# Breaking the HISCO Barrier: AI and Occupational Data Standardization

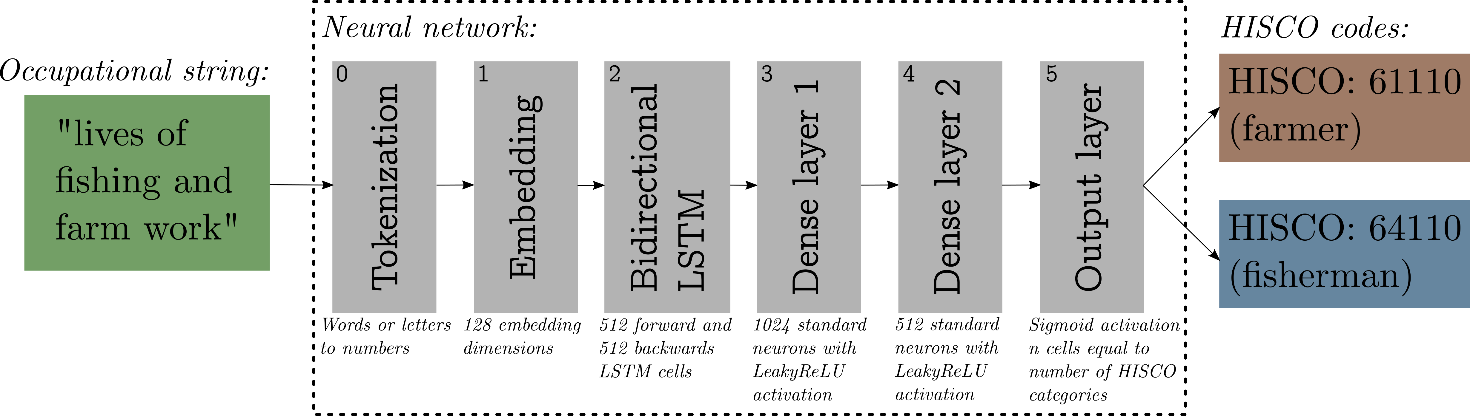
Christian Møller Dahl, Christian Vedel[[1]](#footnote-1),  
University of Southern Denmark

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. Our tool transforms the task of standardising this data from something that previously took several months of work into something that can be done in an afternoon. This is done with the use of a standard neural network to solve the problem, which is demonstrated to perform as well as human labelling. 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.

To show the power of this approach, the existing HISCO labels in the Danish Demographic Database (Clausen, 2015) were used. Most of the original individual-level archives of the Danish censuses have been digitized, and for the years 1787, 1801, 1845 and 1880 many observations have been given manual labels. In total, this yields 3.7 million labels, which can be used as training data. Training an AI on this data enables us to automatically label the remaining census years, which unlocks 8.7 million new individual-level occupational classifications for the Danish population spanning the important 19th century and Denmark’s economic take-off - an important contribution in its own right.

The algorithm consists of a neural network, which is given occupational strings and corresponding HISCO labels. Figure 1 describes the model architecture. Everything is implemented with the broadly popular Keras library for Python (Chollet, 2015), and the architecture is standard for this type of task. As such this project outlines an approach that is available for researchers broadly to implement for their own occupational classification. It is demonstrated for the Danish data, that merely 10,000 labels will give you reasonable performance, while 100,000 labels will generate performance comparable to the hand-labelled data. This dramatically reduces the workload for the classification of millions of observations found in censuses. Furthermore, the potential for transfer learning is large. Training on one set of data can be used to classify occupations in entirely different data of the same period and language, which reduces workload even more.

Figure 1: Network architecture with example



*Notes: The neural network used for the task has the architecture described by the diagram. An occupational string e.g. ‘lives of fishing and farm work’ is (0) tokenized i.e. each word is turned into a unique number for that word or each letter is turned into a unique number for that letter, (1) these tokens are embedded in 128 dimensions, which is then (2) fed into a bidirectional set of 1024 LSTM cells. This is then fed through standard Dense neurons (3) and (4) and finally turned into a prediction. For the given example ‘61110’ for farmer and ‘64110’ for fisherman. After layer (2) there is a global average pooling layer and between layers (2, 3, 4) there is a 50% dropout layer in training.*

Using 270 carefully hand-labelled observations, which were not seen by the algorithm in training, the performance of both human labels and the neural network was evaluated. The results showed that while human accuracy was 97.8%, AI accuracy was 95.9%. Furthermore, when converting to SES scores using the HISCLASS system, the AI performed better with an RMSE of 0.244 compared to 0.296 for human labelling. Even though the model is trained on human-labelled data, it outperforms human accuracy in the SES score conversion, demonstrating the potential of AI to improve upon traditional manual methods of occupational data standardization.

To test the amount of data required to get the algorithm to function adequately it was trained using 100, 1000, 10,000, and 100,000 labels as well as with all the labelled observations from the Danish data. From this exercise, it is demonstrated that 10,000 labels will generate adequate predictive performance while 100,000 labels give a performance which is indistinguishable from using the full 3.7 million observations. Preliminary work with data from English marriage certificates also shows promising performance.

# References

Clausen, N. F. (2015). The Danish Demographic Database—Principles and Methods for Cleaning and Standardisation of Data. Population Reconstruction, 1–22. <https://doi.org/10.1007/978-3-319-19884-2_1>

Chollet, F., & others. (2015). Keras. GitHub. Retrieved from <https://github.com/fchollet/keras>

1. Corresponding author: [Christian-vs@sam.sdu.dk](mailto:Christian-vs@sam.sdu.dk), <https://sites.google.com/view/christianvedel> [↑](#footnote-ref-1)