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 in labelling accuracy. Moreover the underlying model is mulilingual and trained on HISCO data from various sources and various languages. So far the model has shown great performance on Norwegian, Danish, Swedish, English and Dutch data with still more languages being added as of the writing of this abstract. 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. *This also applies to various Norwgegian sources*. 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