SISTEM TANYA JAWAB DETEKSI KANKER DINI MENGGUNAKAN METODE BERT DAN TF-IDF

Diajukan Sebagai Syarat Untuk Menyelesaikan
Pendidikan Program Strata – 1 Pada
Jurusan Teknik Informatika

Oleh :

Louis Garcia
NIM : 09021182126006

Jurusan Teknik Informatika
FAKULTAS ILMU KOMPUTER UNIVERSITAS SRIWIJAYA
2025

HALAMAN PENGESAHAN

SKRIPSI

Sistem Tanya Jawab Deteksi Kanker Dini dengan Metode BERT dan TF - IDF

Sebagai salah satu syarat untuk penyelesaian studi di

Program Studi S1 Teknik Informatika

Oleh:

LOUIS GARCIA

09021182126006

Pembimbing 1 : **Dr. Abdiansah,S.Kom., M.Cs**
NIP. 198410012009121005

Mengetahui

Ketua Jurusan Teknik Informatika

Hadipurnawan Satria, Ph.D

198004182020121001

TANDA LULUS UJIAN KOMPREHENSIF SKRIPSI

Pada hari Jumat tanggal 28 Februari 2025 telah dilaksanakan Ujian Komprehensif skripsi oleh Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Nama : Louis Garcia
NIM : 09021182126006
Judul : Sistem Tanya Jawab Deteksi Kanker Dini dengan Metode BERT dan TF – IDF

dan dinyatakan **LULUS.**

1. Ketua Penguji

   Yunita, M.Cs
   NIP. 197812222006042003 .....

2. Penguji I

   Novi Yusliani, M.T
   NIP. 198912212020122011 .....

3. Pembimbing I

   Dr. Abdiansah, S.Kom., M.CS.
   NIP. 198410012009121005 .....

Mengetahui
Ketua Jurusan Teknik Informatika

Hadipurnawan Satria, Ph.D.
NIP. 198004182020121001

HALAMAN PERNYATAAN BEBAS PLAGIAT

Yang bertanda tangan dibawah ini :

Nama : Louis Garcia

NIM : 09021182126006

Program Studi : Teknik Informatika Reguler

Judul : Sistem Tanya Jawab Deteksi Kanker Dini dengan Metode BERT dan TF – IDF

Hasil Pengecekan Software iThenticate/Turnitin: 3%

Menyatakan bahwa laporan skripsi saya merupakan hasil karya sendiri dan bukan hasil penjiplakan/plagiat. Apabila ditemukan unsur penjiplakan atau plagiat dalam laporan tugas akhir ini, maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya.

Demikian pernyataan ini saya buat dengan sebenarnya dan tidak ada paksaan dari siapapun.

Palembang, 19 Februari 2025

Penulis,

Louis Garcia

NIM. 09021182126006

MOTTO DAN PERSEMBAHAN

"Hidup itu seperti roda yang berputar, Namun yang tepenting teruslah bergerak karena setiap putaran membawa peluang yang baru"

Kupersembahkan Skripsi ini kepada :

- Sang Hyang Adi Buddha, Tuhan Yang Maha Esa
- Orang Tua
- Keluarga Besar
- Sahabat Dekat Penulis
- Almamater Fakultas Ilmu Komputer Universitas Sriwijaya

ABSTRACT

The increasing mortality rate due to cancer, particularly in developing countries like Indonesia, highlights the urgency of developing an effective question-answering detection system. According to data from Globocan 2020, Indonesia recorded 396,914 new cancer cases with 234,511 cancer-related deaths. Additionally, Riskesdas data shows that the prevalence of cancer in Indonesia increased from 1.4 per 1,000 population in 2013 to 1.79 per 1,000 population in 2018. This study aims to develop a cancer early detection question-answering system using BERT (Bidirectional Encoder Representations from Transformers) and TF-IDF (Term Frequency-Inverse Document Frequency) methods. The combination of these two methods is expected to improve the accuracy in understanding cancer symptoms, diagnosis, and treatment. The system was tested using a dataset from Kaggle containing clinical data on various types of cancer, with preprocessing techniques such as case folding, stop word removal, stemming, and tokenization applied to enhance data quality. The system's performance evaluation showed the highest accuracy of 98.85%, achieved with a fine-tuned BERT model. In comparison with the BERT-only model (94.70%) and TF-IDF-only model (96.55%), these results demonstrate that the integration of BERT and TF-IDF is more effective in providing accurate and relevant responses. This study also involved interviews with 10 medical students from Universitas Sriwijaya, class of 2021-2022, to test the validity of the system. Of the 20 questions asked, the system successfully answered 19 correctly, resulting in an accuracy of 95%. The findings of this study contribute to the development of artificial intelligence (AI)-based health technology and support early cancer detection efforts in Indonesia by providing an efficient and reliable cancer detection system.

Key Points : *BERT, TF-IDF, NLP. Model. AI-Based, System, Cancer*

ABSTRAK

Meningkatnya angka kematian akibat kanker, terutama di negara berkembang seperti Indonesia, menunjukkan urgensi pengembangan sistem deteksi tanya jawab yang efektif. Berdasarkan data dari Globocan 2020, Indonesia mencatatkan 396.914 kasus kanker baru dengan 234.511 kematian akibat kanker. Selain itu, data Riskesdas menunjukkan bahwa prevalensi kanker di Indonesia meningkat dari 1,4 per 1.000 penduduk pada tahun 2013 menjadi 1,79 per 1.000 penduduk pada tahun 2018. Penelitian ini bertujuan untuk mengembangkan sistem tanya jawab deteksi dini kanker dengan memanfaatkan metode BERT (*Bidirectional Encoder Representations from Transformers*) dan TF-IDF (*Term Frequency-Inverse Document Frequency*). Kombinasi kedua metode ini diharapkan dapat meningkatkan akurasi dalam memahami gejala, diagnosis, dan pengobatan kanker. Sistem diuji menggunakan dataset dari *Kaggle* yang berisi data klinis mengenai berbagai jenis kanker, dengan penerapan teknik prapemrosesan data seperti case folding, stop word removal, stemming, dan tokenisasi untuk meningkatkan kualitas data. Evaluasi kinerja sistem menunjukkan akurasi tertinggi sebesar 98,85%, dicapai pada pengujian dengan model BERT yang telah *di fine-tuned.* Dalam perbandingan dengan model BERT saja (94,70%) dan TF-IDF saja (96,55%), hasil ini menunjukkan bahwa penggabungan BERT dan TF-IDF lebih efektif dalam memberikan respons yang akurat dan relevan. Penelitian ini juga melibatkan wawancara dengan 10 mahasiswa kedokteran Universitas Sriwijaya angkatan 2021-2022 untuk menguji kevalidan sistem. Dari 20 pertanyaan yang diajukan, sistem berhasil menjawab 19 pertanyaan dengan benar, menghasilkan akurasi 95%. Hasil penelitian ini berkontribusi pada pengembangan teknologi kesehatan berbasis kecerdasan buatan (AI) dan mendukung upaya deteksi dini kanker di Indonesia, dengan memberikan sistem deteksi kanker yang efisien dan handal.

Kata Kunci: BERT, TF-IDF, NLP, Model, Berbasis AI, Sistem, Kanker

KATA PENGANTAR

Dengan mengucapkan puji dan syukur kepada Sang Hyang Adi Buddha, Tuhan Yang Maha Esa, yang telah memberikan bimbingan, rahmat, dan anugerah-Nya, penulis dapat menyelesaikan skripsi ini sebagai bagian dari perjalanan dalam menyelesaikan pendidikan di Program Sarjana Teknik Informatika, Fakultas Ilmu Komputer, Universitas Sriwijaya, dengan judul *"*SISTEM TANYA JAWAB DETEKSI KANKER DINI DENGAN METODE BERT DAN TF-IDF*"*. Penulis menyadari bahwa skripsi ini tidak akan selesai tanpa adanya bantuan dari dosen pembimbing skripsi, dosen pembimbing akademik, serta dukungan dari berbagai pihak lainnya yang telah turut berkontribusi dalam proses penyelesaian skripsi ini. Pada kesempatan ini, penulis ingin menyampaikan rasa terima kasih yang sebesar-besarnya kepada:

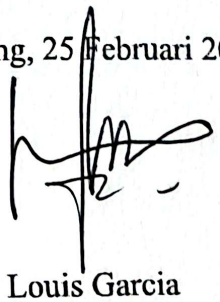
1. Sang Hyang Adi Buddha, Tuhan Yang Maha Esa yang telah melimpahkan bimbingan dan karunia-Nya selama proses pengerjaan skripsi.
2. Kedua orang tua penulis Bapak Tjahyadi Gunawan dan Ibu Halima Hawa yang senantiasa mendampingi dan mendukung penulis untuk menyelesaikan skripsi ini.
3. Kedua adik penulis, Felix Gunawan dan Grace Nathalie yang juga selalu mendukung dan mendampingi penulis dalam menyelesaikan skripsi.
4. Bapak Prof. Dr. Erwin, S.Si., M.Si selaku Dekan Fakultas Ilmu Komputer Universitas Sriwijaya
5. Bapak Hadipurnawan Satria, Ph.D. selaku Kepala Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya

6. Bapak Dr. Abdiansah, S.Kom., M.Cs., selaku dosen pembimbing penulis yang senantiasa hadir dan membimbing penulis sejak awal penulisan skripsi hingga selesai, serta selalu memberikan nasihat dan arahan yang konstruktif selama proses pengerjaan skripsi.
7. Bapak Samsuryadi, M.Kom., Ph.D. selaku dosen pembimbing akademik penulis yang telah mengarahkan penulis dari semester 1 hingga semester akhir mengenai pemilihan konsentrasi dan mata kuliah sehingga penulis dapat memilih konsentrasi sesuai dengan yang penulis inginkan.
8. Bapak dan Ibu Dosen Jurusan Teknik Informatika yang telah memberikan ilmu dan Pelajaran selama penulis melaksanakan perkuliahan.
9. Staf Admin Jurusan Teknik Informatika dan Staf Fakultas Ilmu Komputer yang telah membantu urusan administrasi sekaligus akademis penulis.
10. Teman – teman Badan Pengurus Harian Keluarga Mahasiswa Buddhis Palembang (KMBP) yang selalu mendukung dan mendampingi penulis dari awal hingga akhir penyelesaian skripsi.
11. Teman – teman Badan Pengurus Harian Fasilkom Science Community (FASCO) 2022 – 2023 yang senantiasa mendukung dalam pengerjaan skripsi.
12. Teman – teman seangkatan khususnya TI Reguler – L1 yang selalu mendukung dan menemani penulis dari awal perkuliahan hingga sekarang.

Sebagai penutup, penulis menyadari bahwa skripsi ini masih belum sempurna. Oleh karena itu, penulis sangat mengharapkan kritik dan saran dari

para pembaca demi perbaikan dan pengembangan di masa mendatang. Penulis juga berharap agar skripsi ini dapat memberikan kontribusi yang bermanfaat bagi kemajuan teknologi dan ilmu komputer, khususnya dalam bidang kesehatan di Indonesia.

Palembang, 25 Februari 2025



Louis Garcia

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB 1

PENDAHULUAN

1.1. Pendahuluan

Bab ini akan membahas secara rinci berbagai komponen utama penelitian ini. Latar belakang masalah yang mendasari penelitian, perumusan masalah utama, batas-batas yang diterapkan untuk memperjelas ruang lingkup penelitian, tujuan yang ingin dicapai melalui penelitian, dan manfaat yang diharapkan dari hasil penelitian adalah beberapa topik yang akan dibahas. Kajian penelitian secara keseluruhan akan bergantung pada setiap pokok pikiran yang dibahas.

1.2. Latar Belakang Masalah

Kanker adalah penyebab kematian kedua terbesar di dunia, dengan 9,6 juta kematian per tahun, dan 70% di antaranya terjadi di negara berkembang, termasuk Indonesia. Pada 2020, Indonesia mencatat 396.914 kasus baru dan 234.511 kematian akibat kanker. Kanker payudara dan kanker leher rahim adalah yang paling umum pada perempuan, sedangkan pada laki-laki, kanker paru-paru dan kolorektal mendominasi (Direktorat Jenderal Pelayanan Kesehatan, 2024).

Berdasarkan permasalahan yang ada, salah satu solusi yang dapat ditawarkan adalah sistem tanya jawab. Melalui sistem tanya jawab, pengguna dapat mengajukan pertanyaan dengan menggunakan bahasa alami, dan komputer akan memahami maksud pertanyaan tersebut, kemudian memberikan jawaban yang sesuai. Dengan cara ini, pengguna dapat memperoleh informasi secara langsung tanpa harus melakukan pencarian di mesin pencari atau membuka situs web

kesehatan, yang mengharuskan mereka untuk menavigasi konten yang diinginkan. Pendekatan ini sangat membantu dalam upaya deteksi kanker dini, memberikan akses cepat dan akurat terhadap informasi yang diperlukan (Kusumaningrum et al., 2023). Sistem tanya jawab dalam *Natural Language Processing* (NLP) adalah teknologi yang memungkinkan komputer untuk memahami dan merespons pertanyaan yang diajukan oleh pengguna dalam bahasa alami. Sistem ini menggunakan berbagai teknik NLP untuk menganalisis dan memahami konteks serta maksud dari pertanyaan yang diberikan, kemudian memberikan jawaban yang relevan (Burger et al., 2001).

Term Frequency-Inverse Document Frequency (TF – IDF) adalah metode yang menggabungkan pendekatan *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF). Metode ini diusulkan oleh Salton sebagai kombinasi yang meningkatkan performa, terutama dalam meningkatkan nilai *recall* dan *precision. Weighted Inverse Document Frequency* (WIDF) merupakan perluasan dari IDF (Deolika et al., 2019).

Penelitian lain yang dilakukan oleh Sasmita (2018) mengatakan bahwa algoritma TF – IDF dapat digunakan untuk mengembangkan sistem penentuan jawaban otomatis untuk sistem informasi *E-Complaint.* Selain itu, penelitian yang dilakukan oleh Musfiroh (2013) mengatakan bahwa algoritma *term frequency-inverse document frequency* merupakan salah satu metode yang tepat untuk digunakan dalam pencarian kata ditiap dokumen. Hal ini yang menjadi alasan peneliti untuk menggunakan metode TF – IDF dalam tugas akhir.

Dalam pengembangan sistem tanya jawab, beberapa teknik preprocessing sangat penting untuk meningkatkan kualitas data. Teknik-teknik seperti *case folding, stopword removal, stemming* dan tokenisasi dapat membantu mengurangi kompleksitas data serta meningkatkan representasi teks yang lebih relevan untuk proses klasifikasi. Selain teknik-teknik tradisional ini, pendekatan yang lebih maju seperti BERT (*Bidirectional Encoder Representations from Transformers*) telah terbukti lebih efektif dalam menghasilkan representasi teks kontekstual, terutama ketika data latih terbatas (Subhi et al., 2024). Hal ini dapat meningkatkan performa klasifikasi pada beberapa metode machine learning. Model BERT dikembangkan sebagai model deep learning yang dilatih secara *bidirectional* menggunakan data teks tanpa label. Hal ini dicapai dengan mengintegrasikan lapisan konteks dari sisi kiri dan kanan. Akibatnya, model BERT dapat disesuaikan untuk berbagai tugas pembelajaran mesin, seperti klasifikasi dan penjawaban pertanyaan, hanya dengan menambahkan satu lapisan (Budiman et al., 2024).

Penelitian Diah (2024) menunjukkan bahwa model BERT yang disesuaikan dengan baik berhasil menganalisis sentimen ulasan film "Dirty Vote" dengan skor akurasi, presisi, recall, dan F1 di atas 0,8. Sementara itu, Yudi (2021) menemukan bahwa BERT mencapai F1-score 0,9327 (93,27%) dalam klasifikasi teks berita Bahasa Indonesia. Kedua penelitian ini membuktikan keunggulan BERT dalam klasifikasi teks, yang menjadi alasan penulis menggunakannya dalam tugas akhir.

Penelitian Sabine (2021) menunjukkan bahwa penggabungan model BERT dengan vektorisasi TF-IDF menghasilkan kinerja yang lebih baik dalam tugas pencarian hukum berbasis undang-undang. BERT memberikan pemahaman

kontekstual yang mendalam, sementara TF-IDF menangkap hubungan eksplisit antar kata. Kombinasi keduanya menghasilkan model yang lebih akurat dan informatif. Pengujian terhadap model pra-latih, fine-tuning, serta penambahan pengetahuan eksternal dan augmentasi data juga meningkatkan performa model. Pendekatan terbaik yang digunakan adalah model ensemble, menggabungkan Sentence-BERT dengan dua representasi TF-IDF, yang menghasilkan akurasi lebih tinggi. Fine-tuning pada BERT classifier dan metode similarity scoring berbasis BERTScore semakin memperkuat sistem, menjadikan penggabungan BERT dan TF-IDF efektif dalam memahami teks hukum yang kompleks dan meningkatkan akurasi sistem pencarian dokumen hukum.

Berdasarkan referensi penelitian yang diuraikan, algoritma BERT yang dikombinasikan dengan TF-IDF dipilih sebagai metode dalam pengembangan sistem tanya jawab untuk deteksi kanker dini. Penggunaan BERT sebagai model pengolahan bahasa alami diharapkan dapat meningkatkan kualitas respons, mengingat kemampuannya dalam memahami konteks dan nuansa bahasa secara lebih mendalam dibandingkan algoritma tradisional. Di sisi lain, TF-IDF berfungsi sebagai teknik representasi teks yang memperkuat pemahaman tentang pentingnya istilah dalam dokumen. Kombinasi ini memungkinkan sistem memberikan jawaban yang lebih relevan dan akurat terhadap pertanyaan mengenai deteksi kanker, sehingga mendukung upaya deteksi kanker dini dengan lebih efektif.

1.3. Rumusan Masalah

Berdasarkan latar belakang tersebut, rumusan masalah dapat dirumuskan sebagai berikut.

1. Bagaimana cara mengembangkan sistem tanya jawab deteksi kanker dini menggunakan metode BERT dan TF – IDF?
2. Bagaimana kinerja algoritma BERT dan TF – IDF dalam pengembangan sistem tanya jawab deteksi kanker dini?

1.4. Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dipaparkan, maka terdapat tujuan yang ingin dicapai.

1. Menghasilkan sistem tanya jawab deteksi kanker dini dengan menggunakan metode BERT dan TF – IDF.
2. Mengetahui kinerja algoritma BERT dan TF – IDF dalam pengembangan sistem tanya jawab deteksi kanker dini.

1.5. Manfaat Penelitian

Adapun berbagai manfaat yang dapat diperoleh dari penelitian ini.

1. Memberikan wawasan kepada pengembang aplikasi mengenai efektivitas penggunaan algoritma BERT dan TF-IDF dalam sistem tanya jawab sehingga dapat menjadi sumber inspirasi dalam pengembangan aplikasi lainnya untuk mencapai hasil yang optimal dalam pelaksanaan sistem tanya jawab.

2. Membantu tenaga kesehatan dalam mendiagnosis sel kanker secara cepat dan akurat, sehingga langkah pencegahan penyakit dapat dilaksanakan dengan lebih efektif daripada sebelumnya.

1.6. Batasan Masalah

Adapun beberapa Batasan masalah yang menjadi batas atau limitasi dalam penelitian ini sebagai berikut.

1. Dataset yang digunakan terbatas hanya diambil melalui platform *website Kaggle.*
2. Data yang digunakan adalah data hasil klinis dalam Bahasa inggris yang kemudian akan diterjemahkan ke dalam Bahasa Indonesia.
3. Data kanker yang digunakan terbatas pada 16 jenis kanker saja dan tidak mencakup semua jenis kanker.
4. Dataset yang digunakan hanya dataset Bahasa Indonesia saja.

1.7. Sistematika Penulisan

Sistematika penulisan tugas akhir mengikuti standar penulisan tugas akhir Fakultas Ilmu Komputer Universitas Sriwijaya yaitu sebagai berikut:

BAB I. PENDAHULUAN

Bab ini menguraikan tentang latar belakang, pertanyaan penelitian, tujuan dan manfaat penelitian, batasan masalah, serta sistematika desain penelitian yang menjadi fokus penelitian ini.

BAB II. KAJIAN LITERATUR

Bab ini membahas landasan teori yang digunakan dalam penelitian, meliputi definisi algoritma TF-IDF, BERT, serta konsep sistem tanya jawab, pemrosesan data, dan literatur relevan yang mendukung penelitian ini.

BAB III. METODOLOGI PENELITIAN

Bab ini menjelaskan setiap langkah yang dilakukan selama proses penelitian, mencakup pengumpulan data, analisis data, dan perancangan perangkat lunak. Setiap langkah dijelaskan secara rinci berdasarkan kerangka kerja yang telah disusun.

BAB IV. PENGEMBANGAN PERANGKAT LUNAK

Bab ini membahas perancangan perangkat lunak yang akan dibangun, mulai dari analisis kebutuhan hingga tahap pengujian untuk menilai perkembangan.

BAB V. HASIL DAN PEMBAHASAN

Bab ini akan membahas hasil analisis dari penelitian yang dilakukan menggunakan pendekatan yang telah ditetapkan sebelumnya. Analisis ini digunakan untuk menarik kesimpulan dari penelitian.

BAB VI. PENUTUP

Bab ini akan membahas kesimpulan dari hasil penelitian serta saran untuk meningkatkan penelitian di masa depan.

1.8. Kesimpulan

Pada bab ini telah dijelaskan latar belakang, rumusan masalah, tujuan dan manfaat penelitian, batasan masalah serta sistematika penelitian yang dijadikan

pedoman dan landasan dalam pengembangkan model tanya jawab deteksi kanker dini dengan menggunakan metode BERT dan TF – IDF.

DAFTAR PUSTAKA

Azizah, A. N., Asy'ari, M. F., Prastya, I. W. D., & Purwitasari, D. (2023). Easy Data Augmentation untuk Data yang Imbalance pada Konsultasi Kesehatan Daring. *Jurnal Teknologi Informasi dan Ilmu Komputer*, *10*(5), 1095-1104.

Bahani, K., Moujabbir, M., & Ramdani, M. (2021). An accurate fuzzy rule-based classification systems for heart disease diagnosis. Scientific African, 14, e01019.

Budiman, I., Faisal, M. R., Faridhah, A., Farmadi, A., Mazdadi, M. I., Saragih, T. H., & Abadi, F. (2024). Classification Performance Comparison of BERT and IndoBERT on SelfReport of COVID-19 Status on Social Media. Journal of Computer Sciences Institute, 30, 61-67.

Burger, J., Cardie, C., Chaudhri, V., Gaizauskas, R., Harabagiu, S., Israel, D., ... & Weishedel, R. (2001, July). Issues, tasks and program structures to roadmap research in question & answering (Q&A). In Document Understanding Conferences Roadmapping Documents (pp. 1-35).

Cahyani, D. E., & Patasik, I. (2021). Performance comparison of tf-idf and word2vec models for emotion text classification. Bulletin of Electrical Engineering and Informatics, 10(5), 2780-2788.

Cendana, M., & Permana, S. D. H. (2019). Pra-Pemrosesan Teks Pada Grup Whatsapp Untuk Pemodelan Topik. Jurnal Mantik Penusa, 3(3).

Chandradev, V., Suarjaya, I. M. A. D., & Bayupati, I. P. A. (2023). Analisis Sentimen Review Hotel Menggunakan Metode Deep Learning BERT. Jurnal Buana Informatika, 14(02), 107-1

Deolika, A., Kusrini, K., & Luthfi, E. T. (2019). Analisis pembobotan kata pada klasifikasi text mining. (JurTI) Jurnal Teknologi Informasi, 3(2), 179-184.

Direktorat Jenderal Pelayanan Kesehatan. (2024). Direktorat Jenderal Pelayanan Kesehatan - Kementerian Kesehatan Republik Indonesia. Diakses pada 20 Juli 2024, dari https://kemkes.go.id.

Esairina. (2024, September 22). Mengenal term frequency-inverse document frequency (TF-IDF) pada model NLP. Medium. diakses pada 27 Oktober 2024, dari https://esairina.medium.com/mengenal-term-frequency-inverse-ocument-frequency-tf-idf-pada-model-nlp-e0cc571f7e37/

Gorunescu, F. (2011). Data Mining: Concepts, models and techniques (Vol. 12). Springer Science & Business Media.

Gumilar, F. R., Syahidin, Y. Y., & Sonia, D. (2021). Perancangan Sistem Informasi Kunjungan Pasien Bpjs Rawat Jalan Dengan V-Model. Explor. Sist. Inf. dan Telemat, 12(2), 204

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers. https://www.sciencedirect.com/science/article/abs/pii/B9780123814790000048

Hamidi, M. Z., Purwitasari, D., & Esti Anggraini, R. N. (2023). Kombinasi Ekstraksi Kata Kunci dan Ekspansi Kueri Untuk Deteksi Isu Etik pada Ringkasan Penelitian Kesehatan. Techno. com, 22(1).

Huang, B. W. (2024). Generative large language models augmented hybrid retrieval system for biomedical question answering. *CLEF Working Notes*.

Husin, N. (2023). Komparasi Algoritma Random Forest, Naïve Bayes, dan Bert Untuk Multi-Class Classification Pada Artikel Cable News Network (CNN). J. Esensi Infokom J. Esensi Sist. Inf.

Jackins, V., Vimal, S., Kaliappan, M., & Lee, M. Y. (2021). AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. The Journal of Supercomputing, 77(5), 5198-5219.

Kamel, H., Abdulah, D., & Al-Tuwaijari, J. M. (2019, June). Cancer classification using gaussian naive bayes algorithm. In 2019 international engineering conference (IEC) (pp. 165-170). IEEE.

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. Multimedia tools and applications, 82(3), 3713-3744.

Kulsum, U., Jajuli, M., & Sulistiyowati, N. (2022). Analisis Sentimen Aplikasi WETV di Google Play Store Menggunakan Algoritma Support Vector Machine. Journal of Applied Informatics and Computing, 6(2), 205-212.

Kumar, S. (2024). BERT - Understanding the Basics. Diakses pada 7 Oktober 2024, dari https://sushant-kumar.com/blog/bert/.

Kurniawan, M. H., Handiyani, H., Nuraini, T., & Hariyati, R. T. S. (2023). Artificial Intelligence (AI) in Nursing Services: A Literature Review. Faletehan Health Journal, 10(01), 77-84.

Kusumaningrum, R., Hanifah, A. F., Khadijah, K., Endah, S. N., & Sasongko, P. S. (2023). Long short-term memory for non-factoid answer selection in Indonesian question answering system for health information. International Journal of Advanced Computer Science and Applications, 14(2).

Lavin, M. (2019). Analyzing documents with TF-IDF.

Lee, M. C., Zhu, Q., Mavromatis, C., Han, Z., Adeshina, S., Ioannidis, V. N., ... & Faloutsos, C. (2024). HybGRAG: Hybrid Retrieval-Augmented Generation on Textual and Relational Knowledge Bases. *arXiv preprint arXiv:2412.16311*.

Lubis, A. T. U. B., Harahap, N. S., Agustian, S., Irsyad, M., & Afrianty, I. (2024). Question Answering System pada Chatbot Telegram Menggunakan Large Language Models (LLM) dan Langchain (Studi Kasus UU Kesehatan): Question Answering System on Telegram Chatbot Using Large Language Models (LLM) and Langchain (Case Study: Health Law). MALCOM: Indonesian Journal of Machine Learning and Computer Science, 4(3), 955-964.

Ma'rifah, H., Wibawa, A. P., & Akbar, M. I. (2020). Klasifikasi artikel ilmiah dengan berbagai skenario preprocessing. Ekonomi Bisnis, 29, 23-01.

Mubarok, F., Harliana, H., & Hadijah, I. (2019). Perbandingan Antara Metode RUP dan Prototype Dalam Aplikasi Penerimaan Siswa Baru Berbasis Web. *Creative Information Technology Journal*, 2(2), 114-127.

Musfiroh, N., Hamdani, H., & Indah Fitri, A. (2013). Penerapan Algoritma Term Frequency-Inverse Document Frequency (Tf-Idf) Untuk Text Mining. *e-journals, Jurnal Ilmiah Ilmu Komputer*, *8*(3).

Nurwanda, N., Suarna, N., & Prihartono, W. (2024). Penerapan Nlp (Natural Language Processing) Dalam Analisis Sentimen Pengguna Telegram Di Playstore. Jati (*Jurnal Mahasiswa Teknik Informatika*), 8(2), 1841-1846.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research,* 12, 2825-2830. https://scikit learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Sawarkar, K., Mangal, A., & Solanki, S. R. (2024, August). Blended rag: Improving rag (retriever-augmented generation) accuracy with semantic search and hybrid query-based retrievers. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)* (pp. 155-161). IEEE.

Subhi, Y. A. (2025). Klasifikasi Sentimen Menggunakan Metode Passive Aggressive Dengan Menggunakan Model Bahasa Bert Pada Dataset

Kecil. Klasifikasi Sentimen Menggunakan Metode Passive Aggressive Dengan Menggunakan Model Bahasa Bert Pada Dataset Kecil, 6(3), 1838-1847.

Syah, M. I., Harahap, N. S., & Sanjaya, S. (2024). Penerapan Retrieval Augemented Generation Menggunakan Langchain Dalam Pengembangan Sistem Tanya Jawab Hadis Berbasis Web. Zonasi: *Jurnal Sistem Informasi*, 6(2), 370-379.

Pratiwi, P., Dwifa, D., Desiani, A., Amran, A., & Suprihatin, B. (2024). Klasifikasi Penyakit Kanker Paru-Paru Menggunakan Algoritma Naïve Bayes dan Iterative Dichotomizer 3 (ID3). Electr. J. Rekayasa dan Teknol. Elektro, 18(1), 69-80.

Polikar, R. (2016). A survey of ensemble learning. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 32(1), 1-15. https://doi.org/10.1109/TSMCC.2005.858101

Prawira, A., Chrisnanto, Y. H., & Ningsih, A. K. (2024). Deteksi Kecemasan Di Twitter Menggunakan Fitur Word Embedding Bert Dan Metode Bidirectional-Long Short Term Memory. Jati (*Jurnal Mahasiswa Teknik Informatika*), 8(4), 7795-7800.

Rahardjo, A. (2024). Pengertian RUP (Rational Unified Process). Medium. Diakses pada 22 September 2024. https://medium.com/@andrerahardjo/pengertian-rup-rational-unified-process-1bec9c664458

Ren, F., & Zhou, Y. (2020). Cgmvqa: A new classification and generative model for medical visual question answering. *IEEE Access*, *8*, 50626-50636.

Rosa, Salahuddin, M., 2019, Rekayasa Perangkat Lunak Terstruktur dan Berorientasi Object, Informatika, Bandung

Rozi, F. N., & Sulistyawati, D. H. (2019). Klasifikasi Berita Hoax Pilpres Menggunakan Metode Modified K-Nearest Neighbor Dan Pembobotan Menggunakan Tf-Idf. Konvergensi, 15(1), 1-10.

Sasmita, R. A (2018). Pemanfaatan Algoritma TF/IDF Untuk Sistem Informasi E-Complaint Handling.

Sjoraida, D. F., Guna, B. W. K., & Yudhakusuma, D. (2024). Analisis Sentimen Film Dirty Vote Menggunakan BERT (Bidirectional Encoder Representations from Transformers). Jurnal JTIK (*Jurnal Teknologi Informasi dan Komunikasi*), 8(2), 393-404.

Taslim, T., Handayani, S., & Fajrizal, F. (2023). Kinerja Komparatif Optimasi Algoritma Naive Bayes dalam Klasifikasi Teks untuk Uji Klinis Kanker. *Jurnal Eksplora Informatika*, 13(1), 113-123.

Weng, L. (2020). *How to build an open-domain question answering system?* Diakses pada 7 Oktober 2024, dari https://lilianweng.github.io/posts/2020-10-29-odqa/

Wolpert, D. H. (1992). Stacking: Combining classifiers. *Proceedings of the Third International Conference on Neural Networks*, 241-248.

Xie, Q. (2024). MeDAL: Medical Abbreviation Disambiguation Dataset [Dataset]. Kaggle. Diakses tanggal 30 Juni 2024 from https://www.kaggle.com/datasets/xhlulu/medal-emnlp

Yin, J., Jiang, X., Lu, Z., Shang, L., Li, H., & Li, X. (2015). Neural generative question answering. *arXiv preprint arXiv:1512.01337*.

Zhou, Z.-H. (2012). Ensemble learning: Foundations and algorithms. Springer.

Zulkarnain, M. A., Raharjo, M. F., & Olivya, M. (2020). Perancangan Aplikasi Chatbot Sebagai Media E-Learning Bagi Siswa. Elektron: *Jurnal Ilmiah,* 88-95.