Pencarian Dokumen Rekam Medis dengan TF-IDF dan *Vector Space Model (VSM)*

Salsabila Nurulfarah Mahmudah1, Kartika Rizqi Nastiti2, Dian Novitaningrum3, Lukman Heryawan4

*Program Studi Ilmu Komputer, Departemen Ilmu Komputer dan Elektronika, Universitas Gadjah Mada  
Yogyakarta, Indonesia.*

[1salsabila.nurulfarah0598@mail.ugm.ac.id](mailto:1salsabilanurul0598@mail.ugm.ac.id)

2[kartikarizqinastiti@mail.ugm.ac.id](mailto:1salsabilanurul0598@mail.ugm.ac.id)

3[diannovitaningrum@mail.ugm.ac.id](mailto:1salsabilanurul0598@mail.ugm.ac.id)

4[lukmanh@ugm.ac.id](mailto:1salsabilanurul0598@mail.ugm.ac.id)

*Abstract*— ***The growth of medical record documents is increasing over time and the various types of diseases and therapies needed are increasing. However, this has not been followed by an effective and efficient search process. This study aims to deal with search problems that often take a long time with search results that are not necessarily as expected, by building a search model for medical record documents using the Vector Space Model (VSM) and TF-IDF methods. The VSM method allows retrieval of results that do not have to be exactly the same as the search queries entered by the user, but are expected to still provide results that are relevant to the user's desired needs. The model development process was taken based on the data in the FS\_ANAMNESA and FS\_DIAGNOSA columns, followed by preprocessing which consists of deleting blank lines, lowercase, removing punctuation marks, HTML tags, stopwords, excess spaces between words, and normalizing typo words, then forming a TF-IDF matrix based on the frequency of occurrence of each word feature, and followed by the calculation of the similarity value of the search query compared to medical record documents based on the cosine similarity formula. The retrieval results were all columns of each existing medical record document and are sorted based on 10 rows with the highest similarity value. The model evaluation results based on 1000 medical record documents and tested with 20 search queries in this study gave an average precision value of 0.548 and an average recall value of 0.796.***

***Keywords— Medical Records, Preprocessing, Cosine Similarity, TD-IDF, Evaluation Matrix***

***Abstrak*— Pertumbuhan dokumen rekam medis semakin meningkat seiring waktu dan semakin beragamnya jenis penyakit dan terapi yang diperlukan. Namun, hal ini belum diimbangi dengan proses pencarian yang efektif dan efisien. Penelitian ini bertujuan untuk menangani permasalahan pencarian yang seringkali membutuhkan waktu yang lama dengan hasil pencarian yang belum tentu sesuai harapan, dengan membangun model pencarian dokumen rekam medis dengan metode *Vector Space Model* (VSM)dan TF-IDF. Metode VSM memungkinkan hasil *retrieval* yang tidak harus sama persis dengan *query* pencarian yang dimasukkan oleh pengguna, namun diharapkan tetap memberikan hasil yang relevan dengan kebutuhan yang diinginkan pengguna. Proses pengembangan model diambil berdasarkan data pada kolom FS\_ANAMNESA dan FS\_DIAGNOSA, diikuti dengan *preprocessing* yang terdiri dari penghapusan baris kosong, *lowercase*, menghapus tanda baca, *tag* HTML, *stopwords,* spasi berlebih antar kata, serta melakukan normalisasi kata *typo,* lalu dilakukan pembentukan matriks TF-IDF berdasarkan frekuensi kemunculan tiap fitur kata, dan penghitungan nilai kemiripan *query* pencarian dengan dokumen rekam medis berdasarkan formula *cosine similarity.* Hasil *retrieval* adalah keseluruhan kolom dari tiap dokumen rekam medis yang sudah ada dan diurutkan berdasarkan 10 baris dengan nilai kemiripan tertinggi. Hasil evaluasi model berdasarkan 1000 dokumen rekam medis dan pengujian dengan 20 *query* pencarian dalam penelitian ini memberikan nilai rata-rata *precision* sebesar 0.548 dan nilai rata-rata *recall* sebesar 0.796.**

***Kata Kunci*— Rekam Medis, *Preprocessing, Cosine Similarity,* TF-IDF, *Evaluation Matrix***

1. Pendahuluan

Sebuah data rekam medis disimpan paling singkat 25 tahun sejak tanggal kunjungan terakhir [1]. Salah satu indikator kualitas pelayanan fasilitas kesehatan adalah kecepatan dan ketepatan dalam menyediakan *medical record* [2]. Menurut Peraturan Kementerian Kesehatan Indonesia nomor 129/Menkes/SK/II/2008 menjelaskan bahwa salah satu standar minimum dalam sebuah pelayanan di fasilitas kesehatan adalah dapat menyediakan rekam medis dalam waktu 10 menit dan mempunyai kelengkapan informasi sebesar 100% [3]. Namun fakta pada sebuah rumah sakit di Indonesia masih sebesar 65.75% memerlukan waktu > 10 menit untuk data rekam medis pasien berhasil ditampilkan dan tingkat *accuracy* sebesar 87.5% [4].

Menurut UU Praktik Kedokteran dalam penjelasan pasal 46 ayat (1) yang dimaksud dengan Rekam Medis adalah berkas yang berisi catatan dan dokumen tentang identitas pasien, pemeriksaan, pengobatan, tindakan dan pelayanan lain yang telah diberikan kepada pasien [5]. Pengertian rekam medis diperkuat melalui Peraturan Menteri Kesehatan (Permenkes) No. 269/2008, bahwa jenis data rekam medis dapat berupa teks (baik yang terstruktur maupun naratif), gambar digital (jika sudah menerapkan radiologi digital), suara (misalnya suara jantung), video maupun yang berupa biosignal seperti rekaman Elektrokardiografi (EKG) [6].

*Information Retrieval* (IR) atau Temu Balik Informasi dalam ilmu komputer adalah proses mendapatkan informasi yang relevan, dalam hal ini berkaitan dengan merepresentasikan, mencari, dan memanipulasi data besar kumpulan teks elektronik dan data bahasa manusia lainnya [7]. Pencarian dapat didasarkan pada teks lengkap atau pengindeksan berbasis konten lainnya. Informasi dalam konteks ini dapat terdiri dari teks (termasuk data numerik dan tanggal), gambar, audio, video, dan objek multimedia lainnya [8]. Penelitian dan implementasi IR telah banyak dilakukan di berbagai bidang, salah satunya dalam pencarian dokumen rekam medis. Faridah, et. al (2019) melakukan penelitian untuk menemukan kembali gambar dari sebuah arsip gambar yang besar berdasarkan konten *query* citra di bidang medis [9]. Kang, et. al (2017) melakukan penelitian untuk pencarian rekam medis berdasarkan kesamaan semantik [10].

*Vector Space Model* (VSM) adalah model aljabar untuk merepresentasikan objek apa pun sebagai vektor untuk mengukur tingkat kemiripan. Intuisi di balik model ini adalah bahwa setiap vektor mengkodekan arti dari objek yang sesuai dan untuk istilah yang dekat secara semantik, maka ukuran kemiripan antara vektor mereka direpresentasikan dalam nilai yang tinggi [11]. Salah satu pengukuran tingkat kemiripan yang paling populer untuk VSM adalah *Cosine similarity* dan salah satu cara untuk mengkodekan dokumen tekstual ke dalam vektor adalah *Term Frequency - Inverse Document Frequency* (TF-IDF). Wahyudi, et. al (2019) melakukan penelitian dengan memanfaatkan *cosine similarity* dengan TF-IDF sebagai skema pembobotan untuk pencarian file JSON yang relevan dengan *query* yang diberikan [12]. Mas’udia, et. al (2017) yang memanfaatkan TF-IDF untuk melakukan pencarian dokumen tugas akhir [13].

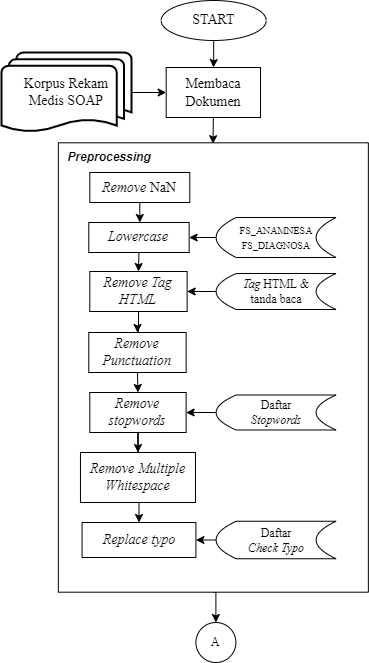
Perangkingan termasuk tahapan dengan menggunakan bantuan *Cosine Similarity.* *Cosine Similarity* merupakan metode yang menganalisa kemiripan yang memiliki fungsi untuk mendapatkan kata atau *term* yang diinginkan berdasarkan *query* serta jarak untuk mengurutkannya [14] [15]. Kedua objek yang dievaluasi dikatakan semakin mirip ketika semakin besar hasil kemiripan yang didapatkan. Terdapat beberapa penelitian sebelumnya yang membahas diantaranya adalah klasifikasi berita *online* dengan pembobotan TF-IDF dan *Cosine Similarity.* Tujuan dilakukan penelitian tersebut adalah untuk memudahkan serta mengurangi adanya kesalahan pemberian kategori oleh *human error.* Nilai evaluasi yang diperoleh adalah 91.25 % untuk tingkat *accuracy*nya [16].

*Precision* adalah rasio jumlah dokumen relevan yang ditemukan dengan total jumlah dokumen yang ditemukan. Sedangkan, evaluasi kedua menggunakan *recall*, *Recall* adalah rasio jumlah dokumen relevan yang ditemukan dengan total jumlah dokumen dalam koleksi yang dianggap relevan [17]. Implementasi lain pada penelitian sebelumnya yaitu pencarian tema yang sejenis pada sinopsis novel bahasa indonesia dengan menggunakan metode *General Vector Space Method.* *Dataset* memiliki jumlah lebih dari 1500 *record.* Kemudian menggunakan data uji untuk 300 data dengan pembagian yaitu 150 data uji pencarian dan sisanya untuk pengujian klasifikasi. Sehingga mendapatkan hasil akhir yaitu *recall* 90% dan *precision* sebesar 85% [18].

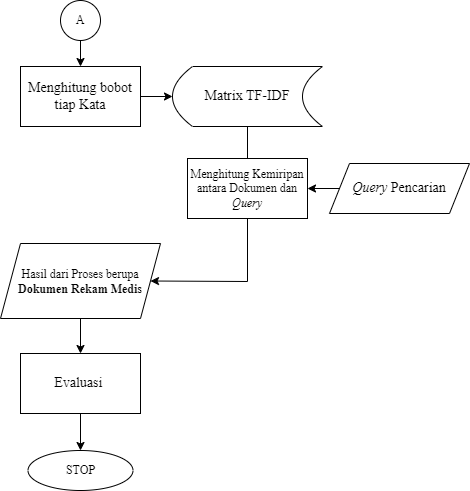
Pada penelitian ini dilakukan proses *retrieve* kumpulan dokumen rekam medis dari *dataset* yang berisi 1000 baris data. *Dataset* tersebut merupakan data Rekam Medis SOAP yang berasal dari REKMED, aplikasi pencatatan rekam medis pasien berbasis *cloud.* *Dataset* ini mempunyai 5 jumlah kolom, yang berisikan FS\_ANAMNESA, FS\_TINDAKAN, FS\_TERAPI, FS\_CATATAN\_FISIK, dan FS\_DIAGNOSA. Metode yang digunakan adalah melakukan *data preprocessing*, penghitungan kemiripan dengan *cosine similarity*, perangkingan berdasarkan *cosine similarity,* dan menggunakan beberapa *evaluation matrix.* Diharapkan hasil dari penelitian ini mampu menawarkan model pencarian dokumen rekam medis yang sesuai dengan *query* yang dimasukkan oleh pengguna.

1. Metodologi

Berikut merupakan penjabaran metodologi yang digunakan di dalam penelitian ini. Gambar 1 menunjukkan proses *preprocessing* dan Gambar 2 menunjukkan proses penghitungan kemiripan.



Gambar 1. Diagram alir tahap *preprocessing*



Gambar 2 Diagram alir penghitungan kemiripan

1. *Preprocessing*

*Preprocessing* adalah tahap pertama yang dilakukan untuk membersihkan dan melakukan normalisasi data rekam medis yang sudah ada. Data rekam medis berisi 5 kolom, yaitu FS\_ANAMNESA, FS\_TINDAKAN, FS\_TERAPI, FS\_CATATAN\_FISIK, dan FS\_DIAGNOSA. Dalam penelitian ini, diambil data teks dari dua kolom, yaitu FS\_ANAMNESA dan FS\_DIAGNOSA, dengan total baris data sebanyak 1000 baris. Berikut adalah rincian dari tiap tahap *preprocessing* yang dilakukan dalam penelitian ini:

1. Menghapus baris kosong, di mana baris yang tidak memiliki data baik di kolom FS\_ANAMNESA atau FS\_DIAGNOSA dihapus untuk tahap berikutnya.
2. *Lowercase*, dimana setiap huruf diubah menjadi huruf kecil.
3. Menghapus tanda baca dan *tag HTML.* Data rekam medis memiliki banyak tanda baca dan *tag HTML* seperti <p>, <br>, dan lain sebagainya sehingga perlu dihapus agar proses selanjutnya menjadi lebih efisien.
4. Menghapus *stopwords*. Kata-kata yang tidak memiliki makna signifikan seperti kata sambung, kata depan, atau kata-kata lainnya dihapus agar tidak terlalu banyak *noise*.
5. Menghapus spasi berlebih antar kata atau karakter.
6. Normalisasi dengan mengganti kata-kata *typo*, seperti ‘peerut’ menjadi ‘perut’, ‘jln’ menjadi ‘jalan’, ‘riwy’ menjadi ‘riwayat’, dan lain sebagainya. Di dalam penelitian ini dihimpun sebanyak 1290 pasang kata yang dijadikan acuan untuk mengganti kata dalam teks rekam medis.
7. *Pembentukan Matriks TF-IDF*

*Term Frequency - Inverse Document Frequency* (TF-IDF) merupakan salah satu cara untuk menghitung bobot kata berdasarkan frekuensi kemunculannya di dalam sebuah korpus. Algoritma TF-IDF digunakan untuk mengevaluasi pentingnya kata-kata dalam korpus tekstual [19]. TF mewakili frekuensi kata, menunjukkan berapa kali mereka muncul dalam korpus persamaan (1). Bobot tiap kata dihitung dari kemunculan kata tersebut dibandingkan dengan jumlah total kata yang ada dalam korpus.

|  | (1) |
| --- | --- |
|  | (2) |
|  | (3) |

Di mana:

frekuensi kemunculan kata *t* dalam dokumen D

korpus dokumen

IDF menunjukkan tingkat kepentingan sebuah kata dalam korpus secara keseluruhan. Nilai IDF didapatkan dengan menghitung logaritma dari nilai *inverse* proporsi dokumen yang mengandung kata tertentu persamaan (2). Nilai proporsi diambil dari jumlah dokumen dalam korpus dibagi dengan jumlah dokumen yang mengandung kata tertentu. Nilai logaritma dari proporsi tersebut merupakan nilai IDF. Bobot TF-IDF dihitung dengan mengalikan dua nilai tersebut sebagaimana terlihat pada persamaan (3). Semakin besar nilai TF-IDF, menunjukkan semakin signifikan kata yang bersangkutan di dalam korpus.

1. *Penghitungan Kemiripan*

*Cosine Similarity* merupakan sebuah ukuran untuk menghitung jarak yang digunakan pada data berupa vektor dari dokumen. Dokumen berisikan sebuah data yang dapat terdiri dari ratusan bahkan ribuan atribut. Setiap atribut mewakili *term* atau kata yang berisi nilai berarti frekuensi kemunculan dalam dokumen tertentu. Vektor pada dokumen merupakan vektor yang berisi frekuensi jumlah kemunculan kata pada suatu dokumen. Perhitungan untuk nilai *term* pada masing-masing dokumen adalah kemiripan pada dua buah vektor dalam ruang dimensi yang didapatkan dari *cosinus* pada sebuah sudut dari perkalian antara kedua buah vektor yang dibandingkan karena *cosinus* pada nilai 0 merupakan 1 dan kurang dari 1 untuk sudut yang lainnya. *Cosine score* yang bernilai 0 mempunyai makna bahwa kedua buah vektor berada pada sumbu 90o secara ortogonaldan tidak memiliki kecocokan, sedangkan *cosine score* yang semakin dekat dengan nilai 1 maka semakin kecil sudut dan semakin besar kecocokan antara 2 vektor [20]. Dua buah vektor dikatakan mirip ketika nilai pada *cosine similarity*nya adalah 1. Penghitungan *cosine similarity* dijelaskan pada persamaan (4), yaitu:

(4)

Dimana:

Vektor Dokumen

Vektor *Query*

Perkalian antara vektor A dan vektor B

Panjang vektor A

Panjang vektor B

*Cross Product* antara |A| dan |B|

Sudut yang terbentuk antara vektor A dan B

1. *Evaluasi*

*Precision* dan *Recall* merupakan salah satu uji evaluasi pada sistem temu balik yang sering dilakukan. *Precision* merupakan ukuran ketepatan dari jumlah dokumen yang dapat ditemukan dan dianggap relevan oleh proses pencarian untuk keperluan pencarian dokumen. Sedangkan, *recall* adalah rasio jumlah dokumen relevan yang ditemukan dengan total jumlah dokumen dalam koleksi yang dianggap relevan. Rumus persamaan (5) menjelaskan mengenai *precision* dan persamaan (6) menjelaskan *recall.*

(5)

(6)

1. Hasil dan Pembahasan

Untuk dapat mengetahui performa dari model *information retrieval* yang dikembangkan, peneliti menggunakan beberapa *query* yang dimasukkan. Berikut adalah *list* 10 *query* yang digunakan untuk menguji model dapat dilihat pada Tabel I dan Tabel II.

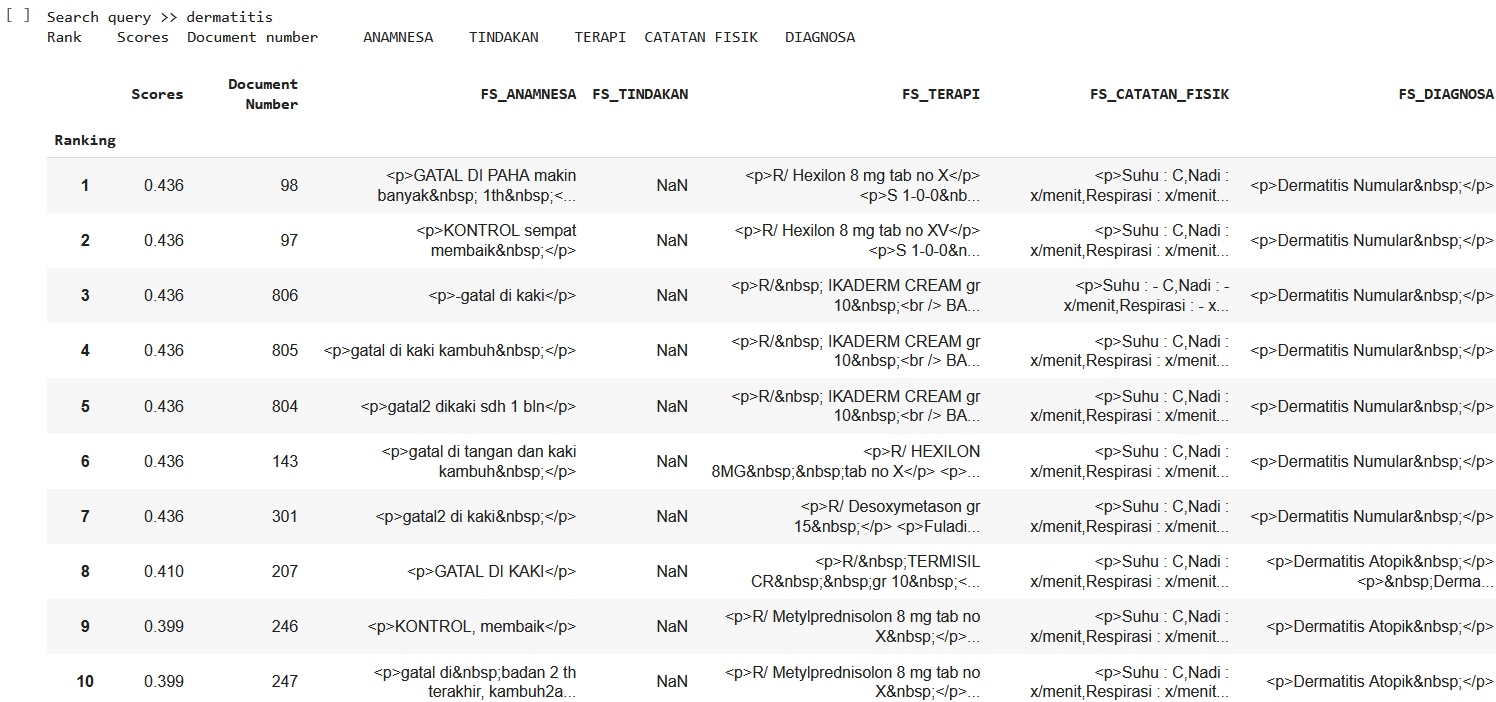
Tabel I  
*LIST* *QUERY* KOLOM FS\_ANAMNESA PADA PENGUJIAN MODEL

| **No** | **FS\_ANAMNESA** | **Jumlah Dokumen** |
| --- | --- | --- |
| 1. | kontrol | 10 |
| 2 | gigi sakit | 10 |
| 3. | ca paru | 10 |
| 4. | nyeri bahu kanan | 6 |
| 5. | nyeri | 4 |
| 6. | telinga nyeri | 4 |
| 7. | lutut kanan sakit | 4 |
| 8. | lutut kiri sakit | 4 |
| 9. | batuk pilek | 2 |
| 10. | batuk | 2 |

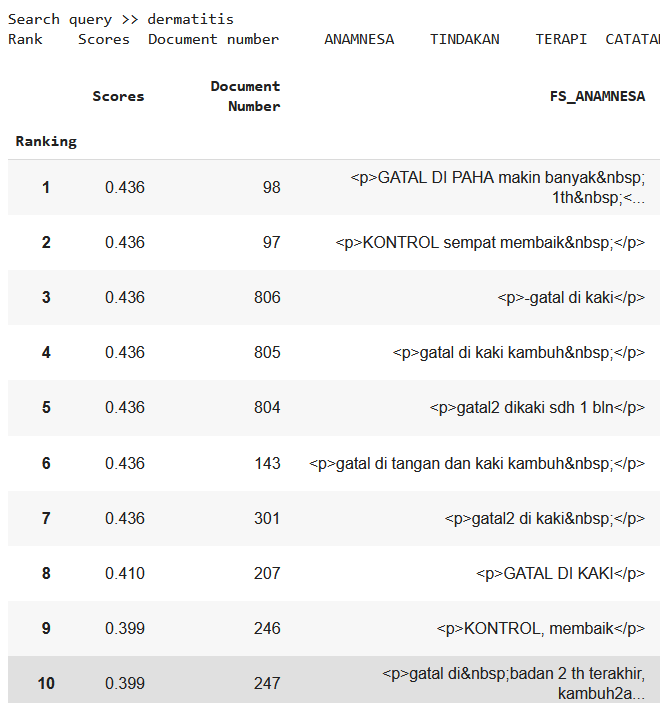
Tabel II  
*LIST* *QUERY* UNTUK KOLOM FS\_DIAGNOSA PADA PENGUJIAN MODEL

| **No** | **FS\_DIAGNOSA** | **Jumlah Dokumen** |
| --- | --- | --- |
| 1. | hamil | 10 |
| 2. | bronkitis akut | 10 |
| 3. | neurosa | 10 |
| 4. | dermatitis | 10 |
| 5. | et | 10 |
| 6. | bph | 10 |
| 7. | abses gigi | 10 |
| 8. | meningioma | 10 |
| 9. | kb | 4 |
| 10. | rhinitis | 4 |

Berikut adalah salah satu contoh dari dokumen yang berhasil di *retrieve.* Gambar 3 menunjukkan semua kolom yang berhasil di *retrieve* sedangkan Gambar 4 menunjukkan lebih jelas pada kolom rangking, *scores, document number,* dan FS\_ANAMNESA.

**

Gambar 3. Semua kolom yang berhasil di *retrieve*



Gambar 4. Kolom rangking, *scores, document number,* dan FS\_ANAMNESA yang berhasil di *retrieve*.

Pada gambar 4 terlihat bahwa setiap dokumen yang berhasil di *retrieve* mempunyai bobot yang didapatkan dari penghitungan *cosine similarity*. Bobot tersebut digunakan untuk menentukan rangking dari dokumen yang berhasil di *retrieve.* Semakin tinggi nilai bobot pada suatu dokumen maka semakin tinggi rangking dokumen tersebut. Semakin tinggi rangking suatu dokumen menunjukkan bahwa dokumen tersebut semakin relevan terhadap *query* yang dimasukkan.

*Average precision* dan *average* *recall* merupakan *evaluation matrix* yang digunakan pada penelitian ini. *Average precision* digunakan untuk mengetahui kemampuan model dalam mendapatkan hasil yang sesuai dengan *query* yang dimasukkan, sedangkan *average* *recall* untuk mendapatkan apa saja dokumen yang mengandung *query* yang dimasukkan.

Dari *query* yang telah dimasukkan, didapatkan *average precision* dan *average recall* seperti yang terdapat pada Tabel III.

Tabel III  
NILAI HASIL PENGUJIAN DENGAN PRECISION DAN RECALL

|  | *Average Precision* | *Average Recall* |
| --- | --- | --- |
| FS\_ANAMNESA | 0.522 | 0.816 |
| FS\_DIAGNOSA | 0.574 | 0.781 |
| ***Average*** | **0.548** | **0.796** |

Dari proses *retrieve* yang dilakukan didapatkan hasil yang cukup baik, yakni dengan *average* *precision* sebesar 54.8 % dan *average recall* sebesar 79.6 %. Jadi didapatkan *low precision* dan *high recall* menunjukkan bahwa model dapat memprediksi lebih banyak *true positive*.

1. Simpulan

Hasil evaluasi model berdasarkan 1000 dokumen rekam medis dan pengujian dengan 20 query pencarian dalam penelitian ini memberikan nilai rata-rata *precision* sebesar 0.548 dan nilai rata-rata *recall* sebesar 0.796. Nilai *low precision* dan *high recall* menunjukkan bahwa model dapat memprediksi lebih banyak *true positive*. Hasil *retrieval* dan nilai evaluasi bergantung pada banyak dokumen dalam korpus dan *query* yang dimasukkan, sehingga sangat mungkin banyak dokumen yang nantinya di-*retrieve* dan nilai *precision* dan *recall* bisa saja berubah jika ada penyesuaian jumlah dokumen dan penggunaan *query* pencarian lainnya.

1. Ucapan Terima-Kasih

Peneliti mengucapkan terima kasih kepada Departemen Ilmu Komputer dan Elektronika Fakultas MIPA Universitas Gadjah Mada, yang telah mendukung dan membantu dalam terlaksananya penelitian ini.

1. Referensi

[1] I. Rosadi and M. I. Purnama, “Analysis Of Time Analysis Of Outstanding Medical Records To Improve The Quality Of Services At Dustira Hospital, Cimahi Imam,” *Int. J. Heal. Sci.*, vol. 1, no. 1, pp. 1–5, 2021.

[2] Menteri Kesehatan Republik Indonesia, *Peraturan Menteri Kesehatan Republik Indonesia Nomor 24 Tahun 2022*, 2022.

[3] Menteri Kesehatan Republik Indonesia, *Peraturan Menteri Kesehatan Republik Indonesia Nomor 129/Menkes/SK/II/2008*, 2008.

[4] E. W. R. P, R. Indrawati, and L. Widjaja, “A Service Quality Review of Medical Record Department In Private Hospital, South Jakarta,” *J. Multidiscip. Acad.*, vol. 05, no. 02, pp. 101–105, 2021, [Online]. Available: <http://www.kemalapublisher.com/index.php/JoMA/article/view/554>.

[5] Indonesia. *Undang-Undang Nomor 29 Tahun 2004 tentang Praktik Kedokteran.* Lembaran Negara RI Tahun 2004 Nomor 116, Tambahan Lembaran RI Nomor 4431. Jakarta: Sekretariat Negara, 2004.

[6] Menteri Kesehatan Republik Indonesia, *Peraturan Menteri Kesehatan Republik Indonesia Nomor 269 Tahun 2008*, 2008.

[7] S. Buttcher, C. L. A. Clarke, and G. V. Cormack, *Information Retrieval Implementing and Evaluating Search Engines*. Cambridge, Massachusetts: The MIT Press, 2010.

[8] N. Hasanah, *Sistem Pencarian Skripsi Berbasis Information Retrieval di FASTIKOM UNSIQ*. Jurnal PPKM, 1(1), 105-113, 2017.

[9] F. Faridah, K. Munadi, and F. Arnia, “Aplikasi Histogram Discrete Cosine Transform (DCT) untuk Sistem Temu Kembali Citra Termal Berbasis Konten,” *Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)*, vol. 2, no. 1, 2019.

[10] Y. Kang, J. Li, J. Yang, Q. Wang, and Z. Sun, “Semantic analysis for enhanced medical retrieval,” *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2017.

[11] R. Shaptala, A. Kyselova, and G. Kyselov, “Exploring the vector space model for online courses,” *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, 2017.

[12] E. Wahyudi, S. Sfenrianto, M. J. Hakim, R. Subandi, O. R. Sulaeman, and R. Setiyawan, “Information retrieval system for searching JSON files with vector space model method,” *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)*, 2019.

[13] P. E. Mas'udia, M. D. Atmadja, and L. D. Mustafa, “Information Retrieval Tugas Akhir dan Perhitungan Kemiripan Dokumen Mengacu pada Abstrak Menggunakan Vector Space Model,” *Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer*, vol. 8, no. 1, pp. 355–362, 2017.

[14] A. T. Adiyanto and D. H. UN, Information Retrieval Sistem Kearsipan Pencarian Dokumen Di Dinas Pemberdayaan Perempuan Dan Perlindungan Anak Kota Semarang Menggunakan Metode Vector Space Model. JURNAL MAHAJANA INFORMASI, 7(1), pp.9-23, https://doi.org/10.51544/jurnalmi.v7i1.2538, 2022.

[15] R. C. N. Santi,S. Eniyati, R. Retnowati, and H. Yulianton, “PENGGUNAAN SISTEM TEMU KEMBALI DALAM PENCARIAN KATA UNTUK TERJEMAHAN AL QURAN”, 2019.

[16] B. Herwijayanti, D. E. Ratnawati, and L. Muflikhah, Klasifikasi Berita Online Dengan Menggunakan Pembobotan TF-IDF dan Cosine Similarity, Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 2, no. 1, p. 306-312, ISSN 2548-964X, Agu. 2017.

[17] M. Widiasri, E. Tjandra and L. M. Chandra, “Peningkatan Kinerja Pencarian Dokumen Tugas Akhir menggunakan Porter Stemmer Bahasa Indonesia dan Fungsi Peringkat Okapi BM25,” TEKNIKA, vol. 6, no. 1, pp. 54-60, 2017.

[18] E. S. Munif, and Y. Kristian, “PENCARIAN TEMA SEJENIS SINOPSIS NOVEL BAHASA INDONESIA DENGAN MENGGUNAKAN GVSM.” Joutica: Journal of Informatic Unisla, 6(2), 492-501, 2021.

[19] M. Chiny, M. Chihab, O. Bencharef, and Y. Chihab, “Netflix Recommendation System based on TF-IDF and Cosine Similarity Algorithms,” *Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning*, 2021.

[20] J. Han, M. Kamber, and J. Pei, *Data mining: Data mining Concepts and Techniques*, 3rd edition. 225 Wyman Street, Waltham, MA 02451, USA: Morgan Kaufmann, 2012.

**Salsabila Nurulfarah Mahmudah,** kelahiran kota Surakarta. Menerima Gelar Sarjana dari Universitas Muhammadiyah Surakarta untuk jurusan Teknik Informatika pada tahun 2020.

**Kartika Rizqi Nastiti,** kelahiran kabupaten Magetan. Menerima Gelar Sarjana dari Universitas Islam Indonesia untuk jurusan Informatika pada tahun 2019.

**Dian Novitaningrum,** kelahiran kota Klaten. Menerima Gelar Sarjana dari Universitas Muhammadiyah Surakarta untuk jurusan Teknik Informatika pada tahun 2020.

**Lukman Heryawan,** merupakan dosen Fakultas Matematika dan Ilmu Pengetahuan Alam pada Universitas Gadjah Mada dan Kepala Lab Rekayasa Perangkat Lunak dan Data. Minat penelitiannya meliputi *medical informatics,* *artificial intelligence,* dan *human-agent interaction*