BINER
Jurnal Ilmiah Informatika dan Komputer
Vol. 5 | No. 2 | Juli 2026

p-ISSN : 2828-0210 | e-ISSN : 2828-0229
https://ojs.unsiq.ac.id/index.php/biner

PERBANDINGAN BI-LSTM DAN SVM UNTUK KLASIFIKASI MULTI-KELAS PERTANYAAN PENGGUNA PADA PLATFORM KONSULTASI KESEHATAN DARING BERBAHASA INDONESIA

Shafana Dhiya Ulhaq Huwaida[1)], Safitri Juanita[2)*]
[1,2)] Sistem Informasi, Fakultas Teknologi Informasi, Universitas Budi Luhur, DKI Jakarta, Indonesia
Email: 2212500959@student.budiluhur.ac.id[1)], safitri.juanita@budiluhur.ac.id[2*)]

Dikirim: 4 Juni 2026; Disetujui: 20 Juni 2026; Dipublikasikan: 31 Juli 2026

ABSTRAK

Pertumbuhan platform konsultasi kesehatan daring di Indonesia menghasilkan volume data teks yang semakin besar sehingga diperlukan sistem klasifikasi otomatis untuk mengategorikan pertanyaan pengguna secara efektif. Penelitian ini bertujuan membandingkan performa Bidirectional Long Short-Term Memory (Bi-LSTM) dan Support Vector Machine (SVM) pada klasifikasi multi-kelas pertanyaan pengguna berbahasa Indonesia. Dataset yang digunakan diekstrak dari 497.974 rekaman konsultasi platform Alodokter periode 2014–2021, dengan 10.659 data valid yang dikelompokkan ke dalam tujuh topik penyakit dengan frekuensi pertanyaan tertinggi. Penelitian ini menerapkan prapemrosesan teks serta tiga strategi penanganan *class imbalance*, yaitu *class weight, oversampling, dan undersampling*. Hasil eksperimen menunjukkan bahwa skenario *class weight* menghasilkan performa terbaik pada kedua model. Bi-LSTM memperoleh *weighted* F1-score sebesar 0,892 dan AUC sebesar 0,985, sedikit lebih tinggi dibandingkan SVM yang memperoleh *weighted* F1-*score* sebesar 0,891 dan AUC sebesar 0,984, sementara kedua model mencapai akurasi yang sama sebesar 0,890. Penelitian ini memberikan bukti empiris perbandingan langsung Bi-LSTM dan SVM pada pertanyaan pengguna platform konsultasi kesehatan daring berbahasa Indonesia, sekaligus mengevaluasi tiga strategi penanganan *class imbalance* dalam satu kerangka eksperimen yang masih terbatas pada penelitian sebelumnya.
Kata Kunci: Bi-LSTM, ketidakseimbangan kelas, klasifikasi teks, konsultasi kesehatan, SVM.

ABSTRACT

The growth of online health consultation platforms in Indonesia has generated an ever-increasing volume of text data, necessitating an automatic classification system to effectively categorize user questions. This study aims to compare the performance of Bidirectional Long Short-Term Memory (Bi-LSTM) and Support Vector Machine (SVM) in the multi-class classification of Indonesian-language user questions. The dataset used was extracted from 497,974 consultation records on the Alodokter platform from 2014 to 2021, with 10,659 valid data points grouped into the seven disease topics with the highest question frequencies. This study applied text preprocessing as well as three strategies for addressing class imbalance: class weighting, oversampling, and undersampling. The experimental results show that the class weight scenario yields the best performance for both models. Bi-LSTM achieved a weighted F1-score of 0.892 and an AUC of 0.985, slightly higher than SVM, which achieved a weighted F1-score of 0.891 and an AUC of 0.984, while both models achieved the same accuracy of 0.890. This study provides empirical evidence of a direct comparison between Bi-LSTM and SVM on user queries from an Indonesian-language online health consultation platform, while also evaluating three strategies for addressing class imbalance within a single experimental framework—an approach that was previously limited in prior research.
***Keywords**: Bi-LSTM, class imbalance, health consultation, SVM, text classification.*

1. PENDAHULUAN

Transformasi digital di bidang kesehatan menjadi semakin penting, khususnya sejak pandemi COVID-19 mendorong percepatan pemanfaatan layanan kesehatan berbasis digital [1]. Salah satu bentuk transformasi tersebut adalah telemedis, yaitu layanan kesehatan jarak jauh yang memanfaatkan teknologi informasi dan komunikasi. Kehadiran telemedis memberikan akses layanan kesehatan yang lebih luas bagi masyarakat, meningkatkan efisiensi waktu pelayanan, serta mengurangi hambatan geografis dan biaya transportasi menuju fasilitas kesehatan [2].

Sebelum pandemi Covid-19, pengguna telemedis dari berbagai aplikasi yang ada di Indonesia mendekati 4 juta. Saat ini, pengguna berbagai aplikasi telemedis sudah melebihi angka 15 juta [3]. Berdasarkan hasil survey terhadap 2.108 responden berusia lebih dari 16 tahun di seluruh Indonesia, Alodokter menjadi salah satu platform telemedis terbesar yang paling banyak digunakan masyarakat Indonesia dengan persentase pengguna sebesar 35,7% [4].

Tingginya penggunaan platform ini menghasilkan ratusan ribu rekaman teks konsultasi kesehatan yang berpotensi dimanfaatkan untuk pengembangan sistem kecerdasan buatan di bidang kesehatan sebagai dasar pengambilan keputusan yang lebih baik [5]. Peningkatan volume data tekstual konsultasi kesehatan ini menimbulkan kebutuhan akan sistem klasifikasi otomatis yang mampu mengelompokkan pertanyaan pengguna platform secara cepat dan akurat.

Pertanyaan yang diajukan pengguna platform konsultasi kesehatan daring umumnya bersifat informal, tidak terstruktur, dan bervariasi dalam penyampaian keluhan, sehingga pengelompokan secara manual menjadi tidak efektif dalam skala besar. Penelitian berbasis *chatbot* menunjukkan bahwa pertanyaan pengguna dalam bahasa alami dapat diolah secara otomatis untuk mendukung pengelompokan kebutuhan pengguna [6], yang memperkuat urgensi pengembangan model klasifikasi multi-kelas otomatis untuk mengategorikan pertanyaan pengguna ke dalam berbagai topik penyakit pada platform konsultasi kesehatan daring.

Berbagai pendekatan *machine learning* dan *deep learning* telah dieksplorasi untuk menjawab tantangan ini, sebagaimana ditunjukkan oleh sejumlah penelitian berikut.

Penelitian terdahulu telah mengeksplorasi penerapan SVM pada klasifikasi teks medis. Penelitian klasifikasi teks berbasis SVM untuk deteksi kanker menunjukkan performa tinggi dengan akurasi 93,83% dan F1-*Score* 94,27%, membuktikan efektivitas SVM sebagai sistem pendukung keputusan awal pada teks medis [7]. Selain itu, penelitian klasifikasi pertanyaan kesehatan konsumen berbahasa Indonesia menunjukkan bahwa SVM dengan representasi TF-IDF mampu mengklasifikasikan pertanyaan kesehatan ke dalam kategori semantik yang relevan dan menjadi *baseline* yang kuat dibandingkan model *deep learning* [8].

Penelitian klasifikasi teks sentimen layanan kesehatan menggunakan *Multiclass* SVM dengan ekstraksi fitur TF-IDF juga menghasilkan akurasi sebesar 97,09% dan F1-score 97,40%, membuktikan efektivitas SVM pada klasifikasi teks multi-kelas di domain kesehatan [9]. SVM juga telah diaplikasikan pada analisis sentimen terkait penyakit hepatitis, yang semakin memperkuat kemampuannya dalam menangani klasifikasi teks pada domain kesehatan [10]. Studi perbandingan metode *machine learning* untuk klasifikasi penyakit turut menegaskan bahwa SVM tetap efektif ketika dikombinasikan dengan ekstraksi fitur yang tepat seperti TF-IDF [11].

Pada pendekatan *deep learning*, penelitian perbandingan LSTM, Bi-LSTM, dan CNN untuk klasifikasi multi-label respons dokter dalam konsultasi kesehatan daring menunjukkan bahwa Bi-LSTM menghasilkan akurasi tertinggi sebesar 0,890, meskipun masih terbatas oleh ukuran dataset yang kecil dan variasi bahasa informal serta singkatan medis [12]. Penelitian lain pada klasifikasi multi-kelas pertanyaan kesehatan berbahasa Indonesia juga menunjukkan bahwa model *deep learning* mampu mengungguli SVM, dengan Bi-LSTM termasuk model RNN berperforma terbaik [13].

Kemampuan Bi-LSTM dalam memahami konteks medis berbahasa Indonesia juga berhasil mengenali entitas medis pada teks

kesehatan tidak terstruktur menggunakan arsitektur Bi-LSTM [14]. Lebih lanjut, penelitian deteksi emosi teks berbahasa Indonesia menunjukkan bahwa Bi-LSTM menghasilkan F1-score 86,3%, mengungguli SVM sebesar 73,8%, yang membuktikan bahwa Bi-LSTM mampu menangkap konteks linguistik lebih baik pada teks berbahasa Indonesia [15]. LSTM juga terbukti mampu mempelajari pola data berurutan dan mempertahankan informasi penting dalam urutan input [16], sehingga menjadi dasar yang kuat bagi pengembangan model Bi-LSTM untuk klasifikasi teks konsultasi kesehatan.

Meskipun berbagai penelitian sebelumnya telah menunjukkan potensi SVM maupun Bi-LSTM dalam klasifikasi teks kesehatan, terdapat beberapa celah penelitian yang masih perlu dikaji lebih lanjut. Pertama, sebagian besar penelitian masih berfokus pada penerapan salah satu model secara terpisah, sehingga belum tersedia evaluasi komparatif yang sistematis antara model pembelajaran mesin konvensional dan model *deep learning* pada domain pertanyaan pengguna platform konsultasi kesehatan daring berbahasa Indonesia. Kedua, penelitian yang ada umumnya menggunakan dataset berukuran relatif kecil [12] atau belum memanfaatkan data konsultasi dalam skala besar yang lebih merepresentasikan keragaman topik dan karakteristik bahasa pengguna. Ketiga, meskipun beberapa studi telah menerapkan strategi penanganan *class imbalance* secara terpisah, efektivitas berbagai strategi seperti *class weight*, *oversampling*, dan *undersampling* pada klasifikasi teks kesehatan berbahasa Indonesia belum pernah dievaluasi dalam satu kerangka eksperimen yang seragam.

Keempat, penelitian klasifikasi pertanyaan kesehatan berbahasa Indonesia yang ada umumnya berfokus pada klasifikasi biner atau jumlah kelas yang terbatas, dan distribusi data yang tidak seimbang masih relatif terbatas. Selain itu, seiring berkembangnya model NLP modern berbasis transformer, diperlukan pemahaman mengenai sejauh mana model yang lebih ringan seperti SVM dan Bi-LSTM masih mampu memberikan kinerja yang kompetitif pada lingkungan dengan sumber daya komputasi yang lebih terbatas. Berdasarkan celah tersebut, penelitian ini bertujuan membandingkan performa Bi-LSTM dan SVM dalam klasifikasi multi-kelas pertanyaan pengguna pada platform konsultasi kesehatan daring berbahasa Indonesia menggunakan 497.974 rekaman konsultasi Alodokter [17]. Pertanyaan pengguna dipilih sebagai fokus utama karena merepresentasikan kebutuhan dan keluhan awal yang menjadi dasar proses konsultasi sebelum dokter memberikan jawaban.

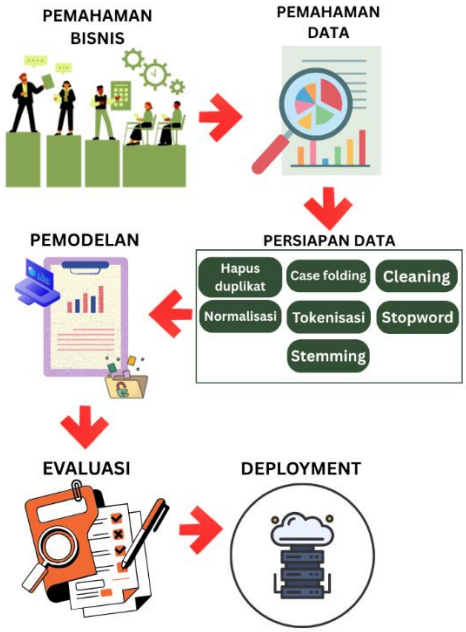
Oleh karena itu, diperlukan model klasifikasi otomatis yang mampu mengelompokkan pertanyaan tersebut ke dalam kategori topik kesehatan secara cepat dan akurat, guna mendukung pengelompokan konsultasi otomatis dan mempercepat distribusi pertanyaan kepada tenaga medis yang sesuai. Klasifikasi dilakukan ke dalam tujuh kategori topik berdasarkan frekuensi pertanyaan tertinggi, yaitu Diabetes, Diare, Gangguan Ginjal, Gangguan Saluran Pernapasan, HIV-AIDS, Stroke, dan Tuberkulosis.

Penelitian ini memberikan tiga kontribusi utama. Pertama, penelitian ini menyajikan evaluasi komparatif antara SVM sebagai representasi model pembelajaran mesin konvensional dan Bi-LSTM sebagai representasi model *deep learning* pada domain pertanyaan pengguna platform konsultasi kesehatan daring berbahasa Indonesia menggunakan dataset Alodokter berskala besar. Kedua, mengevaluasi efektivitas tiga strategi penanganan ketidakseimbangan kelas (*class imbalance)* yaitu *class weight*, *oversampling* (SMOTE pada SVM dan *oversampling* pada Bi-LSTM), dan *undersampling* dalam satu kerangka eksperimen yang konsisten sehingga memungkinkan analisis yang lebih komprehensif mengenai pengaruh masing-masing strategi terhadap performa model.

Ketiga, menyediakan bukti empiris mengenai efektivitas model yang relatif ringan secara komputasi seperti Bi-LSTM dan SVM sebagai alternatif yang layak untuk implementasi pengembangan sistem klasifikasi otomatis pada layanan konsultasi kesehatan daring dengan keterbatasan infrastruktur.

2. METODE

Pada penelitian ini menggunakan metodologi penambangan data CRISP-DM (*Cross Industry Standard Process for Data Mining*) untuk memandu proses analisis data [18]. Langkah-langkah penelitian terdapat pada Gambar 1.



Gambar 1. Tahapan Penelitian

2.1. Pemahaman Bisnis

Tahap pemahaman bisnis dilakukan untuk mengidentifikasi permasalahan pada data konsultasi kesehatan di platform daring. Meningkatnya penggunaan layanan kesehatan digital menyebabkan volume data konsultasi semakin besar dan sulit dianalisis secara manual. Oleh karena itu, diperlukan sistem klasifikasi otomatis untuk mengelompokkan pertanyaan pengguna platform ke dalam topik kesehatan tertentu, terutama karena pertanyaan umumnya menggunakan bahasa informal dan tidak terstruktur.

Penelitian ini bertujuan mengembangkan model klasifikasi multi-kelas pada pertanyaan pengguna pada platform konsultasi kesehatan daring. Dari dataset, dipilih 10 topik dengan frekuensi tertinggi sebagai kandidat label, lalu dilakukan seleksi dan penggabungan berdasarkan kemiripan semantik dan klinis hingga diperoleh 7 kategori final. Kolom pertanyaan pengguna platform konsultasi kesehatan daring dipilih sebagai fokus utama karena merepresentasikan kebutuhan dan keluhan awal sebelum dokter memberikan jawaban. Hasil klasifikasi diharapkan dapat membantu pengelompokan konsultasi dan mempercepat distribusi pertanyaan ke tenaga medis yang sesuai.

2.2. Pemahaman Data

Dataset penelitian terdiri dari 497.974 rekaman konsultasi kesehatan berbahasa Indonesia dari *platform* Alodokter periode Desember 2014 hingga Februari 2021 [17]. Pada tahap pemahaman data, dilakukan penyaringan data pertanyaan pengguna platform konsultasi kesehatan daring berdasarkan nilai *risk* kategori *high* sehingga diperoleh 18.304 data. Selanjutnya diidentifikasi sepuluh *topic_set* dengan frekuensi tertinggi sebagai kandidat label klasifikasi.

Beberapa topik kemudian digabungkan berdasarkan kemiripan semantik dan klinis serta tingginya kesalahan prediksi antar topik yang berdekatan. Topik HIV dan AIDS digabung menjadi HIV-AIDS, batu ginjal dan infeksi ginjal menjadi gangguan ginjal, serta pneumonia dan bronkitis menjadi gangguan saluran pernapasan. Setelah tahap pembersihan data, diperoleh tujuh kategori topik final dengan total 10.659 data valid yang ditampilkan pada Tabel 1.

Tabel 1. Jumlah Dataset yang digunakan

No	Topik	Jumlah Data
1.	Tuberkulosis	3895
2.	HIV-AIDS	1918
3.	Gangguan Saluran Pernapasan	1213
4.	Gangguan Ginjal	1092
5.	Diare	1072
6.	Diabetes	804
7.	Stroke	665
	Total	10,659

2.3. Persiapan Data

Tahap persiapan data bertujuan mengubah data teks mentah menjadi data yang bersih dan siap digunakan pada proses pemodelan menggunakan *machine learning* dan *deep learning*. Tahap ini menggunakan beberapa langkah sebagai berikut:

1) Penghapusan Data Duplikat

Pengecekan duplikasi dilakukan berdasarkan kolom pertanyaan pengguna platform konsultasi kesehatan daring. Dari 14.634 data hasil filter, ditemukan 3.938 data

duplikat yang kemudian dihapus untuk mencegah data *leakage* antara data latih dan data uji sehingga evaluasi model tetap valid.

2) Penghapusan Teks Terlalu Pendek

Teks dengan panjang kurang dari 5 kata dihapus karena dinilai tidak memiliki informasi klinis yang memadai untuk klasifikasi. Sebanyak 14 teks dihapus, termasuk teks singkat seperti "Dok saya", "Dok, kaka", dan alamat email yang tidak relevan.

3) *Case Folding* digunakan untuk mengubah seluruh huruf menjadi *lowercase* agar penulisan kata lebih konsisten.
4) *Cleaning* dilakukan dengan menghapus URL, email, angka, dan karakter khusus serta menggantinya dengan spasi untuk mencegah penggabungan token yang tidak diinginkan.
5) Normalisasi mengonversi kata tidak baku, singkatan, dan slang ke bentuk standar menggunakan kamus manual, termasuk singkatan medis seperti "tb" menjadi "tuberkulosis" dan "dm" menjadi "diabetes melitus".
6) Tokenisasi, teks dipecah menjadi token menggunakan *word tokenizer* dari NLTK.
7) Pada tahap *Stopword removal*, kata yang tidak memiliki kontribusi semantik dihapus menggunakan *stopword* bahasa Indonesia dari Sastrawi yang diperluas dengan 841 *custom Stopword*, termasuk sapaan konsultasi daring, partikel informal, dan nama platform. Filter panjang kata juga diterapkan dengan menghapus token kurang dari 3 karakter atau lebih dari 20 karakter.
8) Tahap terakhir adalah *Stemming* menggunakan Sastrawi Stemmer dengan penerapan *whitelist* istilah medis penting seperti "diabetes", "stroke", dan "tuberkulosis" agar tidak mengalami distorsi makna. Pada Tabel 2 menampilkan contoh semua tahapan ketiga hingga kedelapan.

Tabel 2. Tahap Pembersihan Data

Tahap	Contoh Teks
Teks Asli	Hallo Doc, Saya mau tanya jika perut bawah saya bagian kiri sllu skit doc itu disebabkan oleh apa yah Doc?? #makasih
Case Folding	hallo doc, saya mau tanya jika perut bawah saya bagian kiri sllu skit doc itu disebabkan oleh apa yah doc?? #makasih
Cleaning	hallo doc saya mau tanya jika perut bawah saya bagian kiri sllu skit doc itu disebabkan oleh apa yah doc makasih
Normalisasi	halo dok saya mau tanya jika perut bawah saya bagian kiri selalu sakit dok itu disebabkan oleh apa ya dok makasih
Tokenisasi	['halo', 'dok', 'saya', 'mau', 'tanya', 'jika', 'perut', 'bawah', 'saya', 'bagian', 'kiri', 'selalu', 'sakit', 'dok', 'itu', 'disebabkan', 'oleh', 'apa', 'ya', 'dok', 'makasih']
Stopword Removal	['perut', 'bawah', 'kiri', 'selalu', 'sakit', 'disebabkan', 'apa']
Stemming	['perut', 'bawah', 'kiri', 'selalu', 'sakit', 'sebab', 'apa']

2.4. Pemodelan

Tahap pemodelan dilakukan dengan membandingkan dua algoritma SVM dan Bi-LSTM secara paralel menggunakan *split* data latih dan uji sebesar 80:20 dengan *stratified sampling* untuk menjaga proporsi kelas. Pemodelan dilakukan melalui tiga skenario penanganan *class imbalance* yaitu *class weight*, *oversampling* (SMOTE pada SVM, dan *oversampling* pada Bi-LSTM), dan *undersampling*. Berikut setting parameter pada kedua algoritma:

1) *Support Vector Machine* (SVM) merupakan metode *supervised learning* yang membentuk *hyperplane* optimal untuk memisahkan kelas dengan margin maksimum [19]. SVM diimplementasikan menggunakan LinearSVC dari *library* Scikit-learn dengan parameter random_state=42 dan max_iter=1000, menggunakan representasi teks TF-IDF dengan konfigurasi max_features=5.000, ngram_range = (1,2), min_df = 2, dan max_df=0,8.

(a) Rumus *Hyperplane*

$$f(x) = wTx + b \quad (1)$$

Keterangan:
w = vektor bobot
x = vektor fitur input (TF-IDF)

b = bias

(b) Rumus Margin Maksimum

$$margin = \frac{2}{||\mathbf{w}||} \quad (2)$$

(c) Fungsi Objektif (Minimisasi)

$$\min_{w,b} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i \quad (3)$$

Keterangan:
C = parameter regularisasi
ξi = slack variable untuk data yang tidak terpisah sempurna

(d) Representasi Teks (TF-IDF)

$$\text{TF-IDF}(t,d) = TF(t,d) \times \log\left(\frac{N}{DF(t)}\right) \quad (4)$$

Keterangan:
t = term (kata)
d = dokumen
N = total dokumen
df(t)=jumlah dokumen yang mengandung term t
TF(t,d) = frekuensi term t pada dokumen d

Tabel 3. Arsitektur Model Bi-LSTM

Komponen	Parameter
Vocabulary Size	10.000
Embedding Dimension	128
Maximum Sequence Length	150
SpatialDropout1D	0,3
Bi-LSTM Layer 1	128-unit, dropout 0,3
Bi-LSTM Layer 2	64-unit, dropout 0,3
Dense Layer	64 unit, ReLU
Dropout	0,5
Output Layer	Softmax
Optimizer	Adam
Learning Rate	0,0005
Loss Function	Sparse Categorical Crossentropy
Batch Size	32
Maximum Epoch	15
Early Stopping	patience = 5
Restore Best Weights	True

2) *Bidirectional Long Short-Term Memory* (Bi-LSTM) merupakan pengembangan LSTM yang memproses data dalam dua arah (*forward* dan *backward*), untuk memahami konteks secara menyeluruh [20]. Arsitektur dan parameter model Bi-LSTM yang digunakan pada penelitian ini disajikan pada Tabel 3.

$$h_t^{(\rightarrow)} = LSTM\left(x_t, h_{(t-1)}^{(\rightarrow)}\right)$$
$$h_t^{(\leftarrow)} = LSTM\left(x_t, h_{(t+1)}^{(\leftarrow)}\right)$$
$$h_t = \left[h_t^{(\rightarrow)}; h_t^{(\leftarrow)}\right]$$

(a) Fungsi aktivasi *output* (Softmax):

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad (5)$$

Keterangan:
K = jumlah kelas

(b) Fungsi *loss* (*Sparse Categorical Crossentropy*)

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad (6)$$

Keterangan:
y_i = label sebenarnya
$\hat{y}_i$ = probabilitas prediksi model

Pada penelitian ini, dilakukan beberapa variasi skenario sebagai berikut

1) Skenario 1 (*Class Weight*)

Penanganan *class imbalance* dilakukan dengan memberikan bobot lebih besar pada kelas minoritas selama pelatihan. Pada SVM digunakan parameter *class_weight='balanced'*, sedangkan pada Bi-LSTM digunakan parameter *class_weight* berdasarkan distribusi kelas data latih.

2) Skenario 2 (*Oversampling*)

Pendekatan *oversampling* hanya diterapkan pada data latih untuk menghindari kebocoran data (data *leakage*) ke data uji. Masing-masing model menggunakan teknik *oversampling* yang berbeda sesuai dengan karakteristik representasi datanya.

- **SVM menggunakan SMOTE (*Synthetic Minority Oversampling Technique*)**

SMOTE adalah teknik di mana *oversampling* kelas minoritas dilakukan dengan menghasilkan sampel sintetik baru [21]. Berbeda dengan random *oversampling* yang hanya menduplikasi sampel yang ada, SMOTE menghasilkan data baru yang lebih beragam sehingga mengurangi risiko *overfitting*. Pada penelitian ini, SMOTE diterapkan pada fitur TF-IDF data latih menggunakan *library imbalanced-learn* dengan parameter yang ditampilkan pada Tabel 4.

Tabel 4. Parameter Skenario SMOTE pada SVM

Parameter	Nilai	Keterangan
k_neighbors	5 (*default*)	Jumlah tetangga terdekat untuk interpolasi sampel sintetis
sampling_strategy	'auto' (*default*)	Menyeimbangkan semua kelas minoritas mengikuti kelas mayoritas
random_state	42	Reprodusibilitas hasil

- **Bi-LSTM menggunakan *Random Oversampling***

Random Oversampling merupakan pemberian data dari kelas minoritas ke dalam data *training* secara acak [22]. Pada penelitian ini, target jumlah sampel per kelas ditetapkan berdasarkan rata-rata jumlah sampel seluruh kelas pada data latih (*mean count*), sehingga kelas yang berada di bawah rata-rata akan ditambah sampelnya hingga mencapai nilai tersebut, sementara kelas yang sudah di atas rata-rata dipertahankan. Proses ini diterapkan pada data teks sebelum tokenisasi menggunakan fungsi resample dari *library* sklearn.utils dengan parameter yang ditampilkan pada Tabel 5.

Tabel 4. Parameter Skenario *Random Oversampling* pada Bi-LSTM

Parameter	Nilai	Keterangan
replace	*True*	Mengizinkan duplikasi sampel
N_samples	*Mean count* data latih	Target jumlah sampel per kelas berdasarkna rata-rata seluruh kelas
random_state	42	Reprodusibilitas hasil

3) Skenario 3 (*Undersampling*)

Penanganan *class imbalance* dilakukan dengan mengurangi data pada kelas mayoritas, yaitu Tuberkulosis, dari 3.116 menjadi 1.100 sampel menggunakan *partial random undersampling*, sementara kelas lainnya dipertahankan sesuai distribusi aslinya. Proses ini diterapkan hanya pada data latih menggunakan fungsi resample dari *library* sklearn.utils dengan parameter *replace=False*, n_samples=1.100, dan random_state=42.

2.5. Evaluasi

Evaluasi model dilakukan menggunakan beberapa metrik klasifikasi multi-kelas pada kondisi *class imbalance*. Seluruh metrik dihitung menggunakan pendekatan *macro average* dan *weighted average* untuk memberikan evaluasi yang lebih representatif terhadap distribusi kelas yang tidak seimbang. Adapun metrik evaluasi yang digunakan meliputi:

1) *Confusion* Matrix, digunakan untuk mengevaluasi performa model berdasarkan nilai *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Matriks ini menjadi dasar perhitungan metrik evaluasi klasifikasi [23].

2) *Accuracy*

Accuracy adalah proporsi klasifikasi yang benar terhadap keseluruhan prediksi [24].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

3) *Precision*, yaitu proporsi prediksi benar dari seluruh prediksi pada suatu kelas.

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

4) *Recall*, yaitu proporsi data suatu kelas yang berhasil diklasifikasikan dengan benar.

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

5) F1-*Score*, yaitu rata-rata harmonis antara precision dan *recall*.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

6) AUC (*Area Under the Curve*), yaitu kemampuan model dalam membedakan antar kelas secara keseluruhan meskipun distribusi data tidak seimbang.

2.6. *Deployment*

Bagian ini menjelaskan potensi penerapan model klasifikasi terbaik sebagai sistem pendukung pengelompokan pertanyaan pengguna platform konsultasi kesehatan daring.

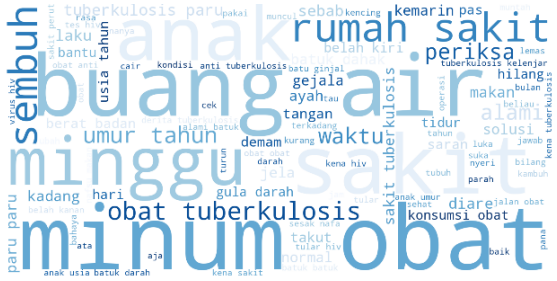
3. HASIL DAN PEMBAHASAN

3.1. Analisis Distribusi Kata Sebelum dan Sesudah Persiapan Data

Analisis distribusi kata dilakukan menggunakan visualisasi *Word Cloud* untuk

memahami karakteristik teks sebelum dan sesudah tahap persiapan data. Berdasarkan Tampilan pada Gambar 2, pada data awal, kata yang dominan masih didominasi oleh sapaan, kata umum, dan istilah kurang informatif. Setelah proses *cleaning*, *stopword removal*, dan *stemming*, kata-kata yang tidak relevan berhasil dikurangi sehingga kata yang berkaitan dengan keluhan dan topik kesehatan menjadi lebih dominan sebagaimana ditampilkan oleh Gambar 3. Hasil ini menunjukkan bahwa tahap persiapan data mampu meningkatkan kualitas representasi teks sebelum proses klasifikasi dilakukan.

Gambar 2. Word cloud Data Kotor



Gambar 3. Word cloud Data Bersih

3.2. Distribusi Data Latih dan Data Uji

Pembagian dataset dilakukan menggunakan metode *Stratified sampling* dengan rasio 80:20 untuk menjaga proporsi setiap kelas pada data latih dan data uji tetap konsisten dengan distribusi dataset asli. Pendekatan ini penting untuk mengurangi bias distribusi kelas selama proses pelatihan dan pengujian model, terutama pada dataset yang bersifat tidak seimbang. Hasil pembagian data pada masing-masing kelas ditampilkan pada Tabel 6.

3.3. Evaluasi Performa Model Klasifikasi

Evaluasi model klasifikasi menggunakan metrik *precision*, *recall*, F1-*score*, dan AUC. Pada klasifikasi multi-kelas, nilai metrik dihitung menggunakan *macro average* dan *weighted average* untuk menggambarkan performa model pada setiap topik serta mempertimbangkan ketidakseimbangan distribusi data. SVM menggunakan representasi TF-IDF berdimensi 8.527 × 5.000 pada data latih dan 2.132 × 5.000 pada data uji, sedangkan Bi-LSTM menggunakan hasil tokenisasi berbentuk sekuens berdimensi 8.527 × 150 pada data latih dan 2.132 × 150 pada data uji.

Penerapan *oversampling* dan *undersampling* mengubah jumlah sampel pada data latih sesuai distribusi kelas yang dihasilkan, sementara jumlah fitur tetap konstan. Adapun dimensi data uji tidak berubah pada seluruh skenario, yaitu 2.132 × 5.000 untuk SVM dan 2.132 × 150 untuk Bi-LSTM.

3.3.1. Hasil Evaluasi SVM

Berdasarkan Tabel 7, metode SVM dengan skenario *class weight* menghasilkan performa terbaik dengan *weighted* F1-*score* sebesar

Tabel 5. Distribusi Data Latih dan Data Uji Seluruh Skenario

No	Topik	Data Latih (80%)				Data Uji 20%
		CW	SMOTE (SVM)	ROS (Bi-LSTM)	U	
1.	Diabetes	643	3.116	1,218	643	161
2.	Diare	858	3.116	1,218	858	214
3.	Gangguan Ginjal	874	3.116	1,218	874	218
4.	Gangguan Saluran Pernapasan	970	3.116	1,218	970	243
5.	HIV-AIDS	1.534	3.116	1,534	1,534	384
6.	Stroke	532	3.116	1,218	532	133
7.	Tuberkulosis	3116	3.116	3,116	1,100	779
	Total data	**8.528**	**21.812**	**10.740**	**6.511**	**2.132**

*CW (*Class Weight*); ROS (*Random Oversampling*); U(*Undersampling*)

0.891 dan *weighted* AUC sebesar 0.984. Topik HIV-AIDS memperoleh performa tertinggi (F1-*score* 0.973), sedangkan Gangguan Saluran Pernapasan menunjukkan performa terendah (F1-*score* 0.685). Perbedaan ini menunjukkan bahwa karakteristik pertanyaan pada beberapa topik lebih mudah dikenali oleh model, sementara pertanyaan pada topik Gangguan Saluran Pernapasan masih menjadi tantangan dalam proses klasifikasi.

Tabel 6. Performa SVM pada Skenario *Class Weight*

Topik	P*	R*	F1*	AUC
Diabetes	0.942	0.907	0.924	0.995
Diare	0.917	0.935	0.926	0.995
Gangguan Ginjal	0.871	0.931	0.900	0.989
Gangguan Saluran Pernapasan	0.660	0.712	0.685	0.953
HIV-AIDS	0.971	0.974	**0.973**	0.997
Stroke	0.890	0.910	0.900	0.997
Tuberkulosis	0.915	0.873	0.894	0.979
MA**	0.881	0.892	0.886	0.986
WA***	0.892	0.890	**0.891**	**0.984**

*P(*Precision*), R(*Recall*), F1(F1-*Score*); **MA (*Macro Average*); ***WA (*Weighted Average*); angka ditebalkan nilai tertinggi.

Berdasarkan Tabel 8, metode SVM dengan skenario SMOTE menghasilkan *weighted* F1-*score* sebesar 0.882 dan *weighted* AUC sebesar 0.981. Topik HIV-AIDS memperoleh performa terbaik (F1-*score* 0.966), sedangkan Gangguan Saluran Pernapasan memiliki performa terendah (F1-*score* 0.669). Dibandingkan skenario *class weight*, hasil ini menunjukkan bahwa penerapan SMOTE belum memberikan peningkatan performa yang signifikan pada model SVM.

Tabel 7. Performa SVM pada Skenario SMOTE

Topik	P*	R*	F1*	AUC
Diabetes	0.942	0.901	0.921	0.994
Diare	0.920	0.911	0.916	0.994
Gangguan Ginjal	0.866	0.922	0.893	0.989
Gangguan Saluran Pernapasan	0.645	0.696	0.669	0.951
HIV-AIDS	0.969	0.964	**0.966**	0.996
Stroke	0.878	0.917	0.897	0.997
Tuberkulosis	0.902	0.869	0.885	0.973
MA**	0.874	0.883	0.878	0.985
WA***	0.884	0.881	**0.882**	**0.981**

*P(*Precision*), R(*Recall*), F1(*F1-Score*); **MA (*Macro Average*); ***WA (*Weighted Average*); angka ditebalkan nilai tertinggi.

Tabel 8. Performa SVM pada Skenario *Undersampling*

Topik	P*	R*	F1*	AUC
Diabetes	0.918	0.901	0.909	0.994
Diare	0.917	0.935	0.926	0.995
Gangguan Ginjal	0.849	0.931	0.888	0.989
Gangguan Saluran Pernapasan	0.600	0.803	0.687	0.957
HIV-AIDS	0.967	0.982	**0.974**	0.997
Stroke	0.873	0.932	0.902	0.997
Tuberkulosis	0.950	0.805	0.871	0.974
MA**	0.868	0.898	0.880	0.986
WA***	0.892	0.878	**0.881**	**0.983**

*P(*Precision*), R*(Recall)*, F1(*F1-Score*), **MA (*Macro Average*); ***WA (*Weighted Average*); angka ditebalkan nilai tertinggi.

Berdasarkan Tabel 9, metode SVM dengan skenario *undersampling* menghasilkan *weighted* F1-*score* sebesar 0.881 dan *weighted* AUC sebesar 0.983. Topik HIV-AIDS tetap menjadi topik dengan performa terbaik (F1-*score* 0.974), sedangkan Gangguan Saluran Pernapasan memperoleh performa terendah (F1-*score* 0.687). Dibandingkan skenario *class weight* dan SMOTE, pendekatan *undersampling* menghasilkan nilai *recall* yang lebih tinggi pada beberapa topik, namun peningkatan tersebut diikuti penurunan *precision*.

Kondisi ini menunjukkan bahwa pengurangan sampel pada kelas mayoritas membuat model SVM lebih banyak mengenali data positif, tetapi pada saat yang sama meningkatkan kemungkinan kesalahan klasifikasi. Oleh karena itu, *undersampling* belum mampu memberikan peningkatan performa yang lebih baik dibandingkan pendekatan *class weight*.

Berdasarkan Tabel 10, Perbedaan performa yang relatif kecil antar skenario menunjukkan bahwa model SVM memiliki kemampuan yang cukup *robust* dalam menghadapi ketidakseimbangan data. Namun, keunggulan konsisten pada skenario *class weight* menunjukkan bahwa penanganan *class*

imbalance melalui penyesuaian bobot lebih efektif dibandingkan teknik resampling pada dataset konsultasi kesehatan. Berbeda dengan SMOTE dan *undersampling* yang memodifikasi distribusi data, *class weight* tetap mempertahankan seluruh data asli sehingga model dapat mempelajari karakteristik setiap topik secara lebih optimal.

Tabel 9. Ringkasan Performa SVM pada Seluruh Skenario

Skenario	A*	WA-F1*	WA-AUC*
Class Weight	**0.890**	**0.891**	**0.984**
SMOTE	0.881	0.882	0.981
Undersampling	0.878	0.881	0.983

*A (*Accuracy*), WA-F1 (*Weighted Average*-F1), WA-AUC (*Weighted Average*-AUC); angka ditebalkan nilai tertinggi pada tiap metrik.

Temuan ini mengindikasikan bahwa kualitas representasi data lebih berpengaruh terhadap performa SVM dibandingkan penambahan atau pengurangan jumlah sampel pada penelitian ini.

Tabel 11 menunjukkan bahwa skenario *class weight* menghasilkan performa terbaik pada sebagian besar topik, yaitu Diabetes, Diare, Gangguan Ginjal, dan Tuberkulosis. Hasil ini mengindikasikan bahwa penyesuaian bobot kelas lebih efektif dalam membantu model SVM dalam menangani ketidakseimbangan data tanpa mengubah distribusi data asli. Sebaliknya, skenario SMOTE tidak menghasilkan F1-*Score* tertinggi pada satu pun topik.

Tabel 10. Perbandingan Model SVM menggunakan F1-*Score* pada Setiap Topik

Topik	CW*	S*	U*
Diabetes	**0.924**	0.921	0.909
Diare	**0.926**	0.916	**0.926**
Gangguan Ginjal	**0.900**	0.893	0.888
Gangguan Saluran Pernapasan	0.685	0.669	**0.687**
HIV-AIDS	0.973	0.966	**0.974**
Stroke	0.900	0.897	**0.902**
Tuberkulosis	**0.894**	0.885	0.871

*CW (*Class Weight*), S (*SMOTE*), U (*Undersampling*); angka ditebalkan nilai tertinggi pada tiap metrik

Temuan ini menunjukkan bahwa penambahan sampel sintetis belum mampu merepresentasikan pola linguistik teks kesehatan secara optimal, terutama karena karakteristik fitur TF-IDF membuat hubungan antar kata pada teks menjadi kurang natural saat dibuat sampel sintetis.

Sementara itu, skenario *undersampling* menghasilkan performa terbaik pada Diare, HIV-AIDS, Stroke, dan Gangguan Saluran Pernapasan. Meskipun demikian, peningkatan performa pada beberapa topik tersebut tidak diikuti oleh peningkatan performa secara keseluruhan. Hal ini menunjukkan bahwa pengurangan jumlah data pada kelas mayoritas dapat menguntungkan topik tertentu, tetapi juga berpotensi mengurangi informasi yang diperlukan untuk membedakan topik lainnya.

3.3.2. Hasil Evaluasi Bi-LSTM

Berdasarkan Tabel 12, skenario *class weight* pada model Bi-LSTM menghasilkan *weighted* F1-*score* sebesar 0.892 dan *weighted* AUC sebesar 0.985. Topik HIV-AIDS menunjukkan performa terbaik dengan F1-*score* sebesar 0.974, sedangkan Gangguan Saluran Pernapasan memperoleh performa terendah dengan F1-*score* sebesar 0.725. Hasil ini menunjukkan bahwa model mampu mengenali sebagian besar topik dengan baik meskipun distribusi data tidak seimbang. Secara umum, penerapan *class weight* efektif dalam meningkatkan perhatian model terhadap kelas minoritas tanpa mengubah distribusi data asli, sehingga menghasilkan performa yang relatif konsisten pada seluruh topik.

Tabel 11. Performa Bi-LSTM pada Skenario *Class Weight*

Topik	P*	R*	F1*	AUC
Diabetes	0.952	0.857	0.902	0.992
Diare	0.942	0.916	0.929	0.992
Gangguan Ginjal	0.824	0.922	0.870	0.983
Gangguan Saluran Pernapasan	0.664	0.798	0.725	0.962
HIV-AIDS	0.976	0.971	**0.974**	0.996
Stroke	0.853	0.917	0.884	0.994
Tuberkulosis	0.937	0.864	0.899	0.983
MA**	0.878	0.892	0.883	0.986
WA***	0.898	0.890	**0.892**	**0.985**

*P(*Precision*), R*(Recall)*, F1(*F1-Score*), **MA (*Macro Average*); ***WA (*Weighted Average*); angka ditebalkan nilai tertinggi pada tiap metrik.

Berdasarkan Tabel 13, penerapan *oversampling* pada model Bi-LSTM menghasilkan *weighted* F1-*score* sebesar 0.877 dan *weighted* AUC sebesar 0.980. Topik

HIV-AIDS tetap menjadi topik dengan performa terbaik (F1-*score* 0.957), sedangkan Gangguan Saluran Pernapasan memiliki performa terendah (F1-*score* 0.648). Penurunan nilai *weighted* F1-*score* dibandingkan skenario *class weight* menunjukkan bahwa penambahan sampel sintetis belum memberikan kontribusi yang signifikan terhadap kemampuan model dalam membedakan seluruh topik secara konsisten.

Temuan ini mengindikasikan bahwa peningkatan jumlah sampel tidak selalu berbanding lurus dengan peningkatan performa klasifikasi, terutama ketika model telah mampu mempelajari representasi teks dengan baik melalui mekanisme pembobotan kelas.

Tabel 12. Performa Bi-LSTM pada Skenario *Oversampling*

Topik	P*	R*	F1*	AUC
Diabetes	0.940	0.870	0.903	0.983
Diare	0.915	0.902	0.908	0.989
Gangguan Ginjal	0.885	0.885	0.885	0.984
Gangguan Saluran Pernapasan	0.741	0.576	0.648	0.938
HIV-AIDS	0.961	0.953	**0.957**	0.994
Stroke	0.837	0.925	0.879	0.996
Tuberkulosis	0.861	0.926	0.892	0.978
MA**	0.877	0.862	0.868	0.980
WA***	0.878	0.880	**0.877**	**0.980**

*P(*Precision*), R*(Recall),* F1(F1-*Score*), **MA (*Macro Average*); ***WA (*Weighted Average*); angka ditebalkan nilai tertinggi pada tiap metrik.

Berdasarkan Tabel 14, metode Bi-LSTM dengan skenario *undersampling* menghasilkan *weighted* F1-*score* sebesar 0.746 dan *weighted* AUC sebesar 0.947. Topik HIV-AIDS tetap menjadi topik dengan performa terbaik (F1-*score* 0.929), sedangkan Gangguan Saluran Pernapasan memperoleh performa terendah (F1-*score* 0.389). Dibandingkan skenario *class weight* dan *oversampling*, pendekatan *undersampling* justru menghasilkan penurunan performa yang cukup besar pada hampir seluruh topik, baik dari sisi *precision* maupun *recall*.

Kondisi ini menunjukkan bahwa pengurangan sampel pada kelas mayoritas membuat model Bi-LSTM kehilangan informasi penting dalam proses pelatihan, sehingga kemampuan generalisasi model menurun. Oleh karena itu, *undersampling* belum mampu memberikan peningkatan performa yang lebih baik dibandingkan pendekatan *class weight*.

Tabel 13. Performa Bi-LSTM pada Skenario *Undersampling*

Topik	P*	R*	F1*	AUC
Diabetes	0.692	0.683	0.688	0.944
Diare	0.898	0.902	0.900	0.979
Gangguan Ginjal	0.465	0.803	0.589	0.956
Gangguan Saluran Pernapasan	0.432	0.354	0.389	0.843
HIV-AIDS	0.925	0.932	**0.929**	0.991
Stroke	0.606	0.647	0.623	0.960
Tuberkulosis	0.878	0.737	0.801	0.945
MA**	0.699	0.723	0.703	0.946
WA***	0.764	0.742	**0.746**	**0.947**

*P(*Precision*), R*(Recall),* F1(*F1-Score*), **MA (*Macro Average*); ***WA (*Weighted Average*). angka ditebalkan nilai tertinggi pada tiap metrik.

Berdasarkan Tabel 15, perbedaan performa yang cukup signifikan antar skenario menunjukkan bahwa model Bi-LSTM cukup sensitif terhadap pendekatan penanganan *class imbalance* yang digunakan. Keunggulan konsisten pada skenario *class weight* menunjukkan bahwa penanganan *class imbalance* melalui penyesuaian bobot lebih efektif dibandingkan teknik resampling pada dataset konsultasi kesehatan.

Berbeda dengan *oversampling* dan *undersampling* yang memodifikasi distribusi data, *class weight* tetap mempertahankan seluruh data asli sehingga model dapat mempelajari karakteristik setiap topik secara lebih optimal. Temuan ini mengindikasikan bahwa kualitas representasi data lebih berpengaruh terhadap performa Bi-LSTM dibandingkan penambahan atau pengurangan jumlah sampel pada penelitian ini.

Tabel 14. Ringkasan Performa Bi-LSTM pada Seluruh Skenario

Skenario	A*	WA-F1*	WA-AUC*
Class Weight	**0.890**	**0.892**	**0.985**
Oversampling	0.880	0.877	0.980
Undersampling	0.742	0.746	0.947

*A (*Accuracy*), WA-F1 (*Weighted Average*-F1), WA-AUC (*Weighted Average*-AUC); angka ditebalkan nilai tertinggi pada tiap metrik

Tabel 16 menunjukkan bahwa skenario *class weight* menghasilkan performa terbaik pada sebagian besar topik, yaitu Diare,

Gangguan Saluran Pernapasan, HIV-AIDS, Stroke, dan Tuberkulosis. Hasil ini mengindikasikan bahwa penyesuaian bobot kelas lebih efektif dalam membantu model Bi-LSTM menangani *class imbalance* tanpa mengubah distribusi data asli.

Sebaliknya, skenario *oversampling* menghasilkan F1-*score* tertinggi pada Diabetes (0.903) dan Gangguan Ginjal (0.885), meskipun selisihnya dengan *class weight* relatif kecil. Temuan ini menunjukkan bahwa penambahan sampel secara acak pada beberapa topik tertentu masih mampu membantu model Bi-LSTM mengenali pola representasi teks, namun belum konsisten di seluruh topik.

Sementara itu, skenario *undersampling* tidak menghasilkan F1-*score* tertinggi pada satu pun topik. Hal ini menunjukkan bahwa pengurangan jumlah data pada kelas mayoritas berpotensi menghilangkan informasi penting yang dibutuhkan model Bi-LSTM untuk membedakan karakteristik linguistik antar topik kesehatan, sehingga performa secara keseluruhan menurun secara konsisten.

Tabel 15. Perbandingan F1-*Score* Model Bi-LSTM pada Setiap Topik

Topik	CW*	O*	U*
Diabetes	0.902	**0.903**	0.688
Diare	**0.929**	0.908	0.900
Gangguan Ginjal	0.870	**0.885**	0.589
Gangguan Saluran Pernapasan	**0.725**	0.648	0.389
HIV-AIDS	**0.974**	0.957	0.929
Stroke	**0.884**	0.879	0.626
Tuberkulosis	**0.899**	0.892	0.801

*CW (*Class Weight*), O (*Oversampling*), U (*Undersampling*); angka ditebalkan nilai tertinggi pada tiap metrik

Tabel 17 menyajikan perbandingan performa terbaik yang dicapai oleh model SVM dan Bi-LSTM dari tiga skenario penanganan *class imbalance* yang diuji. Hasil menunjukkan bahwa kedua model menghasilkan akurasi yang sama, yaitu 0.890. Namun, Bi-LSTM memperoleh *Weighted* F1-*Score* dan *Weighted* AUC yang sedikit lebih tinggi dibandingkan SVM, masing-masing sebesar 0.892 dan 0.985.

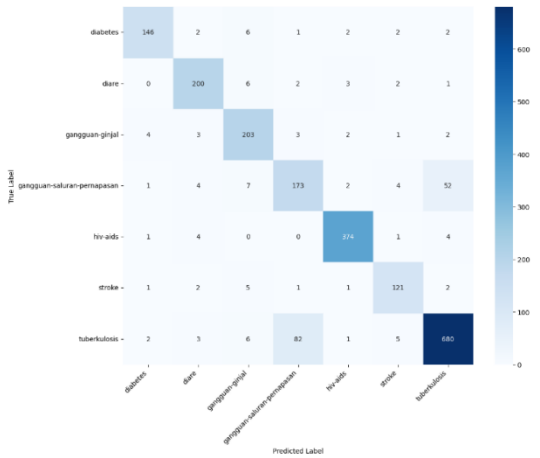
Tabel 16. Perbandingan Performa Terbaik SVM dan Bi-LSTM

Model	Skenario Terbaik	A*	WA-F1*	WA-AUC*
SVM	*Class Weight*	**0.890**	0.891	0.984
Bi-LSTM	*Class Weight*	**0.890**	**0.892**	**0.985**

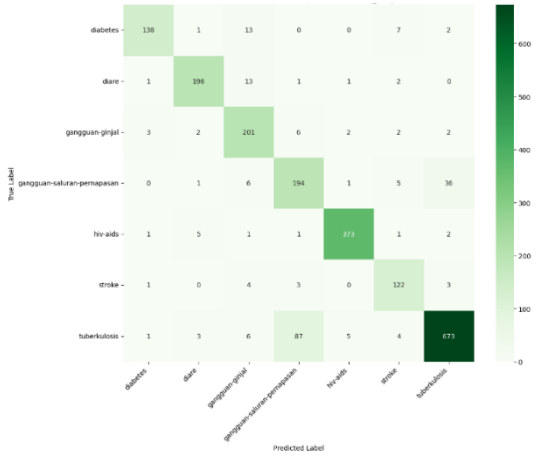
*A(*Accuracy*), *WA-F1(Weighted Average-F1), WA-AUC(Weighted Average-AUC);* angka ditebalkan nilai tertinggi pada tiap metrik.

Selisih Bi-LSTM terhadap SVM hanya sebesar 0.001 pada kedua metrik sehingga kedua model dapat dikatakan memiliki performa yang hampir setara pada dataset ini. Kesetaraan performa kedua model merupakan temuan yang informatif, mengindikasikan bahwa fitur leksikal berbasis TF-IDF pada SVM sudah cukup diskriminatif untuk membedakan topik kesehatan pada dataset ini, sehingga kemampuan pemrosesan kontekstual Bi-LSTM tidak memberikan keunggulan yang jauh lebih besar.

Gambar 4 dan 5 menyajikan visualisasi *confusion matrix* model SVM dan Bi-LSTM pada skenario *class weight*. Secara umum, kedua model menunjukkan prediksi yang baik pada sebagian besar topik, dengan kesalahan klasifikasi tertinggi terjadi pada topik gangguan saluran pernapasan yang cenderung salah diprediksi ke topik dengan gejala serupa.



Gambar 4. Visualisasi Confusion Matrix SVM pada Skenario *Class Weight*



Gambar 5. Visualisasi Confusion Matrix Bi-LSTM pada Skenario *Class Weight*

3.4. Diskusi

Pada kondisi data tidak seimbang seperti dalam penelitian ini, *weighted* F1-*Score* dan *weighted* AUC lebih representatif dibandingkan *accuracy* yang rentan terhadap dominasi kelas mayoritas. *Weighted* AUC khususnya mengukur kemampuan diskriminasi model pada setiap kelas dan relatif tidak dipengaruhi oleh distribusi data, sehingga tidak terpengaruh dominasi tuberkulosis dalam dataset.

Bi-LSTM dengan skenario *class weight* menghasilkan performa terbaik dengan *weighted* F1-*Score* 0.892 dan *Weighted* AUC 0.985, sedikit mengungguli SVM dengan *weighted* F1-*Score* 0.891 dan *weighted* AUC 0.984. Keunggulan Bi-LSTM lebih nyata pada topik yang paling sulit yaitu gangguan saluran pernapasan mencatat F1-*Score* 0.725 dibandingkan 0.685 pada SVM (selisih 0.040), mengindikasikan bahwa kemampuan membaca teks dari dua arah memberikan keunggulan dalam memahami konteks gejala yang ambigu secara linguistik.

Kecilnya selisih performa secara keseluruhan dapat dijelaskan dari karakteristik data. Rata-rata panjang teks hanya sekitar 25 kata setelah prapemrosesan, sehingga Bi-LSTM tidak mendapat ruang cukup untuk memanfaatkan kemampuan *bidirectional*-nya secara optimal. Pada teks pendek, istilah medis spesifik seperti "tuberkulosis" atau "antiretroviral" sudah menjadi sinyal diskriminatif yang kuat dan dapat ditangkap TF-IDF secara efektif, sebagaimana dilaporkan pada penelitian sebelumnya bahwa model *deep learning* pun masih kesulitan membedakan kategori penyakit bergejala mirip pada teks kesehatan berbahasa Indonesia [13].

Dari aspek efisiensi komputasi, SVM bersifat deterministik dengan pelatihan dalam hitungan detik, sedangkan Bi-LSTM membutuhkan akselerasi GPU dengan waktu pelatihan 15-25 menit per skenario dan hasil yang sedikit bervariasi antar eksekusi. SVM lebih praktis untuk lingkungan dengan keterbatasan sumber daya, sementara Bi-LSTM lebih sesuai untuk skenario yang memprioritaskan kemampuan generalisasi pada kasus dengan ambiguitas linguistik tinggi.

Secara praktis, Bi-LSTM *class weight* berpotensi dimanfaatkan sebagai sistem klasifikasi awal pada platform konsultasi kesehatan daring untuk membantu pengelompokan pertanyaan secara otomatis. Namun pemilihan model dalam implementasi nyata sebaiknya mempertimbangkan faktor di luar metrik performa seperti infrastruktur komputasi, waktu respons, dan kemudahan pemeliharaan. Evaluasi lebih lanjut pada data yang lebih beragam tetap diperlukan sebelum diterapkan secara nyata.

3.5. *Deployment*

Berdasarkan hasil evaluasi, model Bi-LSTM dengan skenario *class weight* dipilih sebagai model terbaik karena memperoleh *Weighted F1-Score* sebesar 0.892. Model ini selanjutnya dapat digunakan untuk mengklasifikasikan pertanyaan pengguna ke dalam topik penyakit yang relevan pada platform konsultasi kesehatan daring. Namun model tidak dimaksudkan untuk menggantikan peran dokter, melainkan sebagai sistem pendukung keputusan pada layanan kesehatan digital.

4. PENUTUP

4.1. Kesimpulan

Penelitian ini membandingkan performa *Bidirectional Long Short-Term Memory* (Bi-LSTM) dan *Support Vector Machine* (SVM) untuk klasifikasi teks multi-kelas pertanyaan pengguna pada platform konsultasi kesehatan daring berbahasa Indonesia ke dalam tujuh kategori topik yaitu Diabetes, Diare, Gangguan Ginjal, Gangguan Saluran Pernapasan, HIV-AIDS, Stroke, Tuberkulosis. Hasil penelitian menunjukkan bahwa skenario *class weight* memberikan performa terbaik

pada kedua model. Bi-LSTM memperoleh hasil sedikit lebih unggul dengan *weighted* F1-*Score* sebesar 0.892 dan *weighted* AUC 0.985, dibandingkan SVM.

Selain itu, penelitian ini juga menemukan bahwa penerapan SMOTE tidak memberikan peningkatan performa pada model klasifikasi SVM karena ekstraksi fitur TF-IDF menghasilkan representasi teks yang sangat tersebar dan minim keterkaitan antar kata, sehingga sampel data sintetis yang terbentuk kurang merepresentasikan pola bahasa asli.

Sementara itu pada Bi-LSTM, *oversampling* masih menghasilkan performa yang cukup baik karena penambahan jumlah data membantu model mempelajari pola sekuensial dan konteks kalimat dengan lebih stabil, sedangkan *undersampling* menurunkan performa karena hilangnya informasi penting pada kelas mayoritas. Topik dengan istilah medis spesifik lebih mudah diklasifikasikan, sedangkan Gangguan Saluran Pernapasan secara konsisten memperoleh performa terendah diduga disebabkan oleh kemiripan pola bahasa dengan Tuberkulosis dan jumlah sampel yang relatif terbatas.

Penelitian ini memberikan bukti empiris perbandingan langsung Bi-LSTM dan SVM pada klasifikasi multi-kelas teks konsultasi kesehatan berbahasa Indonesia, sekaligus panduan pemilihan strategi penanganan *class imbalance* untuk domain konsultasi kesehatan daring. Secara praktis, model Bi-LSTM dengan skenario *class weight* berpotensi digunakan sebagai komponen sistem pengelompokan pertanyaan pengguna secara otomatis pada platform konsultasi kesehatan daring guna mempercepat distribusi konsultasi kepada tenaga medis yang sesuai. Namun, penelitian ini memiliki keterbatasan karena dataset hanya berasal dari satu platform sehingga generalisasi ke platform lain perlu divalidasi lebih lanjut, serta belum mengeksplorasi model berbasis transformer yang berpotensi memberikan performa lebih tinggi pada teks kesehatan berbahasa Indonesia.

4.2. Saran

Penelitian selanjutnya disarankan untuk mengeksplorasi model berbasis transformer seperti IndoBERT guna meningkatkan pemahaman konteks bahasa Indonesia pada teks konsultasi kesehatan. Selain itu, teknik data *augmentation* berbasis teks dapat dipertimbangkan sebagai alternatif penanganan *class imbalance* yang lebih sesuai dibandingkan SMOTE.

5. DAFTAR PUSTAKA

[1] M. Y. Samad, F. Rosadi, F. Azzahra, T. Brata, D. A. Permatasari, and N. T. Krisdwiyanto, "Transformasi Digital Sektor Kesehatan Pada Telemedisin Indonesia," *J. Comput. Sci. Contrib.*, vol. 5, no. 1, pp. 13–22, 2025, doi: 10.31599/amzn1a78.

[2] B. S. Santoso, R. T. Budiyanti, and N. Nandini, "Analisis Pemanfaatan Layanan Telemedicine Pasca Pandemi Covid-19 Di Jawa Tengah," *J. Manaj. Kesehat. Indones.*, vol. 12, no. 2, pp. 119–129, 2024, doi: 10.14710/jmki.12.2.2024.119-129.

[3] E. Prima, "Penggunaan Aplikasi Telekonferensi Naik 443 Persen Sejak Pandemi," *Tempo.co*, 2020. [Online]. Available: https://www.komdigi.go.id/berita/sorotan-media/detail/penggunaan-aplikasi-telekonferensi-naik-443-persen-sejak-pandemi. [Accessed: 16-May-2026].

[4] C. M. Annur, "Layanan Telemedicine yang Paling Banyak Digunakan di Indonesia, Apa Saja?," *databoks*, 2022. [Online]. Available: https://databoks.katadata.co.id/layanan-konsumen-kesehatan/statistik/a42d36cd66cec74/layanan-telemedicine-yang-paling-banyak-digunakan-di-indonesia-apa-saja. [Accessed: 16-May-2026].

[5] Widyawati, "Strategi Transformasi Digital Kesehatan 2024 Diluncurkan, Fokus ke Pelayanan Kesehatan bukan Pelaporan untuk Pejabat," *Kemenkes*, 2021. [Online]. Available: https://kemkes.go.id/id/strategi-transformasi-digital-kesehatan-2024-diluncurkan-fokus-ke-pelayanan-kesehatan-bukan-pelaporan-untuk-pejabat. [Accessed: 17-May-2026].

[6] F. Arifin, H. Sibyan, and N. Hasanah, "Rancang Bangun Chatbot Pada Sistem Ekapta Berbasis Natural Language

Processing Dengan Algoritma Artificial Neural Network," *Biner J. Ilm. Inform. dan Komput.*, vol. 4, no. 1, pp. 1–8, 2025, doi: 10.32699/biner.v4i1.7687.

[7] Ginanjar Abdurrahman and H. Oktavianto, "Klasifikasi Teks Mining Untuk Deteksi Kanker Menggunakan Support Vector Machine," *JUSTIFY J. Sist. Inf. Ibrahimy*, vol. 3, no. 1, pp. 16–20, 2024, doi: 10.35316/justify.v3i1.5028.

[8] R. N. Hanami, R. Mahendra, and A. F. Wicaksono, "Semantic classification of Indonesian consumer health questions," *J. Biomed. Semantics*, vol. 16, no. 1, 2025, doi: 10.1186/s13326-025-00334-5.

[9] Moh. Heri Setiawan, I Gede Aris Gunadi, and Gede Indrawan, "Klasifikasi Pelayanan Kesehatan Berdasarkan Data Sentimen Pelayanan Kesehatan menggunakan Multiclass Support Vector Machine," *J. Sist. dan Inform.*, vol. 17, no. 1, pp. 47–54, 2023, doi: 10.30864/jsi.v17i1.512.

[10] J. Adiputra and D. Mahdiana, "Analisis Sentimen Dengan Algoritma Support Vector Machine Terhadap Penyakit Hepatitis Akut Misterius," *J. Ilm. Inform.*, vol. 6, no. 1, pp. 1–8, 2023, doi: 10.36080/idealis.v6i1.2985.

[11] E. S. J. Atmadji, A. P. Wibowo, and E. Faizal, "Comparative Study of Machine Learning Methods for Disease Classification Based on Natural Language Symptom Descriptions," *Int. J. Artif. Intell. Med. Issues*, vol. 3, no. 2, pp. 102–109, 2025, doi: 10.56705/ijaimi.v3i2.361.

[12] S. Juanita, D. Purwitasari, I. K. E. Purnama, M. Raihan, and M. H. Purnomo, "Multi-Label Classification of Bilingual Doctor Responses in Online Medical Consultations Using Deep Learning," *J. Nas. Pendidik. Tek. Inform.*, vol. 14, no. 2, pp. 432–445, 2025, doi: 10.23887/janapati.v14i2.96980.

[13] W. O. Vihikan and I. N. P. Trisna, "Indonesian Health Question Multi-Class Classification Based on Deep Learning," *J. Inf. Syst. Informatics*, vol. 6, no. 3, pp. 1931–1944, 2024, doi: 10.51519/journalisi.v6i3.838.

[14] D. Ignasius, I. Novita Dewi, M. Bernadette Chayeenee Norman, and R. Rakhmat Sani, "Medical Named Entity Recognition from Indonesian Health-News using Bi-LSTM-CRF with Static and Contextual Embeddings," *J. Appl. Informatics Comput.*, vol. 9, no. 6, pp. 2974–2985, 2025, doi: 10.30871/jaic.v9i6.11574.

[15] Y. H. D. Wangsajaya, E. Setyowati, and A. Wibowo, "Deteksi Emosi Teks X Berbahasa Indonesia Menggunakan Bi-LSTM dengan Seleksi Fitur Chi-Square," *J. Algoritm.*, vol. 22, no. 2, pp. 546–555, 2025, doi: 10.33364/algoritma/v.22-2.2658.

[16] A. Rifai, M. Hidayat, and N. Nulngafan, "Prediksi Harga Kentang Di Wonosobo Dengan Menggunaan Metode Long Short Term Memory (Lstm)," *Biner J. Ilm. Inform. dan Komput.*, vol. 4, no. 1, pp. 9–18, 2025, doi: 10.32699/biner.v4i1.7794.

[17] S. Juanita, D. Purwitasari, I. K. E. Purnama, and M. Purnomo, "Doctor's Answer Text Dataset in Indonesian Contains Information on Medical Interview Patterns," 2023. [Online]. Available: https://data.mendeley.com/datasets/p8d5bynh3m/1. [Accessed: 19-May-2026].

[18] D. B. Saputra, V. Atina, and F. E. Nastiti, "Penerapan Model Crisp-Dm Pada Prediksi Nasabah Kredit Menggunakan Algoritma Random Forest," *IDEALIS Indones. J. Inf. Syst.*, vol. 7, no. 2, pp. 240–247, 2024, doi: 10.36080/idealis.v7i2.3244.

[19] I. Siti Aisah, B. Irawan, and T. Suprapti, "Algoritma Support Vector Machine (Svm) Untuk Analisis Sentimen Ulasan Aplikasi Al Qur'an Digital," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 6, pp. 3759–3765, 2024, doi: 10.36040/jati.v7i6.8263.

[20] S. Octaviana and Y. Mentari Indah, "Perbandingan Metode LSTM dan Bi-LSTM untuk Prediksi Data Curah Hujan," *BIAStatistics J. Stat. Theory Apl.*, vol. 19, no. 1, pp. 45–52, 2025.

[21] Ridwan, E. H. Hermaliani, and M. Ernawati, "Penerapan Metode SMOTE Untuk Mengatasi Imbalanced Data Pada Klasifikasi Ujaran Kebencian," *Comput. Sci.*, vol. 4, no. 1, pp. 80–88, 2024, doi: 10.31294/coscience.v4i1.2990.

[22] S. Diantika, "Penerapan Teknik Random Oversampling Untuk Mengatasi Imbalance Class Dalam Klasifikasi Website Phishing Menggunakan Algoritma Lightgbm," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 19–25, 2023, doi: 10.36040/jati.v7i1.6006.

[23] S. Hidayat, N. Herlina, and G. Nurul, "Penerapan Model Support Vector Machine Pada Kasus Klasifikasi Teks Berdasarkan Tujuan SGDS Ke Tiga, Empat, Dan Enam," *SisInfo*, vol. 6, no. 2, pp. 28–37, 2024, doi: 10.37278/sisinfo.v6i2.893.

[24] M. A. Fadilla, M. F. Sholahuddin, and T. Sutabri, "Pengembangan Sistem Klasifikasi Diagnosa Medis Menggunakan Progressive Web Application Terintegrasi Machine Learning," *J. Syntax Admiration*, vol. 5, no. 12, pp. 5488–5503, 2024, doi: 10.46799/jsa.v5i12.1906.