Language Models for Automated Coding in Official Statistics: Risks and Opportunities

Mauro Bruno, Marco Di Zio, Francesco Ortame [1]

Abstract

Statistical classifications, such as NACE, are especially important for National Statistical Offices (NSOs). Automating the process of classification has been a focal point for NSOs over the last decades, implementing deterministic, rule-based methods to directly assign codes to natural language queries. While accurate, these systems are human labour-intensive and generalise poorly across domains, each requiring their own rule-set. Recent advancements in natural language processing, particularly language models, offer an opportunity to improve the efficiecy and scalability of the coding process via semantic search systems. However, they come with their own set of challenges, such as interpretability, hallucinations and uncertainty quantification. In this work, we address these issues conformalising their predictions into a robust statistical framework, namely conformal prediction.

Keywords: Automatic Coding, Language Models, Conformal Prediction.

[1] Mauro Bruno (mbruno@istat.it), Italian National Institute of Statistics (Istat), Italy;
Francesco Ortame (francesco.ortame@uniroma1.it), Sapienza University of Rome, Italy;
Marco Di Zio (dizio@istat.it), Italian National Institute of Statistics (Istat), Italy.