JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.10021
Hal 1335-1347
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Implementasi Model Sentence-BERT untuk Peningkatan Semantic Retrieval pada Sistem Question Answering Layanan Helpdesk Ticketing

Adhityah Anugrah, Julian Supardi*

Fakultas Ilmu Komputer, Magister Ilmu Komputer, Universitas Sriwijaya, Kota Palembang, Indonesia
Email: [1]adhit.nugrah@gmail.com, [2,*]julian@unsri.ac.id
Email Penulis Korespondensi: julian@unsri.ac.id*
Submitted **23-07-2026**; Accepted **09-10-2026**; Published **31-10-2026**

Abstrak

Sistem Question Answering (QA) terutama closed-domain QA memberikan efisiensi dan konsistensi pada layanan helpdesk untuk menyajikan jawaban yang relevan dan akurat. Secara tradisional sistem QA sangat bergantung pada kesamaan kata antara query dan dokumen, sehingga rentan terhadap permasalahan lexical gap. Untuk mengatasi keterbatasan tersebut, pendekatan semantik dikembangkan guna menangkap kesamaan makna secara kontekstual. Penelitian ini bertujuan untuk mengevaluasi dan membandingkan kinerja metode retrieval berbasis leksikal, yaitu Term Frequency-Inverse Document Frequency (TF-IDF) dan Best Matching 25 (BM25), dengan pendekatan semantik menggunakan Sentence Bidirectional Encoder Representations from Transformers (SBERT) pada sistem closed-domain QA. Evaluasi dilakukan menggunakan dataset helpdesk yang dikumpulkan secara mandiri, serta divalidasi menggunakan dataset benchmark, seperti SQuAD v1.1, IDK-MRC, TechQA, dan COVID-QA. Kinerja sistem diukur menggunakan metrik information retrieval standar, meliputi Recall@k, Mean Reciprocal Rank (MRR), dan Normalized Discounted Cumulative Gain (NDCG). Hasil eksperimen menunjukkan bahwa SBERT memberikan kinerja yang sangat kompetitif dan secara umum mencapai hasil terbaik pada dataset yang dievaluasi. Pada dataset Helpdesk, SBERT mencapai Recall@1 99,814%, MRR 99,902%, dan NDCG@3 99,922%. Namun, pada dataset TechQA, TF-IDF dan BM25 menghasilkan Recall@1 yang lebih tinggi dibandingkan SBERT. Temuan ini menunjukkan keunggulan pendekatan semantik tidak selalu berlaku pada setiap dataset, terutama ketika pertanyaan dan jawaban memiliki kemiripan leksikal yang kuat. Secara keseluruhan, hasil penelitian menunjukkan SBERT merupakan pendekatan yang efektif dan kompetitif untuk sistem *closed-domain* QA.

Kata Kunci: Question Answering; Closed-Domain QA; Semantic Retrieval; SBERT; Helpdesk Systems

Abstract

Question Answering (QA) systems, particularly in closed-domain settings, provide efficiency and consistency in helpdesk services by delivering relevant and accurate answers. Traditionally, QA systems heavily rely on lexical similarity between queries and documents, making them vulnerable to the lexical gap problem. To address this limitation, semantic approaches have been developed to capture contextual meaning beyond exact word matching. This study evaluates the effectiveness of lexical-based retrieval methods, namely Term Frequency–Inverse Document Frequency (TF-IDF) and Best Matching 25 (BM25), with a semantic approach using Sentence Bidirectional Encoder Representations from Transformers (SBERT) in a closed-domain QA system. The evaluation is conducted using a self-collected helpdesk dataset and further validated on benchmark datasets, including SQuAD v1.1, IDK-MRC, TechQA, and COVID-QA. Performance is measured using standard information retrieval metrics, namely Recall@k, Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (NDCG). The experimental results show that SBERT achieved highly competitive performance and generally obtained the best results across the evaluated datasets. On the Helpdesk dataset, SBERT achieved a Recall@1 of 99.814%, an MRR of 99.902%, and an NDCG@3 of 99.922%. However, on the TechQA dataset, TF-IDF and BM25 achieved higher Recall@1 values than SBERT. These findings indicate that the advantages of semantic approaches do not necessarily apply to every dataset, particularly when the questions and answers exhibit strong lexical similarity. Overall, the results demonstrate that SBERT is an effective and competitive approach for closed-domain QA systems.

Keywords: Question Answering; Closed-domain QA; Semantic Retrieval; SBERT; Helpdesk Systems

1. PENDAHULUAN

Badan Penyelenggara Jaminan Sosial Ketenagakerjaan (BPJS Ketenagakerjaan) adalah badan hukum publik yang dibentuk berdasarkan Undang-Undang Nomor 24 Tahun 2011 untuk menyelenggarakan program jaminan sosial bagi pekerja Indonesia [1]. BPJS Ketenagakerjaan menggunakan *framework Information Technology Infrastructure Library* (ITIL) untuk membantu organisasi dalam pengelolaan layanan Teknologi Informasi (TI) [2]. Dalam mendukung praktik layanan IT digunakan *helpdesk ticketing* untuk manajemen insiden dan manajemen masalah [3], [4]. *Helpdesk ticketing* digunakan sebagai permintaan bantuan dari pelanggan kepada tim dukungan penyedia layanan, permintaan dilaporkan dalam bentuk teks yang dikenal dengan istilah tiket [5].

Penggunaan *helpdesk ticketing* dalam suatu organisasi dapat meningkatkan produktivitas, kualitas layanan yang lebih baik, serta kepuasan pelanggan yang lebih tinggi [6], [7]. Semakin meningkatnya kompleksitas sistem informasi, *helpdesk* telah berkembang pesat sebagai suatu mekanisme untuk membantu pengguna akhir (*end-user*) dalam menyelesaikan masalah yang dihadapi saat menggunakan teknologi informasi atau sistem informasi tertentu [8]. Masalah yang dilaporkan dalam bentuk tiket menggambarkan permintaan bantuan antara pengguna dan staf yang bertanggung jawab atas pengelolaan suatu layanan [5].

Masalah utama *helpdesk ticketing* terletak pada jumlah tiket yang terus meningkat, waktu tunggu yang lama, dan kurangnya sumber daya manusia [9]. Jumlah tiket sering meningkat secara eksponensial tergantung pada gangguan dari masalah tertentu yang akhirnya memperlambat waktu penyelesaian, tiket dilakukan analisis secara manual dan melihat

Copyright © 2026 The **Adhityah Anugrah**, Page 1335
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.10021
Hal 1335-1347
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

historis untuk memecahkan masalah tertentu [10]. Petugas *helpdesk* sepenuhnya bertanggung jawab untuk mengelola insiden besar, mulai dari identifikasi hingga penyelesaian dan penutupan, karena aspek manualnya, kesalahan manusia secara langsung berdampak pada efisiensi operasi [11]. Untuk mengatasi keterbatasan penanganan tiket secara manual, diperlukan suatu sistem yang mampu menyediakan jawaban secara konsisten berdasarkan historis tiket. Sistem *Question Answering* (QA) merupakan pendekatan otomatis untuk memberikan jawaban yang relevan dan tepat dalam kata-kata yang ringkas terhadap pertanyaan yang diajukan oleh manusia dalam bahasa alami [12], [13]. Sistem QA memungkinkan pengguna memperoleh jawaban yang telah divalidasi kebenarannya secara cepat tanpa perlu membaca dokumen panjang atau mencari secara manual di berbagai sumber informasi [14], [15], [16]. Sistem QA diklasifikasikan ke dalam dua kategori utama, yaitu sistem *open-domain* QA yang menangani tentang hampir semua pertanyaan yang tidak spesifik dan informasi universal seperti *World Wide Web*. Dan kategori kedua yaitu *closed-domain QA* yang menangani pertanyaan di bawah domain tertentu dengan topik spesifik, seperti musik, prakiraan cuaca, dll. [17], [18], [19]. Selain itu tantangan pada *closed-domain QA* adalah mengembangkan model yang hanya memerlukan kumpulan data kecil untuk pelatihan karena kumpulan korpus skala besar mungkin tidak tersedia, hal ini menyebabkan model yang baik pada *open-domain* QA akan kesulitan pada *closed-domain* QA [15].

Salah satu metode yang digunakan dalam sistem QA adalah *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *Best Matching* 25 (BM25) [17], yang menerapkan pendekatan kesamaan leksikal (*lexical similarity*) dengan menghitung frekuensi kemunculan kata antara pertanyaan dan dokumen tepat tanpa mempertimbangkan konteks atau makna kata. Sebaliknya, pendekatan kesamaan semantik (*semantic similarity*) seperti *Bidirectional Encoder Representations from Transformers* (BERT), berfokus pada makna dan konteks kata - kata dengan mempertimbangkan sinonim dan konteks penggunaannya [20]. Hal ini memungkinkan pemahaman yang lebih mendalam terhadap *query* pengguna melampaui pencocokan kata, sehingga metode semantik menjadi lebih cocok untuk menangani beragam masukan bahasa alami [21]. BERT telah menunjukkan kemampuan yang kuat dalam memodelkan kesamaan semantik dan telah diadopsi secara luas untuk menjembatani kesenjangan akibat ketidaksesuaian kosakata [22]. Berdasarkan [23], [24], BERT digunakan dalam berbagai tugas Natural Language Processing (NLP), seperti sistem *question answering*, *machine translation*, dan *sentiment analysis*. Namun, BERT tidak dirancang secara langsung untuk mengukur kemiripan antar kalimat, karena perhitungan *similarity* antara dua kalimat memerlukan inferensi silang (*cross-encoder*) yang mahal secara komputasi.

Penelitian yang dilakukan oleh [25] memperkenalkan Sentence-BERT (SBERT) sebagai metode untuk membentuk *embedding* kalimat menggunakan arsitektur *siamese* berbasis BERT. Dalam metode ini, setiap kalimat diubah menjadi vektor sehingga kesamaan antar kalimat dapat dihitung secara langsung menggunakan *cosine* similarity. Berbeda dengan model BERT standar yang memproses pasangan kalimat secara bersamaan, SBERT memungkinkan perhitungan kemiripan dilakukan dengan lebih efektif. Hasil eksperimen menunjukkan bahwa SBERT mencapai performa yang lebih unggul pada tugas kesamaan semantik efisien [26], [27]. Selain itu, [20] mengembangkan sistem pencarian semantik berbasis SBERT untuk mengakses basis data tertutup yang berisi dokumen klaim malapraktik medis di Jepang. Dalam penelitian tersebut, pertanyaan pengguna dan dokumen dalam basis data diubah menjadi *embedding* menggunakan SBERT, kemudian dicocokkan berdasarkan kesamaan makna menggunakan *cosine similarity*.

Pada penelitian ini, kinerja sistem *closed-domain* QA dianalisis menggunakan *dataset* yang bersumber dari *helpdesk ticketing* BPJS Ketenagakerjaan. Secara khusus, penelitian ini membandingkan pendekatan *lexical retrieval* (TF-IDF dan BM25) dengan *semantic retrieval* (SBERT). Berdasarkan hal tersebut, kontribusi penelitian ini dirumuskan ke dalam tiga hal utama. Pertama, penyusunan dan pemanfaatan data *helpdesk* dunia nyata sebagai representasi permasalahan *closed-domain* QA. Kedua, evaluasi perbandingan kinerja secara objektif antara metode *lexical retrieval* dan *semantic retrieval*. Ketiga, pelaksanaan evaluasi lintas *dataset* secara komprehensif. Penelitian ini menguji konsistensi metode *retrieval* tidak hanya pada data *helpdesk*, tetapi juga pada berbagai *dataset benchmark* publik lintas domain, yaitu SQuAD 1.1, IDK-MRC, TechQA, dan COVID-QA.

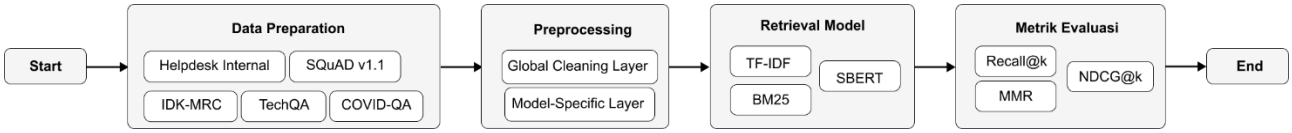
2. METODOLOGI PENELITIAN

2.1 Konteks Penelitian

Penelitian ini dilatarbelakangi oleh kebutuhan untuk meningkatkan efisiensi layanan helpdesk internal di BPJS Ketenagakerjaan yang menangani ribuan tiket layanan terkait sistem kepesertaan, klaim, iuran, dan aplikasi internal. Secara operasional, pengelolaan tiket dilakukan menggunakan platform ManageEngine ServiceDesk Plus sebagai sistem Information Technology Service Management (ITSM). Setiap tiket diproses oleh petugas helpdesk dan diselesaikan dengan solusi yang didokumentasikan di dalam sistem. Dalam praktiknya, petugas helpdesk sering menghadapi pertanyaan yang berulang dengan solusi yang sebenarnya telah tersedia pada tiket historis sebelumnya. Namun, solusi tersebut belum dimanfaatkan secara optimal karena keterbatasan mekanisme pencarian berbasis kata kunci. Oleh karena itu, penelitian ini memodelkan skenario retrieval-based question answering (QA), di mana tiket historis dimanfaatkan sebagai basis pengetahuan agar sistem dapat memberikan rekomendasi solusi secara cepat dan konsisten. Alur keseluruhan penelitian ditunjukkan pada Gambar 1. Seluruh dataset melalui proses global *cleaning* untuk memastikan konsistensi data sebelum diterapkan pada masing-masing model *retrieval*. Kinerja pendekatan yang diusulkan dievaluasi menggunakan metrik standar *information retrieval*, yaitu *Recall*, *Mean Reciprocal Rank* (MRR), dan *Normalized Discounted Cumulative Gain* (NDCG).

Copyright © 2026 The **Adhityah Anugrah**, Page 1336
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.10021
Hal 1335-1347
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom



Gambar 1. Alur penelitian

2.2 Data *Preparation*

Penelitian ini tidak hanya menggunakan dataset utama yang berasal dari sistem helpdesk BPJS Ketenagakerjaan, tetapi juga beberapa dataset benchmark eksternal untuk mengevaluasi konsistensi dan kemampuan generalisasi metode yang diusulkan. Penggunaan beberapa dataset bertujuan untuk memastikan bahwa model retrieval-based QA tidak hanya efektif pada dataset utama, tetapi juga mampu menunjukkan performa yang stabil pada dataset dengan karakteristik yang berbeda. Rincian dataset yang digunakan dalam penelitian ini disajikan pada Tabel 1.

Dataset utama yang digunakan dalam penelitian ini berasal dari sistem *helpdesk* internal BPJS Ketenagakerjaan, yang dikumpulkan selama periode Januari hingga Desember 2025 dengan mencakup tiket berstatus *closed* dan *resolved*. Mengingat data yang digunakan merupakan aset internal organisasi, penggunaannya telah mendapatkan izin resmi dari pihak organisasi. Meskipun anonimisasi atau pseudonimisasi tidak diterapkan pada *dataset* mentah, perlindungan terhadap informasi internal yang sensitif dilakukan melalui protokol kerahasiaan dan pengendalian akses yang ketat. *Dataset* disimpan dan diproses secara terisolasi dalam lingkungan tertutup dan hanya dapat diakses oleh pihak yang berwenang untuk keperluan penelitian dan evaluasi. Data mentah tidak dibagikan kepada pihak eksternal, tidak dipublikasikan, dan tidak didistribusikan melalui repositori publik mana pun. Dengan demikian, penggunaan *dataset* dibatasi pada tujuan penelitian sesuai dengan izin yang diberikan oleh organisasi.

Seluruh hasil penelitian dibatasi hanya pada metrik evaluasi kinerja secara agregat. Secara keseluruhan, *dataset* awal terdiri dari 11.215 tiket mentah yang mencakup atribut *title*, *description*, dan *resolution*. Data mentah tersebut kemudian melalui tahap *preprocessing* untuk diekstraksi menjadi pasangan pertanyaan jawaban yang terstandarisasi. Mengingat penelitian ini murni mengevaluasi kinerja metode *retrieval* tanpa melakukan pelatihan ulang model (*fine-tuning*), skema pembagian data latih dan data uji (*training-testing split*) tidak digunakan. Kinerja sistem dievaluasi secara menyeluruh dengan mengukur seberapa akurat model mampu menemukan dan menempatkan dokumen jawaban yang benar berdasarkan kecocokan *index* pasangan jawaban pertanyaan pada peringkat pencarian teratas untuk setiap *query* yang diberikan.

Skema evaluasi berbasis *retrieval* ini diterapkan secara konsisten pada beberapa *dataset* publik lintas domain. Pada dasarnya, setiap *dataset* publik tersebut memiliki struktur pembagian himpunan data latih (*training split*) dan data uji (*testing split*). Namun, penelitian ini secara khusus hanya mengekstrak data dari himpunan data latih pada masing-masing *dataset*. Pemilihan ini didasari oleh ketersediaan pasangan pertanyaan dan jawaban yang sudah terpetakan secara lengkap. Berdasarkan kriteria ekstraksi tersebut, pertama, digunakan 87.559 pasangan sampel dari Stanford Question Answering Dataset (SQuAD) v1.1 [28] sebagai *benchmark* berbahasa Inggris. Kedua, untuk mengakomodasi konteks bahasa Indonesia, dipilih 5.042 pasangan sampel valid dari Indonesian Machine Reading Comprehension (IDK-MRC) Dataset [29]. Ketiga, sebanyak 600 pasangan sampel diekstrak dari TechQA Dataset [30] untuk merepresentasikan skenario *closed-domain* teknis industri. Terakhir, *dataset* domain spesifik pada bidang medis diwakili oleh ekstraksi 2.019 pasangan sampel dari COVID-QA Dataset [31].

Tabel 1. Ringkasan *dataset* yang digunakan

Dataset	Jumlah Data	Bahasa	Sumber
Helpdesk Internal	11.215	Indonesia	Internal BPJS Ketenagakerjaan
SQUAD 1.1	87.559	Inggris	Wikipedia
IDK-MRC	5.042	Indonesia	General Articles
TechQA	600	Inggris	IBM Documentation
COVID-QA	2.019	Inggris	CORD-19

2.3 *Preprocessing*

Proses *text preprocessing* disusun dalam dua lapis, yaitu: *global cleaning* yang diterapkan ke seluruh pendekatan *retrieval* untuk memastikan konsistensi dan mengurangi *noise* umum, serta tahap *model-specific* yang disesuaikan dengan kebutuhan TF-IDF, BM25, dan SBERT. Desain berlapis ini bertujuan menjaga kompatibilitas lintas model sekaligus meminimalkan *information loss* pada teks domain teknis.

2.3.1 *Global Cleaning Layer*

Tahapan *global cleaning* diterapkan secara konsisten pada seluruh *dataset* yang digunakan dalam penelitian ini. Tujuan utama dari tahap ini adalah menghilangkan artefak non-linguistik dan inkonsistensi format tanpa mengubah struktur semantik utama dari teks. Dengan menerapkan *preprocessing* yang seragam di awal, pendekatan ini memastikan bahwa perbedaan performa antar-model *retrieval* murni dipengaruhi oleh kapabilitas arsitektural model tersebut, bukan akibat

Copyright © 2026 The **Adhityah Anugrah**, Page 1337
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.10021
Hal 1335-1347
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

bias dari perbedaan persiapan data. Hasil dari tahapan *global cleaning* ini dirangkum pada Tabel 2, yang menyajikan jumlah data sebelum dan sesudah *preprocessing* beserta atribut *dataset* yang digunakan.

Tahapan ini diawali dengan restrukturisasi format untuk membentuk kandidat pasangan pertanyaan-jawaban yang seragam. Pada *dataset* SQuAD v1.1, IDK-MRC, dan COVID-QA, pasangan QA dapat diekstraksi langsung dari atribut bawaannya. Sementara itu, pada *dataset* TechQA, atribut *subject* dan *description* digabungkan untuk membentuk representasi *query* yang lebih komprehensif. Pendekatan penggabungan serupa juga diterapkan pada *dataset helpdesk* internal, di mana atribut *title* dan *description* disatukan guna menangkap konteks pertanyaan pengguna secara lebih utuh.

Setelah struktur QA terbentuk, normalisasi teks dilakukan untuk meningkatkan konsistensi representasi data. Normalisasi karakter berbasis *Unicode Normalization Form Compatibility Composition* (NFKC) diterapkan untuk menyeragamkan karakter yang secara visual identik namun memiliki *encoding* berbeda. Selain itu, karakter tersembunyi seperti *zero-width spaces*, *zero-width joiners*, dan *byte order marks* (BOM) dihapus, mengingat artefak hasil *copy-paste* semacam ini berpotensi merusak proses tokenisasi. Pembersihan dilanjutkan dengan menghapus *tag* HTML untuk membuang artefak *markup* yang tidak memiliki nilai linguistik. Karakter titik berulang maupun simbol dekoratif juga dinormalisasi atau dieliminasi untuk mengurangi *noise* format struktural. Selanjutnya untuk memastikan konsistensi lebih lanjut, normalisasi *whitespace* diterapkan dengan mengonversi rentetan spasi, tab, dan *newline* menjadi satu spasi tunggal, diikuti dengan pemotongan (*trimming*) spasi di awal dan akhir teks. Sebagai langkah akhir dari normalisasi global, diterapkan *case folding* yang mengonversi seluruh teks menjadi huruf kecil (*lowercase*). Langkah ini bertujuan untuk menyeragamkan representasi kata dan meningkatkan stabilitas pencocokan, baik pada pendekatan *lexical* maupun *semantic retrieval*.

Secara keseluruhan, lapisan *global cleaning* ini secara efektif mereduksi *noise* tekstual akibat inkonsistensi *encoding* maupun tata letak. Dengan demikian, teks menjadi jauh lebih representatif untuk dilanjutkan ke proses tokenisasi, *indexing*, dan pembentukan *embedding*. Terakhir, pasangan pertanyaan jawaban yang teridentifikasi sebagai duplikat dihapus. Pembersihan duplikasi ini esensial untuk menjaga integritas *dataset*, mengurangi redundansi komputasi, serta memastikan bahwa metrik evaluasi nantinya merepresentasikan performa model secara objektif tanpa bias dari data yang berulang.

Tabel 2. Ringkasan dari tahapan *global cleaning*

Dataset	Jumlah Data Awal	Jumlah Data Setelah Global Cleaning	Struktur Data
Helpdesk Internal	11.215	8.581	*Title*, *Description*, dan *Answer*
SQUAD 1.1	87.559	87.310	*Question* dan *Answer*
IDK-MRC	5.042	4.971	*Question* dan *Answer*
TechQA	600	594	*Subject*, *Description*, dan *Answer*
COVID-QA	2.019	1.967	*Question* dan *Answer*

2.3.2 *Model-Specific Layer*

Setelah melalui tahapan *global cleaning*, teks diproses lebih lanjut menggunakan *pipeline* khusus yang disesuaikan dengan arsitektur representasi dan mekanisme penilaian masing-masing model pada Tabel 3. Perbedaan penyesuaian pada tahap ini dirancang bukan untuk menciptakan keuntungan yang tidak adil bagi model tertentu, melainkan justru untuk memastikan validitas dan keadilan evaluasi. Setiap algoritma memiliki cara kerja fundamental yang berbeda dalam mengukur relevansi teks. Memaksakan satu *pipeline* tokenisasi dan penyaringan yang identik untuk seluruh model justru akan mendegradasi performa model tertentu. Oleh karena itu, penyesuaian spesifik ini bertujuan agar setiap pendekatan dapat dievaluasi pada kondisi arsitektural yang paling optimal.

Pada model TF-IDF, karena algoritma ini murni bertumpu pada bobot frekuensi kemunculan kata, proses dilanjutkan dengan ekstraksi term yang lebih ketat. Berangkat dari teks bersih hasil *global cleaning*, tokenisasi dilakukan menggunakan pola *regular expression* yang disesuaikan dengan karakteristik istilah teknis. Untuk menangkap informasi kontekstual secara lebih baik, fitur teks dibentuk menggunakan kombinasi representasi unigram dan bigram. Selain itu, dilakukan *filtering term* dengan menghapus kata yang terlalu jarang muncul (min_df = 2) dan kata yang terlalu sering muncul (max_df = 0,90). Proses *stop word removal* juga diterapkan menggunakan pustaka Sastrawi (bahasa Indonesia) dan pustaka standar (bahasa Inggris). Rangkaian penyaringan tambahan ini sangat esensial bagi TF-IDF untuk menghilangkan *noise* leksikal dan mencegah dominasi kata umum yang tidak memiliki kontribusi semantik.

Sementara itu, meskipun sama-sama berbasis leksikal, *pipeline* spesifik untuk BM25 dikonfigurasi lebih sederhana. BM25 secara inheren telah memperhitungkan tingkat kepentingan term melalui mekanisme *inverse document frequency* (IDF) yang tersaturasi, di mana kata umum yang terlalu sering muncul otomatis memperoleh bobot penalti yang proporsional. Oleh karena itu, penghapusan *stop word* secara manual tidak diterapkan pada model ini agar tidak merusak keutuhan konteks leksikal asli. BM25 langsung menggunakan token hasil pemisahan berbasis *regular expression* dari teks *global cleaning*, dengan menerapkan *hyperparameter* standar yaitu k1 = 1,5 dan b = 0,75.

Sebaliknya, preprocessing spesifik pada SBERT dilakukan secara paling minimalis tanpa adanya manipulasi leksikal tambahan. Berbeda dengan TF-IDF yang memecah teks menjadi kumpulan kata acak (*bag-of-words*), SBERT memanfaatkan *contextual embedding* yang menghitung hubungan semantik antar kata berdasarkan urutan kalimat secara utuh. Menerapkan *stop word removal* atau penyaringan kosakata pada SBERT justru akan merusak struktur kalimat dan

Copyright © 2026 The **Adhityah Anugrah**, Page 1338
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.10021
Hal 1335-1347
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

menghilangkan informasi linguistik penting. Oleh sebab itu, dokumen hasil *global cleaning* langsung diubah menjadi representasi vektor semantik menggunakan model Multilingual MiniLM via pustaka SentenceTransformer dengan dimensi *embedding* sebesar 384. Pendekatan ini menjustifikasi bahwa keunggulan atau kelemahan SBERT pada pengujian ini murni merepresentasikan kemampuan pemahaman semantiknya yang *out-of-the-box*, bukan akibat intervensi manipulasi data

Tabel 3. Parameter yang digunakan pada TF-IDF, BM25, dan SBERT

Metode	*Parameter*	Nilai
TF-IDF	n-gram	(1,2)
	stop word	Indonesia dan Inggris
	similarity	Cosine
	min_df	2
	max_df	0,90
	Tokenization	*Regular Expression* (Regex)
BM25	k1	1,5
	b	0,75
	tokenization	*Regular Expression* (Regex)
SBERT	model	Multilingual MiniLM
	embedding	384
	similarity	Cosine

2.4 Metode *Retrieval*

Penelitian ini mengimplementasikan tiga pendekatan *retrieval-based* QA, yaitu TF-IDF, BM25**,** dan SBERT. Ketiga metode tersebut dipilih untuk merepresentasikan dua paradigma utama dalam *information retrieval*, yakni pendekatan berbasis *lexical matching* (TF-IDF dan BM25) serta pendekatan berbasis *semantic retrieval* (SBERT). Penelitian ini bertujuan melakukan evaluasi komparatif terhadap performa pencarian jawaban dalam konteks *dataset helpdesk* BPJS Ketenagakerjaan.

2.4.1 TF-IDF

TF-IDF merupakan metode pembobotan kata yang digunakan untuk merepresentasikan dokumen dalam bentuk vektor numerik. Metode ini menggabungkan dua komponen utama, yaitu *term frequency* (TF) *dan inverse document frequency* (IDF). *Term frequency* menggambarkan seberapa sering suatu kata muncul dalam sebuah dokumen, yang secara matematis dihitung sebagai rasio antara jumlah kemunculan kata $f_{t,d}$ terhadap total seluruh kata dalam dokumen tersebut. TF dihitung dengan menggunakan Persamaan (1).

$$TF(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \quad (1)$$

Sementara itu, *inverse document frequency* dihitung dengan membandingkan jumlah total dokumen N dengan jumlah dokumen yang mengandung kata t yaitu $df(t)$, yang dirumuskan pada Persamaan (2).

$$IDF(t) = \log \frac{N}{df(t)} \quad (2)$$

Bobot akhir TF-IDF diperoleh dari perkalian antara nilai TF dan IDF dapat dilihat pada persamaan di bawah ini:

$$TFIDF(t,d) = TF(t,d) \times IDF(t) \quad (3)$$

Dalam implementasinya, setiap dokumen direpresentasikan sebagai vektor TF-IDF dalam ruang vektor berdimensi tinggi. Relevansi antara *query* dan dokumen kemudian dihitung menggunakan *cosine similarity* antara vektor *query* dan vektor dokumen.

2.4.2 BM25

BM25 merupakan model *retrieval* probabilistik yang dikembangkan sebagai penyempurnaan dari TF-IDF. Metode ini memperhitungkan frekuensi kata, panjang dokumen, serta parameter normalisasi untuk menghasilkan estimasi relevansi dokumen yang lebih stabil sehingga mengurangi bias terhadap dokumen yang terlalu panjang serta menghindari dominasi kata yang sangat sering muncul. Skor relevansi antara dokumen *D* dan *query Q* dihitung dengan menjumlahkan kontribusi setiap kata dalam *query* menggunakan persamaan (4).

$$score(D,Q) = \sum_{t \in Q} IDF(t) \frac{f(t,D)(k_1+1)}{f(t,D) + k_1\left(1 - b + b \cdot \frac{|D|}{avgdl}\right)} \quad (4)$$

Pada persamaan tersebut, $f(t,D)$ menyatakan frekuensi kata t dalam dokumen D, $|D|$ merupakan panjang dokumen, dan $avgdl$ merupakan rata-rata panjang dokumen dalam koleksi. Parameter k_1 mengontrol pengaruh frekuensi kata terhadap skor relevansi, sedangkan parameter b digunakan untuk menyesuaikan normalisasi panjang dokumen. Nilai *inverse document frequency* pada BM25 dihitung menggunakan Persamaan (5) di bawah ini:

Copyright © 2026 The **Adhityah Anugrah**, Page 1339
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.10021
Hal 1335-1347
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

$$IDF(t) = \log \frac{N - df(t) + 0.5}{df(t) + 0.5} \quad (5)$$

2.4.3 SBERT

SBERT merupakan metode *semantic retrieval* yang digunakan untuk menghasilkan representasi vektor dari kalimat atau dokumen sehingga memungkinkan pencarian berdasarkan kedekatan makna. Berbeda dengan metode *lexical retrieval* seperti TF-IDF dan BM25 yang murni bergantung pada persinggungan kata, SBERT memetakan teks ke dalam ruang vektor semantik menggunakan arsitektur *transformer* yang telah melalui prapelatihan. Pada arsitektur ini, setiap teks masukan s diproses oleh *encoder* untuk menghasilkan *embedding* vektor v, yang dapat dinyatakan melalui Persamaan (6).

$$v = f_\theta(s) \quad (6)$$

di mana f_θ merupakan fungsi *encoder* dari model SBERT. Setelah *query* dan dokumen diubah menjadi *embedding* vektor, tingkat kesamaan antara keduanya dihitung menggunakan *cosine similarity* yang dirumuskan pada Persamaan (7). Untuk menjamin eksperimen dapat direproduksi, implementasi SBERT dibangun menggunakan pustaka SentenceTransformers versi 5.2.0 dengan memanggil *checkpoint* sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2, yang memproyeksikan teks ke dalam ruang vektor berdimensi 384.

$$sim(q, d) = \frac{q \cdot d}{||q||\,||d||} \quad (7)$$

Sebagai bentuk optimasi komputasi pada tahap implementasi, proses ekstraksi vektor dilakukan dengan menerapkan normalisasi L2 pada seluruh embedding dokumen dan *query*, sehingga magnitudo setiap vektor bernilai 1. Dengan normalisasi ini, perhitungan *cosine similarity* pada Persamaan (7) secara matematis disederhanakan menjadi perhitungan perkalian titik murni ($q \cdot d$). Oleh karena itu, representasi *embedding* dokumen diindeks menggunakan pustaka Facebook AI *Similarity Search* (FAISS) versi 1.13.1 dengan konfigurasi ruang pencarian IndexFlatIP. Kombinasi normalisasi *embedding* dan IndexFlatIP ini secara signifikan mempercepat komputasi pencarian dokumen yang paling relevan tanpa mengurangi presisi *cosine similarity*.

2.5 Metrik Evaluasi

Untuk mengukur kinerja sistem *retrieval-based QA*, penelitian ini menggunakan metrik evaluasi berbasis peringkat (*ranking-based metrics*), yaitu *Recall@*k**,** *Mean Reciprocal Rank* (MRR), dan *Normalized Discounted Cumulative Gain* (NDCG)**.** Pemilihan metrik ini didasarkan pada karakteristik sistem retrieval, di mana tujuan utama bukan hanya menemukan jawaban yang benar, tetapi juga menempatkannya pada posisi peringkat yang tinggi.

2.5.1 Recall@k

Recall@k mengukur kemampuan sistem dalam menampilkan jawaban relevan di antara *k* kandidat teratas. Secara formal, Recall@k didefinisikan pada Persamaan (8).

$$Recall@k = \frac{\text{Jumlah } query \text{ dengan jawaban relevan dalam } top \text{ k}}{\text{Total Query}} \quad (8)$$

Metrik ini bernilai 1 apabila jawaban yang benar muncul dalam *k* posisi teratas, dan 0 jika tidak. Dalam konteks sistem *helpdesk*, Recall@k penting untuk mengevaluasi apakah solusi yang tepat berhasil diretrieval dalam daftar kandidat awal yang akan ditinjau pengguna atau sistem lanjutan.

2.5.2 MRR

MRR mengukur posisi peringkat jawaban relevan pertama untuk setiap *query*. Untuk suatu *query* ke-*i*, nilai *reciprocal rank* dihitung pada Persamaan (9) di bawah ini:

$$RR_i = \frac{1}{rank_i} \quad (9)$$

di mana $rank_i$ adalah posisi jawaban relevan pertama. Nilai MRR merupakan rata-rata dari seluruh reciprocal rank yang dinyatakan pada Persamaan (10).

$$MMR = \frac{1}{N}\sum_{i=i}^{N}\frac{1}{rank_i} \quad (10)$$

dengan *N* adalah jumlah total *query*. MRR memberikan bobot lebih besar pada jawaban yang muncul pada peringkat teratas, sehingga metrik ini sensitif terhadap kualitas pemeringkatan awal.

2.5.3 NDCG

Metrik NDCG digunakan untuk mengevaluasi kualitas pemeringkatan dengan mempertimbangkan seberapa tinggi posisi dokumen relevan yang berhasil ditemukan. Nilai *Discounted Cumulative Gain* (DCG) pada peringkat ke- *k* dinyatakan pada Persamaan (11).

This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.10021
Hal 1335-1347
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

$$DCG@k = \sum_{i=1}^{k} \frac{rel_i}{log_i(i+1)} \quad (11)$$

di mana rel_i merepresentasikan skor tingkat relevansi kandidat dokumen pada peringkat ke-i. Mengingat seluruh pasangan teks pada *dataset* (baik data helpdesk maupun data publik) diformat sedemikian rupa sehingga setiap *query* hanya memiliki tepat satu dokumen jawaban yang relevan (divalidasi berdasarkan kecocokan *index*), maka skema relevansi yang diterapkan bersifat biner. Secara spesifik, nilai rel_i = 1 jika kandidat pada posisi ke-i adalah dokumen jawaban pasangannya yang tepat, dan nilai rel_i = 0 jika dokumen tersebut salah/tidak relevan. Nilai DCG yang diperoleh kemudian dinormalisasi terhadap nilai idealnya sebagaimana dinyatakan pada Persamaan (12).

$$NDCG@k = \frac{DCG@k}{IDCG@k} \quad (12)$$

Normalisasi ini memastikan nilai metrik selalu berada pada rentang 0 hingga 1. Karena batasan skema *dataset* menetapkan maksimal hanya ada satu dokumen yang relevan per *query*, maka pemeringkatan yang paling ideal adalah ketika dokumen jawaban tersebut langsung ditemukan pada peringkat pertama (i=1). Berdasarkan skema ini, nilai kalkulasi IDCG@k akan selalu konstan bernilai 1. Meskipun diimplementasikan pada relevansi biner tunggal, NDCG tetap sangat krusial dalam sistem *retrieval* ini karena secara proporsional memberikan penalti logaritmik yang semakin berat apabila sistem menempatkan jawaban yang benar pada urutan peringkat yang semakin rendah

2.6 *Platform*

Penelitian ini dilakukan menggunakan komputer dengan spesifikasi processor AMD Ryzen 7 5800H dengan Radeon Graphics (16 CPU, ~3.2 GHz) dan RAM 32 GB DDR4, yang menjalankan sistem operasi Windows 11 Pro. Implementasi sistem dikembangkan menggunakan bahasa pemrograman Python versi 3.10.6. Pada metode TF-IDF, penelitian ini memanfaatkan *library* scikit-learn, khususnya modul TfidfVectorizer, untuk membangun representasi vektor dokumen berdasarkan frekuensi kata dan *inverse document frequency*. Metode BM25 diimplementasikan menggunakan *library* rank-bm25 dengan kelas BM25Okapi untuk menghitung skor relevansi antara *query* dan dokumen. Sementara itu, metode *semantic retrieval* menggunakan model SBERT melalui *library* sentence-transformers, yang menghasilkan representasi *embedding* teks. Untuk mempercepat proses pencarian vektor *embedding*, penelitian ini menggunakan *library* FAISS. Selain itu, beberapa library pendukung seperti pandas, NumPy, re, unicodedata, serta modul penyimpanan model seperti joblib dan pickle digunakan untuk proses pengolahan data, preprocessing teks, dan penyimpanan model selama eksperimen berlangsung.

3. HASIL DAN PEMBAHASAN

Dalam menentukan kebenaran hasil pencarian (ground truth) untuk proses evaluasi, penelitian ini memanfaatkan pemetaan satu-ke-satu (one-to-one mapping) berdasarkan keselarasan indeks posisional antara himpunan pertanyaan dan himpunan dokumen jawaban. Secara formal, jika keseluruhan query direpresentasikan sebagai himpunan $Q = \{q_1, q_2, \ldots, q_n\}$ dan keseluruhan korpus jawaban direpresentasikan sebagai $A = \{a_1, a_2, \ldots, a_n\}$, maka sebuah query pada indeks ke-i (q_1) hanya memiliki keterkaitan relevansi mutlak dengan dokumen jawaban yang menempati urutan indeks yang persis sama (a_i). Oleh karena itu, pada saat sistem megambil sekumpulan kandidat dokumen untuk merespons *query* q_i, hasil pencarian tersebut diklasifikasikan sebagai kandidat yang relevan jika dan hanya jika sistem mengembalikan dokumen a_j dengan syarat $i = j$. Pendekatan evaluasi berbasis kecocokan indeks (*index matching*) ini menjamin objektivitas pengujian. Sebagai contoh, jika *query* berasal dari data urutan ke-100, maka target jawaban yang relevan secara eksklusif hanyalah dokumen pada indeks ke-100 di dalam korpus. Pendekatan otomatis berbasis kecocokan indeks ini menjamin bahwa seluruh proses evaluasi terhindar dari bias subjektif.

Kinerja masing-masing metode menggunakan matriks evaluasi Recall@k untuk k=1,2,dan 3 untuk mengukur kemampuan menemukan jawaban relevan pada sejumlah hasil teratas, NDCG@k untuk k=1,2, dan 3 untuk menilai kualitas urutan hasil dengan mempertimbangkan posisi relatif jawaban pada setiap tingkat k, serta MRR untuk mengevaluasi seberapa cepat jawaban benar pertama muncul dalam peringkat yang dihasilkan.

3.1 Evaluasi Dataset Helpdesk BPJS Ketenagakerjaann

Hasil evaluasi performa dari pendekatan *lexical retrieval* dan *semantic retrieval* pada *dataset* Helpdesk BPJS Ketenagakerjaan menggunakan model TF-IDF, BM25, dan SBERT dengan tujuh metrik evaluasi (Recall@1, Recall@2, Recall@3, MRR, NDCG@1, NDCG@2, dan NDCG@3) ditunjukkan pada Tabel 4. Secara umum, hasil tersebut menunjukkan bahwa model berbasis *semantic retrieval*, yaitu SBERT, secara konsisten memberikan performa terbaik dibandingkan pendekatan *lexical retrieval* (TF-IDF dan BM25) pada seluruh metrik yang digunakan. SBERT mencapai nilai Recall@1, Recall@2, dan Recall@3 masing-masing sebesar 99,814%, 99,977%, dan 99,988%, yang mengindikasikan kemampuan sangat tinggi dalam menemukan jawaban relevan, khususnya pada peringkat teratas. Hal ini diperkuat oleh nilai MRR sebesar 99,902%, yang menunjukkan bahwa dokumen relevan hampir selalu muncul pada posisi pertama dalam hasil pencarian. Dari sisi kualitas pemeringkatan, nilai NDCG@1, NDCG@2, dan NDCG@3 yang

Copyright © 2026 The **Adhityah Anugrah**, Page 1341
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.10021
Hal 1335-1347
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

mendekati sempurna (≥99,814%) mengindikasikan bahwa SBERT tidak hanya mampu menemukan dokumen yang relevan, tetapi juga mengurutkannya secara optimal sesuai tingkat relevansi.

Jika dibandingkan dengan pendekatan *lexical retrieval*, terlihat bahwa BM25 memberikan performa lebih baik dibandingkan TF-IDF pada seluruh metrik, yang menunjukkan bahwa mekanisme pembobotan probabilistik pada BM25 lebih efektif dalam menangani distribusi term dibandingkan TF-IDF yang berbasis frekuensi. Namun demikian, kedua metode *lexical* masih memiliki keterbatasan dalam menangkap hubungan semantik antar kata, sehingga performanya tetap berada di bawah SBERT. Keunggulan SBERT dapat dijelaskan dari kemampuannya dalam merepresentasikan teks ke dalam *dense vector embeddings* yang mempertimbangkan konteks dan makna semantik. Hal ini memungkinkan SBERT untuk mengatasi *lexical gap*, yaitu kondisi ketika pertanyaan dan jawaban menggunakan kata yang berbeda namun memiliki makna yang sama (misalnya sinonim atau parafrase). Sebaliknya, TF-IDF dan BM25 cenderung bergantung pada kesamaan kata secara eksplisit (*exact term matching*), sehingga lebih rentan gagal dalam kasus variasi bahasa alami. Lebih lanjut, selisih performa yang relatif kecil namun konsisten antara SBERT dan BM25 (misalnya Recall@1: 99,814% vs 98,835%) menunjukkan bahwa meskipun *dataset* memiliki tingkat kemiripan leksikal yang tinggi, pendekatan semantik tetap memberikan keuntungan tambahan dalam meningkatkan akurasi *retrieval*. Secara keseluruhan, hasil ini menegaskan bahwa pendekatan *semantic retrieval* berbasis SBERT lebih unggul dan stabil dibandingkan pendekatan *lexical retrieval*, baik dari sisi kemampuan menemukan jawaban relevan maupun kualitas pemeringkatan.

Tabel 4. Evaluasi pada *dataset* Helpdesk BPJS Ketenagakerjaan dalam persentase (%)

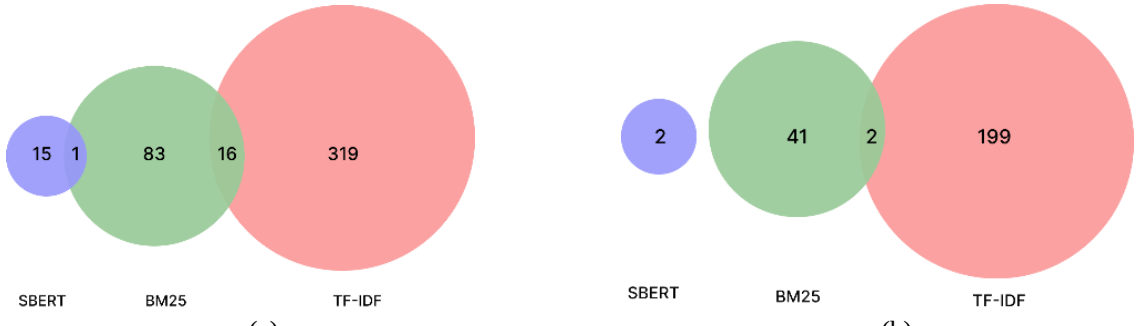
Metode	Recall			MRR	NDCG		
	@1	@2	@3		@1	@2	@3
TF-IDF	96,096	97,658	98,205	97,251	96,096	97,081	97,355
BM25	98,835	99,499	99,732	99,299	98,835	99,254	99,370
SBERT	99,814	99,977	99,988	99,902	99,814	99,917	99,922

Untuk memberikan gambaran performa, hasil Recall@1 masing-masing metode ditampilkan pada Tabel 5. Pada Tabel 5 Jumlah kegagalan dalam menemukan jawaban yang relevan pada peringkat teratas, yang direpresentasikan melalui jumlah kegagalan Recall@k. Secara umum, SBERT memiliki jumlah kesalahan paling rendah pada seluruh nilai k, yaitu 16 pada Recall@1, 2 pada Recall@2, dan hanya 1 pada Recall@3. Jumlah kegagakan Recall@k mengalami penurunan seiring dengan meningkatnya nilai k pada seluruh metode. Hal ini menunjukkan bahwa semakin banyak hasil yang dipertimbangkan, semakin besar kemungkinan jawaban yang relevan dapat ditemukan. SBERT menunjukkan kemampuan yang lebih baik dalam menempatkan jawaban yang relevan pada peringkat yang lebih tinggi.

Tabel 5. Perbandingan jumlah kegagalan retrieval atau miss recall@k (R@k) pada *dataset* Helpdesk

Metode	Miss R@1	Miss R@2	Miss R@3
TF-IDF	335	201	154
BM25	100	43	43
SBERT	16	2	1

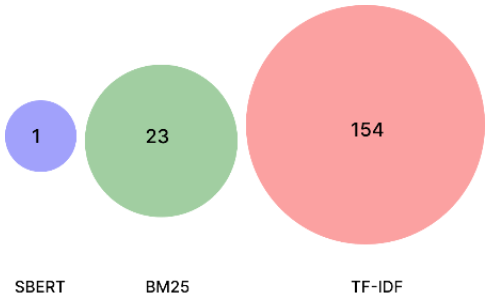
Distribusi kesalahan (*error overlap*) pada evaluasi *Recall@k* antara tiga metode *retrieval*, yaitu TF-IDF, BM25, dan SBERT ditunjukan pada Gambar 2. Melalui Gambar 2, Recall@1, TF-IDF memiliki jumlah kesalahan paling tinggi dengan 319 *query* yang hanya gagal pada metode tersebut, diikuti oleh BM25 dengan 83 *query*, sedangkan SBERT menunjukkan jumlah kesalahan paling rendah yaitu 15 *query*. Selain itu, terdapat 16 *query* yang gagal secara bersamaan pada TF-IDF dan BM25 dan 1 *query* gagal bersamaan pada BM25 dan SBERT. Sebaliknya, tidak adanya *overlap* antara TF-IDF dan SBERT serta tidak ditemukannya kasus yang gagal pada ketiga metode secara bersamaan. Pada Recall@2, TF-IDF 199 *query*, diikuti oleh BM25 41 *query*, sedangkan SBERT memiliki jumlah kesalahan dua *query* yang. Selain itu, hanya terdapat dua *query yang* gagal secara bersamaan pada TF-IDF dan BM25, sementara tidak terdapat overlap antara SBERT dengan metode lainnya maupun *overlap* pada ketiga metode sekaligus.Pada Recall@3, TF-IDF 154 *query*, diikuti oleh BM25 23 query, dan SBERT hanya satu *query* gagal. Tidak terdapat *overlap* kesalahan antar metode, yang mengindikasikan bahwa setiap metode gagal pada *query* yang berbeda.



(a) (b)

Copyright © 2026 The **Adhityah Anugrah**, Page 1342
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.10021
Hal 1335-1347
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom



(c)

Gambar 2. Diagram Venn hasil Recall@1 (a), Recall@2 (b), dan Recall@3 (c) pada dataset Helpdesk

3.2 Perbadingan dengan Dataset Lain

Untuk mengevaluasi konsistensi kinerja metode yang diusulkan, dilakukan pengujian pada beberapa *dataset* dengan karakteristik yang berbeda, yaitu COVID-QA, IDK-MRC, TechQA, dan SQuAD v1.1. Evaluasi lintas *dataset* ini bertujuan untuk menganalisis stabilitas performa TF-IDF, BM25, dan SBERT dalam berbagai konteks domain, serta mengidentifikasi apakah pola keunggulan dan kelemahan masing-masing metode tetap konsisten. Hasil evaluasi Recall@k dengan *dataset* lain dapat dilihat pada Tabel 6. Tabel 6 menunjukkan bahwa SBERT secara konsisten mencapai kinerja tertinggi pada seluruh dataset, dengan nilai Recall@1 berkisar antara 98.990% hingga 100,000%, dan bahkan mencapai 100,000% pada seluruh metrik Recall@k pada dataset COVID-QA, IDK-MRC, dan SQuAD v1.1. Sebaliknya, TF-IDF menunjukkan performa terendah dengan Recall@1 berkisar antara 89,761% (IDK-MRC) hingga 99,495% (TechQA), sementara BM25 memberikan peningkatan signifikan dibandingkan TF-IDF dengan rentang Recall@1 antara 95,510% hingga 99,495%. Selisih performa antara metode *lexical* dan *semantic* terlihat paling jelas pada dataset IDK-MRC, dimana SBERT mencapai 100.000% Recall@1, sedangkan BM25 dan TF-IDF masing-masing hanya mencapai 95,510% dan 89,761%. Pada dataset Helpdesk BPJS Ketenagakerjaan, SBERT juga unggul dengan Recall@1 sebesar 99,814%, dibandingkan BM25 (98,835%) dan TF-IDF (96,096%).

Selain itu hasil MRR dan NDCG@k pada dataset lain dapat dilihat pada Tabel 7. Tabel 7 menunjukkan bahwa SBERT secara konsisten mencapai kinerja tertinggi pada seluruh dataset berdasarkan metrik MRR dan NDCG@k, dengan nilai yang mendekati atau mencapai 100,000% pada hampir semua skenario, khususnya pada COVID-QA, IDK-MRC, dan SQuAD v1.1 yang seluruh metriknya bernilai sempurna. Sebaliknya, TF-IDF menunjukkan performa terendah dengan nilai MRR berkisar antara 92,659% (IDK-MRC) hingga 99,748% (TechQA), serta NDCG@1 terendah sebesar 89,761% pada IDK-MRC. BM25 memberikan peningkatan yang signifikan dibandingkan TF-IDF, dengan MRR antara 97,669% hingga 99,748% dan NDCG@1 antara 95,510% hingga 99,495%, namun masih berada di bawah SBERT. Perbedaan paling mencolok terlihat pada dataset IDK-MRC, dimana SBERT mencapai 100,000% pada seluruh metrik, sedangkan BM25 dan TF-IDF masing-masing hanya mencapai MRR sebesar 97,669% dan 92,659%, serta NDCG@1 sebesar 95,510% dan 89,761%. Pada dataset Helpdesk BPJS Ketenagakerjaan, SBERT juga unggul dengan MRR sebesar 99,902%, dibandingkan BM25 (99,299%) dan TF-IDF (97,251%).

Tabel 6. Hasil evaluasi Recall@k dengan *dataset* lain dalam persentase (%)

Dataset	Metode	Recall@1	Recall@2	Recall@3
Helpdesk Internal	TF-IDF	96,096	97,658	98,205
	BM25	98,835	99,499	99,732
	SBERT	99,814	99,977	99,988
COVID-QA	TF-IDF	94,662	97,153	97,661
	BM25	98,678	99,695	99,797
	SBERT	100,000	100,000	100,000
IDK-MRC	TF-IDF	89,761	94,307	95,152
	BM25	95,510	99,523	99,921
	SBERT	100,000	100,000	100,000
Tech-QA	TF-IDF	99,495	100,000	100,000
	BM25	99,495	100,000	100,000
	SBERT	98,990	99,832	100,000
SQuAD v1.1	TF-IDF	98,686	99,391	99,478
	BM25	99,451	99,962	99,984
	SBERT	100,000	100,000	100,000

Copyright © 2026 The **Adhityah Anugrah**, Page 1343
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.10021
Hal 1335-1347
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

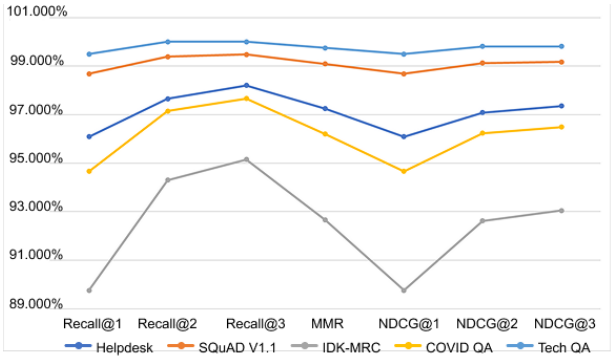
Tabel 7. Hasil evaluasi MRR dan NDCG@k dengan *dataset* lain dalam persentase (%)

Dataset	Metode	MRR	NDCG@1	NDCG@2	NDCG@3
Helpdesk Internal	TF-IDF	97,251	96,096	97,081	97,355
	BM25	99,299	98,835	99,254	99,370
	SBERT	99,902	99,814	99,917	99,922
COVID-QA	TF-IDF	96,196	94,662	96,234	96,488
	BM25	99,252	98,678	99,320	99,371
	SBERT	100,000	100,000	100,000	100,000
IDK-MRC	TF-IDF	92,659	89,761	92,629	93,052
	BM25	97,669	95,510	98,042	98,241
	SBERT	100,000	100,000	100,000	100,000
Tech-QA	TF-IDF	99,748	99,495	99,814	99,814
	BM25	99,748	99,495	99,814	99,814
	SBERT	99,467	98,990	99,521	99,605
SQuAD v1.1	TF-IDF	99,092	98,686	99,131	99,174
	BM25	99,717	99,451	99,774	99,785
	SBERT	100,000	100,000	100,000	100,000

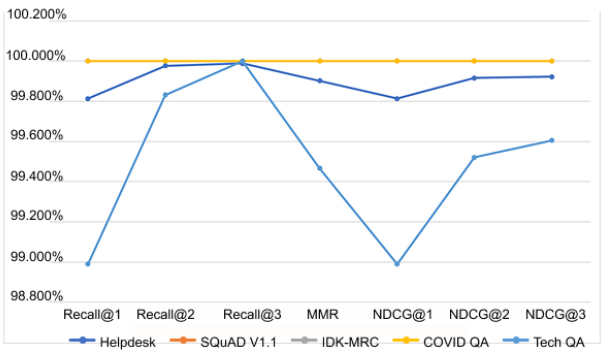
3.3 TF-IDF, BM25 dan SBERT

Visualisasi kinerja metode TF-IDF, BM25, dan SBERT pada berbagai *dataset* berdasarkan metrik Recall@k, MRR, dan NDCG@k disajikan pada Gambar 3, dimana Gambar 3(a), Gambar 3(b), dan Gambar 3(c) masing-masing merepresentasikan hasil evaluasi untuk TF-IDF, BM25, dan SBERT. Berdasarkan Gambar 3(a), terlihat bahwa TF-IDF menunjukkan variasi performa yang cukup besar antar *dataset*, dengan penurunan paling signifikan pada *dataset* IDK-MRC, dimana nilai Recall@1 hanya mencapai 89,761%. Pada Gambar 3(b), BM25 menunjukkan peningkatan performa dibandingkan TF-IDF, dengan nilai Recall@1 pada *dataset* IDK-MRC sebesar 95,510%, meskipun masih lebih rendah dibandingkan *dataset* lain yang mendekati 99%. Sementara itu, Gambar 3(c) menunjukkan bahwa SBERT memiliki performa yang paling stabil pada sebagian besar *dataset*, dengan nilai yang konsisten berada pada kisaran 99% hingga 100%. Namun demikian, pada dataset TechQA, SBERT menunjukkan performa yang sedikit lebih rendah dibandingkan TF-IDF dan BM25. Secara kuantitatif, pada *dataset* TechQA nilai Recall@1 SBERT sebesar 98,990%, lebih rendah dibandingkan TF-IDF dan BM25 yang masing-masing mencapai 99,495%, sehingga terdapat selisih sebesar 0,505 poin persentase.

Berdasarkan hasil evaluasi, penurunan performa pada *dataset* IDK-MRC disinyalir tidak berkaitan langsung dengan metrik panjang *query*, melainkan terindikasi dipengaruhi oleh karakteristik linguistik yang digunakan. Walaupun *query* pada *dataset* ini relatif pendek, formulasi pertanyaannya cenderung menampilkan variasi ekspresi yang dinamis, seperti penggunaan sinonim, perbedaan struktur kalimat, serta disparitas pemilihan kata (*lexical gap*) antara *query* dan dokumen. Kompleksitas semantik inilah yang diestimasikan menjadi faktor penyulit dalam proses pencocokan teks, yang pada gilirannya berkontribusi terhadap penurunan akurasi di peringkat teratas. Sebaliknya, pada *dataset* berdomain teknis seperti TechQA dan Helpdesk BPJS Ketenagakerjaan yang memiliki pola *query* yang lebih konsisten dan banyak mengandung istilah spesifik, seluruh metode menunjukkan performa yang relatif tinggi dan stabil, dengan nilai Recall@1 yang umumnya berada di atas 97%. Selain itu, pada ketiga visualisasi terlihat adanya peningkatan performa dari Recall@1 ke Recall@3, yang menunjukkan bahwa dokumen relevan umumnya masih dapat ditemukan dalam beberapa peringkat teratas, meskipun tidak selalu berada pada posisi pertama.



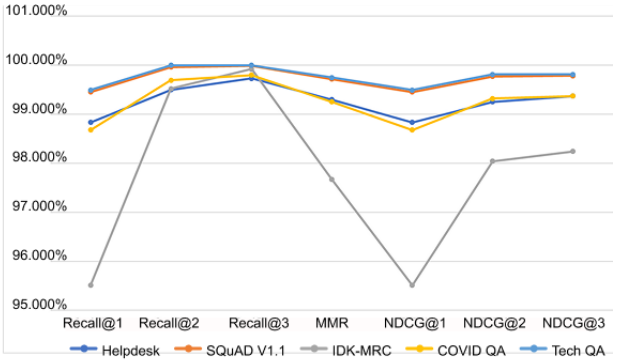
(a)



(b)

Copyright © 2026 The **Adhityah Anugrah**, Page 1344
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.10021
Hal 1335-1347
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

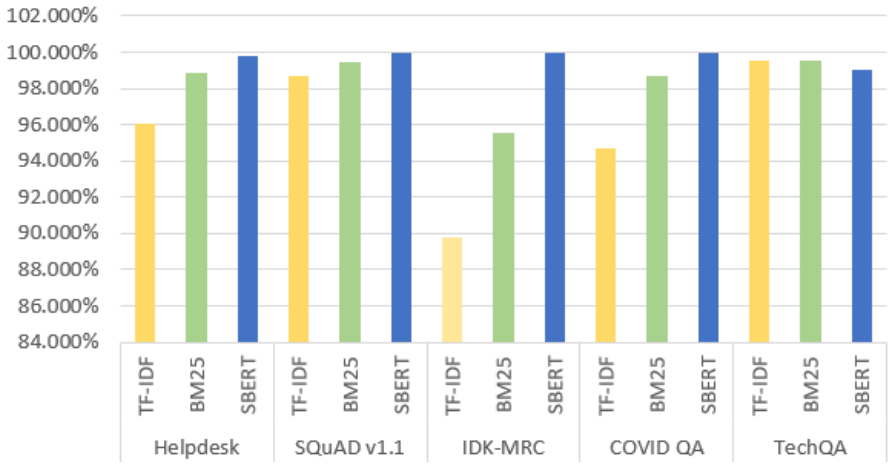


(c)

Gambar 3. Visualisasi kinerja metode retrieval pada berbagai dataset berdasarkan metrik Recall@k, MRR, dan NDCG@k, dimana (a) TF-IDF, (b) BM25, dan (c) SBERT

3.4 Perbandingan *lexical retrieval* dan *semantic retrieval*

Perbandingan kinerja antara metode TF-IDF, BM25, dan SBERT pada berbagai dataset berdasarkan metrik Recall@1 ditunjukkan pada Gambar 4. Visualisasi ini bertujuan untuk memberikan gambaran komprehensif mengenai perbedaan performa antara pendekatan berbasis kecocokan kata (*lexical retrieval*) dan pendekatan berbasis pemahaman makna (*semantic retrieval*). Berdasarkan Gambar 4, terlihat bahwa pendekatan *semantic retrieval* yang direpresentasikan oleh SBERT secara umum memberikan performa yang lebih tinggi dan stabil dibandingkan pendekatan *lexical retrieval* (TF-IDF dan BM25) pada sebagian besar *dataset*. Hal ini terlihat jelas pada *dataset* IDK-MRC, dimana SBERT mencapai nilai Recall@1 sebesar 100,000%, jauh lebih tinggi dibandingkan BM25 sebesar 95,510% dan TF-IDF sebesar 89,761%. Selisih ini menunjukkan bahwa SBERT mampu mengatasi variasi ekspresi bahasa yang tidak dapat ditangkap secara optimal oleh pendekatan berbasis kecocokan kata.



Gambar 4. Perbandingan evaluasi Recall@1 pada lima *dataset* yang digunakan

Pada dataset COVID-QA dan SQuAD v1.1, ketiga metode menunjukkan performa yang sangat tinggi, dengan SBERT tetap mempertahankan nilai sempurna (100,000%), sementara TF-IDF dan BM25 sedikit lebih rendah namun tetap berada di atas 94%. Hal ini menunjukkan bahwa ketika struktur pertanyaan dan jawaban relatif konsisten, kedua pendekatan dapat bekerja dengan baik, meskipun pendekatan berbasis pemahaman makna tetap lebih unggul.

Sebaliknya, pada dataset TechQA terlihat bahwa pendekatan *lexical retrieval* justru sedikit lebih unggul dibandingkan *semantic retrieval*. TF-IDF dan BM25 mencapai nilai Recall@1 sebesar 99,495%, sedangkan SBERT berada pada 98,990%, dengan selisih sebesar 0,505 poin persentase. Kondisi ini mengindikasikan bahwa pada domain yang memiliki terminologi teknis yang konsisten, kecocokan kata secara langsung masih memberikan keuntungan dibandingkan pemodelan berbasis representasi semantik.

Pada dataset Helpdesk BPJS Ketenagakerjaan, SBERT kembali menunjukkan performa tertinggi dengan nilai Recall@1 sebesar 99,814%, dibandingkan BM25 sebesar 98,835% dan TF-IDF sebesar 96,096%. Hal ini menunjukkan bahwa pada data dengan variasi bahasa yang lebih natural dan tidak terstruktur secara ketat, pendekatan berbasis pemahaman makna lebih efektif dalam menemukan dokumen yang relevan.

Secara keseluruhan, hasil ini menunjukkan bahwa pendekatan *semantic retrieval* memiliki keunggulan dalam menangani variasi bahasa dan ekspresi yang beragam, sedangkan pendekatan *lexical retrieval* tetap kompetitif pada

Copyright © 2026 The **Adhityah Anugrah**, Page 1345
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.10021
Hal 1335-1347
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

domain dengan struktur bahasa yang lebih konsisten dan berbasis terminologi. Temuan ini menegaskan bahwa pemilihan metode retrieval perlu mempertimbangkan karakteristik dataset yang digunakan, serta membuka peluang untuk mengombinasikan kedua pendekatan guna memperoleh performa yang lebih optimal.

4. KESIMPULAN

Penelitian ini menyajikan evaluasi komparatif terhadap tiga metode retrieval, yaitu TF-IDF, BM25, dan SBERT, pada berbagai *dataset* dengan karakteristik yang beragam. Hasil eksperimen menunjukkan bahwa SBERT sebagai pendekatan semantic retrieval secara umum memberikan performa terbaik dan paling stabil, khususnya pada *dataset* dengan variasi linguistik yang tinggi. SBERT memperoleh nilai Recall@1 hingga mencapai 100,000% pada beberapa *dataset*, seperti COVID-QA, IDK-MRC, dan SQuAD v1.1. Sementara itu, metode *lexical retrieval* seperti TF-IDF dan BM25 tetap menunjukkan performa yang kompetitif, bahkan sedikit lebih baik dibandingkan SBERT pada domain dengan terminologi teknis yang konsisten seperti TechQA. Temuan utama dari penelitian ini menunjukkan bahwa pendekatan *semantic retrieval* lebih *robust* dalam menangani variasi ekspresi dan makna, sedangkan metode *lexical retrieval* tetap efektif pada domain dengan tingkat kemiripan istilah yang tinggi. Selain itu, nilai evaluasi yang sangat tinggi pada beberapa *dataset* menunjukkan bahwa performa model sangat dipengaruhi oleh karakteristik ruang *retrieval* dan distribusi data. Oleh karena itu, evaluasi pada skenario yang lebih realistis menggunakan data terpisah atau data yang belum pernah dilihat sebelumnya diperlukan untuk menilai kemampuan generalisasi model secara lebih baik. Sebagai kontribusi utama, penelitian ini menyediakan analisis komprehensif mengenai *lexical retrieval* dan *semantic retrieval* pada berbagai domain serta menunjukkan bahwa kombinasi kedua pendekatan memiliki potensi untuk meningkatkan performa sistem *retrieval-based question answering*. Penelitian selanjutnya dapat difokuskan pada pengembangan metode *hybrid retrieval* yang menggabungkan keunggulan kedua pendekatan tersebut, serta evaluasi pada skenario *open-domain* untuk mengukur kemampuan generalisasi secara lebih mendalam.

REFERENCES

[1] Republik Indonesia, *Undang-undang (UU) Nomor 24 Tahun 2011 tentang Badan Penyelenggara Jaminan Sosial*. 2011. [Online]. Available: https://peraturan.bpk.go.id/Details/39268

[2] D. Wang, D. Zhong, and L. Li, "A comprehensive study of the role of cloud computing on the information technology infrastructure library (ITIL) processes," *Libr. Hi Tech*, vol. 40, no. 6, pp. 1954–1975, Dec. 2022, doi: 10.1108/LHT-01-2021-0031.

[3] A. K. Swain and V. R. Garza, "Key Factors in Achieving Service Level Agreements (SLA) for Information Technology (IT) Incident Resolution," *Inf. Syst. Front.*, vol. 25, no. 2, pp. 819–834, Apr. 2023, doi: 10.1007/s10796-022-10266-5.

[4] D. MacLean and R. Titah, "Implementation and impacts of IT Service Management in the IT function," *Int. J. Inf. Manage.*, vol. 70, p. 102628, Jun. 2023, doi: 10.1016/j.ijinfomgt.2023.102628.

[5] A. Zangari, M. Marcuzzo, M. Schiavinato, A. Gasparetto, and A. Albarelli, "Ticket automation: An insight into current research with applications to multi-level classification scenarios," *Expert Syst. Appl.*, vol. 225, p. 119984, Sep. 2023, doi: 10.1016/j.eswa.2023.119984.

[6] F. Al-Hawari and H. Barham, "A machine learning based help desk system for IT service management," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 33, no. 6, pp. 702–718, Jul. 2021, doi: 10.1016/j.jksuci.2019.04.001.

[7] B. M. Osman and A. M. S. Muse, "Enhancing Helpdesk Efficiency through Management Information System – A Resource Allocation Study in Technology Firms," *Qual. Innov. Prosper.*, vol. 28, no. 3, pp. 63–81, Nov. 2024, doi: 10.12776/qip.v28i3.2034.

[8] B. Bernárdez, J. I. Panach, J. A. Parejo, A. Durán, N. Juristo, and A. Ruiz-Cortés, "An empirical study to evaluate the impact of mindfulness on helpdesk employees," *Sci. Comput. Program.*, vol. 230, p. 102977, Aug. 2023, doi: 10.1016/j.scico.2023.102977.

[9] A. Revina, K. Buza, and V. G. Meister, "IT Ticket Classification: The Simpler, the Better," *IEEE Access*, vol. 8, pp. 193380–193395, 2020, doi: 10.1109/ACCESS.2020.3032840.

[10] S. S. Ali Zaidi, M. M. Fraz, M. Shahzad, and S. Khan, "A multiapproach generalized framework for automated solution suggestion of support tickets," *Int. J. Intell. Syst.*, vol. 37, no. 6, pp. 3654–3681, Jun. 2022, doi: 10.1002/int.22701.

[11] S. Ahmed *et al.*, "An Empirical Analysis of State-of-Art Classification Models in an IT Incident Severity Prediction Framework," *Appl. Sci.*, vol. 13, no. 6, p. 3843, Mar. 2023, doi: 10.3390/app13063843.

[12] M. A. Calijorne Soares and F. S. Parreiras, "A literature review on question answering techniques, paradigms and systems," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 32, no. 6, pp. 635–646, Jul. 2020, doi: 10.1016/j.jksuci.2018.08.005.

[13] A. Farea and F. Emmert-Streib, "Understanding question-answering systems: Evolution, applications, trends, and challenges," *Eng. Appl. Artif. Intell.*, vol. 156, p. 110997, Sep. 2025, doi: 10.1016/j.engappai.2025.110997.

[14] G. M. Biancofiore, Y. Deldjoo, T. Di Noia, E. Di Sciascio, and F. Narducci, "Interactive Question Answering Systems: Literature Review," *ACM Comput. Surv.*, vol. 56, no. 9, pp. 1–38, Oct. 2024, doi: 10.1145/3657631.

[15] M. A. Kia, A. Garifullina, M. Kern, J. Chamberlain, and S. Jameel, "Adaptable Closed-Domain Question Answering Using Contextualized CNN-Attention Models and Question Expansion," *IEEE Access*, vol. 10, pp. 45080–45092, 2022, doi: 10.1109/ACCESS.2022.3170466.

[16] A. Formica, I. Mele, and F. Taglino, "A template-based approach for question answering over knowledge bases," *Knowl. Inf. Syst.*, vol. 66, no. 1, pp. 453–479, Jan. 2024, doi: 10.1007/s10115-023-01966-8.

[17] K. Nassiri and M. Akhloufi, "Transformer models used for text-based question answering systems," *Appl. Intell.*, vol. 53, no. 9, pp. 10602–10635, May 2023, doi: 10.1007/s10489-022-04052-8.

Copyright © 2026 The **Adhityah Anugrah**, Page 1346
This Journal is licensed under a Creative Commons Attribution 4.0 International License

[18] Q. Guo, S. Cao, and Z. Yi, "A medical question answering system using large language models and knowledge graphs," *Int. J. Intell. Syst.*, vol. 37, no. 11, pp. 8548–8564, Nov. 2022, doi: 10.1002/int.22955.

[19] M. Laugiwa and E. Yulianti, "Question Answering through Transfer Learning on Closed-Domain Educational Websites," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 9, no. 1, pp. 104–110, Feb. 2025, doi: 10.29207/resti.v9i1.6163.

[20] N. Fujishiro, Y. Otaki, and S. Kawachi, "Accuracy of the Sentence-BERT Semantic Search System for a Japanese Database of Closed Medical Malpractice Claims," *Appl. Sci.*, vol. 13, no. 6, p. 4051, Mar. 2023, doi: 10.3390/app13064051.

[21] M. D. S. Antariksa, A. Sugiharto, and B. Surarso, "BERT Model Fine-tuned for Scientific Document Classification and Recommendation," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 9, no. 4, pp. 754–764, Aug. 2025, doi: 10.29207/resti.v9i4.6789.

[22] K. A. Hambarde and H. Proença, "Information Retrieval: Recent Advances and Beyond," *IEEE Access*, vol. 11, pp. 76581–76604, 2023, doi: 10.1109/ACCESS.2023.3295776.

[23] N. M. Gardazi, A. Daud, M. K. Malik, A. Bukhari, T. Alsahfi, and B. Alshemaimri, "BERT applications in natural language processing: a review," *Artif. Intell. Rev.*, vol. 58, no. 6, p. 166, Mar. 2025, doi: 10.1007/s10462-025-11162-5.

[24] L. R. Aini and Evi Yulianti, "Expertise Retrieval Using Adjusted TF-IDF and Keyword Mapping to ACM Classification Terms," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 9, no. 3, pp. 497–505, May 2025, doi: 10.29207/resti.v9i3.6397.

[25] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 3980–3990. doi: 10.18653/v1/D19-1410.

[26] K. Peyton and S. Unnikrishnan, "A comparison of chatbot platforms with the state-of-the-art sentence BERT for answering online student FAQs," *Results Eng.*, vol. 17, p. 100856, Mar. 2023, doi: 10.1016/j.rineng.2022.100856.

[27] J. Seo, S. Lee, L. Liu, and W. Choi, "TA-SBERT: Token Attention Sentence-BERT for Improving Sentence Representation," *IEEE Access*, vol. 10, pp. 39119–39128, 2022, doi: 10.1109/ACCESS.2022.3164769.

[28] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2016, pp. 2383–2392. doi: 10.18653/v1/D16-1264.

[29] R. A. Putri and A. Oh, "IDK-MRC: Unanswerable Questions for Indonesian Machine Reading Comprehension," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 6918–6933. doi: 10.18653/v1/2022.emnlp-main.465.

[30] V. Castelli *et al.*, "The TechQA Dataset," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 1269–1278. doi: 10.18653/v1/2020.acl-main.117.

[31] T. Möller, A. Reina, R. Jayakumar, and M. Pietsch, "COVID-QA: A question answering dataset for COVID-19," in *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, 2020. [Online]. Available: https://aclanthology.org/2020.nlpcovid19-acl.18/

This Journal is licensed under a Creative Commons Attribution 4.0 International License