

**RANCANG BANGUN SEARCH ENGINE PADA PT XYZ
MENGUNAKAN ALGORITMA BERT-RAG DAN GPT-4**



UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

Skripsi

Virginia Lim
00000055667

**PROGRAM STUDI SISTEM INFORMASI
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA
TANGERANG
2024**

**RANCANG BANGUN SEARCH ENGINE PADA PT XYZ
MENGUNAKAN ALGORITMA BERT-RAG DAN GPT-4**



Diajukan sebagai Salah Satu Syarat untuk Memperoleh

Gelar Sarjana Komputer (S.Kom)

Virginia Lim

00000055667

PROGRAM STUDI SISTEM INFORMASI

FAKULTAS TEKNIK DAN INFORMATIKA

UNIVERSITAS MULTIMEDIA NUSANTARA

TANGERANG

2024

HALAMAN PERNYATAAN TIDAK PLAGIAT

Dengan ini saya,

Nama : Virginia Lim
Nomor Induk Mahasiswa : 00000055667
Program studi : Sistem Informasi

Skripsi dengan judul:

Rancang Bangun *Search Engine* pada PT XYZ Menggunakan Algoritma BERT-RAG dan GPT-4

merupakan hasil karya saya sendiri bukan plagiat dari karya ilmiah yang ditulis oleh orang lain, dan semua sumber, baik yang dikutip maupun dirujuk, telah saya nyatakan dengan benar serta dicantumkan di Daftar Pustaka.

Jika di kemudian hari terbukti ditemukan kecurangan/penyimpangan, baik dalam pelaksanaan skripsi maupun dalam penulisan laporan skripsi, saya bersedia menerima konsekuensi dinyatakan TIDAK LULUS untuk Tugas Akhir yang telah saya tempuh.

Tangerang, 16 Desember 2024



Virginia Lim

UMMN

UNIVERSITAS
MULTIMEDIA
NUSANTARA

HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH

Yang bertanda tangan di bawah ini:

Nama : Virginia Lim

NIM : 00000055667

Program Studi : Sistem Informasi

Jenjang : S1

Judul Karya Ilmiah : Rancang Bangun *Search Engine* pada PT XYZ Menggunakan Algoritma BERT-RAG dan GPT-4

Menyatakan dengan sesungguhnya bahwa saya bersedia* (**pilih salah satu**):

- Saya bersedia memberikan izin sepenuhnya kepada Universitas Multimedia Nusantara untuk mempublikasikan hasil karya ilmiah saya ke dalam repositori Knowledge Center sehingga dapat diakses oleh Sivitas Akademika UMN/Publik. Saya menyatakan bahwa karya ilmiah yang saya buat tidak mengandung data yang bersifat konfidensial.
- Saya tidak bersedia mempublikasikan hasil karya ilmiah ini ke dalam repositori Knowledge Center, dikarenakan: dalam proses pengajuan publikasi ke jurnal/konferensi nasional/internasional (dibuktikan dengan *letter of acceptance*) **.
- Lainnya, pilih salah satu:
 - Hanya dapat diakses secara internal Universitas Multimedia Nusantara
 - Embargo publikasi karya ilmiah dalam kurun waktu 3 tahun.

Tangerang, 7 Januari 2024


(Virginia Lim)

* Pilih salah satu

** Jika tidak bisa membuktikan LoA jurnal/HKI, saya bersedia mengizinkan penuh karya ilmiah saya untuk dipublikasikan ke KC UMN dan menjadi hak institusi UMN.

KATA PENGANTAR


Puji dan Syukur saya panjatkan kepada Tuhan Yang Maha Esa atas segala rahmat-Nya, yang memungkinkan saya menyelesaikan penulisan skripsi ini dengan judul Rancang Bangun Search Engine pada PT XYZ menggunakan Algoritma BERT-RAG dan GPT-4. Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana di Program Studi Sistem Informasi, Fakultas Teknik dan Informatika, Universitas Multimedia Nusantara. Saya menyadari bahwa pencapaian ini tidak lepas dari bimbingan dan dukungan berbagai pihak. Oleh karena itu, saya mengucapkan terima kasih kepada:

1. Bapak Dr. Andrey Andoko, M.Sc, selaku Rektor Universitas Multimedia Nusantara.
2. Bapak Dr. Eng. Niki Prastomo, selaku Dekan Fakultas Teknik Informatika Universitas Multimedia Nusantara.
3. Ibu Ririn Ikana Desanti, S.Kom., M.Kom., selaku Ketua Program Studi Sistem Informasi Universitas Multimedia Nusantara.
4. Ibu Dinar Ajeng Kristiyanti, S.Kom., M.Kom., sebagai Pembimbing yang telah memberikan bimbingan, arahan, dan motivasi atas terselesainya tugas akhir ini.
5. Keluarga saya yang telah memberikan bantuan dukungan material dan moral, sehingga penulis dapat menyelesaikan tugas akhir ini.

Semoga karya ilmiah ini bisa menjadi referensi penelitian yang baik untuk berbagai penelitian di kemudian hari.

Tangerang, 16 Desember 2024

UMMN


Virginia Lim

UNIVERSITAS
MULTIMEDIA
NUSANTARA

Rancang Bangun *Search Engine* pada PT XYZ menggunakan Algoritma BERT-RAG dan GPT-4

Virginia Lim

ABSTRAK

Industri 4.0 telah membawa transformasi besar dalam pengelolaan informasi, dengan tantangan utama berupa peningkatan volume data, termasuk dokumen tidak terstruktur seperti PDF. PT. XYZ menghadapi kesulitan dalam memastikan informasi yang dapat diakses secara cepat, akurat, dan relevan oleh karyawan. Metode pencarian tradisional seperti dokumen dalam bentuk kertas sering kali menghasilkan waktu pencarian yang lama dan hasil yang kurang sesuai, sehingga memengaruhi efisiensi kerja serta pengambilan keputusan. Penelitian ini bertujuan merancang *search engine* berbasis *web* menggunakan *hybrid* model, yaitu *Bidirectional Encoder Representations from Transformers* (BERT) dan *Retrieval-Augmented Generation* (RAG), serta *Application Program Interface* (API) GPT-4 untuk menghasilkan pencarian yang relevan, akurat, dan kontekstual.

Data yang kompleks memerlukan pendekatan terstruktur untuk pengelolaan dan analisis yang efektif. Penelitian ini menggunakan *framework data mining Cross-Industry Standard Process for Data Mining* (CRISP-DM), dimulai dari memahami proses bisnis perusahaan, pemilihan data, *data preparation* dengan tahapan ekstraksi hingga *encoding* PDF. Selanjutnya *data modelling* menggunakan *hybrid* model BERT-RAG dan GPT-4, evaluasi model, dan *deployment*. Pada tahap *deployment* menggunakan metode *prototype* untuk membangun antarmuka *web*, memungkinkan partisipasi aktif pengguna dalam proses pengembangan. Kombinasi CRISP-DM dan *prototyping* memastikan sistem *search engine* yang inovatif, relevan, dan *user-friendly*.

Hasil dari penelitian ini berupa *search engine* berbasis *website* yang dapat meningkatkan kemampuan pencarian data. Hal ini ditandai dengan akurasi model mencapai 86% meningkat 4% dari penelitian sebelumnya dalam kesamaan antara *keyword* yang dimasukkan dengan hasil yang ditampilkan. Hasil akurasi yang didapatkan dihitung dengan menggunakan *cosine similiraty*. Pengujian *website* juga dilakukan menggunakan *User Acceptance Testing* (UAT) dengan evaluasi berhasil.

Kata kunci: API GPT-4, BERT, *Retrieval-Augmented Generation*, *Search Engine*.

UNIVERSITAS
MULTIMEDIA
NUSANTARA

Design and Development of a Search Engine at PT XYZ using BERT-RAG

Algorithm and GPT-4

Virginia Lim

ABSTRACT

Industry 4.0 has brought significant transformations in information management, with a primary challenge being the increasing volume of data, including unstructured documents such as PDFs. PT. XYZ faces difficulties in ensuring information can be accessed quickly, accurately, and relevantly by employees. Traditional search methods, such as paper-based documents, often result in prolonged search times and less relevant results, impacting work efficiency and decision-making. This research aims to design a web-based search engine using a hybrid model, namely Bidirectional Encoder Representations from Transformers (BERT) and Retrieval-Augmented Generation (RAG), along with the GPT-4 Application Programming Interface (API) to deliver relevant, accurate, and contextual searches.

Complex data requires a structured approach for effective management and analysis. This research employs the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, starting with understanding the company's business processes, data selection, and data preparation involving extraction and encoding of PDFs. Subsequently, the data modeling phase uses the hybrid BERT-RAG and GPT-4 model, followed by model evaluation and deployment. The deployment phase applies the prototype method to build a web interface, enabling active user participation in the development process. The combination of CRISP-DM and prototyping ensures an innovative, relevant, and user-friendly search engine system.

The results of this research include a website-based search engine capable of enhancing data search capabilities. This is evidenced by a model accuracy rate of 86%, an improvement of 4% over previous studies, in terms of the similarity between input keywords and displayed results. The accuracy obtained was calculated using cosine similarity. Website testing was also conducted through User Acceptance Testing (UAT), with successful evaluations.

Keywords: API GPT-4, BERT, Retrieval-Augmented Generation, Search Engine.

U N I V E R S I T A S
M U L T I M E D I A
N U S A N T A R A

DAFTAR ISI

HALAMAN PERNYATAAN TIDAK PLAGIAT	ii
HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH	iii
KATA PENGANTAR.....	iv
ABSTRAK	v
ABSTRACT	vi
DAFTAR ISI.....	vii
DAFTAR TABEL	ix
DAFTAR GAMBAR.....	x
DAFTAR RUMUS	xii
DAFTAR LAMPIRAN.....	xiii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	4
1.4 Tujuan dan Manfaat Penelitian.....	5
1.4.1 Tujuan Penelitian.....	5
1.4.2 Manfaat Penelitian.....	5
1.5 Sistematika Penulisan.....	6
BAB II LANDASAN TEORI.....	7
2.1 Penelitian Terdahulu.....	7
2.2 Teori tentang Topik Skripsi.....	11
2.2.1 <i>Database</i>	11
2.2.2 <i>Kecerdasan Buatan (Artificial Intelligence)</i>	11
2.2.3 <i>Basis data terdistribusi</i>	12
2.2.4 <i>Natural Language Processing (NLP)</i>	13
2.2.5 <i>Machine Learning</i>	13
2.2.6 <i>Deep Learning</i>	15
2.2.7 <i>Software Development Life Cycle</i>	15
2.2.8 <i>User Acceptance Testing</i>	15
2.3 Framework dan Algoritma	16
2.3.1 <i>Cross-Industry Standard Process for Data Mining (CRISP-DM)</i>	16
2.3.2 <i>Prototype</i>	20

2.3.3	<i>Unified Modeling Language (UML)</i>	22
2.3.4	<i>Birectional Encoder Representations from Transformers (BERT)</i> ..	25
2.3.5	<i>Retrieval-Augmentasi Generation (RAG)</i>	26
2.4	<i>Tools Penelitian</i>	28
BAB III METODOLOGI PENELITIAN		32
3.1	Gambaran Umum Objek Penelitian	32
3.2	Metode Penelitian	33
3.1	Alur Penelitian	33
3.2	Metodologi <i>Data Mining</i>	35
3.3	Metode Pengembangan Sistem	37
3.4	<i>Data Modelling</i>	39
3.3	Teknik Pengumpulan Data	40
3.3.1	Populasi dan Sampel	42
3.3.2	Periode Pengambilan Data	42
3.4	Teknik Analisis Data	43
BAB IV ANALISIS DAN HASIL PENELITIAN		46
BAB V SIMPULAN DAN SARAN		92
5.1	Simpulan	92
5.2	Saran.....	93
DAFTAR PUSTAKA		95
LAMPIRAN		99



U N I V E R S I T A S
M U L T I M E D I A
N U S A N T A R A

DAFTAR TABEL

Tabel 2. 1 Penelitian Terdahulu	7
Tabel 2. 2 Simbol Use Case Diagram	23
Tabel 2. 3 Simbol Class Diagram	23
Tabel 2. 4 Simbol Activity Diagram	24
Tabel 3. 1 Perbandingan Metodologi Data Mining.....	35
Tabel 3. 2 Perbandingan Metode Waterfall, Prototyping dan RAD	37
Tabel 3. 3 Tabel perbandingan metode problem solving	39
Tabel 3. 4 Perbandingan code editor	43
Tabel 3. 5 Perbandingan framework python.....	44
Tabel 3. 6 Perbandingan database system.....	45
Tabel 4. 1 Tabel kebutuhan fungsionalitas	62
Tabel 4. 2 Tabel kebutuhan non-fungsionalitas	63
Tabel 4. 3 Spesifikasi table admin	71
Tabel 4. 4 Tabel spesifikasi files.....	71
Tabel 4. 5 Hasil UAT user 1	82
Tabel 4. 6 Hasil UAT user 2	84
Tabel 4. 7 Hasil UAT User 3	85
Tabel 4. 8 Tabel hasil akurasi dokumen	88
Tabel 4. 9 Tabel klasifikasi perbandingan penelitian terdahulu	89

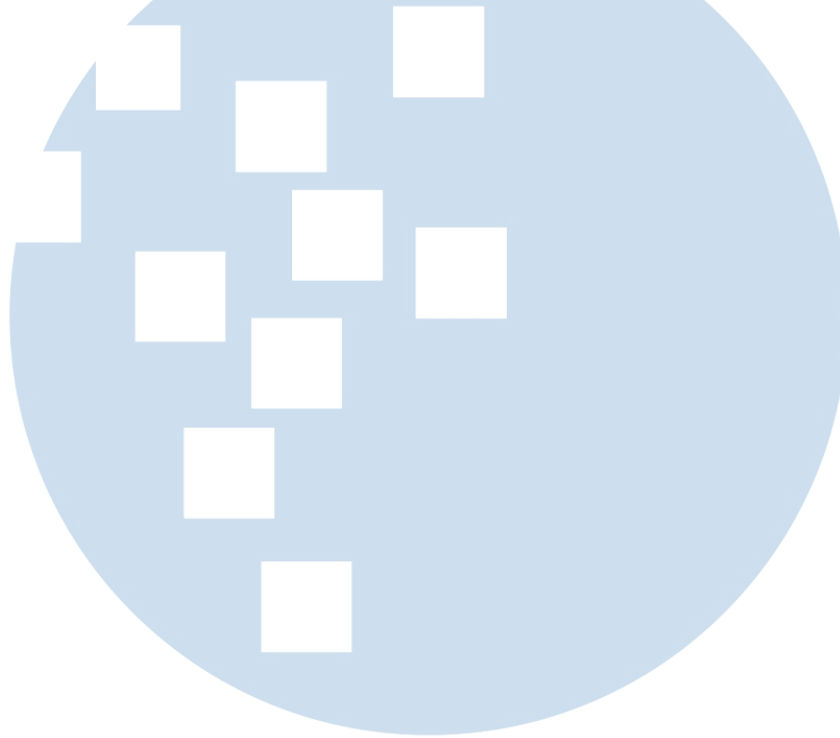
UMMN

UNIVERSITAS
MULTIMEDIA
NUSANTARA

DAFTAR GAMBAR

Gambar 2. 1 Overview RAG.....	27
Gambar 2. 2 RAG with/without Datastore Update	28
Gambar 3. 2 Alur Penelitian	34
Gambar 3. 3 Alur Pengumpulan Data.....	41
Gambar 4. 1 Proses Bisnis Penyimpanan dan Pencarian Dokumen secara Konvensional	46
Gambar 4. 2 Contoh PDF yang dikumpulkan.....	48
Gambar 4. 3 Hasil Eksplorasi Data.....	48
Gambar 4. 4 Kategori Data	49
Gambar 4. 5 Library yang digunakan	49
Gambar 4. 6 Code ekstraksi PDF.....	50
Gambar 4. 7 Hasil Ekstrasi PDF	51
Gambar 4. 8 Tokenize BERT.....	51
Gambar 4. 9 Fungsi Encode text using BERT	52
Gambar 4. 10 Fungsi untuk encoding PDF.....	53
Gambar 4. 11 Fungsi Encode text menggunakan BERT	53
Gambar 4. 12 Hasil Encoding File PDF	54
Gambar 4. 13 Library Transformation.....	55
Gambar 4. 14 Fungsi untuk merging data.....	55
Gambar 4. 15 Fungsi Perhitungan skor dokumen.....	56
Gambar 4. 16 Fungsi Top-k Selection	56
Gambar 4. 17 Generate response menggunakan GPT-4.....	57
Gambar 4. 18 Return respon dari GPT-4	58
Gambar 4. 19 Hasil Data Modelling	58
Gambar 4. 20 Fungsi untuk cosine similarity	59
Gambar 4. 21 Code untuk Longest Common Subsequence.....	60
Gambar 4. 22 Use case diagram.....	64
Gambar 4. 23 Activity diagram login admin	66
Gambar 4. 24 Activity diagram melakukan generate dokumen.....	67
Gambar 4. 25 Activity diagram melakukan upload data.....	68
Gambar 4. 26 Activity diagram hasil	69
Gambar 4. 27 Rancangan class diagram	70
Gambar 4. 28 Tabel Relasi.....	72
Gambar 4. 29 Wireframe halaman search user	73
Gambar 4. 30 Wireframe halaman search admin.....	73
Gambar 4. 31 Wireframe halaman login.....	74
Gambar 4. 32 Wireframe halaman Upload	75
Gambar 4. 33 Wireframe halaman hasil	75
Gambar 4. 34 Flow mockups	76
Gambar 4. 35 Mockups halaman home user.....	77
Gambar 4. 36 Mockups halaman home admin	77
Gambar 4. 37 Mockups halaman upload	78
Gambar 4. 38 Mockups halaman hasil user	78
Gambar 4. 39 Mockups halaman hasil user admin	79
Gambar 4. 40 Tampilan halaman home user	79
Gambar 4. 41 Tampilan halaman home user	80

Gambar 4. 42 Tampilan halaman login.....	80
Gambar 4. 43 Tampilan halaman hasil	81
Gambar 4. 44 Tampilan halaman upload	82
Gambar 4. 45 Tampilan baru salah satu halaman	87

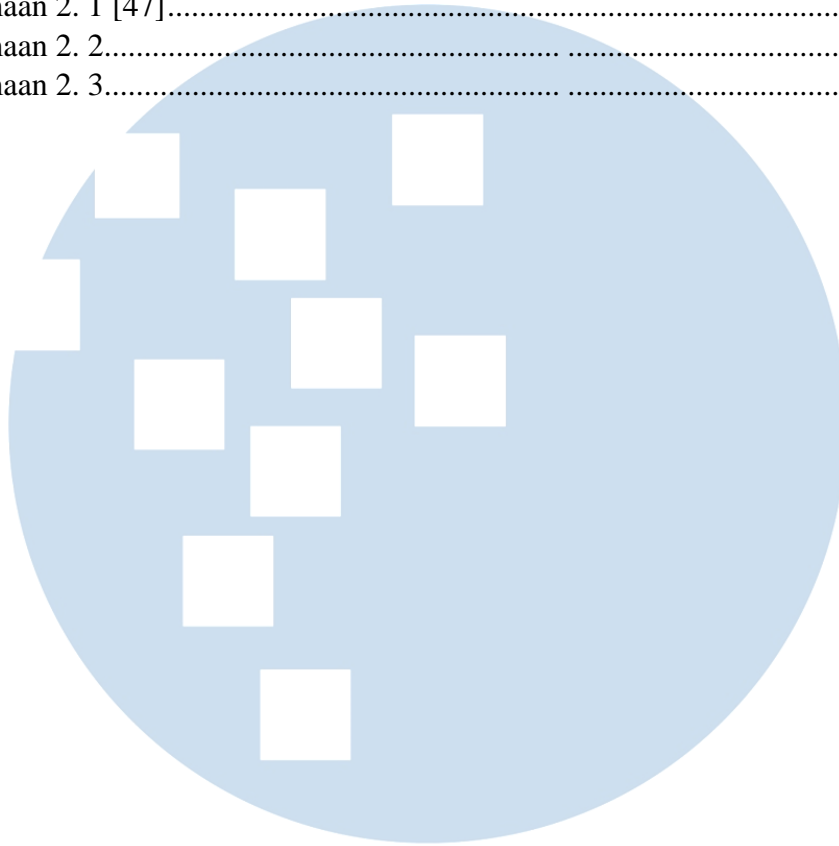


UMMN

UNIVERSITAS
MULTIMEDIA
NUSANTARA

DAFTAR RUMUS

Persamaan 2. 1 [47]..... 19
Persamaan 2. 2..... 20
Persamaan 2. 3..... 20

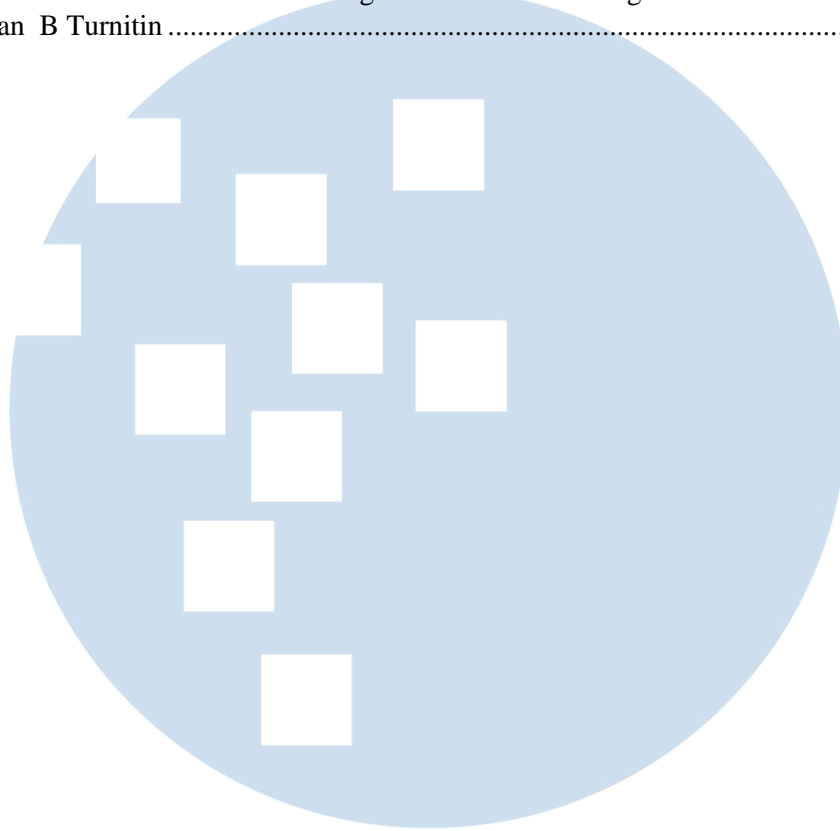


UMMN

UNIVERSITAS
MULTIMEDIA
NUSANTARA

DAFTAR LAMPIRAN

Lampiran A Form Konsultasi Bimbingan Dosen Pembimbing 99
Lampiran B Turnitin 99



UMMN

UNIVERSITAS
MULTIMEDIA
NUSANTARA