BAB I
PENDAHULUAN

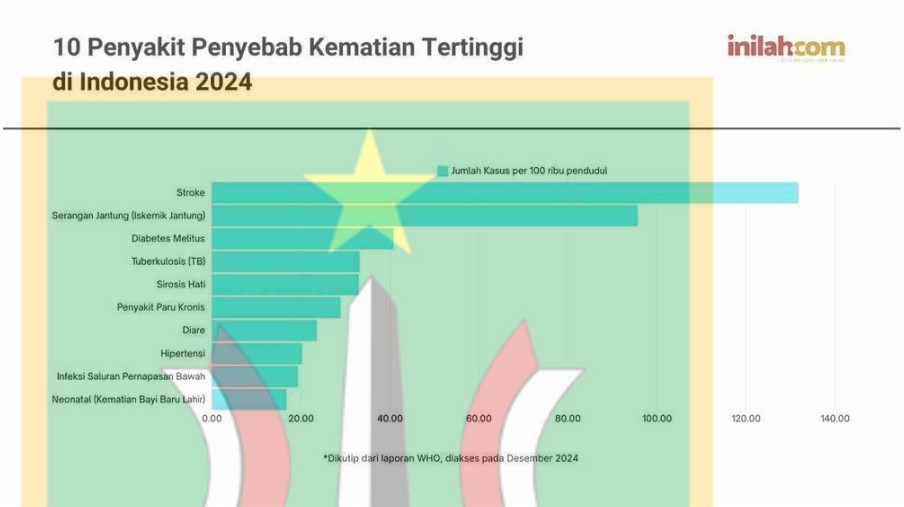
1.1 Latar Belakang Masalah

Transformasi digital telah membawa perubahan signifikan terhadap cara masyarakat mengakses dan menyebarkan informasi kesehatan. Masyarakat saat ini lebih mudah memperoleh pengetahuan mengenai gejala, pencegahan, dan pengobatan penyakit melalui media sosial seperti X (Twitter), Instagram, dan TikTok. Namun, fenomena ini juga menimbulkan dampak negatif berupa peningkatan penyebaran informasi medis yang tidak akurat atau menyesatkan, yang dapat memengaruhi persepsi dan perilaku kesehatan masyarakat (Cinelli et al., 2020).

Di luar domain kesehatan, pemanfaatan data dan teknologi komunikasi nirkabel juga terus dikembangkan untuk mendukung transformasi digital di Indonesia. Salah satu contohnya adalah penggunaan *LoRaWAN* untuk mendukung implementasi digital smart meter di kawasan perkotaan Jakarta, di mana korelasi berbagai variabel teknis seperti RSSI, SNR, elevasi, dan jarak antar gateway dianalisis secara empiris untuk memastikan kualitas konektivitas jaringan dan keandalan infrastruktur digital (Andrianingsih, Wibowo, Wirawan, Enriko, & Harahap, 2024). Penelitian tersebut menegaskan bahwa pengambilan keputusan berbasis data menjadi kunci dalam perancangan sistem digital yang andal, sejalan dengan fokus penelitian ini yang mengoptimalkan akurasi klasifikasi konten kesehatan berbahasa Indonesia berbasis penyakit.

Menurut World Health Organization (World Health Organization, 2019), penyakit tidak menular (PTM) seperti penyakit jantung, hipertensi, dan diabetes melitus menjadi penyebab utama kematian di dunia, dengan kontribusi lebih dari 70% dari total angka kematian global. Di Indonesia, tiga penyakit tersebut termasuk dalam kategori penyakit kronis dengan prevalensi tertinggi sebagaimana dilaporkan dalam Riset Kesehatan Dasar (Kementerian Kesehatan Republik Indonesia, 2019). Prevalensi hipertensi

mencapai 34,1%, diabetes sebesar 6,5%, dan penyakit jantung sebesar 1,5% dari total populasi penduduk. Kondisi ini menunjukkan bahwa ketiga penyakit kronis tersebut masih menjadi ancaman serius bagi kesehatan masyarakat Indonesia (Kementerian Kesehatan Republik Indonesia, 2024).



Gambar 1.1 Sepuluh Penyakit Penyebab Kematian Tertinggi di Indonesia Tahun 2024

Sumber: Laporan WHO, dikutip dari Inilah.com (Desember 2024)

Berdasarkan Gambar 1.1, dapat dilihat bahwa stroke dan penyakit jantung iskemik menempati urutan teratas sebagai penyebab kematian di Indonesia, diikuti oleh diabetes melitus dan hipertensi yang juga menduduki posisi sepuluh besar (WHO, 2024). Fakta ini memperlihatkan bahwa penyakit tidak menular masih mendominasi penyebab kematian masyarakat Indonesia dan menjadi tantangan utama dalam bidang kesehatan masyarakat (Budholiya, Shrivastava, & Sharma, 2022).

Seiring meningkatnya jumlah pengguna media sosial yang membahas isu kesehatan, muncul kebutuhan akan sistem otomatis yang mampu mengklasifikasikan dan mengidentifikasi konten kesehatan berdasarkan topik penyakit secara akurat. Hal ini penting untuk membantu masyarakat memilah informasi medis yang benar dan mencegah penyebaran hoaks kesehatan (Kusuma Ningrum & Maytsa Ismawardi, 2024).

Teknologi *Machine Learning (ML)* dan *Natural Language Processing (NLP)* berperan penting dalam membangun sistem klasifikasi teks otomatis. Berbagai algoritma *gradient boosting* seperti *Extreme Gradient Boosting (XGBoost)* dan *Light Gradient Boosting Machine (LightGBM)* telah terbukti efektif dalam prediksi penyakit dan pemodelan risiko kesehatan berbasis data, termasuk pada kasus yang melibatkan ketidakseimbangan kelas dan optimasi fitur (Yang & Guan, 2022).

Namun, sebagian penelitian masih mengandalkan representasi teks konvensional seperti TF-IDF, yang cenderung berbasis frekuensi sehingga kurang menangkap konteks semantik dan hubungan makna antar kata; karena itu banyak studi mengombinasikan TF-IDF dengan metode semantik/embedding untuk meningkatkan kualitas representasi (B. Wang et al., 2024)

Penerapan *machine learning* dan kecerdasan buatan di Indonesia juga telah dikembangkan pada berbagai domain, seperti kebencanaan dan pertanian. Pada ranah kebencanaan, sistem rekomendasi risiko longsor telah dibangun dengan memanfaatkan informasi kripto-spasial dan pemodelan konteks lingkungan sehingga keputusan mitigasi dapat dilakukan secara lebih akurat dan adaptif (Hindarto, Rachmadi, Hariadi, & Damastuti, 2025). Di bidang pertanian, perbandingan arsitektur *deep learning* seperti *Convolutional Neural Network (CNN)* dan Artificial Neural Network (ANN) menunjukkan bahwa pemilihan model berpengaruh signifikan terhadap performa sistem deteksi otomatis (Suherman, Hindarto, Makmur, & Santoso, 2023).

Untuk mengatasi hal tersebut, penelitian ini menggunakan *IndoBERT (Indonesian Bidirectional Encoder Representations from Transformers)* yang dikembangkan oleh (Wilie et al., 2020) sebagai model pembentuk representasi teks *(embedding)*. *IndoBERT* menghasilkan vektor numerik berisi pemahaman kontekstual dari teks berbahasa Indonesia, sehingga lebih mampu menangkap makna semantik dan struktur sintaksis dibandingkan pendekatan berbasis frekuensi kata seperti TF-IDF.

Representasi *embedding* yang dihasilkan IndoBERT kemudian digunakan sebagai input bagi model *XGBoost* dan *LightGBM* untuk proses klasifikasi.

Pendekatan ini disebut sebagai model *hybrid embedding boosting*, di mana hasil representasi semantik dari IndoBERT menjadi dasar bagi algoritma boosting untuk melakukan klasifikasi yang lebih efisien dan akurat. Metode ini menggabungkan keunggulan pemahaman konteks dari model transformer melalui embedding dengan kemampuan prediksi kuat dari algoritma *gradient boosting* (Liu, Liu, Wang, & Zhu, 2023).

Penelitian lain pada domain finansial digital menunjukkan bahwa algoritma klasifikasi seperti *Naïve Bayes* dan *Random Forest* masih banyak digunakan dalam analisis sentimen ulasan pengguna aplikasi berbahasa Indonesia dan mampu menghasilkan performa yang baik meskipun berbasis representasi teks tradisional (*sparse*) (Kaeren & Andrianingsih, 2025). Pada bidang klimatologi, penerapan algoritma Naïve Bayes untuk mengklasifikasikan curah hujan harian berdasarkan fenomena El-Niño, La-Niña, dan kondisi normal juga dilaporkan menghasilkan akurasi yang tinggi, sehingga menegaskan potensi metode klasifikasi terawasi untuk mendukung pengambilan keputusan pada domain kritis seperti iklim dan cuaca (Erlinda & Andrianingsih, 2025).

IndoBERT memiliki kemampuan untuk memahami hubungan semantik dalam teks Bahasa Indonesia yang kompleks, termasuk penggunaan bahasa informal yang umum ditemukan di media sosial (Lestari et al., 2022). Dalam penelitian terkini, Pendekatan ini disebut sebagai model *hybrid embedding boosting*, di mana hasil representasi semantik dari IndoBERT menjadi dasar bagi algoritma boosting untuk melakukan klasifikasi yang lebih efisien dan akurat. Metode ini menggabungkan keunggulan pemahaman konteks dari model transformer melalui embedding dengan kemampuan prediksi kuat dari algoritma *gradient boosting* (Liu et al., 2023).

Dari sisi sosial, penelitian ini berkontribusi terhadap peningkatan literasi kesehatan digital dan pengendalian misinformasi medis di media

sosial, karena pendekatan klasifikasi otomatis yang menangkap konteks semantik konten dapat membantu memfilter dan mengklasifikasikan informasi yang tidak akurat (Eichinger et al., 2020). Sedangkan dari sisi akademik, penelitian ini memperluas penerapan model transformer IndoBERT pada domain klasifikasi konten kesehatan berbahasa Indonesia, yang masih jarang dilakukan di penelitian terdahulu (Wilie et al., 2020)

Secara keseluruhan, penelitian ini berfokus pada upaya optimasi akurasi kategorisasi konten kesehatan berbasis penyakit menggunakan kombinasi *IndoBERT, XGBoost,* dan *LightGBM*. Pendekatan ini diharapkan mampu meningkatkan akurasi, efisiensi, serta relevansi linguistik sistem klasifikasi konten kesehatan digital di Indonesia, sekaligus memberikan kontribusi nyata terhadap validitas informasi medis di ruang digital.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, dapat diidentifikasi adanya permasalahan mengenai belum optimalnya akurasi sistem klasifikasi konten kesehatan berbahasa Indonesia dalam mengelompokkan topik penyakit seperti jantung, hipertensi, dan diabetes. Hal ini disebabkan oleh keterbatasan metode representasi teks konvensional serta belum diterapkannya model *IndoBERT embedding* yang dikombinasikan dengan algoritma *XGBoost* dan *LightGBM* untuk menghasilkan klasifikasi yang lebih akurat. Oleh karena itu, penelitian ini merumuskan beberapa pertanyaan penelitian sebagai berikut:

1. Bagaimana penerapan model *IndoBERT* dapat menghasilkan representasi *(embedding)* teks berbahasa Indonesia dalam klasifikasi konten kesehatan berbasis penyakit?
2. Bagaimana perbandingan performa algoritma *XGBoos*t dan *LightGBM* dalam mengklasifikasikan konten kesehatan menggunakan hasil *embedding IndoBERT*?

3. Bagaimana hasil optimasi akurasi sistem klasifikasi konten kesehatan berbasis penyakit melalui kombinasi *IndoBERT embedding, XGBoost, dan LightGBM*?

Rumusan masalah ini berfungsi sebagai pedoman utama dalam pelaksanaan penelitian, sekaligus menjadi dasar untuk menentukan arah analisis dan metode yang digunakan dalam menjawab permasalahan optimasi akurasi klasifikasi konten kesehatan digital berbahasa Indonesia.

1.3 Tujuan Penelitian

Penelitian ini bertujuan untuk memberikan pemahaman ilmiah dan komprehensif mengenai optimasi akurasi klasifikasi konten kesehatan berbahasa Indonesia berbasis penyakit dengan menggunakan pendekatan *IndoBERT* embedding yang dikombinasikan dengan algoritma *XGBoost* dan *LightGBM*. Tujuan ini disusun sebagai jawaban atas rumusan masalah yang telah dijabarkan sebelumnya, dengan fokus pada proses representasi teks, perbandingan performa algoritma, serta optimasi hasil klasifikasi konten kesehatan digital. Adapun tujuan penelitian ini adalah sebagai berikut:

1. Menerapkan model *IndoBERT* untuk menghasilkan *embedding* teks berbahasa Indonesia sebagai representasi fitur dalam proses klasifikasi konten kesehatan berbasis penyakit.
2. Menganalisis dan membandingkan performa algoritma *XGBoost* dan *LightGBM* dalam mengklasifikasikan konten kesehatan berdasarkan hasil *embedding IndoBERT* untuk memperoleh model dengan performa terbaik.
3. Mengoptimalkan akurasi sistem klasifikasi konten kesehatan berbasis penyakit melalui kombinasi *IndoBERT embedding*, *XGBoost*, dan *LightGBM*, sehingga dapat menghasilkan model yang lebih efisien dan bermanfaat dalam mendukung validasi serta penyebaran informasi kesehatan yang benar kepada masyarakat.

1.4 Batasan Masalah

Agar penelitian ini memiliki arah yang terfokus dan hasil yang dapat dianalisis secara objektif, maka ruang lingkup penelitian perlu dibatasi agar sesuai dengan tujuan utama, yaitu mengoptimalkan akurasi sistem klasifikasi konten kesehatan berbahasa Indonesia berbasis penyakit menggunakan pendekatan *IndoBERT* embedding dan algoritma *XGBoost* serta *LightGBM*. Pembatasan ini juga dilakukan agar penelitian tetap relevan, terukur, dan dapat direplikasi pada penelitian selanjutnya. Adapun batasan masalah dalam penelitian ini dijelaskan sebagai berikut:

1. **Data Penelitian**

   Data yang digunakan dalam penelitian ini berupa konten teks berbahasa Indonesia yang diambil dari media sosial Twitter (X). Data difokuskan pada topik penyakit jantung, hipertensi, dan diabetes melitus sebagai representasi dari penyakit tidak menular (PTM).

2. **Cakupan Analisis**

   Analisis penelitian dibatasi pada proses klasifikasi konten kesehatan berbasis penyakit, tanpa membahas aspek medis atau validasi kebenaran isi dari konten tersebut. Fokus penelitian adalah pada peningkatan akurasi model klasifikasi teks.

3. **Fokus Kajian Penelitian**

   Penelitian ini difokuskan pada tahapan pembentukan representasi teks (embedding) menggunakan model IndoBERT, serta penerapan dan evaluasi algoritma XGBoost dan LightGBM untuk menghasilkan model klasifikasi terbaik berdasarkan metrik evaluasi yang digunakan.

4. **Metode Analisis**

   Analisis dilakukan menggunakan pendekatan *machine learning supervised classification*. Proses pelatihan model menggunakan metode *K-Fold Cross Validation* untuk memperoleh hasil yang

stabil dan objektif. Evaluasi performa model dilakukan menggunakan metrik *accuracy, precision, recall,* dan *f1-score*.

5. **Batasan Prapemrosesan Teks (Normalisasi)**

   Penelitian ini belum mengakomodasi normalisasi lanjutan terhadap singkatan serta istilah bahasa daerah atau bahasa asing pada teks media sosial.

6. **Perangkat Lunak yang Digunakan**

   Penelitian ini menggunakan bahasa pemrograman Python dengan bantuan beberapa pustaka (library) seperti pandas, numpy, scikit-learn, lightgbm, dan xgboost untuk pengolahan data, pelatihan model, dan visualisasi hasil.

7. **Keluaran Penelitian**

   Hasil akhir penelitian difokuskan pada model klasifikasi konten kesehatan berbasis penyakit yang memiliki akurasi tertinggi. Model ini diharapkan dapat mendukung upaya peningkatan literasi kesehatan digital dan membantu masyarakat dalam mengenali informasi kesehatan yang valid dan akurat.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat baik secara teoritis, akademik, maupun praktis, khususnya dalam pengembangan teknologi pemrosesan bahasa alami *(Natural Language Processing/NLP)* untuk Bahasa Indonesia. Selain itu, hasil penelitian ini diharapkan dapat berkontribusi terhadap peningkatan literasi kesehatan digital masyarakat serta mendukung upaya penyebaran informasi medis yang akurat dan dapat dipercaya di media sosial.

1. **Manfaat Teoritis**

   Secara teoritis, penelitian ini memberikan kontribusi terhadap pengembangan ilmu di bidang *Natural Language Processing (NLP)*, terutama dalam konteks Bahasa Indonesia. Penelitian ini menunjukkan penerapan model *IndoBERT* sebagai pembentuk

embedding teks yang dikombinasikan dengan algoritma *XGBoost* dan *LightGBM* untuk meningkatkan akurasi klasifikasi konten kesehatan berbasis penyakit. Selain itu, penelitian ini dapat menjadi referensi untuk pengembangan model *hybrid embedding boosting* pada penelitian klasifikasi teks di masa mendatang.

2. **Manfaat Praktis**

   Secara praktis, hasil penelitian ini dapat dimanfaatkan untuk mengembangkan sistem klasifikasi otomatis konten kesehatan digital yang lebih akurat dan kontekstual. Sistem ini dapat membantu instansi kesehatan, lembaga pemerintahan, dan masyarakat umum dalam mengenali topik penyakit pada konten teks secara cepat dan objektif. Dengan demikian, penelitian ini dapat mendukung peningkatan validitas dan literasi informasi kesehatan digital, serta berperan dalam upaya mencegah penyebaran hoaks atau informasi medis yang menyesatkan di media sosial.

3. **Manfaat Akademik**

   Secara akademik, penelitian ini dapat menjadi referensi bagi mahasiswa dan peneliti dalam memahami penerapan model *transformer (IndoBERT)* dan integrasinya dengan algoritma gradient boosting seperti *XGBoost* dan *LightGBM*. Penelitian ini juga dapat memperkaya wawasan di bidang data science, pembelajaran mesin, dan informatika kesehatan, serta menjadi contoh implementasi nyata penerapan machine learning untuk analisis teks berbahasa Indonesia dalam konteks akademik.

1.6 Kontribusi Penelitian

Penelitian ini memberikan beberapa kontribusi penting dalam bidang pemrosesan bahasa alami *(Natural Language Processing/NLP)* dan klasifikasi teks berbahasa Indonesia, khususnya di ranah kesehatan digital. Kontribusi utama penelitian ini terletak pada pengembangan model *hybrid IndoBERT–Boosting*, yaitu penggabungan antara model representasi bahasa

berbasis transformer *(IndoBERT)* dengan algoritma pembelajaran mesin *XGBoost* dan *LightGBM*. Pendekatan ini diusulkan untuk meningkatkan akurasi sistem klasifikasi teks berbahasa Indonesia, khususnya pada konten yang berkaitan dengan topik penyakit kronis seperti jantung, hipertensi, dan diabetes melitus.

Selain itu, penelitian ini memperlihatkan kemampuan *IndoBERT* embedding dalam menangkap makna semantik teks kesehatan berbahasa Indonesia secara lebih mendalam dibandingkan metode tradisional seperti TF-IDF. Representasi berbasis *IndoBERT* ini memungkinkan model klasifikasi untuk memahami konteks kata secara dua arah, sehingga hasil klasifikasi menjadi lebih akurat dan relevan. Penelitian ini juga berkontribusi dalam memberikan analisis komparatif terhadap performa algoritma *XGBoost* dan *LightGBM*, dengan tujuan untuk menentukan model *boosting* yang paling efektif dalam memproses hasil *embedding* dari *IndoBERT*.

Dari sisi implementasi, penelitian ini menghasilkan optimasi akurasi sistem klasifikasi konten kesehatan berbasis penyakit yang dapat menjadi dasar bagi pengembangan sistem deteksi dan kategorisasi konten kesehatan digital di Indonesia. Secara sosial, penelitian ini juga berkontribusi terhadap peningkatan literasi dan edukasi kesehatan masyarakat, karena hasil model dapat membantu dalam memfilter serta mengidentifikasi informasi kesehatan yang valid dan relevan. Dengan demikian, penelitian ini tidak hanya memberikan nilai tambah bagi pengembangan metode NLP di bidang akademik, tetapi juga memiliki manfaat praktis dalam mendukung penyebaran informasi kesehatan yang akurat di era digital.