A Picture is Worth a Thousand Definitions:

Validating Company Data using Artificial Intelligence and Satellite Data

Gabriela Alves Werb[1,2], Lisa Reichenbach[1], Ece Yalcin-Roder[1], Susanne Walter[1]

[1] RDSC, Deutsche Bundesbank

[2] Business Information Systems, Frankfurt University of Applied Sciences

The statistics department of the Deutsche Bundesbank offers high-quality data to analysts, researchers, and the public for assessing the financial system and supporting policymaking. However, the company master data, which are crucial for analyses at the microeconomic level, originate from several sources and may exhibit gaps, conflicting information and lack standardization. Therefore, comprehensive processing, including extensive manual checks are required to ensure consistent and reliable data quality. Accurate sector classification and regional location of companies is particularly important for meaningful analyses and valid conclusions. This project aims to leverage contextual information from alternative unstructured data sources to identify and rectify potential misclassifications, fill in missing data, and enhance the overall quality of company master data. By incorporating additional data such as textual descriptions of their activities and images of their products as well as satellite and street view images of their production facilities and office locations, we expect to gain valuable insights for more accurate sector and location determination. To tackle the challenges of handling different data representations in multiple modalities (tables, text, images), we will explore multimodal artificial intelligence (AI) methods such as LLMs and NLP as an innovative approach to automate and supplement current manual data quality checks.

Besides the goal of improving the quality of company data, we aim to extract information from satellite images of the regional company facilities and assets to retrieve a more accurate estimation of the potential damage from climate change imposed by physical risks or wild fires. Physical risk indicators comprise the potential regional climate risk (risk of flooding e.g.) but also the potentially endangered assets. The satellite images of companies' local facilities provide a basis for estimating these endangered assets.

This project aims to demonstrate AI's potential to enhance data quality and integrity and to promote efficiency in data production. The expected findings and insights from this project have implications beyond official statistics, suggesting new research paths for using alternative data sources and AI to improve data quality in various contexts.

Keywords

Data Quality; Multimodal Learning; Artificial Intelligence; Large Language Models; Natural Language Processing; Unstructured Data; Official Statistics; Satellite Images; Street View