Breaking the HISCO Barrier: Automatic Occupational Standardization with OccCANINE*

Christian Møller Dahl, Torben Johansen, Christian Vedel,
University of Southern Denmark

Abstract

This paper introduces a new tool, *OccCANINE*, to automatically transform occupational descriptions into the HISCO classification system. The manual work involved in processing and classifying occupational descriptions is error-prone, tedious, and time-consuming. We finetune a preexisting language model (CANINE) to do this automatically, thereby performing in seconds and minutes what previously took days and weeks. The model is trained on 14 million pairs of occupational descriptions and HISCO codes in 13 different languages contributed by 22 different sources. Our approach is shown to have accuracy, recall, and precision above 90 percent. Our tool breaks the metaphorical HISCO barrier and makes this data readily available for analysis of occupational structures with broad applicability in economics, economic history, and various related disciplines.

JEL codes: C55, C81, J1, N01, N3, N6, O33,
Keywords: Occupational Standardization, HISCO Classification System, Machine Learning in Economic History, Language Models

*We would like to thank Andreas Slot Ravnholt for conscientious RA work in the early stages of this project. This paper has also benefited from conversations and feedback from Simon Wittrock, Paul Sharp, Matthew Curtis, Casper Worm Hansen, Julius Koschnick and Hillary Vipond. Thanks also to Bram Hilkens who checked 200 Dutch HISCO codes generated by our model. This project was presented at the 12th Annual Workshop on "Growth, History and Development" (SDU), the PhD seminar series (SDU), at the 5th ENCHOS meeting (Linz), the 2nd Norwegian Winter Games in Economic History (Oslo), a Workshop on Machine Learning in Economic History (Lund) and the 3rd Danish Historical Political Economy Workshop (SDU). We would like to thank every participant at these events for insightful questions and comments. The codebase behind this paper as well as the paper itself benefited from improvements suggested by ChatGPT.
All code and guides on how to use *OccCANINE* is available on GitHub
https://github.com/christianvedels/OccCANINE

1 Introduction

The study of occupational outcomes requires systematic data on people's occupations and the HISCO (Historical International Standard Classification of Occupations) system has emerged as the standard for categorizing diverse occupational data. However, the manual classification of vast datasets into HISCO codes has been an arduous and time-consuming process for researchers thus hampering progress. A simple back-of-the-envelope exercise demonstrates the problem well: Even a highly experienced researcher might spend 10 seconds recognizing and typing the correct HISCO code for any given occupational description and even for 10,000 unique occupational descriptions this would mean that the researcher would spend in the order of 28 hours coding everything, or 280 hours (11 days - no breaks) for 100,000 observations.

In this paper, we present a solution that transforms the task of coding occupations into something which is done automatically in a few minutes or a couple of hours including verification of the quality. We introduce *OccCANINE* - a transformer language model (J. H. Clark, Garrette, Turc, & Wieting, 2022) (CANINE) - which we fine-tune on 14 million observations of occupational descriptions with associated HISCO codes in 14 different languages. This training data was generously contributed by 22 different research projects, each of which is cited in Table 1. The outcome is a model with an impressive overall accuracy of 93.5 percent,[1] capable of taking a straightforward textual description of an occupation and accurately determining the most applicable HISCO codes associated with it.

The HISCO system was introduced in an effort to produce internationally comparable occupational data (Leeuwen, Maas, & Miles, 2002). It, and its various modifications, has since then become the most widely used classification scheme for historical occupation with the so-called PST system being the most widely spread alternative (Wrigley, 2010). It should be noted, that the model presented here, can be fine-tuned into other classification systems with relative ease.

By significantly reducing the time and effort required for HISCO coding, our tool democratizes access to historical occupational data analysis, enabling researchers to conduct more extensive and diverse studies and dedicate more time to data quality. This breakthrough has the potential to unlock new insights into occupational trends and shifts over time, contributing valuable knowledge to the fields of economics, sociology, political science, history, and many related fields. Furthermore, this paper stands as a recipe on how to solve a wide range of similar problems, where many messy descriptions need to be classified into some system. Similar problems are found in historical customs records, educational descriptions, and much more. All of these can be addressed with a similar setup.

The remainder of this paper proceeds as follows. Section 2 motivates our solution. Section 3 outlines the model architecture, training data and training procedure. Section 4 describes how well our method performs. Section 5 concludes with recommendations on how to use *OccCANINE*.

[1] 95.5 percent precision, 98.7 percent recall and an F1-score of 96.0 percent.

2 Motivation

2.1 The problem of occupational coding

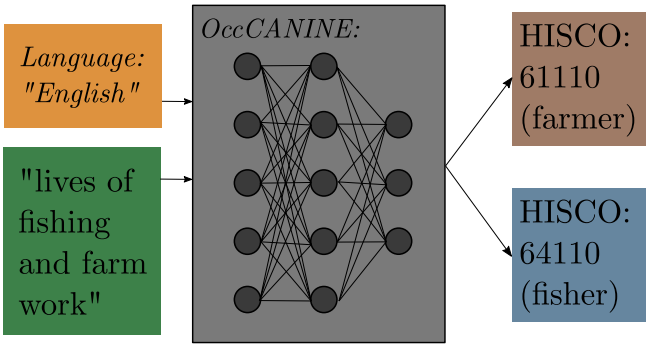
To be able to generate insights into everything from women's empowerment, social mobility, the effect of railways, first nature geography and the origins of the Industrial Revolution to the interplay of technology and development, we explicitly or implicitly need to know what people did for a living historically.[2] Given the value of addressing these inquiries (among numerous others), it becomes worthwhile trying to acquire large amounts of historical occupational data. This data usually comes in the form of large lists of textual descriptions. "Lives of fishing and farm work", is a stereotypical entry found in sources such as censuses, marriage certificates, etc. The task of the researcher is then to take these descriptions and turn them into standardized occupational categories (61110: 'Farmer' and 64100: 'Fisherman' in the HISCO system). With the invention of HISCAM (Lambert, Zijdeman, Leeuwen, Maas, & Prandy, 2013) and its derivations (G. Clark, Cummins, & Curtis, 2022), it has also become common to convert these categories into a single measure of social status based on occupation.

The challenge of transforming raw textual occupational descriptions into standardized categories is not trivial, necessitating either a lot of manual work by error-prone research assistants or sophisticated methods for interpreting and categorizing text data using the classical natural language processing toolbox. In particular, the diversity of occupational descriptions is a problem.[3] The classical approach to HISCO coding involves classical string matching and string cleaning using e.g. regular expressions. Because of negations, changing spelling conventions, typos and transcription errors, this quickly becomes complex and error-prone. Typically the following steps are involved:

1. Domain knowledge is applied in forming rules for cleaning strings: "Srvnt" becomes "servant", "sgt." becomes "sergent", and so on.
2. Stop words are removed: "He is the servant" becomes "servant", "after a long career he retired" becomes "retired".
3. The unique remaining strings are manually matched to the HISCO catalogue

This pipeline needs to be repeated for every single source with little scope for generalisability. *OccCANINE* replaces all of that.

Figure 1: Conceptual model



Notes: This illustrates the conceptual model: A neural network takes occupational descriptions and language as inputs and outputs relevant HISCO codes.

2.2 A faster, better, scalable and replicable solution

The primary barrier to overcome is that previous methods are either slow, inaccurate, lack scalability or do not generalize effectively across different data sources. Our solution addresses all of this. We teach a language model to understand occupational descriptions as a person would. We finetune a preexisting language model on 14 million pairs of descriptions and HISCO codes. As such, we end up with a model, which inherently captures the occupational meaning of inputted strings. This means, that we can input occupational descriptions with typos, spelling mistakes, etc. The model then (similarly to ChatGPT but smaller) draws on a vast knowledge of language and similarity (within and across the languages it is trained on) to output the appropriate HISCO code. In effect, all the steps described in Section 2.1 are replaced by one step: The input consists of a raw occupational description and a language as context. HISCO codes are provided as outputs (see Figure 1). This approach has the following advantages:

1. It requires no string cleaning. The text as transcribed is fed directly into the model.
2. It is as accurate if not more accurate than a human labeller.
3. The model has a general understanding of historical occupations, which means it generalises well to other settings with little or no fine-tuning.
4. It is fully replicable. Given the same inputs, *OccCANINE* will always deliver the same HISCO codes. Replicability has innate scientific value but it also reduces the humanly introduced variance and resultant attenuation bias in downstream analysis.

[2]Allen (2009); Berger (2019); G. Clark (2023); Goldin (2006); Lampe and Sharp (2018); Mokyr (2016); Vedel (2023); Vries (2008).

[3]The Danish censuses 1787-1901 (Clausen, 2015; Robinson, Mathiesen, Thomsen, & Revuelta-Eugercios, 2022) contain no fewer than 17,865 unique descriptions corresponding to the occupation 'farm servant'.

2.3 Literature

The closest related literature is in two strands: Occupational medicine and administrative and survey data. We will introduce the recent advances in this chronologically. The chronological improvement in performances demonstrates the rapid underlying technological development, which now allows the simple method we propose in this paper. A central term in this entire literature is production rate: The number of occupational descriptions one chooses to automatically transcribe (where the rest are left to a human labeller). This is typically done by only automatically transcribing some rate (e.g. 80 percent) of observations for which a label is assigned with the highest probability.

Early approaches only perform well at lower production rates. Patel, Rose, Owens, Bang, and Kaufman (2012) demonstrate 89 percent agreement with a human labeller for 71 percent production rate. They mainly rely on rule-based classical NLP-approaches. Gweon, Schonlau, Kaczmirek, Blohm, and Steiner (2017) suggest combining classical rule-based approaches combined with bag-of-word cosine distance and nearest neighbours matching. For 'fully automatic labelling' (100 percent production rate),[4] they achieve 65 percent accuracy on German survey data for the ISCO-88 system.[5] They suggest the method is used at lower production rates, where higher performance is demonstrated. As such, a large chunk is still left for the human labeller.[6] Schierholz and Schonlau (2020) review this and other machine learning approaches to automatic occupational labelling. Using boosting trees, they demonstrate around 78 percent agreement in 100 percent production rate.

More recent literature builds on the introduction of Transformers (Vaswani et al., 2017) and pre-trained models like BERT (Devlin, Chang, Lee, & Toutanova, 2018). Garcia, Adisesh, and Baker (2021) implement a method that combines traditional exact matching and data cleaning techniques with advanced text analysis methods, including TF-IDF and Doc2Vec (utilizing BERT), for cases that do not match exactly. This approach is then applied within a conventional machine learning framework, resulting in a macro F1-Score of 0.65 and a top-5 per-digit macro F1-Score of 0.76, as evaluated within the context of the Canadian National Occupational Classification Scheme. The research most comparable to ours is conducted by Safikhani, Avetisyan, Föste-Eggers, and Broneske (2023). They fine-tune German BERT and GPT-3 on 47,526 observations of German survey data to classify these into the German KldB system. They end up with a maximum Cohen's kappa of 64.22 percent for the full occupational code in their test data.[7] This should be compared to Schierholz and Schonlau (2020), which achieves only 48.5 percent Cohen's kappa on the same test data. To the extent that our data is comparable, we beat the performance by a large margin by achieving an overall Cohen's kappa of 88.9 percent, accuracy of 93.5 percent, and an F1-score of 96.0 percent on our test data. We do not consider lower levels of 'production rate' because it is barely relevant at this level of performance. Moreover, our method requires no string cleaning, correction of spelling mistakes, or stop word removal. The model implicitly handles all of this.

[4] Which still requires stop word removal, and other tweaks.

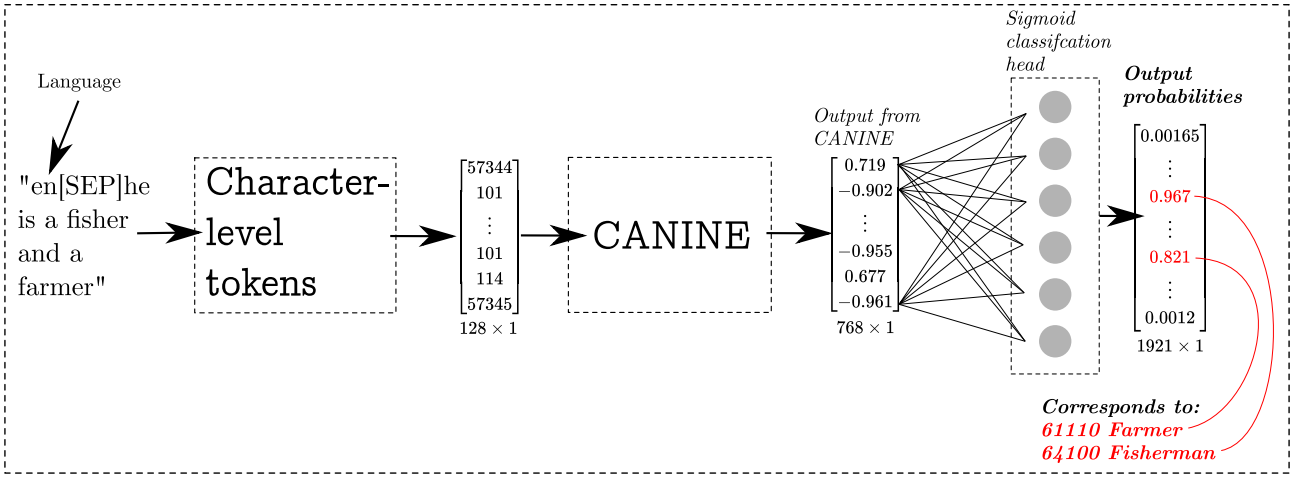
[5] Which is similar to the HISCO system.

[6] To achieve 90 percent accuracy, Gweon et al. (2017) requires around 40 percent manual labelling.

[7] Cohen's kappa is roughly comparable to accuracy but takes into account the agreement that occurs by chance.

3 Architecture, data and training procedure

Figure 2: Architecture of *OccCANINE*



Notes: Inputted text is converted to character-level tokens, which then serves as input into the CANINE architecture. The 768×1 output of this is then passed through a sigmoid classification head. The output is a 1921×1 vector which represents the probability of each available HISCO label. In the given example, the elements corresponding to the code for farmer and fisherman have a high probability.

3.1 Architecture

We make use of the CANINE architecture (J. H. Clark et al., 2022), which is a modestly sized (127 million parameter) language model based on the common transformer architecture (Vaswani et al., 2017). It was pre-trained on 104 languages of Wikipedia data. The choice of this particular architecture has a threefold motivation: *First*, the Wikipedia pre-training ensures multilingual capabilities, and it is reasonable to assume some similarity between historical occupational descriptions and Wikipedia. *Second*, we want this to be a relatively broadly available tool, and the model is small enough that it is possible to run (and even to do some small finetuning) on a common laptop. *Third*, and most importantly in this case, the model is based on character-level tokenization. The most commonly used tokenization approaches work at the word or wordpiece level. But for archival data, the risk is that this will cause unnecessary model variance. When 'farmer' is mistyped as 'frmter', this would put strain on traditional methods, due to tokenization weirdness, but the CANINE architecture is more robust to this.

On top of the CANINE model, we add a classification head of size $[1 \times 1921]$, one output for each of the 1921 potential codes in the HISCO system.[8] The entire model architecture is illustrated in Figure 2. We input both the language (as context) and an occupational description. This is passed through the CANINE architecture and the classification head to get a vector of probabilities. These are then turned into specific predictions based on an optimal threshold derived in Section 4. We chose a sigmoid classification head over softmax, since we do not want the probabilities to be normalized. Thus the model independently assigns a probability to each potential HISCO code. This has the advantage, that there is no implicit penalty for the model to predict more than one occupation in cases where this is appropriate. However, the model does not distinguish between cases where multiple HISCO codes fit because of ambiguity and cases where multiple HISCO codes fit because that person had multiple actual occupations. If it is desired in some applications it is trivial to normalize the output to sum to one by dividing the output by its sum.

3.2 Data

This project utilizes training data that was obtained from public sources or provided by fellow researchers, for which we express our sincere gratitude. We developed a semi-standardized framework to process all of this data and make the best use of it.[9] This involves four steps. *First*, we replaced all non-English characters from the occupational descriptions with a standard English equivalent.[10] *Second*, we manually checked the data for thoroughly for peculiarities.[11] *Third*, we made sure that only standardized HISCO codes were used and removed any observations with non-standard HISCO codes.[12] *Fourth*, a common practice is for manual labellers to only include one occupation, even when a description corresponds to two different occupations or more (such as fisherman and farmer). To enhance our model's chance of picking this up, descriptions were linked by the conjunction that serves the same function as 'and' in each particular language. E.g. 'he is a farmer' + 'he fishes' becomes 'he is a farmer *and* he fishes'.[13] In total, our data consists of 18 million observations. Of this, we use 14 million observations in training (of which 50,000 observations are used in cross-validation during training). The rest is divided among post-training validation data (10 percent) - which is what we draw on for Section 4 for now - and final model testing to be performed when we are entirely sure that no more model specification decisions are to be made (5 percent). Moreover, we do out-of-distribution (OOD) evaluation with entirely different data sources in English, Danish, Swedish and Dutch.[14]

[8] As defined by `https://github.com/cedarfoundation/hisco`.

[9] Details of which can be seen in 'Data_cleaning_scripts/' in the GitHub repository of this project.

[10] E.g. 'æ' became 'ae', 'ø' became 'oe', etc.

[11] An example of this, is that often there would be a 'raw' occupational description and a 'clean' occupational description. In this cases both would become training data.

[12] IPUMS (MPS, 2020) have their own adaptation of HISCO, for which reason we took an extra step and only used data for which the HISCO codes are unaltered from the original standard. For this we used the cross-walk provided by Mourits (2017).

[13] Such combinations were randomly drawn within each data source marked with * in Table 1. For each unique occupational description, 10 random combinations were drawn.

[14] Schneider and Gao (2019), Copenhagen Burial Records from Robinson et al. (2022), Enflo, Molinder, and Karlsson (2022), Soetermeer (1674).

Table 1: Training data

Shorthand name	Observations	Percent	Language	Source
DK_census	5,391,656	29.794%	da	Clausen (2015); The Danish National Archives
EN_marr_cert	4,046,203	22.359%	en	G. Clark et al. (2022)
EN_uk_ipums	3,026,859	16.726%	en	MPS (2020); Office of National Statistics
SE_swedpop	1,793,557	9.911%	se	SwedPop Team (2022)
JIW_database*	966,793	5.343%	nl	Moor and van Weeren ("2021")
EN_ca_ipums	818,657	4.524%	unk	MPS (2020); Statistics Canada
CA_bcn*	644,484	3.561%	ca	Pujades Mora and Valls (2017)
HISCO_website*	392,248	2.168%	mult	HISCO database (2023)
HSN_database	184,937	1.022%	nl	Mandemakers et al (2020)
NO_ipums	147,255	0.814%	no	MPS (2020)
FR_desc*	142,778	0.789%	fr	HISCO database (2023)
EN_us_ipums	139,595	0.771%	en	MPS (2020); Bureau of the Census
EN_parish	73,806	0.408%	en	de Pleijt, Nuvolari, and Weisdorf (2019)
DK_cedar	46,563	0.257%	da	Ford (2023)
SE_cedar*	45,581	0.252%	se	Edvinsson and Westberg (2016)
DK_orsted	36,608	0.202%	da	Ford (2023)
EN_oclack*	24,530	0.136%	en	Mourits (2017)
EN_loc*	23,179	0.128%	en	Mooney (2016)
IS_ipums	20,459	0.113%	is	MPS (2020)
SE_chalmers	14,426	0.08%	se	Ford (2023)
DE_ipums	8,482	0.047%	ge	MPS (2020); German Federal Statistical Office
IT_fm*	4,525	0.025%	it	Fornasin and Marzona (2016)

Notes: This is a comprehensive overview of the data used for training our model. The shorthand name is the name we use in the remainder of this paper. Observations are the effective number of observations we have after cleaning procedures. The different languages found in the training data are also listed. Data marked with a * is data where combined occupations were created as described in Section 3.2.

3.3 Training

The model was trained for 26 days, 21 hours, 3 minutes, and 7 seconds on an NVIDIA A100 80GB GPU using the AdamW optimizer implemented in PyTorch (Loshchilov & Hutter, 2019; Paszke et al., 2019). It was trained on batches of 256 observations for a total of 42 epochs. Every time the model improved in validation accuracy over earlier instances, it was saved. CANINE has 10 percent dropout between all layers by default. We further regularized the training procedure with simple string augmentation in the form of random character changes and random word insertions.[15] This approach (although here in a much simpler implementation) is inspired by TextAttack by Morris et al. (2020). In the training procedure, the language is provided first, followed by a separator and then the occupational description. To improve cross-lingual performance, we randomly set the language for an observation to 'unk' (for *unknown*) with a 25 percent probability. As a result, we have a finetuned version of CANINE that not only has an intrinsic understanding of historical occupations but is also multilingual and capable of leveraging language context. Moreover, it is resilient to spelling errors and exhibits strong performance overall, as will be showcased in the following section.

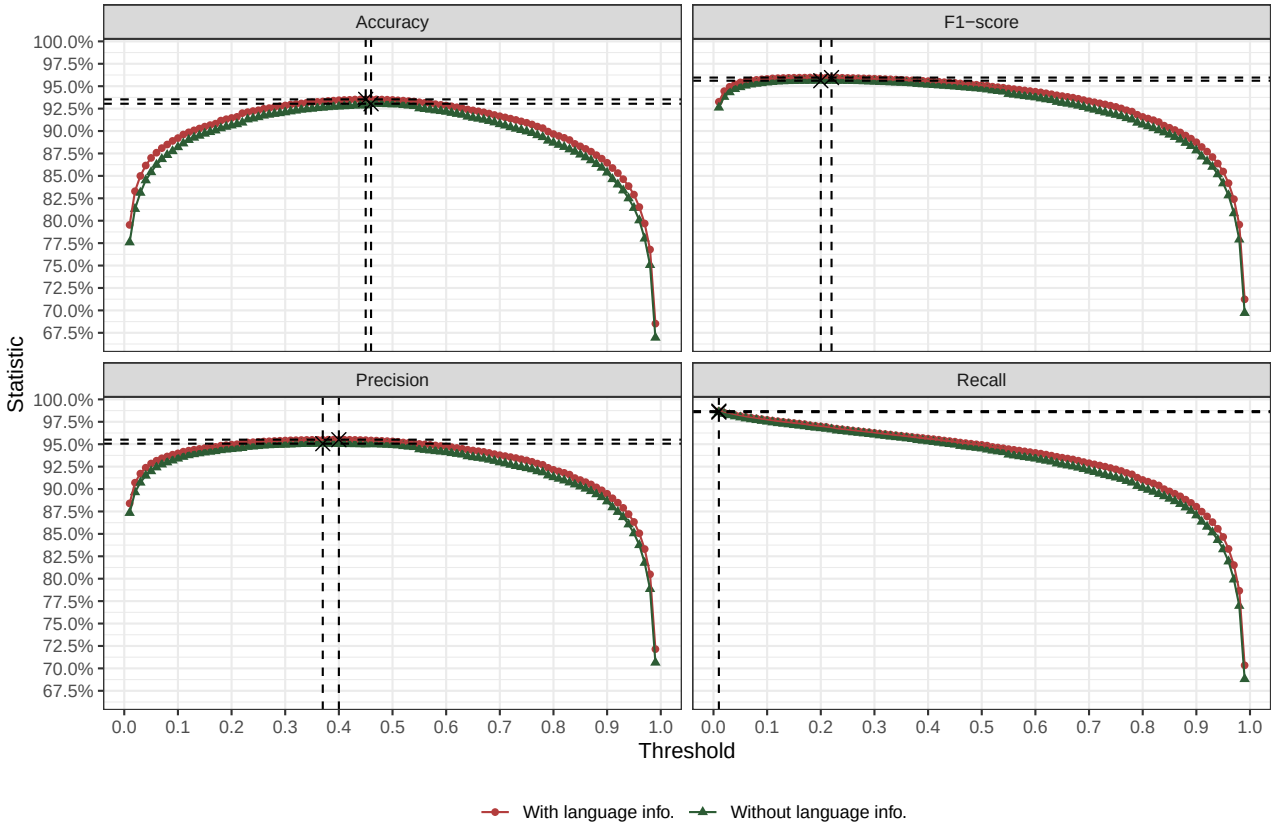
4 Performance

The performance of our model was evaluated on 1 million observations, which were not used for training. Using this data, we test performance at classification thresholds ranging from 0.01 to 0.99[16] in terms of Accuracy (exact match), Precision, Recall, and F1 score. The best overall performance for each of these metrics is presented in Table 2 and Figure 3. The result is 93.6 percent overall accuracy, 95.5 percent precision, 98.2 percent recall and an F1-score of 0.960 when optimal classification thresholds are used.

[15] Each input had a 10 percent chance of random word insertion and a separate 10 percent chance of random character alterations, where each character then had a 10 percent chance of being replaced by a random character.

[16] The threshold controls when we decide that a certain probability of a HISCO code should be turned into a prediction of that HISCO code. We find the optimal threshold with a grid search with a precision of 0.01.

Figure 3: Optimal Threshold



Notes: Model performance at various classification thresholds, depicting Accuracy, F1 score, Precision, and Recall. The red line represents performance with language information and the green line without language information. The dashed vertical lines indicate the optimal thresholds for each metric.

Table 2: Best overall performance

Metric	Lang. info.	Value	Optimal thr.
Accuracy	No	0.930	0.46
	Yes	0.935	0.45
F1 score	No	0.956	0.20
	Yes	0.960	0.22
Precision	No	0.951	0.37
	Yes	0.955	0.40
Recall	No	0.986	0.01
	Yes	0.987	0.01

Notes: This table illustrates the peak performance of our model across 1 million validation observations. Metrics include Accuracy, F1 score, Precision, and Recall, with and without language context. Optimal thresholds (thr.) indicate the point where each metric is maximized. A similar table for each language is available in Appendix A.

Moreover, we manually check four datasets which are entirely out of distribution and not seen during training: The Copenhagen Burial Records (1861-1911) (Robinson et al., 2022) the Indefatigable Training Ship data (1865-1995) (Schneider & Gao, 2019), a dataset of Swedish strikes 1859 to 1902 (Enflo et al., 2022) and a dataset of a Dutch Wealth Tax in 1674 (Soetermeer, 1674). For these tests, we use language-wise accuracy-optimal classification thresholds found in Table A1 in the appendix. We manually reviewed 200 random predictions from each dataset to evaluate the accuracy. The results are shown in Table 3. *OccCANINE* achieves higher than 90 percent accuracy and *substantial agreement* in all cases. The Training Ship data and the Swedish strike data already contain HISCO codes and we had the Dutch Wealth Tax data checked by an expert.[17] Among these, there is a rate of exact agreement (same occupation being assigned the exact same HISCO code) of 82, 80.5 and 67 percent, respectively. However, most cases of disagreement have very similar HISCO codes. For each observation, we checked whether the disagreement was substantial. Here, we define substantial agreement as two labels which are similar enough that the difference would not matter in most applications.[18] From this we know that there is *substantial agreement* between *OccCANINE* and the original labels in 96.5, 93.5 and 92.5 percent of cases respectively. Moreover, *OccCANINE* can be finetuned easily. In the case of the Swedish Strikes data, the model can be finetuned for this purpose in fewer than 10 minutes,[19] after which *OccCANINE* achieves 95.5 percent agreement with the original source on observations, which are not seen during finetuning.

[17] We are grateful to Bram Hilkens for this.

[18] An example of an unsubstantial difference is whether "foreman aircraft repairs" should be labelled as *"22690 Other Production Supervisors and General Foremen"* or *"22610 Production Supervisor or Foreman, General"*.

[19] On a 1080 Ti 11GB GPU. Alternatively, fewer than two hours on a laptop without a GPU.

Table 3: Out of Distribution Testing Accuracy

Dataset	N checked	Accuracy	Exact agreement	Subst. agreement
Copenhagen Burial Records	200	0.950	-	-
Training Ship Data	200	0.985	0.820	0.965
Swedish Strikes	200	0.945	0.805	0.935
Dutch Familiegeld	200	0.925	0.670	0.925

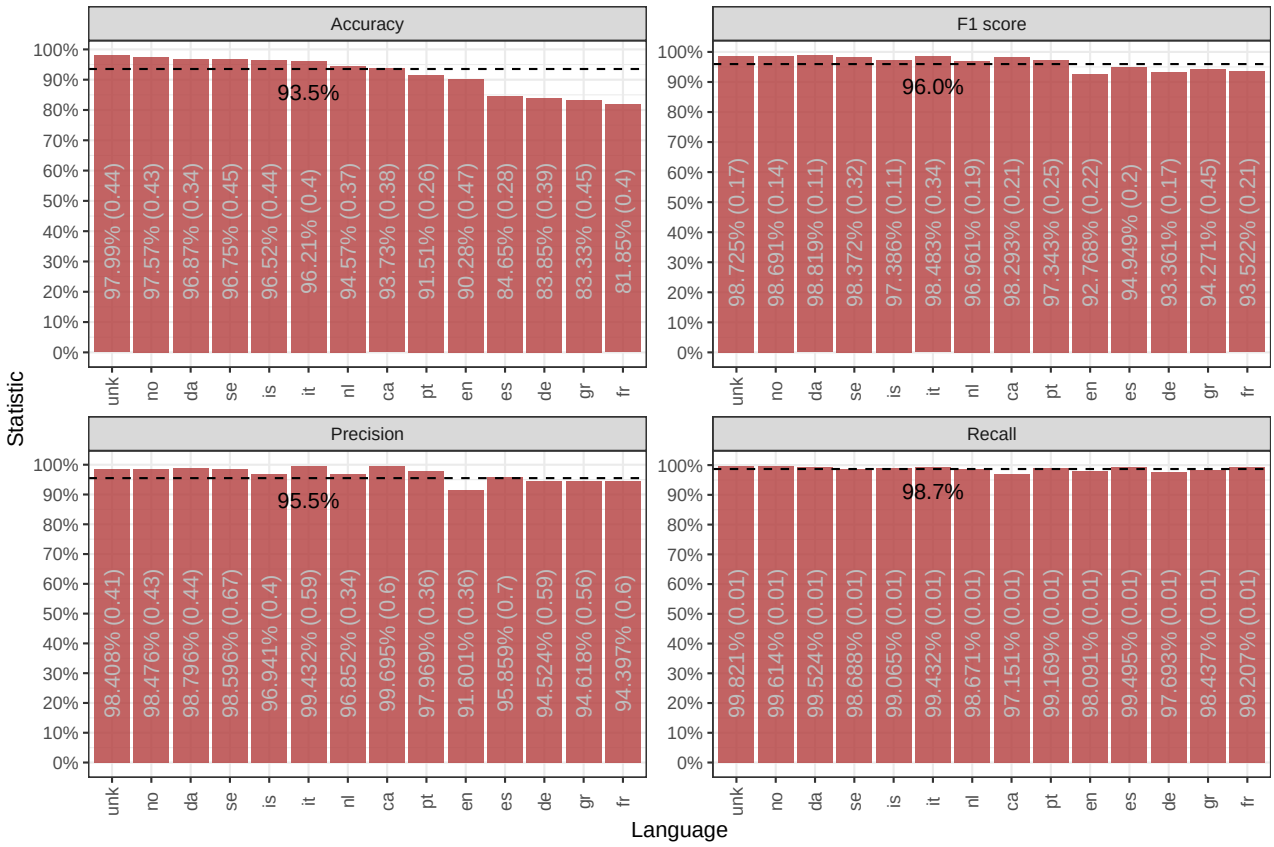
Notes: This table reports accuracy metrics for out-of-distribution tests, emphasizing the model's adaptability. The Copenhagen Burial Records, Training Ship Data, Swedish Strikes data and Dutch Familiegeld each with 200 randomly selected observations, were used to benchmark performance. The language-wise accuracy-optimal classification thresholds found in Table A1 in the appendix were used.

We test the method with and without language context. The model is trained such that 25 percent of training observations will randomly have no language context. As such, we test performance in two different settings: One where the model is not explicitly told anything about the language of the occupational description, and one where it is. There is a small but notable improvement when language information is provided, but in the absence of language information, the model still performs robustly, indicating a multilingual conceptual understanding of occupations (see Table 2). The performance for each separate language is also tested and demonstrated in Figure 4. The method works well for all languages it has been trained on. Appendix A presents a detailed table of these metrics for each separate language (Table A1).

We also test how well the model performs on rare versus frequent occupational categories. As expected the model works best for the most frequent occupational categories.[20] We estimate this relationship non-parametrically and demonstrate high performance is maintained for at least the most frequent 99 percent of occupational data, which accounts for the 524 most common HISCO codes (see Appendix B for more details). A natural concern is whether there is a correlation between outcomes of interest and this accuracy. The rarity of certain occupations could correlate with their socio-economic status (SES), and in turn, the rarity could drive low accuracy of our method. This could potentially introduce systematic bias when using our method in applied settings. Appendix C tests whether such correlation exists. Our analysis reveals neither a statistically significant nor a practically meaningful correlation that would cause concern.

[20] Which is typically also the case for a skilled human labeller.

Figure 4: Performance by Language



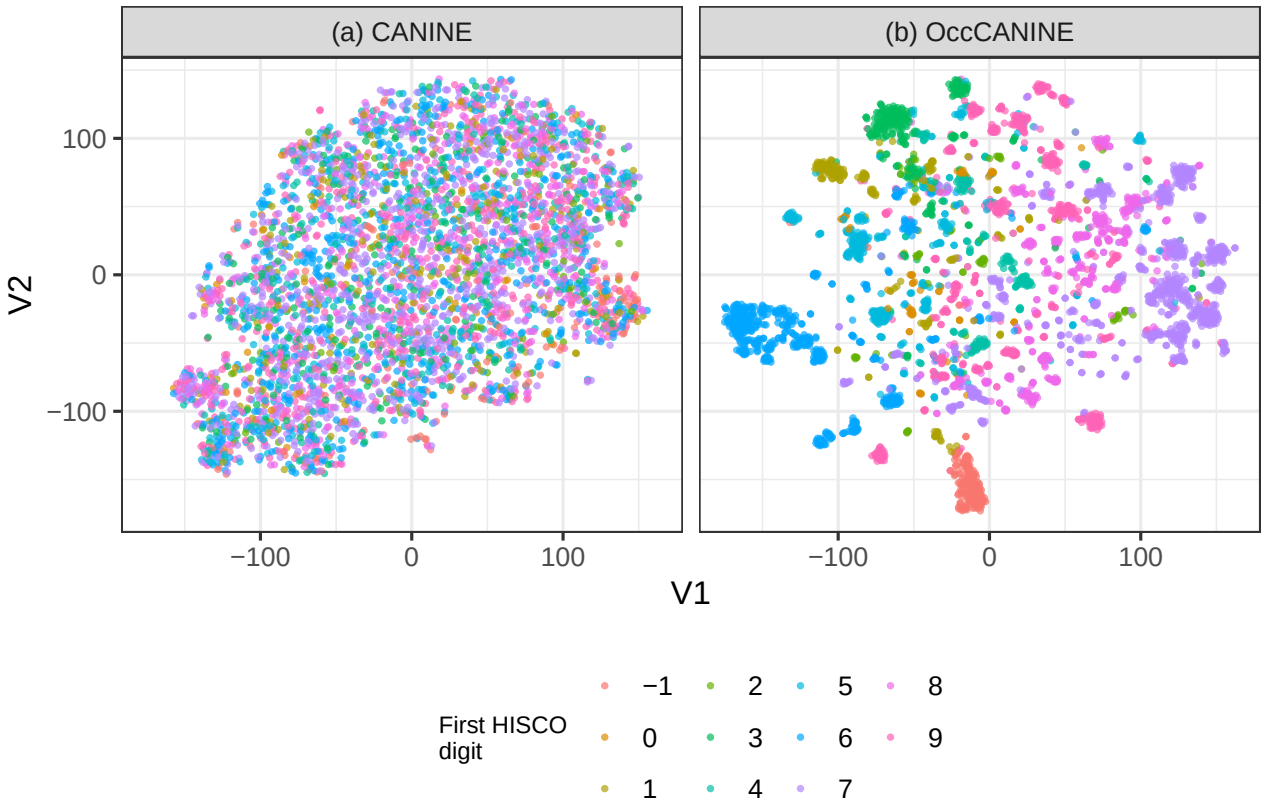
Notes: Performance metrics by language, showing Accuracy, F1 score, Precision, and Recall. Each bar represents a language, with the dashed line indicating the highest value achieved across all languages. Optimal performance per language is annotated on the respective bars. The optimal threshold is shown in the parentheses.

To investigate the underlying semantic knowledge that the model has obtained in training we passed 10,000 validation observations through the model to get embeddings.[21] This 768 dimensional output is ought to represent the meaning of each occupational description. If similar descriptions cluster closely, particularly across different languages, it suggests that the model has acquired a structural comprehension of occupations. Figure 5 shows a low-dimensional representation of these embeddings (using t-SNE to reduce the dimensionality). Panel A shows results from the original CANINE (J. H. Clark et al., 2022) finetuned on Wikipedia. Panel B shows results from *OccCANINE*, which is further finetuned on the historical occupational training data presented. The colours represent the first digit of the HISCO code according to the source. As such, these roughly represent different sectors of the economy. It should be noted that occupations which are closer together tend to have the same colour, and this is in contrast to the results of Panel A. This suggests that the OccCANNINE picks up similar occupations as being semantically similar. This result shows the potential for generalised high performance across different domains, and in turn, it suggests that *OccCANINE* is a valuable starting point for other applications related to occupational descriptions in a historical setting.[22]

[21] The output from the final layer before the Sigmoid classification head.

[22] The model is openly available via `https://huggingface.co/Christianvedel/OccCANINE`.

Figure 5: t-SNE Visualizations of Occupational Embeddings



Notes: Panel (a) illustrates the t-SNE visualization for embeddings derived from the original CANINE model, trained on Wikipedia data. Panel (b) shows embeddings from our version of CANINE further finetuned on historical occupational data. The colours correspond with the first digit of the HISCO code, roughly indicative of economic sectors. The data depicted was not seen during model training.

5 Conclusion and Recommendations

Our comprehensive evaluation demonstrates that the *OccCANINE* model is a powerful tool for automatically transforming occupational descriptions into standardized HISCO codes. We generally recommend a classification threshold of 0.22 to optimize the F1 score and a threshold of 0.45 to maximize accuracy, which we base on the model's performance on 1 million validation observations. Appendix A contains language-specific optimal thresholds. These thresholds should serve as a starting point for researchers, providing optimal accuracy, and a balance between precision and recall that is suitable for most analytical purposes. The public repository contains a step-by-step guide on how to use *OccCANINE*.[23]

[23] See `https://github.com/christianvedels/OccCANINE`.

We hope that the freed resources can be used to do more research, but also increase the quality of research involving standardized data on historical occupations. It is easy to get a measure of accuracy every time this is applied; it takes in the order of an hour to manually verify 100 random observations. We strongly encourage researchers who want to use our method to check at least 100 observations and report the accuracy obtained from this in publications where our method is applied. This practice not only provides an additional layer of validation but also equips researchers with a concrete measure of the model's performance. This contributes to the transparency of the inherent uncertainties in data work.

For projects with unique requirements or when applying the model to underrepresented languages, we provide an interface for task-specific finetuning. This customization is especially recommended if existing domain-specific training data is available, if the application domain diverges significantly from the training data, or if HISCO codes are of prime interest to the research question at hand. If it is necessary to generate training data, users can also leverage the existing model to generate preliminary predictions, which can then quickly be refined manually, significantly reducing the time and effort compared to manual labelling from scratch.

In conclusion, *OccCANINE* represents a significant stride in historical occupational data processing, effectively breaking the HISCO barrier which has long stood. By automating the translation of occupational descriptions into HISCO codes with high accuracy, our model streamlines research in historical social science and paves the way for answering important research questions.

References

Allen, R. C. (2009). The high-wage economy of pre-industrial britain. In *The british industrial revolution in global perspective* (p. 25–56). Cambridge University Press.

Berger, T. (2019). Railroads and rural industrialization: evidence from a historical policy experiment. *Explorations in Economic History*, *74*, 101277. Retrieved from https://www.sciencedirect.com/science/article/pii/S0014498318302080 doi: https://doi.org/10.1016/j.eeh.2019.06.002

Clark, G. (2023). The inheritance of social status: England, 1600 to 2022. *Proceedings of the National Academy of Sciences*, *120*(27), e2300926120. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.2300926120 doi: 10.1073/pnas.2300926120

Clark, G., Cummins, N., & Curtis, M. (2022, Jun). *Three new occupational status indices for england and wales 1800-1939.* (Available at: https://www.mjdcurtis.com/pdfs/ccc-occupational-indices-draft.pdf)

Clark, J. H., Garrette, D., Turc, I., & Wieting, J. (2022). ¡scp¿canine¡/scp¿: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, *10*, 73–91. Retrieved from http://dx.doi.org/10.1162/tacl_a_00448 doi: 10.1162/tacl_a_00448

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

de Pleijt, A., Nuvolari, A., & Weisdorf, J. (2019, 03). Human Capital Formation During the First Industrial Revolution: Evidence from the use of Steam Engines. *Journal of the European Economic Association*, *18*(2), 829-889. Retrieved from https://doi.org/10.1093/jeea/jvz006 doi: 10.1093/jeea/jvz006

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding.* Retrieved from https://arxiv.org/abs/1810.04805

Edvinsson, S., & Westberg, A. (2016). *Swedish Occupational Titles from CEDAR.* IISH Data Collection. Retrieved from https://hdl.handle.net/10622/KNGX6B doi: 10622/KNGX6B

Enflo, K., Molinder, J., & Karlsson, T. (2022). *Sweden, 1859-1902.* IISH Data Collection. Retrieved from https://hdl.handle.net/10622/TAVJXR doi: 10622/TAVJXR

Ford, N. M. (2023). *In the footsteps of chalmers and Ørsted: The expansion of higher technical education in sweden and denmark, 1829-1929.* Unpublished. Retrieved from https://www.nickford.com/ (Unpublished)

Fornasin, A., & Marzona, A. (2016). *HISCO Italian Formasin Marzona 2006.* IISH Data Collection. Retrieved from https://hdl.handle.net/10622/SRVW6S doi: 10622/SRVW6S

Garcia, C. A. S., Adisesh, A., & Baker, C. J. O. (2021). S-464 Automated Occupational Encoding to the Canadian National Occupation Classification Using an Ensemble Classifier from TF-IDF and Doc2Vec Embeddings. *Occupational and Environmental Medicine*, *78*, A161. doi: 10.1136/OEM-2021-EPI.442

Goldin, C. (2006, May). The quiet revolution that transformed women's employment, education, and family. *American Economic Review*, *96*(2), 1-21. Retrieved from https://www.aeaweb.org/articles?id=10.1257/000282806777212350 doi: 10.1257/000282806777212350

Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M., & Steiner, S. (2017). Three methods for occupation coding based on statistical learning. *Journal of Official Statistics*, *33*(1), 101–122. Retrieved from https://doi.org/10.1515/jos-2017-0006 doi: doi:10.1515/jos-2017-0006

HISCO database. (2023). *Hisotry of Work Information System.* https://historyofwork.iisg.amsterdam/. (Accessed: 2023-12-10)

Lambert, P. S., Zijdeman, R. L., Leeuwen, M. H. D. V., Maas, I., & Prandy, K. (2013). The construction of hiscam: A stratification scale based on social interactions for historical comparative research. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *46*(2), 77-89. Retrieved from https://doi.org/10.1080/01615440.2012.715569 doi: 10.1080/01615440.2012.715569

Lampe, M., & Sharp, P. (2018). *A Land of Milk and Butter.* University of Chicago Press. Retrieved from https://ideas.repec.org/b/ucp/bkecon/9780226549507.html

Leeuwen, M., Maas, I., & Miles, A. (2002). *Hisco: Historical international standard classification of occupations.* Leuven University Press. Retrieved from https://books.google.dk/books?id=EMPtAAAAIAAJ

Loshchilov, I., & Hutter, F. (2019). *Decoupled weight decay regularization.*

Mokyr, J. (2016). *A culture of growth: The origins of the modern economy.* Princeton University Press. Retrieved 2024-02-03, from http://www.jstor.org/stable/j.ctt1wf4dft

Mooney, G. (2016). IISH Data Collection. Retrieved from https://hdl.handle.net/10622/ERGY0V (UNF:6:+wMiF+S0BuzIBIXW3zrclA==) doi: 10622/ERGY0V

Moor, T. D., & van Weeren, R. ("2021", "6"). Dataset Ja, ik wil - Amsterdam marriage banns registers 1580-1810. datarepository.eur.nl. Retrieved from "https://datarepository.eur.nl/articles/dataset/Dataset_Ja_ik_wil_-_Amsterdam_marriage_banns_registers_1580-1810/14049842" doi: "10.25397/eur.14049842.v1"

Morris, J. X., Lifland, E., Yoo, J. Y., Grigsby, J., Jin, D., & Qi, Y. (2020). *Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp.* Retrieved from https://arxiv.org/abs/2005.05909

Mourits, R. (2017). *HISCO - OCC1950 CROSSWALK.* DANS Data Station Social Sciences and Humanities. Retrieved from https://doi.org/10.17026/dans-zap-qxmc doi: 10.17026/dans-zap-qxmc

MPS. (2020). *Integrated public use microdata series, international: Version 7.3.* https://doi.org/10.18128/D020.V7.3. Minneapolis, MN: IPUMS.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. Retrieved from http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Patel, M., Rose, K., Owens, C., Bang, H., & Kaufman, J. (2012, Mar). Performance of automated and manual coding systems for occupational data: a case study of historical records. *American Journal of Industrial Medicine*, *55*(3), 228–231. doi: 10.1002/ajim.22005

Pujades Mora, J. M., & Valls, M. (2017). *Barcelona Historical Marriage Database.* IISH Data Collection. Retrieved from `https://hdl.handle.net/10622/SDZPFE` doi: 10622/SDZPFE

Robinson, O., Mathiesen, N. R., Thomsen, A. R., & Revuelta-Eugercios, B. (2022, Jun). *Link-lives guide: version 1.* Online. (Available at: `https://www.rigsarkivet.dk/udforsk/link-lives-data/`)

Safikhani, P., Avetisyan, H., Föste-Eggers, D., & Broneske, D. (2023). Automated occupation coding with hierarchical features: a data-centric approach to classification with pre-trained language models. *Discover Artificial Intelligence*, *3*(1), 6. Retrieved from `https://doi.org/10.1007/s44163-023-00050-y` doi: 10.1007/s44163-023-00050-y

Schierholz, M., & Schonlau, M. (2020, 11). Machine Learning for Occupation Coding—A Comparison Study. *Journal of Survey Statistics and Methodology*, *9*(5), 1013-1034. Retrieved from `https://doi.org/10.1093/jssam/smaa023` doi: 10.1093/jssam/smaa023

Schneider, E., & Gao, P. (2019). *Indefatigable training ship, growth patterns of children 1865-1995.* UK Data Service. Retrieved from `https://reshare.ukdataservice.ac.uk/853251/` doi: 10.5255/UKDA-SN-853251

Soetermeer, H. G. O. (1674). *Familiegeld rijnland, 1674.* `https://zoeken.geheugenvanzoetermeer.nl/detail.php?nav_id=12-1&ref=archiefcategorie&containersoortid=25271613&id=21194854`. (Catalog number: 1179)

SwedPop Team. (2022, 8). Documentation of SwedPop IDS [Computer software manual]. Sweden. Retrieved from `https://swedpop.se/wp-content/uploads/2021/08/Documentation-SwedPop-IDS.pdf`

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). *Attention is all you need.* Retrieved from `https://arxiv.org/abs/1706.03762`

Vedel, C. (2023). *A perfect storm and the natural endowments of trade-enabling infrastructure* (PhD dissertation, University of Southern Denmark. The Faculty of Business and Social Science). (Chapter 1 in "Natural Experiments in Geography and Institutions: Essays in the Economic History of Denmark") doi: 10.21996/jt34-zc23

Vries, J. d. (2008). *The industrious revolution: Consumer behavior and the household economy, 1650 to the present.* Cambridge University Press.

Wrigley, E. A. (2010). The pst system of classifying occupations. *Unpublished paper, Cambridge Group for the History of Population and Social Structure, University of Cambridge*.

Appendix:

Breaking the HISCO Barrier:
Automatic Occupational Standardization with OccCANINE

Christian Møller Dahl, Torben Johansen, Christian Vedel,
University of Southern Denmark

https://github.com/christianvedels/OccCANINE

A Optimal threshold

Table A1 shows the optimal threshold value for Accuracy, F1, Precision and Recall for each separate language. We recommend using these thresholds when using the model without further finetuning.

Table A1: Optimal thresholds for all languages

Language	N. test obs.	Statistic	Value	Optimal thr.
ca	36679	Accuracy	0.9372938	0.38
		F1 score	0.9829256	0.21
		Precision	0.9969465	0.60
		Recall	0.9715096	0.01
da	287338	Accuracy	0.9687058	0.34
		F1 score	0.9881936	0.11
		Precision	0.9879555	0.44
		Recall	0.9952420	0.01
de	1257	Accuracy	0.8385044	0.39
		F1 score	0.9336095	0.17
		Precision	0.9452400	0.59
		Recall	0.9769292	0.01
en	421676	Accuracy	0.9028022	0.47
		F1 score	0.9276790	0.22
		Precision	0.9160129	0.36
		Recall	0.9809083	0.01

Table A1: Optimal thresholds for all languages (continued):

Language	N. test obs.	Statistic	Value	Optimal thr.
es	495	Accuracy	0.8464646	0.28
		F1 score	0.9494947	0.20
		Precision	0.9585859	0.70
		Recall	0.9949495	0.01
fr	16339	Accuracy	0.8185323	0.40
		F1 score	0.9352232	0.21
		Precision	0.9439684	0.60
		Recall	0.9920742	0.01
gr	96	Accuracy	0.8333333	0.45
		F1 score	0.9427083	0.45
		Precision	0.9461806	0.56
		Recall	0.9843750	0.01
is	1177	Accuracy	0.9651657	0.44
		F1 score	0.9738623	0.11
		Precision	0.9694138	0.40
		Recall	0.9906542	0.01
it	264	Accuracy	0.9621212	0.40
		F1 score	0.9848339	0.34
		Precision	0.9943182	0.59
		Recall	0.9943182	0.01
nl	66335	Accuracy	0.9457149	0.37
		F1 score	0.9696100	0.19
		Precision	0.9685221	0.34
		Recall	0.9867114	0.01
no	9065	Accuracy	0.9757308	0.43
		F1 score	0.9869104	0.14
		Precision	0.9847582	0.43
		Recall	0.9961390	0.01
		Accuracy	0.9150508	0.26

Table A1: Optimal thresholds for all languages (continued):

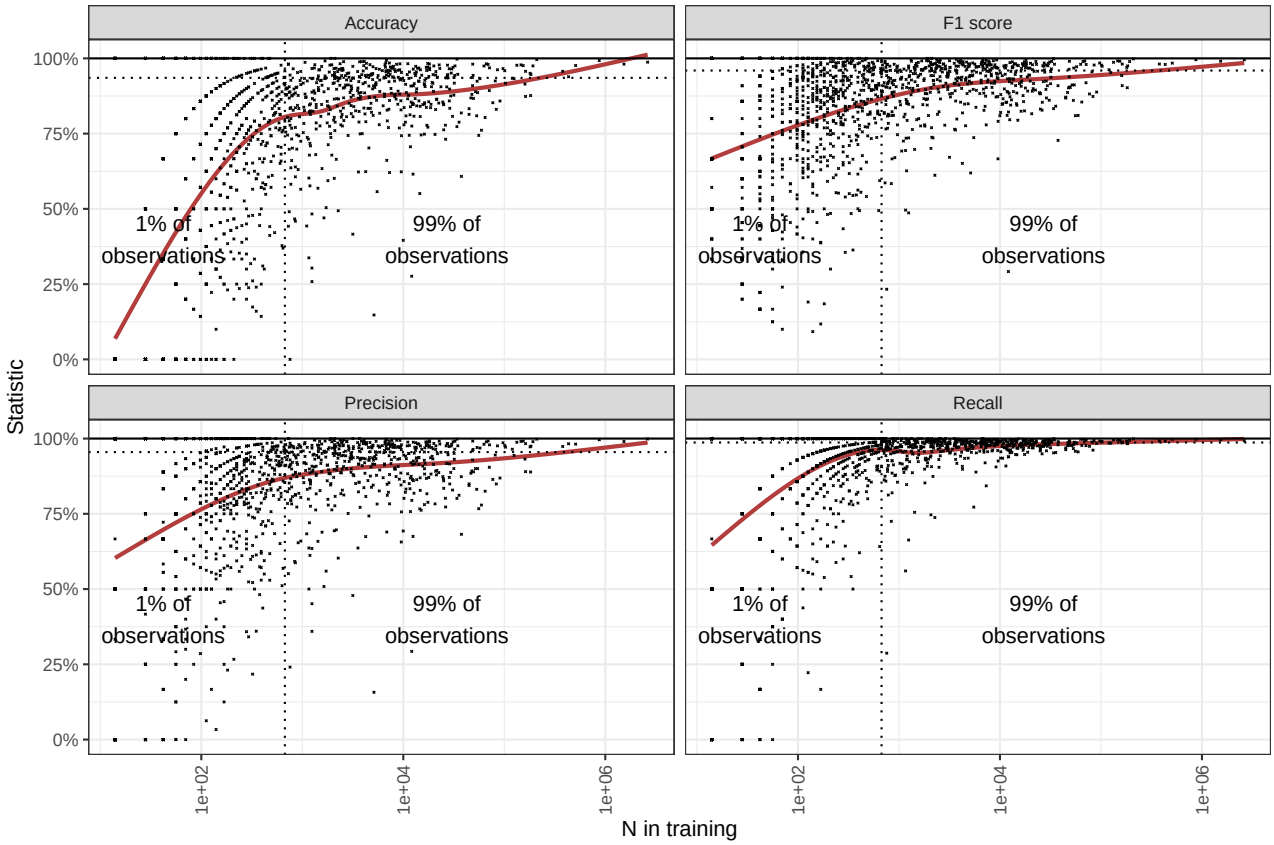
Language	N. test obs.	Statistic	Value	Optimal thr.
pt	1083	F1 score	0.9734292	0.25
		Precision	0.9796861	0.36
		Recall	0.9916898	0.01
se	111817	Accuracy	0.9675184	0.45
		F1 score	0.9837151	0.32
		Precision	0.9859629	0.67
		Recall	0.9868833	0.01
unk	46379	Accuracy	0.9798831	0.44
		F1 score	0.9872493	0.17
		Precision	0.9840768	0.41
		Recall	0.9982104	0.01

Notes: This is a complete table of the optimal threshold according to Accuracy, F1, Precision and Recall. We recommend using these thresholds when using the method without any fine-tuning for any specific language.

B Label frequency and performance

The reliability of our model across different HISCO codes was evaluated to understand its predictive consistency. A plot illustrating this performance, stratified by each HISCO code, is presented in Figure A1. It reveals that while the model performs well across the board, there is a tendency for the error rate to increase for rarer occupations, as is common in any multiclass classification problem. For illustrative purposes, we divide these classes into the 99th percentile and the 1st percentile according to their relative share in the training data. 524 HISCO codes (of a total of 1452 ever observed in a million observations) account for 99 percent of the observations.

Figure A1: Peformance for each HISCO code



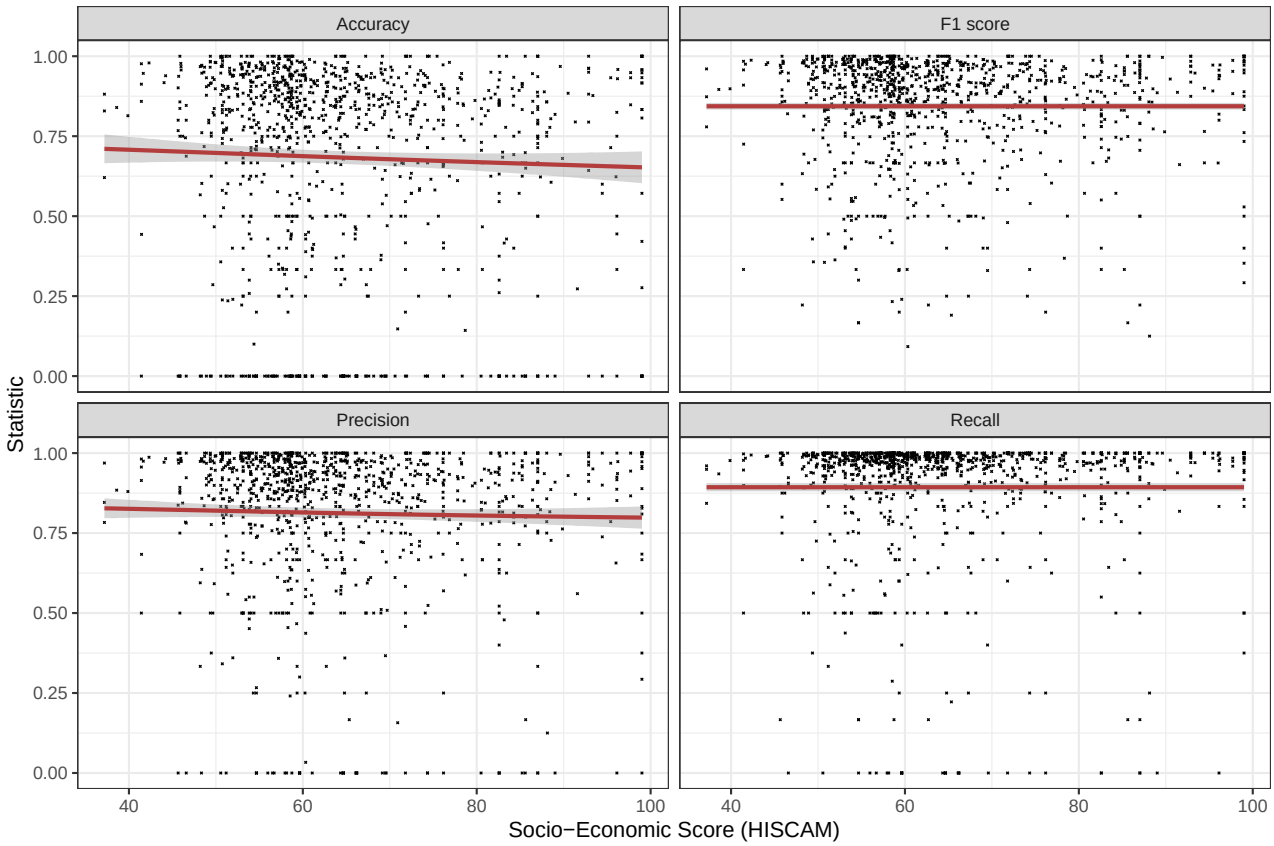
Notes: The model's performance stratified by HISCO codes in terms of Accuracy, Precision, Recall, and F1 score. Each point represents a HISCO code, with the position along the x-axis indicating its frequency in the training data. The red line indicates a smoothed trend across the data points.

The GAM-estimated[24] trend line in Figure A1 indicates that HISCO codes that are underrepresented in the training data (shown by the vertical lines indicating the 1% and 99% cumulative frequency of observations) tend to have lower performance metrics. To account for this variation in performance, users may consider adjusting the classification threshold. A lower threshold might increase the recall of rare occupations at the cost of more false positives. In some cases, the false positives might be less problematic. When visually inspecting them, they tend to be related to the true occupation. Furthermore, the finetune-method provided with OccCANINE can be utilized to improve the model's performance in identifying specific occupations of interest. This can be achieved by finetuning the model with an oversampling of rare occupations.

[24]Generalized Additive Model

C Performance by SES

Figure A2: Model performance and SES



Notes: The scatter plot depicts the relationship between model performance (Accuracy, Precision, Recall, and F1 score) and HISCAM socio-economic scores. Each point corresponds to a HISCO code, with the red line indicating a smooth trend across the data points. No discernible pattern suggests a non-biased performance of the model across different socio-economic strata.

In historical occupational data, the rarity of certain occupations could correlate with the socio-economic status (SES) associated with the occupation, and in turn, the performance of our model might systematically vary with the socioeconomic status of occupations.[25] This potentially introduces systematic bias when using our method in applied settings. To investigate this, we first visualize the relationship between the SES, derived from HISCO codes using the HISCAM score, and the model's performance metrics. This plot, shown in Figure A2, allows us to examine if there is any correlation between SES values and accuracy, precision, recall, or F1 score.

[25]Note that this problem might also affect squishy wet neural networks; also known as human labellers

Table A2: Performance metrics and socio-economic status

	Accuracy	F1 score	Precision	Recall
	(1)	(2)	(3)	(4)
Panel A				
HISCAM score	-0.0010	-0.0003	-0.0003	-3.69×10^{-5}
	[-0.0025; 0.0005]	[-0.0012; 0.0005]	[-0.0012; 0.0005]	[-0.0007; 0.0007]
Panel B				
HISCAM score	0.0008	0.0005	0.0003	0.0005
	[-0.0005; 0.0021]	[-0.0004; 0.0013]	[-0.0006; 0.0011]	[-0.0001; 0.0012]
log(n)	0.0752	0.0328	0.0244	0.0230
	[0.0678; 0.0826]	[0.0282; 0.0374]	[0.0195; 0.0294]	[0.0187; 0.0272]
Observations	1,005	1,005	1,005	1,005

Notes: This table presents the correlation between the socioeconomic score of an HISCO code (HISCAM) and the model performance for that HISCO code. The coefficients of the HISCAM score are small across all specifications and the confidence intervals always include zero. This is also the case when controlling for the logarithm of the number of observations in Panel B. 95 percent confidence intervals in brackets based on heteroskedasticity-robust standard errors.

The trend line in Figure A2 is estimated using GAM, which imposes no linearity, but we end with a remarkably linear relationship for which reason we also find it reasonable to run simple linear regressions to test the relationship. The results from these regressions reveal no significant effect, with small standard errors, suggesting that the model's performance is not systematically correlated with the socio-economic status implied by the occupational codes. We show this in Table A2: Panel A shows the simple regression. It is contestable whether the number of training observations is a confounder or a mediator. In either case, it is included in Panel B. Both panels show qualitatively the same result: There is practically no correlation between HISCAM and the performance of OccCANINE.