# Pilot Study Report

Organisation: Statistics Canada

Author(s): Isaac Ross and Justin J. Evans

Date: 2020/05/28

Version: 2.0

## 1. Background and why and how this study was initiated

Statistics Canada's Generalized Coding tool (G-Code) is currently used for most of the agencies' automated coding activities. Operation and Integration Division's Business Process and Technical Analysis Section (BPTAS) has been using G-Code to develop WordMatching solutions. In 2019 we have successfully auto-coded +120K records, while maintaining an error rate at or below that of human coders (<5%). However, WordMatching solutions, which can be composed of hundreds of thousands of reference text entries, require regular maintenance to remain up to date. Recent G-Code development has enabled the integration of machine-learning (ML) algorithms, such as FastText and XG-Boost, for the use of coding in production. The BPTAS team has since shifted to focus on developing models to code the North American Industry Classification System (NAICS) and National Occupational Classification (NOC) using the FastText algorithm. However, WordMatching will continue to be used to code classifications / surveys that do not warrant the expense of developing a machine learning model. For example, simple classifications such as language or country, or small surveys that do not have enough historical data to create a viable model, and do not code enough records to counter to upfront investment or expense of developing a viable model.

The majority of surveys at Statistics Canada are coded manually. Statistics Canada is therefore supporting the operationalization of WordMatching or ML model to realize coding efficiencies. However, implementation of an ML solution would introduce additional processes including feature selection, feature engineering, hyper-parameter tuning, training, and quality analysis that do not form part of the existing business process. As a result, Operation and Integration Division's position is that any cost savings realized by the BPTAS team will be re-invested into the research, development, and implementation of machine learning initiatives.

Recently, the Canadian Community Household Survey (CCHS) sought to implement a ML solution to code NAICS and NOC in their most recent quarters (2019 Q3, Q4). As the Coding and Corrections Environment (CCE) cannot currently consume results from G-Code ML models, we developed a temporary production pipeline to code CCHS. As a result, a number of manual processes – file conversions, file merges, and file loads – were implemented (Appendix, Figure X). This lack of integration also impacts the quality control (QC) sampling of auto-coded records, which will be expanded upon later. However, the upcoming CCE release will enable seamless integration of ML solutions.

# 2. Data

## 2.1 Input Data

Table 1 – Model training data sources.

| Survey | Year | Record # |
|---|---|---|
| Labour Force Survey (LFS) | 2016-2018 | 425,958 |
| Canadian Community Household Survey (CCHS) | 2019 | 88,782 |
| NOC Index Entries | 2016 | 114,161 |
| NAICS Index Entries | 2017 | 38,256 |

In order to create models to code NAICS and NOC we initially used CCHS historical records as training data. However, further analysis showed that the addition of Labour Force Survey (LFS) historical records and Index Entries provided by Standards Division (Table 1) improved our ability to code CCHS (1). Factors that were considered include: similarity of survey question, whether pseudocodes were used during classification, and data quality (ex. estimated error rate). We are currently testing the addition of records from Job Vacancy and Wage Survey (2016-2019), Canadian Health Survey on Children and Youth (CHSCY), and CENSUS (2016). We intend to include the results of that testing in the next report.

## 2.2 Data Preparation

In order to create NAICS and NOC models several techniques were explored. The BPTAS team tested multiple bag of words (concatenation, reference tagging, Caesar Cipher), stemming, lemmatization, up-sampling of minority classes, utilizing pre-trained FastText embeddings, as well as separate French and English models. The combination of an expanded training dataset and the following preprocessing techniques resulted in the best model performance:

1. Removal of Stop Words
2. Lowercasing character conversion
3. Merging of the variables “Business Name” and “Name of Employer” to create “Company”
4. Application of a Caesar Cipher to differentiate text inputs from the variables “Company”, “Industry”, “Job Title” and “Job Description”
5. Addition of Classification Index Entries (40-80K records) & Labour Force Survey (440K records) to CCHS’s training datasets (89K records)

This process has been previously documented (1). The script is available through Github (https://github.com/UNECE/CodingandClassification_Statcan).

## 2.3 Feature Selection

We selected four main text variables that relate to respondent economic participation and are common through a number of different surveys at Statistics Canada. These variables are “Company”, “Industry”, “Job Title” and “Job Description”. These variables were selected because, in addition to being part of numerous surveys, they are the most significant indicators when attempting to classify different occupations (NOC) and industries (NAICS).

## 2.4 Output data

The intent of our pilot to date has been to use machine learning to code NAICS and NOC for the CCHS, which we have now successfully done for the last two quarters of 2019 (Figure 1). After implementing our production pipeline we were able to automate the coding for 12.6% (Q3) and 13.3% (Q4) of records. Records where our NAICS and NOC models predicted a code below a confidence threshold, were manually coded through the same 'Interactive' process in the CCE.

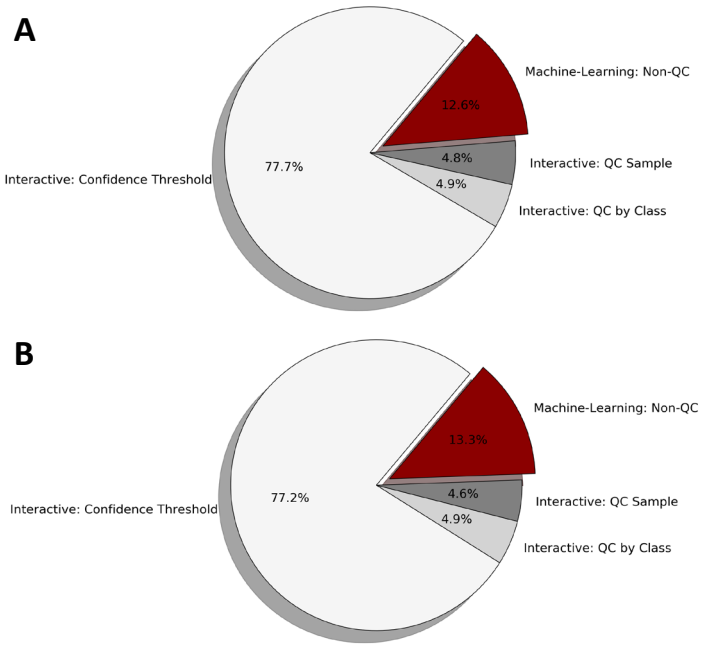


**Figure 1.** Production pipeline using CCHS data (2019 Q3, Q4). The percent of records in each stage of the pipeline. (A) In CCHS Q3 records which did not obtain a high enough confidence score were sent to be interactively coded (77.7%) (Interactive: Confidence Threshold). A per class quality validation (4.9%) was applied to remove problematic classes and these records were sent to be interactively coded (Interactive: QC by Class). The remaining records were then split into a quality control sample (4.8%) (Interactive: QC by Class), to be verified by expert coders, and the remaining records sent to output without manual verification (12.6%) (Machine-Learning: Non-QC). Q3 Records = 7430. (B) The same process was applied for CCHS Q4. Q4 Records = 7404.

# 3. Machine Learning Solution

## 3.1 Models tried

To standardize processes and ensure version control across coding projects, Statistics Canada requires models implemented in production to be run through G-Code. Our mandate was to attempt to use existing algorithms integrated in G-Code (FastText, XGBoost) in order to create production quality models and implement them into our existing business and system processes. As such, no other machine learning solutions were investigated by our team. Analysis of the viability of other technologies is being conducted by the Data Science Division team at Statistics Canada. Since we began our efforts with FastText, XG Boost has also been integrated into G-Code, and discussions are currently underway regarding which additional technologies to pursue. Currently the next likely candidates appear to be Tensor Flow, PyTorch, and Support Vector Machines (SVMs).

## 3.2 Model(s) finally selected and the criterion

**Table 2.** A comparison of NOC models based on the composition of their training data. Each model was then tested on 157,527 records from multiple surveys which share similar write-ins. Overall Accuracy, F1, Precision, and Recall were calculated on the entire testing dataset.

| Measure | CCHS | CCHS + JVWS + Index Entries |
|---|---|---|
| Record # | 157,527 | 157,527 |
| Overall Accuracy (%) | 57.9 | 64.4 |
| Weighted Average F1-Score | 40.7 | 51.5 |
| Weighted Average Precision | 43.7 | 53.3 |
| Weighted Average Recall | 40.7 | 51.8 |

Models were compared against each other based on their overall accuracy and weighted F1 scores, precision, and recall. As seen in the table above, and described in Section 2.1 of this report, the addition of historical data from Job Vacancy and Wage Survey (JVWS) and Index entries improved all NOC model performance metrics (Table 2). Despite low overall metrics, the application of a confidence threshold and quality control plan allowed us to implement the model in production while maintaining an error rate below 5%. The process to obtain these results is available in our production report (2).

## 3.3 Hardware used

The following hardware is being used to develop ML solutions.

- Processor: Intel Core i5-3570
- CPU: 3.0GHz
- RAM: 16.0 GB

## 3.4 Runtime to train the model

The runtime, based on the hardware specifications provided, to create and save a FastText model with 88,782 records is 27.6 minutes.

# 4. Results

**Table 3.** Error rate of NAICS and NOC in CCHS production pipeline. The row 'Both' indicates when a record had either a NAICS or NOC error, or both a NAICS and NOC error. (A) As the error rate of our QC sample was below our 5% in Q3 and (B) Q4, we were able to return a QC'd portion. (*) Indicates the error rate of manual coders in production before verification is applied.

A

| Classification | Interactive: Confidence Threshold | Interactive: QC by Class | Interactive: QC Sample |
|---|---|---|---|
| NAICS 2017 | 3.0* | 5.5 | 2.2 |
| NOC 2017 | 4.7* | 6.3 | 2.5 |
| Both | 3.9* | 10.6 | 4.2 |

B

| Classification | Interactive: Confidence Threshold | Interactive: QC by Class | Interactive: QC Sample |
|---|---|---|---|
| NAICS 2017 | 1.1* | 3.3 | 0.0 |
| NOC 2017 | 1.6* | 5.5 | 1.8 |
| Both | 1.3* | 7.1 | 1.8 |

**Table 4.** NAICS and NOC model metrics for the CCHS production pipeline. Overall Accuracy, F1, Precision, and Recall were calculated on the 'Interactive: QC Sample'. (A) Q3 Record number = 343. (B) Q4 Record number = 358.

A

| Measure | NAICS 2017 | NOC 2016 |
|---|---|---|
| Error Rate (%) | 2.2 | 2.5 |
| Weighted Average F1-Score | 97.5 | 96.9 |
| Weighted Average Precision | 97.4 | 96.7 |
| Weighted Average Recall | 97.8 | 97.5 |

B

| Metric | NAICS 2017 | NOC 2016 |
|---|---|---|
| Error Rate (%) | 0 | 1.8 |
| Weighted Average F1-Score | 100 | 98.5 |
| Weighted Average Precision | 100 | 98.5 |
| Weighted Average Recall | 100 | 98.8 |

In order to evaluate whether the models implemented in production were accurate we selected a quality control (QC) sample from the ML coded records. These records were then manually coded by expert coders to determine an error rate for NAICS and NOC. This coding activity was done fully blind of the original ML assigned code to determine an unbiased error rate. In both Q3 and Q4 of CCHS the error rate of our QC sample was below our target of 5% (Table 3). The following standard model metrics (F1-score, Precision, Recall) were also run on the QC Sample (Table 4).

# 5. Code/programming language

In order to provide an example of the code used to create and test our FastText models we used Statistic Poland's EOICOP open-source data. The scripts are available through Github (https://github.com/UNECE/CodingandClassification_Statcan), and follow a three step process:

1. UNECE_Fasttext_Step1_Format_clean_splitdata
2. UNECE_Fasttext_Step2_HyperparameterTune_Model
3. UNECE_Fasttext_Step3_TrainModel
4. UNECE_Fasttext_Step4_TestModelMetrics

However, it is important to note that implementation of models was required to be run through Statistics Canada's approved coding platform G-Code. Therefore, the scripts used in production are not the same as those provided for the example.

## 6. Evolution of this study inside the organisation

In the course of our efforts to develop high performing models for coding NAICS and NOC we have been engaging with a number of other partners within Statistics Canada, such as the Data Science Accelerator (now Data Science Division), the ML Community of Practice at Statistics Canada, the methodology team assigned to G-Code, the G-Code IT team, our subject matter area partners in the Canadian Community Household Survey, Census, and Labour Statistics Division. However, most of our engagement with such partners has only been consultative, and the majority of the efforts put in to develop our models has been our own. That being said, while we have been leading the efforts to develop models for NAICS and NOC, we have been sharing our results in a series of presentations and long form reports with all of our internal and external partners, which have generally been well received and has solicited a significant amount of interest. Not only has our progress been of interest to partners on their own journey towards production, but many of the challenges that we have faced have prompted other units to become more engaged with certain facets of the business process, such as engaging more closely with methodology about a QC strategy to address sampling and the exclusion of certain classes that have an elevated predicted error rate.

## 7. Is it a proof of concept or is it already used in production?

We have successfully moved our models for NAICS and NOC into production for two different surveys, the CCHS and the Canadian Health Measures Survey (CHMS). This was facilitated by the development and implementation of our QC strategy to mitigate problematic/minority classes with higher error rates, and the extensive consultations with clients to get their buy-in. Both of these factors were not guaranteed from the onset of our project. In order to overcome these hurdles we consulted extensively with methodology in order to develop a strategy to mitigate problematic classes and pursued open communication with our clients and a measured implementation in order to assuage client reservations. In the end we did obtain the approval from our clients by performing a number of parallel runs and comparison analyses to clearly demonstrate the quality of the outputs that our models are able to achieve.

While we were able to overcome these challenges, there remain a number of barriers to further advancing our efforts to ramp up the adoption of machine learning solutions to other surveys. We continue to be limited by minimal staff and IT resources, and we need to enhance some of our systems, such as G-Code and the CCE, in order to integrate additional technologies and to facilitate more nuanced quality sampling. We would also like to enhance our analysis of model performance to more clearly assess variance in final estimates, commonly associated with standard statistical processes. In order to address these challenges we will be pursuing a number of different solutions.

We have requested more powerful IT hardware, and will be looking in to leveraging some of the more powerful hardware that has been acquired by our internal partners in Data Science Division. We will be creating a business case for expanding staff resources in order to obtain management support for expanding our team and therefore our ability to accelerate our implementation progress. Work to prioritize the integration of new technologies into our G-Code platform is currently underway and work will be undertaken to enhance G-Code in the coming fiscal year. The new version of the CCE is slated to be released into Production in June which will integrate our models more seamlessly into our systems and business processes, and further enhancements to the CCE to facilitate improved quality sampling will be done in the new fiscal year. Finally, we will continue to work with our partners in methodology to enhance our capacity to analyse and assess model performance going forward.

## 7.1 What is now doable which was not doable before?

By introducing machine learning models into our business processes we have been able to achieve cost savings for two different surveys, and have reduced the time it takes to process records for those surveys. This has allowed resources that can be re-allocated to other initiatives, and increased the speed at which our clients are able to complete the total activities for their surveys. We have demonstrated the potential cost savings if ML were to be implemented for some of the surveys that process a greater volume of records, such as JVWS and LFS. We have also demonstrated that ML solutions can be as accurate, or more accurate, than human coders. In addition to realized efficiencies, our integration of machine learning technologies into our systems has opened the door to future clients to be able to take advantage of these technologies for other surveys.

## 7.2 Is there already a roadmap/service journey available how to implement this?

There is not a roadmap/service journey documented per say, but this is something that we are looking in to formalizing. Our unit had the good fortune of starting our journey from a point where we already had a number of systems and business processes in place for both manual and automated coding. We have implemented a quality control strategy that should ensure that the models perform as anticipated, and that we can monitor performance over time.

There remains a number of outstanding key elements that we need to address, such as the development of an integrated QC solution into our systems that is tailored to the specific needs of machine learning solutions, the enhancement of our quality validation strategy to include analysis of a coefficient of variation, and the enhancement of our systems to be able to use additional machine learning technologies. Statistics Canada is also in the process of developing a quality validation framework that we will be adopting, that includes a checklist to ensure that policies and directives have been complied with, that proper documentation has been completed, that quality has been properly evaluated and will be monitored, that a peer review has been completed, etc. Needless to say the adoption of this framework will drastically impact our business processes, but we believe that it will be important in order to ensure that confidentiality, quality, and explainability are properly addressed.

## 7.3 Who are the stakeholders?

The stakeholders that are implicated in our pilot study are:

-Business Process and Technical Analysis Section – Operations and Integration Division (our unit)
-G-Code Methodology Team – Statistical Integration Methods Division
-Data Science Accelerator – Data Science Division
-Canadian Community Household Survey – Centre for Population Health Data
-Canadian Health Measures Survey – Centre for Population Health Data
-CCE Dev Team – Centre for Social Data Integration and Development
-G-Code Dev Team – IT Solution Lifecycle Management

Our stakeholders have been involved in consultations regarding modelling techniques (Data Science Accelerator), quality control and validation (G-Code Methodology Team), integrated systems development (CCE and G-Code Dev Teams), and implementation into production (Canadian Community Household Survey and Canadian Health Measures Survey).

## 7.4 Robustness

Ensuring ML meets the quality standards necessary for production is crucial. In the production pipeline that was implemented for CCHS we took a quality control sample to validate that the models were predicting at an accuracy comparable to human coders (<5.0%). If this QC sample exceeds our acceptable outgoing quality control level then certain ML predicted records would be sent for manual coding. Continual assessment of model performance in production is necessary in order to assess drift and indicate if retraining the model is necessary.

## 7.5 Fall Back

The potential fall back plan if our machine learning solution were to fail would be to code all of the records manually, perhaps with small amount of automation using WordMatching. Resources are in place to continue that manual work in the short term, however we are operating in an ecosystem of tight budgets and increased demand for the creation of statistics in a timely manner in order to stay relevant in a time when private companies are able to release data much faster than we are. Therefore, while we are able to survive without machine learning solutions in place, the ramifications of such a failure would be the unrealized operational and cost efficiencies that could otherwise be obtained.

# 8. Conclusions and lessons learned

In the course of developing NAICS and NOC models for CCHS we have learned a great deal. First and foremost we have learned what it takes to build high performing models in terms of pre-processing, feature selection, model balancing, etc. In addition we have acquired a greater understanding of the investment in time required to develop a functional machine learning solution and to transition it into production, the number of records required for a model to be viable for coding a large classification, and how to approach the problem of minority classes.

Furthermore, we have developed a great appreciation for the interdependencies required to transition into production. We had to perform extensive consultations with Methodology in order to determine an appropriate quality control strategy. We also had to dedicate time and resources in order for our suite of coding systems to be enhanced in order to move our models into a production environment, such as the integration of FastText into G-Code and the enhancement of the CCE to be able to consume machine learning results from G-Code. Some of these challenges remain, but we will be continuing to work in the coming months to make our quality control and validation even more robust, to incorporate additional technologies into G-Code, and to enhance the CCE to perform more nuanced quality sampling.

## 9. Potential organisation risk if ML solution not implemented

The potential risk if our machine learning solutions are not implemented for text classification activities are three-fold. In the absence of improved automated solutions the organization would continue to devote funding to manual coding processes. Machine learning has the potential to improve data quality by reducing the amount of human intervention and therefore potential for human error. Finally, implementing operational efficiencies would improve our ability to produce rapid official statistics. In a time when private companies are able to produce data for public consumption very quickly, it is vitally important that NSOs improve operations in order to stay relevant.

## 10. Has there been collaboration with other NSIs, universities, etc?

We have collaborated with a number of other NSIs through the HLG-MOS ML Project, but have not been directly collaborating with other external institutions. A number of different technologies and pre-processing techniques were suggested to us during the last Sprint in Belgrade. However, due to limited resources we are only now beginning to explore the potential of these other approaches to improve the performance of our models. Examples of some of the suggested methods would be: the use of other technologies such as SVMs, and the use of libraries of synonyms such as Fuzzy Wuzzy and JellyFish.

## 11. Next Steps

*Applicability of NAICS and NOC models for future surveys*

Based on our success in coding CCHS's recent cycles (2019 Q3, Q4), we will consider using these models to code NAICS and NOC for other similar surveys. Given the complexity of the existing pipeline and risk involved with manual file manipulation outside of the CCE (Appendix, Figure A1), a cost-benefit analysis is warranted when deciding which auto-coding strategies (word-matching or ML) are appropriate for smaller surveys. However, once the CCE is able to consume results from G-Code ML models, the barrier to implement our existing ML solutions will be reduced. In addition to deploying our ML solutions for CCHS, CHMS, and other smaller surveys in the short term, we will also be working in collaboration with our subject matter and methodology partners to develop viable models for the JVWS and LFS in 2020. JVWS currently codes ~250k records per year and LFS currently codes ~220k records per year. These two large surveys represent our largest current manual coding clients that stand the most to benefit from introducing ML solutions in their operations.

*Expanding machine-learning oversight in production*

In order to comply with Statistics Canada's upcoming Framework for Responsible Machine Learning Processes, the BPTAS team will be expanding its ML production documentation. A consideration of how our modeling efforts promote fairness, transparency, privacy, and security amongst other factors, will serve to ensure our coding activities follow the Government of Canada's Principles of Responsible Artificial Intelligence. As we seek to expand our modeling activities, a framework that provides guidance through Peer Review / Committee Sign-Off at Statistics Canada will ensure robust quality validation. The framework is still under development and will be implemented on an ongoing basis as the framework and business processes are put in place.

## 12. References

1. **YanPeng Gao, Isaac Ross, Justin J. Evans.** Statistiscs_Canada_FastText_Techniques_Report. [Online] 2019. https://statswiki.unece.org/pages/viewpage.action?pageId=256969394.

2. **Justin J. Evans, Isaac Ross, Julie Portelance.** StatisticsCanada_CCHS_ML_Production_Report. [Online] 2020. https://statswiki.unece.org/pages/viewpage.action?pageId=256969394.