

**IMPLEMENTASI Pencarian Semantik dalam
Tafsir Al-Quran dengan Algoritma
Cosine Similarity dan Large
Language Models**

SKRIPSI

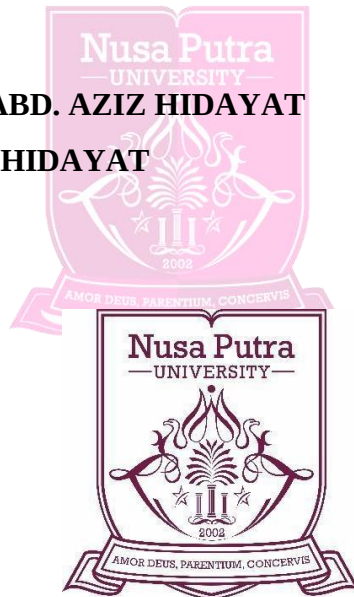
*Diajukan Untuk Memenuhi Salah Satu Syarat Dalam Menempuh
Gelar Sarjana Teknik Informatika*

MOH. ABD. AZIZ HIDAYAT

20220040014

TOFIK HIDAYAT

20200040056



**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNIK, KOMPUTER DAN DESAIN
UNIVERSITAS NUSA PUTRA
SUKABUMI
JUNI 2024**

**IMPLEMENTASI PENCARIAN SEMANTIK DALAM
TAFSIR AL-QURAN DENGAN ALGORITMA
COSINE SIMILARITY DAN LARGE
LANGUAGE MODELS**

SKRIPSI

*Diajukan Untuk Memenuhi Salah Satu Syarat Dalam Menempuh
Gelar Sarjana Teknik Informatika*

MOH. ABD. AZIZ HIDAYAT

20220040014

TOFIK HIDAYAT

20200040056



**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNIK, KOMPUTER DAN DESAIN
UNIVERSITAS NUSA PUTRA
SUKABUMI
JUNI 2024**

PERNYATAAN PENULIS

JUDUL : IMPLEMENTASI PENCARIAN SEMANTIK DALAM TAFSIR
AL-QURAN DENGAN ALGORITMA *COSINE SIMILARITY*
DAN *LARGE LANGUAGE MODELS*

NAMA : MOH. ABD. AZIZ HIDAYAT

NIM : 20220040014

NAMA : TOFIK HIDAYAT

NIM : 20200040056

“Saya menyatakan dan bertanggungjawab dengan sebenarnya bahwa Skripsi ini adalah hasil karya saya sendiri kecuali cuplikan dan ringkasan yang masing-masing telah saya jelaskan sumbernya. Jika pada waktu selanjutnya ada pihak lain yang mengklaim bahwa Skripsi ini sebagai karyanya, yang disertai dengan bukti-bukti yang cukup, maka saya bersedia untuk dibatalkan gelar Sarjana Komputer/Sarjana Teknik saya beserta segala hak dan kewajiban yang melekat pada gelar tersebut”

Sukabumi, 20 Juni 2024



Moh. Abd. Aziz Hidayat

Penulis 1



Tofik Hidayat

Penulis 2

PERSETUJUAN SKRIPSI

JUDUL : IMPLEMENTASI Pencarian Semantik dalam Tafsir
AL-QURAN dengan Algoritma *Cosine Similarity*
dan *Large Language Models*

NAMA : MOH. ABD. AZIZ HIDAYAT

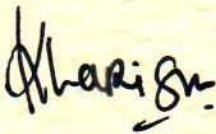
NIM : 20220040014

NAMA : TOFIK HIDAYAT

NIM : 20200040056

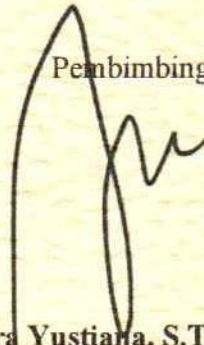
Skripsi ini telah di periksa dan di setujui
Sukabumi, 07 Juni 2024

Pembimbing 1



Ivana Lucia Kharisma, M.Kom
NIDN : 0429038002

Pembimbing 2



Indra Yustiana, S.T., M. Kom
NIDN. 0409017604

Ketua Program Studi Teknik Informatika



Ir. Somantri, S.T., M.Kom
NIDN.0419128801

PENGESAHAN SKRIPSI

JUDUL : IMPLEMENTASI PENCARIAN SEMANTIK DALAM TAFSIR
AL-QURAN DENGAN ALGORITMA *COSINE SIMILARITY*
DAN *LARGE LANGUAGE MODELS*

NAMA : MOH. ABD. AZIZ HIDAYAT

NIM : 20220040014

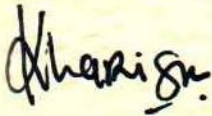
NAMA : TOFIK HIDAYAT

NIM : 20200040056

Skripsi ini telah diujikan dan dipertahankan di depan Dewan Penguji pada Sidang Skripsi tanggal 20 Juni 2024 Menurut pandangan kami, Skripsi ini memadai dari segi kualitas untuk tujuan penganugerahan gelar Sarjana Komputer (S.Kom).

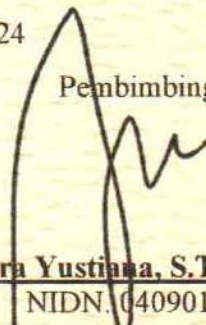
Sukabumi, 20 Juni 2024

Pembimbing 1



Ivana Lucia Kharisma, M.Kom
NIDN : 0429038002

Pembimbing 2



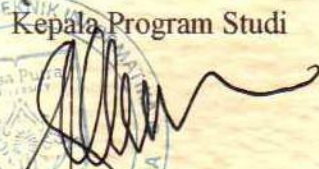
Indra Yustiana, S.T., M. Kom
NIDN. 0409017604

Ketua Penguji



Alun Sujjada, S.Kom., M.T
NIDN.0718108001

Kepala Program Studi



Ir. Somantri, S.T., M.Kom
NIDN.0419128801

Plh. Dekan Fakultas Teknik, Komputer dan Desain

Ir. Paikun, S.T., M.T., IPM., ASEAN Eng
NIDN. 0402037401

HALAMAN PERSEMBAHAN

*Alhamdulillah Rabbil Alamin, puji dan syukur saya panjatkan kepada Allah SWT.
Terimakasih atas karunia-Mu yang telah membantu saya menyelesaikan skripsi
ini dengan lancar.*

*Saya mepersembahkan skripsi ini kepada orang tua tercinta saya, Ibu dan Ayah.
Saya ingin mengucapkan terima kasih kepada mereka atas doa, dorongan,
inspirasi, pengorbanan, nasihat, dan kasih sayang yang tidak pernah berhenti
sampai saat ini.*

*Selain itu, saya berterima kasih kepada dosen pembimbing yang telah
memberikan bimbingan dan petunjuk selama proses penyusunan skripsi ini.*

*Tidak lupa untuk berterima kasih kepada teman-teman dan rekan kerja yang telah
mendukung dalam berbagai cara. Mudah-mudahan skripsi ini akan membantu
kemajuan ilmu pengetahuan dan memberikan kontribusi positif bagi masyarakat.*

*Saya ingin mengucapkan terima kasih kepada semua orang yang telah
memberikan dukungan dan inspirasi dalam perjuangan ini,*

*Dan terakhir, saya ingin mengucapkan terima kasih kepada diri saya sendiri yang
telah melakukan upaya dan perjuangan selama ini. Saya berterima kasih atas
upaya yang telah dilakukan. Ini bukan akhir dari segalanya, ada tantangan lain
yang menanti untuk diatasi. Mari kita berdoa, berusaha, dan tidak pernah
menyerah.*

ABSTRACT

The search for relevant and accurate information is a crucial aspect of understanding the interpretation of the Quran. Implementing semantic search can help users find interpretations that align with their context and needs. The cosine similarity algorithm can be used to measure the similarity between a query and interpretation data. After using the cosine similarity algorithm, Large Language Models (LLMs) can be employed to deeply understand the meaning of queries and interpretations, resulting in more relevant and accurate searches. Therefore, this research aims to implement cosine similarity and LLMs for semantic search. The proposed semantic search system is built following the ML-SDLC method, which consists of the stages: planning, data collection, data preprocessing, model development, evaluation, and deployment. Users can perform searches by entering inputs into the search engine, the input will be converted into vectors, then the system will search the Quran interpretation dataset that has been converted into vectors and stored in a vector database. Cosine similarity is used to measure the semantic relevance between the query representation and the vector representation of the Quran interpretations. The GPT-4 model will summarize the semantic search results and present them to the user. Based on the model evaluation results using F1-Score vs. Threshold and ROC-AUC Score, the F1-Score yielded a value of 0.819 with a threshold of 0.410, while the ROC-AUC Score result was 0.587. The model testing results showed an accuracy of 96%, and for relevance testing, the results showed an accuracy of 89%.

Keywords: Semantic Search, Cosine Similarity, LLM, Quran Interpretation, Vector

ABSTRAK

Pencarian informasi yang relevan dan akurat merupakan aspek penting dalam memahami tafsir Al-Quran. Implementasi pencarian semantik dapat membantu pengguna menemukan tafsir yang sesuai dengan konteks dan kebutuhan. Algoritma *cosine similarity* dapat di gunakan untuk mengukur kesamaan antara query dan data tafsir. Kemudian setelah penggunaan algoritma *cosine similarity*, *Large Language Models* (LLM) dapat di gunakan untuk memahami makna query dan tafsir secara mendalam, sehingga menghasilkan pencarian yang lebih relevan dan akurat. Oleh karena itu, pada penelitian ini bertujuan untuk menerapkan algoritma *cosine similarity* dan LLM untuk pencarian semantik. Sistem pencarian semantik yang di usulkan di bangun dengan mengikuti metode ML-SDLC yang terdiri dari tahapan : perencanaan, pengumpulan data, *preprocessing* data, pengembangan model, evaluasi dan deployment. Pengguna dapat melakukan pencarian dengan memasukan inputan pada mesin pencarian, hasil inputan akan di jadikan vektor kemudian sistem akan melakukan pencarian ke Dataset tafsir Al-Quran yang sudah di ubah ke dalam bentuk vektor dan di simpan dalam database vektor kemudian *cosine similarity* di gunakan untuk mengukur relevansi semantik antara representasi query dan representasi vektor dari tafsir Al-Quran, Model GPT-4 akan merangkum hasil pencarian semantik dan di tampilkan ke pengguna. Berdasarkan hasil evaluasi model dengan menggunakan *F1-Score Vs Threshold* dan *ROC-AUC Score*. *F1-Score* memberikan nilai 0,819 dan untuk *Threshold* 0,410 semntara itu hasil dari *ROC-AUC Score* 0,587. Berdasarkan hasil pengujian model menunjukan akurasi 96% dan untuk pengujian relevansi hasil menunjukan akurasi 89%.

Kata kunci : Pencarian Semantik, Cosine Similarity, LLM, Tafsir Al-Quran, Vektor

KATA PENGANTAR

Alhamdulillah, puji dan syukur penulis panjatkan kehadirat Allah SWT atas segala rahmat dan karunia-Nya yang telah memberikan kekuatan serta kesabaran bagi penulis untuk menyelesaikan skripsi ini. Skripsi ini disusun sebagai salah satu syarat untuk kelulusan dan mendapatkan gelar Sarjana Komputer (S.Kom) pada Program Studi Teknik Informatika Universitas Nusa Putra Sukabumi.

Pada skripsi ini penulis mengambil judul **“Implementasi Pencarian Semantik Dalam Tafsir Al-Quran Dengan Algoritma Cosine Similarity Dan Large Language Models”**. Penelitian ini bertujuan untuk menggali potensi dan efektivitas penggunaan algoritma *cosine similarity* dan *large language models* dalam pencarian semantik dalam tafsir Al-Quran, sehingga dapat mempermudah untuk melakukan pencarian tafsir Al-Quran dan memberikan pemahaman yang lebih mendalam terhadap makna-makna dalam Al-Quran.

Dalam penyusunan skripsi ini, penulis pun menemui banyak kesulitan yang disebabkan keterbatasan pengetahuan. Namun, berkat pencerahan dari Allah Yang Maha kuasa dan kemurahan hati dari setiap orang yang memberikan dukungan serta bimbingan dari berbagai sudut pemikiran, penyusunan skripsi ini dapat terselesaikan. Oleh karena itu, penulis mengucapkan terima kasih yang sedalam - dalamnya kepada semua pihak yang memberikan dukungan dalam berbagai bentuk.

Sehubungan dengan itu penulis menyampaikan penghargaan dan ucapan terima kasih yang sebesar-besarnya kepada :

1. Dr. Kurniawan, S.T., M.Si., M.M. Selaku rektor Universitas Nusa Putra Sukabumi beserta jajarannya yang telah memberikan kesempatan kepada penulis untuk menuntut ilmu di Universitas Nusa Putra.
2. Wakil Rektor I Bidang Akademik Universitas Nusa Putra Sukabumi Bapak Anggu Pradiftha Junfithrana, S.Pd., Mt beserta jajarannya yang telah memberikan kesempatan kepada penulis untuk menuntut ilmu di Universitas Nusa Putra.
3. Kepala Program Studi Teknik Informatika Bapak Somantri, S.T, M.Kom yang telah memberikan dukungan, dan pengarahan demi kelancaran penelitian ini.

4. Dosen Pembimbing I Ibu Ivana Lucia Kharisma, M.Kom yang telah menyediakan waktu, tenaga, dan pikiran untuk memberikan panduan, masukan, dan dorongan dalam penyusunan skripsi ini.
5. Dosen Pembimbing II Bapak Indra Yustiana, ST., M. Kom yang telah menyediakan waktu, tenaga, dan pikiran untuk memberikan panduan, masukan, dan dorongan dalam penyusunan skripsi ini.
6. Dosen Penguji Bapak Allun Sujjada, M.Kom Yang telah memberikan waktu dan perhatian dalam menguji skripsi ini serta memberikan masukan yang membangun.
7. Para Dosen Program Studi Teknik Informatika di Universitas Nusa Putra Sukabumi yang telah berjasa kepada penulis atas ilmu yang telah diberikan.
8. Kepada ke Orang Tua kami yang telah memberikan doa dan sangat mendukung kami dalam segala hal dan selalu mengusahakan yang terbaik.
9. Kepada pemilik nama Nurhasanah, selaku istri dari penulis Moh. Abd. Aziz Hidayat. Terima kasih untuk segala dukungan dan partisipasinya selama penulisan skripsi ini. Semoga Allah SWT selalu menjaga dan melindungi dimanapun kamu berada.
10. Kepada pemilik nama Rahma Kamilah, selaku istri dari penulis Tofik Hidayat. Terima kasih telah membantu dan membersamai selama proses penulisan skripsi serta memberikan semangat dan dukungan sampai terselesaikannya skripsi ini.
11. Kepada Bapa Henda, M. Ag Sebagai kepala seksi Pendidikan Diniyah dan Pondok Pesantren di kemenag kabupaten sukabumi beserta staff, pengurus dan jajarannya yang telah memeberikan ijin untuk penelitian.
12. Kepada Bapa Dr. H.M Deddy Soe Aiddy, MSI selaku kepala Yayasan Rahmatan Lil-Alamien beserta pengurus dan jajarannya yang telah memeberikan ijin untuk penelitian.
13. Teman seperjuangan kelas TI20C angkatan 2020, terima kasih untuk kebersamaan yang telah diberikan.
14. Teman seperjuangan kelas TI22C angkatan 2022, terima kasih untuk kebersamaan yang telah diberikan.

15. Kepada rekan-rekan sepejuangan kami yaitu mahasiswa Program Studi Teknik Informatika yang telah menyemangati sampai selesainya skripsi ini.
16. Kepada semua pihak terkait lainnya, kami ucapkan terima kasih sebesar-besarnya.

Penulis menyadari bahwa skripsi ini masih jauh dari kesempurnaan, oleh karena itu, penulis sangat mengharapkan kritik dan saran yang mampu membangun pemikiran-pemikiran baru untuk penulisan-penulisan selanjutnya.

Sukabumi, 20 Juni 2024

Moh. Abd. Aziz Hidayat

Tofik Hidayat

Penulis 1



Penulis 2

**HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI
TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS**

Sebagai sivitas akademik UNIVERSITAS NUSA PUTRA, kami yang bertanda tangan di bawah ini:

Nama : MOH. ABD. AZIZ HIDAYAT
NIM : 20220040014
Nama : TOFIK HIDAYAT
NIM : 20200040056
Program Studi : Teknik Informatika
Jenis karya : Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Nusa Putra *Hak Bebas Royalti Noneksklusif (Non-exclusive Royalty- Free Right)* atas karya ilmiah kami yang berjudul:

“IMPLEMENTASI PENCARIAN SEMANTIK DALAM TAFSIR AL-QURAN DENGAN ALGORITMA COSINE SIMILARITY DAN LARGE LANGUAGE MODELS”. Beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Noneksklusif ini Universitas Nusa Putra berhak menyimpan, mengalih media/format-kan, mengelola dalam bentuk pangkalan data (database), merawat, dan memublikasikan tugas akhir kami selama tetap mencantumkan nama kami sebagai penulis/pencipta dan sebagai pemilik Hak Cipta. Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Universitas Nusa Putra

Pada tanggal : 20 Juni 2024

Yang Menyatakan:



Moh. Abd. Aziz Hidayat

Penulis 1



Tofik Hidayat

Penulis 2

DAFTAR ISI

HALAMAN SAMPUL	i
HALAMAN JUDUL.....	ii
PERNYATAAN PENULIS.....	iii
PERSETUJUAN SKRIPSI.....	iv
PENGESAHAN SKRIPSI	v
HALAMAN PERSEMBAHAN	vi
ABSTRACT.....	vii
ABSTRAK.....	viii
KATA PENGANTAR	ix
HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI.....	xii
DAFTAR ISI.....	xiii
DAFTAR TABEL.....	xvi
DAFTAR GAMBAR	xvii
DAFTAR LAMPIRAN	xix
BAB I PENDAHULUAN	1
16.1 Latar Belakang	1
16.2 Rumusan Masalah	3
16.3 Batasan Masalah.....	3
16.4 Tujuan Penelitian.....	3
16.5 Manfaat Penelitian.....	4
16.6 Sistematika Penulisan.....	4
BAB II TINJAUAN PUSTAKA.....	6
2.1 Penelitian Terkait	6
2.2 Tafsir Al-Quran	9
2.3 Pencarian Semantik	10
2.4 <i>Cosine Similarity</i>	11
2.5 <i>Large Language Models (LLM)</i>	13
2.6 GPT	14
2.7 Transformer	15
2.8 <i>Vector Database</i>	16
2.9 Black Box Testing.....	17
2.10 Kerangka Pemikiran.....	18

BAB III METODOLOGI PENELITIAN.....	22
3.1 Tahapan Penelitian	22
3.2 Identifikasi Masalah	23
3.3 Metode Pengumpulan data.....	25
3.3.1 Studi Pustaka.....	26
3.3.2 Wawancara.....	26
3.3.3 Seed Data	27
3.4 Data <i>Preproessing</i>	28
3.5 <i>Feature Engineering</i>	28
3.6 Model Selection.....	29
3.7 <i>Model Training</i>	31
3.8 <i>Model Evaluation</i>	31
3.9 <i>Model Testing</i>	31
3.10 <i>Deployment</i>	31
3.10.1 Desain UI	32
3.10.2 Evaluasi/Testing	33
3.11 Maintenance	33
BAB IV HASIL DAN PEMBAHASAN.....	34
4.1 Wawancara	34
4.2 <i>Literatur Review</i>	35
4.3 Seed Data.....	36
4.4 Data Preprocessing	39
4.5 Vektorisasi Data.....	40
4.6 Cosine Similarity.....	45
4.7 <i>Large Language Models</i>	48
4.8 Pelatihan Model.....	49
4.9 Model Evaluation	49
4.10 Model Testing	56
4.11 Deployment	58
4.12 Pengujian Sistem.....	60
BAB V KESIMPULAN DAN SARAN	62
5.1 Kesimpulan.....	62
5.2 Saran.....	63
DAFTAR PUSTAKA	64

LAMPIRAN	67
----------------	----



DAFTAR TABEL

Tabel 2. 1 Penelitian Terkait	6
Tabel 4. 1 Tabel Skenario model testing.....	57
Tabel 4. 2 Hasi Pengujian.....	61



DAFTAR GAMBAR

Gambar 2. 1 Flowchart Pencarian Semantik	11
Gambar 2. 2 Flowchart Algoritma Coisne Similarity	13
Gambar 2. 3 Flowchart LLM.....	14
Gambar 2. 4 Arsitektur Transformer	16
Gambar 2. 5 Kerangka pemikiran	19
Gambar 3. 1 Tahapan Penelitian.....	22
Gambar 3. 2 Gambaran Umum Sistem	23
Gambar 3. 3 proses search engine	24
Gambar 3. 4 Proses seed data.....	27
Gambar 3. 5 Desain Tampilan Awal	32
Gambar 3. 6 Desain Tampilan Hasil Pencarian.....	32
Gambar 4. 1 Data JSON Surah.....	37
Gambar 4. 2 Data JSON Ayat	37
Gambar 4. 3 kode scraping data	38
Gambar 4. 4 Hasil data scraping.....	38
Gambar 4. 5 kode clearing data.....	39
Gambar 4. 6 Kode penerjemahan	39
Gambar 4. 7 Contoh hasil penerjemahan yang di lakukan	40
Gambar 4. 8 Flowchart Vektorisasi	40
Gambar 4. 9 kode <i>embedding</i> word.....	41
Gambar 4. 10 kode model bert	42
Gambar 4. 11 konfigurasi vektor database	42
Gambar 4. 12 Raw data sebelum di vektorisasi.....	43
Gambar 4. 13 data Vektor	44
Gambar 4. 14 Kode input data ke Qdrant.....	45
Gambar 4. 15 Tampilan UI Qdrant.....	45
Gambar 4. 16 Proses vektorisasi inputan	46
Gambar 4. 17 Kode penerpan <i>cosine similarity</i>	46
Gambar 4. 18 Hasil dari perhitungan cosine similarity	47
Gambar 4. 19 Contoh hasil pencarian	47

Gambar 4. 20 Proses penerapan Large Language Models	48
Gambar 4. 21 Hasil ringkasan.	48
Gambar 4. 22 kode pengurutan Score	49
Gambar 4. 23 Kode evaluasi model	50
Gambar 4. 24 Hasil F1-Score Vs Threshold.....	51
Gambar 4. 25 Hasil ROC-AUC Score.....	51
Gambar 4. 26 Gambar ringkasan hasil dari semua evaluasi.....	52
Gambar 4. 27 Hasil <i>F1-Score Vs Threshold</i> dimensi 768.....	53
Gambar 4. 28 Hasil ROC-AUC Score dimensi 768.....	54
Gambar 4. 29 Hasil F1-Score Vs Threshold dimensi 384.....	55
Gambar 4. 30 Hasil ROC-AUC Score dimensi 384.....	55
Gambar 4. 31 Proses berjalan nya server	59
Gambar 4. 32 Halaman Awal	59
Gambar 4. 33 Halaman Hasil Pencarian.....	60



DAFTAR LAMPIRAN

Lampiran 1 Kegiatan Penelitian Di Kemenag Kab. Sukabumi	67
Lampiran 2 Kegiatan Penelitian Di Yayasan Rahmatan Lil-Alamien	67
Lampiran 3 Hasil Pencarian Al-Baqarah.....	68
Lampiran 4 Hasil Pencarian Al-Baqarah.....	68
Lampiran 5 Hasil Pencarian An-Nisa.....	69
Lampiran 6 Hasil Pencarian Tafsir Al-Baqarah	69
Lampiran 7 Hasil Pencarian Tafsir Puasa	70
Lampiran 8 Hasil Pencarian Tafsir An-Nisa	70
Lampiran 9 Hasil Pencarian Tafsir Al-Baqarah 21	71
Lampiran 10 Hasil Pencarian Tafsir Tentang Puasa	71
Lampiran 11 Hasil Pencarian Tafsir An-Nisa 4	72
Lampiran 12 Hasil Pencarian Ayat Apa Perintah Puasa	73
Lampiran 13 Hasil Pencarian Ayat Apa Perintah Zakat	73
Lampiran 14 Hasil Pencarian Ayat Apa Perintah Solat	74
Lampiran 15 Hasil Pencarian Ayat Tentang Perintah Puasa.....	74
Lampiran 16 Hasil Pencarian Ayat Tentang Perintah Zakat	75
Lampiran 17 Hasil Pencarian Ayat Tentang Perintah Solat.....	75
Lampiran 18 Hasil Pencarian Hukum Pembelian Dengan Crypto	76
Lampiran 19 Hasil Pencarian Hukum Bagi Yang Mengidolakan Korea	76
Lampiran 20 Hasil Pencarian Apa hukum puasa di hari tasyrik	77
Lampiran 21 Hasil Pencarian Carikan Tafsir tentan kurban	77
Lampiran 22 Hasil Pencarian Jelaskan tafsir dari surat al-ikhlas.....	78
Lampiran 23 Hasil Pencarian Carikan Tafsir dari surat Ar-Rum ayat 21	78
Lampiran 24 Hasil Pencarian Carikan Tafsir tentang tenggelamnya firaun	79
Lampiran 25 Hasil Pencarian Carikan tentang kaum Nabi Nuh AS	79
Lampiran 26 Apa surat pertama yang di turunkan	80
Lampiran 27 Nabi pertama yang di angkat	80
Lampiran 28 Kurban di zaman Nabi Ibrahim AS	81
Lampiran 29 Tafsir tentang toleransi beragama	81
Lampiran 30 Tafsir tentang qadha puasa Ramadhan	82
Lampiran 31 Dari apa manusia di ciptakan.....	82

Lampiran 32 Tafsir tentang keesaan Allah SWT	83
Lampiran 33 Keutamaan salat tahajud	83
Lampiran 34 Seberapa sabar nabi daud.....	84
Lampiran 35 Hasil Skenario Model Testing Validator 1	85
Lampiran 36 Hasil Skenario Model Testing Validator 2	86
Lampiran 37 Hasil Skenario Model Testing Validator 3.....	87
Lampiran 38 Hasil Skenario Model Testing Validator 4.....	88
Lampiran 39 Hasil Pengujian Validator 1.....	89
Lampiran 40 Hasil Pengujian Validator 2.....	90
Lampiran 41 Hasil Pengujian Validator 3.....	91
Lampiran 42 Hasil Pengujian Validator 4.....	92
Lampiran 43 Biodata Penulis 1	93
Lampiran 44 Biodata Penulis 2	94



BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan era digitalisasi informasi saat ini, membuat pencarian dan pemahaman makna terhadap Al-Qur'an tidak hanya terbatas pada pencarian *konveksional*. Al-Qur'an merupakan sumber utama ajaran dan pedoman bagi umat islam di seluruh dunia. Memahami makna dan tujuan dari setiap ayat Al-Qur'an merupakan hal yang penting bagi umat islam. Namun, dalam proses pencarian tafsir Al-Qur'an, seringkali orang mengalami kendala dalam menemukan konteks yang relevan dengan pertanyaan atau kebutuhan mereka. Hal ini disebabkan oleh variasi istilah, gaya bahasa, dan kompleksitas struktur kalimat Al-Quran.

Untuk memahami makna dari Al-Quran di perlukan sebuah ilmu tafsir. Tafsir yaitu ilmu yang mempelajari cakupan tentang keadaan turunya ayat Al-Qur'an, asbab al-nuzul, urutan makiyyah-madaniyyah, muhkam-mustasyabbih, nasikh-mansukh, khās-'am, mutlaq-muqāyyād, mujmal, halal-harām, janji-janji, perintah-larangan, i'tibar perumpamaan di dalam Al-Qur'an[1]. Dalam mempelajari Al-Quran biasanya orang – orang hanya fokus dalam mempelajari ilmu tajwid untuk membaca Al-Quran dan merasa itu sudah cukup. Untuk itu dengan adanya penelitian ini bisa menambah wawasan dan pemahaman bahwa mempelajari Al-Quran bukan hanya terfokus pada membaca saja melainkan memahami juga makna dari setiap ayat-ayat yang ada. Dengan adanya penerapan pencarian semantik dalam tafsir Al-Quran di penelitian ini memungkinkan mesin pencari untuk memahami intents dan konteks dari pertanyaan pencarian, sehingga hasil pencarian yang dihasilkan menjadi lebih relevan dan akurat.

Metode yang digunakan untuk mengambil informasi Al-Qur'an dapat dikategorikan menjadi tiga jenis yaitu : berbasis semantik, berbasis kata kunci dan *Cross-Language Information Retrieval* (CLIR). Metode berbasis semantik adalah teknik pencarian berbasis konsep yang menghasilkan hasil berdasarkan makna kata atau kecocokan konsep. Metode berbasis kata kunci menghasilkan hasil berdasarkan kueri kata yang cocok dengan huruf. Metode CLIR mengambil

informasi yang ditulis dalam bahasa yang berbeda dari bahasa permintaan pengguna. [2]. Kebanyakan alat pencarian Alquran menggunakan teknik pencarian kata kunci. Metode pencarian semantik dalam Al-Quran yang ada saat ini adalah metode berbasis ontologi dan himpunan sinonim. Teknik berbasis ontologi mencari konsep/topik yang sesuai dengan *query* pengguna. Metode ini kemudian menghasilkan ayat-ayat yang berkaitan dengan topik tersebut. Metode himpunan sinonim menggunakan sinonim untuk kata kueri. Kemudian, menemukan semua ayat Al-Quran yang cocok dengan sinonim dari sinonim kata-kata tersebut. CLIR menerjemahkan kata-kata dari kueri masukan ke dalam bahasa berbeda dan kemudian mengambil ayat-ayat yang cocok dengan kata-kata yang diterjemahkan [2].

Penelitian kali ini akan membahas mengenai penggunaan pencarian semantik pada tafsir Al-Quran dengan menggunakan algoritma *cosine similarity* dan *large language models*. *Cosine similarity* adalah perhitungan kesamaan antara dua vektor n dimensi dengan mencari kosinus dari sudut diantara keduanya dan sering digunakan untuk membandingkan dokumen dalam text mining[3]. LLM (*Large Language Model*) merupakan Salah satu Kecerdasan Buatan generatif yang paling populer saat ini, LLM yaitu model jaringan syaraf yang didasarkan pada data teks dengan jumlah yang besar, dan dirancang untuk menghasilkan output yang menyerupai manusia, termasuk namun tidak terbatas pada prosa, puisi, dan bahkan kode program[4].

Walaupun pencarian semantik telah berkembang dalam berbagai bidang. Akan tetapi, pengaplikasiannya dalam tafsir Al-Quran masih terbatas. Penelitian ini berusaha mengatasi kesenjangan tersebut dengan mengintegrasikan teknologi pencarian semantik menggunakan algoritma *cosine similarity* dan *large language models*. Untuk memperoleh hasil yang lebih efektif, pengembangan aplikasi ini menggunakan *Vector Database*. Tujuan spesifik dari penggunaan *Vector Database* meliputi peningkatan dalam akurasi pencarian tafsir dan kemudahan dalam mengidentifikasi dan memahami tafsir, dataset yang akan di gunakan di ambil dari quran.kemenag.go.id dan quran.com.

Berdasarkan latar belakang yang telah dipaparkan maka penulis tertarik untuk melakukan penelitian dengan judul “Implementasi Pencarian Semantik Dalam

Tafsir Al-Quran Dengan Algoritma *Cosine Similarity* Dan *Large Language Models*". Diharapkan dengan penelitian ini akan memberikan kontribusi yang berarti dalam bidang studi tafsir Al-Quran, dan juga dengan algoritma yang digunakan dapat mengoptimalkan pencarian semantik.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan di atas, maka permasalahan yang didapat adalah sebagai berikut :

1. Bagaimana cara mengintegrasikan algoritma *cosine similarity* dan *large language models* untuk melakukan pencarian semantik dalam teks tafsir Al-Quran?
2. Bagaimana akurasi dari pencarian semantik dengan algoritma yang digunakan?
3. Bagaimana pengaruh penggunaan *vector* database dalam algoritma yang digunakan?

1.3 Batasan Masalah

Agar pembahasan penelitian ini berfokus pada ruang lingkup masalah yang diinginkan, maka ada batasan masalah yang akan dibatasi sebagai berikut :

1. Pemfokusan pencarian semantik hanya pada tafsir Al-Quran.
2. Tafsir yang digunakan adalah tafsir Ibnu Katsir, Tafsir Wajiz, Tafsir Tahlili dan Tafsir Mufti Muhammad Shafi, juga beberapa data tambahan seperti Kesimpulan Surah, Asbabun Nuzul, Kosa Kata, Outro Surah dan Intro Surah.
3. Database *vector* yang digunakan berdimensi 384, 768, dan 3072.
4. Pencarian tidakdi gunakan bisa untuk melakukan perhitungan.
5. Hasil atau urutan dari hasil pencarian tidak selalu sama.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah di atas, maka penelitian ini bertujuan untuk :

1. Menerapkan algoritma *cosine similarity* dan *large language models* untuk melakukan pencarian semantik dalam teks tafsir Al-Quran.
2. Mengetahui kinerja pencarian semantik dengan algoritma yang digunakan.

3. Mengetahui pengaruh penggunaan *vector* database dalam algoritma yang digunakan.

1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah:

1. Bagi Penulis
 - a. Memberikan pengalaman dalam menerapkan algoritma *cosine similarity* dan *large language models* untuk pencarian semantik, yang dapat meningkatkan pemahaman dan keterampilan penulis di bidang tersebut.
 - b. Hasil dari penelitian ini akan menjadi bahan informasi yang bermanfaat bagi penulis.
 - c. Mendapatkan pengalaman dalam sebuah penelitian dan penulisan penelitian.
2. Bagi Instansi
 - a. Aplikasi yang dihasilkan bisa digunakan untuk media pembelajaran tafsir Al-Quran.
 - b. Aplikasi yang dihasilkan bisa digunakan mempermudah melakukan pencarian tafsir Al-Quran.
3. Bagi Pembaca
 - a. Hasil dari penelitian ini akan menjadi bahan informasi yang bermanfaat bagi pembaca sehingga dapat menambah referensi dan wawasan.
 - b. Memberikan pemahaman tentang penerapan algoritma *cosine similarity* dan *large language models* untuk pencarian semantik.

1.6 Sistematika Penulisan

Sistematika penulisan skripsi ini disusun untuk memberikan gambaran secara umum tentang penulisan skripsi yang dikerjakan. Adapun sistematika skripsi ini adalah sebagai berikut :

BAB I: Pendahuluan

Pada bab ini dijelaskan tentang latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian dan sistematika penulisan

BAB II: Tinjauan Pustaka

Pada bab ini membahas tentang landasan umum mengenai konsep dasar dan teori yang berkaitan dengan penelitian diantaranya mengenai literatur-literatur terkait tentang metode pencarian semantik, algoritma *cosine similarity*, *Large Language Models*, vektor database, serta penelitian terdahulu yang relevan.

BAB III: Metodologi Penelitian

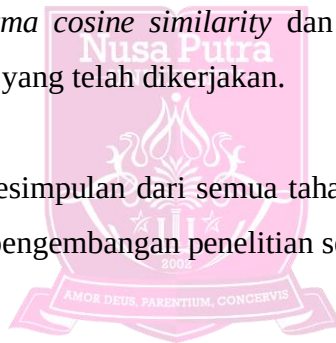
Pada bab ini mencakup berbagai tahapan penelitian, meliputi tahapan pengumpulan data, analisis, perancangan, dan pengembangan sistem. Dibahas juga teknik dan alat penelitian yang digunakan sepanjang tahapan tersebut, serta lokasi subjek penelitian, data/informasi yang diperlukan, dan waktu yang dialokasikan untuk setiap tahapan.

BAB IV: Hasil Dan Pembahasan

Pada bab ini berisi tentang hasil dari penelitian yang dilakukan, pembahasan dari penelitian yang dilakukan dan pembahasan dari implementasi pencarian semantik dengan *algoritma cosine similarity* dan *large language models*, serta pengujian dari penelitian yang telah dikerjakan.

BAB V : Penutup

Pada bab ini berisi kesimpulan dari semua tahapan penelitian yang dilakukan dan saran – saran untuk pengembangan penelitian selanjutnya.





BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan uraian yang telah di paparkan pada BAB sebelum nya, penelitian ini dapat menjawab pertanyaan rumusan masalah dari BAB 1 yang dapat di simpulkan sebagai berikut :

1. Penelitian telah menghasilkan sistem pencarian semantik tafsir Al-Quran dengan mengintegrasikan algoritma *cosine similarity* yang di gunakan untuk menghitung kesamaan dua buah vektor antara vektor dari inputan (*query*) dan vektor dari database Qdrant. Penelitian ini juga berhasil menerapkan *large language models* dengan menggunakan GPT-4 untuk melakukan proses peringkasan dari hasil pencarian sehingga menghasilkan sebuah rangkuman dan juga di gunakan untuk mencari jawaban terbaik dari hasil perhitungan *cosine similarity*.
2. Berdasarkan pengujian yang telah di lakukan kinerja sistem pencarian semantik dengan algoritma *cosine similarity* dan *large language models* menunjukan hasil yang baik. Yaitu, 96% untuk akurasi dari model testing dan 89% untuk akurasi relevansi hasil.
3. Dimensi vektor 3072 mencapai F1-Score tertinggi sebesar 0.819 dan memiliki kinerja terbaik dalam hal ROC-AUC Score sebesar 0.582. Sebagai perbandingan, dimensi vektor 768 hanya menghasilkan F1-Score sebesar 0.6 dan ROC-AUC Score sebesar 0.41, sementara dimensi vektor 384 menghasilkan F1-Score sebesar 0.65 dan ROC-AUC Score sebesar 0.45. Hasil ini menunjukkan bahwa dimensi vektor 3072 memberikan performa yang paling unggul. Penggunaan dimensi vektor yang lebih tinggi memungkinkan representasi teks yang lebih detail dan kaya, penting untuk menangkap nuansa dan konteks dari tafsir Al-Quran. Oleh karena itu, untuk mencapai kinerja optimal dalam pencarian semantik tafsir Al-Quran, penggunaan dimensi vektor 3072 sangat direkomendasikan.

5.2 Saran

Berdasarkan hasil penelitian yang telah dilakukan terdapat beberapa saran perbaikan yang dapat dilakukan untuk meningkatkan akurasi model sistem pencarian semantik tafsir ayat Al-Quran, antara lain:

1. Memperbanyak data tafsir : Meningkatkan jumlah data yang digunakan untuk model mesin pencarian dapat memberikan referensi yang lebih banyak sehingga hasil pencarian akan lebih baik.
2. Meningkatkan kualitas data: Meningkatkan kualitas data yang digunakan untuk model pencarian sehingga model dapat belajar dari data yang lebih akurat.
3. Mengembangkan algoritma pencarian: Mengembangkan algoritma pencarian yang lebih canggih sehingga model dapat lebih memahami maksud pengguna.
4. Melakukan evaluasi terhadap sistem yang telah dibuat secara berkala untuk memastikan bahwa sistem pencarian selalu mampu menjawab *query* dengan tepat.



DAFTAR PUSTAKA

- [1] H. Hidayat, "Sejarah Perkembangan Tafsir Al-Qur'an," *Al-Munir J. Stud. Ilmu Al-Qur'an dan Tafsir*, vol. 2, no. 01, pp. 29–76, 2020.
- [2] E. H. Mohamed and E. M. Shokry, "QSST: A Quranic Semantic Search Tool based on word embedding," *J. King Saud Univ. Inf. Sci.*, vol. 34, no. 3, pp. 934–945, 2022.
- [3] D. N. Lindang, A. Y. Muniar, A. Halid, M. Muhajirin, and A. Amiruddin, "Sistem Penentuan Kemiripan Antar Skripsi Menggunakan Metode Cosine Similarity Pada Perpustakaan," in *Seminar Nasional Teknik Elektro dan Informatika (SNTEI)*, 2023, pp. 321–324.
- [4] P. Denny *et al.*, "Computing education in the era of generative AI," *Commun. ACM*, vol. 67, no. 2, pp. 56–67, 2024.
- [5] F. Beirade, H. Azzoune, and D. E. Zegour, "Semantic query for Quranic ontology," *J. King Saud Univ. Inf. Sci.*, vol. 33, no. 6, pp. 753–760, 2021.
- [6] M. Alshammeri, E. Atwell, and M. ammar Alsalka, "Detecting semantic-based similarity between verses of the Quran with Doc2vec," *Procedia Comput. Sci.*, vol. 189, pp. 351–358, 2021.
- [7] Y. Han, C. Liu, and P. Wang, "A Comprehensive Survey on Vector Database: Storage and Retrieval Technique, Challenge," *arXiv Prepr. arXiv2310.11703*, 2023.
- [8] D. Schumacher, "V3CTRON| Data Retrieval & Access System For Flexible Semantic Search & Retrieval Of Proprietary Document Collections Using Natural Language Queries.," *Available SSRN*, 2023.
- [9] N. Oh, G.-S. Choi, and W. Y. Lee, "ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models," *Ann. Surg. Treat. Res.*, vol. 104, no. 5, p. 269, 2023.
- [10] E. D. Saputra, N. S. Harahap, J. Jasril, and Y. Yusra, "Question Answering Al-Qur'an Menggunakan Generative Pre-Trained Transformer 3.5 Berbasis Chatbot Telegram," *Jutisi J. Ilm. Tek. Inform. dan Sist. Inf.*, vol. 13, no. 1, pp. 550–563, 2024.
- [11] V. Setlur, A. Kanyuka, and A. Srinivasan, "Olio: A Semantic Search Interface for Data Repositories," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023, pp. 1–16.
- [12] F. AZMI, A. Saleh, and N. P. Dharshinni, "Face Identification on Login Security Using Algorithm Combination of Viola-Jones and Cosine Similarity," *J. Informatics Telecommun. Eng.*, vol. 4, no. 1, pp. 203–211, 2020.
- [13] M. Yunus, I. Mohamad, Z. Wildan Budiawan, and D. Wahyudin, "Similarity detection for hadith of Fiqh of women using cosine similarity and boyer

- moore method,” *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 1, pp. 63–75, 2020.
- [14] W. X. Zhao *et al.*, “A survey of large language models,” *arXiv Prepr. arXiv2303.18223*, 2023.
- [15] H. Naveed *et al.*, “A comprehensive overview of large language models,” *arXiv Prepr. arXiv2307.06435*, 2023.
- [16] R. Abdurrohman, “Uji performa chatbot dengan retrieval augmented generation dan model gpt-4 untuk domain taharah berdasarkan empat imam mazhab fikih (studi kasus kitab rahmah al ummah fi ikhtilaf al a’immah).”
- [17] E. Waisberg *et al.*, “GPT-4 and medical image analysis: strengths, weaknesses and future directions,” *J. Med. Artif. Intell. Vol 6 (December 30, 2023) J. Med. Artif. Intell.*, 2023, [Online]. Available: <https://jmai.amegroups.org/article/view/8338>
- [18] S. M. Jain, “Hugging Face BT - Introduction to Transformers for NLP: With the Hugging Face Library and Models to Solve Problems,” S. M. Jain, Ed., Berkeley, CA: Apress, 2022, pp. 51–67. doi: 10.1007/978-1-4842-8844-3_4.
- [19] G. I. P. Djuharis, “Prediksi harga bitcoin menggunakan transformer: Mengintegrasikan harga, sentimen, tren, dan volume Dalam analisis deret waktu multivariat.” Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta.
- [20] T. Taipalus, “Vector database management systems: Fundamental concepts, use-cases, and current challenges,” *Cogn. Syst. Res.*, vol. 85, p. 101216, 2024, doi: <https://doi.org/10.1016/j.cogsys.2024.101216>.
- [21] C. Perone, “Bachelor’s Thesis Assignment”.
- [22] C. Galli, N. Donos, and E. Calciolari, “Performance of 4 Pre-Trained Sentence Transformer Models in the Semantic Query of a Systematic Review Dataset on Peri-Implantitis,” *Information*, vol. 15, no. 2, p. 68, 2024.
- [23] W. Wang and Y. Yang, “Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models,” *arXiv Prepr. arXiv2403.06098*, 2024.
- [24] S. A. Elisya and I. L. Kharisma, “Pembuatan sistim informasi arsip surat berbasis website (studi kasus: kelurahan sriwidari),” *J. CoSciTech (Computer Sci. Inf. Technol.*, vol. 4, no. 1, pp. 136–145, 2023.
- [25] L. D. Iklima and I. Yustiana, “Aplikasi Belajar Huruf Hijaiyah Berbasis Web TPA Uswatun Hasanah,” *J. CoSciTech (Computer Sci. Inf. Technol.*, vol. 3, no. 3, pp. 364–370, 2022.
- [26] M. N. Adlini, A. H. Dinda, S. Yulinda, O. Chotimah, and S. J. Merliyana, “Metode penelitian kualitatif studi pustaka,” *Edumaspul J. Pendidik.*, vol. 6, no. 1, pp. 974–980, 2022.
- [27] N. Septian, N. Suarna, and D. R. Amalia, “Sistem Informasi Akademik

- Berbasis Website untuk Meningkatkan Prestasi Belajar Siswa MAN 3 Cirebon di Masa Pandemi Covid-19,” *Intern. (Information Syst. Journal)*, vol. 5, no. 1, pp. 1–11, 2022.
- [28] I. L. Kharisma, D. A. Septiani, A. Fergina, and K. Kamdan, “Penerapan Algoritma Decision Tree untuk Ulasan Aplikasi Vidio di Google Play,” *J. Nas. Teknol. dan Sist. Inf.*, vol. 9, no. 2, pp. 218–226, 2023.



