TRANSFORMTIKA, Vol.18, No.1, July 2020, pp. 63 - 70
ISSN: 1693-3656,
journals.usm.ac.id/index.php/transformatika

Sensitivitas Sistem Pencarian Artikel Bahasa Indonesia Menggunakan Metode n-gram Dan Tanimoto Cosine

Candra Supriadi[1], Hidriyanto Dwi Purnomo[2], Irwan Sembiring[3]
[1]Magister Sistem Informasi,Fakultas Teknologi Informasi, Universitas Kristen Satya Wacana, Salatiga,email: mascandraduand@gmail.com
Jln. Dr. O. Notohamidjojo Blotongan Sidorejo, Kota Salatiga, 50715, Indonesia
[2]Magister Sistem Informasi, Fakultas Teknologi Informasi, Universitas Kristen Satya Wacana, Salatiga, email: hindriyanto@uksw.edu
Jln. Dr. O. Notohamidjojo Blotongan Sidorejo, Kota Salatiga, 50715, Indonesia,
[3]Magister Sistem Informasi, Fakultas Teknologi Informasi, Universitas Kristen Satya Wacana, Salatiga, email: irwan@uksw.edu
Jln. Dr. O. Notohamidjojo Blotongan Sidorejo, Kota Salatiga, 50715, Indonesia,

ARTICLE INFO

Article history:

Received 13 April 2020
Received in revised form 21 Arpil 2020
Accepted 26 April 2020
Available online 8 July 2020

ABSTRACT

The human need for technology and the availability of adequate infrastructure is evidence that technology is now a part of basic human needs. Increasing number of journals and scientific papers, it must be more selective in selecting and sorting even though there are already many online service providers and journal portals. Researcher that used search engines and plagiarism and recommendation systems has been carried out with various methods deemed appropriate to improve the performance of the system itself, this paper has the purpose of calculating the similarity between one article with another article by implementing n-gram and tanimoto cosine. The number of articles tested was forty-three titles and abstracts, tested fifty times with randomly selected keywords, by breaking down each title and abstract sentence into n characters (n = 2 to 8) including spaces and punctuation, then counted similarity with the query or keyword used for system testing. The Result of the Ngram and Tanimoto Cosine Experiments was using several threshold variations from n = 2 to 8. After observing fifty times the threshold test of **0.15** has the highest accuracy at **n = 4** at **0.92,** the highest precision at n = 3 at **0.42** and the highest recall at the test **n = 2** = **0.44**. From the results of some of the above tests it can be concluded that for solving words using the n gram algorithm and tanimoto cosine can shorten the time in searching online-based Indonesian articles quickly

Keywords: Search Engine, Tanimoto cosine, N-gram

Received April 13, 2020; Revised April 29, 2020; Accepted July 12, 2020

1. Introduction

Perkembangan teknologi memiliki dampak yang sangat signifikan dalam kehidupan sehari-hari, mulai dari kegiatan yang sederhana hingga kegiatan yang membutuhkan tingkat ketelitian yang tinggi. Kegiatan yang umum dilakukan oleh sebuah instansi adalah kegiatan pengarsipan dokumen, baik dokumen dalam bentuk fisik maupun elektronik. Umumnya kegiatan pengarsipan melibatkan dokumen dengan jumlah yang cukup besar, sehingga diperlukan suatu metode yang praktis dan efisien dalam pengelolaanya. Salah satu metode yang digunakan dalam pengelolaan dokumen adalah pengklasteran atau pengklasifikasian dokumen. Pencarian suatu dokumen dalam kumpulan dokumen yang sesuai dengan kebutuhan bukan hal yang mudah untuk dilakukan. Pengguna harus mencari satu persatu, membaca setiap dokumen, dan menganalisis apakah dokumen tersebut sesuai dengan yang dibutuhkan atau tidak [1]. Untuk melakukan semuanya itu membutuhkan waktu yang lama dan tidak efisien. Apalagi jika terdapat jumlah dokumen yang sangat banyak. *Information Retrieval* merupakan proses pemisahan dokumen dari sekumpulan dokumen untuk menentukan dokumen mana yang harus diambil agar dapat memenuhi kebutuhan user akan informasi[2]. Algoritma Ranking digunakan *dalam Information Retrieval* agar menghasilkan urutan dokumen yang relevan dengan kebutuhan user. Karena itu dibutuhkan suatu sistem untuk melakukan pencarian suatu dokumen dalam sekumpulan dokumen menggunakan kata kunci (keyword) yang sesuai dengan kebutuhan user dengan menggunakan algoritma *n gram* dan *Tanimoto Cosine*. Penelitian ini mengimplementasikan n-gram pada dibatasi pada n=2 sampai n=3 pada level pencarian. Untuk penentuan threshold yang diterapkan pada sistem tersebut yaitu 0.5, 0.3 dan 0.15serta teks sumber yang akan diuji hanya diambil abstrak saja pada tiap tiap artikel untuk metode pengoreksian atau pembagian n-gram tidak memperhatikan grammar, struktur kata, maupun tanda baca[3].

2. Research Method

a. N-Gram

N-Gram adalah serangkaian n-item yang berurutan dari sebuah data, biasanya berupa teks. Nilai N tersebut bisa bervariasi tergantung dari kebutuhan, mulai dari satu hingga sepanjang teks yang ada. Posisi *n-gram* berikutnya dihitung dari posisi yang sebenarnya bergeser sesuai dengan *offset* yang diberikan[4]. Nilai *offset* bergantung pada pembagian yang digunakan dalam *n-gram*. Pembagian *n-gram* dapat bervariasi tergantung dari pendekatan dalam membagi teks menjadi bentuk *n-gram*. *N-gram* untuk setiap *string* dihitung dan kemudian dibandingkan satu per satu [5]. Ngram dapat berupa *unigram* (n=1), *bigram* (n=2), *trigram* (n=3), dan seterusnya.sebagai contoh penerapan n-gram, sebuah susunan kata "SAYA MAKAN", maka *n-gram* yang dihasilkan dapat dilihat pada Tabel. 2.1. Contoh N-Gram Dari Susunan Kata "SAYA MAKAN" berikut ini:

Tabel. 1 Contoh N-Gram Dari Susunan Kata "SAYA MAKAN"

Susunan Kata	n	n-gram yang dihasilkan
SAYA MAKAN	2	[SA], [AY], [YA],[A(spasi)], [(spasi)M], [MA], [AK], [KA], [AN]
	3	[SAY], [AYA], [YA(spasi)], [A(spasi)M], [(spasi)MA], [MAK], [AKA], [KAN]
	5	[SAYA(spasi)], [AYA(spasi)M], [YA(spasi)MA], [A(spasi)MAK], [(spasi)MAKA], [MAKAN]
	7	[SAYA(spasi)MA], [AYA(spasi)MAK], [YA(spasi)MAKA], [A(spasi)MAKAN]

b. Tanimoto Cosine

Dalam penelitian ini proses searcing mengimplementasikan metode *tanimoto cosine. Tanimoto cosine* merupakan gabungan antara *tanimoo similarity* dan *consine similarity*, *cosine similarity dan tanimoto similarity* dapat dilihat pada persamaan berikut ini [5]:

$$\sum\left(\frac{(s_i)(w_i+w_i')}{2}\right) \qquad \text{..... (1)}$$

CosSim(I,J) merupakan *cosine similarity* dari *subtree* i dan j dengan jumlah vektor k. ai dan aj adalah bobot dari vektor term yang sama dalam *subtree* I dan J [6]. Perhitungan *tanimoto similarity* dapat dilihat pada persamaan berikut:

$$T(I,J) = \frac{\sum_k a_i a_j}{\left(\sqrt{\sum_k a_i^2}\right)+\sqrt{\sum_k a_j^2}-\sum_k a_i a_j} \qquad \text{.....(2)}$$

T(I, J) adalah *tanimoto similarity* dari subtree I dan J dengan jumlah vektor k. ai dan aj adalah bobot dari vektor term yang sama dalam subtree I dan J[7]. Perhitungan *Tanimoto Cosine* dapat dirumuskan sebagai berikut:

$$TC(Z_i, Z_j) = T\,(Z_i, Z_j\,) \text{ x } CosSim(Z_i,Z_j) \quad \text{.....(3)}$$

Dalam persamaan tersebut T(Zi, Zj) adalah perhitungan *similarity* menggunakan *tanimoto similarity* dan CosSim(Zi, Zj) adalah perhitungan *similarity* menggunakan *cosine similarity*. Penggabungan keduanya terbukti lebih akurat [8].

c. *Confusion matrix*

Confusion matrix adalah alat *(tools)* visualisasi yang biasa digunakan pada *supervised learning*. Pada tiap kolom pada matrix adalah contoh kelas prediksi, sedangkan tiap baris mewakili kejadian dikelas yang sebenarnya [1]. *Confusion matrix* digunakan untuk mempermudah penguji dalam mencari nilai akurasi presisi dan recall pada pengujian ini[4]. Pada Tabel 2. *Confusion Matrix* adalah contoh *confusion matrik* yang menujukan 2 kelas (Prediksi dan Aktual)

Tabel 2.Confusion Matrix

	RELEVAN	*TIDAK RELEVAN*
RETRIEVE	TP	FP
NOT RETREIVE	FN	TN

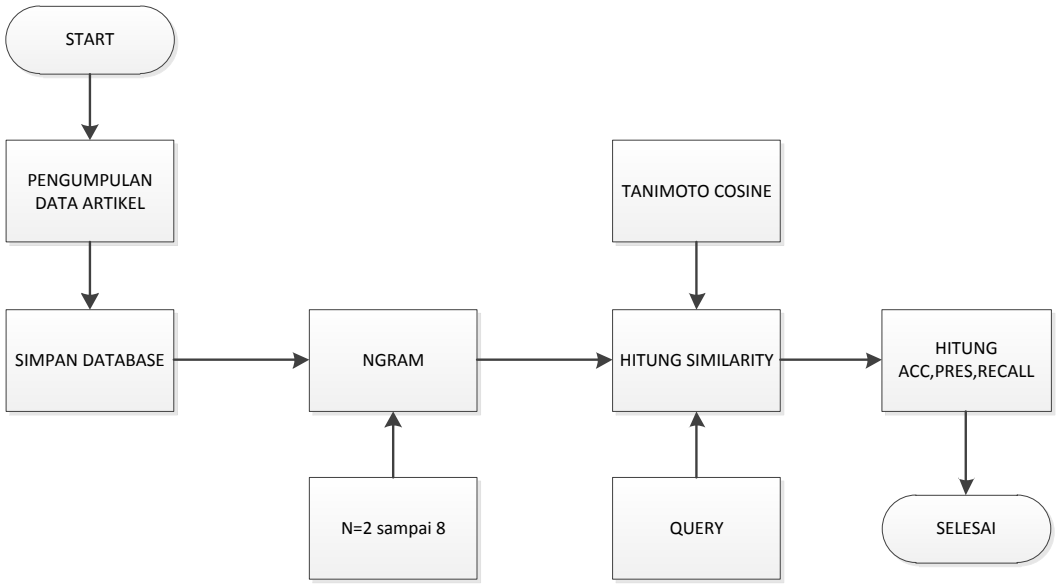
Keterangan:

TP *(True Positif)* = Sama dengan kata kunci dan Muncul
FP (*False Positif*) = Tidak sama dengan kata kunci Tapi muncul
FN (*False Negatif)* = Tidak Relevan dengan kata kunci dan muncul
TN (*True Negatif)* = Relevan dengan kata kunci tapi tidak muncul

Dari Tabel 2 *Confusion Matrix* maka dapat dihitung nilai Akurasi, Presisi dan Recall pada pengujian[9].

- Akurasi adalah perbandingan kasus yang diidentifikasi benar dengan total semua kasus. Nilai persamaannya adalah Akurasi = *(TP=TN)/(TP+FP+FN+TN)*
- Presisi adalah proporsi kasus dengan hasil positif yang benar, dengan maksud lain presisi merupakan kualitas pada sistem yang akan diujikan. Nilai persamaannya adalah Presisi = *TP/(TP+FP)*
- Recall proporsi kasus positif yang diidentifikasi dengan benar. Dengan maksud lain recall merupakan kualitas hasil relevan yang disajikan dalam sistem tersebut. Nilai persamaannya adalah Recall = *TP/(TP+FN)*

d. Tahap Penelitian



Gambar 1 Alur Penelitian

Perancangan sistem dimulai dengan pengumpulan data artikel beserta abstrak yang dipilih secara acak dengan jumlah total empat puluh tiga judul dan abstrak, dan lima puluh kata kunci yang juga dipilih secara acak. Kemudian dari data tersebut, judul dan abstrak langsung dikelompokkan kedalam tiga kategori yang kemudian disimpan kedalam database[10]. Langkah selanjutnya yaitu implementasi n-gram, yang dalam penelitian ini menggunakan n=2 sampai 8, semua data yang ada didalam database dipecah kedalam n-karakter termasuk tanda baca, titik dan koma. Setelah semua data dipecah kedalam n-karakter, kemudian dicocokkan menggunakan algoritma *tanimoto coefficient* yang selanjutnya bisa diketahui nilai *similarity* antara data artikel dengan kata kunci. Nilai similarity naik turun sesuai dengan seberapa besar jumlah kemiripan n-karakter dengan kata kunci. Pengujian sistem menggunakan beberapa variasi *akurasi presisi dan recall* yang kemudian dibandingkan dari beberapa variasi n yang lain yaitu 2 sampai 8 sehingga bisa ditarik kesimpulan[11].

3. Results and Analysis

3.1 *Teknik Analisa Data*

Perhitungan yang akan dilakukan yaitu dengan membagi *n-gram*, n=2, n=3, n=4, n=5 n=6, n=7 dan n=8. Dalam penelitian ini menggunakan 3 threshold yaitu 0.5, 0.3 dan 0.15. Dalam pengujian digunakan 50 percobaan untuk tiap tiap n-gram. penggunaan threshold ini bertujuan untuk membatasi hasil yang akan muncul dalam pengujian sistem tersebut. Tabel 3 Daftar Kata Kunci yang dipakai dalam pengujian ini Berikut ini adalah kata kunci yang akan diujikan pada sistem.

Tabel 3 Daftar Kata Kunci

No	Query	No	Query
1	usaha bisnis desain grafis	26	rumus akurasi presisi recall
2	perancangan sistem informasi	27	sistem penjualan tunai dan kredit
3	dasar desain grafis	28	sistem informasi penjualan
4	analisa keuangan	29	perancagan sistem pengelolaan keuangan
5	aplikasi perkantoran	30	analisa desain interior dan eksterior
6	kelebihan penggunaan sistem	31	analisa faktor yang mempengaruhi kinerja
.....	.....	....	.....
25	standar kemiringan atap	50	pengaruh gaji terhadap kinerja karyawan

Dari ke 50 pengujian yang dilakukan pada tiap-tiap n-gram maka akan didapatkan hasil sebagai berikut:

Tabel 4. Tabel Uji coba pada n=2

		usaha bisnis desain g...	ngan sistem informa	dasar desain grafis	analisa keuangan	aplikasi perkantoran	an penggunaan	n pendapatan	perancangan animasi	.....	terhadap kinerja
no	Judul	1	2	3	4	5	6	7	8	.....	50
1	usaha bisnis desain grafis	1,09	0,09	0,52	0,09	0,03	0,12	0,15	0,09	.....	0,08
2	peluang bisnis desain grafis	0,70	0,21	0,47	0,21	0,10	0,21	0,25	0,26	.....	0.09
3	desain grafis	0,65	0,12	0,72	0,08	0,03	0,09	0,13	0,08	.....	0,04
4	bisnis desain grafis dan pemasaran	0,65	0,27	0,63	0,14	0,16	0,21	0,34	0,29	.....	0,13
5	peluang bisnis di bidang desain grafis	0,62	0,19	0,45	0,18	0,12	0,19	0,28	0,23	.....	0,12
.... .	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
43	peluang bisnis desain grafis dan percetakan	0,54	0,20	0,39	0,16	0,17	0,20	0,33	0,23	.....	0,15

Untuk mencari nilai akurasi presisi dan recall dari pengujian tersebut menggunakan confusion matrix, maka akan didapat data sebagai berikut. Perhitungan nilai akurasi, presisi dan recal dapat dilihat pada tabel 5 Uji coba n=2 dan *threshold* 0.5

Tabel 5 Uji coba n=2 dan *threshold* 0.5

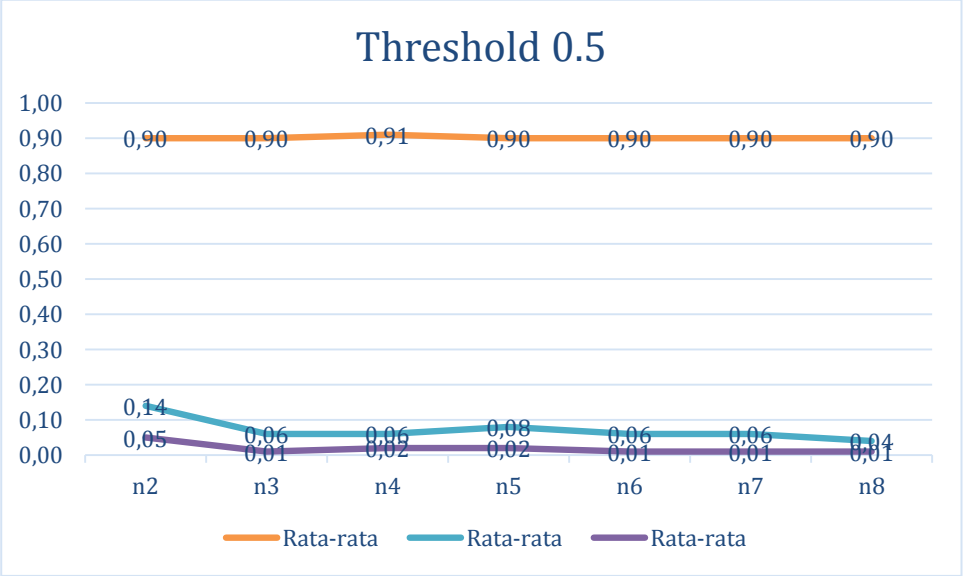
	RELEVAN	NOT RELEVAN		
RETREIVE	7	0		Akurasi = 0.90
NOT RETREIVE	3	21		Presisi = 1
				Recall = 0.70

Dari hasil pengujian sebanyak 50 kali untuk setiap n-gram (n=2 s/d n=8) maka akan didapat hasil rata-rata dari pengujian tersebut dengan threshold yang ditentukan. Tabel 6 merupakan perhitungan akurasi presisi dan recall pada tiap-tiap percobaan.

Tabel 6 Tabel Rata-rata perhitungan.

n-gram	**Rata-rata**								
	Threshold 0.5			Threshold 0.3			Threshold 0,15		
	Akurasi	Presisi	Recall	Akurasi	Presisi	Recall	Akurasi	Presisi	Recall
n=2	0.90	0.14	0.05	0.85	0.29	0.35	0.35	0.14	0.72
n=3	0.90	0.06	0.01	0.91	0.13	0.05	0.91	0.42	0.32
n=4	0.91	0.06	0.02	0.90	0.14	0.04	0.92	0.30	0.18
n=5	0.90	0.08	0,02	0.90	0.14	0.05	0.91	0.26	0.12
n=6	0.90	0.06	0,01	0.90	0.10	0.04	0.91	0.28	0.10
n=7	0.90	0.06	0.01	0.90	0.10	0.03	0.91	0.26	0.08
n=8	0.90	0.04	0.01	0.90	0.06	0.02	0.91	0.23	0.07

Dari Tabel 6 diatas dapat dibuat grafik perbandingan antara Akurasi, Presisi dan Recall, pada setiap pengujian n-gram. Gambar 2 merupakan grafik rata-rata perhitungan akurasi, presisi dan recall pada threshold 0.5

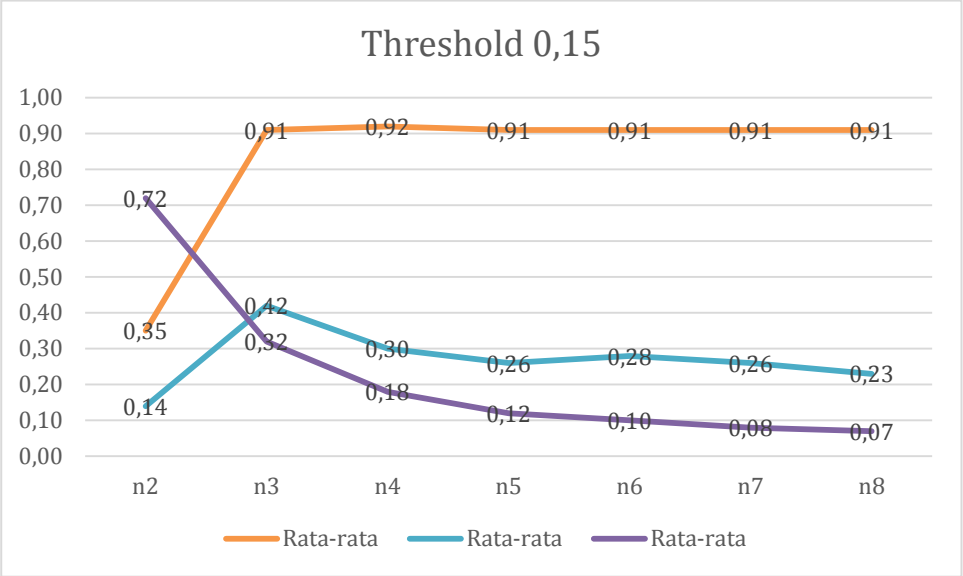


Gambar 2 Grafik pengujian menggunakan Threshold 0.5

Merujuk pada Tabel 6 Tabel Rata-rata perhitungan diatas gambar 2 merupakan grafik rata-rata perhitungan akurasi presisi dan recall pada threshold 0.5.

Merujuk pada Tabel 6 Tabel Rata-rata perhitungan diatas gambar 3 merupakan grafik rata-rata perhitungan akurasi presisi dan recall pada threshold 0.3.

Gambar 4 Grafik pengujian menggunakan Threshold 0.15



Gambar 3 Grafik pengujian menggunakan Threshold 0.3

Merujuk pada Tabel 6 Tabel Rata-rata perhitungan diatas gambar 4 merupakan grafik pengujian threshold 0.15

4. Conclusion

Dari hasil Penelitian diatas dapat memdapatkan kesimpulan yaitu :

a. Dari pengujian diatas bisa kita lihat bahwa ada nilai akurasi, nilai presisi dan nilai recall terbesar. Nilai akurasi terbesar yaitu pada pengujian n=4 adalah 0.92. dan nilai presisi tertinggi yaitu pada pengujian n=3 adalah 0.42, sedangkan nilai recall tertinggi pada pengujian n=2 adalah 0.72.

b. Dari ketiga grafik diatas tersebut terlihat bahwa terdapat nilai akurasi tertinggi yaitu 0.92 yang terdapat pada n=4 dengan threshold 0.15, sedangkan presisinya 0.3. data lain menunjukan nilai yang signifikan juga seperti pengujian n=3 dengan nilai presisi terbesar yaitu 0.42 tetapi dengan nilai akurasi 0.91

c. Pada pengujian n=4 pada threshold 0.15 yang berarti bahwa pengujian ini memiliki nilai ketepatan tertinggi pada sistem rekomendasi ini,

d. Dari Hasil beberapa pengujian diatas dapat disimpulkan bahwa untuk pemecahan kata dengan menggunakan algoritma n gram dan tanimoto cosine dapat mempersingkat waktu dalam pencarian artikel bahasa indonesia berbasis online secara cepat

References

[1] A. A. Gustiawan, W. Ramansyah, and M. Risnasari, “PENGEMBANGAN SISTEM PENCARIAN INFORMASI BUKU BERBASIS WEB MENGGUNAKAN MOVING CONTRACTING WINDOW PATTERN ALGORITHM di PERPUSTAKAAN SMKN 3 BANGKALAN,” vol. 4, no. 1, pp. 28–35, 2017.

[2] A. Indriani, T. Informatika, S. Informasi, M. T. Informatika, and U. A. Dahlan, “IMPLEMENTASI

JACCARD INDEX DAN N-GRAM PADA REKAYASA," pp. 95–101.
[3] J. Informatika, F. Matematika, and P. Alam, "Prosiding SNST ke-5 Tahun 2014 Fakultas Teknik Universitas Wahid Hasyim Semarang 79," pp. 79–84, 2014.
[4] M. Pencari and S. E. Optimization, "Analisis Pengaruh Kata Kunci Kompetitif Pada Search Engine Optimization (SEO) Terhadap Pemasaran Online Untuk Produk Notebook," pp. 73–80, 2010.
[5] I. Much, I. Subroto, and M. Khosyi, "Indonesian Articles Recommender System Using N- Gram and Tanimoto Coefficient," vol. 6, no. 1, pp. 15–20, 2018.
[6] A. Yudhana *et al.*, "IMPLEMENTASI DETEKSI PLAGIARISME MENGGUNAKAN METODE N-GRAM DAN JACCARD SIMILARITY TERHADAP ALGORITMA WINNOWING," no. 3, pp. 2–7, 2018.
[7] R. T. Wahyuni, D. Prastiyanto, and E. Supraptono, "Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi," vol. 9, no. 1, 2017.
[8] A. A. Prasanti, M. A. Fauzi, and M. T. Furqon, "Klasifikasi Teks Pengaduan Pada Sambat Online Menggunakan Metode N- Gram dan Neighbor Weighted K-Nearest Neighbor (NW-KNN)," vol. 2, no. 2, pp. 594–601, 2018.
[9] P. Algoritma and P. K. Kmp, "Penerapan algoritma pencarian knuth-morris-pratt (kmp) dalam sistem informasi perpustakaan smk ti pratama," pp. 112–115, 2018.
[10] D. Tampubolon, B. Nadeak, and M. Panjaitan, "PENERAPAN ALGORITMA ZHU-TAKAOKA UNTUK PENCARIAN NAMA HOTEL PADA APLIKASI PEMESANAN HOTEL DI KOTA," vol. 14, no. September, pp. 1–4, 2019.
[11] A. S. Rafika, H. Y. Putri, and F. D. Widiarti, "SEBAGAI SUMBER BARU UNTUK KUTIPAN," vol. 3, no. 2, pp. 193–205, 2004.