Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 9, No. 5, Mei 2025, hlm. x-x http://j-ptiik.ub.ac.id

Sistem Tanya Jawab Closed-Domain terhadap Dokumen Fatwa menggunakan Retrieval Augmented Generation dan Large Language Model

Firhan Imam Haekal[1], Rizal Setya Perdana[2], Putra Pandu Adikara[3]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]firhanih@student.ub.ac.id, [2]rizalespe@ub.ac.id, [3]adikara.putra@ub.ac.id

Abstrak

Agama Islam menjunjung tinggi kesesuaian antara hukum serta prinsip syariah dengan pengambilan keputusan dan tanggapan seorang Muslim terhadap setiap masalah atau fenomena sosial yang terjadi. Permasalahan dan berbagai fenomena sosial baru seiring berkembangnya teknologi serta peradaban memunculkan kebutuhan fatwa sebagai dasar pengambilan keputusan seorang Muslim dalam menghadapi kedua hal tersebut. Berbagai figur religi serta Majelis Ulama Indonesia (MUI) berupaya memenuhi permintaan ini dengan sejumlah aktivitas dakwah, namun permintaan masif dan berulang dari masyarakat memunculkan risiko penyampaian materi dakwah secara dangkal, inkonsistensi materi terhadap sumber, misinformasi, dan minimnya etika dalam menyampaikan materi. Solusi yang diusulkan dalam penelitian ini berupa rancangan sistem tanya jawab yang menghasilkan jawaban komprehensif berdasarkan dokumen fatwa yang sudah ada. Solusi ini dapat diwujudkan menggunakan rangka kerja Retrieval Augmented Generation (RAG) yang terdiri dari Large Language Model (LLM) sebagai generator, yaitu penghasil jawaban, serta *retriever* sebagai pencari dokumen sumber dari jawaban. Untuk memastikan sistem dapat membentuk jawaban dengan kemiripan yang tinggi terhadap dokumen fatwa, penelitian ini juga membandingkan metode pencarian untuk *information retrieval* seperti *cosine similarity* dan Maximum Inner Product Search (MIPS). Hasil pengujian terhadap komponen *retriever* menunjukkan bahwa metode pencarian *cosine similarity* dan MIPS mencapai nilai kinerja terbaik pada setiap metrik pada jumlah K sebanyak 25, salah satunya F1-Score@K dengan nilai 0,52. Nilai rerata terbaik dari dua metrik kinerja *generator* tercapai dengan integrasi RAG, yaitu 0,67 untuk P-BERT dan 0,63 untuk F-BERT.

Kata kunci: *retrieval augmented generation, large language model, fatwa, syariah*

Abstract

Islam upholds the compatibility between the law and the principles of sharia with the decision-making and response of a Muslim to any problems or social phenomena that occur. Problems and various new social phenomena along with the development of technology and civilization have led to the need for fatwas as the basis for Muslim decision-making in response towards these two things. Various religious figures and the Indonesian Ulema Council are trying to fulfill this demand with a number of da'wah activities, but massive and repeated requests from the public raise the risk of delivering da'wah material superficially, inconsistency of material with sources, misinformation, and lack of ethics in delivering material. The solution proposed in this research is the design of a question and answer system for fatwa documents that produces comprehensive answers based on reliable sources. This solution can be realized using the Retrieval Augmented Generation (RAG) framework which consists of a Large Language Model (LLM) as the answer generator and a retriever as the source document finder of the answer. To ensure the system can form answers with high similarity towards appropriate fatwa documents, this research also compares search methods for information retrieval such as cosine similarity and Maximum Inner Product Search (MIPS). The results of testing the retriever component show that the cosine similarity and MIPS search methods achieve the best performance value on each metric at a K number of 25, one of which is F1-Score@K with a value of 0,52. The best average value of the two generator performance metrics is achieved with RAG integration, which is 0,67 for P-BERT and 0,63 for F-BERT.

Keywords: *retrieval augmented generation, large language model, fatwa, syariah*

1. PENDAHULUAN

Umat Islam Indonesia menduduki 80.3% dari bagian seluruh populasi negara dengan jumlah muslim terbanyak di dunia yang didukung dengan capaian 230 juta penduduk pada tahun 2000 (Yung, 2003). Jumlah tersebut membawa pengaruh pada berbagai sektor, seperti sektor politik yang ditunjukkan dari preferensi dari 52% masyarakat Indonesia untuk memilih pemerintah Muslim dibandingkan non-Muslim berdasarkan data Lembaga Survei Indonesia (LSI) (Salim, 2022).

Fenomena tersebut menunjukkan keterikatan antara identitas muslim dengan perilaku dan perspektif muslim di Indonesia. Perspektif ini memengaruhi pengambilan keputusan masyarakat dalam menyikapi permasalahan yang dihadapi dalam kehidupan sehari hari, sehingga memunculkan kebutuhan akan petunjuk pengambilan keputusan berdasarkan landasan yang sesuai dengan hukum Islam dan konsep syariah. Variabilitas dari permasalahan dalam sejumlah topik yang bersumber dari interaksi muslim di Indonesia pada berbagai sektor tersebut menjadi pemicu proses penerbitan fatwa sebagai dasar hukum kehidupan Islam di negara Indonesia.

Dalam upaya memenuhi kebutuhan tersebut, pemuka agama dan Majelis Ulama Indonesia (MUI) menghadiri solusi yang relevan dalam bentuk konten dan dokumen digital, sehingga memudahkan publik mengakses jawaban yang sesuai dengan permasalahan yang dihadapi (Sebihi dan Moazzam, 2024), namun permintaan secara konstan, masif, dan berulang dari masyarakat terhadap permasalahan yang dihadapi menambah beban dari penyampaian materi (Saragih, Sagala, dan Effendi, 2023).

Permasalahan tersebut dapat diselesaikan dengan pengumpulan dokumen fatwa pada satu tempat terpusat dengan sistem yang menyediakan antarmuka untuk mengakses dokumen tersebut. Solusi ini diusulkan dalam penelitian yang dilakukan oleh Hariri (2021), melalui perancangan sistem dengan metode *Fuzzy C-Means* untuk melakukan ekstraksi informasi dokumen fatwa MUI, namun metode ini tidak dapat menghasilkan rangkuman atau jawaban berdasarkan dokumen fatwa yang tersedia, melainkan hanya pengaksesan informasi dari dokumen fatwa tersebut.

Penulis berupaya untuk menyelesaikan permasalahan tersebut dalam penelitian ini dengan rangka kerja Retrieval Augmented Generation (RAG) yang memiliki kapabilitas untuk menghasilkan jawaban menggunakan Large Language Model (LLM) berdasarkan dokumen fatwa yang sudah ada, sehingga tidak terjadi misinformasi ataupun halusinasi berupa fatwa baru. Penelitian ini juga membandingkan metode pencarian dokumen yang digunakan pada penelitian sebelumnya terkait RAG oleh Vaswani, et al., (2020) yaitu Maximum Inner Product Search (MIPS) dan *cosine similarity* berdasarkan kapabilitasnya dalam mencari dokumen pada ruang penyimpanan vektor terlepas dari ukuran dokumen, serta membandingkan kinerja generator berupa LLM dalam menghasilkan jawaban dengan dan tanpa dokumen fatwa yang dikembalikan *retriever* untuk memastikan sistem dapat membentuk jawaban dengan kemiripan yang tinggi terhadap dokumen fatwa.

2. DASAR TEORI

2.1. Fatwa

Menurut Awass (2019), fatwa adalah opini tidak mengikat yang diberikan oleh seorang ahli hukum Islam (mufti) terhadap permintaan dari praktisi religi yang membutuhkan pendapat terhadap urusan religi atau praktik sosial. Proses penerbitan fatwa yang juga diketahui sebagai *ifta* berawal dari pertanyaan terkait kepercayaan dan praktik dari prinsip agama oleh umat Islam pada masa hidup Rasulullah. Permintaan tersebut dipenuhi dengan pendapat Rasulullah berdasarkan jawaban yang diturunkan Al Qur'an, sehingga menjadikannya petunjuk dari tindakan setiap muslim dalam setiap ranah hidupnya.

2.2. Question Answering System

Menurut Hirschman dan Gaizauskas (2001), Question Answering System (QAS) adalah sistem yang dapat menganalisis pertanyaan dalam konteks interaksi yang sedang berlangsung, kemudian mencari satu atau lebih jawaban, dan memberikan jawaban kepada pengguna dalam bentuk yang layak. QAS dapat memberikan jawaban secara langsung, relevan, dan lebih akurat terhadap permintaan pengguna dibandingkan daftar dokumen atau URL layaknya pada mesin pencarian umum (Kashish,

et al., 2022). QAS dibedakan dalam dua kategori, yaitu sistem *domain-independent* dan *domain-specific*. Penelitian yang dilakukan menggunakan jenis QAS *closed-domain* atau *domain-specific* berdasarkan sifat dokumen fatwa yang terbatas pada *domain* agama Islam.

2.3. Retrieval Augmented Generation

Retrieval Augmented Generation (RAG) adalah sebuah *framework* yang menggabungkan memori parametrik dan non parametrik terhadap model generatif (Lewis, et al., 2021). RAG terdiri dari dua komponen utama, yaitu *retriever* dan *generator*. RAG memanfaatkan ruang penyimpanan vektor berisi *embedding* dari data dokumen yang dimasukkan sebagai basis data pencarian dokumen berdasarkan metode yang digunakan. Teks respon kemudian dihasilkan oleh *generator* sesuai dengan informasi tambahan.

2.4. Maximum Inner Product Search

Maximum Inner Product Search (MIPS) adalah permasalahan *searching* untuk mencari nilai tertinggi dari hasil perkalian antara dua vektor. MIPS direpresentasikan dalam pencarian nilai maksimum antara *embeddings* kueri dengan setiap nilai *embeddings* dari dokumen. Berbeda dengan *cosine similarity*, metode pencarian dalam MIPS memiliki kekurangan dalam biaya komputasi.

2.5. *Cosine Similarity*

Cosine Similarity adalah suatu ukuran jarak antara dua vektor berdasarkan nilai kosinus dari arah dua vektor. Ukuran ini memiliki kelebihan dalam perhitungan kemiripan antara dua dokumen dikarenakan sudut dari dua vektor yang merepresentasikan dokumen tersebut memiliki nilai yang sama dengan *dot product* dari kedua vektor (Singhal, n.d.). Berdasarkan mekanisme tersebut, pertimbangan dari ukuran dokumen memiliki kontribusi minimum terhadap hasil akhir dari nilai kemiripan terhadap vektor yang dibandingkannya. Rumus persamaan *cosine similarity* ditunjukkan pada Persamaan 1.

$$cos(A, B) = \frac{A \cdot B}{||A|| \cdot ||B||} \quad (1)$$

Keterangan:
A = Vektor dokumen A
B = Vektor dokumen B

2.6. *Confusion Matrix*

Confusion matrix adalah metrik kinerja yang digunakan pada tugas klasifikasi multi kelas untuk mengukur kinerja dari sistem bedasarkan nilai *precision*, *recall*, dan f1-score. Menurut Sokolova dan Lapalme (2009), *recall* adalah metrik kinerja model berdasarkan probabilitas prediksi model terhadap kelas, *precision* adalah metrik kinerja model berdasarkan probabilitas prediksi kelas positif dari data dengan kelas positif, dan f1-score adalah nilai rata-rata harmonik dari *precision* dan *recall* untuk menyatakan hubungan antar kelas berdasarkan data dan prediksi model. Penelitian yang dilakukan menggunakan dua jenis *confusion matrix* khusus untuk *retriever* dan *generator*.

2.6. *Confusion Matrix Retriever*

Evaluasi kinerja *retriever* menggunakan metrik Precision@K, F1-Score@K, dan Recall@K, dengan perbandingan antara sejumlah K dokumen S yang dikembalikan dengan dokumen relevan. Rumus persamaan Precision@K, Recall@K, dan F1-Score@K ditunjukkan pada Persamaan 2, 3, dan 4.

$$Precision@K = \sum_{i=1}^{k} \frac{rel(s_i)}{k} \quad (2)$$

Keterangan:
k = Jumlah dokumen
s_i = Dokumen i dari irisan dokumen S

$$Recall@K = \sum_{i=1}^{k} \frac{rel(s_i)}{|S|} \quad (3)$$

Keterangan:
k = Jumlah dokumen
s_i = Dokumen i dari irisan dokumen S
S = Himpunan dokumen terurut

$$F1\ Score@K = 2\frac{Precision@K \cdot Recall@K}{Precision@K + Recall@K} \quad (4)$$

2.7. *Confusion Matrix Generator*

Evaluasi kinerja *generator* menggunakan metrik BERTScore, yaitu metrik evaluasi yang menggunakan *embeddings* kontekstual untuk mengukur tingkat kesamaan antara teks keluaran dengan teks referensi dalam

tingkat semantik dibandingkan kemiripan dari pemetaan kata dengan kata dari metode tradisional (Zhang, et al., 2020). Rumus persamaan P-BERT, R-BERT, dan F-BERT dari BERTScore ditunjukkan pada Persamaan 5, 6, dan 7.

$$P_{BERT} = \frac{1}{|\hat{x}|}\sum_{\hat{x}_i \in \hat{x}} \max_{x_j \in x} x_i^\top \hat{x}_j \quad (5)$$

Keterangan:
x = Token teks referensi
$\hat{x}$ = Token teks luaran

$$R_{BERT} = \frac{1}{|\hat{x}|}\sum_{x_i \in x} \max_{\hat{x}_i \in \hat{x}} x_i^\top \hat{x}_j \quad (6)$$

Keterangan:
x = Token teks referensi
$\hat{x}$ = Token teks luaran

$$F_{BERT} = 2\frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}} \quad (7)$$

3. METODOLOGI PENELITIAN

3.1. Tahap Penelitian

Uraian metodologi penelitian terdiri atas rangkaian tahapan yang akan dilaksanakan dalam perancangan sistem QAS. Penelitian dimulai dengan melakukan studi kepustakaan mengenai QAS, rangka kerja RAG, metode pencarian *cosine similarity* dan MIPS, LLM, serta metode evaluasi kinerja dari komponen *retriever* dan *generator*. Tahap berikutnya adalah memeroleh dan mempersiapkan data *question-answer* fatwa yang akan digunakan dalam penelitian. Sistem kemudian dirancang berdasarkan tujuan dari penelitian dan diimplementasikan sesuai rancangan. Sistem yang telah dirancang kemudian diujikan untuk dilakukan evaluasi kinerja hingga menghasilkan analisis sebagai pemenuhan tujuan penelitian dan penunjang kesimpulan serta saran dari penelitian.

3.2. Studi Literatur

Tahap awal dari penelitian berupa penghimpunan sumber literatur terkait implementasi *pre-trained* model dengan *dataset* dokumen *question-answer* serta aplikasi dari teknik RAG pada model yang akan diimplementasi. Tahap berikutnya adalah perancangan QAS berdasarkan model tersebut untuk diujicoba dan dianalisis dalam segi kinerja *retriever* dalam mengembalikan dokumen relevan menggunakan Precision@K, Recall@K, dan F1-Score@K dengan metode *cosine similarity* dan MIPS. Evaluasi kinerja dari *generator* berupa LLM menggunakan *precision*, *recall*, dan f1-score berdasarkan rangka kerja BERTScore.

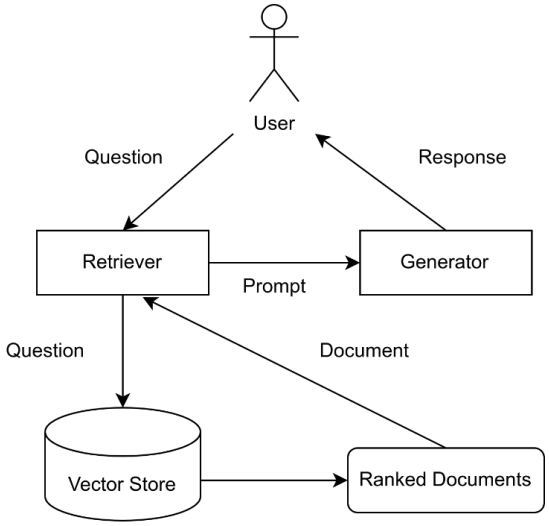
3.3. Persiapan Data

Pemerolehan data dilakukan dengan proses *scraping* pada situs islamqa.info versi bahasa Indonesia dari pasangan pertanyaan dan fatwa. Data yang diperoleh berjumlah 5556 berisi pasangan pertanyaan, fatwa terhadap pertanyaan, dan topik dalam format fail .pdf. Data tersebut kemudian diterapkan pra pemrosesan sebagai pembersihan tingkat *token*. Tahapan tersebut terdiri dari pengambilan teks dari dokumen .pdf, penghapusan tanda *newline*, penghapusan bahasa arab, penghapusan angka, dan melakukan normalisasi huruf kapital menjadi huruf kecil.

3.4. Rancangan Sistem

Sistem memiliki tipe *closed-domain* yang didasari pada rangka sistem QA rancangan Hairschman dan Gaizauskas (2001). Komponen dari sistem terdiri dari *retriever* yang berfungsi untuk mencari dokumen relevan terhadap kueri masukan dan *generator* yang berfungsi untuk menghasilkan jawaban berdasarkan dokumen yang dikembalikan *retriever*.

Kueri berupa pertanyaan yang dimasukkan oleh pengguna diubah menjadi *embeddings* oleh *sentence transformers*, sehingga dapat digunakan dalam proses *similarity search* oleh *retriever* terhadap dokumen yang tersedia pada ruang penyimpanan vektor. Hasil dari proses ini berupa dokumen terurut berdasarkan nilai kemiripan hasil pencarian menggunakan MIPS atau *cosine similarity* dengan kueri. Fatwa dari satu dokumen dengan nilai kemiripan tertinggi kemudian diambil untuk diberikan ke *generator* bersama pertanyaan dari pengguna dalam bentuk *prompt*, sehingga jawaban dapat dihasilkan. Rancangan sistem dapat dilihat pada Gambar 1.



Gambar 1. Diagram Rancangan Sistem

3.5. Rancangan Pengujian Sistem

Pengujian dilakukan terhadap dua bagian dari sistem, yaitu *retriever* dan *generator*. Pengujian terhadap *retriever* menggunakan 30% data yang tidak dimasukkan dalam ruang penyimpanan vektor dari sistem dan 70% data yang tersimpan pada ruang penyimpanan vektor. Pertanyaan dari data uji kemudian dijadikan kueri terhadap sistem dengan sejumlah *K*. Sistem menggunakan metode MIPS dan *cosine similarity* untuk mencari dokumen jawaban berdasarkan kueri yang diberikan. Topik dari *K* dokumen yang dikembalikan *retriever* dibandingkan dengan topik sesungguhnya dari kueri pada data uji. Pengujian terhadap *generator* dilakukan dengan memasukkan bagian jawaban dari seluruh data penelitian ke dalam ruang penyimpanan vektor dan menjadikan bagian pertanyaan sebagai kueri terhadap sistem. Sistem menghasilkan dua jenis jawaban, yaitu jawaban berdasarkan data pada ruang penyimpanan vektor dan jawaban tanpa menggunakan data pada ruang penyimpanan vektor untuk dibandingkan kemiripannya.

3.6. Analisis Hasil Pengujian Sistem

Analisis hasil pengujian *retriever* dilakukan dengan membandingkan kinerja dari metode MIPS dan *cosine similarity* pada *retriever* menggunakan metrik kinerja Recall@K, Precision@K, dan F1-Score@K. Analisis hasil pengujian *generator* dilakukan dengan mengukur nilai P-BERT, R-BERT, dan F-BERT dari kedua jenis jawaban yang dihasilkan terhadap bagian jawaban sesungguhnya pada data.

3.7. Data Pengujian

Data yang digunakan dalam proses pengujian memiliki jumlah 5556 data yang terdiri dari pasangan pertanyaan dan fatwa terhadap pertanyaan tersebut dengan jumlah topik sebanyak 14 topik. Data pengujian untuk *retriever* dipisah berdasarkan dua kategori, yaitu data tersimpan pada ruang vektor dan data uji. Pemisahan dilakukan secara acak dengan presentase 30% data pada setiap topik untuk data uji dan 70% data pada setiap topik untuk disimpan pada ruang vektor. Jumlah data untuk setiap topik ditunjukkan pada Tabel 1.

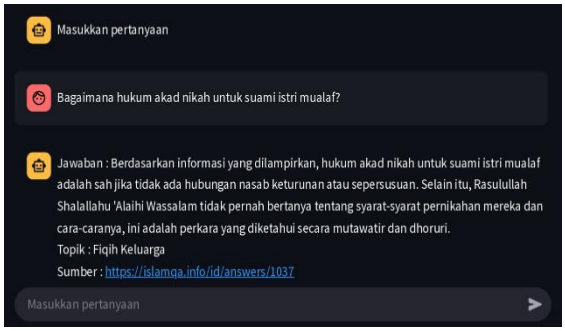
Tabel 1. Distribusi Data Sumber Pengujian berdasarkan Topik

Topik	Jumlah
Fiqih dan Usul Fiqih	3639
Fiqih Keluarga	554
Akidah	538
Adab, Akhlak Dan Pensucian Jiwa	261
Al-Qur'an dan Ilmu Al-Qur'an	151
Hadis dan ilmu-ilmunya	116
Sejarah dan Biografi	66
Ilmu dan Dakwah	64
Pendidikan	57
Politik Islam	35
Problematika Kejiwaan dan Sosial	32
Hak Penerbitan	4
Judi dan Undian	3
Tidak Didefinisikan	36

4. HASIL DAN ANALISIS

4.1. Implementasi Sistem

Penelitian ini mengimplementasikan sistem menggunakan program Python dengan antarmuka yang disediakan oleh pustaka *streamlit*. Antarmuka dari Implementasi sistem ditunjukkan pada Gambar 2.



Gambar 2. Antarmuka Implementasi Sistem dengan Contoh Pasangan Kueri serta Jawabannya

Operasi dari sistem dimulai dengan memberikan kueri berupa pertanyaan, kemudian sistem akan menelusuri ruang penyimpanan vektor untuk mencari dokumen dengan konten paling relevan terhadap kueri yang diberikan, hingga akhirnya menghasilkan jawaban berdasarkan konten dari dokumen yang ditemukan disertai dengan topik dokumen dan URL situs diunduhnya dokumen.

4.2. Hasil Pengujian Terhadap *Retriever*

Pengujian terhadap *retriever* dilakukan berdasarkan dua metode pencarian, yaitu *cosine similarity* dan MIPS. Data yang digunakan dalam pengujian adalah sub bagian pasangan pertanyaan dan jawaban dari data utama dengan porsi 30%. Skema pengujian dimulai dengan melakukan kueri berdasarkan pertanyaan yang belum terdapat jawabannya menggunakan metrik Precision@K, Recall@K, dan F1-Score@K. Skema ini didasari pada tujuan evaluasi kinerja kedua metode menggunakan *embeddings pre-trained* dalam mengembalikan dokumen yang relevan terhadap topik dari kueri tanpa kemiripan leksikal atau semantik sempurna. Jumlah *K* yang digunakan dalam pengujian memiliki nilai dalam rentang 5 hingga 25. Hasil pengujian dari kedua metode ditunjukkan pada Tabel 2, 3, 4, dan 5.

Tabel 2. Kinerja Macro Average Pengambilan Topik oleh Vector Store berdasarkan Maximum Inner Product Search

K	Macro Average		
	Precision	Recall	F1-Score
5	0.47	0.14	0.19
10	0.47	0.28	0.31
15	0.47	0.43	0.40
20	0.47	0.58	0.47
25	0.47	0.73	0.52

Tabel 3. Kinerja Weighted Average Pengambilan Topik oleh Vector Store berdasarkan Maximum Inner Product Search

K	Weighted Average		
	Precision	Recall	F1-Score
5	0.48	0.14	0.19
10	0.48	0.28	0.32
15	0.48	0.43	0.41
20	0.48	0.58	0.47
25	0.48	0.73	0.52

Tabel 4. Kinerja Macro Average Pengambilan Topik oleh Vector Store berdasarkan Cosine Similarity

K	Macro Average		
	Precision	Recall	F1-Score
5	0,47	0,14	0,19
10	0,47	0,28	0,31
15	0,47	0,43	0,40
20	0,47	0,58	0,47
25	0,47	0,73	0,52

Tabel 5. Kinerja Weighted Average Pengambilan Topik oleh Vector Store berdasarkan Cosine Similarity

K	Macro Average		
	Precision	Recall	F1-Score
5	0,48	0,14	0,19
10	0,48	0,28	0,32
15	0,48	0,43	0,41
20	0,48	0,58	0,47
25	0,48	0,73	0,52

4.3. Hasil Pengujian Terhadap *Generator*

Pengujian terhadap *generator* bertujuan untuk mengukur kinerja *generator* dalam menghasilkan teks dengan dokumen referensi sebagai *ground truth*. Data yang diujikan terhadap *generator* mencakup keseluruhan *dataset* pertanyaan dan fatwa. Hasil pengujian *generator* dengan dan tanpa informasi tambahan oleh *retriever* ditunjukkan pada Tabel 6 dan 7.

Tabel 6. Kinerja Penghasilan Jawaban oleh Generator dengan Prior Knowledge

Topik	P-BERT	R-BERT	F-BERT
Adab, Akhlak, dan Pensucian Jiwa	0.64	0.61	0.63
Akidah	0.64	0.61	0.63
Al-Qur’an dan Ilmu Al-Qur'an	0.64	0.61	0.62
Fiqih Keluarga	0.64	0.60	0.62
Fiqih dan Usul Fiqih	0.65	0.61	0.63
Hadis dan ilmu-ilmunya	0.63	0.59	0.61
Hak Penerbitan	0.61	0.67	0.64
Ilmu dan Dakwah	0.64	0.62	0.63
Judi dan Undian	0.61	0.61	0.61
Pendidikan	0.64	0.60	0.62
Politik Islam	0.64	0.63	0.63
Problematika Kejiwaan dan Sosial	0.63	0.60	0.61
Sejarah dan Biografi	0.65	0.60	0.62
Tidak Didefinisikan	0.64	0.60	0.62

Tabel 7. Kinerja Penghasilan Jawaban oleh Generator dengan Retrieved Knowledge

Topik	P-BERT	R-BERT	F-BERT
Adab, Akhlak Dan Pensucian Jiwa	0.66	0.66	0.66

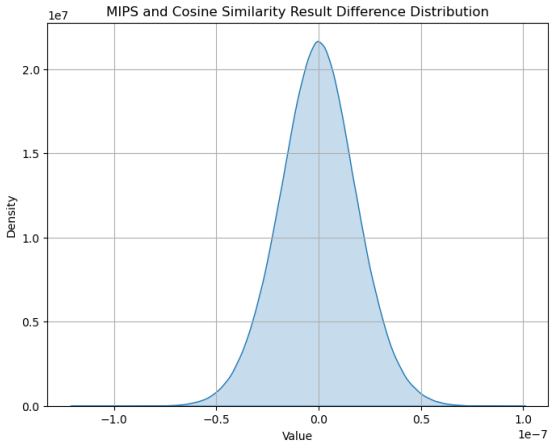
Akidah	0.59	0.59	0.59
Al-Qur'an dan Ilmu Al-Qur'an	0.62	0.62	0.62
Fiqih Keluarga	0.67	0.67	0.67
Fiqih dan Usul Fiqih	0.60	0.60	0.60
Hadis dan ilmu-ilmunya	0.63	0.63	0.63
Hak Penerbitan	0.68	0.68	0.68
Ilmu dan Dakwah	0.60	0.60	0.60
Judi dan Undian	0.63	0.63	0.63
Pendidikan	0.67	0.67	0.67
Politik Islam	0.59	0.59	0.59
Problematika Kejiwaan dan Sosial	0.63	0.63	0.63
Sejarah dan Biografi	0.68	0.68	0.68
Tidak Didefinisikan	0.59	0.59	0.59

Tabel 8. Rerata Makro Kinerja Generator pada Setiap Topik

Topik	P-BERT	R-BERT	F-BERT
Prior Knowledge	0,64	0,61	0,62
Retrieved Knowledge	0,67	0,59	0,63

4.4. Analisis Hasil Pengujian Pengambilan Topik dengan *Cosine Similarity* dan Maximum Inner Product Search

Hasil pengujian menunjukkan peningkatan nilai dari setiap metrik seiring dengan ditingkatkannya jumlah K dokumen yang dikembalikan. Kedua metode mencapai nilai terbaik dari seluruh metrik pada K dengan jumlah 25, salah satunya pada f1-score dengan nilai 0,52. Nilai metrik kinerja menunjukkan tidak adanya perbedaan antara metode MIPS dan *cosine similarity*. Distribusi dari selisih nilai kemiripan oleh MIPS dan *cosine similarity* dari kueri seluruh data uji terhadap data tersimpan ditunjukkan pada Gambar 3.

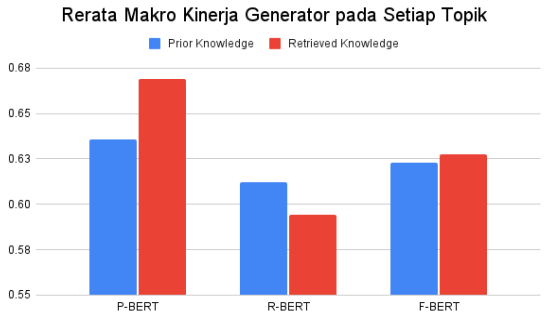


Gambar 3. Distribusi Selisih Nilai Kemiripan Metode Maximum Inner Product Search dan Cosine Similarity

Visualisasi distribusi menunjukkan bahwa selisih nilai kemiripan antara metode *cosine similarity* dan MIPS cenderung bernilai 0 dan pada rentang yang sangat kecil, yaitu 10^{-7}. Berdasarkan visualisasi tersebut, dapat disimpulkan bahwa kemiripan kinerja retriever menggunakan MIPS dan *cosine similarity* disebabkan oleh kecilnya nilai kemiripan antara kueri dengan data tersimpan yang dihasilkan oleh kedua metode, sehingga menunjukkan bahwa normalisasi terhadap hasil *dot product* pada *cosine similarity* antara kueri dengan vektor dokumen setelah melalui transformasi oleh *sentence transformer* tidak memiliki kontribusi signifikan pada perubahan nilai kemiripan.

4.4. Analisis Hasil Penghasilan Teks dengan *Retrieved* dan *Prior Knowledge*

Hasil pengujian terhadap *generator* pada Tabel 8 menunjukkan kecenderungan peningkatan nilai pada metrik kinerja P-BERT dan F-BERT ketika *generator* diberikan informasi tambahan berdasarkan dokumen yang dikembalikan oleh *retriever* dibandingkan penghasilan jawaban tanpa informasi tambahan, sehingga mengandalkan *pre-trained embeddings*. Grafik dari rerata perbandingan kinerja *prior knowledge* dan *retrieved knowledge* ditunjukkan pada Gambar 4.



Gambar 4. Rerata Makro Kinerja Generator pada Setiap Topik

Kinerja *generator* dengan penghasilan jawaban berdasarkan dokumen yang dikembalikan sebagai informasi tambahan atau *retrieved knowledge* lebih unggul dalam metrik P-BERT dengan nilai 0,67 dan F-BERT dengan nilai 0,63 dibandingkan penghasilan jawaban tanpa informasi tambahan atau *prior knowledge*. Pengaruh aplikasi *retriever* terhadap LLM membuat jawaban yang dihasilkan lebih unggul dalam nilai kemiripan terhadap dokumen fatwa

referensi, namun kurang mampu untuk melakukan generalisasi jawaban dibandingkan dengan penghasilan jawaban tanpa adanya *retriever*. Unggulnya nilai P-BERT pada jawaban dengan *retrieved knowledge* sesuai dengan kondisi ideal sistem untuk menghasilkan jawaban semirip mungkin dengan dokumen referensi fatwa dibandingkan menghasilkan jawaban secara umum, sehingga mempertahankan kesesuaian ketentuan dan hukum pada sumber fatwa dalam jawaban yang dihasilkan.

5. PENUTUP

5.1. Kesimpulan

Berdasarkan penelitian yang telah dilakukan pada penerapan pada rancangan sistem tanya jawab *closed-domain* terhadap dokumen fatwa menggunakan RAG dan LLM, dapat ditarik beberapa kesimpulan.

Pertama, Rancangan sistem QAS untuk dokumen fatwa berbahasa Indonesia harus mengimplementasikan metode pencarian pada *retriever* dan *generator* berupa LLM untuk menghasilkan jawaban berdasarkan dokumen yang dikembalikan. Nilai setiap metrik kinerja terbaik dari *retriever* dicapai dalam jumlah K sebanyak 25, salah satunya adalah nilai F1 Score@K yang berjumlah 0,52. Pemilihan salah satu dari kedua metode pencarian tidak akan memengaruhi sistem secara signifikan dikarenakan selisih nilai metrik kinerja kedua metode yang cenderung bernilai 0 dan dalam rentang 10^{-7}, sehingga sistem dapat mengimplementasikan *cosine similarity* maupun MIPS sebagai metode pencarian.

Kedua, Pengaruh metode RAG pada LLM ditunjukkan dari unggulnya nilai rerata makro dengan nilai 0,67 untuk P-BERT dan 0,63 untuk F-BERT. Hal ini disebabkan oleh penggunaan dokumen sumber yang dikembalikan oleh *retriever* sebagai penunjang proses penghasilan jawaban oleh *generator*. Keunggulan metrik kinerja P-BERT sesuai dengan kondisi ideal sistem untuk menghasilkan jawaban semirip mungkin dengan dokumen referensi fatwa dibandingkan menghasilkan jawaban secara umum.

5.2. Saran

Berdasarkan hasil penelitian, disarankan untuk melakukan ujicoba penelitian serupa dengan pendekatan rangka sistem sesuai yang merujuk pada penelitian pertama terkait RAG, yaitu dengan adanya Dense Passage Retriever (DPR) khusus berbahasa Indonesia sebagai *retriever* dan Bidirectional and Auto-regressive Transformers (BART) sebagai *generator* monolingual berbahasa Indonesia. *Dataset* fatwa juga dapat ditambahkan dari sumber lainnya selain islamqa.info, sehingga memperluas kemampuan sistem dalam menjawab pertanyaan yang berbeda-beda.

Penelitian selanjutnya dapat diarahkan ke perancangan metode yang tepat dalam sintesis jawaban dari sejumlah dokumen fatwa sebagai penunjang sistem yang lebih adaptif terhadap pertanyaan yang diberikan oleh pengguna.

6. DAFTAR PUSTAKA

Awass, O., 2019. Fatwa, Discursivity, and the Art of Ethical Embedding. Journal of the American Academy of Religion, 87(3), pp.765–790. https://doi.org/10.1093/jaarel/lfz031.

Hariri, F.R., 2021. Implementation of Fuzzy C-Means for Clustering the Majelis Ulama Indonesia (MUI) Fatwa Documents. Jurnal Online Informatika, 6(1), p.79. https://doi.org/10.15575/join.v6i1.591.

Hirschman, L. dan Gaizauskas, R., 2001. Natural language question answering: the view from here. Natural Language Engineering, 7(4), pp.275–300. https://doi.org/10.1017/S1351324901002807.

Kashish, P., Mohammed Arshad, S., N, S., and Department of Electronics and Communication BNMIT Bangalore, Karnataka State, India, 2022. VOICE ENBALED Q & A SYSTEM. International Journal of Engineering Applied Sciences and Technology, 7(7), pp.78–85. https://doi.org/10.33564/IJEAST.2022.v07i07.015.

Kim, S.H., Schramm, S., Wihl, J., Raffler, P., Tahedl, M., Canisius, J., Luiken, I., Endrös, L., Reischl, S., Marka, A., Walter, R., Schillmaier, M., Zimmer, C., Wiestler, B. dan Hedderich, D.M., 2024. Boosting LLM-Assisted Diagnosis: 10-Minute LLM Tutorial Elevates Radiology Residents' Performance in Brain MRI Interpretation.

https://doi.org/10.1101/2024.07.03.24309779.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S. dan Kiela, D., 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Available at: <http://arxiv.org/abs/2005.11401> [Diakses 27 September 2024].

Munshi, A.A., Al Sabban, W.H., Farag, A.T., Rakha, O.E., Al Sallab, A. dan Alotaibi, M., 2022. Automated Islamic Jurisprudential Legal Opinions Generation Using Artificial Intelligence. Pertanika Journal of Science and Technology, 30(2), pp.1135–1156. https://doi.org/10.47836/pjst.30.2.16.

Salim, D.P., 2022. Islamic political supports and voting behaviors in majority and minority Muslim Provinces in Indonesia. Indonesian Journal of Islam and Muslim Societies, 12(1), pp.85–110. https://doi.org/10.18326/ijims.v12i1.85-110.

Saragih, A.F.F., Sagala, R.F. dan Effendi, E., 2023. Peran Media Sosial Dalam Membangun Dakwah Islam yang Efektif. Khidmatussifa: Journal of Islamic Studies, 2(1), pp.31–41. https://doi.org/10.56146/khidmatussifa.v2i1.57.

Sebihi, A., Moazzam, A., 2024. ISLAM IN THE DIGITAL AGE: NAVIGATING FAITH AND TECHNOLOGY. EPRA International Journal of Research & Development (IJRD), pp.77–80. https://doi.org/10.36713/epra15075.

Singhal, A., n.d. Modern Information Retrieval: A Brief Overview.

Sokolova, M. dan Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), pp.427–437. https://doi.org/10.1016/j.ipm.2009.03.002.

Yung, H., 2003. Islam in South East Asia and Christian Mission. Transformation. [online] https://doi.org/10.1177/026537880302000406.

Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q. dan Artzi, Y., 2020. BERTScore: Evaluating Text Generation with BERT. Available at: <http://arxiv.org/abs/1904.09675> [Diakses 28 Oktober 2024]