] [1672790,12]
 [1578302,14]

Nilai fingerprint yang dibentuk dari algoritma winnowing digunakan untuk mengukur prosentase kemiripan teks pada persamaan Jaccard Coefficient. Persamaan Jaccard Coefficient digunakan untuk menghitung kemiripan (similarity) dari kumpulan kata-kata yang telah dihitung nilai hash nya. Berikut ini rumus persamaan Jaccard Coefficient.

$$\text{Similarity} = \frac{\text{Jumlah fingerprint sama}}{\text{Total seluruh fingerprint}} \times 100$$

6.1. Algoritma Manber

Proses Algoritma Manber hampir sama dengan Algoritma Winnowing, yang membedakan adalah Algoritma winnowing menentukan posisi dari fingerprint, sedangkan Algoritma Manber tidak menampilkan informasi posisi fingerprint [2][3].

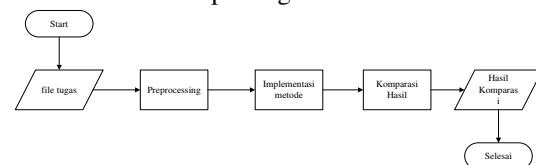
1. Penghapusan karakter-karakter yang tidak relevan (whitespace insensitivity), antara lain spasi atau tanda baca.
2. Pembentukan rangkaian gram dengan ukuran k.

3. Penghitungan nilai hash dari setiap gram.
 4. Membagi ke dalam window tertentu.
 5. Pemilihan beberapa nilai hash menjadi document fingerprinting
 6. Prosentase kesamaan antara 2 dokumen dengan persamaan Jaccard Coefficient
- Pendekatan Manber memilih hasil dari proses hashing dengan cara memilih semua hasil hashing dengan yang memenuhi kriteria 0 mod p. Dengan cara ini fingerprints yang terpilih tidak tergantung dari posisinya. Pendekatan ini mudah untuk diimplementasikan.

7. HASIL DAN PEMBAHASAN

7.1. Perancangan Sistem

Perancangan sistem yang dibuat merupakan sistem untuk mendeteksi kemiripan tugas mahasiswa satu dengan yang lainnya. Sistem ini dapat menerima inputan berupa text yang merupakan tugas dari mahasiswa. Tahap awal sistem melakukan preprocessing terlebih dahulu terhadap tugas mahasiswa. Setelah melalui tahapan preprocessing tugas mahasiswa akan dicocokkan dengan tugas lainnya dengan menggunakan masing-masing metode yaitu Winnowing dan Manber. Masing-masing metode akan menghasilkan persentase kemiripan tugas mahasiswa yang satu dengan yang lainnya. Kemudian dari hasil tersebut dianalisis untuk mengetahui metode mana yang memiliki unjuk kerja yang paling baik. (Gipp, B. and Meuschke, N. 2011). Gambar 2 merupakan gambaran arsitektur sistem komparasi algoritma text mining dalam mendeteksi kemiripan tugas mahasiswa.

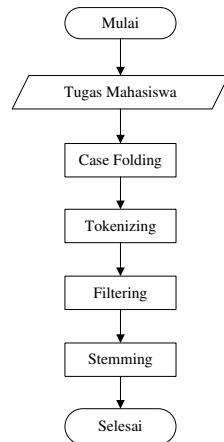


Gambar. 2 Arsitektur Sistem

7.2. Flowchart Preprocessing

Pada tahapan preprocessing dilakukan beberapa tahapan yaitu diawali dengan case folding yaitu mengkonversi seluruh huruf dalam tugas

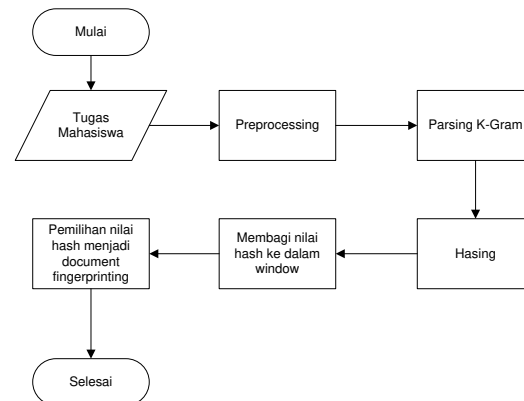
mahasiswa menjadi huruf kecil. Kemudian dilanjutkan dengan tahap tokenizing yaitu proses memisahkan setiap kata yang menyusun dokumen tugas mahasiswa dengan memanfaatkan karakter spasi. Tahap selanjutnya yaitu filtering dimana tahap ini akan mengambil kata yang penting dari tahap tokenizing. Tahap terakhir yaitu stemming dimana pada tahap ini mengubah sebuah kata menjadi kata dasar dengan menghilangkan imbuhan. Gambar 3 merupakan flowchart preprocessing [4][5].



Gambar.3 Flowchat Preprocessing

7.3. Flowchart Winnowing

Pada Algoritma Winnowing setelah melakukan tahap preprocessing dilanjutkan dengan membentuk teks hasil langkah preprocessing ke dalam rangkaian k-gram. Selanjutnya dilakukan proses rolling hash untuk menghasilkan nilai hash dari setiap gram yang terbentuk. Langkah selanjutnya setelah mendapat nilai hash dari setiap gram dilanjutkan dengan membentuk window, proses pembentukan window sama seperti proses k-gram dari nilai-nilai yang dihasilkan dengan besar window = 3



Gambar 4. Flowchart Winnowing

7.4. Implementasi Perbandingan Algoritma

Berikut ini merupakan tahapan-tahapan dalam mengimplementasikan perbandingan algoritma winnowing dengan algoritma manber untuk mendeteksi presentase plagiarisme pada tugas mahasiswa yaitu :

Tabel 1. Data Uji Coba

Percobaan 1. Tugas Mahasiswa dengan tingkat kemiripan 100 %	
Bahasa pemrograman ini merupakan suatu himpunan dari aturan sintaks dan semantik yang dipakai untuk mendefinisikan program komputer. Bahasa ini memungkinkan seorang programmer dapat menentukan secara persis data mana yang akan diolah oleh komputer, bagaimana data ini akan disimpan/diteruskan, dan jenis langkah apa yang akan diambil dalam berbagai situasi secara persis.	Bahasa pemrograman ini merupakan suatu himpunan dari aturan sintaks dan semantik yang dipakai untuk mendefinisikan program komputer. Bahasa ini memungkinkan seorang programmer dapat menentukan secara persis data mana yang akan diolah oleh komputer, bagaimana data ini akan disimpan/diteruskan, dan jenis langkah apa yang akan diambil

	dalam berbagai situasi secara persis.
Percobaan 2. Tugas Mahasiswa dengan tingkat kemiripan 50 %	
Bahasa pemrograman ini merupakan suatu himpunan dari aturan sintaks dan semantik yang dipakai untuk mendefinisikan program komputer. Bahasa ini memungkinkan seorang programmer dapat menentukan secara persis data mana yang akan diolah oleh komputer, bagaimana data ini akan disimpan/diteruskan, dan jenis langkah apa yang akan diambil dalam berbagai situasi secara persis.	Bahasa pemrograman adalah himpunan perintah sintaks semantik yang dipakai membuat program komputer. programmer dapat menentukan data yang akan diolah komputer untuk membuat program komputer. Ada beberapa Bahasa pemrograman yang sering digunakan diantaranya javascript, php, c++, asp.net dan masih banyak lagi yang lainnya. Setiap bahasa pemrograman memiliki kekurangan dan kelebihan masing-masing.
Percobaan 3. Tugas Mahasiswa dengan tingkat kemiripan 30 %	
Bahasa pemrograman ini merupakan suatu himpunan dari aturan sintaks dan semantik yang dipakai untuk mendefinisikan program komputer. Bahasa ini memungkinkan seorang programmer dapat menentukan secara persis data mana yang akan	Kemampuan seorang programmer bisa didapat dari pendidikan formal atau informal. Programmer membuat program komputer dengan bahasa pemrograman. Ada beberapa Bahasa pemrograman yang sering digunakan

diolah oleh komputer, bagaimana data ini akan disimpan/diteruskan, dan jenis langkah apa yang akan diambil dalam berbagai situasi secara persis.	diantaranya javascript, php, c++, asp.net dan masih banyak lagi yang lainnya. Setiap bahasa pemrograman memiliki kekurangan dan kelebihan masing-masing.
--	--

Tabel.2 Percobaan 1. Tingkat kemiripan 100 % dengan winnowing

uji coba	n-gram (n)	window (w)	kemiripan (%)
1	2	3	100
2	3	3	100
3	4	3	100
4	5	3	100
5	6	3	100
6	7	5	100
7	8	5	100
8	9	5	100
9	10	5	100

Tabel.3 Percobaan 2. Tingkat kemiripan 50 % dengan winnowing

uji coba	n-gram (n)	window (w)	kemiripan (%)
1	2	3	47.3
2	3	3	33.3
3	4	3	19.5
4	5	3	19.5
5	6	3	16.0
6	7	5	13.7
7	8	5	11.1
8	9	5	9.18
9	10	5	8.02

Tabel.4 Percobaan 3. Tingkat kemiripan 30 % dengan winnowing

uji coba	n-gram (n)	window (w)	kemiripan (%)
1	2	3	38.9
2	3	3	23.6
3	4	3	12.0
4	5	3	10.7
5	6	3	7.81
6	7	5	6.93
7	8	5	5.68
8	9	5	5.05
9	10	5	5.05

Tabel.5 Percobaan 1.Tingkat kemiripan 100 % dengan manber

uji coba	n-gram (n)	kemiripan (%)
1	2	100
2	3	100
3	4	100
4	5	100
5	6	100
6	7	100
7	8	100
8	9	100
9	10	100

Tabel.6 Percobaan 2.Tingkat kemiripan 50 % dengan manber

uji coba	n-gram (n)	kemiripan (%)
1	2	45.2
2	3	32.1
3	4	17.6
4	5	17.5
5	6	15.4
6	7	11.7
7	8	10.3
8	9	8.08
9	10	7.01

Tabel.7 Percobaan 3.Tingkat kemiripan 30 % dengan manber

uji coba	n-gram (n)	kemiripan (%)
1	2	45.2
2	3	32.1
3	4	17.6
4	5	17.5
5	6	15.4
6	7	11.7
7	8	10.3
8	9	8.08
9	10	7.01

Tabel.8 Hasil Perbandingan algoritma winnowing dan manber

Tingkat kemiripan (%)	Hasil Winnowing (%)	Hasil Manber (%)
100	100	100
50	47.3	45.2
30	23.6	39.8

KNOWN	PREDICTED	
	Masker	TanpaMasker
Masker	0.96	0.04
TanpaMasker	0.26	0.74
* Average Accuracy is 0.85.		

Gambar 2. Tingkat Akurasi Pengujian

4.

KESIMPULAN

Berdasarkan hasil penelitian yang dilakukan pada proses pelatihan maka dapat disimpulkan bahwa, metode bag of visual word dapat melakukan proses klasifikasi pada penggunaan masker. Proses yang dilakukan adalah mendeteksi citra input menggunakan metode SURF untuk mendapatkan nilai interest point, kemudian dikelompokkan dengan metode K-Means. Hasil dari k-means dibuatkan sebuah matriks yang menampung nilai histogram pada masing-masing citra input. Hasil tersebut dilakukan klasifikasi menggunakan metode SVM untuk mendapatkan model proses pelatihan. Pada proses pelatihan metode bag of visual Words dapat memberikan hasil yang baik dimana tingkat akurasi yang diperoleh mencapai 94% dan pada proses pengujian dapat diperoleh tingkat akurasi sebesar 85%. Adapun saran yang dikemukakan peneliti pada penelitian ini adalah mengambil sampel lebih banyak lagi dalam proses pelatihan dan diperlukan metode pembandingan untuk mengetahui tingkat keberhasilan antara masing-masing metode yang digunakan.

Berdasarkan hasil percobaan yang telah dilakukan pada masing-masing algoritma yaitu algoritma winnowing dengan algoritma manber terhadap tugas mahasiswa, maka dapat ditarik kesimpulan sebagai berikut;

1. Algoritma winnowing memiliki kinerja yang lebih baik dibandingkan dengan algoritma manber.
2. Terdapat perbedaan mendasar dari kedua algoritma yaitu pada algoritma winnowing

memiliki window dan k-gram sedangkan algoritma manber hanya memiliki k-gram.

DAFTAR

PUSTAKA

- [1] Wali, M., & Safrizal, S. (2018). Similar text sebagai Pengkodean Aplikasi Plagiarisme. *Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi)*, 2(1), 11-19.
- [2] Astutik, Sariyanti, et al. "Sistem Penilaian Esai Otomatis Pada E-learning Dengan Algoritma Winnowing." *Jurnal Informatika University Petra Kristian*, vol. 12, no. 2, Nov. 2014, pp. 47-52
- [3] Parewe, Andi Maulidinnawati Ak; Aman, Andryanto; Dewang, Dody Putra M. Perbandingan Algoritma Winnowing Dan Algoritma Manber Dalam Mendeteksi Berita Hoax Di Media Sosial. *PROSIDING SEMANTIK*, 2021, 3.1: 41-46.
- [4] Andhika, Fatardhi R. 2010. Penerapan String Suggestion dengan Algoritma Levenshtein Distance dan Alternatif Algoritma Lain dalam Aplikasi, Sekolah Tinggi Teknik Informatika, Institut Teknologi Bandung (ITB). Bandung.
- [5] Setiawan, A. (2017). Implementasi Algoritma Winnowing Untuk Deteksi Kemiripan Judul Skripsi Studi Kasus Stmik Budidarma. *Majalah Ilmiah INTI (Informasi dan Teknologi Ilmiah)*, 12(1).