



HAL
open science

Business-RAG: Information Extraction for Business Insights

Muhammad Arslan, Christophe Cruz

► **To cite this version:**

Muhammad Arslan, Christophe Cruz. Business-RAG: Information Extraction for Business Insights. 21st International Conference on Smart Business Technologies, Jul 2024, Dijon, France. pp.88 - 94, 10.5220/0012812800003764 . hal-04862172

HAL Id: hal-04862172

<https://hal.science/hal-04862172v1>

Submitted on 9 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Business-RAG: Information Extraction for Business Insights

Muhammad Arslan^a and Christophe Cruz^b

Laboratoire Interdisciplinaire Carnot de Bourgogne (ICB), Université de Bourgogne, Dijon, France

Keywords: Business Intelligence (BI), Decision-Making, Information Extraction (IE), Large Language Models (LLMs), Natural Language Processing (NLP), Retrieval-Augmented Generation (RAG).

Abstract: Enterprises depend on diverse data like invoices, news articles, legal documents, and financial records to operate. Efficient Information Extraction (IE) is essential for extracting valuable insights from this data for decision-making. Natural Language Processing (NLP) has transformed IE, enabling rapid and accurate analysis of vast datasets. Tasks such as Named Entity Recognition (NER), Relation Extraction (RE), Event Extraction (EE), Term Extraction (TE), and Topic Modeling (TM) are vital across sectors. Yet, implementing these methods individually can be resource-intensive, especially for smaller organizations lacking in Research and Development (R&D) capabilities. Large Language Models (LLMs), powered by Generative Artificial Intelligence (GenAI), offer a cost-effective solution, seamlessly handling multiple IE tasks. Despite their capabilities, LLMs may struggle with domain-specific queries, leading to inaccuracies. To overcome this challenge, Retrieval-Augmented Generation (RAG) complements LLMs by enhancing IE with external data retrieval, ensuring accuracy and relevance. While the adoption of RAG with LLMs is increasing, comprehensive business applications utilizing this integration remain limited. This paper addresses this gap by introducing a novel application named Business-RAG, showcasing its potential and encouraging further research in this domain.


1 INTRODUCTION


Enterprises depend heavily on data from various sources like invoices, surveys, and legal documents for their operations (Abdullah et al., 2023). Extracting relevant information from this data, known as IE, is essential for informed decision-making. NLP, a field of AI, automates IE tasks such as NER, RE, EE, TE, and TM (Abdullah et al., 2023). NER identifies and classifies entities like people and organizations, RE finds relationships between entities, EE extracts information about events, TE identifies key contextual terms, and TM uncovers prevalent topics within text documents. These tasks streamline operations across sectors like finance, healthcare, and manufacturing (Martinez-Rodriguez et al., 2020).

While studies have explored NER, RE, EE, TE, and TM within business, implementing these methods often demands expertise and resources (Abdullah et al., 2023; Martinez-Rodriguez et al., 2020). Rather than employing distinct methods for

each IE task, LLMs grounded in NLP and powered by GenAI offer a superior solution. GenAI, a subdivision of AI, possesses the capability to generate diverse content types including text, images, etc. (Feuerriegel et al., 2024).

LLMs are trained on vast datasets, they excel in discerning connections between words and phrases, thereby enhancing contextual understanding and producing the most pertinent information (Feuerriegel et al., 2024). By leveraging LLMs, organizations can swiftly develop AI-powered applications (Arslan & Cruz 2024) that efficiently extract vital information from diverse sources, thereby streamlining business processes and minimizing manual labor and model development costs (Feuerriegel et al., 2024). However, LLMs face challenges with domain-specific queries and may produce inaccurate or irrelevant information (Kandpal et al., 2023). To address this, RAG integrates external data retrieval into the generative process, improving output accuracy and relevance (Lewis et al., 2020).

^a  <https://orcid.org/0000-0003-3682-7002>

^b  <https://orcid.org/0000-0002-5611-9479>

Although the integration of RAG with LLMs has experienced significant growth recently, there is a notable scarcity of studies focusing on developing applications using this integration for business-related IE domains. To address this gap, this paper begins by reviewing existing studies on Business IE and associated NLP tasks, followed by an overview of RAG with LLMs applications. Finally, it introduces a novel application, Business-RAG, showcasing the implementation of RAG with the LLM in developing a business IE solution.

The paper follows this structure: Section 2 explores relevant literature on Business IE and existing applications of RAG with LLMs. In Section 3, we introduce our innovative business application, Business-RAG, which utilizes RAG with the LLM for Business IE. Section 4 discusses the advantages and limitations of the system. Finally, Section 5 provides concluding remarks.

2 BACKGROUND

The literature relevant to this study can be divided into two main sections. The first section introduces Business IE, emphasizing its significance in informing business decisions and reviewing previous studies utilizing different NLP techniques for Business IE tasks. The second section introduces applications of RAG with LLMs. Finally, the study identifies a gap by underscoring the scarcity of research on RAG with LLMs specifically tailored for Business IE.

2.1 Business IE

IE encompasses retrieving structured data from various sources like text documents, emails, web pages, and databases (Martinez-Rodriguez et al., 2020). This structured data includes entities, relationships, events, and other pertinent facts, extracted via techniques from NLP, machine learning, and computational linguistics (Martinez-Rodriguez et al., 2020). Specifically tailored to meet the needs of businesses and organizations, Business IE focuses on extracting insights from diverse business-related sources such as financial reports, customer feedback, market research, and legal documents (de Almeida Bordignon et al., 2018). Extracted information may comprise key performance indicators, market trends, customer sentiments, competitive analysis, and regulatory compliance data, crucial for informed decision-making.

Five primary IE tasks are identified within Business IE: NER, RE, EE, TE, and TM, all grounded in NLP principles (Martinez-Rodriguez et al., 2020). Various techniques are employed for Business IE tasks, including Bidirectional Encoder Representations from Transformers (BERT)-based models (Devlin et al., 2018; Arslan & Cruz 2024), ontology-based methods (Arendarenko & Kakkonen 2012), machine learning algorithms (Sun, 2022; Piskorski et al., 2021), syntactic and semantic rules, Term Frequency - Inverse Document Frequency (TF-IDF) approaches (Bzhalava et al., 2024), rule-based approaches (Korger & Baumeister, 2021), and deep learning models (Bellan et al., 2022). These methods enable businesses to automate processes, make informed decisions, and derive valuable insights (Arslan & Cruz 2022) from their data. However, challenges arise such as the need for separate systems for each IE task, difficulty in selecting and validating accuracy, and limited accessibility to proprietary data for model training.

2.2 RAG with LLMs

Considering above-mentioned obstacles, the potential of RAG with LLMs rooted in advanced NLP principles is under investigation. LLMs, represent sophisticated machine learning models extensively trained on textual data to replicate human-like text generation (Raiaan et al., 2024). RAG enhances LLMs by incorporating an initial step where they retrieve relevant information from external sources before generating text (Lewis et al., 2020). This integration improves the accuracy and relevance of the output, reducing errors and enhancing overall information quality (Lewis et al., 2020).

The application of RAG with LLMs spans a diverse array of domains and tasks, encompassing fields such as biomedical research, finance, healthcare, education, software development, and humanitarian aid (Gao et al., 2023). This advanced technology facilitates tasks ranging from Question Answering (QA) in medical and financial contexts to summarizing medical texts, generating book reviews guided by reference documents, and aiding clinical decision-making (Zhao et al., 2023; Li et al., 2022).

Additionally, it supports educational decision making, enterprise search, sentiment classification, health education, technical product information QA, and software development and maintenance (Li et al., 2022). Moreover, it plays a role in generating realistic images, combating online hate speech, classifying scientific documents, generating entity descriptions, translating text to SQL, and facilitating open-domain

question answering and fact verification (Zhao et al., 2023). Its applications extend to professional knowledge QA, multicultural enterprise QA, personalized dialogue systems, event argument extraction, intelligence report generation, short-form open-domain QA, automated cash transaction booking, question answering with private data, scientific document classification, clinical-related writing, and pharma industry regulatory compliance QA (Zhao et al., 2023; Li et al., 2022). This breadth of applications underscores the versatility and potential impact of RAG with LLMs in enhancing information processing, decision-making, and automation across various sectors.

It is evident that there is a lack of research on RAG with LLMs specifically designed for business IE, a gap that could significantly enhance IE for deriving business insights. To address this shortfall, the following section introduces Business-RAG, showcasing the application of this integrated approach of RAG with LLMs within the business domain.

3 BUSINESS-RAG

To develop a RAG system tailored for Business IE, we have adopted the foundational principles of EE. Within this frame, each business event encapsulates a unique collection of named entities, intricately linked through RE algorithms. Moreover, every business event incorporates a spectrum of business terms, carefully categorized under relevant business topics. For instance, consider a Merger and Acquisition (M&A) event as an exemplar of a business event. In this scenario, named entities could include the companies involved, key executives, financial figures, and regulatory bodies, while relation extraction algorithms discern the nature and implications of the transaction. Concurrently, business terms such as "stock purchase agreement", "synergy", and "integration strategy" are categorized under appropriate topics like "corporate finance" or "strategic management". This structured approach serves as the backbone for our Business-RAG system, enabling the extraction of nuanced insights from vast datasets.

To create an interactive system tailored for business insights, we developed a specialized business chatbot assistant with a user-friendly interface (UI) using Chainlit.io, integrated with the LLM (see Fig.1 and Fig. 2). Our exploration encompassed various publicly available LLM releases, such as BLOOM (Le Scao et al., 2022),

Falcon (Zhang et al., 2023), GPT-4 (Achiam et al., 2023), Llama2 (Touvron et al., 2023), and Chinchilla (Hoffmann et al., 2022). Ultimately, we chose Llama2 (Touvron et al., 2023) for its superior performance, attributed to its comprehensive training on diverse datasets.

To tailor Llama2 for Business IE-related tasks and harness insights from business news articles, we integrated RAG technology into our solution. This strategic fusion enables the chatbot to handle business inquiries precisely using extracted data. The chatbot's capability depends on the richness of the provided business dataset. Our dataset comprised 2,845 news articles from online platforms in March 2023. In the following discussion, we will explore the utility of news articles in generating various types of business insights (see Table 1).

- 1) **Business Prospects:** Utilizing Business-RAG, businesses can analyze news articles to identify emerging market trends, potential investment opportunities, and upcoming business ventures. By extracting information related to new product launches, partnerships, and market expansions, companies can gain insights into promising business prospects and strategic opportunities.
- 2) **Retail Customization:** With Business-RAG, retailers can analyze customer sentiment, preferences, and purchasing patterns from news articles to personalize their offerings. By extracting data on consumer trends, product reviews, and shopping behaviors, retailers can tailor their product assortments, pricing strategies, and marketing campaigns to better meet the needs and preferences of their target audience.
- 3) **Financial Prediction:** Business-RAG enables financial institutions to extract relevant data from news articles to predict market trends, stock performance, and economic indicators. By analyzing information on financial reports, industry analyses, and regulatory changes, organizations can make informed decisions regarding investments, asset allocations, and risk management strategies.
- 4) **Competitive Assessment:** Businesses can use Business-RAG to gather competitive intelligence from news articles, enabling them to assess competitor strategies, market positioning, and industry trends. By extracting data on competitor product launches, partnerships, and acquisitions, companies can identify areas for differentiation, market opportunities, and potential threats to their business.
- 5) **Customer Opinion Analysis:** By leveraging Business-RAG, organizations can extract and analyze customer opinions, feedback, and reviews

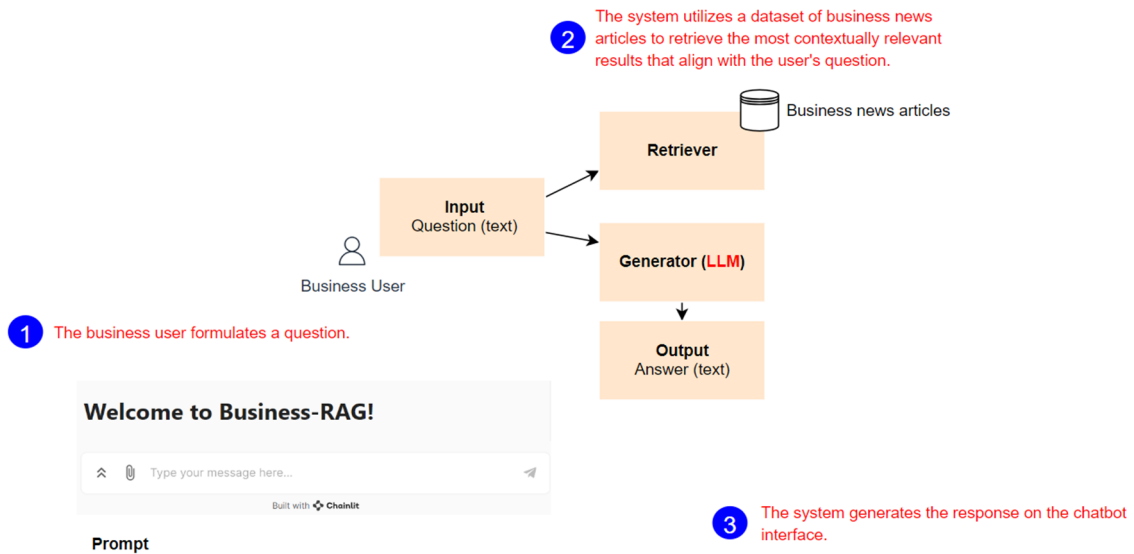


Figure 1: Illustrates the graphical representation of a Business-RAG Application: Here, a business user interacts with the RAG system powered by the LLM, querying it with external datasets of business news articles to generate a response.



Figure 2: Showcases the design of a business chatbot assistant, facilitating business users to inquire about news articles-related queries.

from news articles to understand customer sentiments and preferences. By extracting data on customer experiences, satisfaction levels, and brand perceptions, businesses can identify areas for improvement, develop targeted marketing strategies, and enhance customer engagement.

6) Regulatory Compliance and Risk Mitigation: Business-RAG enables organizations to extract regulatory updates, compliance requirements, and legal developments from news articles to ensure regulatory compliance and mitigate legal risks.

7) Brand Reputation Tracking: Businesses can use Business-RAG to monitor news articles for mentions of their brand, products, or services, enabling them to track brand reputation and public perception. By extracting data on brand mentions, sentiment analysis, and media coverage, organizations can identify potential reputation risks, respond to emerging issues, and protect their brand image.

8) Supply Chain Optimization: With Business-RAG, organizations can extract supply chain data from news articles to optimize logistics, mitigate risks,

and improve operational efficiency. By analyzing information on supplier relationships, market trends, and supply chain disruptions, companies can identify opportunities for cost savings, inventory management, and supply chain optimization strategies.

4 DISCUSSION

Extracting pertinent business data from diverse textual sources presents a significant challenge due to the inherent variability in data structures and formats. This complexity necessitates the application of a diverse set of IE tasks, including NER, RE, EE, TE, and TM, tailored to the specific requirements of the organization. Each of these tasks plays a crucial role in distilling valuable insights from the vast array of textual information available.

Traditionally, implementing these IE tasks often involves the deployment of separate systems or tools, each specializing in a particular aspect of IE. However, managing and integrating multiple systems can be complex, resource-intensive, and time-consuming. Leveraging GenAI with LLMs presents a promising alternative, offering sophisticated language understanding capabilities across a wide range of domains.

While LLMs are not inherently domain-specific, their versatility and ability to process vast amounts of text data make them attractive candidates for various

applications, including business information extraction. To tailor LLMs for business-specific tasks, we integrated RAG technology, which enhances the generative capabilities of LLMs by incorporating an initial step of information retrieval from external sources.

Our system offers flexibility in adapting to evolving business requirements, dynamic updates in real-time, and transparent source attribution, which enhances the credibility and reliability of the extracted information. However, it is essential to acknowledge the limitations of our approach, such as reliance on a single LLM model, namely Llama2. Looking ahead, future research could focus on exploring alternative LLM models tailored specifically for the business domain could further enhance the performance and effectiveness of information extraction systems in real-world business settings.

5 CONCLUSION

The integration of RAG with LLMs represents a promising advancement in IE, particularly within the business realm. By leveraging LLMs rooted in NLP principles, this integration offers an efficient and cost-effective solution capable of performing multiple IE tasks simultaneously. Despite the considerable growth in the integration of RAG with LLMs in recent years, the existing literature lacks

Table 1: Sample Records of User Queries.

Business Area		Related Questions
1	Business Prospects	- What new business ventures are being launched? - Which companies are expanding into new markets or industries? - Who are the key stakeholders involved in upcoming business initiatives?
2	Retail Customization	- Which companies are implementing personalized shopping experiences? - Who are the key figures driving innovation in retail customization? - What customer feedback or preferences are influencing retail customization strategies?
3	Financial Prediction	- Who are the financial analysts making predictions about market trends? - Where are the emerging opportunities for investment in various sectors? - Who are the key players in the financial industry making significant predictions?
4	Competitive Assessment	- Who are the major competitors in a specific market segment? - Where are the locations of new competitors entering the market? - Who are the key executives leading competitive initiatives within companies?
5	Customer Opinion Analysis	- What are customers saying about specific products or services? - Who are the influencers shaping customer opinions in certain industries? - Where are the locations with the highest customer satisfaction ratings?
6	Regulatory Compliance and Risk Mitigation	- What new regulations are being implemented in various industries? - Who are the regulatory authorities overseeing compliance standards? - What compliance violations or risks are companies facing?
7	Brand Reputation Tracking	- What are the sentiments surrounding specific brands in the market? - Where are the locations with the strongest brand presence or recognition? - What strategies are companies using to enhance brand reputation?
8	Supply Chain Optimization	- What disruptions or innovations are affecting supply chains? - Who are the key players optimizing supply chain operations? - What trends are emerging in supply chain management practices?

comprehensive studies specific to the business domain. To address these gaps, this paper conducts a review of existing research on business IE and associated tasks, shedding light on the potential applications of RAG with LLMs. Furthermore, it introduces a novel real-world application that showcases the practical implementation of this integration in developing a Business IE application. While this application is still in its developmental stages, thorough evaluation is necessary, which constitutes the future work of this study. The primary aim is to illustrate the adaptability and efficiency of RAG with LLMs within the realm of business operations.

DATASET AVAILABILITY

<https://drive.google.com/file/d/18UB-TamXvCFpqq0edfH7EVPjB19Ec34dC/view?usp=sharing>

ACKNOWLEDGEMENTS

The authors express gratitude to the French government for the National Research Agency (ANR) funding and extend appreciation to Cyril Nguyen Van (company: FirstEco) for generously providing the dataset.

REFERENCES

- Abdullah, M. H. A., Aziz, N., Abdulkadir, S. J., Alhussian, H. S. A., & Talpur, N. (2023). Systematic literature review of information extraction from textual data: recent methods, applications, trends, and challenges. *IEEE Access*, 11, 10535-10562.
- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.
- Arendarenko, E., & Kakkonen, T. (2012). Ontology-based information and event extraction for business intelligence. In *Artificial Intelligence: Methodology, Systems, and Applications: 15th International Conference, AIMS 2012, Varna, Bulgaria, September 12-15, 2012. Proceedings 15* (pp. 89-102). Springer Berlin Heidelberg.
- Arslan, M., & Cruz, C. (2022). Extracting Business Insights through Dynamic Topic Modeling and NER. In *KDIR* (pp. 215-222).
- Arslan, M., & Cruz, C. (2024). Business text classification with imbalanced data and moderately large label spaces for digital transformation. *Applied Network Science*, 9(1), 11.
- Bellan, P., Dragoni, M., & Ghidini, C. (2022, September). Extracting business process entities and relations from text using pre-trained language models and in-context learning. In *International Conference on Enterprise Design, Operations, and Computing* (pp. 182-199). Cham: Springer International Publishing.
- Bzhalava, L., Kaivo-oja, J., & Hassan, S. S. (2024). Digital business foresight: Keyword-based analysis and CorEx topic modeling. *Futures*, 155, 103303.
- de Almeida Bordignon, A. C., Thom, L. H., Silva, T. S., Dani, V. S., Fantinato, M., & Ferreira, R. C. B. (2018, June). Natural language processing in business process identification and modeling: a systematic literature review. In *Proceedings of the XIV Brazilian Symposium on Information Systems* (pp. 1-8).
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative ai. *Business & Information Systems Engineering*, 66(1), 111-126.
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997.
- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... & Sifre, L. (2022). Training compute-optimal large language models. arXiv preprint arXiv:2203.15556.
- Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2023, July). Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning* (pp. 15696-15707). PMLR.
- Korger, A., & Baumeister, J. (2021, September). Rule-based Semantic Relation Extraction in Regulatory Documents. In *LWDA* (pp. 26-37).
- Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., ... & Al-Shaibani, M. S. (2022). Bloom: A 176b-parameter open-access multilingual language model.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.
- Li, H., Su, Y., Cai, D., Wang, Y., & Liu, L. (2022). A survey on retrieval-augmented text generation. arXiv preprint arXiv:2202.01110.
- Martinez-Rodriguez, J. L., Hogan, A., & Lopez-Arevalo, I. (2020). Information extraction meets the semantic web: a survey. *Semantic Web*, 11(2), 255-335.
- Piskorski, J., Stefanovitch, N., Jacquet, G., & Podavini, A. (2021, April). Exploring linguistically-lightweight keyword extraction techniques for indexing news articles in a multilingual set-up. In *Proceedings of the EACL Hackashop on news media content analysis and automated report generation* (pp. 35-44).
- Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., ... & Azam, S. (2024). A

review on large Language Models: Architectures, applications, taxonomies, open issues and challenges. IEEE Access.

Sun, T. (2022). Relation Extraction from Financial Reports (Doctoral dissertation, University of York).

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Zhao, R., Chen, H., Wang, W., Jiao, F., Do, X. L., Qin, C., ... & Joty, S. (2023). Retrieving multimodal information for augmented generation: A survey. arXiv preprint arXiv:2303.10868.

ZXhang, Y. X., Haxo, Y. M., & Mat, Y. X. (2023). Falcon llm: A new frontier in natural language processing. AC Investment Research Journal, 220(44).

