United Nations

ECE/CES/2024/22

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

Artificial intelligence and generative artificial intelligence experiences at Turkish Statistical Institute

Prepared by Türkiye

Summary

This document describes the experience of the Turkish Statistical Institute (TurkStat) in leveraging artificial intelligence to transform data collection, processing, and analysis for official statistics. By adopting new techniques, TurkStat has enhanced the accuracy, efficiency, and depth of its statistical insights, improving data quality across various applications and modernizing key areas such as consumer price index.

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. Introduction

1. At Turkish Statistical Institute (TurkStat), unlocking the power of technological innovation and data science for official statistics represents a paradigm shift in how we collect, process and interpret data. By harnessing the latest advancements in machine learning, Artificial Intelligence (AI) and data processing techniques, TurkStat has revolutionized the accuracy, speed and depth of its statistical insights.

2. TurkStat specialized in leveraging AI and advanced data analysis techniques to extract meaningful insights from large and complex datasets. Currently, TurkStat actively utilizes web scraping and scanner data for consumer price statistics and job openings. Prices of white goods, cars, furniture, bus and flight tickets, clothing and job postings from career websites are collected via web scraping. The collected prices are used in the calculation of consumer price indexes, while job advertisements contribute to labour force statistics. Classification of Individual Consumption According to Purpose (COICOP) and International Standard Classification of Occupations (ISCO-08) codes are automatically predicted and assigned to the web scraped data using machine learning algorithms.

3. TurkStat also operates its own big data environment for extracting meaning from vast datasets. In its big data environment, TurkStat has developed a system for collecting online prices through web scraping in this big data environment. Internet prices are scraped from carefully selected businesses, considering certain criteria such as representativeness, reproducibility, volume, content source of the website, etc., and processed for use in Consumer Price Index (CPI) calculations. This innovative approach, along with barcode scanner data, is reducing costs and improving the data used for CPI calculations.

4. TurkStat is modernizing CPI calculations by incorporating scanner data collected at stores. This data, including sales information for various products, allows for more accurate, efficient, and timely price statistics without the need for physical data collection. This approach, alongside web scraping represents a significant step forward for TurkStat.

5. In addition, TurkStat utilizes speech-to-text technology, which is an offline model, to handle different languages and achieve the most accurate results possible. Speech-to-text technology, like a voice recorder that writes things down, can convert spoken words into text. This technology is useful for tasks like transcribing meetings or controlling devices by voice.

6. Furthermore, TurkStat is working on predicting the statistical Classification of Products by Activity (CPA) codes for e-invoices data. This project aims to be a breakthrough in monitoring the course of a product from its production until it reaches the final consumer and in tracking the added value created. TurkStat possesses a massive amount of daily e-invoice data (7.5 million) that requires product category codes to be assigned. A project is underway to build a model using machine learning to automatically assign these codes based on product names and descriptions. While the models have achieved some success on pre-labelled data, their accuracy drops significantly when applied to real-world e-invoice data. This highlights the challenge of training models with clean and relevant data.

7. TurkStat has also explored the use of Large Language Models (LLMs) to assign product category codes to e-invoices. While some LLMs have shown promise, such as ChatGPT with manual prompting, their overall accuracy has not been high enough. Additionally, security risks and resource limitations make using online LLMs impractical for TurkStat. The project concluded that further research is needed to effectively utilize LLMs for text classification tasks.

8. TurkStat is in the process of building a classification prediction portal to assist users in assigning classification codes to various types of data, including occupations, economic activity, and individual consumption. This portal aims to improve internal data processing and assist external users with classification tasks. By providing recommendations and reducing errors, the portal will ultimately enhance data quality and efficiency.

II. Using artificial intelligence to improve labour market analysis

9. The accurate classification of job vacancies is crucial for informed policy decisions and a thorough understanding of labour market trends. An innovative study by the Turkish Statistical Institute (TurkStat) tackles this challenge by integrating cutting-edge techniques like administrative data collection, web scraping and machine learning.

10. The study begins by carefully compiling a comprehensive and high-quality dataset by acquiring job vacancy data from official sources as well as the internet. To ensure consistency and reliability, data from these diverse sources are then meticulously harmonized. Through a series of data preprocessing steps, the information is structured for rigorous analysis.

11. Utilizing powerful neural networks, the study predicts ISCO-08 for each job vacancy. This model demonstrates promising accuracy in assigning appropriate occupational classifications to a wide range of job descriptions, highlighting the effectiveness of these innovative methodologies in handling complex datasets.

12. Furthermore, TurkStat makes use of Business Intelligence (BI) tools to create an intuitive and user-friendly dashboard that encapsulates these predictions. This dashboard provides a detailed breakdown of job vacancies categorized by both occupation and region. By visually representing this data, stakeholders gain valuable insights into the regional distribution of jobs across various occupational categories, leading to a more nuanced understanding of labour market dynamics.

13. This convergence of administrative data integration, advanced machine learning for classification and subsequent data visualization through BI tools results in a powerful resource for policymakers, economists and researchers. This holistic approach not only streamlines the classification process but also empowers data-driven decision-making in labour market analysis.

14. In conclusion, this study by TurkStat exemplifies the effectiveness of innovative approaches in enriching official statistics. It provides a robust framework for analysing job vacancies, predicting occupational classifications and presenting regional breakdowns through user-friendly dashboards. This accomplishment is built upon a strong foundation of high-quality data, paving the way for a deeper understanding of the labour market.

III. Web scraping and scanner data for automating price collection

15. TurkStat is dedicated to leveraging cutting-edge technologies to enhance its data collection and analysis processes. In recent years, TurkStat has actively implemented web scraping and scanner data collection methods to improve the accuracy and efficiency of consumer price statistics.

A. Web scraping

16. In 2020, TurkStat collaborated with the Scientific and Technological Research Council of Türkiye on a project aimed at establishing an infrastructure for web scraping to collect online prices. Thus, a big data environment was built at TurkStat to conduct big data and advanced analytical studies.

17. The businesses to collect online prices included in the web scraping project were selected based on the following criteria:

- Representativeness
- Volume
- Content
- Reproducibility

- Technical features
- Methodological considerations
- Target variables
- Metadata.

18. TurkStat staff conducted extensive data analysis to transform the raw data into a format suitable for price collection. They assigned COICOP codes to products that matched the defined item descriptions and developed software codes to ensure consistency in the prices collected.

19. Web scraping was initially used to collect prices for white goods, electronics, clothing, furniture, new cars and bus tickets. In 2022, these prices were incorporated into the index calculations for the first time. As of 2024, web scraping accounts for approximately 5.1 per cent of the total number of prices used in CPI calculations.

20. The use of web scraping and scanner data in CPI calculations offers several benefits, including:

- Reduced costs
- Decreased respondent burden
- Increased data collection frequency and volume.

B. Scanner data

21. In addition to web scraping, TurkStat has begun implementing other alternative methods for compiling the CPI. One of the most promising methods is the use of scanner data.

22. Scanner data is a large data source that can be used to automate price collection processes. By utilizing barcode information, TurkStat aims to modernize price statistics, leverage technological advancements and enhance the quality of statistical production.

23. The scanner data project aims to incorporate sales information collected by optical scanners at retail points of sale into the CPI calculation. While the primary focus is on price information, the project also collects detailed data on sales locations.

24. The use of scanner data presents a significant opportunity to improve the CPI. The data includes information such as the sales value, quantity and price of products sold in all stores of a retail chain. This information has the potential to enhance the price, weight and representativeness elements used in the CPI.

25. After completing the methodological and technological infrastructure preparations, TurkStat has begun incorporating scanner data into the CPI calculation in 2021. The share of scanner data in the CPI calculation was 21 per cent in at that time and has increased to 42.6 per cent as of January 2024. This implementation has reduced costs in terms of field organization, as the data used for sales is obtained digitally without the need for field visits. Integrating scanner data into the existing CPI calculations has also reduced response burden and physical data collection costs, focused on products that have been sold and provided more comprehensive price information over a longer period.

26. The scanner data project represents a significant step forward in the modernization of price statistics in Türkiye. By using cutting-edge technologies and data sources, TurkStat is enhancing the accuracy, efficiency, and timeliness of CPI calculations, providing valuable insights for policymakers and researchers.

27. The implementation of web scraping and scanner data collection methods has significantly enhanced TurkStat’s ability to collect and analyse consumer price data. These methods have improved the accuracy, efficiency and timeliness of CPI calculations, providing valuable insights for policymakers and researchers.

IV. Speech-to-Text

28. Speech-to-Text technology, also known as speech-to-text conversion (from a video or audio file), is a system that converts voice utterances into written text. This technology detects speech or voice commands and translates them into a comprehensible text format, usually displayed on a screen or used for processing in other systems.

29. Speech-to-Text technology is widely used today to convert audio content from different sources such as voice recordings, meetings, speeches or presentations into written text. This technology allows users to communicate directly by speaking without typing or to convert voice data into text-based data. Speech-to-Text technology is known for its ability to accurately transcribe spoken expressions into text. In this way, it offers many practical uses, such as people directing their devices with voice commands, transferring speech content to text format through transcript tools.

30. At TurkStat, a general-purpose offline speech recognition model is used for speech-to-text. This model is trained on a large dataset of various voices and is also a multitasking model capable of multilingual speech recognition, speech translation and language identification, including Turkish. This allows us to transcribe meetings with high accuracy. The higher the quality of the audio, the higher the accuracy.

V. Predicting Classification of Products by Activity for products in e-invoices with artificial intelligence

31. There are large amounts of e-invoice data that TurkStat can access (approximately 7.5 million per day on average). Subject matter units need Classification of Products by Activity (CPA) codes for these data. The project aims to create a model that will predict the CPA codes using the name and description of the products in e-invoice data and automatically assign CPAs to this model in data transfer.

A. Prediction with machine learning

32. A total of six studies were conducted (five machine learning and one semantic model). CPA classification contains a large number of classes (3,218 classes) compared to classical supervised machine learning problems. The machine does the learning process from data. Therefore, it is very sensitive to omissions and errors in the dataset. There are inconsistencies, errors and problems with CPA code representation in the studied datasets.

33. The training datasets used are different from e-invoice data because there is no labelled e-invoice data. This causes the model to fail in e-invoice data, even if it is successful in the data set on which it is trained.

34. Among a total of five model attempts, Support Vector Machines (SVM) was the most successful. Additionally, another model with successful results on e-invoice data is the sentence transformer with cosine similarity. There was a significant increase in success when a threshold value was set when making predictions. However, in this case, the coverage value decreases.

35. Two datasets were used to measure the accuracy of the models. First, a human attempt was made to assign CPA to 500 completely randomly selected records. The results were as follows:

Situation	*Frequency*	*%*
A CPA is assigned	111	22.2%
Lower digits or more than one CPA can be assigned	58	11.6%
CPA not assignable	331	66.2%
Total number of samples	500	

36. The second dataset is a dataset of 715 records from e-invoices, again assigned CPA by humans.

37. By making assignments to these two datasets with the two mentioned models, the following results were obtained:

				Predicted correctly			
		Predictable		*#*		*%*	
Data set	*Number of records*	*SVM*	*Cosine*	*SVM*	*Cosine*	*SVM*	*Cosine*
Random	500	97	500	30	20	6.00%	4.00%
CPA assigned	715	663	714	162	177	22.70%	24.80%

38. As can be seen from the table, the accuracy rate of the models for randomly selected records is between 4 per cent and 6 per cent, while the correct prediction rate for the CPA assigned dataset reaches 25 per cent.

B. Prediction with Large Language Models

39. In order to assign CPA to products in e-invoices with AI, Harmonized System (HS) and Combined Nomenclature (CN) classifications, which are more common than CPA, were used. These were then converted to CPA using corresponding tables.

40. Online LLM models work very quickly. However, the infrastructure in which they work is unknown.

41. To use online LLM models in codes, an Application Programming Interface (API) key is needed and obtaining this API key requires creating a user. When creating this user, each service provider has its own contracts that must be accepted, some even require a phone number. While testing LLMs, first of all, the prompt was worked on so that the language model could answer in the desired format, and the prompts were used, which were tried and obtained the desired result.

42. **ChatGPT**: ChatGPT has a limited free API service and experiments have been conducted using this service. 715 records were predicted smoothly at once.

43. **Gemini**: Experiments were conducted with the API key received by the staff for Gemini. No response was received when a total of 715 records were given, so answers were obtained by sending 100 records each with a prompt in a loop. However, while it was expected to complete a total of eight cycles, an error was received after the sixth cycle. The last three cycles were run again and all results were obtained. As a comment here, we can think that when so many requests are sent to this free service, it is probably restricted somewhere.

44. **LLama 2**: The model via Hugging Face was run with Google Colab using Graphics Processing Unit (GPU). When research was conducted, it was seen that the model was optimized for English and may have limitations in other languages. Out-of-memory error continued for 50 and 20 records. Since the output for 10 records was not complete, 7 was chosen as the appropriate number. However, while getting results in 7's, an out of memory error was received again after 2 cycles. For each memory error, it was necessary to restart the entire notebook. After only 63 records were assigned, the study was terminated when it was realized that resource usage was high. Among the 100 compute units provided by Colab with monthly paid subscription, 39 units were used.

45. **Anthropic Claude**: Anthropic Claude API keys are free to obtain, but they are subject to various restrictions. Claude API can be obtained after signing up via email and provides a $5 initial credit. Using this credit, HS estimation was made based on the Google Sheets for the 715 records. Since Claude gave relatively better and more fluent answers, an annotated prediction was taken instead of just the HS code. In this case, alternative predictions or six-digit predictions were also obtained. It took approximately 6 hours to estimate the 715 records.

46. **ChatGPT 4**: With ChatGPT 4, two predictions are made on Google sheets: short and long. ChatGPT4 worked very fast and gave results within minutes. API key is not used here. Since the API key is not required and predictions were made very quickly. For this reason, 5,000 more randomly selected records from 2022 and 2023 sales e-invoices were assigned HS codes. These will be checked by humans later and the success rate of the models will be better understood.

47. As a result, the following results were obtained for the dataset of 715 records assigned to CPA:

Model	*Number of 6-fold HS/CN that the model can predict (a)*	*The model predicts at least one of the CPAs is correct (b)*	*% (b/a)*	*% (b/715)*	*Prediction by the model is correct (c)*	*% (c/a)*	*% (c/715)*
ChatGPT (manual)	611	418	68%	58%	362	59%	51%
ChatGPT (API)	595	371	62%	52%	326	55%	46%
Claude	539	300	56%	42%	264	49%	37%
ChatGPT4	340	181	53%	25%	169	50%	24%
Gemini	557	225	40%	31%	178	32%	25%
llama	56	0	0%	0%			

48. When the above table is analysed, the most successful results were achieved with ChatGPT, which is entered manually by a human commanding one by one. However, this success rate, which reaches 60 per cent among the records that can be assigned CPA, is not satisfactory in terms of both performance and cost, as it will decrease even more when all e-invoice dataset is considered, whether CPA can be assigned or not.

49. As a result, TurkStat currently does not have the environment to run offline LLM. Nevertheless, some trials were made by forcing the conditions – by paying for computation power. Online LLMs put data security at risk and require the creation of a corporate user.

50. During these studies, research has been conducted to use LLMs with maximum efficiency in text classification, but the desired success rates of LLM models in text classification have not been achieved. Our research on this subject continues.

VI. Classification prediction portal

51. One of the most important functions of AI is to classify data. Like other National Statistics Offices TurkStat also uses many classifications for:

- Occupation
- Product
- Activity
- Trade and
- Education etc.

52. A portal is being prepared to develop predictions according to the classification selected for single or bulk inputs to be provided for use in corporate studies. This portal is designed to meet both internal needs (classifying relevant fields in survey and administrative data) and external needs (assisting users in solving classification problems). The portal will be interoperable (with classification portal, data collection portal, etc.).

53. The most important objectives in the preparation of this prediction portal are

- To create a recommendation system to support the code assignment processes
- To play an assisting role in the code assignment processes of specialized staff

- To increase labour efficiency by using the portal, especially in studies where a large amount of code needs to be assigned
- To prevent errors in code assignment processes and
- To contribute to improving data quality.

54. So far, forecasting models have been developed for ISCO-08, COICOP and Nomenclature of Economic Activities (NACE). In the future, other classifications will be added and the infrastructure will be prepared to serve the portal.