

BIBB-Kolloquium zur Berufsbildungsforschung

29.9.2026

14.15 Uhr, Bundesinstitut für Berufsbildung, Bonn,
Raum 5.4.08 – Botanischer Garten

Dr. Erdem Yilmaz

Brussels Institute for Social and Population Studies

zum Thema:

From 4-Digit Codes to Social Class: Benchmarking Supervised Classification, Semantic Retrieval, and LLM Reranking for Multilingual Occupational Coding and Downstream Stratification

This report documents the design, implementation, and evaluation of a fully local pipeline that codes multilingual free-text occupation responses into the International Standard Classification of Occupations (ISCO-08) and derives from those codes the social-class and status measures (e.g., ESeC, ISEI, Oesch) on which much of stratification research depends. This coding is costly, slow, and only moderately consistent when done by hand. The pipeline is local (hence GDPR-friendly), reproducible, and runs on consumer-grade hardware. We benchmark and compare three approaches to the coding step: (M1) a supervised multilingual transformer classifier; (M2) semantic retrieval against a 361k-document multilingual occupation index; and (M3) the supervised classifier with an added large-language-model reranking step. On held-out European Social Survey (ESS) data the supervised classifier reaches 59.2% exact four-digit accuracy, comparable to human inter-coder agreement for this intrinsically ambiguous task, which is itself low at the four-digit level (reported by past research to be $\approx 42\text{--}53\%$ between individual coders; $\kappa \approx 0.5$ between agencies). More importantly for stratification research, given that coding errors are mostly near-misses that coarser schemes absorb, the sociologically relevant aggregate measures are recovered far more accurately, namely 80.4% for the 9-class ESeC (87.8% for its 5-class collapse), 80.5% for the 7-group ESeG, 77.9% for the 16-category Oesch scheme, and a Pearson correlation of 0.87 for the ISEI socio-economic index. The same class-and status-level accuracy holds when the model is trained and tested on canonical occupation text, so it is not an artefact of the ESS sample. Retrieval alone (M2) is substantially weaker, and contrary to the intuitive expectation, adding an LLM reranker (M3) reduces accuracy on noisy multilingual survey responses (-3.7 pp), while on clean canonical titles its effect is approximately neutral, i.e., $+0.5$ pp over the backbone it reranks, with no gain over the best supervised model. Instead, the supervised model's (M1) well-calibrated confidence ($\text{ECE} = 0.038$) supports a human-in-the-loop workflow in which only genuinely uncertain cases are escalated to human coders. We report bootstrap confidence intervals and paired significance tests for every comparison, a train/test leakage audit, and a decomposition of geographic variation showing that the cross-country accuracy spread (in the ESS corpus) reflects response style and occupational-title ambiguity rather than language resources, whereas a genuine resource effect appears only across languages on canonical text in the normative corpus. The pipeline runs entirely offline on a single consumer-grade workstation, making automated, auditable occupation coding feasible without sending respondent data to external services.