JNATIA Volume 1, Nomor 4, Agustus 2023 p-ISSN: 2986-3929
Jurnal Nasional Teknologi Informasi dan Aplikasinya

Analisis Data Berbentuk Teks dalam Sistem Diagnosis Penyakit dengan Supervised Learning

I Gusti Ngurah Bagus Ferry Mahayudha[a1], Ida Bagus Gede Dwidasmara[a2]

Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Udayana
Jalan Raya Kampus Udayana, Bukit Jimbaran, Kuta Selatan, Badung, Bali Indonesia
[1]gunggus2367@gmail.com
[2]dwidasmara@unud.ac.id

Abstract

In computer science, text refers to a sequence of characters that can be represented and processed by a computer. It is the basic unit of data for representing human-readable information, such as letters, numbers, symbols, and spaces. In computer programming, text is typically represented as a string of characters. Textual data can be stored in variables, manipulated using various string operations, and displayed to users through input/output operations.Text plays a crucial role in many areas of computer science, including natural language processing, information retrieval, data mining, and text-based communication systems like email, chat applications, and social media. It serves as a fundamental component for storing, analyzing, and processing vast amounts of textual information in various applications.

Keywords: *string of characters, textual data, NLP, information retrieval, data mining*

1. Pendahuluan

Dokter adalah sebutan untuk seorang profesional medis yang memiliki kualifikasi dan pelatihan yang memadai dalam ilmu kedokteran. Mereka memiliki pengetahuan dan keahlian untuk mendiagnosis, merawat, dan mengelola berbagai masalah kesehatan manusia. Tugas utama seorang dokter adalah untuk merawat pasien, mendiagnosis kondisi medis, dan memberikan pengobatan yang sesuai. Dalam penangan pasien yang memerlukan obat atau pengobatan tertentu, dokter akan memberi resep dokter yang dapat digunakan untuk membeli obat di apotek. Sistem pelayanan dokter, atau juga dikenal sebagai sistem perawatan Kesehatan mencakup berbagai komponen, seperti dokter, rumah sakit, klinik, pusat kesehatan, asuransi kesehatan, dan lembaga lain yang terlibat dalam perawatan kesehatan. Sistem pelayanan dokter dapat dilakukan secara otomatis dengan penggunaan machine learning, sehingga dapat membantu dokter dalam memberi pelayanan berupa diagnosis penyakit serta resep dokter kepada pasien selama 24 jam. Rekomendasi obat yang dilakukan dapat melalui proses awal, yaitu mengetahui penyakit yang dialami pasien jika pasien tidak melakukan konsultasi terlebih dahulu ke dokter. Cara agar mengetahui penyakit yang dialami pasien adalah menggunakan machine learning. Penggunaan machine learning dalam mendiagnosis penyakit sudah termasuk ke dalam data mining, karena diagnosis penyakit merupakan proses untuk mencari relasi dan informasi yang terdapat di dalam data.

1.1. Data Mining

Data mining adalah proses pencarian informasi dalam kumpulan data yang memiliki kapasitas yang besar.[6] Tujuan dari data mining adalah untuk mencari pola tersembunyi atau korelasi yang bisa menyediakan pengetahuan dan membantu dalam pengambilan keputusan. Secara garis besar metode dalam data mining dapat dibagi kedalam lima bagian yaitu klasifikasi, prediksi, asosiasi, estimasi dan klastering.[6]

1.2. Natural Language Processing

Natural Language Processing (NLP) adalah cabang dari kecerdasan buatan dan linguistik komputasional yang berfokus pada pemahaman, interpretasi, dan generasi bahasa manusia secara alami. NLP mencakup berbagai teknik dan algoritma yang dirancang untuk memungkinkan komputer untuk berinteraksi dengan, memahami, dan memanipulasi teks atau data bahasa manusia. Tujuan utama NLP adalah untuk memungkinkan komputer untuk memahami dan menganalisis bahasa manusia dalam bentuk teks atau ucapan.

1.3. Supervised Learning

Supervised Learning (Pembelajaran Terawasi) adalah model machine learning yang mempelajari data yang sudah diketahui labelnya (misalnya, klasifikasi gambar dengan label kucing atau anjing). Model ini belajar untuk menghubungkan input dengan output yang diinginkan dan kemudian dapat melakukan prediksi pada data baru.

1.4 Linear SVM

Linear SVM (Support Vector Machine) adalah salah satu algoritma pembelajaran mesin yang digunakan untuk klasifikasi dan pemisahan data. SVM adalah algoritma yang efektif untuk pemisahan dua kelas atau lebih dengan membangun hyperplane (bidang pemisah) dalam ruang fitur. Keuntungan utama dari Linear SVM adalah kemampuannya untuk mengatasi masalah klasifikasi yang kompleks dan penanganan dimensi yang tinggi. Algoritma ini juga efisien dalam penggunaan memori dan memiliki sifat matematis yang kuat.

2. Metode Penelitian

Metode penelitian sangat menentukan hasil penelitian yang akan dilakukan atau dikerjakan, karena terkait cara yang baik dan benar dalam proses pengumpulan data, analisis data dan juga dalam pengambilan keputusan dari hasil penelitian. Adapun metode penelitian yang digunakan adalah sebagai berikut:

2.1. Teknik Pengumpulan Data

Teknik pengumpulan data yang digunakan untuk memperoleh data adalah studi pustaka. Penulis melakukan studi pustaka pada website Kaggle.com untuk mencari data yang akan dipecah menjadi data latih dan data uji yang dapat digunakan untuk membuat data model machine learning. Namun sebelum itu, jumlah data akan dibatasi untuk mengoptimalkan komputasi.

2.2. Model Pengembangan Sistem

Prototyping adalah proses merancang sebuah prototype dimana prototype sendiri adalah sebuah model dari sebuah model produk yang mungkin belum memiliki semua fitur produk sesungguhnya namun sudah memiliki fitur –fitur utama dari produk sesungguhnya dan biasa digunakan untuk keperluan testing/uji coba untuk bahan uji coba sebelum berlanjut ke fase pembuatan produk sesungguhnya. [4]

2.3. Analisis Kebutuhan Sistem

Sistem yang dapat berjalan secara otomatis tentunya memerlukan metode pembelajaran mesin yang mumpuni. Dengan kata lain, sistem yang dibuat berupa model machine learning yang dapat memprediksi secara akurat. Model tersebut akan digunakan untuk memprediksi data uji dari inputan aplikasi. Data user melalui inputan aplikasi akan diproses terlebih dahulu agar mudah diprediksi. Setelah data uji diprediksi, data uji dan hasil prediksi data tersebut akan dimasukkan ke dalam database untuk disimpan. Urutan kebutuhan sistem tersebut meliputi:

a. Aplikasi

Untuk mendapat data dari user, diperlukan adanya aplikasi. Aplikasi yang dibuat berupa web application. Aplikasi ini akan digunakan oleh user untuk mendiagnosis penyakit yang diderita oleh user.

b. Application Programming Interface

API digunakan sebagai alat komunikasi dengan sistem diagnosis. API akan mengirim data dari aplikasi berupa JSON kepada sistem diagnosis yang kemudian akan menghasilkan diagnosis penyakit.

c. Sistem Diagnosis

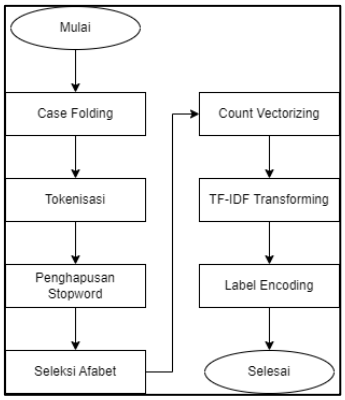
Sistem diagnosis yang dibuat meliputi fungsi pembersihan fitur dan simpanan berupa objek ekstraksi fitur dan model klasifikasi machine learning yang memiliki akurasi dengan skor 75% atau lebih. Dengan data bersih yang sudah melalui pra-pemrosesan dan sistem yang akurat, diagnosis dapat dilakukan dengan baik.

d. Database

Untuk menyimpan data uji beserta hasil prediksi data uji dari API, diperlukan database dengan kapasitas penyimpanan yang cukup. Data dalam database dapat dikirim kembali ke aplikasi sehingga user dapat menikmati sistem diagnosis aplikasi.

2.4. Pra-pemrosesan Data

Setelah data berhasil dikumpul, dilakukan pra-pemrosesan data yang bertujuan untuk mempermudah pembobotan menjadi vektor atau matriks TF-IDF.



Gambar 1. Flowchart Pra-pemrosesan Data

Berbagai tahapan yang dilakukan dalam pra-pemrosesan data adalah:

a. Pembersihan Data

Data yang akan di-preprocessing harus dibersihkan terlebih dahulu agar tidak terjadi kesalahan dalam preprocessing. Alur pembersihan data adalah sebagai berikut:

1. Case-Folding

Case-folding digunakan untuk mengubah huruf kapital menjadi huruf kecil untuk mempermudah tokenisasi.

2. **Tokenisasi**

   Tokenisasi dilakukan untuk mempermudah stemming setiap kata/token dalam suatu kalimat yang sudah menjadi list kata.

3. **Penghapusan Stopword**

   Stopword harus dihilangkan agar meminimalisir komputasi, sehingga pembuatan model klasifikasi dapat dilakukan secara lebih efisien.

4. **Seleksi Alfabet**

   Karena setelah tokenisasi masih ada substring berupa angka, maka dilakukan seleksi alfabet menggunakan metode regular expression sebagai tahap terakhir dalam pembersihan data.

b. Ekstraksi Fitur

Pembobotan data bersih penting dilakukan agar data bersih dapat dikomputasi oleh algoritma machine learning. Alur pembobotan data adalah sebagai berikut:

1. **Count Vectorizing**

   Vektorisasi data teks yang sudah bersih dilakukan dengan metode Count Vectorizer. Vektor yang dihasilkan memiliki besaran panjang yang sama dengan seluruh kata unik dalam data. Selanjutnya, akan dilakukan ekstraksi fitur TF-IDF.

2. **TF-IDF Vectorizing**

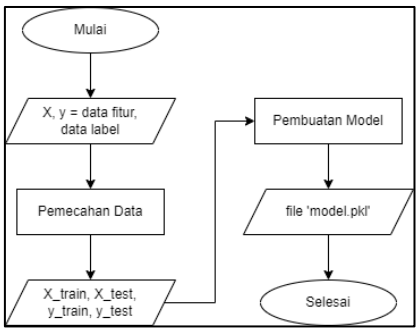
   Setelah data mengalami pembobotan dengan metode Count Vectorizer, data akan dibobot dengan metode TF-IDF Vectorizer untuk mendapat ekstraksi fitur berupa vektor TF-IDF yang akan digunakan untuk melatih model.

3. **Label Encoding**

   Label yang masih betipe kategorik harus diubah menjadi numerik dengan metode Label Encoding. Label yang sudah bertipe numerik akan mempermudah komputasi yang dilakukan untuk melatih model.

2.5. Pembuatan Model Klasifikasi

Dalam pembuatan model digunakan modul pickle. Model akan diubah menjadi file dengan ekstensi .pkl. File yang memuat model akan disisipkan ke dalam sistem sehingga proses prediksi atau diagnosis dapat dilakukan.



Gambar 2. Flowchart Pembuatan Model Klasifikasi

Langkah-langkah dalam pembuatan model adalah sebagai berikut:

1. **Pemecahan Data**

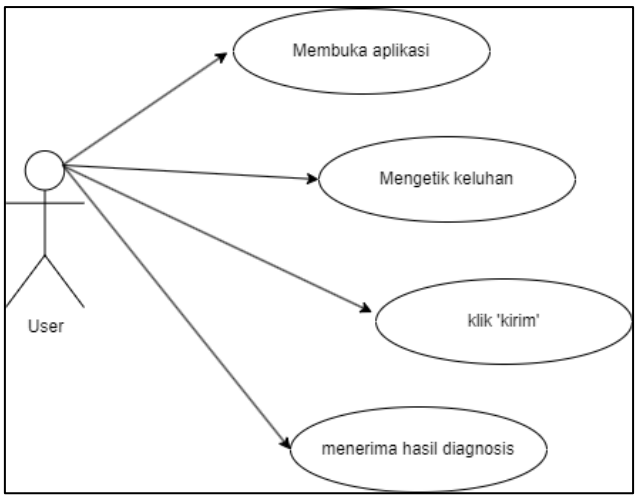
   Dataset dipecah menjadi 2 bagian terlebih dahulu, yaitu data fitur dan data label. Lalu, masing-masing data fitur dan data label akan dipecah menjadi 2, sehingga terdapat 4 bagian dataset, yaitu data fitur latih, data fitur uji, data label latih dan data label uji.

2. **Pembuatan Model**

   Model yang dibuat menggunakan algoritma Linear SVM karena memiliki waktu fitting yang relative cepat untuk di-fitting dengan data dengan banyak fitur serta dengan akurasi yang memuaskan. Fitting model dilakukan menggunakan data fitur latih dan data label latih.

3. Hasil dan Pembahasan

3.1. Use Case Diagram



Gambar 3. Use Case Diagram

Tabel 1. Deskripsi Use Case Diagram

Use Case Diagnosis	
Sasaran	User dapat mendapat diagnosis penyakit
Persyaratan	User memberi keluhan tentang gejala yang dialami
Pasca Kondisi	Website memberi diagnosis penyakit yang benar
Kondisi Akhir yang Gagal	Website memberi diagnosis penyakit yang tidak benar
Aliran Utama/Jalur Dasar	1. User membuka aplikasi 2. User mengetik keluhan pada text input 3. User mengklik tombol ‘kirim’

3.2. Hasil Prediksi dan Evaluasi Model

Skor akurasi dari model yang dibuat mencapai angka 77.4 %. Dengan demikian, akurasi dianggap baik dan dapat melakukan diagnosis. Berikut merupakan hasil diagnosis dari model klasifikasi diagnosis penyakit.

Tabel 2. Hasil Prediksi Data Baru

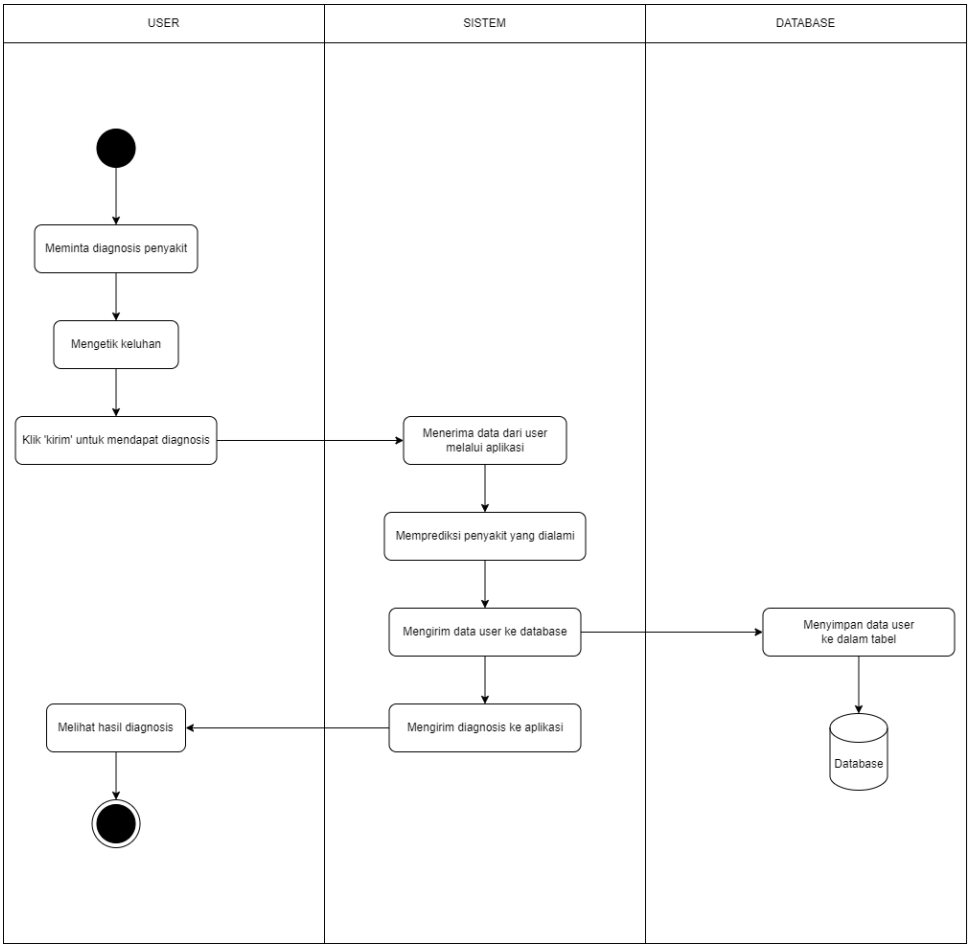
No	Keluhan	Diagnosis
1.	I am so anxious and i often panic	Anxiety
2.	I am so stressed because of the problem which hit me	Anxiety and Stress
3.	dad has been coughing for 4 hours	Cough
4.	my gf can’t sleep at night, and she has been sleepless for 2 days	Insomnia
5.	my sister needs to lose some weight	Obesity

Evaluasi model dilakukan dengan metode metric scoring yang meliputi accuracy score, precision score, dan recall score. Berikut merupakan hasil evaluasi model dengan metode metric scoring:

Tabel 3. Hasil Metric Scoring

Accuracy Score	Precision Score	Recall
77.442807	74.504627	71.03638

3.3. Activity Diagram



Gambar 4. Activity Diagram

Dari gambar activity diagram, dapat dijelaskan bagaimana aplikasi bekerja. User yang ingin mendapat diagnosis penyakit dapat menggunakan aplikasi. Dari API aplikasi, sistem dapat mendapat data dari user dan mengolah data sehingga menghasilkan prediksi berupa diagnosis penyakit. Setelah itu, sistem dapat mengirim diagnosis ke aplikasi.

3.4. Rancangan Tabel pada Database

Database yang dibuat hanya memerlukan satu tabel saja untuk menampung data dari user beserta prediksi dari sistem. Data atau atribut yang disimpan di dalam tabel meliputi:

Tabel 4. Deskripsi Atribut Pada Tabel

Nama Atribut	Tipe Data	Status
id_keluhan	Varchar (8)	Unique
keluhan	Text (500)	-
prediksi	Varchar (20)	-

4. Kesimpulan

Berdasarkan penelitian tersebut, model memiliki akurasi prediksi sebesar 77,44%, skor precision sebesar 74,50%, dan skor recall sebesar 71.03%. Dari skor akurasi tersebut, model machine learning dengan algoritma Linear SVM dapat digunakan untuk mendiagnosis penyakit dari pasien. Model tersebut juga dapat diterapkan ke dalam aplikasi berbasis web. Adanya penelitian ini diharapkan dapat membantu pengelola apotek dalam memberikan pelayanan 24 jam tanpa adanya tenaga manusia, sehingga pasien dapat dilayani kapanpun dan dimanapun.

Daftar Pustaka

[1] Muhammadin, A., & Sobari, I. A., “Analisis Sentimen Pada Ulasan Aplikasi Kredivo Dengan Algoritma Svm Dan Nbc”, Reputasi: Jurnal Rekayasa Perangkat Lunak, vol. 2, no. 2, pp. 85-91, 2021.

[2] D. Vonega, A. Fadila, and D. Kurniawan, “Analisis Sentimen Twitter Terhadap Opini Publik Atas Isu Pencalonan Puan Maharani dalam PILPRES 2024”, JAIC, vol. 6, no. 2, pp. 129-135, 2022.

[3] AINIZAR, Muhammad Alif; NISA, Khoirun. “Aplikasi Informasi Pendaftaran Member dan Penjualan Merchandise pada Komunitas Manchester City Supporters Club Indonesia Chapter Purwokerto”. Jurnal Nasional Teknologi Informasi dan Aplikasnya, vol. 1, no. 2, pp. 771-780, 2023.

[4] Fitria Nur Hasanah, M.Pd, Rahmania Sri Untari, M.Pd., Rekayasa Perangkat Lunak, UMSIDA Press, pp. 23, 2020.

[5] Muhammadin, A., & Sobari, I. A., “Analisis Sentimen Pada Ulasan Aplikasi Kredivo Dengan Algoritma Svm Dan Nbc”, Reputasi: Jurnal Rekayasa Perangkat Lunak, vol. 2, no. 2, pp. 85-91, 2021.

[6] Wijaya Kusuma Sandi, Ida Bagus Gede Dwidasmara, “Implementasi Algoritma K-Means Clustering dalam Penentuan Klasifikasi Tingkat Pembangunan Perekonomian di Provinsi Bali”, Jurnal Nasional Teknologi Informasi dan Aplikasnya, vol. 1, no. 2, pp. 761-770, 2023.

Halaman ini sengaja dibiarkan kosong