Breaking the HISCO Barrier: AI and Occupational Data Standardization

Christian Møller Dahl, Christian Vedel[1],
University of Southern Denmark

How do we standardize large collections of historical occupational data? For decades, the answer has been HISCO codes. However, the manual work involved in processing and classifying occupational descriptions is error-prone, tedious, and time-consuming. This paper introduces a new tool powered by a language model. The neural network takes occupational descriptions as inputs and outputs HISCO codes, thereby transforming the task of HISCO coding into something that takes seconds rather than months. This approach is shown to have similar, if not better, than human performance — for instance, 98 percent precision and 97 percent recall in Danish census data. Moreover, the method is shown to apply (with similar performance) across different languages and a diverse collection of sources. Millions of individual-level occupational descriptions found in sources such as censuses and marriage certificates contain valuable information that can be used to gain new insights. Our tool breaks the metaphorical HISCO barrier and makes this data readily available for analysis of occupational structures with broad applicability in economic history, labor economics, and economics more broadly.

[1] Corresponding author: Christian-vs@sam.sdu.dk, https://sites.google.com/view/christianvedel