United Nations ECE/CES/2024/21

Economic and Social Council

Distr.: General
30 May 2024

English only

Economic Commission for Europe

Conference of European Statisticians

Seventy-second plenary session
Geneva, 20 and 21 June 2024
Item 5 of the provisional agenda
Use of Artificial Intelligence and Large Language Models in official statistics and authoritative geospatial data

Adopting artificial intelligence in the production and dissemination of official statistics

Prepared by Norway

Summary

Will the use of artificial intelligence (AI) affect the production and availability of facts in the future? The answer is obviously yes. The development and use of AI takes place at a rapid pace and on an increasing scale in all parts of society, and in Statistics Norway we expect AI to play an increasingly important role in the production and dissemination of statistics. It is too early to foresee all the consequences and how we will be affected by AI. But, as in other organizations, AI has the potential of significantly improving products impacts, operational efficiency and workflow optimization in Statistics Norway.

At Statistics Norway, our initiative to integrate AI particularly through the development of internal chatbots represents a strategic approach to leveraging technology for enhancing internal processes. The decision to prioritize chatbot development is driven by several considerations, each aimed at bolstering operational efficiency, safeguarding data, and supporting our staff more effectively.

The document is presented to the Conference of European Statisticians' session on "Use of artificial intelligence and large language models in official statistics and authoritative geospatial data" for discussion.

I. How artificial intelligence will affect our organization

1. For almost 150 years, Statistics Norway (SSB) has described the Norwegian society in numbers. Every year, nearly 300 statistics are published that provide insight into different areas, such as population trends, price trends, house prices, the labour market, climate emissions, business and industry and name trends. We have identified several areas in Statistics Norway that we believe will be strongly influenced by artificial intelligence (AI) technology.

A. Production of statistics

2. Data is the most important raw material in the production of statistics and is also an important product for research and analysis. For some statistics, machine learning has been used for several years, for example for the purpose of classification and coding. Going forward, we believe that AI can help automate or simplify traditional methods for collecting, editing, and analysing data. These processes can be both time-consuming and resource intensive. AI can replace manual processes and reduce the risk of errors and inaccuracies.

3. Machine learning algorithms can also help automate data collection from various sources. Because algorithms can identify patterns in large data sets, it is possible to improve the quality of the final products and gain increased timeliness because the statistics can be produced faster.

B. Dissemination

4. With the increasing use of AI in all segments of society, the population's expectations of government agencies will also change. Users will expect personalized services, increased automation, better accessibility, and support. This development will require adaptation and investment in new technology and expertise in all public agencies.

5. Official statistics can be made more easily accessible, and dissemination can be improved through interactive interfaces based on AI. Instead of users having to spend time searching for statistics and analysis on our website ssb.no, a chatbot can answer questions, provide guidance on data sources and present statistical results in an adapted and even more engaging way.

6. This will make it easier for businesses, organizations, researchers, and the public to benefit from the statistics, thereby increasing the use and value of official statistics.

C. Data sharing and training of algorithms

7. AI requires access to large and reliable datasets that can be used to train and evaluate the models on which it is based. A fair AI model produces results that are not discriminatory or biased. An AI reflects the patterns in the data it is trained on and can therefore absorb biases or prejudices in these. To achieve fair algorithms, the training data must be representative and not contain biases. The national statistical offices have such data. Registry data and data collected through surveys over many years cover a wide range of economic, demographic, social and environmental issues that are critical to developing equitable AI algorithms that can understand and solve complex problems.

D. Statistics on the use of artificial intelligence

8. The National Programme for Official Statistics (2024–2027) [1], provides the framework for the Norwegian statistical system for the next few years. It specifies which statistical areas are covered, which public authorities are responsible for the statistics, and development areas for improving the statistics. The programme does not directly describe statistics on the use or development of AI in the private sector, the public sector or society, but includes statistics describing innovation in the private sector, and statistics on information

and communication technology (ICT) use in private households, the private sector, and the public sector.

9. Initially, these statistics may be able to capture some of the development in the use of AI. The rapid development in the use of AI and the diversity of methods and applications may nevertheless make it challenging to adequately describe the use and significance AI has gained so far in various areas of society.

10. A prerequisite for producing official statistics on AI, is the development of common definitions and classifications for AI and associated technologies. The rapid development also makes this challenging.

E. Research

11. Statistics Norway's research activities include further developing, strengthening, updating, and documenting several different economic models. Among other things, the models are used to prepare forecasts and analyse the consequences of policy measures.

12. AI can contribute to the continuous development of these models, and hopefully also simplify the work processes. For example, machine learning can be used to improve predictions and forecasts, identify complex relationships, automate model fitting and updating, and optimize model parameters. This can strengthen the accuracy, robustness, and relevance of the models, and provide a basis for better decisions and policy formulation.

II. An assistant for official statistics

13. Using commercially available tools, such as ChatGPT, requires little of local information technology (IT) systems. But to fully exploit the possibilities of AI, it is necessary to be able to store and process large amounts of data quickly. There must be access to computational power that makes it possible to train, use and continuously update advanced AI models.

14. Statistics Norway is in the process of establishing our statistics production on a new and cloud-based data platform. The restructuring work is extensive, as production runs for almost 300 statistics will be rebuilt. The platform is a prerequisite for Statistics Norway to be able to use AI in the production of official statistics.

15. As discussed above, AI can be used directly to solve some of the work processes in the production of statistics. But AI can also be an effective work tool, or "assistant", for employees in the work of creating and communicating the statistics. For example, AI can provide good and effective support in coding efforts, even for employees with very high coding skills.

16. While the potential for AI as an assistant is great, there are several risks associated with its use. This applies in particular to the use of commercial AI assistants. Examples include problems with verifying whether the output is correct, copyright uncertainty, and how information fed into commercial AI services can be shared with other users of the services through the algorithms. Such implicit and unintended sharing may have consequences that we now do not fully understand.

17. Many organizations, including Statistics Norway, are therefore looking into the possibilities of developing their own AI assistants. The purpose is for employees and users of ssb.no to have access to the technology safely, without the risk of information going astray.

A. Using open-source tools

18. Our development process for these chatbots is grounded in the use of Python, Langchain, open-source Large Language Models (LLMs) from Hugging Face, Chainlit user interfaces, and Docker containers. Python serves as the primary programming language due to its versatility and the extensive support it receives from a broad developer community. LangChain and LLMs from Hugging Face are critical for enabling natural language

processing capabilities, making our chatbots capable of understanding and generating responses that closely mimic human interaction. Chainlit provides the graphical user interface and deployment is facilitated through Docker containers, ensuring that our chatbots can operate reliably on our data platform.

19. To address the urgent need for user-centric data retrieval and handling mechanisms, we have developed several chatbot prototypes using a blend of advanced capabilities from open-source LLMs. Our system's architecture integrates Python and the LangChain library, providing a seamless interaction with the LLMs and ensuring robust performance.

20. We have created prototypes for several chatbots based on the Retrieval-Augmented Generation (RAG) architecture. This innovation allows for enhanced response accuracy by integrating a knowledge base directly into the generation process. The knowledge base is a vector database containing "vectorized" internal documentation. This allows the chatbots to answer questions about the knowledge base and summarize information in it.

21. We have also developed a specialized chatbot that connects to the SSB external Application Programming Interface (API) so that users can chat directly with SSB's published data. This chatbot framework employs a range of specialized tools and functionalities that interface with the SSB's API, enabling users to efficiently navigate through various data operations. Key capabilities include searching for relevant data tables based on user inputs, retrieving metadata to understand table structures, and executing structured queries to obtain the necessary data.

22. Furthermore, we are testing PandasAI technology to create a "methodology bot". This new chatbot aims to assist researchers and data analysts by offering methodological guidance, automating data manipulations, and providing simpler on-demand methodological support to our statistics producers.

23. A pivotal feature of our chatbots is their user-oriented interfaces, powered by the ChainLit library. This component provides an intuitive and interactive platform that allows users to articulate their queries and receive prompt and accurate responses. The system's asynchronous configuration not only enhances performance but also scales effectively to handle concurrent user requests without compromising on speed.

B. Trade-offs

24. The trade-off between using the proprietary LLMs and smaller, open-source models highlights the nuanced decisions that organizations must make in deploying AI technologies.

25. Proprietary models, developed by large tech companies, offer distinct advantages in terms of performance and reliability. They are trained on vast, diverse, often proprietary datasets, allowing them to understand and generate human language with high accuracy. Additionally, these models often come with professional support and maintenance, which can alleviate the operational burden on organizations, particularly those lacking in-house AI expertise.

26. However, proprietary models have their drawbacks. They can be costly due to licensing fees, and their "black box" nature restricts customization. Organizations are also tied to the vendor's strategic decisions, which may not always align with their needs.

27. On the other hand, using open-source models offers significant flexibility. These models can be tailored to meet unique organizational requirements and optimized for specific tasks, allowing for modifications and enhancements by in-house teams or through community collaboration. Open-source models also reduce financial barriers, making advanced AI technologies accessible to a wide array of organizations. This democratization of technology encourages innovation and experimentation within the field, enabling projects with limited budgets to incorporate cutting-edge AI capabilities without incurring prohibitive costs.

28. However, also the use of open-source models may present certain challenges. One drawback is the potential lack of support and maintenance. While community collaboration can foster improvements, it also implies that the responsibility for development and upkeep

falls largely on the users or the community. Additional concerns can be the suitability and user friendliness of the interfaces, as well as the potential security issues posed by the open nature of the models.

C. Advantages of in-house chatbot development

29. The in-house development of chatbots affirms our commitment to the highest data security standards. Given the sensitive nature of our data, it is critical that our technological solutions are designed with inherent security measures to counter potential threats. Managing the development process internally allows for the implementation of strict security protocols, ensuring our chatbots meets rigorous data protection criteria.

30. Opting for internal chatbot development over the acquisition of proprietary software solutions offers significant cost advantages. Utilizing open-source technologies and leveraging our internal expertise minimizes reliance on external providers, eradicates licensing fees, and allows for specific customization without added expenditure. This approach represents a cost-effective strategy for developing tailored technological solutions.

31. Internal development facilitates the ongoing refinement and adaptation of chatbot functionalities to meet changing organizational needs and incorporate user feedback. This flexibility ensures that chatbot services remain pertinent, effective, and aligned with our strategic goals, highlighting the benefits of an adaptable development strategy over the static off-the-shelf solutions.

III. Putting artificial intelligence to work

32. Currently, our AI chatbots are in the initial phase of development, presenting a "minimal viable product" that demonstrates substantial potential for future enhancements. We plan to conduct an in-depth analysis of various LLMs, both open-source and proprietary, to explore their capabilities, biases, and overall suitability. Our aim is to refine the chatbot's functionalities and overcome any existing limitations. Areas where we plan on exploring our AI chatbots are briefly described below.

A. Internal documentation access

33. The implementation of chatbots provides Statistics Norway with instant access to internal documentation. This accessibility is facilitated through intuitive conversational interfaces, minimizing the need to manually sift through documents. The ability to quickly query and retrieve information not only speeds up the research process but also ensures that our staff can access and leverage the vast repositories of internal knowledge more effectively.

B. Manual search efficiency

34. Chatbots excel in navigating the complexities of extensive documentation, enabling users to conduct targeted searches with precision. This capability is particularly valuable in environments where time is of the essence and accuracy is paramount. By understanding natural language queries, chatbots streamline the process of locating specific data points, sections, or methodologies within comprehensive manuals, significantly enhancing productivity.

C. Methodological assistance

35. Customized to provide guidance on the unique statistical methods and analytical practices of Statistics Norway, chatbots can acts as on-demand advisors. This personalized support ensures that staff can access basic methodological assistance tailored to their specific operational context at any time. Chatbot outputs can be coordinated with the contents of our GitHub library of preferred methodological packages and functions.

D. “No unnecessarily boring task”

36. Beyond direct operational benefits, AI also positively impacts the working environment by automating routine tasks and facilitating easier access to information. Reducing the cognitive burden on employees allows for greater focus on complex and intellectually rewarding activities, enhancing job satisfaction, and fostering an innovative and learning-centric organizational culture.

IV. Quality assurance in the “age of artificial intelligence”

37. The system for quality assurance of Norwegian official statistics is based on quality requirements in the Norwegian Statistics Act [2] and in the guidelines given in the European Statistics Code of Practice [3]. The 16 principles in the guidelines in the were developed in 2005 and revised in 2011 and 2017. Technological developments since the last revision of the guidelines have been rapid and several of the principles are being challenged by the new technologies [4].

38. Ensuring transparency and sufficient documentation when using new methods, such as machine learning or other forms of artificial intelligence, is challenging because the algorithms are complex and because the software does not always have code that is openly available. The development of frameworks for describing uncertainty is crucial.

39. Fair algorithms are crucial for AI to foster positive outcomes and prevent the reinforcement of societal biases. The upcoming European Union AI regulation seeks to address these concerns by mandating responsible development and usage of AI systems, focusing on data quality, transparency, human oversight, accuracy, and robustness.

V. References

[1] National programme for official statistics: https://www.ssb.no/en/omssb/nasjonalt-program-for-offisiell-statistikk

[2] The Statistics Act: https://www.ssb.no/en/omssb/ssbs-virksomhet/styringsdokumenter/statistikkloven

[3] European Statistics Code of Practice: https://ec.europa.eu/eurostat/web/quality/european-quality-standards/european-statistics-code-of-practice

[4] Puts, M. and Daas, P. (2021) ML from the Perspective of Official Statistics. The Survey Statistician 84, 12–17 https://www.destatis.de/EN/About-Us/Events/Machine-Learning/Slides/p2_puts.pdf?__blob=publicationFile

In working on the text, the authors used ChatGTP as a discussion partner.