PETIR: Jurnal Pengkajian dan Penerapan Teknik Informatika
Vol. 17, No. 1, Maret 2024, P-ISSN 1978-9262, E-ISSN 2655-5018
DOI: https://doi.org/10.33322/petir.v17i1.2341

Pencarian Fungsional Obat Menggunakan Algoritma Tf-Idf Dan Cosine Similarity

Karina Djunaidi[1)]; Rahma Farah Ningrum[1]; Dine Tiara Kusuma[1]; Khesya Medhyna Saradiba[1]*

1. Institut Teknologi PLN, Menara PLN, Jl. Lingkar Luar Barat, Duri Kosambi, Cengkareng, Jakarta Barat, DKI Jakarta 11750, Indonesia

[*)]Email: karina@itpln.ac.id

Received: 15 Januari 2024 | Accepted: 15 Januari 2024 | Published: 7 Juni 2024

ABSTRACT

Pharmacists have an important role in ensuring that patients receive the right treatment. One that can support the pharmacist's task is a drug search feature that can provide detailed information about drugs, indications and other information needed. However, the drug search feature still has shortcomings due to the limitations and complexity of the information obtained. This research is made with the aim of applying TF/IDF and cosine similarity algorithms into a drug search system to help pharmacists find similarities in keywords and drug indications. The data used in this research is 425 drug data obtained from web scrapping, the data is then processed with text preprocessing then the results are applied to the TD/IDF algorithm for weighting, the results are then processed using the cosine similarity algorithm. The results of this study are in the form of drug search applications that were tested on experts with the results of the suitability of drug functions with indications of 79%.

Keywords: *Cosine Similarity, TD/IDF Algorithm, drugs searching, text mining*

ABSTRAK

Apoteker mempunyai peran penting untuk memastikan pasien menerima perawatan yang tepat. Salah satu yang dapat menunjang tugas apoteker adalah fitur pencarian obat yang dapat memberikan informasi detil tentang obat, indikasi dan informasi lain yang dibutuhkan. Namun fitur pencarian obat masih memiliki kekurangan karena keterbatasan dan kompleksitas informasi yang diperoleh. Penelitian ini dibuat dengan bertujuan menerapkan algoritma TF/IDF dan cosine similarity ke dalam sistem pencarian obat untuk membantu apoteker mencari kesamaan kata kunci dan indikasi obat. Data yang digunakan pada penelitian ini sebanyak 425 data obat yang diperoleh dari web scrapping, data kemudian diproses dengan text preprocessing kemudian hasilnya diterapkan algoritma TD/IDF untuk pembobotan, hasilnya kemudian diolah dengan menggunakan algoritma cosine similarity. Hasil dari penelitian ini berupa aplikasi pencarian obat yang diujicobakan ke pakar dengan hasil kesesuaian fungsi obat dengan indikasi sebesar 79%.

Kata kunci: *Cosine Similarity, Algoritma TD/IDF, pencarian obat, teks mining*

PETIR: Jurnal Pengkajian dan Penerapan Teknik Informatika
Vol. 17, No. 1, Maret 2024, P-ISSN 1978-9262, E-ISSN 2655-5018
DOI: https://doi.org/10.33322/petir.v17i1.2341

1. PENDAHULUAN

Di dunia medis, pencarian obat merupakan bagian yang sangat penting dalam pengobatan pasien. Namun, pencarian obat masih banyak mengalami kendala karena keterbatasan sumber informasi dan kompleksitas informasi yang diperoleh. Oleh karena itu, text mining cosine similarity dapat menjadi solusi untuk meningkatkan efektivitas pencarian obat [1] karena dengan text mining kompleksitas informasi yang terkumpul dapat dianalisa keterhubungannya [2]. Text mining cosine similarity adalah teknik analisis teks yang digunakan untuk mencari kemiripan antara dokumen berdasarkan kosinus dari sudut pandang vector [3],[4]. Dalam pencarian obat, teknik ini dapat digunakan untuk mencari obat-obatan yang memiliki kemiripan dalam komposisi kimia, efek terapeutik, atau indikasi medis. Tujuan dari penelitian ini adalah menerapkan *text mining* metode *cosine similarity* dalam sistem pencarian obat untuk membantu apoteker dalam mencari kesamaan kata kunci (*keywords)* dan indikasi obat. Ada 4 tahapan dalam *text preprocessing* yaitu *Case Folding*, *Filtering*, *Stemming*, dan *Tokenizing* [5], [6], [7] kemudian hasil dari tahap sebelumnya dilanjutkan dengan menggunakan algoritma TD/IDF untuk pembobotan. Algoritma TD/IDF adalah salah satu metode penghitungan relevansi dokumen yang populer digunakan dalam mesin pencari dan sistem informasi. Pada dasarnya, algoritma ini berfungsi untuk memberikan bobot pada kata-kata yang muncul dalam sebuah dokumen, sehingga dokumen yang lebih relevan dengan kata kunci pencarian akan muncul lebih tinggi dalam hasil pencarian [8], [9]. Proses algoritma TD/IDF dimulai dengan menghitung frekuensi kemunculan setiap kata dalam sebuah dokumen (term frequency/TD), kemudian dibandingkan dengan frekuensi kemunculan kata yang sama dalam seluruh dokumen yang ada dalam koleksi data *(inverse document frequency/IDF)*. Bobot akhir dari setiap kata dalam dokumen dihitung dengan cara mengalikan TD dan IDF [10], [11]. Penelitian ini menggunakan 425 data obat dimana untuk data uji menggunakan 10 data obat. Sebelum mengolah data, dilakukan tahap *preprocessing* dan perhitungan TF/IDF. Hasil dari tahap *preprocessing*, data kemudian diolah ke perhitungan TF/IDF, setelah diperoleh hasil perhitungan TF/IDF kemudian dilakukan evaluasi menggunakan cosine similarity dan validasi oleh pakar.

Selain dapat diterapkan dalam pencarian obat, teks mining dapat diaplikasikan dalam berbagai bidang salah satunya analisis opini pengguna aplikasi [12], analisa sentimen terhadap vaksin [13], serta untuk mendeteksi spam pada email [14]. Sedangkan cosine similarity yang digunakan pada penelitian ini untuk mencari komposisi kimia obat juga digunakan pada penelitian lain berupa klasifikasi bidang pekerjaan [15], Analisa sitasi pada artikel ilmiah [16] dan deteksi perbedaan gambar hasil scan pada android [17].

2. METODE/PERANCANGAN PENELITIAN

Gambar 1 menunjukkan perancangan penelitian yang sudah dilakukan pada penelitian ini.

Gambar 1. Perancangan Penelitian

Vol. 17, No. 1, Maret 2024, P-ISSN 1978-9262, E-ISSN 2655-5018
DOI: https://doi.org/10.33322/petir.v17i1.2341

a. Text Preprocessing

Tahap awal dari penelitian ini adalah mengumpulkan data obat dengan cara melakukan *web scraping* pada website AiCare: ai-care.id untuk memperoleh data obat. Sebanyak 425 data diperoleh dari hasil *web scraping* yang kemudian data tersebut diolah ke tahap berikut yaitu *text preprocessing. Text preprocessing* ini terbagi menjadi *case folding, filtering, stemming*, dan *tokenizing*. Pada tahap *case folding*, dilakukan proses untuk mengubah kalimat menjadi huruf kecil serta menghilangkan angka, simbol dan tanda baca serta menyisakan huruf atau *string* saja. Kemudian dilanjutkan ke tahap *filtering* yaitu dengan membuang kata yang dianggap tidak penting. Hasil dari tahap *filtering* ini kemudian dilanjutkan ke tahap *stemming* yaitu mengubah kata berimbuhan menjadi kata dasar. Pada tahap *tokenizing,* dilakukan proses pemotongan data dari kalimat menjadi data tunggal. Hasil dari tahap *text preprocessing* ditunjukkan pada tabel 1 dibawah ini.

Tabel 1. Hasil *text preprocessing*

Data awal	Case Folding	Filtering	Stemming	Tokenizing
Kondisi medis berupa infeksi virus HIV atau AIDS.	kondisi medis berupa infeksi virus hiv atau aids	kondisi medis infeksi virus hiv aids	kondisi medis infeksi virus hiv aids	[‘kondisi’,’medis’,’infeksi’, ‘virus’,’hiv’,’aids’]
Obat ini diberikan pada penderita diabetes melitus tipe 2.	obat ini diberikan pada penderita diabetes melitus tipe 2	obat penderita abetes melitus tipe	obat derita diabetes melitus tipe	[‘obat’,’derita’,’diabetes’,’melitus’,’tipe,]

b. Pembobotan *Term Frequency-Inverse Document Frequency* (TF/IDF)

Selanjutnya tahap pembobotan TF/IDF, dimana pada tahap ini dilakukan normalisasi dan perhitungan kemunculan kata dalam kalimat untuk mencari perangkingan pada data. TF/IDF merupakan teknik pengolahan teks dengan memberikan bobot pada kata dari sebuah dokumen[18]. Untuk melakukan pembobotan TF/IDF, kata yang diperoleh pada tahap preprocessing, dihitung jumlah kemunculannya pada kalimat, setelah itu kemudian hasil tiap kemunculan kata yang sama dijumlahkan. Hasil TF-IDF diperoleh dari perhitungan dengan menggunakan rumus:

$$TF = \frac{Jumlah\ kemunculan\ kata}{Jumlah\ kata\ dalam\ dokumen} \quad (1)$$

$$IDF = log\frac{n}{df} \quad (2)$$

Dimana n adalah jumlah dokumen dan df adalah hasil kemunculan setiap kata yang sama. Hasil dari perhitungan TF/IDF ditunjukkan pada tabel 2 berikut ini.

Vol. 17, No. 1, Maret 2024, P-ISSN 1978-9262, E-ISSN 2655-5018
DOI: https://doi.org/10.33322/petir.v17i1.2341

Tabel 2. Hasil TF/IDF

Kata	DF	D1	D2	D3	D4	…	D10
aids	0.00000	1.00000	0.00000	0.00000	0.00000	…	0.00000
akibat	0.00000	0.00000	0.00000	0.00000	0.00000	…	0.00000
alami	0.00000	0.00000	0.00000	0.00000	0.00000	…	1.00000
…	…	…	…	…	…	…	…
kondisi	0.00000	0.221849	0.221849	0.221849	0.00000	…	0.00000
kongestif	0.00000	0.00000	0.00000	0.00000	0.00000	…	0.00000

c. Cosine Similarity

Data dari hasil perhitungan TF/IDF kemudian dievaluasi menggunakan cosine similarity untuk melihat seberapa mirip indikasi yang diberikan oleh pengguna dengan kata kunci yang ada pada data menggunakan rumus sebagai berikut [19] [20]:

$$Simillarity = \cos\theta \frac{A.B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (3)$$

Dimana:
A · B = Vektor dot product dari A dan B
||A|| = Panjang dari vektor A
||B|| = Panjang dari vektor B
||A||B|| = *cross product* antara |A| dan |B|
Ai = bobot query pada term i
Bi = bobot dokumen pada term i dan dokumen j
n = jumlah vektor

Hasil perhitungan dengan menggunakan rumus cosine similarity pada data disajikan pada tabel 3 dibawah ini.

Tabel 3. Cosine Simillarity

Obat	*Cosine Similarity*
Abacavir	0, 5045
Adefovir	0,1109
Adapalene	0,1109
Acarbose	0.0172
Acetylcysteine	0,0093
Abiraterone	0.0085
Acetazolamide	0.0065
Aclidinium-bromide	0,0057
Abemaciclib	0.0000
Adalimumab	0.0000

Pada tabel 3 dapat dilihat bahwa dataset ke-1 merupakan data set dengan tingkat kemiripan tertinggi dengan nilai 0,504545 yaitu tingkat kemiripannya 50,45% dibandingkan dengan tingkat

Vol. 17, No. 1, Maret 2024, P-ISSN 1978-9262, E-ISSN 2655-5018
DOI: https://doi.org/10.33322/petir.v17i1.2341

minimum dataset ke-8. Kesamaan dengan kata kunci, yaitu nilai 0.

3. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan data sejumlah 425 data obat yang diperoleh dari *web-scraping* pada *website Ai-Care*. Dari sejumlah data tersebut data yang diujicobakan hanya 10 data indikasi obat yang diproses melalui *text preprocessing*. Setelah diperoleh hasil dari *text preprocessing* kemudian diolah dengan TF/IDF dan akan dievaluasi menggunakan *cosine similarity*. Data yang diperoleh kemudian dievaluasi untuk melihat kata kunci dengan 10 data indikasi hasil dari pembobotan dan *cosine similarity* menggunakan kata kunci "Mengobati infeksi hiv" ditunjukkan pada tabel 4 dan tabel 5 berikut ini.

Tabel 4. Hasil evaluasi TF/IDF

Data	*Cosine Similarity*
Abacavir (D1)	3,7447%
Abemaciclib (D2)	5,2218%
Abiraterone (D3)	11,076%
Acarbose (D4)	2,699%
Acetazolamide (D5)	18,907%
Acetylcysteine (D6)	8,8539%
Aclidinium-bromide (D7)	24,696%
Adalimumab (D8)	25,143%
Adapalene (D9)	8,6198%
Adefovir (D10)	8,6198%

Tabel 5. Hasil evaluasi *cosine similarity*

Data	*Cosine Similarity*
Abacavir	50.45%
Adefovir	11.09%
Adapalene	11.09%
Acarbose	0.172%
Acetylcysteine	0.93%
Abiraterone	0.85%
Acetazolamide	0.65%
Aclidinium- bromide	0.57%
Abemaciclib	0%
Adalimumab	0%

Hasil evaluasi yang ditunjukkan oleh tabel 4 dan 5 menunjukkan bahwa nilai cosine similarity pada data 1 lebih tinggi dari nilai cosine similarity pada data 8, hal ini berarti penerapan cosine similarity pada kata kunci memberikan bobot yang cukup baik dalam pemberian peringkat jenis obat dengan fungsi obat. Hasil dari penelitian ini kemudian diujicobakan dengan pakar yang merupakan seorang apoteker pada sebuah klinik di Tangerang. Hasil dari ujicoba ini dapat dilihat pada tabel 6 berikut ini.

Vol. 17, No. 1, Maret 2024, P-ISSN 1978-9262, E-ISSN 2655-5018
DOI: https://doi.org/10.33322/petir.v17i1.2341

Tabel 6. Hasil Validasi

No.	Pertanyaan	Penilaian Pakar
1.	Berapa persentase hasil pencarian yang dilakukan?	80%
2.	Berapa persentase penjelasan tentang obat?	80%
3.	Berapa persentase kesesuaian data obat	80%
4.	Berapa persentase kesesuaian indikasi obat dengan fungsi obat?	79%

Tabel 6 menunjukkan bahwa, hasil presentase hasil pencarian, penjelasan obat, dan kesesuaian data obat sebesar 80% akurat. Sedangkan hasil kesesuaian indikasi obat dengan fungsi obat 79%. Hal ini disebabkan terdapat kesalahan karena kata kunci memiliki banyak kecocokan, sehingga sistem memilih nilai terbesar dari perhitungan *cosine similarity* yang ada.

4. KESIMPULAN DAN SARAN

Cosine Similarity berhasil diterapkan dalam sistem pencarian obat dan sistem dapat menghasilkan keluaran berupa obat yang sesuai dengan indikasi. Hasil cosine similarity 50,45% pada data pertama kata kunci menunjukkan hasil pemeringkatan jenis obat dengan fungsi obat. Sistem pencarian obat ini telah diujicobakan pada pakar dengan hasil keakuratan 80% untuk hasil pencarian, penjelasan obat dan kesesuaian data obat, sedangkan untuk kesesuaian indikasi obat dengan fungsi obat menunjukkan hasil 79%.

Pengembangan penelitian ini selanjutnya dapat dilakukan dengan menambahkan pengelompokkan fungsi obat serta data obat yang digunakan. Selain itu juga dapat menggunakan algoritma lain untuk meningkatkan tingkat akurasi pencarian indikasi obat dengan fungsi obat.

UCAPAN TERIMAKASIH

Terima kasih kepada Institut Teknologi PLN dan Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) ITPLN atas pendanaan penelitian ini.

DAFTAR PUSTAKA

[1] Indarto E, Widiasri M, Siswantoro J. “PEMBUATAN WEBSITE BERBAHASA INDONESIA UNTUK PENCARIAN RESEP MASAKAN DENGAN METODE COSINE SIMILARITY. CALYPTRA”. 2019 Sep 1;8(1):2301-16.

[2] Ristanti PY, Wibawa AP, Pujianto U. “Cosine similarity for title and abstract of economic journal classification”. In 2019 5th International Conference on Science in Information Technology (ICSITech) 2019 Oct 23 (pp. 123-127). IEEE.

[3] Syamsuddin S, Alloto'dang K. “Perancangan Sistem Klasifikasi Surat Elektronik (E-Mail) Menggunakan Metode Cosine Similarity”. 2020 Sep 23;1(5):594-606.

[4] Kumar S, Kar AK, Ilavarasan PV. Applications of text mining in services management: A systematic literature review. International Journal of Information Management Data Insights. 2021 Apr 1;1(1):100008

[5] Hassani H, Beneki C, Unger S, Mazinani MT, Yeganegi MR. Text mining in big data analytics. Big Data and Cognitive Computing. 2020 Jan 16;4(1):1.

[6] Kowsari, Jafari Meimandi, Heidarysafa, Mendu, Barnes, Brown. Text Classification Algorithms: A Survey. Information [Internet]. 2019 Apr 23;10(4):150. Available from: http://dx.doi.org/10.3390/info10040150

Vol. 17, No. 1, Maret 2024, P-ISSN 1978-9262, E-ISSN 2655-5018
DOI: https://doi.org/10.33322/petir.v17i1.2341

[7] Apriani A, Zakiyudin H, Marzuki K. Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF System Penerimaan Mahasiswa Baru pada Kampus Swasta. Jurnal Bumigora Information Technology (BITe). 2021 Jul 10;3(1):19-27.

[8] Asmarajati D. Analisis Perbandingan Algoritma Tf-Idf Dengan Sql Query Untuk Kasus Pencarian Pada Sistem Informasi Dokumentasi Arsip (Sidokar). Device. 2020 May 31;10(1):1-8.

[9] Naf'an MZ, Burhanuddin A, Riyani A. Penerapan Cosine Similarity dan Pembobotan TF-IDF untuk Mendeteksi Kemiripan Dokumen. Jurnal Linguistik Komputasional. 2019 Mar 26;2(1):23-7

[10] Lolyta SN, Dillak RY, Laumal FE. Sistem deteksi plagiarisme lintas bahasa menggunakan algoritma tf-idf. Jurnal Ilmiah Flash. 2019 Jun 1;5(1):29-32.

[11] Fajriani A. Aplikasi ISO (Informasi Spesialite Obat) Indonesia Berbasis Web Menggunakan Metode Pencarian Binary Search: Web-Based Indonesian ISO (Informasi Spesialite Obat) Application Using Binary Search Search Method. Decode: Jurnal Pendidikan Teknologi Informasi. 2021 Mar 31;1(1):8-16.

[12] H. B. Tambunan and T. W. D. Hapsari, "Analisis Opini Pengguna Aplikasi New PLN Mobile Menggunakan Text Mining", petir, vol. 15, no. 1, pp. 121–134, Dec. 2021

[13] Fathonah Fira, Herliana Asti, "Penerapan Text Mining Analisis Sentimen Mengenai Vaksin Covid-19 menggunakan Naive Bayes", Jurnal Sains dan Informatika, vol.7, no. 2, pp.155-164

[14] Hidayat A, "Aplikasi Text Mining untuk mendeteksi Spam pada Email berbasis Naive Bayes", Teknologi Pintar, vol. 2, no. 8, 2022

[15] P. C. Siswipraptini, "Klasifikasi Pekerjaan Bidang Teknologi Informasi Menggunakan Algoritma Cosine Similarity", kilat, vol. 12, no. 1, pp. 38–48, Apr. 2023

[16] U. Mardatillah, W. B. Zulfikar, A. R. Atmadja, I. Taufik and W. Uriawan, "Citation Analysis on Scientific Articles Using Cosine Similarity," 2021 7th International Conference on Wireless and Telematics (ICWT), Bandung, Indonesia, 2021, pp. 1-4, doi: 10.1109/ICWT52862.2021.9678402

[17] K. Telaumbanua and L. Nababan, "Implementasi Metode Cosine Similarity Dalam Mendeteksi Kemiripan Dan Perbedaan Gambar Hasil Scan Berbasis Android", IEED, vol. 1, no. 1, pp. 27–36, Dec. 2022.

[18] Anugrah IG. Penerapan Metode N-Gram dan Cosine Similarity Dalam Pencarian Pada Repositori Artikel Jurnal Publikasi. Building of Informatics, Technology and Science (BITS). 2021 Dec 31;3(3):275-84.

[19] T. M. Fahrudin, M. H. Hartanto, A. S. Paramita, A. Aulia, R. A. Maulana, and I. R. Anniswa, "TEMU KEMBALI INFORMASI BERITA KEGIATAN PROGRAM STUDI MENGGUNAKAN ALGORITMA PEMBOBOTAN TF-IDF DAN COSINE SIMILARITY", sitasi, vol. 2, no. 1, pp. 270-279, Sep. 2022.

[20] M. Lestari, K. Djunaidi, R. F. Ningrum, D. T. Kusuma dan K. M. Saradiba, "Implementasi Cosine Similarity dalam Pencarian Fungsional Obat," dalam Prosiding Seminar Nasiona Energi, Kelistrikan, Teknik dan Informatika, Jakarta, 2023