A Hybrid TF-IDF and Knowledge Graph-Enhanced Retrieval-Augmented Generation Framework with Large Language Models for Domain-Aware Question Answering

Lilyani Asri Utami[1,*], Hilda Rachmi [2], Syarif Hidayatulloh[3]

[1,3]*Faculty of Information Technology, Universitas Nusa Mandiri, Jakarta, Indonesia*

[2]*Faculty of Engineering and Informatics, Universitas Bina Sarana Informatika, Jakarta, Indonesia*

(Received: September 1, 2025; Revised: October 25, 2025; Accepted: February 8, 2026; Available online: March 17, 2026)

Abstract

This study aims to develop a domain-aware legal Question-Answering (QA) system tailored for Indonesia's Micro, Small, and Medium Enterprises (MSMEs) by proposing a hybrid Retrieval-Augmented Generation (RAG) framework that integrates Term Frequency–Inverse Document Frequency (TF-IDF), Knowledge Graph (KG), and Large Language Model (LLM) components. In this framework, TF-IDF contributes by performing lexical-level retrieval to identify the most relevant documents based on keyword weighting; the KG enriches this retrieval by providing semantic relationships among legal entities, enabling deeper contextual understanding; and the LLM generates coherent responses conditioned on both lexical and semantically grounded evidence. Together, these components work synergistically to strengthen factual grounding during retrieval and improve contextual reasoning during generation. Methodologically, the system processes a curated dataset of 1,400 legal question–answer pairs collected from national legal repositories, including legislation, government regulations, and MSME digitalization guidelines. The process includes text preprocessing, keyword extraction using TF-IDF, semantic enrichment through a KG that maps legal entities and their relationships, and answer generation via an LLM powered by the RAG pipeline. The system was evaluated using Precision, Recall, F1-Score, Bilingual Evaluation Understudy (BLEU), and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metrics, validated by five legal experts. Results show an accuracy improvement from 76.5% to 83.5% after integrating KG, with Precision of 0.853, Recall of 0.877, and F1-Score of 0.865. The generative evaluation yielded a BLEU score of 0.9276 and ROUGE-L of 0.9301, indicating strong linguistic and semantic alignment between system outputs and expert-authored references. The study concludes that this approach offers a practical foundation for building AI-based legal assistance tools and highlights future opportunities for expansion to other legal domains and multilingual RAG applications.

Keywords: Retrieval-Augmented Generation, TF-IDF, Knowledge Graph, Large Language Models, MSME

1. Introduction

The rapid growth of artificial intelligence developed systems of natural language processing with an ability to generate responses resembling human speech. There are still serious challenges with response consistency, particularly for long dialogues and specialist domains. Traditional information retrieval methods, such as keyword or statistical search, often fail to ensure precision because they rely heavily on surface-level word co-reference rather than semantic meaning. Conversely, newer generative models are able to generate text even sounding more natural but are still beset with flaws such as poor access to recent information and the tendency for generating false or fictional responses. Even venerable older language models such as BERT, RoBERTa, and GPT-2, although useful as applied to Natural Language Processing (NLP) applications, are still subject to unreality and insufficient domain knowledge when applied to specialized documents [1].

To address these challenges, keyword extraction is applied to identify key terms more efficiently [2], and this process is integrated into a RAG framework that combines information retrieval with generative modeling. RAG enhances language models by linking prompt engineering with database querying, enabling the system to produce context-rich

***Corresponding author: Lilyani Asri Utami (lilyani.lau@nusamandiri.ac.id)**
DOI: https://doi.org/10.47738/jads.v7i2.1136

and adaptive responses [3], while also helping maintain conversational context and incorporate domain-specific knowledge [4]. In practice, RAG is commonly used for the injection of domain-specific knowledge into software for which specialization is the priority [5]. In the RAG model, techniques such as Term Frequency–Inverse Document Frequency (TF-IDF) are at the core of recognizing important terms in documents. TF-IDF has been widely applied in diverse domains, including news articles, social media, and biomedical literature, to enable applications such as text retrieval and data mining [6]. The method describes each word in a document with a numerical score that indicates its importance [7]. Specifically, TF-IDF calculates the Term Frequency (TF) and Inverse Document Frequency (IDF) of a document. The weights are used to excavate document similarity, calculate the significance of certain words and keywords, and rank search results [8], [9]. The TF-IDF score is directly proportional to a term's frequency in a document but normalized by its frequency in the entire corpus [10]. This makes TF-IDF both efficient in response behavior-related key feature extraction [11] and in weighting words based on frequency of occurrence [12]. However, TF-IDF alone does not capture semantic or domain-specific nuances. To overcome this, Knowledge Graphs (KGs) are used to enhance TF-IDF weighting with structured semantic relationships [13]. Represented as networks of entities and relations [14], KGs enable deeper contextual understanding beyond surface-level term frequency, improving retrieval quality and reducing the risk of misinformation in generative outputs [15] [16]. Despite these benefits, KG development still requires human oversight to ensure adequate coverage and accuracy [17].

The results obtained from retrieval can then be combined with the capability of Large Language Models (LLMs) to achieve both natural and informative answers. LLMs have progressed rapidly in the past few years with great generalization and reasoning abilities [18], as well as a remarkable increase in adoption between both academic and industrial environments [19]. TF-IDF algorithm is generally utilized for text retrieval and data mining. Meanwhile, LLMs have been proven to achieve state-of-the-art performance on a variety of NLP tasks and are thus the default models employed in the majority of experimental setups [20]. The discovery of LLMs has revolutionized the paradigm of NLP towards improved classification, generation, and text understanding [3]. Trained on vast amounts of information, LLMs can execute myriad NLP tasks such as language comprehension and question-answering, and have even been employed in industrial applications [21]. But they suffer from transparency, knowledge update, and unreality problems. That's where RAG enters the picture: it enhances LLMs by bringing in pertinent information from outside knowledge bases and integrating it into the generation process [22]. By responding on the basis of dynamically up-to-date content, RAG increases the precision and trustworthiness of output [23]. LLMs also possess the ability to respond to questions beyond the scope of a Knowledge Graph (KG), for instance, speculative reasoning and understanding of unknown entities or relationships, because of their extensive training corpora and zero-shot inference capabilities [24]. They have been outstanding in both natural language understanding and generation tasks [25]. When combined with an LLM that has been boosted with retrieval and a KG as an external knowledge base, an LLM can facilitate precise case retrieval, comparative analysis, and logical reasoning [26].

In the RAG architecture, LLMs produce embeddings, text parsing, and natural language outputs [27]. The core of RAG is its retrieval component that proactively pulls data from large text corpora and injects knowledge of facts into prompt engineering and thereby boosting the relevance and accuracy of reasoning and generation in LLMs [28]. In essence, an LLM is an advanced artificial intelligence (AI) that can both understand and generate human-like language [29]. They are learned from gigantic text databases, enabling them to acquire patterns and structure of natural language. Their applications span machine translation and speech recognition, content generation and text classification, and they are applied extensively in many industries such as finance, healthcare, and marketing, where speedy processing and analysis of large-scale textual data are of paramount importance [30].

Legal question answering for Micro, Small, and Medium Enterprises (MSMEs) in Indonesia presents unique challenges that cannot be addressed by TF-IDF, KG, or LLM components when used in isolation. MSME regulations are dispersed across numerous statutes, licensing guidelines, and administrative procedures, making retrieval difficult due to high lexical similarity but low semantic clarity. TF-IDF assists in identifying relevant legal segments, yet it cannot capture hierarchical relations such as business classifications, licensing dependencies, or obligations across regulatory layers. Knowledge Graphs help bridge this gap by modeling explicit legal entity relationships; however, they lack the generative capability needed to produce natural-language legal explanations. Meanwhile, LLMs can articulate coherent responses, but without domain-grounded retrieval and structured semantic context, their outputs risk being generic,

incomplete, or legally inaccurate. These limitations collectively create a strong motivation for a hybrid approach that tightly integrates TF-IDF, KG, and LLM components specifically for the MSME legal domain.

2. Related Works

Recent advancements in artificial intelligence have fostered significant progress in information retrieval and Question-Answering (QA) systems through the integration of statistical, semantic, and generative approaches. These hybrid methods aim to enhance accuracy, contextual comprehension, and explainability in automated reasoning processes.

Wei Shi et al. [44] introduced the KGRank method, a keyphrase extraction approach that combines Knowledge Graphs (KGs) and semantic clustering using DBpedia and the Personalized PageRank (PPR) algorithm. The model achieved an F-measure of approximately 0.35, outperforming traditional approaches such as TF-IDF, SingleRank, and ExpandRank (below 0.3). Their findings demonstrated that embedding semantic knowledge into statistical frameworks reduces information loss and strengthens conceptual relationships between key terms within a single document. Expanding this direction, Saat et al. [45] applied KGs to content-based recommender systems (CB-RS) by integrating semantic representations derived from datasets such as MovieLens, MPST, and Serendipity-SAC2018, further enhanced with Latent Dirichlet Allocation (LDA) topic modeling. Their approach employed a two-hop semantic path to discover contextually relevant yet novel (*serendipitous*) items, mitigating overspecialization and filter bubble effects. The KG-based recommender significantly outperformed five baselines (TF-IDF, LDA, rTF-IDF, rLDA, and dTF-IDF) in both precision and serendipity, proving the potential of semantic networks to deliver meaningful and diverse recommendations.

Further conceptual advancements were made by Peng et al. [46], who provided a comprehensive overview of Knowledge Graphs as the next generation of knowledge bases. They highlighted that unlike traditional, static databases, KGs support semantic reasoning, adaptive learning, and contextual inference across heterogeneous data sources. Their study outlined five major research challenges in KG development—knowledge acquisition, representation and embedding, graph completion, knowledge fusion, and reasoning—which together define the roadmap toward dynamic, explainable, and adaptive knowledge systems. This conceptual framework positioned KGs as a foundational component for building AI systems capable of contextual understanding and transparent decision-making. Meanwhile, Hou et al. [47] proposed the KG-EGV (Knowledge Graph–Enhanced Generative Verification) framework, integrating Knowledge Graph reasoning with Large Language Models (LLMs) to ensure evidence-based answer generation. Experiments revealed an accuracy improvement of 5–9% and a substantial reduction in hallucination errors compared to baseline models. The study underscored how combining structured knowledge with generative reasoning yields more factual, coherent, and explainable QA outcomes, particularly in knowledge-intensive domains such as law and healthcare.

Although these studies demonstrate the effectiveness of Knowledge Graphs in improving retrieval and generative reasoning, they primarily operate in open domains and rely on general-purpose entity graphs such as DBpedia or large heterogeneous knowledge bases. In contrast, the KG developed in this study is constructed specifically for the MSME legal domain, capturing regulatory hierarchies, licensing dependencies, business classifications, and cross-reference relationships that are not represented in prior works. This domain specialization enables more precise semantic retrieval and reduces noise from irrelevant entities.

In a different application, Gan et al. [43] developed an LLM-agent framework for automated resume screening. After fine-tuning the LLaMA2 model, the system achieved an F1-score of 87.73%, while evaluations using BLEU and ROUGE metrics indicated strong linguistic coherence and alignment with human assessments. The model operated 11 times faster than manual screening, showcasing the efficiency and consistency of LLM-based automation in structured decision-making tasks. Similarly, Chen et al. [42] designed an LLM-based intelligent Q&A and maintenance standardization framework for railway locomotive systems by integrating the Universal Information Extraction (UIE) model with ChatGLM. The model attained precision of 93.65%, recall of 93.28%, and an F1-score of 93.46%, while the Q&A module achieved BLEU-4 = 86.87%, ROUGE-1 = 89.60%, ROUGE-2 = 87.54%, and ROUGE-L = 94.26%. Moreover, the system accelerated maintenance data standardization by more than tenfold compared to manual methods, demonstrating the adaptability of domain-specific LLMs to complex industrial environments.

Although previous studies have applied Knowledge Graphs and LLM-based frameworks across various domains, these works rely on general-purpose knowledge sources and do not integrate lexical retrieval, domain-specific semantic structures, and RAG-based generation into a unified pipeline. In contrast, this study introduces a novel combination of TF-IDF, a specialized MSME legal Knowledge Graph, and RAG-enhanced LLMs, enabling domain-grounded retrieval and legally coherent answer generation. This domain-focused hybrid design addresses gaps left by prior KG and LLM systems, which have not been tailored for the regulatory complexity of MSME legal question answering.

Building upon these advancements, the present research introduces a hybrid framework that integrates TF-IDF, KGs, and RAG powered by LLMs to develop a domain-aware legal QA system tailored for Micro, Small, and Medium Enterprises (MSMEs). The framework combines statistical term weighting through TF-IDF for extracting relevant legal concepts, semantic enrichment via Knowledge Graphs to model relationships among entities, concepts, and regulations, and generative reasoning through RAG-based LLMs to produce coherent and evidence-grounded responses. Furthermore, a legal QA system was implemented based on this architecture to enable interactive legal information retrieval and reasoning. The system's performance was rigorously evaluated using BLEU and ROUGE metrics, demonstrating strong semantic alignment and linguistic coherence between the generated responses and expert-validated legal reference answers.

3. Methodology

3.1.Experimental Dataset

This experiment evaluates the performance of a Domain-Aware RAG system integrating TF-IDF, KG, and LLM for legal question answering in the context of Micro, Small, and Medium Enterprises (MSMEs). The primary objective is to assess how this integration enhances the accuracy, relevance, and coherence of legal responses compared to conventional text-based retrieval approaches. The dataset consists of approximately 1,400 legal question–answer collected from open-access legal repositories such as *peraturan.go.id*, *hukumonline.com*, Indonesian legal documents including laws, government regulations, ministerial decrees, and MSME digitalization guidelines. The data are categorized into fourteen legal domains: general issues, business legality and licensing, intellectual property, contracts and agreements, halal product assurance, taxation, export–import, business competition, digital law and e-commerce, personal data protection, digital transactions and payments, legal strategies and disputes, employment, and consumer protection. Each entry underwent preprocessing before being utilized by the TF-IDF and KG modules for relevant legal context retrieval, which was subsequently processed by the LLM to generate coherent, accurate, and contextually appropriate legal responses within the MSME legal domain.

3.2.Experimental Procedures

The architecture of the proposed system adopts a Domain-Aware RAG framework integrating TF-IDF and Knowledge Graph as domain-specific retrievers before generating responses through a LLM, as illustrated in figure 1.

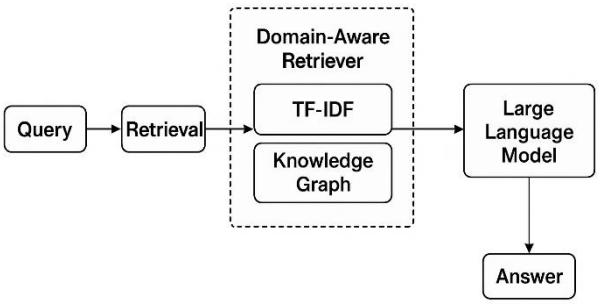

Figure 1. Experimental Procedures

As shown in figure 1, this architecture illustrates the proposed Domain-Aware RAG framework that integrates TF-IDF, KG, and LLM for domain-specific legal question answering in the context of Micro, Small, and Medium Enterprises (MSMEs). The process begins with a user query, which is sent to the retrieval module to extract potentially relevant legal documents. These candidates are then processed by the Domain-Aware Retriever in a two-stage

 ISSN 2723-6471

sequence: the TF-IDF component first performs lexical retrieval by assigning statistical weights to significant legal terms, after which the Knowledge Graph refines and enriches these TF-IDF results by evaluating semantic relationships among legal entities such as regulatory hierarchies, licensing dependencies, and definitional connections. This interaction allows the KG to re-rank or filter TF-IDF candidates, ensuring that the retrieval step captures not only frequent terms but also domain-specific contextual associations that TF-IDF alone cannot identify. Before feature extraction and KG construction, the corpus underwent a text preprocessing pipeline to ensure data cleanliness and uniformity The integrated representation is then passed to the LLM, which generates coherent and context-aware responses by combining statistical weighting and structured domain knowledge. The final Answer module presents natural, accurate, and legally consistent responses to the user. This architecture ensures that the system not only retrieves relevant legal information but also understands domain semantics and produces explainable, human-like answers.

An enormous amount of unstructured text is generated daily from various sources. With the rapid explosion of information, preprocessing for some domains has emerged as a major task in performing analysis, identifying correlations, and building NLP models [31]. Preprocessing refers to the cleaning of text by following a series of steps before further analysis [32]. In this study, preprocessing was applied uniformly to the entire MSME legal corpus to ensure lexical consistency and reduce noise prior to TF-IDF weighting and Knowledge Graph construction. It is a process used to transform raw data into a form that is efficient and useful. It is needed because raw data is incomplete and irregular in structure. Text preprocessing was conducted using standard NLP procedures to normalize the MSME legal corpus prior to TF-IDF weighting and Knowledge Graph construction. The steps included case folding [33], removal of punctuation and non-alphabetic characters [33], [34], tokenization [35], stopword filtering [36], and stemming/lemmatization [37] to reduce morphological variations. These procedures follow widely adopted preprocessing practices in legal and domain-specific text analysis. Uniform preprocessing across the corpus improves the quality of lexical features for TF-IDF, enhances entity detection for KG construction, and ultimately contributes to more accurate and semantically aligned retrieval and generation in the overall system.

3.3.Term Frequency – Inverse Document Frequency (TF-IDF)

TF-IDF is used to measure word significance within a document [38]. The model combines two metrics: Term Frequency (TF), which measures the frequency of occurrence of a word within a document, and Inverse Document Frequency (IDF), which penalizes highly occurring words across many documents [39]. Term Frequency (TF) is calculated by dividing the frequency of the word occurrence in a document by the document's word frequency. It shows the importance of a word within a document. Inverse Document Frequency (IDF) is calculated as the logarithm of the total number of documents in the corpus divided by the number of documents containing the word. It measures the word's importance within the entire corpus. TF-IDF is the product of TF and IDF. This value indicates the significance of a word in a document relative to the entire corpus. TF-IDF is calculated using the following formulas:

$$tf_{i.j} = \frac{n_{i.j}}{\sum_k nk.j} \quad (1)$$

$$idf_i = log \frac{|D|}{|\{j:t_i \in d_j\}|} \quad (2)$$

In formula (1), the numerator represents the number of times a word appears in a document, and the denominator represents the total number of words in that document. In formula (2), |D| is the total number of documents in the corpus, and the numerator is the number of documents containing the word. Multiplying TF by IDF gives the TF-IDF value of a feature word, which can also be considered the weight of that feature word, calculated as shown in formula (3):

$$tfidf_{i.j} = tf_{i.j} \times idf_i \quad (3)$$

3.4. Knowledge Graph (KG)

Knowledge Graph (KG) is an innovative approach with the potential to replace the step of feature selection to mitigate the drawbacks of high-dimensional data by removing redundant and unnecessary information and thereby improving classification processes [40]. Knowledge Graphs help reduce the subjectivity of manual feature selection by offering

 ISSN 2723-6471

a structured and explicit representation of domain knowledge. Although their construction involves human judgment, the resulting graph enables more consistent use of prior knowledge and supports systematic extraction of implicit legal relationships [41]. From the point of view of graph theory, a knowledge graph is described as follows: entities in a knowledge graph are represented as vertices (nodes), the relations between entities are represented as edges, and the whole knowledge graph can be strictly expressed by equations (4), (5), and (6).

$$V = \{V_0, V_1, V_2, V_3, \ldots\ldots V_n\} \tag{4}$$

$$E = \{(V_0, V_1, R_1), (V_0, V_2, R_2), \ldots, (V_n, V_{n-1}, R_n)\} \tag{5}$$

$$G = (V, E) \tag{6}$$

In formula (5), R is for the relationship between two entities. A knowledge graph consists of triples in the form of (entity–relation–entity). An entity of the knowledge graph can have properties of its own, and the network results through the relationships. A knowledge graph is essentially a conceptual network—a symbolic modeling of the actual domain which it models.

3.5. KG Improves TF-IDF

Word frequency measures from the knowledge graph have the capability to encode salience of entities in the text. In this study, TF-IDF is calculated to measure the importance of information relating to MSME law. In general, TF-IDF takes into consideration how often an entity appears in the legal knowledge graph. It is a statistical method to determine the significance of a word within a corpus. The worth of a word increases with its frequency in a document but decreases in proportion to its frequency across the entire corpus.

3.6. LLM

In this study, an LLM serves as the main communication interface between users and the AI system. The model used is OpenAI GPT-4o mini, equipped with a 128,000-token context window and a 16,384-token output limit, enabling it to process lengthy legal texts and generate well-structured MSME-related responses. User queries are passed through an API webhook to the backend, where the model processes them using prompt conditioning enriched by TF-IDF–ranked passages and entities detected in the Knowledge Graph. This ensures that the LLM receives both lexical and semantic cues before generating an answer. The output is returned in HTML format for immediate rendering, allowing the system to deliver more contextual, accurate, and responsive legal information compared to conventional keyword-based techniques.

3.7. Retrieval Performance Evaluation

The retrieval performance of the proposed RAG system was evaluated using the standard metrics of Precision, Recall, and F1-score, calculated based on the results of expert validation conducted by five legal professionals. Each output retrieved by the system—whether a legal document, article, or regulatory clause—was independently reviewed by the experts to assess its relevance to the user query. The evaluation process classified each retrieval outcome into four categories: True Positive (TP) for results identified as relevant by both the system and experts, False Positive (FP) for results retrieved by the system but deemed irrelevant by experts, False Negative (FN) for relevant items missed by the system, and True Negative (TN) for results correctly excluded as irrelevant by both parties. All five legal experts possess domain expertise in Micro, Small, And Medium Enterprise (MSME) regulations and conducted their assessments independently. To ensure the reliability of these expert judgments and reduce subjective bias, inter-annotator agreement was measured using Fleiss' Kappa, yielding $\kappa = 0.75$, which indicates substantial agreement [48] among the evaluators. The aggregated validation results were then used to compute the retrieval performance metrics as follows:

$$\text{Precision} = \frac{Number\ of\ relevant\ results}{Number\ of\ responses} \tag{7}$$

Precision shows the proportion of relevant answers among all responses returned by the system.

$$\text{Recall} = \frac{Number\ of\ relevant\ results}{Total\ number\ of\ test\ data} \tag{8}$$

Recall measures the ability of the system in retrieving all the relevant answers.

$$\text{F1-Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (9)$$

F1-Score balances Precision and Recall to reflect a better picture of the overall accuracy.

3.8. Answer Generation Quality Evaluation

The Bilingual Evaluation Understudy (BLEU) metric serves as an automated index for evaluating the similarity between a generated text and its reference, offering a quantitative and objective basis for assessment. This approach enhances fairness in evaluation and allows for the rapid analysis of multiple outputs, thereby significantly improving evaluation efficiency. Meanwhile, the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric assesses how well the generated text captures the key information contained in the reference text by focusing on the recall rate. This metric emphasizes the completeness and informativeness of the generated content, enabling a more holistic and comprehensive evaluation of text generation quality [42], [43].

The quality of the legal answers generated by the system was evaluated using and ROUGE metrics, with expert-validated reference answers serving as the ground truth. Each legal question was paired with a single reference answer prepared and approved by five legal experts specializing in MSME regulations to ensure factual and legal accuracy. The system-generated and expert reference answers were then compared using automated Python-based scripts employing the Natural Language Toolkit (NLTK) and ROUGE library. BLEU measured lexical and phrasal overlap (n-gram similarity) between the system and expert responses, applying smoothing functions to stabilize results across varying sentence lengths, while ROUGE assessed semantic and content similarity through ROUGE-1, ROUGE-2, and ROUGE-L variants. All scores were normalized between 0 and 1 for interpretability. This evaluation technique provides an objective quantitative measure of how closely the RAG-based LLM reproduces expert-level legal reasoning, demonstrating its capability to generate linguistically coherent and semantically accurate answers aligned with human legal expertise.

4. Results and Discussion

This section presents the implementation results and evaluation of the proposed Domain-Aware RAG system developed for Indonesian MSME legal question answering. Evaluation metrics include Precision, Recall, F1-Score, and BLEU/ROUGE to measure linguistic and semantic similarity between system outputs and reference answers.

4.1. Retrieval Performance Using TF-IDF

Applying the TF-IDF method to the SME law dataset provided various weights to each term, dependent on the extent of its relevance in the document and in the entire corpus. The certain legal terms that are more specific are found to possess greater TF-IDF weight values. This step identifies the most relevant legal keywords related to the user's query while minimizing the influence of frequent but less meaningful terms. The TF-IDF results are shown in table 1.

Table 1. TF-IDF Results

Ranking	Index	Text	TF-IDF Results
1	430	Final Income Tax calculation based on gross turnover and tariff.	0.284869
2	524	SME exemption with requirements.	0.25567
3	354	SME obligation under Final Income Tax with turnover threshold.	0.243517
4	302	Partnership taxation through Final Income Tax on business turnover.	0.241135
5	510	VAT as consumption, PPh as income tax.	0.239191
6	429	Simplified Final Income Tax scheme for SMEs with moderate turnover.	0.228042
7	497	Turnover ceiling for eligibility.	0.226792
8	522	Income Tax imposition based on taxpayer status.	0.226246
9	502	SME tax by turnover and tariff.	0.209702
10	118	Mandatory Final Income Tax and VAT obligations for SMEs.	0.195418

Each score shown in table 1 represents the TF-IDF weight obtained by applying Equations (1)–(3) to the processed MSME legal corpus. For instance, entries with terms such as 'Final Income Tax' or 'turnover threshold' produce higher TF-IDF scores because they appear more frequently in MSME tax regulations while remaining relatively uncommon in unrelated documents. This explains why these items rank at the top of the retrieval results.

4.2. Enhanced Retrieval with TF-IDF and Knowledge Graph Integration

This work constructed a corpus of SME law with legal text, rules, and relevant case studies. The corpus was used as the foundation to construct a legal knowledge graph on SME law as a domain-specific knowledge graph to improve the TF-IDF results. The SME legal knowledge graph primarily consists of information relating to SME legal cases and legal terminologies. It includes entities as cases and the relationships between sample cases and other cases, and the relationships between cases and SME legal provisions, all of which are depicted as edges in the graph. The resulting knowledge graph is shown in table 2.

Table 2. Knowledge Graph Results

Text	Match_in_KG	Match_Count	Matched_Tokens
How is SME Final Income Tax calculated?	1	2	smes, final income tax
What is the regulation on SMEs?	1	2	smes, regulation
what is NIB?	1	1	nib
How to establish a CV or PT for SMEs?	1	3	cv, pt, smes

As shown in table 2, the KG successfully identifies specific legal entities from user queries for example, “smes,” “nib,” or “cv/pt” and these matched entities are then encoded into vector embeddings using Equations (4)–(6). This embedding process strengthens semantic retrieval beyond TF-IDF’s lexical matching, providing a clearer signal for downstream LLM generation. The Knowledge Graph (KG) constructed from the corpus is shown in figure 2.

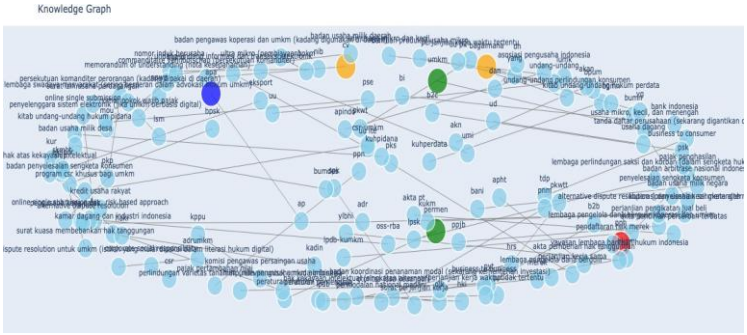

Figure 2. Knowledge Graph Based on the Corpus

Figure 2 illustrates the Knowledge Graph (KG) constructed from the Indonesian MSME legal corpus, consisting of entities such as business forms (e.g., CV, PT), licensing requirements (e.g., NIB, NPWP), government institutions, regulatory concepts, and legal procedures. Each node represents a legal entity, while edges encode semantic relations derived from two main processes: (1) clause-level co-occurrence within the same legal article, indicating conceptual association, and (2) rule-based linguistic patterns, such as *regulated-by*, *issued-by*, *required-for*, or *related-to*. Edge weights increase proportionally with the frequency of co-occurrence, allowing the KG to reflect the strength of legal relationships across documents. Figure 2 demonstrates how these enriched semantic relations guide the RAG retrieval pipeline by strengthening domain-relevant signals unseen by TF-IDF alone. Together, these visuals clarify how the KG is constructed, weighted, and operationalized within the system, thereby improving methodological transparency and replicability. The results of TF-IDF combined with the KG are shown in table 3.

Table 3. KG Improve TF-IDF Results

Text	Avg_TFIDF	Max_TFIDF	KG_Token_Found	Total_Token
SME final income tax	0.3957	1	2	6
Regulations regarding SMEs	0.6667	1	2	3
what is nib	0.6667	1	2	3

ISSN 2723-6471

Text	Avg_TFIDF	Max_TFIDF	KG_Token_Found	Total_Token
Procedures for establishing a CV or PT for SMEs	0.5161	1	3	8
Export prerequisites for SMEs	0.6	1	3	5
what is the ITE Law	0.6887	1	2	4
is QRIS legal for SMEs	0.2748	1	1	5
differences between pkwt and pkwtt	0.8	1	4	5
what is nil tax return	0.25	1	1	4
Automatic linkage of NIB and NPWP	0.4	1	2	5
does export require a special license	0.4029	1	1	5

Based on TF-IDF values and token frequency in the Knowledge Graph (KG), as indicated in table 3, all queries demonstrate a specific level of representation. The query "what is the difference between PKWT and PKWTT" posted the highest mean TF-IDF value (0.8) with 4 of 5 tokens present in the KG, indicating high relevance to the existing knowledge base. The remaining questions with comparatively higher scores are "what is the ITE Law" with Avg_TFIDF of 0.6887 and 2 matching tokens in KG, "what is NIB" and "regulations for SMEs," both with scores of 0.6667 and 2 corresponding tokens. On the other hand, "is QRIS legal for SMEs" and "what is a Nil Tax Return (SPT Nihil)" received lower mean TF-IDF scores of 0.2748 and 0.25, respectively, but also had some tokens matching in the KG. "does export require a special license" returned a mid-range score of 0.4029 with 1 matching token. These findings suggest that questions having legal or regulatory terms such as PKWT, PKWTT, NIB, and the ITE Law tend to have greater TF-IDF levels with the Knowledge Graph. More generic or functional queries, such as those pertaining to QRIS or SPT Nihil, also have lower TF-IDF values and lower token counts linked in the KG. In this framework, the KG functions as a semantic reweighting filter, where TF-IDF–ranked tokens are boosted when they appear as entities within the KG.

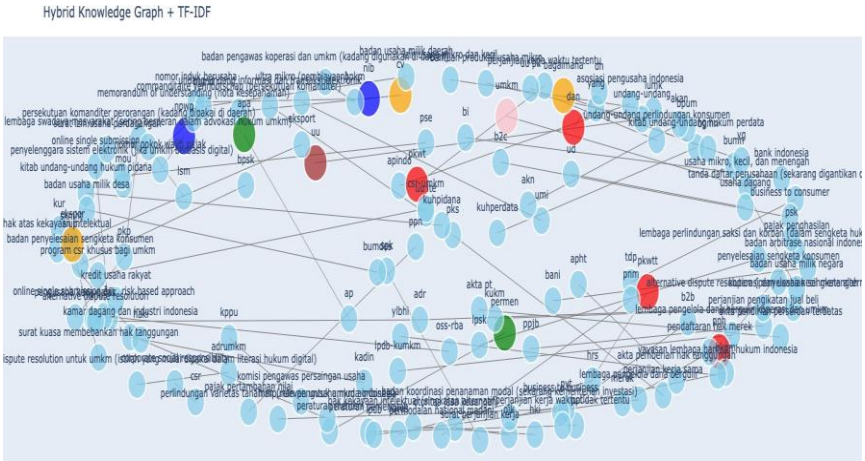

Figure 3. Combination of Knowledge Graph and TF-IDF

Figure 3 presents the hybrid Knowledge Graph + TF-IDF layer, where TF-IDF–weighted tokens are overlaid onto the KG structure to highlight domain-relevant legal entities. Colored nodes indicate terms with high TF-IDF scores that also appear as KG entities, receiving a semantic boosting effect, while blue nodes represent general legal concepts. The KG maintains its relational structure, but TF-IDF highlights strengthen entities that are both statistically frequent and semantically linked. This hybrid visualization shows how TF-IDF and KG interact as a semantic filtering mechanism to produce more context-aware retrieval signals before being passed to the LLM in the RAG pipeline.

Table 4. Accuracy Evaluation Results of Retrieval Models Using TF-IDF and TF-IDF + Knowledge Graph (KG)

Evaluation Metric	TF-IDF Model	TF-IDF + KG Model	Improvement (%)
Accuracy	0.7650	0.8350	+9.1503%
Precision	1.0000	1.0000	0%
Recall	0.5300	0.6700	+26.415%
F1-Score	0.6928	0.8024	+15.819%

 ISSN 2723-6471

As shown in table 4, the integration of the Knowledge Graph (KG) led to noticeable improvements in Recall, F1-Score, and Accuracy, while Precision remained relatively stable. This outcome is expected, as the KG broadens the search space by including semantically related legal entities. Consequently, the system retrieves more documents that are contextually relevant but not always precisely matched to the query terms. Because the number of false positives does not decrease proportionally, Precision remains unchanged even as true positives increase, explaining the stable precision despite overall performance gains. Nevertheless, the higher recall indicates that the system can capture broader legal contexts, which is particularly valuable in legal question answering tasks for MSMEs.

4.3. Generative Response Evaluation Using LLM

In the final stage, the LLM generates narrative, context-aware answers by using the integrated TF-IDF and Knowledge Graph (KG) results as retrieval-based contextual input. TF-IDF functions as a relevance filter by prioritizing key legal terms, while the KG contributes structured semantic relationships that enrich the retrieved context. The LLM then synthesizes these hybrid signals to produce coherent, legally meaningful, and easily understandable responses within the chatbot-based Question Answering system. This integration demonstrates that TF-IDF guides relevance, the KG strengthens domain knowledge, and the LLM enhances explanation quality, resulting in outputs that are not only accurate but also more interpretable and informative for MSME users.

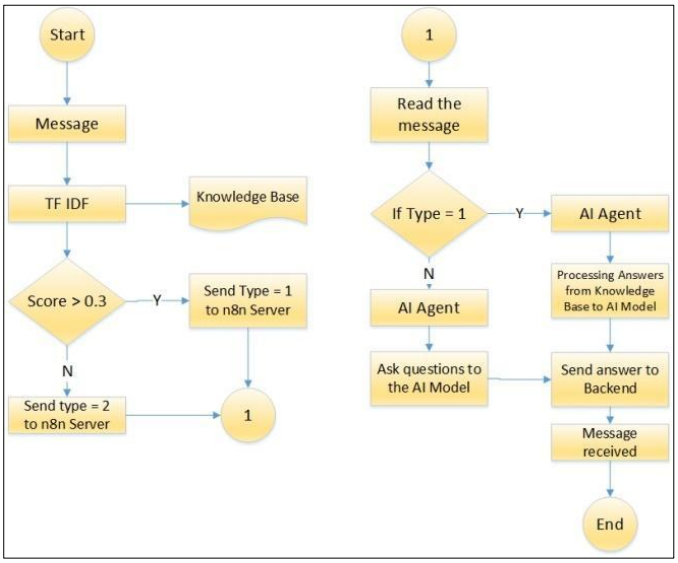

Figure 4. Workflow Architecture of Legal Question Answering System Integrating TF-IDF, KG, and LLM

Figure 4 illustrates the workflow of an enhanced hybrid legal chatbot system that integrates TF-IDF, Knowledge Graph (KG), and AI Agent components to provide contextually rich and accurate responses. The process begins when a user submits a question, which is first processed using the TF-IDF method to measure textual similarity with legal documents stored in the Knowledge Base. If the relevance score exceeds the defined threshold (e.g., 0.3), the system retrieves information directly from the knowledge base, which is further enriched by the Knowledge Graph, a semantic structure that maps relationships among legal entities, concepts, and regulations. This integration allows the system to enhance factual precision and contextual understanding beyond keyword matching. When the score is below the threshold, the message is passed to the AI Agent, which leverages the LLM to generate a response by combining retrieved data and graph-based knowledge. The generated or retrieved response is then sent to the backend and displayed to the user. This workflow demonstrates how the combination of TF-IDF retrieval, Knowledge Graph semantic enrichment, and LLM-based generation results in a more intelligent, legally coherent, and context-aware chatbot system. The example of chatbot responses is shown in figure 5.

ISSN 2723-6471

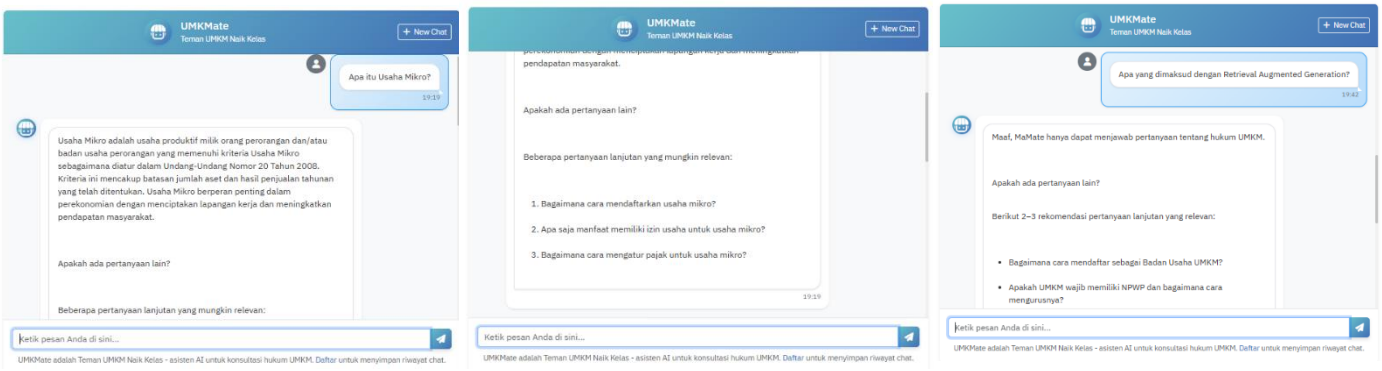

Figure 5. Example of Chatbot Responses to Legal Questions

Based on figure 5, the narrative and interactive responses generated by the chatbot are presented in table 5 and table 6.

Table 5. Sample Correct Dialogue Generated by the Legal Chatbot

User	"What is a Micro Business?"
System	"A Micro Business is a productive economic activity owned by an individual or an individual business entity that meets the criteria defined under Law Number 20 of 2008. These criteria include government-regulated thresholds for net assets and annual turnover, which facilitate the identification and classification of micro-scale enterprises."
System	"Do you have another question?" "You may also be interested in asking:" "1. How do I register a Micro Business?" "2. What are the benefits of obtaining a business license for a Micro Business?" "3. How do I manage tax obligations for a Micro Business?"

This transcript demonstrates the system's ability to retrieve and articulate legally accurate information based on TF-IDF and KG retrieval, produce coherent and context-aware explanations through the LLM generation process, and support conversational continuity by suggesting relevant follow-up questions.

Table 6. Sample Incorrect Dialogue Generated by the Legal Chatbot

User	"What is meant by Retrieval-Augmented Generation?"
System	"Sorry, I can only answer questions related to MSME legal matters."
System	"Do you have another question?" "Here are 2–3 relevant follow-up suggestions:" "You may also be interested in asking:" "1. How do I register an MSME as a legal business entity?" "2. Is an MSME required to have a Tax Identification Number (NPWP), and how do I apply for one?"

As shown in table 6, when both TF-IDF and the KG fail to detect domain-relevant terms, the system identifies the query as out-of-scope and returns a restricted response. This filtering step ensures domain consistency and prevents unnecessary LLM generation. The retrieval performance evaluation was conducted to assess the ability of the RAG system to identify and retrieve legally relevant documents based on user queries. The assessment employed three key performance metrics—Precision, Recall, and F1-score—calculated from the classification results of True Positive (TP), False Positive (FP), and False Negative (FN), derived from the expert-validated ground truth. Table 7 presents the detailed retrieval performance of the RAG system.

Table 7. RAG System Retrieval Performance Results

Metric	True Positive (TP)	False Positive (FP)	False Negative (FN)	Precision	Recall	F1-Score
RAG System	1400	240	196	0.853	0.877	0.865

 ISSN 2723-6471

As shown in table 7, the RAG system achieved 1,400 True Positives (TP), 240 False Positives (FP), and 196 False Negatives (FN), resulting in a Precision of 0.853, Recall of 0.877, and F1-score of 0.865. These values are computed at the retrieved-document level, where each of the 1,400 input legal questions may generate multiple retrieved items. The high Precision value indicates that most documents retrieved by the system were indeed relevant to the legal context of the queries, while the strong Recall score demonstrates the system's capability to capture the majority of relevant documents identified by experts. The balanced F1-score further reflects the robustness and consistency of the retrieval mechanism, confirming the system's reliability in performing accurate and comprehensive information retrieval. These results suggest that the RAG framework effectively captures semantic relationships among legal entities and concepts, going beyond simple keyword matching. Consequently, the RAG system proves to be efficient in supporting the retrieval of legally relevant, context-aware information tailored to the MSME domain.

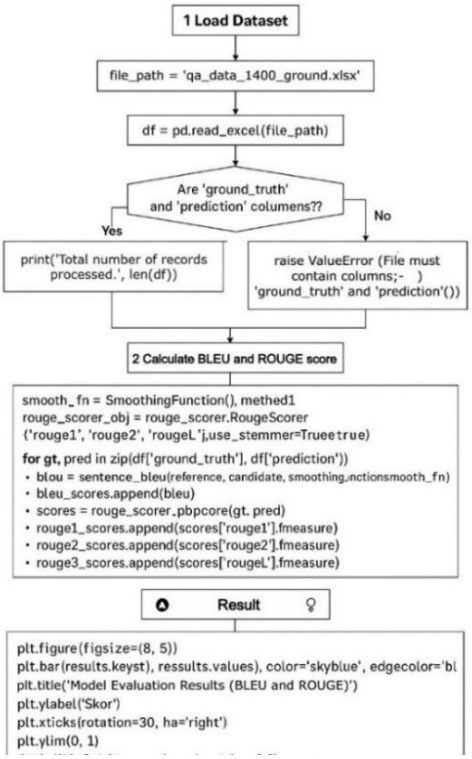

Figure 6. Workflow of BLEU and ROUGE Score Calculation for Model Evaluation

The figure 6 illustrates the evaluation workflow of the legal chatbot using BLEU and ROUGE metrics. The process begins with loading the dataset containing ground_truth and prediction columns, followed by linguistic similarity calculation between system-generated and reference answers. The results are visualized in a bar chart, showing n-gram and sequence overlap levels that reflect the model's generative quality in producing relevant, semantically accurate legal responses. The quality of the generated legal answers produced by the RAG-based system was evaluated using the BLEU and ROUGE metrics to measure linguistic and semantic alignment with expert-validated reference answers.

Table 8. Summary of BLEU and ROUGE Evaluation Metrics for Model Performance

Evaluation Metric	Score
Bleu	0.9276
ROUGE-1	0.9301
ROUGE-2	0.9275
ROUGE-L	0.9301

As shown table 8, the system achieved a BLEU score of 0.9276, indicating a high level of lexical and phrasal similarity between the system-generated and reference responses. Additionally, the ROUGE-1, ROUGE-2, and ROUGE-L scores

were 0.9301, 0.9275, and 0.9301, respectively, demonstrating that the model effectively preserves both surface-level word overlaps and deeper semantic coherence. The consistently high ROUGE values suggest that the generated legal answers closely match expert-authored responses not only in wording but also in contextual structure and meaning. These results confirm that the integration of RAG and LLM mechanisms enables the system to produce contextually accurate, linguistically coherent, and semantically aligned legal responses, making it suitable for assisting MSMEs in understanding regulatory information with near-human precision.

5. Conclusion

This study successfully developed a hybrid Domain-Aware RAG framework that integrates the strengths of TF-IDF, KG, and LLM to improve the performance of legal question-answering systems for MSMEs in Indonesia. The integration of KG significantly enhanced semantic relationships among legal entities, resulting in a 26.4% improvement in Recall and a 15.8% increase in F1-Score compared to the baseline TF-IDF model. Meanwhile, incorporating LLMs enabled the system to generate more narrative, relevant, and comprehensible responses for non-expert users, achieving BLEU and ROUGE scores above 0.92. The system remains limited by its reliance on a static, domain-specific Indonesian legal corpus, restricted transferability to other legal areas, and its absence of multilingual capabilities. Future research may extend this framework to domains such as digital contracts, data protection, or online litigation. However, such expansion poses technical challenges, including the availability of high-quality legal datasets, the effort required to construct domain ontologies for KG development, and variations in legal interpretation across jurisdictions. Addressing these challenges through automated corpus updating, semi-automated KG construction, and jurisdiction-aware modeling will be crucial for scaling the system. Overall, the findings of this study establish a foundation for AI-powered legal chatbots that can support legal literacy and regulatory compliance among MSMEs in the digital economy era.

6. Declarations

6.1. Author Contributions

Conceptualization: L.A.U, H.R, and S.H.; Methodology: L.A.U; Software: S.H.; Validation: L.A.U and H.R; Formal Analysis: L.A.U, H.R., and S.H.; Investigation: H.R.; Resources: L.A.U; Data Curation: H.R.; Writing Original Draft Preparation: L.A.U, H.R., and S.H.; Writing Review and Editing: L.A.U, and H.R; Visualization: S.H.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

This research received financial support in 2025 from the Ministry of Higher Education, Science, and Technology (Kemdiktisaintek), Indonesia.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Ajay Mukund and K. S. Easwarakumar, “Optimizing Legal Text Summarization Through Dynamic Retrieval-Augmented Generation and Domain-Specific Adaptation,” *Symmetry (Basel).*, vol. 17, no. 5, pp. 1-12, 2025, doi:

10.3390/sym17050633.

[2] W. Zhuohao, W. Dong, and L. Qing, "Keyword Extraction from Scientific Research Projects Based on SRP-TF-IDF," *Chinese J. Electron*., vol. 30, no. 4, pp. 652–657, 2021.

[3] I. Radeva, I. Popchev, L. Doukovska, and M. Dimitrova, "Web Application for Retrieval-Augmented Generation: Implementation and Testing," *Electron*., vol. 13, no. 7, pp. 1-12, 2024, doi: 10.3390/electronics13071361.

[4] N. A. Akbar, R. Dembani, B. Lenzitti, and D. Tegolo, "RAG-Driven Memory Architectures in Conversational LLMs—A Literature Review With Insights Into Emerging Agriculture Data Sharing," *IEEE Access*, vol. 13, no. June, pp. 123855–123880, 2025, doi: 10.1109/ACCESS.2025.3589241.

[5] L. C. Chen, M. S. Pardeshi, Y. X. Liao, and K. C. Pai, "Application of retrieval-augmented generation for interactive industrial knowledge management via a large language model," *Comput. Stand. Interfaces*, vol. 94, no. September 2024, pp. 1-15, 2025, doi: 10.1016/j.csi.2025.103995.

[6] L. Ching Chen, "An extended TF-IDF method for improving keyword extraction in traditional corpus- based research: An example of a climate change corpus," *Data Knowl. Eng*., vol. 153, no. 1, pp. 1-12, 2024, doi: https://doi.org/10.1016/j.datak.2024.102322.

[7] Y. Nuri and E. Senyurek, "Filtering articles based on their abstracts using TF-IDF," *Int. J. Adv. Eng. Manag*., vol. 6, no. 08, pp. 364–368, 2024, doi: 10.35629/5252-0608364368.

[8] H. J. Kim, J. W. Baek, and K. Chung, "Optimization of associative knowledge graph using TF-IDF based ranking score," *Appl. Sci.*, vol. 10, no. 13, pp.1-12, 2020, doi: 10.3390/app10134590.

[9] Y. Wang, "Research on the TF–IDF algorithm combined with semantics for automatic extraction of keywords from network news texts," *J. Intell. Syst*., vol. 33, no. 1, pp.1-12, 2024, doi: 10.1515/jisys-2023-0300.

[10] H. Liu, X. Chen, and X. Liu, "A Study of the Application of Weight Distributing Method Combining Sentiment Dictionary and TF-IDF for Text Sentiment Analysis," *IEEE Access,* vol. 10, no. 1, pp. 32280– 32289, 2022, doi: 10.1109/ACCESS.2022.3160172.

[11] J. Zhou, Z. Ye, S. Zhang, Z. Geng, N. Han, and T. Yang, "Investigating response behavior through TF- IDF and Word2vec text analysis: A case study of PISA 2012 problem-solving process data," *Heliyon*, vol. 10, no. 16, pp. 1-15, 2024, doi: 10.1016/j.heliyon.2024.e35945.

[12] A. R. Lubis, M. K. M. Nasution, O. S. Sitompul, and E. M. Zamzami, "The effect of the TF-IDF algorithm in times series in forecasting word on social media," *Indones. J. Electr. Eng. Comput. Sci*., vol. 22, no. 2, pp. 976–984, 2021, doi: 10.11591/ijeecs.v22.i2.pp976-984.

[13] Y. Wang, D. Zhang, Y. Yuan, K. Liu, and Y. Yang, "Improvement of TF-IDF Algorithm Based on Knowledge Graph," *in 2018 IEEE 16th International Conference on Software Engineering Research, Management and Applications (SERA)*, vol. 1, no. 1, pp. 19–24, 2018, doi: 10.1109/SERA.2018.8477196.

[14] M. M. Hussien, A. N. Melo, A. L. Ballardini, C. S. Maldonado, R. Izquierdo, and M. Á. Sotelo, "RAG- based explainable prediction of road users behaviors for automated driving using knowledge graphs and large language models," *Expert Syst. Appl.*, vol. 265, no. July 2024, pp. 1-14, 2025, doi: 10.1016/j.eswa.2024.125914.

[15] I. Harrando and R. Troncy, "Combining Semantic and Linguistic Representations for Media Recommendation," *HAL Open Sci.*, vol. 2022, no. Jan., pp. 1–10, 2022.

[16] X. Kehan, Z. Kun, L. Jingyuan, and W. Yuanzhuo, "CRP-RAG: A Retrieval-Augmented Generation Framework for Supporting Complex Logical Reasoning and Knowledge Planning," *Electron*., vol. 14, no. 1, pp. 1–34, 2025, doi: 10.3390/electronics14010047.

[17] Y. Li, V. Zakhozhyi, D. Zhu, and L. J. Salazar, "Domain Specific Knowledge Graphs as a Service to the Public: Powering Social-Impact Funding in the US," *in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 2020, no. 1, pp. 2793–2801, 2020, doi: 10.1145/3394486.3403330.

[18] Z. Wang, Z. Liu, W. Lu, and J. Lu, "Improving knowledge management in building engineering with hybrid retrieval-augmented generation framework," *J. Build. Eng*., vol. 103, no. 1, pp. 1-12, 2025, doi: https://doi.org/10.1016/j.jobe.2025.112189.

[19] V. Armant et al., "Can Knowledge Graphs and Retrieval-Augmented Generation be combined to Explain Query / Answer

Relationships Truthfully ?" *HAL Open Sci.*, vol. 2025, no. Jan., pp. 1–15, 2025.

[20] A. Zubiaga, "Natural language processing in the era of large language models," *Front. Artif. Intell*., vol. 6, no.1, pp. 1-12, 2023, doi: 10.3389/frai.2023.1350306.

[21] L. Bahr, C. Wehner, J. Wewerka, J. Bittencourt, U. Schmid, and R. Daub, "Knowledge graph enhanced retrieval-augmented generation for failure mode and effects analysis," *J. Ind. Inf. Integr.*, vol. 45, no. February, pp. 1-12, 2025, doi: 10.1016/j.jii.2025.100807

[22] Z. Li, Z. Wang, W. Wang, K. Hung, H. Xie, and F. L. Wang, "Retrieval-augmented generation for educational application: A systematic survey," *Comput. Educ. Artif. Intell*., vol. 8, no. May, pp. 1-17, 2025, doi: 10.1016/j.caeai.2025.100417.

[23] B. Han, T. Susnjak, and A. Mathrani, "Automating Systematic Literature Reviews with Retrieval- Augmented Generation: A Comprehensive Overview," *Appl. Sci*., vol. 14, no. 19, pp.1-12, 2024, doi: 10.3390/app14199103.

[24] L. Xu, L. Lu, M. Liu, C. Song, and L. Wu, "Nanjing Yunjin intelligent question-answering system based on knowledge graphs and retrieval augmented generation technology," *Herit. Sci*., vol. 12, no. 1, pp. 1–23, 2024, doi: 10.1186/s40494-024-01231-3.

[25] A. O. M. Saleh, G. Tur, and Y. Saygin,"SG-RAG MOT : SubGraph Retrieval Augmented Generation with Merging and Ordering Triplets for Knowledge Graph Multi-Hop Question Answering," *AI Res. J.*, vol. 2025, no. Jan., pp. 1–24, 20252025.

[26] Y. Shang, "Empowering knowledge graphs with hybrid retrieval-augmented generation for the intelligent mix scheme of mass concrete," *Case Stud. Constr. Mater*., vol. 23, no. May, pp. 1-19, 2025, doi: 10.1016/j.cscm.2025.e04979.

[27] M. DeBellis, N. Dutta, G. Jacob, and A. Balaji, "Integrating Ontologies and Large Language Models to Implement Retrieval Augmented Generation," *Appl. Ontol.*, vol. 19, no. 4, pp. 389–407, 2025, doi: 0.1177/15705838241296446.

[28] J. Liang, "GeoGraphRAG: A graph-based retrieval-augmented generation approach for empowering large language models in automated geospatial modeling," *Int. J. Appl. Earth Obs. Geoinf*., vol. 142, no. June, pp. 1-12, 2025, doi: 10.1016/j.jag.2025.104712.

[29] J. Miao, C. Thongprayoon, S. Suppadungsuk, O. A. Garcia Valencia, and W. Cheungpasitporn, "Integrating Retrieval-Augmented Generation with Large Language Models in Nephrology: Advancing Practical Applications," *Med.*, vol. 60, no. 3, pp. 1–15, 2024, doi: 10.3390/medicina60030445.

[30] J. Brand, A. Israeli, and D. Ngwe, "Using LLMs for Market Research," *J. Mark. Res.*, vol. 2024, no. Jan., pp. 1–10, 2024.

[31] A. Petukhova and N. Fachada, "TextCL: A Python package for NLP preprocessing tasks," *SoftwareX,* vol. 19, no. 1, pp. 1-12, 2022, doi: 10.1016/j.softx.2022.101122.

[32] B. Probierz and J. Kozak, "Knowledge graphs to an analysis and visualization of texts from scientific articles," *Procedia Comput. Sci*., vol. 225, no.1, pp. 4324–4333, 2023, doi: 10.1016/j.procs.2023.10.429.

[33] A. Erfina and M. R. N. R. Alamsyah, "Implementation of Naive Bayes classification algorithm for Twitter user sentiment analysis on ChatGPT using Python programming language," *Data Metadata*, vol. 2, no.1, pp. 2–11, 2023, doi: 10.56294/dm202345.

[34] H. Mehta and K. Passi, "Social media hate speech detection using explainable AI," *Algorithms*, vol. 15, no. 8, pp. 1-12, 2022.

[35] L. R. Halim and A. Suryadibrata, "Cyberbullying Sentiment Analysis with Word2Vec and One- Against-All Support Vector Machine," *Int. J. New Media Technol.*, vol. 8, no. 1, pp. 57-71, 2021.

[36] M. A. Palomino and F. Aider, "Evaluating-the-Effectiveness-of-Text-PreProcessing-in-Sentiment- AnalysisApplied Sciences-Switzerland.pdf," *Appl. Sci.*, vol. 12, no. Jan., pp. 8765–8775, 2022.

[37] A. Jabbar, M. I. Tamimy, A. Akhunzada, S. Iqbal, and S. Hussain, "Empirical evaluation and study of text stemming algorithms," *Artif. Intell. Rev*., vol. 53, no. 1, pp. 5559–5588, 2020, doi: 10.1007/s10462-020-09828-3.

[38] A. Farhan AlShammari, "Implementation of Keyword Extraction using Term Frequency-Inverse Document Frequency (TF-IDF) in Python," *Int. J. Comput. Appl*., vol. 185, no. 35, pp. 975–8887, 2023.

[39] L. Zhang, "Features extraction based on Naive Bayes algorithm and TF-IDF for news classification," *PLoS One*, vol. 20, no. 7 July, pp. 1–17, 2025, doi: 10.1371/journal.pone.0327347.

[40] M. Atef Mosa, "Predicting Semantic Categories in Text Based on Knowledge Graph Combined with Machine Learning Techniques," *Appl. Artif. Intell.*, vol. 35, no. 12, pp. 933–951, 2021, doi: 10.1080/08839514.2021.1966883.

[41] Q. Wu, "Integrating Knowledge Graph and Machine Learning Methods for Landslide Susceptibility Assessment," *Remote Sens.*, vol. 16, no. 13, pp.1-12, 2024, doi: 10.3390/rs16132399.

[42] A. Chen, Y. Tian, J. Zhang, C. Li, and H. Zhang, "LLM-based intelligent Q & A system for railway locomotive maintenance standardization," *Sci. Rep.*, vol. 2025, no. Jan., pp. 1–12, 2025, doi: 10.1038/s41598-025-96130-3.

[43] C. Gan, Q. Zhang, and T. Mori, "Application of LLM Agents in Recruitment : A Novel Framework for Automated Resume Screening," *J. Inf. Process.*, vol. 32, no. 1, pp. 881–893, 2024, doi: 10.2197/ipsjjip.32.881.

[44] W. Shi, W. Zheng, J. X. Yu, H. Cheng, and L. Zou, "Keyphrase Extraction Using Knowledge Graphs," *Data Sci. Eng.*, vol. 2, no. 4, pp. 275–288, 2017, doi: 10.1007/s41019-017-0055-z.

[45] N. U. R. Izyan, Y. Saat, M. Mohd, S. Azman, M. Noah, and S. M. Al-ghuribi, "Beyond Relevance : Enhancing Serendipity in Content-Based Recommendations With Knowledge Graphs," *IEEE Access,* vol. 13, no. August, pp. 142980–142989, 2025, doi: 10.1109/ACCESS.2025.3598342.

[46] C. Peng, F. Xia, M. Naseriparsa, and F. Osborne, "Knowledge Graphs : Opportunities and Challenges," *Artif. Intell. Rev.*, vol. 56, no. 11, pp. 13071–13102, 2023, doi: 10.1007/s10462-023-10465-9.

[47] K. Hou, J. Li, Y. Liu, S. Sun, H. Zhang, and H. Jiang, "KG-EGV: A Framework for Question Answering with Integrated Knowledge Graphs and Large Language Models," *Electronics*, vol. 13, no. 23, pp. 1–23, 2024, doi: 10.3390/electronics13234835.

[48] F. Moons and E. Vandervieren, "Measuring agreement among several raters classifying subjects into one or more (hierarchical) categories : A generalization of Fleiss ' kappa," *Behav. Res. Methods*, vol. 57, no. 1, pp. 1–19, 2025, doi: 10.3758/s13428-025-02746-8.