Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No 2, April 2026, Page. 1886-1904
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5723

Job Recommendation for Fresh Graduates to Reduce Competency Gaps Using Content-Based Filtering and Retrieval-Augmented Generation

Iftitah Yanuar Rahmawati[1], Felda Mufarihati[2], Christian Sri Kusuma Aditya*[3]

[1,2,3]Informatics, Universitas Muhammadiyah Malang, Indonesia

Email: [3]christianskaditya@umm.ac.id

Received : Feb 7, 2026; Revised : Feb 25, 2026; Accepted : Feb 25, 2026; Published : Apr 18, 2026

Abstract

Job recommendation systems are frequently used to help job seekers find suitable positions. Nevertheless, many existing systems focus primarily on accuracy and provide limited justification. This lack of openness can erode user confidence, particularly among recent grads who need a clear explanation of how their individual experiences fit the recommendations. Furthermore, these systems frequently lack sophisticated methods to explain the reasoning behind the recommendations, such as Retrieval-Augmented Generation (RAG), which makes them seem impersonal and difficult to trust. The purpose of this research is to develop an explainable job recommendation system that generates employment suggestions based on language comprehension by integrating RAG and Content-Based Filtering (CBF). User profiles and open positions are displayed using TF-IDF and Sentence-BERT, and then the experience level-based cosine similarity is calculated. To measure competency coverage, matching and absent *skill*s are identified in an explicit *skill*-gap analysis. The Large Language Model and FAISS-based RAG modules leverage the explanations that are produced by finding matched and missing abilities as context. The CBF approach was used to evaluate recommendation relevance, while BLEU and ROUGE on ten test documents were used by HR specialists for validation. The system's mean ROUGE-1 F1 score was 0.4659, and its mean ROUGE-L score was 0.2918, based on 10 evaluation cases. Results show that the proposed recommendation system provides accurate and adequate guidelines based on HR references. This paper enriches Informatics by consolidating semantic similarity modeling, explicit competency-gap reasoning, and grounded text generation together to form a cohesive explainable recommendation framework targeted to cold-start job seekers.

Keywords : content-based filtering, explainable AI, retrieval-augmented generation, sentence-BERT, *skill* gap analysis

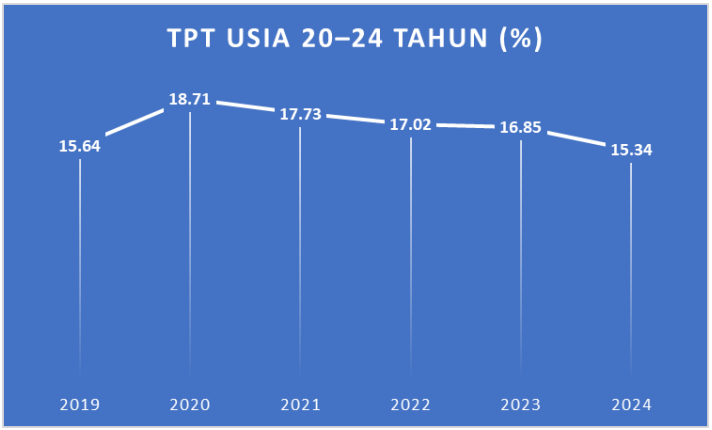
1. PENDAHULUAN

Dalam bidang ketenagakerjaan di wilayah Indonesia masih menghadapi berbagai masalah besar ketika lulusan baru tidak memenuhi kebutuhan pasar kerja. Disebabkan keahlian mereka belum sepenuhnya mampu menyesuaikan dengan kebutuhan perusahaan, banyak lulusan perguruan tinggi menghadapi kesulitan dalam persaingan [1], [2], [3]. Sebagaimana mestinya, kurikulum tidak sesuai dengan kebutuhan pasar kerja, perubahan ke dunia kerja menjadi lambat dan tidak berhasil [4]. Tingkat pengangguran terbuka atau dapat disingkat TPT nasional sebesar 4.91% pada tahun 2024, dengan 15.34% pada kelompok usia 20 hingga 24 tahun, yang sebagian besar adalah lulusan baru [5]. Data berikut telah memperlihatkan dampak TPT untuk kelompok usia 20 hingga 24 tahun dari tahun 2019 hingga 2024, yang menggambarkan naik turunnya angka pengangguran di kalangan *fresh graduate.*

Grafik pada Gambar 1 telah memperlihatkan dampak pengangguran di kalangan usia 20 hingga 24 tahun dari tahun 2019 hingga 2024. TPT menurun, melaikan angka pengangguran tetap tinggi, terutama di tahun 2020. Ini menunjukkan bahwa lulusan baru menghadapi tantangan besar dalam mencari pekerjaan yang sesuai dengan kemampuan mereka.

Vol. 7, No 2, April 2026, Page. 1886-1904
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5723

Sebaliknya, sejumlah besar pekerjaan tersedia melalui platform pencarian kerja online. Tidak hanya situasi ini memberi pencari kerja lebih banyak pilihan, tetapi juga menyebabkan banyak informasi, terutama bagi mahasiswa baru yang belum memiliki banyak pengalaman [6], [7]. Sebagian besar platform hanya menampilkan informasi deskriptif dan daftar lowongan, tanpa memberikan konteks yang jelas tentang seberapa dekat profil kompetensi seseorang dengan persyaratan pekerjaan [8]. Akibatnya, lulusan baru sering mengalami kesulitan menemukan pekerjaan yang benar-benar mereka butuhkan, sementara perusahaan juga sulit mendapatkan kandidat yang memenuhi persyaratan khusus pekerjaan. Ini menunjukkan bahwa sistem rekomendasi pekerjaan yang akurat dan jelas [8], [9], [10].



Gambar 1. Tingkat Pengganggguran (TPT) Usia 20-24, 2019-2024.
Sumber : Diolah dari BPS data [5] (Diakses Jan 2026)

Sistem rekomendasi pekerjaan dapat dibuat dengan berbagai cara. Untuk meningkatkan transparansi rekomendasi pekerjaan, Upadhyay dkk. [11] menyarankan *job posting explainable* yang berbasis *Knowledge Graph* dan *Named Entity Recognition* (NER). Namun, metode ini memiliki banyak masalah dengan pengolahan data dan tidak praktis untuk penggunaan skala besar. Hussein dkk. [12] menggunakan SVD++ untuk meningkatkan akurasi dengan mempertimbangkan hubungan sosial eksplisit dan implisit. Namun, metode ini bergantung pada riwayat interaksi pengguna, jadi tidak cocok untuk siswa baru yang belum memiliki pengalaman kerja atau interaksi sebelumnya. Untuk menyesuaikan profil pengguna dan deskripsi pekerjaan, Ro'uf dkk. [13] mengembangkan metode berdasarkan kemiripan konten dengan *Multilayer Perceptron* (*MLP Classifier*). Model yang dihasilkan sering kali bersifat kotak hitam dan tidak memberikan penjelasan yang jelas tentang rekomendasi. Penjelasan yang dapat dipahami pengguna adalah penting, menurut penelitian tentang sistem rekomendasi yang dijelaskan dan kecerdasan buatan yang dijelaskan (XAI). Namun, evaluasi penjelasan masih banyak dilakukan secara kualitatif dan belum secara sistematis disejajarkan dengan penilaian pakar domain [1]. Namun, evaluasi penjelasan masih banyak dilakukan secara kualitatif dan belum disejajarkan secara sistematis dengan penilaian pakar domain seperti praktisi *Human Resource* (HR). selain itu, penelitian yang mengintegrasikan pencocokan semantik, analisis kesenjangan keterampilan, dan generasi penjelasan berbasi *Retrieval-Augmented Generation* (RAG) yang divalidasi oleh pakar HR masih terbatas [14], [15], [16], [17].

Dalam pratik rekrutmen, praktisi *Human Resource* (HR) memainkan peran penting dalam menilai relevansi kandidat untuk posisi berdasarkan kombinasi pengalaman, kualifikasi formal, dan keterampilan. Agar sistem menjadi akurat secara komputasional dan dapat diterima dalam pengambilan keputusan nyata, sangat penting untuk memberikan penjelasan rekomendasi yang sesuai dengan pola penilaian HR [8], [9]. Sebuah studi terkini juga menegaskan bahwa penerapan model dasar/LLM dalam prosedur penerimaan pegawai harus dibarengi dengan perangkat mekanisme klarifikasi sehingga hasil

Vol. 7, No 2, April 2026, Page. 1886-1904
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5723

dari sistem menjadi lebih gamblang, dapat diuji, serta memperkuat keyakinan dari para pemakai dan praktisi Sumber Daya Manusia [18]. Dengan memanfaatkan representasi teks seperti TF-IDF, Word2Vec, dan BERT *embedding*, *Content-Based Filtering* (CBF) dapat mencocokkan *profile* pengguna dengan deskripsi pekerjaan sehingga mampu mengidentifikasi kemiripan leksikal [19]. Di sisi lain, metode *Retrieval-Augmented Genration* (RAG) memungkinkan integrasi antara komponen *retrieval* dan model generatif untuk menghasilkan penjelasan naratif relevan [20], [21], [22]. Pendekatan CBF juga tetap efektif pada skenario *cold-start* karena tidak bergantung pada riwayat interaksi pengguna [23], sedangkan konsep *rationalization* dalam RAG memungkinkan sistem menjelaskan alasan di balik rekomendasi berbasis teks [6].

Penelitian ini menyajikan suatu metode rekomendasi lowongan yang memadukan *Content-Based Filtering* berbasis representasi semantik Sentence-BERT dengan analisis kesenjangan keterampilan melalui identifikasi *matched* dan *missing skill*, serta modul generasi penjelasan berbasis RAG [19], [24], [25]. Integrasi ini memungkinkan sistem merepresentasikan kesesuaian kandidat tidak hanya dari aspek kata kunci, tetapi juga makna dan kompetensi secara simultan. Dari perspektif Informatika dan Ilmu Data, penelitian ini berkontribusi pada pengembangan kerangka kerja sistem rekomendasi pekerjaan berbasis teks yang dapat dijelaskan dan diukur melalui evaluasi kuantitatif menggunakan metrik BLUE dan ROUGE yang divalidasi oleh pakar HR [15], [16], [17]. Ini menghasilkan nilai kuantitatif tentang seberapa sesuai penjelasan sistem dengan penilaian profesional. Oleh karena itu, diharapkan bahwa penelitian ini akan menghasilkan sistem rekomendasi pekerjaan yang akurat, dapat dijelaskan, dan berfokus pada kebutuhan karyawan baru. Ini juga akan membantu mengurangi disparitas kompetensi dan meningkatkan penyerapan tenaga kerja muda di Indonesia [5], [20], [26].

2. METODE PENELITIAN

Metode pada penelitian ini disusun dengan mengadaptasi rancangan metodologi yang telah dirumuskan pada proposal pra-skripsi, kemudian disesuaikan dengan implementasi sistem dan eksperimen yang telah dilakukan. Metodologi penelitian dirancang untuk memastikan bahwa sistem rekomendasi pekerjaan yang dikembangkan tidak hanya relevan bagi *fresh graduate*, tetapi juga mampu memberikan penjelasan yang dapat dipahami dan sejalan dengan cara penilaian *Human Resource* (HR) [1], [8], [9], [23].

2.1. Arsitektur Sistem

Sistem rekomendasi pekerjaan *explainable* yang diusulkan dirancang dalam beberapa tahapan terintegrasi, yaitu: (1) pengumpulan dan praproses data, (2) deteksi *fresh graduate*, (3) pembentukan profil dan vektorisasi, (4) *content-based filtering* dan pemeringkatan, serta (5) generasi penjelasan berbasis Retrieval-Augmented Generation (RAG) dan evaluasi. Data bersumber dari dua dataset open source, yaitu LinkedIn Job Postings (±124.000 lowongan) dan Resume Dataset (Structured) (±54.000 resume) yang berisi informasi pendidikan, keterampilan, dan pengalaman kandidat [12]. Dalam *Job Recommendation System* modern, pendekatan berbasis profil teks (resume-lowongan) umum digunakan karena memungkinkan pencocokan kandidat dengan pekerjaan tetap berjalan walaupun riwayat interaksi pengguna terbatas (*cold-start*) [6], [10].

Tabel 1 menyajikan lima contoh mengenai profil pengguna yang telah dikelompokkan sebagai lulusan baru melalui sistem penilaian multi-sinyal. Karakteristik yang ditunjukkan mencakup total pengalaman kerja, pengalaman magang, status sebagai pegawai senior, serta keterampilan dan kemampuan yang telah melalui tahap pembersihan teks. Data tersebut dimanfaatkan untuk menyusun teks profil terpadu yang kemudian akan diproses dalam langkah vektorisasi dan penghitungan kesamaan sematik.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No 2, April 2026, Page. 1886-1904
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5723

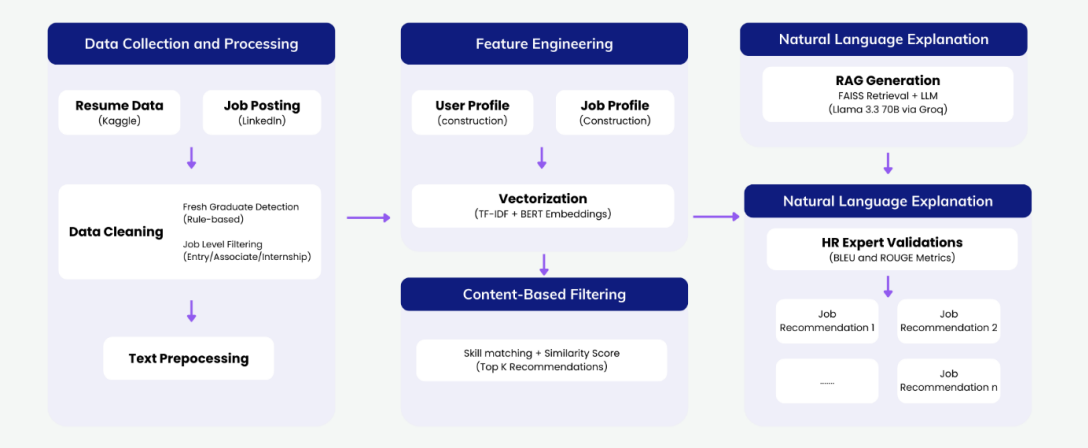
Tabel 1. Sample dari User Profiles (Fresh Graduate Subset)

Person ID	Fresh Graduate	n_jobs	Has Intern	Has Senior Title	Cleaned *Skill*s	Cleaned Abilities
132	True	3	False	False	Oracle Golden Gate, Perl Scripting, Database Administration, Troubleshooting, Performance Tuning	Oracle Golden Gate support, replication, automation, performance tuning
166	True	2	False	False	SQL, HTML, JavaScript, Python	Programming, analytics, problem-solving
396	True	5	False	False	Social Media, WordPress, SEO, Adobe Illustrator, Graphic Design	Website management, digital marketing, analytics

Tabel 2. Sampel dari Job Postings Dataset

Job_ID	Job Title	Company	Location	Required *Skill*s	Experience Level
921716	Marketing Coordinator	Corcoran Sawyer Smith	Princeton, NJ	Marketing, Sales, Adobe Creative Cloud, Multitasking	Entry/Not Specified
1829192	Mental Health Therapist	-	Fort Collins, CO	Health Care Provider, Counseling, MSW, LPC	Entry/Not Specified
10998357	Assistant Restaurant Manager	The National Exemplar	Cincinnati, OH	FOH Management, Customer Service, Communication	Entry/Not Specified

Tabel 2 memperlihatkan lima contoh posisi pekerjaan yang diambil dari dataset Linkedln Job Postings. Kolom *Required Skill* adalah kombinasi antara persyaratan pekerjaan dan tag keterampilan yang digunakan dalam analisis kesenjangan keterampilan. Data ini sangat penting dalam menghitung keterampilan yang cocok, keterampilan yang menjadi elemen dari mekanisme *Content-Based Filtering*.



Gambar 1. Arsitektur Sistem Rekomendasi Pekerjaan

Tahapan awal dalam pemrosesan teks mencakup normalisasi huruf (*lowercasing*), *lemmatization*, menghilangkan tanda baca, digit yang tidak sesuai, penghapusan kata umum, dan penghapusan *token* ganda untuk mengurangi *noise* dan meningkatkan mutu representasi teks [6], [23]. Pemeriksaan *fresh graduate* dilakukan dengan pendekatan *heuristic rule-based*, dilanjutkan dengan penyaringan level jabatan *(Entry/Associate/Internship*) untuk memastikan kesesuaian rekomendasi bagi lulusan baru [8], [9]. Proses tersebut dapat diamati pada Gambar 2 yang menggambarkan arsitektur sistem rekomendasi pekerjaan.

2.2. Deteksi Fresh Graduate

Deteksi *fresh graduate* dilakukan dengan skema *multi-signal scoring* yang menggabungkan beberapa indikator positif dan negatif pada resume pengguna. Indikator positif meliputi: tidak memiliki pengalaman kerja formal, memiliki pengalaman magang atau *trainee*, memiliki gelar sarjana, dan adanya proyek atau *capstone project*; indikator negatif mencakup keberadaan kata kunci senioritas seperti "*senior", "lead", "manager", atau "director"* pada *job title* maupun deskripsi [27].
Skor tersebut ditentukan melalui rumus yang dijelaskan dalam persamaan (1):

$$\text{score}_{fresh} = 2[\neg exp] + 1\,[intern] + 1[bachelor] + 1[capstone] - 2[senior] \quad (1)$$

dengan [.] adalah *Iverson bracket* yang bernilai 1 jika kondisi terpenuhi dan 0 jika tidak. Variabel *exp* menggambarkan adanya pengalaman kerja yang resmi, *intern* menunjukkan bahwa seseorang memiliki pengalaman sebagai magang atau peserta pelatihan, *bachelor* menandakan pemilikan gelar pendidikan sarjana, *capstone* mencerminkan keberadaan proyek akhir atau proyek *capstone*, dan *senior* menunjukkan adanya istilah senioritas seperti *senior, lead, manager, atau director* dalam profil pengguna.

Setelah langkah menentukan lulusan baru dengan cara multi-sinyal (Persamaan 1) dan seleksi tahap jabatan, sukses didapatkan 192 profil pemakai yang memenuhi ketentuan sebagai lulusan baru dari jumlah sekitar 54.000 daftar riwayat hidup dalam arsip awal. Jumlah ini menggambarkan angkatan yang nyata tidak mempunyai pengalaman kerja taraf senior, menampakkan petunjuk pendidikan sarjana, dan tidak ada tanda senioritas di nama atau deskripsi tugas sebelumnya. Kelompok ini dipakai sebagai populasi pokok di seluruh prosedur pemberian saran dan evaluasi sistem.

2.3. Prapemrosesan Teks dan Rekayasa Fitur

Profil pengguna dibentuk dengan menggabungkan informasi keterampilan, pendidikan, dan riwayat pengalaman menjadi satu *unified text profile*, sedangkan profil pekerjaan dibentuk dari judul pekerjaan, deskripsi, dan daftar keterampilan yang dipersyaratkan. Untuk mempertahankan makna istilah teknis multikata (misalnya "*machine learning*", "*natural language processing*"), diterapkan *phrase detection* dan pemetaan *multiword term* sehingga frasa penting diperlakukan sebagai satu token. Dalam penelitian ekstraksi *skill* terkini, adanya keragaman bentuk penyebutan *skill* mampu memengaruhi mutu dalam pengenalan *skill* dan *downstream matching*, memerlukan perlakuan khusus terhadap gabungan keterampilan yang dapat menyesuaikan [10].

Pipeline praproses mengikuti praktik terhadap standar *information retrieval* dan sistem rekomendasi berbasis teks, meliputi *lowercasing, lemmatization, stopword removal,* penghapusan tanda baca dan angka yang tidak sesuai, serta penghapusan token duplikat. Perolehannya akan diproses sebelumnya, kemudian digunakan sebagai masukan bagi modul vektorisasi TF-IDF dan SentenceBERT.

2.4. Vectorisasi dan Perhitungan Kesamaan

Penelitian pada tahap tersebut dapat menerapkan dua pendekatan representasi teks, yang meliputi penggambaran leksikal pada penggunaan TF-IDF dan penggambaran semantik yang

menggunakan *embedding* SentenceBERT, dengan fokus utama untuk menilai perbedaan ciri khas serta kinerja keduanya dalam mengukur kesamaan antara profil pengguna dan lowongan kerja. Meskipun sejumlah penelitian terdahulu mengangkat sebuah pendekatan *hybrid* dengan mencampurkan nilai TF-IDF dan *embedding* secara bersamaan [6],[22], penerapan sistem dalam penelitian ini menjadikan embedding SentenceBERT sebagai pendekatan utama, sedangkan TF-IDF digunakan sebagai *baseline* pembanding dalam penilaian akhir pada peningkatan mutu kesesuaian yang diperoleh dari penggambaran semantik yang kontekstual.

Pendekatan ini bersumber dari berbagai penemuan terbaru yang memperlihatkan adanya model *embedding* pada penggunaan *transformer* seperti BERT dan turunannya lebih unggul dalam memeriksa kesamaan arti antar teks, terutama pada tugas yang mengikuti keragaman redaksi, kesamaan, dan susunan kalimat yang lengkap, yang sering ditemukan pada ringkasan dan deskripsi pekerjaan [10], [27]. Selain itu, penggunaan BERT sebagai representasi sematik utama semakin marak dalam metode pada berbagai topik rekomendasi, termasuk ketika digabungkan dengan komponen pembelajaran mesin untuk meningkatkan pengambilan keputusan [28]. Melalui pendekatan pada penggunaan *transformer* dengan kesadaran diri juga terbukti berhasil dalam meningkatkan mutu untuk menganjurkan pekerjaan *JobFormer*, yang menghubungkan penggambaran semantik dengan informasi tentang keterampilan untuk pemeringkatan anjuran dalam pekerjaan [29].

TF-IDF yang justru sebaliknya bersifat leksisal dan bergantung pada kemunculan istilah eksplisit, guna mampu memiliki keterbatasan dalam menangkap hubungan semantik laten dan bahasan arti yang lebih dalam. Akibatnya, TF-IDF lebih baik digunakan sebagai metode dasar atau pembanding daripada sebagai representasi utama dalam pencocokan teks yang kompleks [4], [21]. Dalam penelitian yang dilakukan oleh Syah dkk. [31], para peneliti membandingkan TF-IDF dan BERT cenderung lebih akurat daripada TF-IDF dalam tugas evaluasi jawaban berbasis teks. Selain itu, penelitian yang dilakukan pada domain pencarian informasi menunjukkan bahwa representasi hybrid TF-IDF dan BERT melalui fusi tingkat fitur menghasilkan kinerja yang lebih baik daripada masing-masing TF-IDF dan BERT. Ini menunjukkan bahwa penggunaan e*mbedding transformer* sebagai representasi utama dan TF-IDF sebagai dasar pembanding dalam evaluasi diperlukan [32].

Representasi leksikal dihitung menggunakan TF-IDF dengan seleksi fitur berbasis variansi untuk mengurangi dimensi dari sekitar 50.000 istilah menjadi 5.000 fitur paling informatif, diikuti oleh reduksi dimensi menggunakan *Singular Value Decomposition* (SVD) hingga 200 dimensi. Bobot TF-IDF merepresentasikan pentingnya sebuah term dalam dokumen relatif terhadap koleksi, dan kemiripan antara profil pengguna u dan pekerjaan j dihitung menggunakan *cosine similarity*, seperti yang dijelaskan dalam persamaan (2):

$$\text{sim}_{\text{TF-IDF}}(u, j) = \frac{\vec{u} \cdot \vec{j}}{\|\vec{u}\| \, \|\vec{j}\|} \qquad (2)$$

dengan u merefleksikan *profile* pengguna (resume), dan j mereflesikan profile pekerjaan. Simbol $\vec{u}$ dan $\vec{j}$ masing-masing merupakan gambaran vektor TF-IDF setelah proses pengurangan dimensi dengan menggunakan *Singular Value Decomposition* (SVD). Notasi ||.|| menunjukkan norma L2, sementara operasi titik (.) menunjukkan produk titik antara vektor.

Dalam penelitian ini, hasil perhitungan TF-IDF tidak digunakan sebagai skor utama rekomendasi, melainkan dianalisis secara terpisah sebagai baseline pembanding. Pendekatan ini sejalan dengan studi perbandingan representasi teks yang menunjukkan bahwa TF-IDF tetap berguna sebagai metode referensi awal, namun kurang memadai untuk menangkap kemiripan semantik pada teks kompleks dibandingkan *embedding* modern.

Sebagai representasi utama, penelitian ini menggunakan SentenceBERT dengan model all-MiniLM-L6-v2 untuk menghasilkan *embedding* berdimensi 384 dari profil pengguna dan deskripsi

pekerjaan. Model ini dirancang khusus untuk mengoptimalkan pengukuran kemiripan kalimat dan dokumen, sehingga lebih sesuai untuk tugas pencocokan teks dibandingkan dengan BERT standar [10]. *Embedding* yang dihasilkan kemudian dinormalisasi menggunakan *L2-normalization* agar operasi *dot product* ekuivalen dengan *cosine similarity.* Kemiripan semantik antara pengguna dan pekerjaan didefinisikan sebagai persamaan (3):

$$\text{sim}_{BERT}(u, j) = \vec{e}_u \cdot \vec{e}_j \quad (3)$$

dengan $\vec{e}_u$dan $\vec{e}_j$masing-masing *embedding* Sentence-BERT profil pengguna dan pekerjaan yang telah dinormalisasikan menggunakan *L2-normalization*. Dikarenakan embedding sudah mengalami normalisasi, proses *dot product* setara dengan *consine similiarity*. Kedua jenis vektor (TF-IDF dan BERT) diindeks menggunakan FAISS untuk mendukung pencarian kemiripan vektor berskala besar secara efisien.

SentenceBERT dipilih sebagai metode utama karena kemampuannya dalam menangkap kesamaan makna meskipun terdapat perbedaan terminologi, penggunaan sinonim, maupun variasi struktur kalimat, yang merupakan karakteristik umum pada resume dan deskripsi pekerjaan [10], [27]. Studi empiris terbaru juga menunjukkan bahwa *embedding* berbasis *transformer* secara konsisten memberikan performa yang lebih baik dibandingkan TF-IDF dalam tugas kemiripan teks dan rekomendasi berbasis konten, sehingga lebih representatif untuk digunakan sebagai fondasi utama sistem rekomendasi pekerjaan [31].

2.5. Content-Based Filtering dan Penilaian

Content-Based Filtering diterapkan dengan menghitung kemiripan antara setiap *user profile* dan *job profile* menggunakan $\text{sim}_{\text{TF-IDF}}$dan sim_{BERT}, kemudian menggabungkannya dengan analisis kesenjangan keterampilan. Diberikan himpunan keterampilan pengguna $skills_u$dan keterampilan yang dipersyaratkan pekerjaan $required_j$, *matched skills* dan *missing skills* didefinisikan sebagai persamaan (4) dan persamaan (5):

$$\text{matched\_}skills = skills_u \cap required_j \quad (4)$$

$$\text{missing\_}skills = required_j \setminus skills_u \quad (5)$$

dengan $skills_u$ merupakan sekumpulan kemampuan yang dimiliki oleh pengguna, sedangkan $required_j$ adalah sekumpulan kemampuan yang diperlukan dalam suatu pekerjaan. Operasi $\cap$ menunjukkan pertemuan antara dua himpunan yang menghasilkan kemampuan yang relevan. Lalu pada persamaan (5) dengan operator $\setminus$ menunjukkan perbedaan antar himpunan, yaitu kemampuan yang diperlukan untuk pekerjaan namun belum dimiliki oleh pengguna. Derajat kesesuaian kompetensi diukur dengan *skill coverage*, yang didefinisikan pada persamaan (6):

$$skill\_\text{coverage}(u, j) = \frac{|matched\_skills|}{|required_j|} \quad (6)$$

Dengan menyatakan kardinalitas (jumlah elemen) dalam sebuah kelompok. Nilai $skill\_coverage(u, j)$ terletak pada kisaran 0 sampai 1 dan menggambarkan persentase kemampuan dalam pekerjaan yang sudah dimiliki oleh pengguna. Skor akhir rekomendasi dihitung menggunakan *weighted linear combination* yang menggabungkan kemiripan semantik, kemiripan leksikal, dan cakupan keterampilan, seperti yang dijelaskan pada persamaan (7):

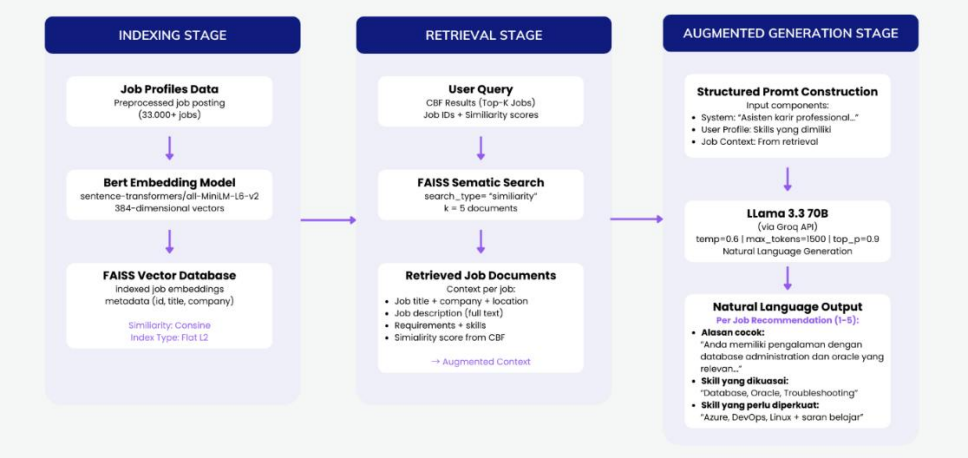
$$\text{Score}(u, j) = 0.5 \cdot sim_{BERT}(u, j) + 0.25 \cdot sim_{TF\text{-}IDF}(u, j) + 0.25 \cdot skill\_coverage(u, j) \quad (7)$$

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No 2, April 2026, Page. 1886-1904
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5723

Dengan $Score(u, j)$ adalah nilai akhir dari rekomendasi untuk kombinasi antara pengguna u dan pekerjaan j. Dengan bobot 0.5 pada sim_{BERT}diberikan untuk memprioritaskan relevansi semantik, sedangkan bobot 0.25 pada $sim_{TF\text{-}IDF}$dan *skill coverage* memastikan bahwa kecocokan kata kunci eksplisit dan kesesuaian kompetensi tetap dipertimbangkan. Untuk setiap *fresh graduate*, sistem mengeluarkan TopK pekerjaan dengan *Score* tertinggi setelah menerapkan *experience level filtering* untuk menghapus posisi *senior/managerial* yang tidak sesuai.

2.6. Membuat Penjelasan Menggunakan RAG

Aspek *explainability* diwujudkan melalui modul Retrieval-Augmented Generation (RAG), yang mengintegrasikan sistem pengambilan FAISS dengan *Large Language Model* (LLM) sebagai generator penjelasan. Pemilihan RAG ini dikarenakan menyediakan *external grounding* (konteks dari dokumen yang di *retrieve*) sehingga keluaran LLM lebih terikat pada bukti, yang dalam berbagai penelitian dianggap penting untuk meningkatkan akurasi dan mengurangi kecenderungan halusinasi pada sistem generatif [33]. Untuk setiap pasangan pengguna pekerjaan yang direkomendasikan, FAISS mengambil sejumlah dokumen pekerjaan teratas yang paling relevan misalnya, lima dokumen teratas sebagai konteks eksternal yang disediakan untuk LLM.



Gambar 2. Kerangka Kerja RAG untuk Sistem Rekomendasi Pekerjaan

Struktur RAG yang dipakai pada sistem rekomendasi pekerjaan diperlihatkan di Gambar 3. Prosedur diawali dengan langkah indeksasi dokumen pekerjaan, dilanjutkan dengan prosedur pencarian (*retrieval*) guna menemukan dokumen yang terkait berdasarkan kesamaan makna, dan ditutup dengan langkah pembuatan penjelasan memakai LLM berdasarkan latar belakang yang didapatkan.

Input ke LLM disusun berdasarkan *prompt* terstruktur. *Prompt* ini mencakup ringkasan profil pengguna, ringkasan lowongan pekerjaan, daftar keterampilan yang cocok serta yang belum terpenuhi, serta instruksi untuk menghasilkan penjelasan dalam bahasa Indonesia. Selanjutnya, LLM (keluarga LlaMA) menghasilkan penjelasan naratif. Penjelasan ini menjelaskan alasan kecocokan, keterampilan yang telah memenuhi persyaratan, keterampilan yang perlu ditingkatkan, serta saran pengembangan karier bagi lulusan baru. Berdasarkan literatur terkini, LLM biasanya menghasilkan penjelasan yang lebih alami dan mudah dipahami oleh pengguna, namun tetap memerlukan pengendalian agar penjelasan sesuai dengan bukti (pencocokan keterampilan dan konteks lowongan) [34], sehingga konteks RAG dan daftar keterampilan yang cocok/tidak digunakan sebagai pengikat (batasan *grounding*). Di samping itu, penelitian tentang sistem rekomendasi yang dapat dijelaskan berdasarkan LLM menunjukkan bahwa kesesuaian kualitas penjelasan dapat diperbaiki melalui mekanisme pelatihan atau penilaian kualitas penjelasan, yang memperkuat argumen untuk menggunakan *prompt* terstruktur dan evaluasi mendalam terhadap kualitas penjelasan [35].

Demi menjamin LLM menyajikan penjelasan yang tepat dan tidak memihak, beberapa cara pengikatan diterapkan pada bagian RAG. Pertama, model dibatasi hanya memakai konteks yang didapat dari hasil pencarian FAISS dan daftar kompetensi yang telah dihitung sebelumnya, jadi tidak boleh memasukkan informasi dari luar batasan itu. Kedua, penjelasan wajib merujuk terang pada kompetensi yang termuat dalam daftar tersebut, menjadikan setiap pernyataan tentang kesesuaian atau kurangnya keahlian bisa dilacak balik ke data sistem. Ketiga, jika ada keterangan yang tidak ditemukan di dalam konteks yang tersedia, model diminta menyatakan bahwa keterangan itu tidak disebutkan, supaya potensi mengarang-ngarang berkurang. Keempat, petunjuk terstruktur dipakai yang mewajibkan keluaran memuat empat unsur pokok, yaitu: (1) dasar kesesuaian, (2) kompetensi yang telah cocok, (3) kompetensi yang perlu ditingkatkan, dan (4) usul untuk peningkatan kemampuan. Metode ini menjamin bahwa penjelasan tetap berlandaskan fakta dan selaras dengan temuan analisis sistem.

Model generatif yang diterapkan dalam studi ini adalah LLaMA-3.3-70B melalui API Groq dengan pengaturan temperature = 0.6 dan top_p = 0.9, bertujuan untuk memastikan adanya keseimbangan antara konsistensi hasil dan keragaman bahasa dalam penciptaan teks.

2.7. Evaluasi dan Validasi HR

Hal ini dilaksanakan pada dua dimensi pokok, yang meliputi mutu pada saran serta penjelasan. Kualitas anjuran dinilai dengan menggunakan metrik *content based filtering* yang meliputi rerata nilai dengan kemiripan (*Average Similarity*), muatan keterampilan, dan panduan lainnya seperti diversity TopK, sedangkan muu penjelasan dibuktikan secara kuantitatif oleh ahli *Human Resource* (HR) dengan menggunakan BLEU dan ROUGE.

Mutu penjelasan tersebut diuji secara kuantitatif oleh para ahli Human Resource (HR) menggunakan BLEU dan ROUGE. Akan tetapi, karena penelitian terbaru menunjukkan bahwa metrik berbasis *n-gram overlap* (seperti BLEU) tidak selalu memiliki korelasi yang kuat dengan penilaian manusia dalam berbagai tugas generasi teks, maka evaluasi dari pakar HR dijadikan sebagai validasi utama untuk memastikan keselarasan dengan cara HR menilai kecocokan [36]. Terdapat sepuluh dokumen pengujian yang berisi pasangan tentang penyampaian sistem dan penjelasan HR dibandingkan dengan memanfaatkan BLEUn (n = 1–4) guna mengukur ketepatan -ngram dan -BLEU-avg sebagai ringkasan, serta ROUGE-1, ROUGE-2, dan ROUGE-L F1 untuk mengukur muatan informasi dan kesamaan struktur teks. Secara umum, skor BLEU dinyatakan sebagai persamaan (8):

$$\text{BLEU} = BP \cdot exp(\sum_{n=1}^{4} w_n \, log \, p_n) \quad (8)$$

dengan $p_n$adalah precision ngram- ke-n, $w_n = \frac{1}{4}$, dan BP adalah *brevity penalty*, sedangkan ROUGEL memanfaatkan *longest common subsequence* (LCS) antara penjelasan sistem dan referensi untuk menghitung precision, recall, dan F1 berbasis LCS. Penguatan ROUGE dengan bentuk semantik pada penelitian terkini juga memperlihatkan adanya penilaian *overlap leksikal* murni dapat ditingkatkan dengan memikirkan kedekatan sebuah arti, guna perolehan hasil ROUGE perlu ditampilkan bersama penilaian para ahli [37]. Penelitian terbaru lain juga memfokuskan adanya ROUGE dapat memberi sinyal yang berguna, melainkan juga tetap perlu kewaspadaan dalam meringkas mutu tanpa pembuktian manusia [38].

Selain penilaian secara langsung, juga akan dilakukan pengamatan pada tampilan yang bersumber kelompok penilaian HR yang meliputi “Sangat Cocok”, “Cukup Cocok”, “Kurang Cocok” untuk menilai ketetapan antara nilai BLEU/ROUGE dan *judgement* pakar. Pendekatan nilai yang bergantung pada penilaian manusia termasuk pemanfaatan para penilai dengan menggunakan LLM dalam beberapa penelitian juga sering diterapkan untuk melengkapi metrik secara langsung ketika fokusnya adalah menilai mutu pada penjelasan rekomendasi [39]. Hasil ini digunakan untuk menilai sejauh mana penjelasan yang dihasilkan sistem selaras dengan cara HR menilai kecocokan kandidat terhadap lowongan pekerjaan.

3. HASIL

Hasil implementasi sistem rekomendasi pekerjaan yang disarankan disajikan di bagian ini. Ini termasuk contoh keluaran sistem, hasil rekomendasi *Top-K*, evaluasi kuantitatif berbasis teks *(*BLEU dan ROUGE) yang divalidasi oleh pakar HR, dan analisis komperhensif terhadap performa sistem.

3.1. Contoh Hasil Rekomendasi dan Penjelasan RAG

Pengujian dilakukan pada profil pengguna yang berbeda untuk menunjukkan kemampuan sistem yang diusulkan untuk dijelaskan. Tabel 3 menunjukkan contoh hasil rekomendasi pekerjaan untuk salah satu pengguna (ID pengguna 166) yang menggunakan metode filtrasi berbasis konten berbasis sentence-BERT yang dikombinasikan dengan modul Generasi *Retrieval-Augmented* (RAG).

Sebelum memberikan anjuran dan penerangan, sistem lebih dulu mencocokkan profil pengguna dan ragam pekerjaan dengan mengukur kesamaan makna memakai dan mengkaji kecocokan keahlian yang nampak. Keahlian yang dipunyai pemakai didapatkan dari kolom *cleaned_skills* dan *cleaned_abilities,* sementara keahlian yang diminta pekerjaan didapat dari kolom *required_skills* dan *job_description*. Proses pencocokan aksara dikerjakan lewat fungsi himpunan untuk memperoleh matched_*skill*s dan *missing_skills* seperti tertera dalam Persamaan (4) dan (5). Derajat kesamaan makna diukur memakai *cosine similarity* pada *embedding* yang sudah dibakukan dan disatukan dalam nilai akhir menurut Persamaan (7). Keluaran ini lantas menjadi dasar untuk modul RAG guna menciptakan uraian berupa cerita. Hasil keseluruhan dari proses ini dapat dilihat pada Tabel 3.

Tabel 3. Contoh Hasil Rekomendasi dan Penjelasan Berbasis RAG (User 166)

Components	Description
User ID	**166**
***Skill* Utama**	SQL, Python, HTML, JavaScript
Model CBF	Sentence-BERT
Job Title	Volunteer: Data Engineer
Company	VolunteerMatch
Location	Spokane, WA
Similarity Score	0.4110
Matched *Skill*s	Python, SQL
Missing *Skill*s	Data Analysis, MySQL
RAG Explanation (Ringkasan)	Rekomendasi diberikan karena kesesuaian kemampuan Python dan SQL dengan kebutuhan posisi Data Engineer. Sistem juga menyarankan penguatan keterampilan analisis data dan MySQL.

3.2. Top-K Rekomendasi Pekerjaan

Sistem juga memperoleh daftar Top-K pada anjuran pekerjaan yang bersumber pada nilai yang hamper sama. Tabel 4 menampilkan lima rekomendasi teratas untuk User ID 166.

Tabel 4. Top-5 Rekomendasi Pekerjaan untuk User 166

Rank	Job Title	Company	Similarity (Bert) %
1	Volunteer: Data Engineer	VoluteerMatch	41.10
2	Automation Engineer	Tata Consultancy Services	35.81
3	Python Developer	Collabera	32.61
4	Control Room Compliance Officer	Partnership Employment	31.48
5	Marketing Specialist	Sales Empowerment Group	31.28

Vol. 7, No 2, April 2026, Page. 1886-1904
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5723

Hasil peringkat pada Tabel 4 menunjukkan bahwa sistem berhasil mengurutkan pekerjaan yang bersumber pada tingkat kesesuaian terhadap profil pengguna. Posisi tertinggi *(Data Engineer*) memiliki skor kemiripan 0.4110, sementara posisi kelima (*Marketing Specialist*) memiliki skor 0.3128, menggambarkan sebuah penurunan sebesar 23.9%. Meskipun terdapat perbedaan nilai, semua anjuran dalam Top-5 tetap sesuai dengan profil pengguna, menandakan adanya model Sentence-BERT mampu menangkap bahasan semantik dengan baik.

Meskipun cakupan area terlihat berbeda dari peran teknis seperti *Data Engineer*, posisi *Marketing Specialist* tetap ada di urutan lima teratas karena Sentence-BERT memahami kesamaan makna berdasarkan konteks penerapan keahlian, contohnya Python untuk menelaah data dan SQL bagi manajemen data pelanggan. Hal tersebut memperlihatkan bahwa representasi embedding tidak cuma berpegangan pada kecocokan kata kunci yang eksplisit, namun juga mengerti kaitan ideologis antar keahlian dalam lingkup pekerjaan yang memusatkan perhatian pada informasi.

Distribusi skor yang menurun secara gradual (0.4110 → 0.3128) memperlihatkan bahwa model dapat membedakan pekerjaan dengan tingkat kesesuaian yang berbeda secara bertahap. Kehadiran Marketing Specialist dalam Top-5, meskipun secara semantik jauh dari Data Engineer, dapat disampaikan melalui overlap keterampilan yang dapat diterima seperti Python untuk pengamatan data dan SQL untuk pengelola database pelanggan, yang semakin sesuai dalam peran pemasaran dengan menggunakan data. Penelitian ini membuktikan pendekatan nilai campuran yang menggabungkan kemiripan semantik dengan kecocokan keterampilan yang jelas, mencegah sistem menjadi terlalu sempit dalam pencocokan kata kunci teknis sekaligus mempertahankan keseuaian melalui pemahaman semantik.

3.3. Evaluasi Kualitas Penjelasan Menggunakan BLEU dan ROUGE

Untuk menilai kualitas penjelasan yang dihasilkan oleh modul RAG, pembuktiannya dilakukan dengan membandingkan penjelasan cara kerja terhadap penjelasan rujukan yang disusun oleh para ahli HR. Hal ini juga menggunakan metrik BLEU dan ROUGE pada perolehan nilai yang ditunjukkan pada Tabel 5.

Tabel 5. Ringkasan Hasil Evaluasi HR Menggunakan BLEU dan ROUGE

Metric	Mean	Std	Min	Max
BLEU-1	0.3474	0.0693	0.2051	0.4150
BLEU-2	0.1706	0.0606	0.0735	0.2510
BLEU-3	0.0812	0.0520	0.0203	0.1504
BLEU-4	0.0442	0.0317	0.0090	0.0987
BLEU-avg	0.1608	0.0499	0.0795	0.2227
ROUGE-1	0.4659	0.0847	0.2712	0.5532
ROUGE-2	0.1321	0.0678	0.0351	0.2446
ROUGE-L	0.2918	0.0556	0.2034	0.3688

Hasil pada Tabel 5 menunjukkan bahwa skor BLEU relatif rendah (rerata BLEU-avg = 0.1608), terutama pada n-gram berorde tinggi (BLEU-3 = 0.0812 dan BLEU-4 = 0.0442). Pola ini mencerminkan penurunan tajam seiring peningkatan orde: BLEU-1 (0.3474) → BLEU-2 (0.1706) → BLEU-3 (0.0812) → BLEU-4 (0.0442), dengan penurunan berturut-turut sebesar 50.9%, 52.4%, dan 45.6%. Penurunan ini menandakan bahwa penjelasan sistem sering kali menggunakan frasa dan struktur kalimat yang berbeda dari penjelasan referensi HR. BLEU mengukur kecocokan n-gram yang tepat, sehingga variasi parafrase, penggunaan sinonim, atau konstruksi kalimat yang berbeda dapat menurunkan skor meskipun konten semantik tetap sesuai. Skor BLEU-4 yang rendah (0.0442) secara khusus memperlihatkan

Vol. 7, No 2, April 2026, Page. 1886-1904
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5723

adanya generator LLM (keluarga LLaMA) memperoleh bahasa yang lebih beragam dan alami daripada hanya menyalin templat yang tetap.

Simpangan baku yang tinggi pada semua metrik BLEU (berkisar antara 0.0317 hingga 0.0693) memperlihatkan variabilitas yang berarti dalam mutu penjelasan di berbagai pasangan *user-job*. Hal ini dapat dihubungkan dengan perbedaan kompleksitas profil, besarnya ketimpangan keterampilan, dan mutu pada bahasan yang diambil dari FAISS.

Sebaliknya, skor ROUGE-1 F1 jauh lebih tinggi (rerata = 0.4659), yang menunjukkan bahwa sistem berhasil mencakup kosakata dan konsep kunci penting dalam penilaian kesesuaian HR. ROUGE-1 mengukur tumpang tindih unigram, dengan fokus pada cakupan konten daripada frasa yang tepat. Perbedaan substansial antara ROUGE-1 (0.4659) dan ROUGE-2 (0.1321), yang mencerminkan penurunan 71.7%, menandakan bahwa meskipun sistem menangkap istilah-istilah kunci secara pribadi dengan baik, urutan bigram spesifik berbeda dari referensi HR.

ROUGE-L F1 (0.2918) mengukur suburutan terpanjang yang sama, memberikan wawasan tentang kesamaan struktural. Nilai menengah antara ROUGE-1 dan ROUGE-2 menyarankan bahwa penjelasan sistem mempertahankan kesamaan secara tersusun sebagian dengan referensi HR. Sistem ini melestarikan beberapa urutan frasa umum sambil memperkenalkan variasi dalam organisasi keseluruhan.

Diskrepansi ini sesuai dengan ciri khas inheren metrik BLEU dan ROUGE: BLEU lebih sensitif terhadap pencocokan n-gram literal, sedangkan ROUGE mengutamakan cakupan informasi keseluruhan. Dalam konteks rekomendasi pekerjaan, ROUGE-1 F1 merupakan panduan yang lebih sesuai karena memfokuskan informasi asli daripada frasa yang tepat. Temuan bahwa ROUGE-1 (0.4659) hampir tiga kali lebih tinggi daripada BLEU-avg (0.1608) memvalidasi hipotesis bahwa modul RAG berhasil menangkap konten penjelasan esensial, meskipun menggunakan ekspresi linguistik yang berbeda dari pakar HR.

3.4. Performa Sistem Berdasarkan Kategori Penilaian HR

Analisis performa berdasarkan kategori penilaian HR ditunjukkan pada Tabel 6.

Tabel 6. Kinerja Sistem Berdasarkan Evaluasi Human Resource (HR)

HR	BLEU-avg (Mean)	ROUGE-1 F1 (Mean)	ROUGE-L F1 (Mean)
Sangat Cocok	0.1613	0.4878	0.2920
Cukup Cocok	0.1806	0.4870	0.3136
Kurang Cocok	0.0795	0.2712	0.2034

Hasil pada Tabel 6 menunjukkan pola yang signifikan dan konsisten di seluruh kategori penilaian. Kategori "Sangat Cocok" mencapai skor ROUGE-1 F1 sebesar 0.4878, sedangkan kategori "Kurang Cocok" hanya mencapai 0.2712, yang menunjukkan penurunan sebesar 44.4%. Pola serupa terlihat pada ROUGE-L F1, dengan perbedaan 30.3% antara kategori tertinggi dan terendah (0.2920 versus 0.2034). Kesenjangan substansial ini menunjukkan bahwa sistem memiliki sensitivitas tinggi dalam membedakan rekomendasi berkualitas tinggi ("Sangat Cocok" dan "Cukup Cocok") dari rekomendasi berkualitas rendah ("Kurang Cocok"), berdasarkan kualitas penjelasan yang dihasilkan. Sensitivitas ini menandakan bahwa komponen retrieval RAG (FAISS) berhasil mengidentifikasi dokumen kontekstual yang lebih relevan untuk pasangan pengguna-pekerjaan dengan kecocokan yang lebih baik.

Meskipun struktur kalimat penjelasan sistem berbeda dari referensi HR, konten substantif tetap selaras, sebagaimana tercermin dalam skor ROUGE-1 F1 yang menunjukkan cakupan topik. Skor yang hampir identik antara "Sangat Cocok" (0.4878) dan "Cukup Cocok" (0.4870) dengan perbedaan hanya 0.16% menunjukkan bahwa sistem mempertahankan kualitas penjelasan yang konsisten di seluruh

rekomendasi yang baik. Kualitas ini baru menurun secara signifikan ketika kecocokan lemah. Penelitian ini memfokuskan kesamaan yang kuat antara luaran sistem dan pandangan penilaian HR, guna membuktikan adanya meskipun LLM menggunakan campuran kata yang unik, satu hal informasi tetap menyesuaikan dan dapat dipahami oleh praktisi HR.

Modul RAG berhasil mengintegrasikan konteks dari dokumen pekerjaan yang relevan (top-5 hasil retrieval FAISS) ke dalam *prompt* terstruktur, sehingga menghasilkan penjelasan komprehensif yang didasarkan pada data faktual. Penurunan tajam pada kategori "Kurang Cocok" (ROUGE-1: 0.2712; BLEU-rata-rata: 0.0795) menunjukkan bahwa ketika kecocokan pengguna-pekerjaan lemah, kualitas konteks yang diambil dan penjelasan yang dihasilkan menurun secara proporsional. Hal ini memperagakan ketetapan *end-to-end* dari sistem. Penelitian ini juga membuktikan adanya bukti pendekatan RAG terletak pada kemampuannya untuk menyesuaikan mutu pada penjelasan secara dinamis yang bersumber pada mutu panduan yang mendasarinya, bukan menghasilkan penjelasan templat yang pasif dan tidak tanggap terhadap pembahasan yang berbeda.

Menariknya, skor BLEU-rata-rata untuk "Sangat Cocok" (0.1613) dan "Cukup Cocok" (0.1806) hampir identik, dengan perbedaan hanya 0.39%. Kesamaan pengurangan ini memfokuskan adanhya anjuran yang sesuai dari baik maupun cukup dalam mempertahankan gaya bahasa dan frasa yang konsisten. Namun, kategori "Kurang Cocok" mengalami penurunan dramatis sebesar 56.1% menjadi 0.0795, yang menunjukkan adanya kecocokan buruk menghasilkan penjelasan yang menyimpang secara berarti dari frasa rujukan HR. Tata cara ini menejelaskan adanya nilai BLEU dalam bahasan ini berfungsi lebih sebagai panduan biner yang berarti cocok melawan tidak cocok daripada skala kualitas yang berkelanjutan. Ketidakmampuan metrik ini untuk membedakan antara kategori "Sangat Cocok" dan "Cukup Cocok", dengan membedakan tajam kelompok "Kurang Cocok", memperkuat pemahaman adanya BLEU menangkap kesamaan bahasan pada permukaan, bukan kesesuaian semantik**.**

3.5. Analisis Kesenjangan Keterampilan dan Konten Penjelasan

Untuk memperdalam pemahaman mengenai interaksi antara kualitas rekomendasi dan ketepatan dalam penjelasan, sebuah pengamatan tambahan mengenai hubungan antara ruang keterampilan dan mutu penjelasan. Meskipun tidak ditampilkan dalam tabel terpisah, pengamatan dari 10 dokumen yang diuji memperlihatkan tata cara pola yang tepat di tiga kategori cakupan keterampilan. Pasangan pencari kerja yang memiliki kecocokan keterampilan lebih dari 70% dari yang dibutuhkan secara konsisten mendapat penilaian "Sangat Cocok" dari pihak HRD. Pasangan ini juga memperoleh nilai ROUGE-1 F1 di atas 0.45. Penjelasan untuk pasangan ini terutama menyoroti pengakuan terhadap kompetensi yang sudah ada dan rekomendasi untuk peningkatan keterampilan kecil. Hal ini mencerminkan keyakinan sistem dalam merekomendasikan posisi yang memiliki kesesuaian tinggi.

Pasangan dalam kategori cakupan keterampilan sedang (40-70%) biasanya mendapat penilaian "Cukup Cocok" dan menunjukkan skor ROUGE-1 F1 antara 0.40 hingga 0.50. Penjelasan untuk kategori ini menyeimbangkan antara menekankan keterampilan yang relevan dan memberikan panduan detail untuk mengatasi kesenjangan keterampilan. Akibatnya, narasi yang dihasilkan menjadi lebih panjang dan kompleks. Sebaliknya, pasangan dengan tumpang tindih keterampilan terbatas (kurang dari 40%) mendapat penilaian "Kurang Cocok" dan menunjukkan skor ROUGE-1 F1 yang jauh lebih rendah (0.25-0.30). Penjelasan untuk kategori ini mengalami kesulitan dalam menemukan hubungan yang berarti. Penjelasan tersebut berulangkali menggunakan saran karier yang bersifat umum yang sekan memberikan anjurna secara mendalam yang bersumber keterampilan. Hal ini menampilkan sebuah keterbatasan sistem dalam menemukan kesesuaian ketika keunggulan dasar sangat lemah.

Pola ini menunjukkan bahwa metode evaluasi dengan bobot (0.5 × kemiripan semantik + 0.25 × kemiripan TF-IDF + 0.25 × cakupan keterampilan) yang berhasil mengutamakan pekerjaan yang sesuai dengan keterampilan. Pendekatan ini memungkinkan modul RAG untuk memperoleh penyampaian

yang lebih sesuai dan terfokus. Pengamatan ketimpangan pada keterampilan yang jelas (*matched_skills* dan *missing_skills*) memberikan input yang terkelola. Hal ini mendukung penjelasan yang dihasilkan oleh model bahasa besar pada penilaian keunggulan yang nyata, bukan hanya kemiripan semantik abstrak. . Tata cara ini berhasil memindahkan sebuah penjelasan yang lebih mendalam ketika tingkat kesesuaian keunggulan yang tinggi.

3.6. Kontribusi Penelitian dan Kebaruan

Penelitian ini mengembangkan sumber bacaan mengenai anjuran terhadap pekerjaan dengan fokus tidak hanya pada ketepatan pada perkiraan kualitas peringkat, melainkan juga pada *explainability* dan kesamaan dengan praktik penilaian HR. Gabungan *Content-Based Filtering*, pengamatan kesenjangan keterampilan, dan penjelasan dengan menggunakan RAG merupakan gabungan yang masih jarang dijelajahi dalam bidang rekomendasi pekerjaan. Pada penellitian yang pertama telah menerapkan penilaian *explainability* secara kuantitatif melalui metrik BLEU dan ROUGE yang digunakan bersumber pada rujukan para ahli HR. Hal ini memberikan ukuran mutu yang dapat diperoleh ulang dan bersifat objektif. Pendekatan ini berbeda dari kebanyakan penelitian sebelumnya yang menilai *explainability* melalui penelitian pengguna atau penilaian kualitatif. Dengan demikian, hasil-hasil ini dapat direplikasi dan dibandingkan di antara berbagai sistem.

Ditahap kedua pengguna *Retrieval-Augmented Generation* dalam wilayah rekomendasi pekerjaan mengatasi isu khayalan yang sering muncul pada penjelasan pada penggunaan LLM murni. Pendekatan ini melandasai sebuah teks yang diperoleh pada bahasan lowongan pekerjaan yang diambil. Ketepatan secara fakta terjamin dan kemungkinan anjuran yang tidak bersumber diminimalkan. Ditahap ketiga memfokuskan pada lulusan baru melalui pemeriksaan dengan menggunakan heuristik dan penyaringan tingkat pengalaman yang lebih lanjut memastikan adanya anjuran dan penyampaian sesuai dengan kebutuhan pencari kerja tingkat pemula. Pendekatan ini memenuhi kebutuhan secara mendalam bagi pribadi dengan pengalaman kerja yang terbatas. Keempat, penilaian multi-sinyal yang menggabungkan kemiripan semantik (Sentence-BERT), kemiripan leksikal (TF-IDF dengan SVD), dan cakupan keterampilan yang eksplisit memberikan evaluasi relevansi yang lebih kuat dibandingkan dengan pendekatan tunggal. Hal ini dibuktikan oleh rekomendasi Top-K yang bervariasi namun tetap relevan dan sejalan dengan penilaian HR.

Secara umum, hasil penelitian mengindikasikan bahwa meskipun nilai BLEU tetap rendah akibat perbedaan Bahasa antara deskripsi sistem dan HR, nilai ROUGE terutama ROUGE-1 F1 menunjukkan adanya keselarasan konten yang signifikan serta cakupan topik yang tepat. Pola yang konsisten di berbagai kategori penilaian HR (nilai tinggi untuk kesesuaian yang relevan, nilai rendah untuk kesesuaian yang tidak relevan) memvalidasi bahwa modul *explainability* sistem menghasilkan output yang sensitif terhadap kualitas rekomendasi dan konsisten dengan penilaian dari para ahli. Dengan Oleh karena itu, penelitian ini berhasil menjembatani kesenjangan penting dalam sistem rekomendasi pekerjaan yang sering kali berfungsi sebagai "*black box*" tanpa penjelasan yang dapat dipahami. Penelitian ini memberikan transparansi yang diperlukan untuk membangun kepercayaan pengguna dan meningkatkan adopsi sistem di kalangan praktisi HR dalam konteks rekrutmen dunia nyata.

4. DISKUSI

Hasil eksperimen menunjukkan bahwa sistem mampu mengidentifikasi poin-poin yang signifikan yang dimanfaatkan oleh para profesional HR. Nilai ROUGE-1 F1 yang lebih tinggi dibandingkan dengan BLEU-avg menunjukkan bahwa sistem memiliki kekuatan lebih dalam hal jangkauan informasi ketimbang kesamaan frasa secara harfiah. Perbedaan signifikan antara ROUGE-1 (0.4659) dan BLEU-avg (0.1608) sebagaimana terlihat pada table 5 menandakan bahwa komponen RAG berguna dalam menyajikan pokok pembahasan dari penjelasan HR meskipun ada variasi dalam susunan kata. Temuan

Vol. 7, No 2, April 2026, Page. 1886-1904
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5723

ini konsisten dengan keterbatasan metrik berbasis n-gram yang sering kali kurang peka terhadap parafrase dan kesamaan makna.

Perbedaan dalam skor berdasarkan kategori evaluasi HR menunjukkan konsistensi antara mutu rekomendasi dan mutu penjelasan. Kategori "Sangat Cocok" dan "Cukup Cocok" menghasilkan nilai ROUGE yang lebih baik dibandingkan dengan kategori "Kurang Cocok". Sebagaimana terlihat pada Tabel 6, kategori "Sangat Cocok" (0.4878) dan "Cukup Cocok" (0.4870) memperlihatkan kesamaan skor yang baik, sedangkan "Kurang Cocok" mengalami penurunan lumayan besar (0.2712). Hal ini membuktikan bahwa mutu narasi mengikuti tingkat kesesuaian yang riil, bukan semata-mata capaian yang terbuat secara awam.

Sistem ini dapat berfungsi dengan baik karena menyatukan tiga sinyal utama yang berbeda. Sentence-BERT mengidentifikasi kesamaan arti antara profil dan tawaran pekerjaan. Analisis *matched skills* dan *missing skills* memberikan kejelasan pada alasan yang diberikan. Modul RAG kemudian menyusun alasan-alasan tersebut menjadi narasi yang sesuai dengan konteks. Paduan ini mendukung upaya rekomendasi yang dapat dijelaskan, yang menekankan pentingnya transparansi serta memungkinkan pengguna dan penilai di bidang itu melakukan penelusuran. Contoh konkret dapat diamati di Tabel 3, yang memperlihatkan keahlian yang ada (contohnya Python dan SQL) beserta keahlian yang kurang (misalnya Analisis Data dan MySQL) yang berfungsi sebagai landasan jelas untuk merangkai uraian. Sebaliknya, lima hasil teratas pada Tabel 4 memperlihatkan bahwa mekanisme masih mempertahankan ragam peran yang pas tanpa mengurangi kesamaan makna.

Supaya posisi penelitian ini lebih gampang dipahami dibanding penelitian terdahulu, Tabel 7 menyajikan perbandingan dengan cara yang sering dipakai dalam sistem rekomendasi pekerjaan.

Table 7. Perbandingan Studi Sistem Rekomendasi Terdahulu dan Penelitian Ini

Penelitian	**Domain**	**Metode Utama**	**Explainability**	**Evaluasi**	**Perbandingan Keterbatasa Penelitian**
Upadhyay et al. [11]	Job postings	Knowledge Graph + template + NER	Template-based (static)	BLEU, ROUGE-L	Tidak menggunakan generative RAG; *skill*-gap tidak eksplisit; fleksibilitas bahasa terbatas
Hussein et al. [12]	General RS	Social-aware SVD++	Tidak ada	MAE, RMSE	Bergantung histori/interaksi; tidak cocok cold-start; tidak ada penjelasan
Ro'uf et al. [13]	Job recommendation (alumni)	Content-based + MLP classifier	Tidak ada (black-box)	Accuracy, Precision, Recall	Tidak menyediakan penjelasan matched/missing *skill*s; berbasis tabular; tidak HR-aligned
This Research	Job recommendation (fresh graduates)	Sentence-BERT + *Skill*-gap + RAG	Narrative grounded explanation	BLEU, ROUGE + HR validation	Integrasi semantic matching + explainable generation

Vol. 7, No 2, April 2026, Page. 1886-1904
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5723

Atas dasar perbandingan pada Tabel 7, studi ini fokus pada keakuratan peringkat juga aspek *explainability* yang cocok dengan cara penilaian HR. Berlawanan dengan desain template yang kaku [7], rancangan ini memakai RAG guna menciptakan penjelasan naratif yang luwes serta berdasarkan situasi. Berbeda dari teknik collaborative *filtering* yang bergantung pada riwayat kegiatan [8], metode ini tetap berguna pada keadaan cold-start sebab didasarkan konten dan arti kata. Tambahan pula, penyatuan studi perbedaan kemampuan secara gamblang membedakan studi ini dari model *classifier black-box* [9], yang umumnya kurang memberikan kejelasan alasan saran.

Dari perspektif Informatika dan Ilmu Komputer, sumbangan penting dari kajian ini berada pada penyatuan temu kembali informasi semantik, pembentukan model kesenjangan kompetensi, dan penciptaan bahasa alami berdasarkan grounding dalam satu perangkat rekomendasi. Pendekatan ini melebarkan fungsi sistem rekomendasi dari semata-mata penataan urutan atau penggolongan menjadi perangkat penunjang pengambilan keputusan yang bisa diverifikasi. Penyatuan ini memamerkan bagaimana metode embedding terkini dan LLM dapat disesuaikan secara terarah untuk diterapkan dalam situasi praktis pekerjaan.

Akan tetapi, evaluasi ini masih menghadapi beberapa tantangan. Jumlah dokumen yang digunakan untuk validasi HR masih tidak memadai. Jumlah penilai HR perlu ditingkatkan untuk memperoleh hasil yang lebih konsisten. Selain *BLEU* dan *ROUGE*, studi selanjutnya seharusnya memperkenalkan metrik berbasis kesamaan makna serta kriteria penilaian manusia yang lebih komprehensif untuk mengukur akurasi alasan, kelengkapan dari kekurangan keterampilan, dan kejelasan penjelasan yang diberikan. Selain itu, melakukan uji coba pada koleksi data pekerjaan yang lebih beragam, yang memuat konteks industri yang berlaku di Indonesia, akan sangat membantu mengukur sejauh mana model dapat diterapkan secara umum serta menaikkan kesesuaian ketika dipakai.

5. KESIMPULAN

Dengan memanfaatkan *sentence-BERT* serta *Retrieval-Augmented Generation* (RAG), studi ini mengembangkan sistem rekomendasi pekerjaan yang dapat dipahami untuk lulusan baru. Sistem ini memanfaatkan mekanisme pengambilan konteks berbasis FAISS dan LLM dari lini LLaMA untuk menghasilkan penjelasan kontekstual serta rekomendasi Top-K dengan menghitung kemiripan semantik. Penjelasan tersebut menyoroti keterampilan yang telah dimiliki, keterampilan yang masih kurang, dan saran untuk pengembangan kompetensi yang terkait dengan posisi yang diusulkan.

Percobaan yang dilakukan pada dataset LinkedIn Job Postings dan Resume menunjukkan saran dan penjelasan yang sesuai dari perspektif profesional HR. Perbedaan skor yang konsisten di antara kategori penilaian HR menunjukkan bahwa kemampuan sistem ini bisa membedakan rekomendasi yang lebih relevan dibandingkan dengan yang kurang relevan berdasarkan kualitas penjelasannya. Hasil evaluasi HR menunjukkan bahwa sistem lebih efektif dalam menangkap poin penting daripada hanya meniru struktur kalimat dari referensi HR dengan nilai rata-rata ROUGE-1 F1 sebesar 0.4659 dan ROUGE-L 0.2918 yang mengindikasikan bahwa kualitas penjelasan mengikuti tingkat kesesuaian kandidat pekerjaan yang dihitung oleh sistem.

Secara keseluruhan, penelitian ini menunjukkan bahwa integrasi pencocokan semantik Sentence-BERT, analisis kesenjangan keterampilan eksplisit, dan generasi penjelasan berbasis RAG efektif menghasilkan sistem rekomendasi pekerjaan yang relevan sekaligus transparan bagi lulusan baru. Untuk membuat sistem ini lebih unggul dan siap untuk digunakan dalam proses perekrutan yang sebenarnya, penelitian selanjutnya diharapkan perlu memperluas sampel pengujian, jumlah penilaian HR, serta metrik evaluasi sematik dan penilaian manusia yang komprehensif.

REFERENCES

[1] E. Mendez Guzman, V. Schlegel, and R. Batista-Navarro, "From outputs to insights: a survey of rationalization approaches for explainable text classification," *Frontiers in Artificial Intelligence*, vol. 7, Art. no. 1363531, 2024, doi: 10.3389/frai.2024.1363531.

[2] V. Anderson dan M. Tomlinson, "Signaling standout graduate employability: The employer perspective," *Human Resource Management Journal*, vol. 31, no. 3, pp. 675–693, 2021, doi: 10.1111/1748-8583.12334.

[3] A. Eimer and C. Bohndick, "Employability models for higher education: A systematic literature review and analysis," *Social Science & Humanity Open*, vol. 8, no. 1, Art no. 100588, 2023, doi: 10.1016/j.ssaho.2023.100588.

[4] C. Chalid, "Tingkat Kompetensi Mahasiswa Fresh Graduate dalam Menghadapi Persaingan Dunia Kerja," *Indones. J. Teach. Teach. Educ.*, vol. 1, no. 1, pp. 10–13, 2021, doi: 10.58835/ijtte.v1i1.58.

[5] Badan Pusat Statistik (BPS), "Keadaan Angkatan Kerja di Indonesia Agustus 2025," Badan Pusat Statistik (BPS), Jakarta, Indonesia, Statistik 04100.25018, Desember 2025. [Daring]. Tersedia pada: https://www.bps.go.id/id/publication/2025/12/19/42a75ee61332755586fdfcdd/keadaan-angkatan-kerja-di-indonesia-agustus-2025.html

[6] D. Çelik Ertuğrul and S. Bitirim, "Job recommender systems: a systematic literature review, applications, open issues, and challenges," *Journal of Big Data*, vol. 12, no. 1, pp. 140, 2025, doi: 10.1186/s40537-025-01173-y.

[7] S. Sarsenbay, A. Kabdiyev, I. Varlamis, C. Sardianos, C. Turan, B. Razhametov, and Y. Kazym, "Generating job recommendations based on user personality and Gallup tests," *Algorithms*, vol. 18, no. 5, Art. no. 275, 2025, doi: 10.3390/a18050275.

[8] F. Tang, R. Zhu, F. Yao, J. Wang, L. Luo, and B. Li, "Explainable person–job recommendations: challenges, approaches, and comparative analysis," *Frontiers in Artificial Intelligence*, vol. 8, Art. no. 1660548, 2025, doi: 10.3389/frai.2025.1660548.

[9] J. Govea, R. Gutierrez, and W. Villegas-Ch, "Transparency and precision in the age of AI: evaluation of explainability-enhanced recommendation systems," *Frontiers in Artificial Intelligence*, vol. 7, Art. no. 1410790, Sep 2024, doi: 10.3389/frai.2024.1410790.

[10] A. Akkasi, "Job description parsing with explainable transformer-based ensemble models to extract the technical and non-technical *skills*," *Nat. Lang. Process. J.*, vol. 9, Art. no. 100102, 2024, doi: 10.1016/j.nlp.2024.100102.

[11] C. Upadhyay, H. Abu-Rasheed, C. Weber, and M. Fathi, "Explainable Job-Posting Recommendations Using Knowledge Graphs and Named Entity Recognition," dalam *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Melbourne, Australia: IEEE, 2021, pp. 3291–3296. doi: 10.1109/SMC52423.2021.9658757.

[12] M. H. Hussein, A. A. Alsakaa, and H. A. Marhoon, "Adopting explicit and implicit social relations by SVD++ for recommendation system improvement," *TELKOMNIKA Telecommunication, Computing, Electronics and Control*, vol. 19, no. 2, pp. 471, 2021, doi: 10.12928/telkomnika.v19i2.18149.

[13] A. Rouf, Y. M. Pranoto, and E. Setyati, "Sistem Rekomendasi Pekerjaan Menggunakan Content Based Similarity," *JUTISI: Jurnal Ilmiah Teknik Informatika dan Sistem Informasi*, vol. 12, no. 2, pp. 618, 2023, doi: 10.35889/jutisi.v12i2.1229.

[14] M. Rani, B. K. Mishra, D. Thakker, and M. N. Khan, "To Enhance Graph-Based Retrieval-Augmented Generation (RAG) with Robust Retrieval Techniques," *in Proceedings of the 2024 18th International Conference on Open Source Systems and Technologies (ICOSST)*, Lahore, Pakistan: IEEE, 2024, pp. 1–6. doi: 10.1109/ICOSST64562.2024.10871140.

[15] W. Yuan, G. Neubig, and P. Liu, "BARTScore: Evaluating Generated Text as Text Generation,", arXiv preprint arXiv:2106.11520, 2021, doi: 10.48550/ARXIV.2106.11520.

[16] M. Zhong, Y. Liu, D. Yin, Y. Mao, Y. Jiao, P. Liu, et al., "Towards a Unified Multi-Dimensional Evaluator for Text Generation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, 2022, pp. 2023–2038, doi: 10.18653/v1/2022.emnlp-main.131.

Vol. 7, No 2, April 2026, Page. 1886-1904
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5723

[17] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment," *in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore: Association for Computational Linguistics, 2023, pp. 2511–2522. doi: 10.18653/v1/2023.emnlp-main.153.

[18] V. S. Pendyala, N. B. Thakur, and R. Agarwal, "Explainable Use of Foundation Models for Job Hiring," *Electronics*, vol. 14, no. 14, Art. no. 2787, 2025, doi: 10.3390/electronics14142787.

[19] R. Karlović, M. Rovis, A. Smajić, L. Sever, and I. Lorencin, "Context-Aware Tourism Recommendations Using Retrieval-Augmented Large Language Models and Semantic Re-Ranking," *Electronics*, vol. 14, no. 22, Art. no. 4448, 2025, doi: 10.3390/electronics14224448.

[20] A. Y. A. Ardhana, H. N. U. Syazeedah, R. I. Fitriyaningrum, and A. Gunawan, "Analisis Ketidaksesuaian antara Pendidikan dengan Kebutuhan Dunia Kerja di Indonesia," *Kompeten Jurnal Ilmiah Ekonomi dan Bisnis*, vol. 3, no. 4, pp. 1020–1026, 2025, doi: 10.57141/kompeten.v3i4.156.

[21] N. Yadav and D. Gopinathan, "Retrieval Augmented Generation Model for Paper Recommendation System," *SN Computer Science*, vol. 6, no. 6, Art. no. 579, 2025, doi: 10.1007/s42979-025-04095-x.

[22] S. He, Y. Zhao, Q. Li, and Y. Ma, "RGPRec: A RAG-Enhanced GNN for Personalized Task Recommendations in Open-Source Communities," *Software Practice and Experience*, vol. 56, no. 1, pp. 3–25, 2026, doi: 10.1002/spe.70022.

[23] L. M. De Campos, J. M. Fernández-Luna, and J. F. Huete, "An explainable content-based approach for recommender systems: a case study in journal recommendation for paper submission," *User Modelling and User-Adapted Interaction*, vol. 34, no. 4, pp. 1431–1465, 2024, doi: 10.1007/s11257-024-09400-6.

[24] S. Luo, J. Xu, X. Zhang, L. Wang, S. Liu, H. Hou, and L. Song, "RALLRec+: Retrieval-augmented large language model recommendation with reasoning," *Expert Systems with Applications*, vol. 297, Art. no. 129508, 2026, doi: 10.1016/j.eswa.2025.129508.

[25] J. Munson, T. Cuezze, S. Nesar, and D. Zosso, "A review of large language models and the recommendation task," *Discover Artificial Intelligence*, vol. 5, no. 1, Art. no. 203, 2025, doi: 10.1007/s44163-025-00334-5.

[26] F. G. F. Putranto, C. Natalia, and N. K. D. Pitriyani, "Closing the Gap Between Education and Labor Market Requirement: Do Vocational Education Matter?," *Journal Indonesian Sustainable Development Planning*, vol. 5, no. 3, pp. 181–191, 2024, doi: 10.46456/jisdep.v5i3.614.

[27] P. Singla, "An Intelligent Job Recommendation System based on Semantic Embeddings and Machine Learning," *Journal of Information Systems Engineering and Management*, vol. 10, no. 5s, pp. 520–542, 2025, doi: 10.52783/jisem.v10i5s.681.

[28] A. N. Hasoon, S. K. Abdulateef, R. S. Abdulameer, and M. L. Shuwandy, "An Intelligent Hybrid AI Course Recommendation Framework Integrating BERT Embeddings and Random Forest Classification," *Computers*, vol. 14, no. 9, Art. no. 353, 2025, doi: 10.3390/computers14090353.

[29] Z. Guan, J.-Q. Yang, Y. Yang, H. Zhu, W. Li, and H. Xiong, "JobFormer: *Skill*-Aware Job Recommendation with Semantic-Enhanced Transformer," *ACM Transactions on Knowledge Discovery from Data*, vol. 19, no. 1, pp. 1–20, 2025, doi: 10.1145/3701735.

[30] M.-H. Ajjam and H. S. Al-Raweshidy, "AI-driven semantic similarity-based job matching framework for recruitment systems," *Information Science*, vol. 724, Art. no. 122728, 2026, doi: 10.1016/j.ins.2025.122728.

[31] M. P. Syah, Ajeng Puspa Wardani, Mohammad Idhom, and Trimono, "Perbandingan Representasi Teks Tf-Idf Dan Bert Terhadap Akurasi Cosine Similarity Dalam Penilaian Otomatis Jawaban Berbasis Teks," *Data Science Indonesia*, vol. 5, no. 1, pp. 47–59, 2025, doi: 10.47709/dsi.v5i1.6021.

[32] P. Aprilio, M. Felix, P. S. Nugraha, and H. Fahmi, "Hybrid Feature Combination of TF-IDF and BERT for Enhanced Information Retrieval Accuracy," *JISA: Jurnal Informatika dan Sains*, vol. 8, no. 1, pp. 8–15, 2025, doi: 10.31326/jisa.v8i1.2179.

[33] M. Klesel and H. F. Wittmann, "Retrieval-Augmented Generation (RAG)," *Bussiness & Information Systems Engineering*, vol. 67, no. 4, pp. 551–561, 2025, doi: 10.1007/s12599-025-00945-3.

Vol. 7, No 2, April 2026, Page. 1886-1904
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5723

[34] A. Said, "On explaining recommendations with Large Language Models: a review," *Frontiers Big Data*, vol. 7, Art. no. 1505284, 2025, doi: 10.3389/fdata.2024.1505284.

[35] M. Yang, M. Zhu, Y. Wang, L. Chen, Y. Zhao, X. Wang, *et al.*, "Fine-Tuning Large Language Model Based Explainable Recommendation with Explainable Quality Reward," *Proceeding of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 8, pp. 9250-9259, 2024, doi: 10.1609/aaai.v38i8.28777.

[36] M. Evtikhiev, E. Bogomolov, Y. Sokolov, and T. Bryksin, "Out of the BLEU: How should we assess quality of the Code Generation models?," *Journal of Systems and Software*, vol. 203, Art. no. 111741, 2023, doi: 10.1016/j.jss.2023.111741.

[37] M. Zhang, C. Li, M. Wan, X. Zhang, and Q. Zhao, "ROUGE-SEM: Better evaluation of summarization using ROUGE combined with semantics," *Expert Systems with Applications*, vol. 237, Art. no 121364, 2024, doi: 10.1016/j.eswa.2023.121364.

[38] A. Auriemma Citarella, M. Barbella, M. G. Ciobanu, F. De Marco, L. Di Biasi, and G. Tortora, "Assessing the effectiveness of ROUGE as unbiased metric in Extractive vs. Abstractive summarization techniques," *Journal of Computer Science*, vol. 87, Art. no. 102571, 2025, doi: 10.1016/j.jocs.2025.102571.

[39] X. Zhang, Y. Li, J. Wang, B. Sun, W. Ma, M. Zhang, *et al.*, "Large Language Models as Evaluators for Recommendation Explanations," *Proceedings of the 18th ACM Conference on Recommender Systems*, Bari, Italy, 2024, pp. 33–42, doi: 10.1145/3640457.3688075.