Towards AI-Based Data Analytics for Environmental Monitoring

Angelica Lo Duca[1,*,†], Rosa Lo Duca[2,†]

[1]Institute of Informatics and Telematics of the National Research Council, via G. Moruzzi 1, 56124 Pisa, Italy
[2]ARPA Lazio, Representative Office, 00187, Roma, Italy

Abstract

Environmental monitoring is essential to intervene promptly in environmental disasters or emergencies. Therefore, in the environmental field, as in other fields, it is essential to create reports quickly. In this article, we describe a possible system that uses Artificial Intelligence (AI) to automatically create environmental reports, always considering the supervision of the environmental expert. The system involves two phases, one for testing and the other for production. During the testing phase, the AI system is calibrated on a subset of the available data, while during the production phase, the AI system is fully operational and works on all the available data. We describe a practical case study using the temperature data provided by ARPA Lazio.

Keywords

Environmental Monitoring, Artificial Intelligence, Data Analytics

1. Introduction

In recent years, Artificial Intelligence (AI) has made substantial progress thanks to the diffusion of Large Language Models (LLMs) [1]. These models offer new opportunities for professionals, speed up the execution times of specific tasks, and provide increasingly advanced models for solving complex tasks [2, 3, 4]. Applying its potential to environmental data analysis means opening a window onto a scenario of considerable global interest. Institutions, associations, companies in the sector, and citizens need instructions on how to read environmental data. Institutions are called upon to make targeted decisions to safeguard the environment. Associations and companies in the sector have the task of calibrating their products and services based on specific criteria that guarantee sustainable development. With their daily actions and choices, simple citizens can make a difference in safeguarding the entire ecosystem. Our work fits into this perspective: the objective is to put AI at the service of the aforementioned categories to help them understand environmental data and make responsible and concrete choices in the environmental context. In this paper, we propose using LLMs to build environmental reports automatically. The proposed system relies on a preliminary phase, called the testing phase, which calibrates the AI model utilizing the support of an environmental expert. In the second production phase, the AI system automatically answers the environmental expert's questions about the data. We also describe a preliminary case study applied to the historical series of temperatures provided by ARPA Lazio [5]. Cao et al. have proposed a similar framework for analyzing environmental data [6]. Compared to their work, our system is based on a calibration of the AI model based on human contribution.

2. Our Approach

We propose a system combining manual supervision with automatic analytics. There are two actors: the environmental expert, a human with deep knowledge of environmental data, and the AI model,

Discovery Science - Late Breaking Contributions 2024
*Corresponding author.
†These authors contributed equally.
angelica.loduca@iit.cnr.it (A. Lo Duca); rosa.loduca@arpalazio.it (R. Lo Duca)
0000-0002-5252-6966 (A. Lo Duca)

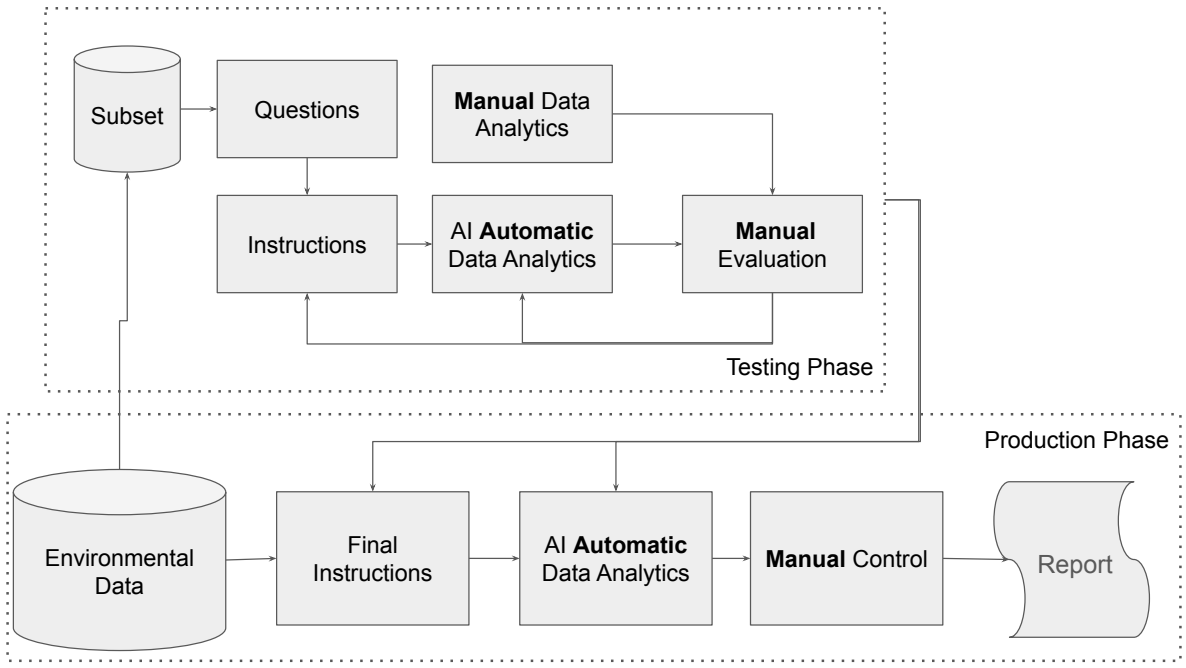

Figure 1: The workflow of the proposed system.

based on LLMs, which is tuned with environmental data through a Retrieval Augmented Generation (RAG) process [7]. At the moment, the AI model ingests only the environmental data, but in the future, it could also inject other environmental documents and reports, helping it better understand the environmental context in which it must operate. Our system comprises two phases, as shown in Figure 1: the testing and production phases. The AI system is calibrated during testing to define the input instructions correctly. This involves subsequent steps. The workflow starts from the upper right of the figure and follows the flow defined by the arrows. Given the database containing all the environmental data, the environmental expert carefully extracts a data subset representative of all the data. Subsequently, the environmental expert defines a series of questions to ask the data. These questions are translated into instructions for the AI model. At the same time, the AI model and the environmental expert proceed with a separate analysis of the data sample. Both respond to the questions posed by the environmental expert manually, through standard data analysis techniques and the AI model through automatic analysis. At the end of the analysis, the environmental expert compares the results obtained by AI with their own results. If the environmental expert detects problems in the answers provided by the AI to the questions, they proceed to reformulate the instructions for the AI model. This process continues until the AI correctly answers the questions asked. Once the evaluation phase is complete, the model is calibrated so we can move on to the production phase, where only the AI model is used for analysis of the entire database of data. At the end of the analysis process, the AI model generates a report with all the answers to the questions. The environmental expert evaluates the final results. We implemented the AI model as a RAG-based system using LangChain [8], ChromaDB [9], and OpenAI GPT-4 [10].

3. Case Study

To test the effectiveness of our proposed approach, we analyzed the historical series of temperature data recorded in the decade 2013-2023 from the AL007 station of the Micro Meteorological Network (RMM) managed by ARPA Lazio and located in via Boncompagni 101 in Rome [5]. The RMM is made up of 9 stations located throughout the Lazio region, each made up of classic meteorological sensors (temperature, humidity, pressure, and precipitation) associated with instrumentation dedicated to the

dispersion of pollutants (sonic anemometers, pyranometers, and pyrgeometers). These tools are digital. In particular, an HMP 45 AC type thermo-hygrometer measures temperature and can detect the air's relative humidity. The temperature values are taken every 30 minutes from each other. Starting from this data, we have focused only on temperature.

Through traditional data pre-processing techniques, the environmental expert (one of the authors) extracted average temperature values, maximum and minimum values, and mode for each month and considered each year. Next, they formulated the following questions, with an increasing level of complexity: 1) In which month and year was the highest temperature value recorded? 2) What was the hottest month for each year? 3) What is the percentage increase in temperature over the decade 2013-2023? We transformed the previous questions into instructions for the AI model. We defined the following basic structure for all the questions:

```
Consider {context} where:
    - Year is the year
    - Month is the month
    - Mean is the average monthly temperature
    - Max is the maximum monthly temperature
    - Min is the minimum monthly temperature
    - Mode is the most frequent monthly temperature value
Answer the following question: {question}
```

The context variable contains the input data, and the question variable is the specific question to be answered. We implemented the AI model as a RAG-based system using LangChain [8], ChromaDB [9], and OpenAI GPT-4 [10].

In the first phase of the test, the AI model read and extracted data from only four years (2013, 2014, 2015, and 2020) compared to the entire available sample (2013-2023). Considering the entire data sample, the AI model answered the first two questions incorrectly. However, if we narrow it down to four years, the answers are correct. Regarding the last question, the AI model did not respond by providing a numerical value, but it offered an interesting result: it precisely described the useful steps to answer the question, adding that it was not able to apply this procedure as it lacked a significant data, that of 2023, which does not fall within the data of the years known to it.

As a second step, we modified the structure of the questions to force the model to read all the years. We defined the following structure for all the questions:

```
I have many datasets from 2013 to 2023.
Consider {context} where:
    - Year is the year
    - Month is the month
    - Mean is the average monthly temperature
    - Max is the maximum monthly temperature
    - Min is the minimum monthly temperature
    - Mode is the most frequent monthly temperature value
Answer the following question considering all the data from 2013 to 2023: {question}
```

Compared to the previous version of the questions, we added more details related to the dataset range (highlighted in the previous code). The AI model answered the third question correctly, while it could not answer completely the first and second questions because it did not find data related to 2021, 2022, and 2023. The refining process throughout the testing phase should continue until the AI model can answer all the questions correctly. Our preliminary experiments demonstrated that although the AI model could not answer all the questions correctly, we can conclude that preliminary results are encouraging compared to the questions we answered manually. As a future work, we plan to improve the testing phase and also implement the production phase of the project.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in Neural Information Processing Systems 33 (2020) 1877–1901.

[2] M. F. Sani, M. Sroká, A. Burattin, Llms and process mining: Challenges in rpa: Task grouping, labelling and connector recommendation, in: International Conference on Process Mining, Springer, 2023, pp. 379–391.

[3] S. A. Gebreab, K. Salah, R. Jayaraman, M. H. ur Rehman, S. Ellaham, Llm-based framework for administrative task automation in healthcare, in: 2024 12th International Symposium on Digital Forensics and Security (ISDFS), IEEE, 2024, pp. 1–7.

[4] G. Olaoye, H. Jonathan, The evolving role of large language models (llms) in banking, EasyChair Preprint no. 13367, EasyChair, 2024.

[5] A. Lazio, Dati rete micro-meteorologica, 2013-2023, Rete micro-meteorologica - ARPA Lazio, 2023. Https://www.arpalazio.it/rete-micro-meteorologica (Last Access 2024/07/15).

[6] C. Cao, J. Zhuang, Q. He, Llm-assisted modeling and simulations for public sector decision-making: Bridging climate data and policy insights, in: AAAI-2024 Workshop on Public Sector LLMs: Algorithmic and Sociotechnical Design, 2024.

[7] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, in: H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33, Curran Associates, Inc., 2020, pp. 9459–74.

[8] LangChain, https://langchain.com/, 2024. (Last Access 2024/07/15).

[9] Chroma core/chroma, The AI-native open-source embedding database, https://github.com/chroma-core/chroma, 2024. (Last Access 2024/07/15).

[10] OpenAI, GPT-4 Technical Report, arXiv.org, 2023. Https://doi.org/10.48550/arXiv.2303.08774.