Breaking the HISCO Barrier: Automatic Occupational Standardization with Occupational CANINE*

Christian Møller Dahl, University of Southern Denmark

Christian Vedel, University of Southern Denmark

Preliminary draft. Please do not cite or redistribute.

This version: 2024-02-08

Abstract

This paper introduces a new tool, *Occupational CANINE*, 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 14 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.

Keywords: HISCO, occupational structure, labour, history, language model

*We would like to thank Andreas Slot Ravnholt for conscientious RA work in the early stages of this project. This paper has also benefitted from conversations and feedback from Torben Johansen, Simon Wittrock and Paul Sharp. 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 Eocnomic History (Oslo) and the Workshop on Machine Learning in Economic History (Lund). 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 benefitted from improvements suggested by ChatGPT.

1 Introduction

The study of occupational outcomes requires systematic data on peoples 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 'Occupational CANINE' - a transformer language model Clark et al. (2022b) (CANINE) - which we fine-tuned on 14 million observations of occupational descriptions with associated HISCO codes in 14 different languages. This was generously contributed by 22 different research projects, each of which is cited in table 1. The result is an algorithm of 93.5 percent overall accuracy[1] which takes a simple string describing an occupation and outputs the most likely HISCO codes associated with it.

The HISCO system was introduced in an effort to produce internationally comparable occupational data (Leeuwen, 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 results presented here easily generalize to any occupational classification system. With a small amount of training data, our algorithm can be fine-tuned to any classification system.

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.

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 our Occupational CANINE.

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

2 Motivation

2.1 The problem of occupational coding

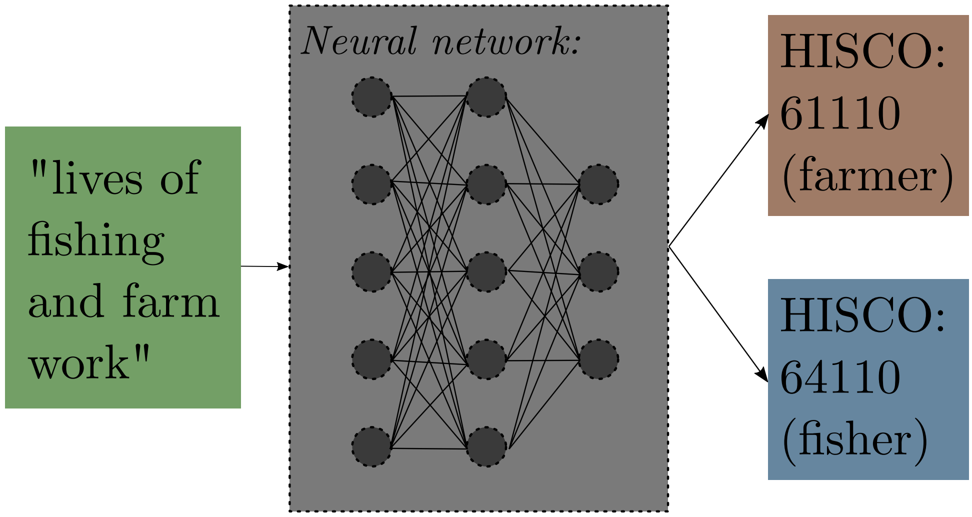
To be able to generate insights into everything from women's empowerment, social mobility, the effect of railways, first nature geography and the origins of the Industrial Revolution to the interplay of technology and development we explicitly or implicitly need to know what people did for a living historically.[2] To the extent that these questions (and many more) are worth answering, it is worth 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 et al. (2013) and its derivations Clark et al. (2022a), 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 classifcal 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 or error-prone. Typically the following steps are involved [4]:

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. *Occupational CANINE* replaces all of that.

Figure 1: Conceptual model



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

2.2 A faster, better and scaleable 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. That means, that we can input whatever occupational description with typos, spelling mistakes, etc. The model then (similarly to chatGPT but smaller) draws on 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. Raw occupational description goes in. HISCO code comes out (see figure 1). We finetune a preexisting language model on 14 million observations. As such we end up with a model, which inherently captures the occupational meaning of inputted strings. 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.

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

[3]The Danish censuses 1787-1901 (Robinson et al., 2022; Clausen, 2015) contain no less than 17,865 unique descriptions corresponding to the occupation 'farm servant'

[4]An example of this is (salgo60, 2021)

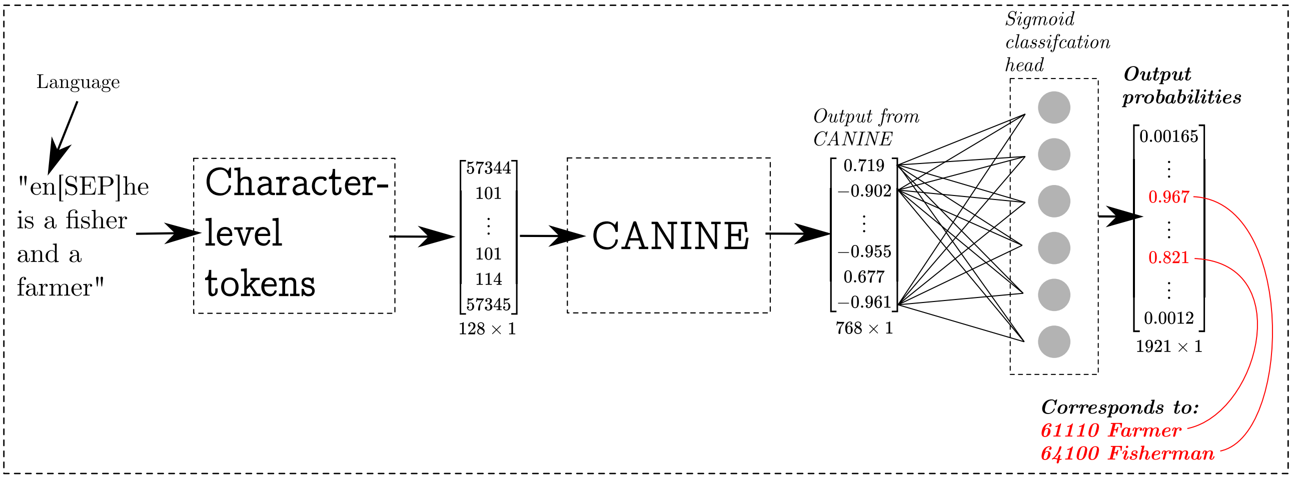
2.3 Literature

The closest related literature is in two strands: Occupational medicine and occupational 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 e.g. the 80 percent of observations for which a label is assigned with the highest probability.

Early approaches only perform well at lower production rates. Patel et al. (2012) demonstrates 89 percent agreement with a human labeller for 71 percent production rate. They mainly rely on rule-based classical NLP-approaches. Gweon et al. (2017) suggests combining classical rule-based approaches combined with bag-of-word cosine distance and nearest neighbours matching. For 'fully automatic labelling'[5] (100 percent production rate) they achieve a 65 percent accuracy in German survey data for the ISCO-88 system.[6] 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.[7] Schierholz and Schonlau (2020) reviews 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 et al., 2018). Garcia et al. (2021), uses classical exact matching, data cleaning and supplements this with TF-IDF and Doc2Vec (based on BERT) for unmatched cases, which is fed through a classical machine learning framework to get a macro F1-Score of 0.65 and a top-5 Per-Digit Macro F1-Score of 0.76 in the Canadian National Occupational Classification Scheme. The most similar paper to the present is Safikhani et al. (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. 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 accuracy of 93.5 percent and an F1-score of 0.960. We do not consider lower levels of 'production rate' because it is barely relevant to this level of performance. Moreover over method requires no string cleaning, correction of spelling mistakes or stopword removal. The model is explicitly trained to handle all of this.

Figure 2: Architecture of Occupational CANINE



Notes: Inputted text is converted to character-level tokens, which then served 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 Architecture, data and training procedure

3.1 Architecture

We make use of the CANINE architecture (Clark et al., 2022b), 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 had a threefold motivation. *First*, 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. Most commonly tokenization happens at the word or wordpiece level. But for archival data, the risk is, that this will cause unnecessary model variance, when e.g. 'farmer' is mistyped as 'frmter'. Such (expected) variation in the data would put strain on traditional methods, but the CANINE architecture is more robust to this.

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

[6] Which is similar to the HISCO system

[7] To achieve e.g. 90 percent accuracy Gweon et al. (2017) requires around 40 percent manual labelling

On top of the CANINE model, we add a classification head of size [1×1921], one output for each of the 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 the following section.

3.2 Data

Table 1: Training data

Shorthand name	Observations	Percent	Language	Source
DK_census	5391656	29.794%	da	Clausen (2015); The Danish National Archives
EN_marr_cert	4046203	22.359%	en	Clark et al. (2022a)
EN_uk_ipums	3026859	16.726%	en	MPS (2020); Office of National Statistics
SE_swedpop	1793557	9.911%	se	SwedPop (2022)
JIW_database*	966793	5.343%	nl	Moor and van Weeren (2021)
EN_ca_ipums	818657	4.524%	unk	MPS (2020); Statistics Canada
CA_bcn*	644484	3.561%	ca	Pujades Mora and Valls (2017)
HISCO_website*	392248	2.168%	mult	HISCO website
HSN_database	184937	1.022%	nl	Mandemakers et al (2020)
NO_ipums	147255	0.814%	no	MPS (2020)
FR_desc*	142778	0.789%	fr	historyofwork.iisg.nl
EN_us_ipums	139595	0.771%	en	MPS (2020); Bureau of the Census
EN_ship_data[1]*	103023	0.569%	en	Schneider & Gao (2019)
EN_parish	73806	0.408%	en	de Pleijt, Nuvolari, Weisdorf (2020)
DK_cedar	46563	0.257%	da	Ford (2023)
SE_cedar*	45581	0.252%	se	Ford (2023)
DK_orsted	36608	0.202%	da	Ford (2023)
EN_oclack*	24530	0.136%	en	O-clack
EN_loc*	23179	0.128%	en	Mooney (2016)
IS_ipums	20459	0.113%	is	MPS (2020)
SE_chalmers	14426	0.08%	se	Ford (2023)
DE_ipums	8482	0.047%	ge	MPS (2020); Statistics Netherlands
IT_fm*	4525	0.025%	it	Fornasin & 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. [1]'EN_ship_data' was made available to us right after we started training the present version of our model. It is used in model evaluation to follow. Data marked with a * is data, where combined occupations were created as described below.

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

This project relies on training data, which was either publicly available or shared with us by researchers. We are grateful for all of this. 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 of all, we replaced all non-English character e.g. 'æ', 'ø', 'å' from occupational descriptions. Secondly, we manually checked the data for thoroughly for peculiarities.[10] Third, we made sure that only standardized HISCO codes were used and removed any observations with non-standard HISCO codes.[11] 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 models chance of picking this up, we artificially combined descriptions within the same source, with that languages word for 'and'. E.g. 'he is a farmer' + 'he fishes' becomes 'he is a farmer *and* he fishes'.[12] 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-traning 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).

3.3 Training

The model was trained for 26 days, 21 hours, 3 m, and 7s on an NVIDIA A100 80GB GPU. It was trained on batches of 256 observations for a total 42 epochs. Every time the model improved in validation accuracy over earlier instances, it was saved. CANINE has 10 per cent 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.[13] This approach (all though 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' with a 25 percent probability. The result of this is a finetuned version of CANINE, which inherently understands something about historical occupations, is multilingual, but also able to utilize language context, is robust to spelling mistakes, and generally performs well, as demonstrated in the next section.

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

[10]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

[11]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 https://github.com/rlzijdeman/o-clack

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

[13]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.

4 Performance

The performance of our model was evaluated in 1 million observations, which were stored separately before training. Using this data, we test performance at classification thresholds ranging from 0.01 to 0.99[14] in terms of Accuracy (exact match), Precision, Recall, and F1 score (harmonic mean of precision and recall). 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 prescion, 98.2 percent recall and an F1-score of 0.960 when optimal classification thresholds are used. Moreover, we manually check two datasets, which are entirely out of distribution and not seen in training: The Copenhagen Burial Records (Robinson et al., 2022), a historical dataset in Danish, and the Indefatigable Training Ship data (Schneider and Gao, 2019), a dataset covering boys admitted to the TS Indefatigable and their parent's occupation. For these tests, we set a fixed classification threshold of 0.5. We manually reviewed 200 random predictions from each dataset to evaluate the accuracy. The results are shown in Table 3. The model achieved 94 percent accuracy on the Copenhagen Burial Records and 95 percent accuracy on the Training Ship Data.[15]

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 for. Appendix A1 presents a detailed table of these statistics for each separate language. Moreover, these results underscore our method's applicability to unseen data, a critical aspect of real-world deployment.

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.[16] We estimate this relationship non-parametrically and demonstrate high performance is maintained for the upper 99 percent of occupational data and more (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 whehter such correlation exists. We find no statistically significant, nor any practically meaningful correlation to generate any worry.

[14]The thresholds 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.

[15]The data from Schneider and Gao (2019) already contained HISCO labels. The labels of our model agreed with these labels in 82 percent of cases. Regarding the cases of disagreement, only 3.5 percent of the observations had substantially different occupations leaving the remainder 96.5 percent of the observations with reasonable HISCO codes. In future version of this paper we will extend with more OOD tests.

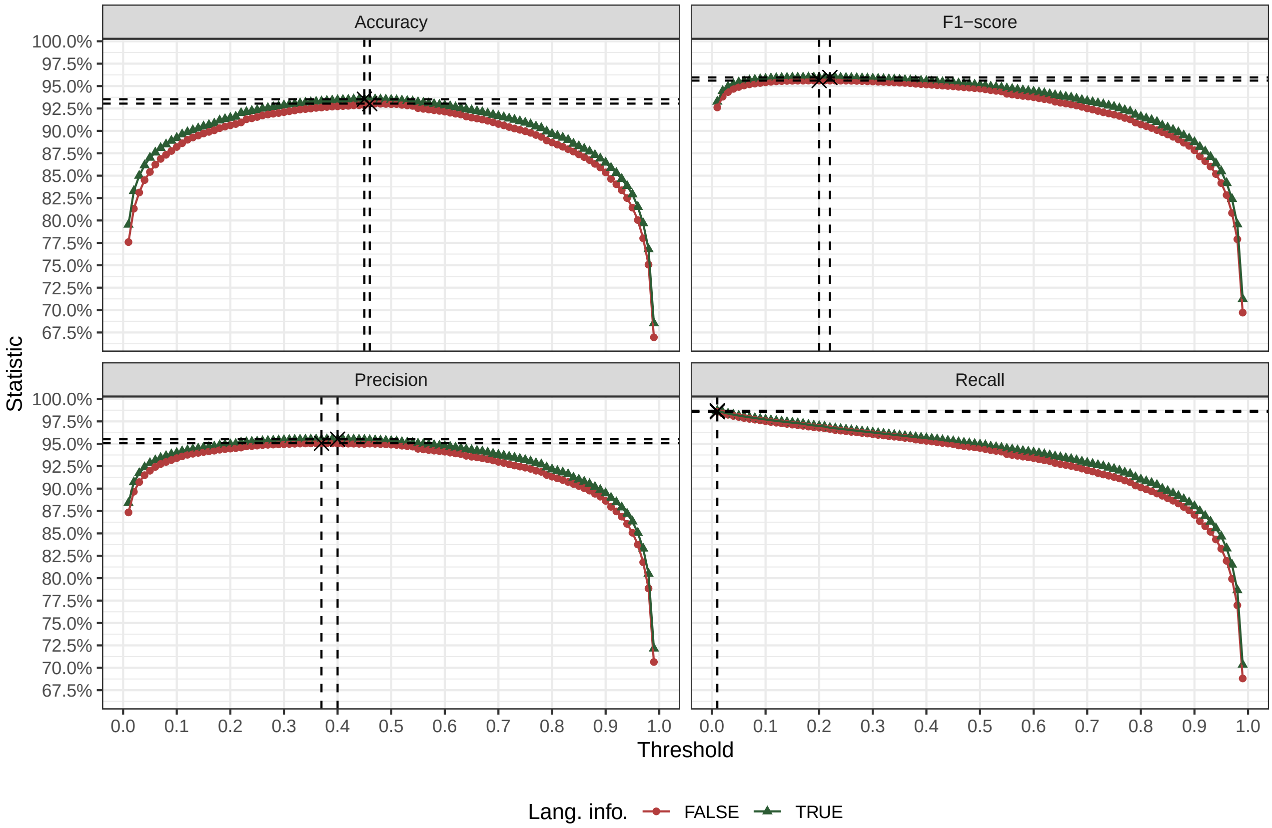
[16]Which is typically also the case for a skilled human labeller.

Table 2: Best overall performance

Statistic	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 indicate the point where each metric is maximized. Language context availability consistently improves metric scores. But only slightly. A similar table for each language is available in Appendix A1.

Figure 3: Optimal Threshold



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

Table 3: Out of Distribution Testing Accuracy

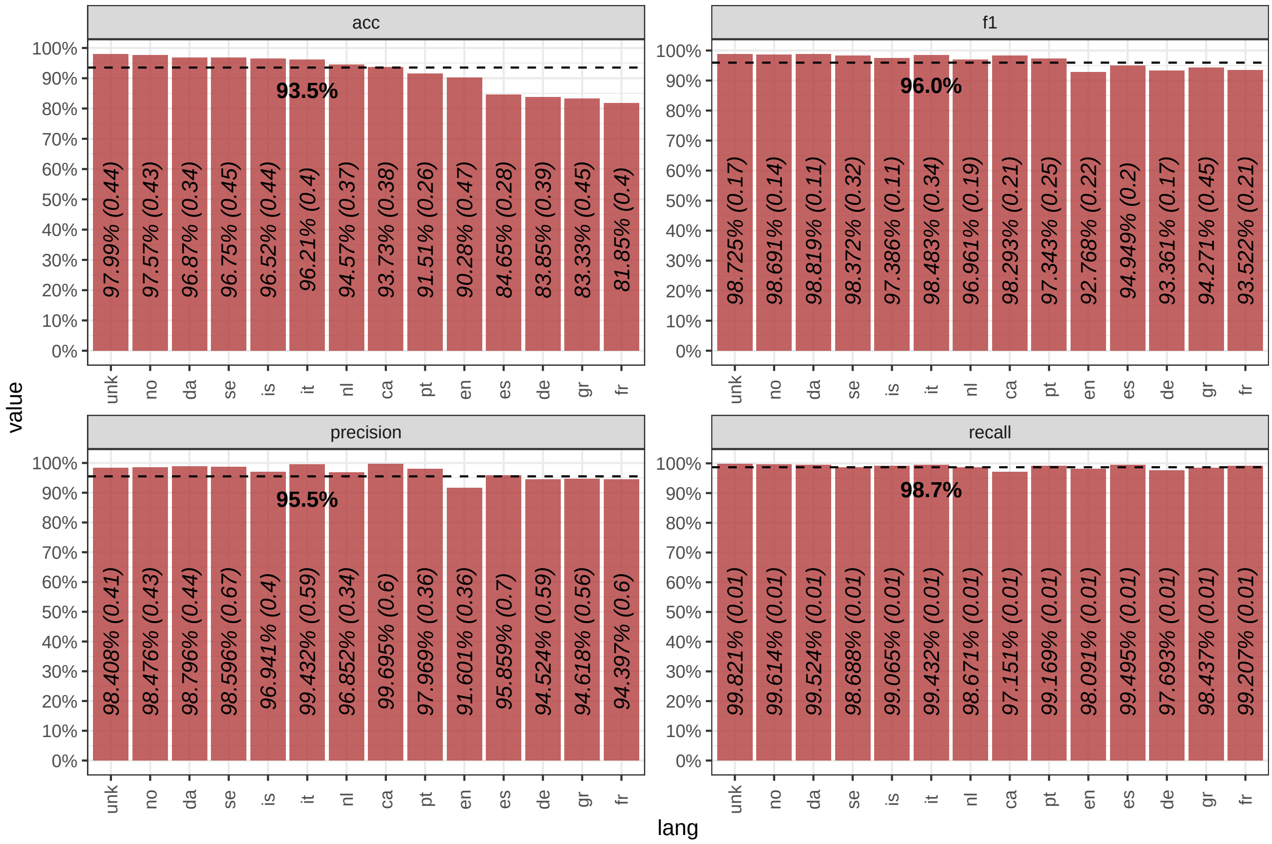
Dataset	Checked observations	Accuracy
Copenhagen Burial Records	200	0.940
Training Ship Data	200	0.952

Notes: This table reports accuracy metrics for out-of-distribution tests, emphasizing the model's adaptability. The Copenhagen Burial Records and Training Ship Data, each with 200 randomly selected observations, were used to benchmark performance. A classification threshold of 0.5 was uniformly applied. These tests simulate the model's deployment in varied historical and linguistic contexts, demonstrating its utility in practical settings.

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.[17] This 768 dimensional output is ought to represent the meaning of each occupational description. If similar descriptions are closer together - especially across languages, this is evidence, that the model has picked up a structural understanding of occupations. Figure 5 shows a low-dimensional representation of these embeddings. Panel A shows results from the original CANINE (Clark et al., 2022b) finetuned on Wikipedia. Panel B shows results from our version of CANINE, 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 Occupational CANNINE picks up similar occupations as being semantically similar. This result shows the potential for generalised high performance across different domains; Our finetuned version of CANINE is a valuable starting point for other applications related to occupational descriptions in a historical setting.[18]

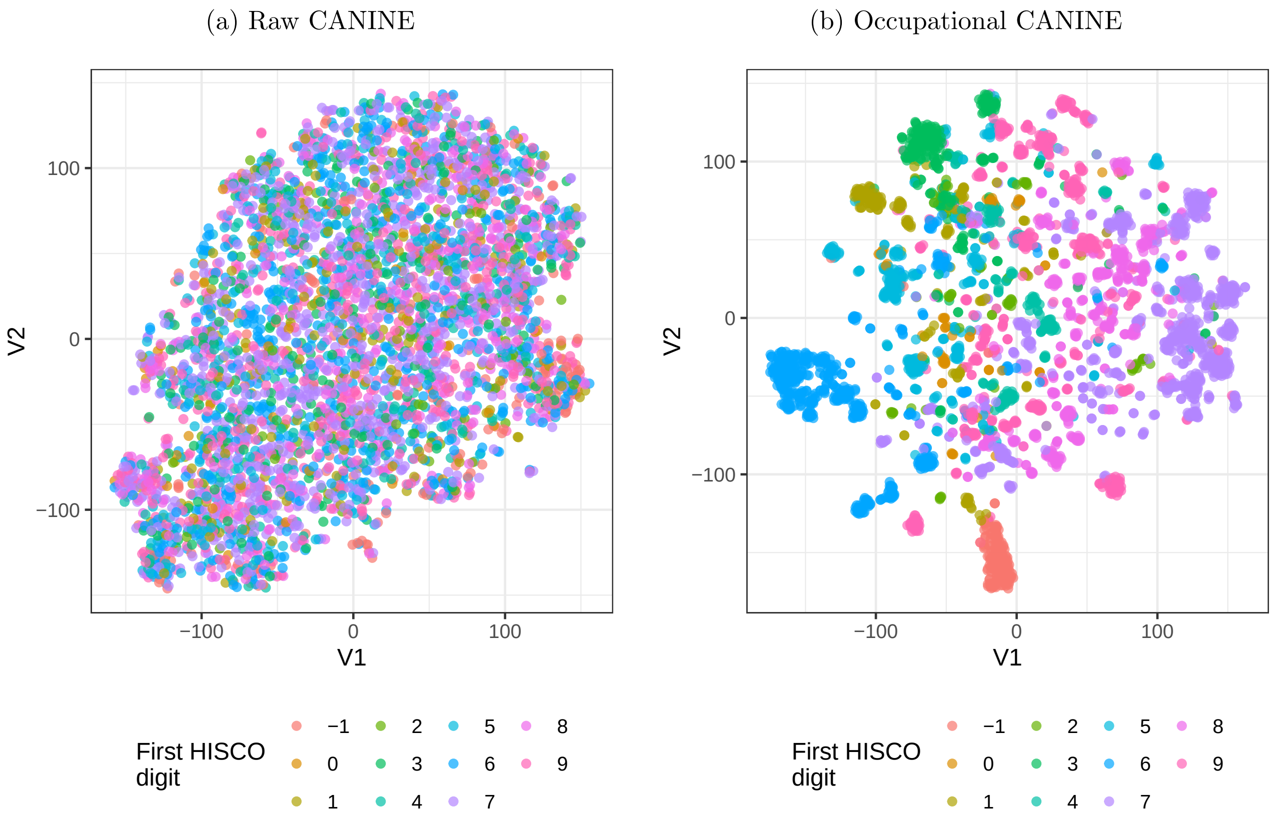
[17]In practice, the output from the final layer before the Sigmoid classification head.
[18]And will be available on HuggingFace

Figure 4: Performance by Lanugage



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 parenthesis.

Figure 5: T-SNE Visualizations of Occupational Embeddings



Notes: Panel (a) illustrates the t-SNE visualization for embeddings derived from the original CANINE model, finetuned on Wikipedia data. Panel (b), on the other hand, shows embeddings from our version of CANINE, further finetuned on historical occupational data. The colours correlate with the first digit of the HISCO code, roughly indicative of economic sectors. The clustering of similar colours in Panel B, as opposed to Panel A, indicates that our finetuned model captures a meaningful structure of occupational descriptions across languages. This demonstrates the enhanced ability of our model to understand and categorize occupations. The data depicted was not seen in model training.

5 Conclusion and Recommendations

Our comprehensive evaluation demonstrates that the 'Occupational CANINE' model is a powerful tool for automatically transforming occupational descriptions into standardized HISCO codes. We recommend a classification threshold of 0.22 to optimize the F1 score and a threshold of 0.45 to maximize accuracy, based on the model's performance with 1 million validation observations. 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 'Occupational CANINE' is adept at delivering high-quality HISCO codes in seconds or minutes. 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, the provided .finetune method facilitates further optimization. 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 key observations in the study. 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 coding from scratch.

In conclusion, 'Occupational CANINE' represents a significant stride in historical occupational data processing, effectively breaking the HISCO barrier that 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*, page 25–56. New Approaches to Economic and Social History. Cambridge University Press.

Berger, T. (2019). Railroads and rural industrialization: evidence from a historical policy experiment. *Explorations in Economic History*, 74:101277.

Clark, G. (2023). The inheritance of social status: England, 1600 to 2022. *Proceedings of the National Academy of Sciences*, 120(27):e2300926120.

Clark, G., Cummins, N., and Curtis, M. (2022a). 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., and Wieting, J. (2022b). ¡scp¿canine¡/scp¿: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Clausen, N. F. (2015). The Danish Demographic Database—Principles and Methods for Cleaning and Standardisation of Data. *Population Reconstruction*, pages 1–22.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

Garcia, C. A. S., Adisesh, A., and 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.

Goldin, C. (2006). The quiet revolution that transformed women's employment, education, and family. *American Economic Review*, 96(2):1–21.

Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M., and Steiner, S. (2017). Three methods for occupation coding based on statistical learning. *Journal of Official Statistics*, 33(1):101–122.

Lambert, P. S., Zijdeman, R. L., Leeuwen, M. H. D. V., Maas, I., and 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.

Lampe, M. and Sharp, P. (2018). *A Land of Milk and Butter*. University of Chicago Press Economics Books. University of Chicago Press.

Leeuwen, M. H. v. (2002). *HISCO: Historical International Standard Classification of Occupations*. Leuven U.P.

Mokyr, J. (2016). *A Culture of Growth: The Origins of the Modern Economy*. Princeton University Press.

Moor, T. D. and van Weeren, R. (2021). Dataset Ja, ik wil - Amsterdam marriage banns registers 1580-1810.

Morris, J. X., Lifland, E., Yoo, J. Y., Grigsby, J., Jin, D., and Qi, Y. (2020). Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp.

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

Patel, M., Rose, K., Owens, C., Bang, H., and Kaufman, J. (2012). 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.

Pujades Mora, J. M. and Valls, M. (2017). Barcelona Historical Marriage Database.

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

Safikhani, P., Avetisyan, H., Föste-Eggers, D., and 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.

salgo60 (2021). Hiscokoder: Historiska yrken. https://github.com/salgo60/HISCOKoder. Accessed: [2024-01-04].

Schierholz, M. and Schonlau, M. (2020). Machine Learning for Occupation Coding—A Comparison Study. *Journal of Survey Statistics and Methodology*, 9(5):1013–1034.

Schneider, E. and Gao, P. (2019). Indefatigable training ship, growth patterns of children 1865-1995.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.

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".

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 Occupational CANINE

Christian Møller Dahl, Christian Vedel,
University of Southern Denmark

Contents

A Optimal threshold

Table A1 shows the optimal threshold value for Accuracy, F1, Precision and Recall for each sepperate language. We recommend using these, 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
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
		Accuracy	0.8333333	0.45

Table A1: Your Caption Here Continued:

Language	N. test obs.	Statistic	Value	Optimal thr.
gr	96	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
pt	1083	Accuracy	0.9150508	0.26
		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
		Accuracy	0.9798831	0.44

Table A1: Your Caption Here Continued:

Language	N. test obs.	Statistic	Value	Optimal thr.
unk	46379	F1 score	0.9872493	0.17
		Precision	0.9840768	0.41
		Recall	0.9982104	0.01

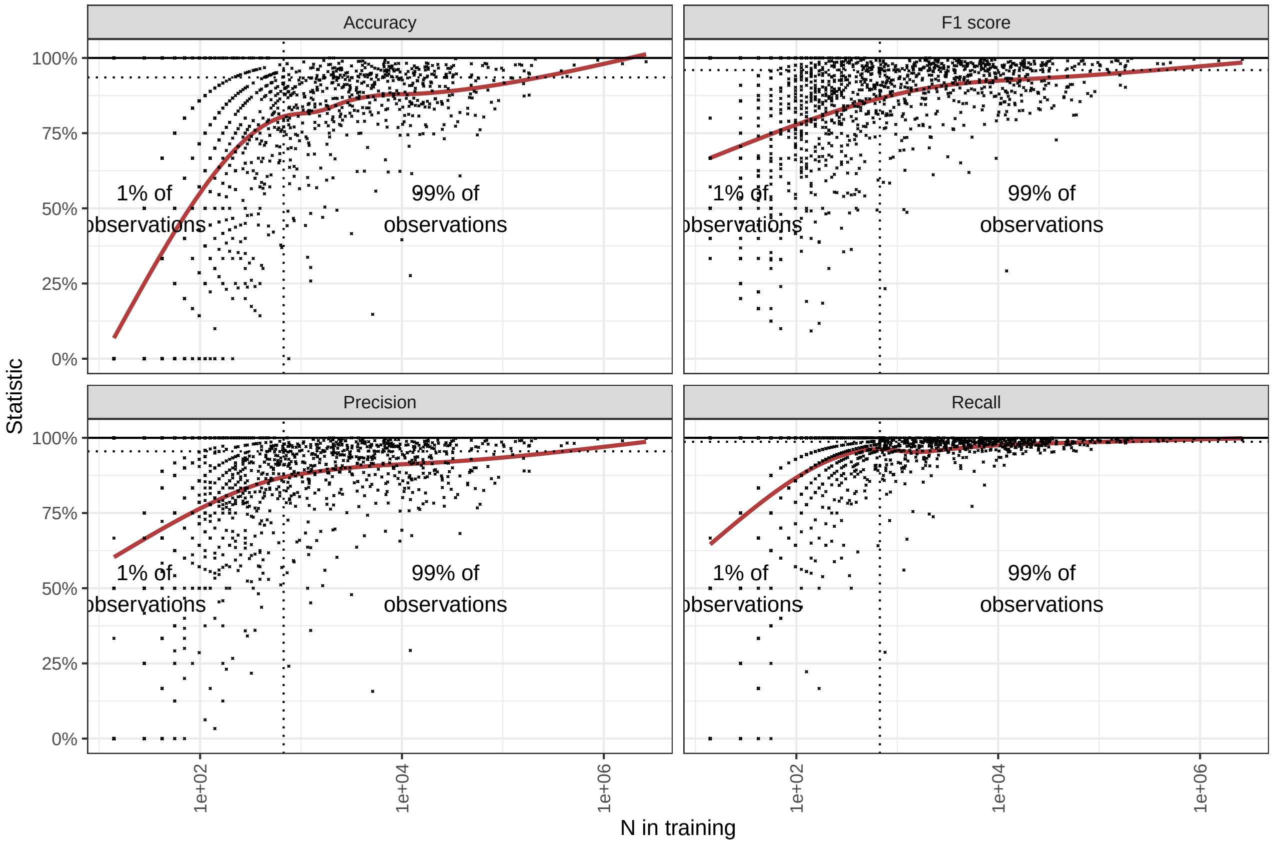
Notes: This is a complete table of the optimal threshold according to Accuracy, F1, Precision and Recall. We recommend using these values when using the method without any fine-tuning for any specific language.

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.

The GAM-estimated 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 variance in performance, users may consider adjusting the classification threshold. A lower threshold might increase the recall of rare occupation 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, using the .finetune method provided in the model code can be applied to enhance the model's ability to capture specific occupations of interest.

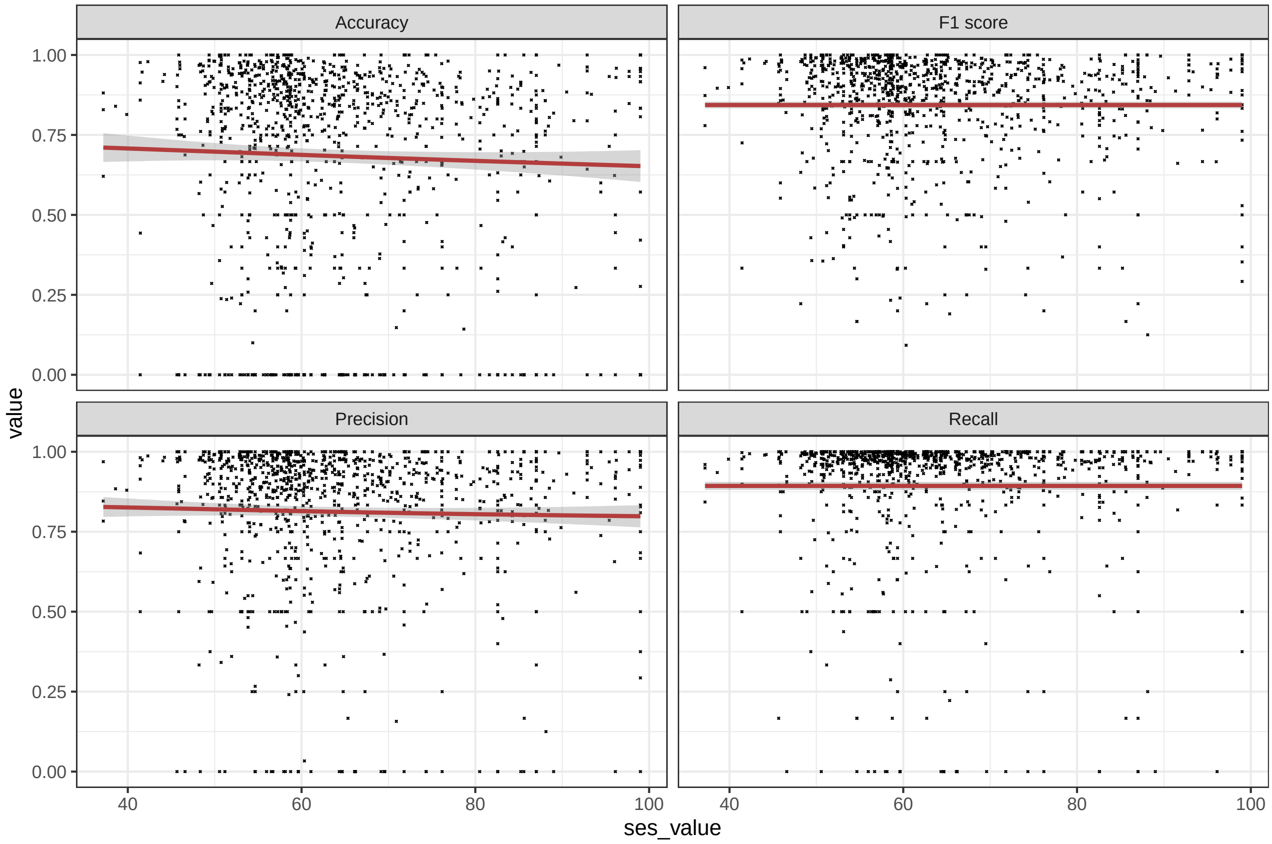
Figure A1: Peformance by each HISCO code



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.

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 their socio-economic status (SES), and in turn, the rarity could drive low accuracy of our method.[1] 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.

[1]Note that this problem might also affect squishy wet neural networks as well - otherwise know research assistants

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: 95% confidence intervals in brackets based on heteroskedasticity-robust covariance matrix. This table presents the correlation between HISCAM socio-economic score of a specific HISCO code performance metrics for that HISCO code. The coefficients of the HISCAM score are small across all specifications and the confidence intervals always includes zero. Even when controlling for the logarithm of the number of observations in Panel B.

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 indeed not systematically correlated the socio-economic status implied by the occupational codes. We show this in Table A2. The first panel 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 no correlation between SES and the performance of Occupational CANINE.

References