Towards Collaborative Business Intelligence with Conversational Agents: A User-Centric Approach

Olga Cherednichenko* and Muhammad Fahad**

Univ Lyon, Univ Lyon 2, UR ERIC, 5 avenue Mendès France, 69676 Bron Cedex, France
*olga.cherednichenko@univ-lyon2.fr
**fahad.muhammad@univ-lyon2.fr
https://eric.univ-lyon2.fr/bi4people/

Résumé. In the dynamic landscape of contemporary business, the demand for cutting-edge methodologies integrating intelligent technologies to facilitate effective decision-making is on the rise. Business Intelligence systems have become indispensable tools for decision-making across diverse organizations, and the emergence of Collaborative Business Intelligence (CBI) extends the horizons of informed decision-making by tapping into external information from varied sources. The central aim of this research centers on the development of a model tailored for CBI processes within distributed virtual teams. This is achieved through the interaction between users and a CBI Virtual Assistant. The key role of the virtual assistant is to augment the accessibility of data exploration for a wider user base, concurrently minimizing the time and effort required for data analysis. The proposed model, through its integration of intelligent technologies, seeks to propel collaborative decision-making to new heights, emphasizing accessibility, efficiency, and user empowerment.

1 Introduction

The decision-making process is intricately woven, often hinging on the critical information possessed by the decision-maker. In the contemporary world, copious amounts of data are amassed across diverse domains. While these data holds immense potential for informed decision-making, their utility is hindered by disparate systems storing data in varied formats, offering differing levels of access and security. Additionally, the data may suffer from incompleteness, contradictions, and unreliability. The management of vast datasets often falls into the hands of data analysts, and this is where Business Intelligence (BI) becomes indispensable. Business intelligence tools play a pivotal role in extracting valuable insights and aiding in strategic decision-making. These tools sift through historical and current data, presenting results in visually intuitive formats. However, a notable challenge lies in the communication gap between technical specialists, decision-makers, and business process analysts. As underlined in (Cherednichenko et al., 2023), dridging this divide is where Collaborative Business Intelligence (CBI) comes into play. As demonstrated by our analysis, the necessity for data research extends beyond the realms of business profitability or government problem-solving ; it is equally relevant to society and individual citizens for informed decision-making. However, organizing the collaboration between potential users, decision-makers, and technical specialists in data analysis poses a considerable challenge. The BI4people project [1] has been initiated to address these challenges. Its objective is to democratize Business Intelligence by implementing a data warehousing process in software-as-a-service mode. This spans from integrating data from various sources to intuitive analysis and data visualization. The unique approach of BI4people encourages individuals to contribute their opinions and comments, fostering a collaborative atmosphere and giving rise to what can be termed as general BI or Collaborative BI. Expanding on our previous research, this paper delves into the potential of conversational chatbots in enhancing Collaborative Business Intelligence tools. Building on our earlier work, which introduced a reference model for a virtual collaborative assistant in CBI, comprising conversational, data exploration, and recommendation agents, we now focus on the role of conversational agents in aiding data exploration. This research explores features crucial for converting user queries into executed commands, especially for non-experienced users. Furthermore, the study addresses the interface challenges between conversational and data exploration components, opening avenues for improved collaboration and accessibility within CBI systems.

2 State of the art

The data exploration process systematically scrutinizes and analyzes datasets to elicit insights, identify patterns, and unveil meaningful information (Jebb et al., 2017). It is an iterative and interactive procedure, combining manual exploration with automated analysis techniques. Data exploration assumes a pivotal role in the overarching Business Intelligence (BI) framework, encompassing the examination, scrutiny, and visualization of data to extract significant insights (Çağatay Demiralp et al., 2017). Employing statistical tools and visualizations facilitates the identification of patterns, trends, and relationships within the data, contributing to the formulation of hypotheses and guiding further investigations (Francia et al., 2022). Python, with its comprehensive library ecosystem including NumPy, Seaborn, Matplotlib, and Pandas, provides robust tools for Exploratory Data Analysis (EDA). However, users must possess expertise in data analysis and proficiency in relevant tools and methodologies to effectively manipulate and explore the data.

The current landscape of collaborative business intelligence reflects an expanding field characterized by the continuous development of tools and techniques, all aimed at facilitating effective collaboration within decision-making processes for teams (Muhammad et al., 2022). Notable advancements within the Collaborative Business Intelligence (CBI) domain include the integration of social media functionalities, enhanced accessibility through mobile devices, and the adoption of cloud-based solutions. These innovations empower users to engage in collaborative work, access data from any location, using any device, and at their convenience. Moreover, the integration of Natural Language Processing (NLP) and Machine Learning (ML) simplifies user interactions with data, extracting valuable insights and enhancing decision-making efficiency (Wolf et al., 2019; Otter et al., 2020). An additional benefit lies in the utilization of Natural Language Querying (NLQ), streamlining non-technical users' access to and

1. https ://eric.univ-lyon2.fr/bi4people/index-en.html

analysis of data, thereby enhancing their involvement in decision-making processes (Caldarini et al., 2022; Gamboa-Cruzado et al., 2023).

The incorporation of robotic assistants in e-commerce platforms and web portals with diverse functionalities has become increasingly prevalent (Gamboa-Cruzado et al., 2023). This study seeks to explore the potential advantages offered by conversational chatbots in enhancing the data exploration process and introducing novel prospects within CBI. A chatbot, a computer program designed to simulate human conversation through text or voice-based interactions (citation), offers notable benefits, including round-the-clock availability, expedited response times, personalized interactions, scalability, and cost-effectiveness.

The primary objective of our research centers around developing a model for CBI processes within distributed virtual teams, facilitated through the interaction between users and a CBI Virtual Assistant. The virtual assistant aims to broaden the accessibility of data exploration for a diverse user base while concurrently reducing the time and effort required for data analysis. As a result, the following research questions are posed :

RQ1 : How can pretrained generative language models be effectively utilized for data query construction to support non-expert users ?

RQ2 : What are the essential components of a prompt created for a generative language model to construct data queries effectively ?

To answer the research questions it is needed firstly to investigate and delineate effective methodologies for utilizing pretrained generative language models in the construction of data queries, with a specific focus on enhancing accessibility for non-expert users. This involves assessing the capabilities of generative language models in formulating natural language queries and discerning strategies that cater to the needs of users lacking specialized expertise in data analysis. Then, to identify and define the critical elements integral to the formulation of prompts tailored for generative language models in the context of constructing data queries. This objective involves scrutinizing the components that contribute to the efficacy of prompts, ensuring their ability to guide generative language models in generating relevant and insightful data queries.

By addressing these objectives, the research aims to contribute insights into the practical application of generative language models in the realm of data exploration, particularly in the context of CBI. The overarching goal is to enhance the usage of such models for recommender systems, thereby streamlining and augmenting the data exploration process within non-experts.

3 Materials and methods

Our investigation commences with a focus on data exploration, specifically exploring the advantages of integrating a virtual assistant to aid users in this intricate process. The initial phases involve defining the domain of study, identifying relevant data sources, and consolidating the acquired data. The data loading process entails acquiring and importing pertinent data from diverse sources into the CBI system. Subsequently, the exploration phase is initiated, allowing users to analyze data, discern patterns, and comprehend underlying trends. Metadata, encompassing descriptive information about the data, is pivotal. This includes structural attributes such as format, semantics, and relationships.

User interaction involves formulating queries to extract specific information or perform calculations. Constructing these queries necessitates adept syntactical formulation and the selection of relevant parameters aligned with the desired analytical or informational requisites. Post formulation, queries must be aligned with corresponding commands within the collaborative business intelligence system to ensure accurate interpretation of user intent.

The generation of prompts, integral to user interaction, necessitates an understanding of the user's context, preferences, and analytical objectives. The research question centers on the efficacy of a generative language model in facilitating data exploration for non-expert users. The experimental pursuit seeks to delineate a linguistic framework for prompt construction and the development of a user query reference model.

The selected dataset pertains to traffic accidents in France [2], specifically extracted from the National Interministerial Road Safety Observatory "ONISR." This dataset encompasses information about bodily accidents on public roads, detailing accident characteristics, location, involved vehicles, and victims.

To construct prompts, the GPT 3.5 model [3] in the English language is employed. The focal business process is Exploratory Data Analysis (EDA), comprising common steps applicable across diverse datasets. These EDA steps include Summary Statistics, Univariate Analysis, Bivariate Analysis, and Multivariate Analysis.

Summary Statistics offer a succinct overview of essential dataset features, encompassing central tendency, variability, and distribution. Measures such as Mean, Median, Mode, Variance, Standard Deviation, Range, Skewness, and Kurtosis are employed.

Univariate Analysis scrutinizes individual variables, while Bivariate Analysis explores relationships between two variables. Multivariate Analysis extends this examination to three or more variables, unveiling intricate relationships and patterns. For our experimental endeavors, a deliberate emphasis is placed on univariate and bivariate analyses, predicated on the presumption that the non-expert user is the focal investigator. The intent is to streamline the analysis, making it user-friendly and conducive to the informed inquiry and interpretation by individuals with varying degrees of analytical acumen.

The experiment involves manipulating collocations from EDA steps, generating prompts, user queries, and employing them to interact with the dataset. The generated grades are analyzed to build a prompt framework facilitating non-expert users in initiating EDA.

4 Results

In our experimentation, we focused on utilizing the GPT-3.5 model to generate data exploration prompts for a dataset concerning accidents in France from 2006 to 2015. The dataset contains statistics on accidents categorized by date. Our objective was to understand how effectively we could prompt the model to provide relevant questions for data analysis tasks, given that the user is a non-expert. Our interactions with the model were framed as a task for a data analysis assistant providing questions to a non-expert user about the dataset. The generated prompts included a description of the dataset, the feature to investigate ("trajet"), and a command (e.g., value counts()) to use for analysis. We conducted 114 rounds of experiments, varying feature names, commands, and dataset information, while occasionally introducing specific instructions like "*you cannot name the command or feature*."

2. https ://www.data.gouv.fr
3. https ://platform.openai.com/docs/model-index-for-researchers

We hypothesized that including comprehensive explanations about the dataset, specifying the feature of interest, and providing a command in the prompt would enhance the relevance of the generated questions. The task for generating questions was explicitly defined to guide the model. For example, a typical prompt was : "*You are a data analysis assistant. The user is not an expert. Suggest a question to the data...*"

Our results revealed that, initially, only 30 per cent of the generated answers were deemed relevant to the task. However, upon introducing additional instructions and refining the communication after non-relevant answers, we achieved a notable improvement in the model's response accuracy. Example of a Relevant Answer : "*Question for the data : "Which reasons for travel are most frequently associated with accidents in France from 2006 to 2015 ?" This information may be valuable for understanding the correlation between different reasons for travel and accident frequencies, potentially guiding safety measures or interventions specific to certain types of trips*."

Example of a Non-Relevant Answer : "*Could you please find out the distribution of reasons for travel (feature : "trajet") in the dataset ? I'm interested in understanding how frequently different reasons for travel occur in the reported accidents. By using value counts() on the "trajet" feature, the user can explore the count of each unique value (reason for travel) in the dataset. The result will provide a breakdown of the frequency of each reason for travel, allowing the user to gain insights into the patterns of accidents based on different travel purposes*." This question explicitly names the command and feature, deviating from the experimental protocol. It does not adhere to the constraints specified in the prompt.

Our findings underscore the significance of detailed dataset descriptions and feature names in generating relevant prompts for data exploration with non-expert users. Contrary to our initial hypothesis, the specific command proved less critical. Additionally, it became evident that thoroughly describing the experimental scenario, highlighting limitations, and refining prompts in response to initial non-relevant answers significantly improved the efficacy of the model in generating pertinent questions. We conclude that establishing a well-defined Exploratory Data Analysis (EDA) pipeline in advance and progressively building the dialogue with the user are crucial for successful interaction with generative language models in data exploration scenarios.

5 Conclusion and Discussion

Our investigation into RQ1 centered on the effective utilization of pretrained generative language models to construct data queries tailored for non-expert users. The results of our experiments demonstrated that pretrained models, such as GPT-3.5, can be harnessed to facilitate data query construction for individuals lacking expertise in data analysis. The key factor influencing the success of this utilization lies in providing a dataset description, emphasizing the significance of the feature under investigation, and guiding the model with a clear task formulation. By framing the interaction within the context of a data analysis assistant catering to a non-expert user, we achieved meaningful and relevant prompts that contributed to a more user-friendly data exploration experience.

In addressing RQ2, we delved into the critical components necessary for constructing prompts that effectively guide generative language models in data query creation. Our experiments revealed that the most pivotal elements include a detailed description of the query and explicit identification of the feature to be analyzed. Surprisingly, the specific command played a less decisive role in comparison to the dataset information and feature name. Additionally, we found that incorporating constraints, such as prohibiting the explicit naming of the command or feature, further enhanced the relevance of the model-generated questions.

In conclusion, our research provides valuable insights into the effective use of pretrained generative language models for supporting non-expert users in data exploration. By understanding the nuanced requirements of constructing prompts, researchers and practitioners can harness the potential of such models to enhance the accessibility and usability of data analysis tools for a broader audience.

6 Acknowledgements

The research study depicted in this paper is funded by the French National Research Agency (ANR), project ANR-19-CE23-0005 BI4people (Business intelligence for the people).

Références

Caldarini, G., S. Jaf, et K. McGarry (2022). A literature survey of recent advances in chatbots. *Information 13*(1), 41.

Cherednichenko, O., F. Muhammad, J. Darmont, et C. Favre (2023). Reference model for collaborative business intelligence virtual assistant. In *CEUR Workshop Proceedings*, Volume Vol-3403, pp. 114–125.

Francia, M., E. Gallinucci, et M. Golfarelli (2022). Cool : A framework for conversational olap. *Inf. Syst. 104*(C).

Gamboa-Cruzado, J. et al. (2023). Use of chatbots in e-commerce : a comprehensive systematic review. *Journal of Theoretical and Applied Information Technology 101*(4).

Jebb, A. T., S. Parrigon, et S. E. Woo (2017). Exploratory data analysis as a foundation of inductive research. *Human Resource Management Review 27*(2), 265–276.

Muhammad, F., J. Darmont, et C. Favre (2022). The collaborative business intelligence ontology (cbiont). In *18e journées Business Intelligence et Big Data (EDA-22)*, Volume B-18, Clermont-Ferrand.

Otter, D. W., J. R. Medina, et J. K. Kalita (2020). A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 604–624.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et J. Brew (2019). Transformers : State-of-the-art natural language processing. In *Conference on Empirical Methods in Natural Language Processing*, pp. 38–45.

Çağatay Demiralp, P. J. Haas, S. Parthasarathy, et T. Pedapati (2017). Foresight : Rapid data exploration through guideposts. *ArXiv abs/1709.10513*.

Summary

In the dynamic landscape of contemporary business, the demand for cutting-edge methodologies integrating intelligent technologies to facilitate effective decision-making is on the rise. Business Intelligence systems have become indispensable tools for decision-making across diverse organizations, and the emergence of Collaborative Business Intelligence (CBI) extends the horizons of informed decision-making by tapping into external information from varied sources. The central aim of this research centers on the development of a model tailored for CBI processes within distributed virtual teams. This is achieved through the interaction between users and a CBI Virtual Assistant. The key role of the virtual assistant is to augment the accessibility of data exploration for a wider user base, concurrently minimizing the time and effort required for data analysis. The proposed model, through its integration of intelligent technologies, seeks to propel collaborative decision-making to new heights, emphasizing accessibility, efficiency, and user empowerment.