Catalogue no. 11-522-x
ISSN: 1709-8211

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

Need for Speed: Using fastText (Machine Learning) to Code the Labour Force Survey

by Justin Evans and Javier Oyarzun

Release date: November 5, 2021

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

Need for Speed: Using fastText (Machine Learning) to Code the Labour Force Survey

Justin Evans & Javier Oyarzun[1]

Abstract

Statistics Canada's Labour Force Survey (LFS) plays a fundamental role in the mandate of Statistics Canada. The labour market information provided by the LFS is among the most timely and important measures of the Canadian economy's overall performance. An integral part of the LFS monthly data processing is the coding of respondent's industry according to the North American Industrial Classification System (NAICS), occupation according to the National Occupational Classification System (NOC) and the Primary Class of Workers (PCOW). Each month, up to 20,000 records are coded manually. In 2020, Statistics Canada worked on developing Machine Learning models using fastText to code responses to the LFS questionnaire according to the three classifications mentioned previously. This article will provide an overview on the methodology developed and results obtained from a potential application of the use of fastText into the LFS coding process

Key Words: Machine Learning; Labour Force Survey; Text classification; fastText.

1. Introduction

The Labour Force Survey (LFS) is a mission-critical survey at Statistics Canada as it plays a fundamental role in the estimate of labour market conditions in Canada. LFS provides employment estimates by industry, occupation, public and private sector, and hours worked amongst others. The survey results are also used to make decisions regarding job creation, education and training, retirement pensions and income support. One of the main components of LFS is the monthly coding of the North American Industrial Classification System (NAICS), National Occupational Classification (NOC) and Primary Class of Workers (PCOW) records. Each records contains numeric and text descriptions, collected from survey respondents, which are used to assign a code in each classification. At Statistics Canada, the Operations and Integration Division coders manually code between 15,000 and 20,000 records in 7.5 days. Due to the limited time and large volume of coding required, a machine learning (ML) solution was sought.

National Statistical Organizations (NSOs) have been using ML for a variety of tasks including imputation, estimation, and text classification (United Nations Economic Commission for Europe's Machine Learning Team, 2018). In text classification, descriptions provided by survey respondents and associated with a variable of interest (NAICS, NOC, PCOW classifications) can be used in supervised ML tasks. Across NSOs, both the algorithms used for supervised ML tasks– Random Forest, Naïve Bayes, fastText, Support Vector Machine, Convolutional Neural Network, Multilayer perceptrons, etc. - and preprocessing steps can vary based on the type and availability of data (Sthamer, 2020).

At Statistics Canada (StatCan), the majority of automated coding activities are carried out using the internally developed coding application G-Code. G-Code's recent integration of fastText, a library for text representation and text classification created by Facebook's AI Research Lab (Joulin et al., 2016), and connection to StatCan coding platform Coding and Corrections Environment (CCE) has facilitated the use of machine learning models in the production of official statistics. This article will detail the LFS's classifications to be coded, the machine learning framework to do so, and the evaluation of the results.

[1]Justin Evans, Statistics Canada, 170 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 (justin.evans@statcan.gc.ca); Javier Oyarzun, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 (javier.oyarzun@statcan.gc.ca)

2. NOC, NAICS and PCOW Classifications

2.1 NOC

The NOC[2] was developed and is maintained by Employment and Social Development Canada (ESDC) and Statistics Canada. NOC is based on the organizational framework of occupations in the Canadian labour market. The criteria used to group occupations in NOC are generally the tasks, duties, and responsibilities of the occupation, and can include the degree of responsibility and complexity of work, as well as the products made and services provided. Therefore, NOC has been set up in a four-tiered hierarchical arrangement from broad categories to unit groups, leading to 500 distinct classes.

Figure 2.1-1. NOC hierarchical class structure

0 - Management occupations
- 00 - Senior management occupations
    - 001 - Legislators and senior management
        - 0011 - Legislators
        - 0012 - Senior government managers and officials
        - 0013 - Senior managers - financial, communications and other business services

2.2 NAICS

The NAICS[3] was developed collaboratively by the statistical agencies of Canada, Mexico, and the United States. NAICS is based on supply-side or production-oriented principles, to ensure that industrial data, classified to NAICS, are suitable for the analysis of production-related issues such as industrial performance. The criteria used to group industries in NAICS are input structures, labour skills, and production processes. Like NOC, NAICS has been set up in a hierarchical structure with the highest level dividing the economy into 20 sectors and the lowest dividing industries by economic activity, leading to 324 distinct classes.

2.3 PCOW

The PCOW is used to classify workers as either employed in a public or private business. Therefore, PCOW has two classes (1: Public, 2: Private).

3. Machine Learning Framework

3.1 Algorithm

FastText has been shown to achieve comparable results at a faster speed than traditional deep learning classifiers using sub-word n-gram representations and hierarchical softmax. FastText allows for each word to be represented as a bag of character n-grams; for example, 'text' with n=3, is converted to '<te, tex, ext, xt>'. Word vectors can therefore include misspelled or made-up words, and concatenated words (Joulin et al., 2016). In survey response data, which commonly contains misspelled words and industry specific terminology, we have found fastText to be able to produce high quality predictions. In addition, we find that given LFS large number of records and large class counts (NAICS-

[2] National Occupational Classification (NOC) 2016 Version 1.3 : https://www23.statcan.gc.ca/imdb/p3VD.pl?Function=getVD&TVD=1267777

[3] North American Industry Classification System (NAICS) Canada 2017 Version 3.0 : https://www23.statcan.gc.ca/imdb/p3VD.pl?Function=getVD&TVD=1181553

4, NOC-4), fastText's use of hierarchical softmax, an approximation of the softmax loss function, helped to reduce model training time.

3.2 Training

To evaluate the model fit, the previously coded available data was split into a training and a validation dataset. To account for monthly variations in reported occupations, one year of LFS was retained as a validation dataset (June, 2019 – May, 2020, records = 200,000). As LFS reviews its historical data each time a new classification year is introduced - for example moving from NAICS 2012 to NAICS 2017v1 - we were able to include training data from January 2013 to May 2019 (records = 1,500,000). Selection of auxiliary input text variables within the training dataset was evaluated using chi-squared and information gain. Minimal text preprocessing was done upon concatenation of selected text variables: uppercasing, accent removal, and masking. Random search hyper-parameter tuning, using k-fold cross validation, was done on all model variants before being run on the validation dataset.

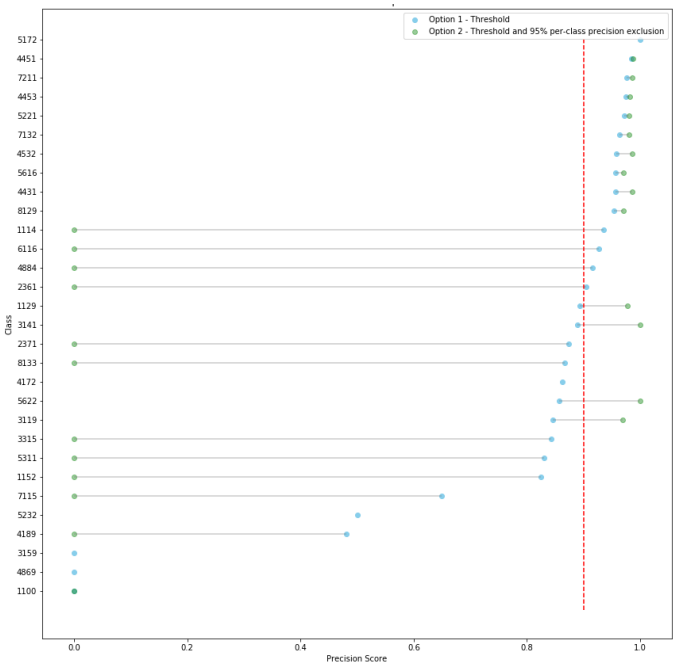
3.3 Threshold & Class Selection

In model prediction, fastText returns the class with the highest activation score after the softmax layer; returning a confidence score between 0 and 1. While this score is weakly associated with the model's probability in predicting a positive class, advanced calibration methods can be used to tune the classifier's confidence scores (Zadrozny & Elkan, 2002). Here, we show that while excluding all ML predictions with a confidence score below a given threshold can result in a high coding rate and overall precision score (Table 3.3-1, Option 1), many classes above the threshold fall below an acceptable level to be used in official statistics (Figure 3.3-1, Option 1). Therefore, to avoid systematic error and optimize the coding rate, we selected a threshold after looping through all available confidence scores while maintaining a 95% per class precision on our validation dataset. ML predictions in classes which fell below 95% precision were excluded from the proposed coding process (Option 2). Further work will investigate the impact of calibrating fastText confidence scores on the number of accepted classes in our process.

Table 3.3-1. Options Analysis for Model Prediction Thresholding

Options	Confidence Threshold	Coding Rate	Overall Precision	Classes > 95% precision
1 – Threshold	0.88	62.08%	95.10%	100
2 – Threshold and 95% per-class precision exclusion	0.96	41.41%	98.29%	143

Figure 3.3-1. Precision score of a subset of NAICS classes in model prediction thresholding. Each dot represents one class. Red line indicates a 95% class precision.



4. Results

Based on the model framework described in Section 3, NAICS, NOC, and PCOW models were created and tested on one year of LFS (validation dataset) (Table 4-1.). In the following sub-sections, we will address the impact of using those models in the coding process and on LFS's published estimates.

Table4-1.
Overall autocoding rate & precision score at the classification and record level, on the validation dataset.

Classification	Classification Level		Record Level	
	Autocoding Rate	Precision Score	Autocoding Rate	Precision Score
NAICS	40.9%	98.3%	18.5%	99.0%
NOC	30.7%	97.5%	18.5%	98.2%
PCOW	100%	97.6%	18.5%	98.6%

4.1 Comparison to manual coding

Evaluating the quality of a coding process can be time-consuming and costly. Here, we were able to utilize LFS's migration from its prior coding platform (CODCO) to its new coding platform (CCE) to compare manually coded records with those coded using ML. To do so, we compared the final outputted classes of a record coded in one coding platform to the same record in a different coding platform or to the classes predicted using ML; in which the error rate is the rate of disagreement on the final outputted class. We show that on the subset of records that would have been fully coded by ML (~ 18.5% of records), the error rate between two manually coded months (CODCO vs. CCE) is the same as if ML (before quality control) had been used in either coding platforms (ML vs. CODCO, ML vs. CCE). We further show that randomly selected ML records within the CCE's quality control (QC) process exhibit the same low error rate (ML QC'd in CCE) (Table 4.1-1). These results demonstrate that the threshold and class selection criteria (Section 3.3) produces NAICS, NOC, and PCOW predictions that are of the same quality as the manual process alone.

Table 4.1-1
Comparison of ML to manual coding error rates (ML subset) in different coding platforms (CODCO, CCE)

	CODCO vs. CCE	ML vs. CODCO	ML vs. CCE	ML QC'd in CCE
Year-Month	2020-09	2020-09	2020-09	2021-06
NAICS	2.2%	2.2%	1.7%	0.9%
NOC	2.3%	2.3%	2.0%	1.2%
PCOW	1.9%	1.9%	1.7%	1.3%

4.2 Time-series break analysis

In order to assess whether ML would impact the LFS trends which are key outputs for users , we sought to determine if there would be any observable changes in the distribution of classes over time. Each NAICS, NOC, and PCOW class was expressed as a series of data points ordered in time (time-series). For a time-series an abrupt change at a point in time is called a structural break or time-series break. Structural breaks would indicate that there was a significant change in LFS's time-series data and would have to be communicated with data users.

For each month in the year period, a count of units per class was calculated for Manual Coding (without ML) and the Integrated Manual & ML (with ML) to assess potential time-series breaks. Due to the low-error rate of the ML models the distribution of classes does not appear to be drastically different when ML predictions are integrated into the manual process (Table 4.2-1). Here, the largest difference in class counts are 'Full-service restaurants and limited-service eating places' (NAICS - 7225), with 6 more records assigned to this class.

Figure4.2-1.
Class distribution change when integrating ML predictions. Reference period: May, 2020.

Store shelf stockers, clerks and order fillers

Food counter attendants, kitchen helpers …

Janitors, caretakers and building superintendents

NOC-4	# Coded Manual	# Coded Int. ML	Absolute Difference
6622	258	263	5
6711	388	392	4
6733	218	214	4
0631	127	130	3
4112	73	76	3
6731	283	286	2
7452	177	174	2
6341	88	90	2
6511	63	61	2
…	…	…	…

Full-service restaurants and limited-service eating places

NAICS-4	# Coded Manual	# Coded Int. ML	Absolute Difference
7225	874	880	6
7224	15	13	2
4471	61	63	2
5616	91	93	2
5617	399	401	2
7139	266	268	2
8113	72	70	2
9120	274	272	2
1112	33	34	1
…	…	…	…

LFS uses the bootstrap variance estimation method to report on monthly total employment and unemployment (‘Statistics Canada, 2017). We therefore sought to determine if there would be any structural break in these published estimates if ML had been used. Two different criteria (listed below) were used to identify any potential differences between the Manual estimates and the Integrated Manual & ML estimates. A similar approach was taken to evaluate time-series breaks at the NAICS-6 digit when the Capital Expenditures Survey (CES) moved to the Integrated Business Statistics Program (IBSP) (Oyarzun, 2016).

- Criterion 1: Flag the estimate if the 90% confidence interval for the Manual estimate does not overlap the 90% confidence interval for the Integrated Manual & ML estimate.
- Criterion 2: Flag the estimate if the Manual point estimate does not fall in the 90% confidence interval (CI) for the Integrated Manual & ML estimate.

Here we compared class estimates at each digit level (ex. NAICS-4, NAICS-3, NAICS-2, NAICS-1) over a 12 month period (October, 2019 – September, 2020). We find that integrating ML does not result in variance estimates that fall outside the manual estimate confidence limits (Criteria 1) nor have a manual process point estimate outside of its confidence limit (Criteria 2). While we didn’t observe breaks in the 13,665 estimates compared (NAICS: 5,332; NOC: 8,309; PCOW: 24), we can expect some volatility in classes with low record counts, but this should not represent a concern for data users.

4.3 Quality control analysis

Quality control is a technique used to ensure that quality is above an established level by measuring the quality of the characteristics of interest, comparing it to a quality target and taking corrective action if the standard is not achieved (‘Statistics Canada, 2003). In the CCE, both manual coding and ML are subject to quality control through a Simple Random Sample (denoted ‘Regular QC’ in Table 4.3-1). The QC rate for each coder is determined based on the previous month’s error rate, and directly impacts the out-going error rate. Based on the average number of records received for coding and the QC rates assigned, coders are able to complete a maximum number of records each month (approximately 24,200 records). Here, we show two scenarios where using ML can shift manual coders from front-line coding to verification. Further, by reducing the QC on ML going from Scenario 2 to Scenario 3 (indicated in ‘Total Manual Codes assigned (without regular QC)’), we can increase QC on the harder, more error prone records that remain for manual coders (denoted ‘Total workload remaining (regular QC), non-ML’); thereby reducing the out-going error rate. If these ML efficiencies are re-invested under a short-term ML QC of 50% and long-term ML QC of 10%, the out-going error rate would drop to 10.4% and 9.5% respectively (Table 4.3-1).

Table4.3-1
Scenarios for re-investing ML efficiencies in manual QC and the impact on the estimated out-going error rate.

#	Metric	Scenario 1 (no ML)	Scenario 2 (ML at 50% QC)	Scenario 3 (ML at 10% QC)
1	Coding Centre Record "Budget"	24,200	24,200	24,200
2	Estimated Number of ML Codes	0	3,000	3,000
3	Total Manual Codes assigned (without regular QC)	22,000	20,545	19,155
4	Total workload remaining (regular QC), non-ML	2,200	3,655	5,046
5	Records not part of QC, non-ML	12,929	8,889	7,896
6	Records not part of QC, ML	0	1,500	2,850
7	Estimated # errors out-going, non-ML	2,586	1,778	1,579
8	Estimated # errors out-going, ML	0	45	85.5
9	**Estimated out-going error rate**	**14.78%**	**10.42%**	**9.51%**

5. Conclusion

We have demonstrated a model framework to code high quality NAICS, NOC, and PCOW predictions for the LFS. With tunable confidence thresholding and class exclusions rules, we can ensure ML codes at a quality required for official statistics. Moreover, our examination of the QC and coding process suggest ML will shift the nature of manual coding work and improve LFS's out-going error rate. Finally it is worth noting that the ML coding was successfully implemented in the LFS production process in October 2021.

Acknowledgements

The authors would like to thank the following people for their contributions: Laura Wile, Khushnood Khan, Anthony Yeung, Danielle Lebrasseur, Julie Portelance, Isaac Ross, Kartik Vashisth, Sreenija Koya, Sylvie White, Meghan Fulford, David Menyah, and Charles Mitchell.

References

Joulin, E. Grave, P. Bojanowski, T. Mikolov. (2016), "Bag of Tricks for Efficient Text Classification", arXiv: 1607.01759.

Oyarzun, J. (2016). "Needle in a haystack, how to detect break in series: the capital expenditure survey integrated business statistics program experience." paper presented at SSC Annual Meeting - Proceedings on the Survey Methods Section.

Statistics Canada (2017), "Methodology of the Canadian Labour Force Survey", Catalogue no. 71-526-X2017001.

Statistics Canada (2003), "Survey Methods and Practices." 309-319. Catalogue no. 12-587-XPE.

Sthamer, C. (2020) "UNECE–HLG-MOS Machine Learning Project Classification and Coding Theme Report", report

United Nations Economic Commission for Europe's Machine Learning Team (2018) "The use of machine learning in official statistics", report

Wood, S., Muthyala, R., Jin, Y., Qin, Y., Rukadikar, N., Rai, A., & Gao H. (2017), "Automated Industry Classification with Deep Learning", paper presented at IEEE International Conference on Big Data

Zadrozny, B., & Elkan, C. (2002), "Transforming Classifier Scores into Accurate Multiclass Probability Estimates", KDD, pp. 694-699.