Pengolahan Bahasa Alami untuk Anamnesis Penyakit pada Anak

Muchtadho Dhila Arafat[1], Sri Kusumadewi[2], Chanifah Indah Ratnasari[3]
[1,2,3]Universitas Islam Indonesia
E-mail: 21917032@students.uii.ac.id

Article Info
Article History
Received: 2025-01-10
Revised: 2025-02-20
Published: 2025-03-09

Keywords:
Anamnesis;
Cosine Similarity;
Natural Language Processing;
Pediatric Diseases;
TF-IDF.

Abstract
This study explores the application of Natural Language Processing (NLP) in the anamnesis process for pediatric diseases, particularly for fever-related conditions such as Dengue Fever, Acute Respiratory Infections (ARI), Typhoid Fever, and Urinary Tract Infections (UTI). Anamnesis is a crucial step in medical diagnosis that relies on verbal communication between patients and doctors. However, communication challenges can hinder accurate and timely diagnoses, leading to suboptimal medical treatment. This research applies NLP to interpret patients' verbal communication using a voice-to-text approach. The data, sourced from RSUD Bagas Waras Klaten, is in audio (.wav) format and processed using the speechRecognition library. After conversion to text, preprocessing techniques such as case folding, tokenizing, filtering, and stemming are applied. The study employs Term Frequency-Inverse Document Frequency (TF-IDF) and Cosine Similarity to identify symptom patterns. The research follows the Cross Industry Standard for Data Mining (CRISP-DM) framework. The outcome is a web-based system that allows users to consult regarding their symptoms. After NLP analysis, doctors can make faster and more accurate diagnoses. The anamnesis diagnosis results will be accessible to doctors, users, and administrators, enhancing efficiency in clinical decision-making.

Artikel Info
Sejarah Artikel
Diterima: 2025-01-10
Direvisi: 2025-02-20
Dipublikasi: 2025-03-09

Kata kunci:
Anamnesis;
Kesamaan Kosinus;
Pemrosesan Bahasa Alami;
Penyakit Anak;
TF-IDF.

Abstrak
Penelitian ini membahas penerapan Natural Language Processing (NLP) dalam proses anamnesis penyakit pada anak, khususnya untuk gejala demam seperti DBD, ISPA, Demam Tifoid, dan Infeksi Saluran Kemih. Anamnesis merupakan langkah awal dalam diagnosis yang bergantung pada komunikasi verbal antara pasien dan dokter. Namun, tantangan dalam komunikasi dapat menghambat diagnosis yang cepat dan tepat, berisiko pada penanganan medis yang kurang optimal. Penelitian ini menerapkan NLP untuk menginterpretasikan komunikasi verbal pasien menggunakan metode voice-to-text. Data yang digunakan berasal dari Klinik Margorejo dalam format suara (.wav), yang kemudian diproses menggunakan library speechRecognition. Setelah diubah ke dalam teks, data melalui tahap preprocessing seperti case folding, tokenizing, filtering, dan stemming. Selanjutnya, metode Term Frequency-Inverse Document Frequency (TF-IDF) dan Cosine Similarity digunakan untuk mengenali pola gejala. Kerangka kerja yang digunakan adalah Cross Industry Standard for Data Mining (CRISP-DM). Hasil penelitian ini berupa sistem berbasis website yang memungkinkan pasien melakukan konsultasi gejala. Setelah NLP menganalisis keluhan, dokter dapat melakukan diagnosis yang lebih akurat dan cepat. Hasil diagnosis anamnesis dapat diakses oleh dokter, pengguna, dan admin, sehingga meningkatkan efisiensi dalam pengambilan keputusan klinis.

I. PENDAHULUAN

NLP merupakan salah satu kecerdasan buatan (AI) yang berfokus kepada bagaimana bahasa manusia berinteraksi dengan komputer (Toruan et al., 2024) . Dalam kehidupan sehari-hari NLP dapat dimanfaatkan untuk memahami konteks dari teks dengan baik dan memberikan dampak yang signifikan pada berbagai aplikasi atau sistem seperti analisis teks, penerjemah bahasa, dan sistem interaksi manusia dan komputer (Samosir et al., 2022).

Meskipun sudah banyak diterapkan dalam bidang medis, NLP belum juga diterapkan terhadap proses anamnesis penyakit pada anak. Anamnesis merupakan salah satu langkah awal dalam diagnosis penyakit yang sangat bergantung pada komunikasi verbal antara pasien dan anak (Purba et al., 2024). Namun komunikasi verbal antara pasien dan dokter terkait gejala dari penyakit yang dialami seringkali menjadi tantangan karena volume informasi yang diterima beragam (Adista et al., 2023) sehingga dari permasalahan ini dapat

menghambat proses diagnosa yang cepat dan tepat. Ketidaktepatan dalam mendiagnosis dapat berakibat dalam penanganan media yang kurang optimal, memperpanjang waktu diagnosis, dan pada dampak yang serius dapat membahayakan keselamatan pasien khususnya anak-anak (Nurachman & Fitrianingrum, 2022).

Untuk membantu pakar dalam mengambil tindakan diagnosis yang cepat dan tepat diperlukan solusi yang dapat menginterpretasikan komunikasi verbal antara pasien dan dokter yaitu dengan menerapkan NLP (Cannavaro, 2023). Penerapan teknologi NLP dalam proses anamnesis tersebut dapat berfungsi untuk menginterpretasikan pertanyaan ataupun pernyataan yang kompleks sehingga dapat memahami kalimat tersebut dan memberikan jawaban yang lebih tepat serta meningkatkan kualitas komunikasi.

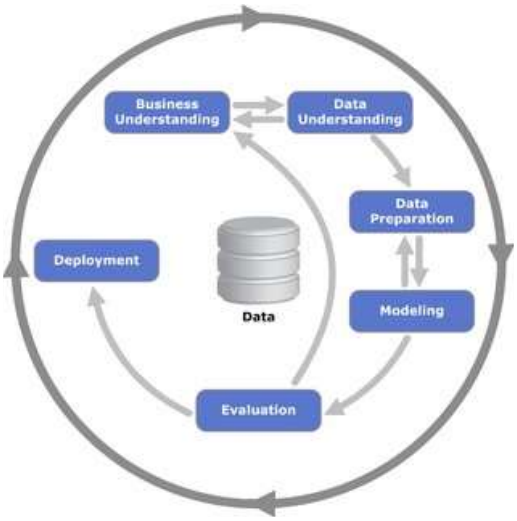
NLP sudah banyak diterapkan dalam bidang medis. Hal ini dikarenakan NLP mampu menangkap konteks dan makna dari bahasa alami yang seringkali sulit diproses melalui metode konvensional, serta pokok utama pemanfaatan NLP di bidang medis dapat meningkatkan akurasi dalam pengenalan istilah-istilah yang penting dan mampu memberikan dukungan dalam pengambilan keputusan klinis, serta mempercepat waktu atau meningkatkan efisiensi dalam melakukan analisis data besar pada catatan medis elektronik (EHR) (Gholipour et al., 2023).

Penelitian ini bertujuan untuk mengimplementasikan sistem *voice-to-text* dalam anamnesis penyakit pada anak menggunakan library *SpeechRecognition*. Sistem ini memungkinkan identifikasi gejala dari data anamnesis berbasis suara yang kemudian diolah dengan model NLP untuk mendeteksi penyakit dengan gejala demam. Selain itu, penelitian ini juga mengevaluasi efektivitas pengolahan bahasa alami berbasis website dalam mendukung proses diagnosis medis.

Diharapkan penelitian ini dapat membantu pihak terkait dalam mengekstrak informasi penyakit dengan gejala demam pada anak dari anamnesis berbentuk audio. Selain itu, penelitian ini dapat menjadi referensi dalam pengembangan metode identifikasi dan ekstraksi informasi penyakit menggunakan NLP, sehingga dapat meningkatkan efisiensi dan akurasi proses anamnesis berbasis teknologi.

II. METODE PENELITIAN

Langkah-langkah penelitian yang dilakukan dalam identifikasi ekstraksi informasi penyakit dengan gejala demam dapat dilihat pada Gambar 1. Penelitian ini menggunakan CRISP-DM sebagai kerangka kerja yang terstruktur untuk mengembangkan, mengimplementasikan, dan mengevaluasi model NLP yang digunakan untuk menganalisis teks medis atau anamnesis pasien (Zia et al., 2022).



Gambar 1. Tahapan CRIPS-DM

Dari Gambar 1. (Dhewayani et al., n.d.) dapat dijelaskan tahapan-tahapan CRISP-DM adalah sebagai berikut:

1. Business Understanding
   Business Understanding adalah proses menentukan tujuan bisnis, memahami situasi dan kondisi pada saat penelitian dan menetapkan sebuah tujuan dari penelitian yang dilakukan ke dalam permasalahan yang diselesaikan dengan data mining.
2. Data Understanding
   Data *understanding* adalah tahap persiapan, melakukan pengecekan terhadap data yang digunakan, mengumpulkan data awal serta melakukan identifikasi pada kualitas data. Dalam data *understanding*, data yang digunakan akan melalui proses deskripsi dari setiap fiturnya.
3. Data Preparation
   Data *preparation* merupakan proses yang dilakukan setelah data telah dikumpulkan. Pada tahap ini, data akan melalui proses identifikasi, pemilihan data, pembersihan data dan transformasi data.
4. Modelling
   Modeling merupakan tahap implementasi algoritma yang akan digunakan untuk melakukan pencarian, identifikasi, serta menghasilkan pola yang akan digunakan pada data penelitian.

5. Evaluation

Evaluation adalah suatu proses untuk melakukan pengukuran hasil evaluasi dari model yang telah diimplementasikan sebelumnya di tahap modelling. Hasil evaluasi tersebut menggambarkan proses dari data mining yang telah dilakukan dan mengukur model yang paling baik untuk digunakan.

6. Deployment

Deployment atau penyebaran merupakan proses menggunakan model yang dihasilkan sebelumnya. Dalam tahap ini terdapat 2 jenis kegiatan penyebaran yaitu melakukan perencanaan dan pemantauan dari penyebaran hasil yang dilakukan. Jenis kegiatan selanjutnya menyelesaikan tugas penutup dengan membuat laporan akhir dan melakukan tinjauan proyek. Kedua kegiatan ini dapat dilakukan semua atau menyelesaikan salah satunya.

III. HASIL DAN PEMBAHASAN

1. Business Understanding

Business understanding melakukan pemahaman tujuan berdasarkan pada perspektif bisnis kemudian diubah ke dalam definisi masalah *machine learning* hingga dilakukan penentuan solusi yang akan diusulkan untuk menangani permasalahan yang ada. Dalam penelitian ini tujuan penelitian yaitu mengimplementasikan NLP untuk anamnesis penyakit dengan gejala demam pada anak menggunakan algoritma TF IDF - *cosine similarity* dalam sebuah *website.*

Selanjutnya, dilakukan pengumpulan data yang terdiri dari data primer dan data sekunder. Data sekunder merupakan data terkait penelitian terdahulu, teori-teori yang digunakan sebagai bahan referensi dalam penelitian, seperti teori gejala demam, NLP, CRISP-DM, *text processing,* TF-IDF, *cosine similarity, flowchart,* UML, bahasa pemrograman PHP, *framework laravel, database,* MySQL, *blackbox testing, user acceptance test* (UAT). Sedangkan, data primer merupakan dataset anamnesis yang digunakan untuk ekstraksi informasi dari penyakit dalam gejala demam yang bersumber dari Klinik Margorejo berbentuk *voice*.

2. Data Understanding

Data understanding merupakan tahap untuk memahami data yang digunakan. Dalam penelitian ini menggunakan dataset yang bersumber dari data Klinik Margorejo, nantinya data tersebut akan berbentuk voice (.wav).

Dataset tidak menggunakan data uji karena pada penelitian ini akan menerapkan NLP dengan menggunakan *cosine similarity*. NLP dapat digolongkan kedalam *unsupervised learning* (Perdana et al., 2021). Data anamnesis yang ditanyakan dokter kepada pasien akan mencakup "*The Sacred Seven of history taking*". *The Sacred Seven of history* taking merupakan teknik memberikan informasi yang paling krusial dalam menentukan diagnosis dan pengobatan pasien.

Berdasarkan *The Sacred Seven of History Taking*, anamnesis dilakukan dengan beberapa parameter, yaitu *chronology* untuk menelusuri perjalanan penyakit sejak awal hingga saat wawancara, *bodily location* untuk menentukan lokasi serta penyebaran keluhan, dan *quality* untuk memahami bentuk serta karakteristik gejala. Selain itu, *quantity* digunakan untuk mengukur intensitas atau keparahan gejala, sementara setting mengidentifikasi situasi yang memicu munculnya gejala. *Aggravating or alleviating factors* membantu menentukan faktor yang memperburuk atau meredakan keluhan, sedangkan *associated manifestations* digunakan untuk mengidentifikasi gejala lain yang mungkin terkait dengan kondisi pasien.. (Idrus, 2020).

3. Data Preparation

Data *preparation* merupakan proses melakukan persiapan data dengan menyesuaikan dataset agar dapat sesuai dengan kebutuhan pada saat tahap pemodelan. Dataset yang digunakan berbentuk *voice* (.wav) dari hasil anamnesis pasien dan dokter. Data tersebut dilakukan deteksi suara dan bahasa oleh sistem yang nantinya akan dikonversi dari suara ke teks, dan hasil akhir akan menampilkan *output* (keluaran) berupa teks.

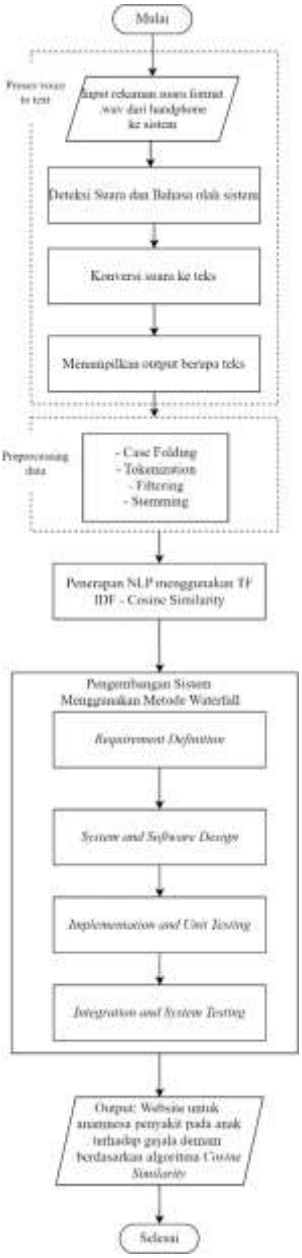
Setelah data berbentuk teks, selanjutnya dilakukan text *pre-processing*. *Preprocessing* yang digunakan dalam penelitian ini meliputi tahap *case folding* untuk mentransformasi kata yang memiliki huruf besar menjadi huruf kecil. Tahap *tokenize* untuk membagi atau memecah kata yang awalnya dalam bentuk kalimat utuh menjadi penggalan kata perkata. Tahap filtering, yang bertujuan untuk menghilangkan kata-kata umum yang sering muncul dalam jumlah banyak. Tahap *stemming* adalah tahap dimana setiap kata akan diubah dari imbuhan menjadi kata dasar, tahap ini diperlukan untuk mengurangi

jumlah indeks yang berbeda dari satu data sehingga kata yang mendapat imbuhan atau awalan akan kembali ke bentuk dasarnya.

4. Modelling

Pada tahap modelling, proses pembentukan model dilakukan untuk mendukung anamnesis penyakit pada anak menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *Cosine Similarity*. Model NLP diterapkan dalam sistem untuk mengolah data teks yang telah dikonversi dari suara pasien, sehingga memungkinkan identifikasi gejala berdasarkan pola yang telah dipelajari.

Proses ini mencakup pengolahan data dari tahap input hingga menghasilkan output dari sistem. Ilustrasi lebih lanjut mengenai alur kerja model dapat dilihat pada Gambar 2 berikut.



Gambar 2. Tahapan Modeling

a) Input Rekaman

Penelitian ini diawali dengan input data berupa *voice* berformat .wav yang diunggah dari handphone ke sistem, lalu diproses menggunakan library *Speech Recognition* untuk mendeteksi suara dan bahasa. Proses dimulai dengan pembuatan objek *Recognizer*, kemudian sistem membaca file audio menggunakan *sr.AudioFile()* dan mengonversinya menjadi teks dengan *r.recognize_google(audio)* melalui API Google Speech Recognition. Hasil konversi teks ditampilkan oleh sistem dan digunakan dalam tahap *text preprocessing* untuk analisis lebih lanjut, dengan dataset gejala yang telah dikonversi disajikan dalam Tabel 1 sebagai dasar pengolahan data selanjutnya.

Tabel 1. Dataset Gejala

Gejala	Kode Gejala
Demam Intermittent	G01
Demam Menggigil	G02
Lama Demam	G03
Sakit Kepala	G04
Mual	G05
Muntah	G06
Nyeri Sendi	G07
Nyeri Perut	G08
Nyeri Ulu Hati	G09
Bintik Merah	G10
Lidah Kotor	G11
Kejang	G12
Batuk	G13
Pilek	G14
Sesak Nafas	G15
Sulit Bab	G16
Bab Cair	G17
Bak Kemerahan	G18
Bak Sakit	G19
Nafsu Makan Turun	G20
Lemas	G21
Mimisan	G22

Setelah melakukan input Dataset, maka selanjutnya adalah input Dataset yang akan dijadikan *Query*, Untuk Dataset *Query* disini yang akan digunakan adalah Input keluhan yang terdapat pada Tabel 2.

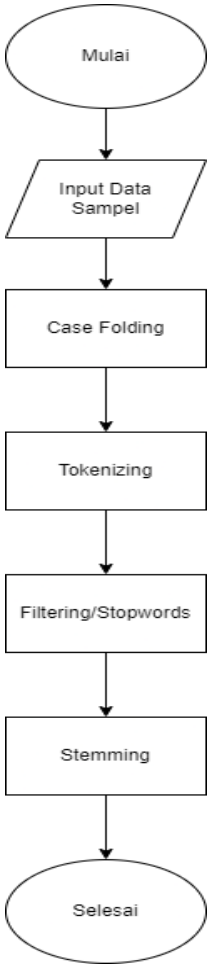
Tabel 2. Input Keluhan

No	Keluhan
1	Anak saya panas naik turun semalam, ada nyeri perut dan mual, kadang muntah, sekarang anak saya pusing dan lemas
2	Anak saya buang air kecil sakit dan kadang keluar darah
3	Anak saya demam sudah 4 hari dok, setiap makan minum selalu mual dan muntah. Badannya dingin
4	Demam sudah 3 hari lamanya, terkadang mual mual dok

5	Anak saya sering mual dan kalau kencing sakit dok, ini sedang demam tinggi

b) Preprocessing Data

Langkah *preprocessing* yang dilakukan adalah Pengolahan kata yang terdiri dari tahapan *case folding, tokenizing, filtering*, dan *stemming*, langkah tersebut dapat di lihat pada Gambar 3 berikut.



Gambar 3. Preprocessing

1) Case Folding

Case Folding adalah tahapan yang bertujuan untuk mengubah semua respon menjadi huruf kecil, pada proses ini, karakter 'A'-'Z' yang terdapat pada data diubah menjadi karakter 'a'-'z'. Sementara itu, karakter lain yang bukan huruf dan angka, seperti tanda baca dan spasi, dianggap sebagai pembatas. Tabel 3 adalah perbandingan Original Dataset dan hasil *Case Folding.*

Tabel 3. Case Folding

Index	Original	Casefolding
0	Anak saya panas naik turun semalam, ada nyeri perut dan mual, kadang muntah, sekarang anak saya pusing dan lemas	anak saya panas naik turun semalam ada nyeri perut dan mual kadang muntah sekarang anak saya pusing dan lemas
1	Anak saya buang air kecil sakit dan kadang keluar darah	anak saya buang air kecil sakit dan kadang keluar darah
2	Anak saya demam sudah 4 hari dok, setiap makan minum selalu mual dan muntah. Badannya dingin	anak saya demam sudah hari dok setiap makan minum selalu mual dan muntah badannya dingin
3	Demam sudah 3 hari lamanya, terkadang mual mual dok	demam sudah hari lamanya terkadang mual mual dok
4	Anak saya sering mual dan kalau kencing sakit dok, ini sedang demam tinggi	anak saya sering mual dan kalau kencing sakit dok ini sedang demam tinggi

2) Tokenizing

Tabel 4 menunjukkan hasil tokenisasi.

Tabel 4. Tokenizing

Index	Casefolding	Tokenizing
0	anak saya panas naik turun semalam ada nyeri perut dan mual kadang muntah sekarang anak saya pusing dan lemas	['anak', 'saya', 'panas', 'naik', 'turun', 'semalam', 'ada', 'nyeri', 'perut', 'dan', 'mual', 'kadang', 'muntah', 'sekarang', 'anak', 'saya', 'pusing', 'dan', 'lemas']
1	anak saya buang air kecil sakit dan kadang keluar darah	['anak', 'saya', 'buang', 'air', 'kecil', 'sakit', 'dan', 'kadang', 'keluar', 'darah']
2	anak saya demam sudah hari dok setiap makan minum selalu mual dan muntah badannya dingin	['anak', 'saya', 'demam', 'sudah', 'hari', 'dok', 'setiap', 'makan', 'minum', 'selalu', 'mual', 'dan', 'muntah', 'badannya', 'dingin']
3	demam sudah hari lamanya terkadang mual mual dok	['demam', 'sudah', 'hari', 'lamanya', 'terkadang', 'mual', 'mual', 'dok']
4	anak saya sering mual dan kalau kencing sakit dok	['anak', 'saya', 'sering', 'mual', 'dan', 'kalau',

	ini sedang demam tinggi	'kencing', 'sakit', 'dok', 'ini', 'sedang', 'demam', 'tinggi']

3) Filtering

Tabel 5. menunjukkan hasil filtering.

Tabel 5. Filtering

Index	Tokenizing	Stopwords
0	['anak', 'saya', 'panas', 'naik', 'turun', 'semalam', 'ada', 'nyeri', 'perut', 'dan', 'mual', 'kadang', 'muntah', 'sekarang', 'anak', 'saya', 'pusing', 'dan', 'lemas']	['anak', 'saya', 'panas', 'naik', 'turun', 'semalam', 'ada', 'nyeri', 'perut', 'dan', 'mual', 'kadang', 'muntah', 'sekarang', 'anak', 'saya', 'pusing', 'dan', 'lemas']
1	['anak', 'saya', 'buang', 'air', 'kecil', 'sakit', 'dan', 'kadang', 'keluar', 'darah']	['anak', 'saya', 'buang', 'air', 'kecil', 'sakit', 'dan', 'kadang', 'keluar', 'darah']
2	['anak', 'saya', 'demam', 'sudah', 'hari', 'dok', 'setiap', 'makan', 'minum', 'selalu', 'mual', 'dan', 'muntah', 'badannya', 'dingin']	['anak', 'saya', 'demam', 'sudah', 'hari', 'dok', 'setiap', 'makan', 'minum', 'selalu', 'mual', 'dan', 'muntah', 'badannya', 'dingin']
3	['demam', 'sudah', 'hari', 'lamanya', 'terkadang', 'mual', 'mual', 'dok']	['demam', 'sudah', 'hari', 'lamanya', 'terkadang', 'mual', 'mual', 'dok']
4	['anak', 'saya', 'sering', 'mual', 'dan', 'kalau', 'kencing', 'sakit', 'dok', 'ini', 'sedang', 'demam', 'tinggi']	['anak', 'saya', 'sering', 'mual', 'dan', 'kalau', 'kencing', 'sakit', 'dok', 'ini', 'sedang', 'demam', 'tinggi']

4) Stemming

Tabel 6. Stemming

Index	Tokenizing	Stopwords
0	['anak', 'saya', 'panas', 'naik', 'turun', 'semalam', 'ada', 'nyeri', 'perut', 'dan', 'mual', 'kadang', 'muntah', 'sekarang', 'anak', 'saya', 'pusing', 'dan', 'lemas']	turun malam ada nyeri perut dan mual kadang muntah sekarang anak saya pusing dan lemas']
1	['anak', 'saya', 'buang', 'air', 'kecil', 'sakit', 'dan', 'kadang', 'keluar', 'darah']	['anak saya buang air kecil sakit dan kadang keluar darah']
2	['anak', 'saya', 'demam', 'sudah', 'hari', 'dok', 'setiap', 'makan', 'minum', 'selalu', 'mual', 'dan', 'muntah', 'badannya', 'dingin']	['anak saya demam sudah hari dok tiap makan minum selalu mual dan muntah badan dingin']
3	['demam', 'sudah', 'hari', 'lamanya', 'terkadang', 'mual', 'mual', 'dok']	['demam sudah hari lama terkadang mual mual dok']
4	['anak', 'saya', 'sering', 'mual', 'dan', 'kalau', 'kencing', 'sakit', 'dok', 'ini', 'sedang', 'demam', 'tinggi']	['anak saya sering mual dan kalau kencing sakit dok ini sedang demam tinggi']

5) Hasil Preprocessing

Tabel 7 menunjukkan hasil preprocessing.

Tabel 7. Hasil Preprocessing

Index	Hasil Preprocessing
0	anak saya panas naik turun malam ada nyeri perut dan mual kadang muntah sekarang anak saya pusing dan lemas
1	anak saya buang air kecil sakit dan kadang keluar darah
2	anak saya demam sudah hari dok tiap makan minum selalu mual dan muntah badan dingin
3	demam sudah hari lama terkadang mual mual dok
4	anak saya sering mual dan kalau kencing sakit dok ini sedang demam tinggi

c) TF – IDF

Setelah tahap *preprocessing*, melalui beberapa langkah untuk

mengimplementasikan pemfilteran berbasis konten, termasuk menggunakan *cosine similarity* dan TF-IDF untuk menentukan kemiripan dan tinggi angka yang mirip (Rianti et al., 2024).

TF-IDF merupakan kombinasi dari *term frequency* (TF) dan *inverse document frequency* (IDF), di mana TF merupakan salah satu *term weighting scheme* (TWS) paling sederhana karena hanya bergantung pada jumlah kemunculan term dalam dokumen tertentu, sehingga kurang efektif dalam membedakan dokumen relevan dari yang tidak relevan (Jiang et al., 2021). Untuk mengatasi keterbatasan ini, IDF diperkenalkan dengan mempertimbangkan frekuensi koleksi guna meningkatkan kapasitas diskriminatif suatu term dalam klasifikasi teks. IDF dikembangkan dari *document frequency* (DF), yang mengukur jumlah dokumen tempat suatu term muncul, dengan asumsi bahwa term yang lebih jarang muncul di berbagai dokumen dianggap lebih penting dibandingkan term yang sering muncul.

$$TF - IDF = tf_t * IDF_t \quad (1)$$

$$IDF = \log\left(\frac{D}{DF}\right) \quad (2)$$

Keterangan:

tf_t = *Term Frequency*

IDF = *Inverse Document Frequency*

D = Total Dokumen

DF = Total kemunculan kata pada dari seluruh dokumen

Pada Sklearn, formula IDF yang digunakan agak berbeda. Formula IDF pada sklearn dapat dilihat pada persamaan (3).

$$IDF = \ln\left(\frac{1+D}{1+DF}\right) + 1 \quad (3)$$

Selanjutnya pada sklearn, nilai TF-IDF juga akan dilakukan normalisasi L2 yang memiliki formula dapat dilihat pada (2.4).

$$L2 - Norm = \sqrt[2]{TF - IDF_1^2 + TF - IDF_2^2 + \cdots + TF - IDF_i^2} \quad (4)$$

Sehingga hasil dari TF-IDF L2 Norm adalah seperti pada persamaan (5)

$$TF - IDF\ L2 - Norm = \frac{TF - IDF\ sebelum\ Norm}{L2 - Norm} \quad (5)$$

Hasil TF-IDF Bisa dilihat pada Tabel berikut:

Tabel 8. Hasil TF-IDF

	bab	bak	...	ulu
Query	0	0	...	0
Nyeri Ulu Hati	0	0	...	0.577
Sulit Bab	0.707	0	...	0
Bab Cair	0.707	0	...	0
Bak Kemerahan	0	0.707	...	0
Bak Sakit	0	0.707	...	0

d) Cosine Similarity

Pada tahapan ini akan menghitung similaritas dengan jenis *Cosine Similarity*. Penggunaan *cosine similarity* dikarenakan dari hasil TF-IDF yang digunakan dimana formula *Cosine Similarity* tidak memanfaatkan nilai rata-rata suatu item.

Cosine Similarity ditentukan sebagai kosinus sudut θ yang terbentuk antara vektor-vektor. Formula *Cosine Similarity* memiliki rumus seperti pada persamaan (6) berikut (Al Rasyid et al., 2024):

$$Sim(A,B) = \frac{\sum n(A) * n(B)}{\sqrt{\sum n(A)^2} * \sqrt{\sum n(B)^2}} \quad (6)$$

Keterangan:

A = Item A (contoh D1)

B = Item B (contoh D2)

Perhitungan Similarity dilakukan pada *Query* terhadap Semua Gejala, dan didapatkan hasil perhitungan *Similarity Query* terhadap 22 gejala terlihat pada Tabel 9.

Tabel 9. Hasil Similaritas

Gejala	Similaritas
Mual	0.149
Muntah	0.149
Nyeri Sendi	0.089
Nyeri Perut	0.177
Nyeri Ulu Hati	0.065
...	....
Nafsu Makan Turun	0.065
Lemas	0.149

Pada penelitian ini di tetapkan untuk ambang batas yaitu > 0.15, hal ini dikarenakan pada manualisasi ini maksimal nilai *Cosine* adalah 1 dan apabila

ambang batas ditetapkan berdasar penelitian (Jepriana & Hanief, n.d.) yang memiliki ambang batas 2.75 maka apabila nilai batas ambang ditetapkan 2.75 tidak akan mendapatkan gejala similaritas, dan menghasilkan Gejala Similah berdasarkan sorting Top N-Recommend terlihat pada Tabel 10.

Tabel 10. Hasil Top N-Recommend

Gejala	Similaritas
Nyeri Perut	0.177
Mual	0.149
Muntah	0.149
Lemas	0.149
Nyeri Sendi	0.089
Nyeri Ulu Hati	0.065
Nafsu Makan Turun	0.065
...	...

5. Evaluation

Pada tahap *evaluation* dilakukan evaluasi terhadap sistem yang dibuat, dalam hal ini berupa *website* dengan melakukan pengujian menggunakan *black box testing*

6. Deployment

Melakukan penyebaran hasil penelitian dengan menyebarkan *website* untuk digunakan pengguna dan laporan hasil penelitian serta artikel jurnal hasil penelitian yang berisi hasil ekstraksi informasi dari data anamnesis tersebut dalam bentuk ilmu pengetahuan.

IV. SIMPULAN DAN SARAN

A. Simpulan

Berdasarkan analisis menggunakan TF-IDF dan *Cosine Similarity*, pernyataan anamnesis yang diberikan memiliki tingkat kemiripan dengan beberapa gejala yang telah ditentukan. Dari hasil analisis ini, dapat disimpulkan bahwa penggunaan metode *Natural Language Processing* (NLP) dengan TF-IDF dan *Cosine Similarity* memungkinkan sistem untuk mengenali pola gejala dalam teks input pasien secara lebih akurat. Dengan pendekatan ini, sistem dapat mengolah data *voice-to-text* menjadi representasi numerik yang membantu dalam klasifikasi gejala berdasarkan tingkat relevansinya. Hal ini menunjukkan bahwa metode NLP dapat digunakan secara efektif untuk mendukung proses anamnesis penyakit berbasis teknologi, sehingga dapat membantu tenaga medis dalam melakukan deteksi awal penyakit dengan lebih cepat dan akurat.

Untuk meningkatkan akurasi dan penerapan metode ini, dapat dilakukan pengujian lebih lanjut dengan dataset yang lebih besar dan lebih bervariasi serta penerapan algoritma lain untuk memperkaya pemahaman sistem terhadap konteks gejala medis, sehingga dapat memberikan hasil yang lebih optimal dalam mendukung keputusan medis.

B. Saran

Pembahasan terkait penelitian ini masih sangat terbatas dan membutuhkan banyak masukan, saran untuk penulis selanjutnya adalah mengkaji lebih dalam dan secara komprehensif tentang Pengolahan Bahasa Alami untuk Anamnesis Penyakit pada Anak.

DAFTAR RUJUKAN

Adista, A., Biokimia, B., Kedokteran Universitas Syiah Kuala, F., Aceh, B., Anatomi Histologi, B., Pendidikan Dokter, B., & Neurologi, B. (2023). TINJAUAN BIOMOLEKULAR PATOFISIOLOGI DAN PENEGAKAN DIAGNOSIS CANCER INDUCED BONE PAIN (CIBP). In *Jurnal Sinaps* (Vol. 6, Issue 3).

Al Rasyid, R., Handayani, D., & Ningsih, U. (2024). Penerapan Algoritma TF-IDF dan Cosine Similarity untuk Query Pencarian Pada Dataset Destinasi Wisata. *Jurnal Teknologi Informasi Dan Komunikasi)*, *8*(1), 2024. https://doi.org/10.35870/jti

Cannavaro, N. (2023). Aplikasi Chatbot untuk Layanan Akademik Menggunakan Platform RASA Open Source dengan Fitur Two Stage Fallback. *Jurnal Ilmu Komputer Dan Informatika*, *3*(1), 53–64. https://doi.org/10.54082/jiki.73

Dhewayani, F. N., Amelia, D., Alifah, D. N., Sari, B. N., Jajuli, M., HSRonggo Waluyo, J., Telukjambe Timur, K., Karawang, K., & Barat, J. (n.d.). Implementasi K-Means Clustering untuk Pengelompokkan Daerah Rawan Bencana Kebakaran Menggunakan Model CRISP-DM. *Jurnal Teknologi Dan Informasi*. https://doi.org/10.34010/jati.v12i1

Gholipour, M., Khajouei, R., Amiri, P., Hajesmaeel Gohari, S., & Ahmadian, L. (2023). Extracting cancer concepts from clinical notes using natural language processing: a systematic review. *BMC Bioinformatics*,

24(1). https://doi.org/10.1186/s12859-023-05480-0

Idrus, F. (2020). *MANUAL-CSL-1-Panduan-Pembelajaran-Komunikasi-Dokter*. Departemen Ilmu Kedokteran Jiwa (Psikiatri) Fakultas Kedokteran Universitas Hasanuddin Makassar.

Jepriana, W., & Hanief, S. (n.d.). *METODE ITEM-BASED COLLABORATIVE FILTERING UNTUK MODEL SISTEM REKOMENDASI KONSENTRASI PENJURUSAN DI STMIK STIKOM BALI*.

Jiang, Z., Gao, B., He, Y., Han, Y., Doyle, P., & Zhu, Q. (2021). Text Classification Using Novel Term Weighting Scheme-Based Improved TF-IDF for Internet Media Reports. *Mathematical Problems in Engineering*, *2021*. https://doi.org/10.1155/2021/6619088

Nurachman, M. T., & Fitrianingrum, I. (2022). Pengaruh Komunikasi Efektif Terhadap Kejadian Tidak Diharapkan (KTD). *Jurnal Cerebellum*, *8*(2), 12–15. https://doi.org/10.26418/jc.v%vi%i.54151

Perdana, K., Pricillia, T., & Zulfachmi. (2021). *Optimasi TextBlob Menggunakan Support Vector Machine untuk Analisis Sentimen (Studi Kasus Layanan Telkomsel)*. 13–15.

Purba, M., Penmaley, F., & Dame Panjaitan, J. (2024). Dugaan Pelanggaran Disiplin Terbanyak Akibat Kurangnya Komunikasi Dokter dan Pasien. *Jurnal Global Ilmiah*, *1*(5).

Rianti, A., Majid, N. W. A., & Fauzi, A. (2024). MACHINE LEARNING JOURNAL ARTICLE RECOMMENDATION SYSTEM USING CONTENT BASED FILTERING. *JUTI : Jurnal Ilmiah Teknologi Informasi*, *22*. https://doi.org/http://dx.doi.org/10.12962/j24068535.v22i1.a1193

Riany, A. F., & Testiana, G. (2023). Penerapan Data Mining untuk Klasifikasi Penyakit Stroke Menggunakan Algoritma Naïve Bayes. *Jurnal SAINTEKOM*, *13*(1), 42–54. https://doi.org/10.33020/saintekom.v13i1.352

Samosir, F. V. P., Toba, H., & Ayub, M. (2022). BESKlus : BERT Extractive Summarization with K-Means Clustering in Scientific Paper. *Jurnal Teknik Informatika Dan Sistem Informasi*, *8*(1). https://doi.org/10.28932/jutisi.v8i1.4474

Toruan, A. M. L., Panjaitan, B. M., Tumangger, E. M. K., Ulfa, N. R., & Panjaitan, G. D. (2024). Penggunaan NLP dalam Analisis Sentimen untuk Kepuasan Pelanggan pada Penggunaan E-commerce : Lazada. *SAINTEK: Jurnal Sains, Teknologi & Komputer*, *1*, 18–20.

Zia, A., Aziz, M., Popa, I., Khan, S. A., Hamedani, A. F., & Asif, A. R. (2022). Artificial Intelligence-Based Medical Data Mining. In *Journal of Personalized Medicine* (Vol. 12, Issue 9). MDPI. https://doi.org/10.3390/jpm12091359