Proceedings of Statistics Canada Symposium 2018
Combine to Conquer: Innovations in the Use of Multiple Sources of Data

The Imitation Game: An Overview of a Machine Learning Approach to Code the Industrial Classification

Javier Oyarzun[1]

Abstract

Statistics Canada's Business Register (BR) plays a fundamental role in the mandate of Statistics Canada. The BR is a database that includes all businesses operating in Canada. Close to two hundred business surveys use the BR in various ways. The Business Register has a direct impact on the efficiency of the business survey process, the reliability of data produced by business statistics programs and the coherence of the national accounting system. One of its key attributes is the industrial code. In early 2018, Statistics Canada started developing a new methodology to probabilistically code the industrial classification of businesses. This methodology, which uses text mining and machine learning, will provide Statistics Canada with a tool to code missing industrial classifications and improve the overall quality of the industrial classifications on the BR.

This article explains the North American Industrial Classification System (NAICS), its usage in statistical programs at Statistics Canada, the current and new approaches to NAICS coding, as well as a discussion on the challenges related to uncoded businesses and examples of complex cases of NAICS coding.

Key Words: Business Register; Machine Learning; Text Mining; NAICS; Coding, industrial, classification.

1. Introduction

Statistics Canada's Business Register (BR) plays a fundamental role in the mandate of Statistics Canada. The BR is a database that includes all businesses operating in Canada. Close to two hundred business surveys use the BR in various ways, mainly for establishing survey frames, sampling, collecting and processing data, and producing estimates. The Business Register has a direct impact on the efficiency of the business survey process, the reliability of data produced by business statistics programs and the coherence of the national accounting system. Such an endeavour demands a reliable quality control process. One of its key attributes is the industrial code from the North American Industrial Classification System (NAICS) assigned to each business. The NAICS is a six-digit code which is assigned based on the main business activity. Currently about 500,000 of the near 7,000,000 active businesses remain uncoded, the vast majority being relatively small businesses. In early 2018, Statistics Canada began developing machine learning algorithms to code these units to increase coverage of the industrial classification of businesses with the goal to improve economic estimates for all economic survey programs.

This article is divided in five sections and explores the multiple facets of this new coding methodology. The first provides a brief introduction to the BR. The second is a discussion on the NAICS and its usage at Statistics Canada. The third presents BR's current NAICS coding approaches and the fourth new text mining and machine learning NAICS coding approaches. Finally, a discussion on the BR's population of uncoded entities (backlog) and the application of the new NAICS coding methods is presented in the last section.

2. NAICS

The NAICS is an industry classification system developed by the statistical agencies of Canada, Mexico and the United States in 1997. Created against the background of the North American Free Trade Agreement, it is designed to provide common definitions for the industrial structure of the three countries and a common statistical framework to facilitate

[1] Javier Oyarzun, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 (javier.oyarzun@canada.ca)

the analysis of the three economies. 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.[2]

The NAICS is designed to classify businesses and other organizations engaged in the production of goods and services. They include incorporated (corporations, T2) and unincorporated (T1) businesses. They also include government institutions and agencies engaged in the production of marketed and non-marketed services, as well as organizations such as farms, professional associations, unions, charitable or non-profit organizations, and the employees of households. NAICS is a comprehensive six-digit code with 928 options assigned to each business based on its main activity. The structure of NAICS is hierarchical. It is composed of sectors (two-digit codes), subsectors (three-digit codes), industry groups (four-digit codes), and industries (five-digit codes). The sixth digit is used to designate national industries. Table 2-1 presents the NAICS classification at the two-digit level.

Table 2-1
NAICS at the two-digit level

NAICS	Sector
11	Agriculture, forestry, fishing and hunting
21	Mining, quarrying, and oil and gas extraction
22	Utilities
23	Construction
31-33	Manufacturing
41	Wholesale trade
44-45	Retail trade
48-49	Transportation and warehousing
51	Information and cultural industries
52	Finance and insurance
53	Real estate and rental and leasing
54	Professional, scientific and technical services
55	Management of companies and enterprises
56	Administrative and support, waste management and remediation services
61	Educational services
62	Health care and social assistance
71	Arts, entertainment and recreation
72	Accommodation and food services
81	Other services (except public administration)
91	Public administration

3. Current NAICS Coding Approaches

The BR has the responsibility of coding and maintaining the NAICS for all Canadian businesses. Assigning a NAICS to Canadian businesses has been a challenge ever since its creation. Throughout the years, the BR has examined different approaches for NAICS coding. The main approaches currently used by the BR are the following:

- **Administrative sources – mainly from the Canada Revenue Agency (CRA)**
  - Business self-coded industrial classification (NAICS), and
  - Business activity descriptions (as provided on business tax returns) – these are then processed with Statistics Canada's G-Code[3] based on a reference file of common industrial terms.
- **Survey feedback**
  - Businesses are surveyed or pre-contacted.
- **Profiling activities**

[2] Statistics Canada, (2018), *NAICS 2017 version 3.0*.
[3] Statistics Canada, (2018), *G-Code version 3.0.*

- Businesses are contacted on a periodic basis to be profiled for BR maintenance.

These approaches have strengths and weaknesses. For example, self-coded NAICS have been criticized for its poor quality in some industrial sectors. However, it is often the only source of information available (especially for T1s). Coding based on CRA's activity description field using G-Code requires the business activity description to be filled in, which is not always the case. Coding through survey is expensive and increases response burden. Manual coding, through survey or profiling, can enhance quality, but requires a lot of time and training.

4. New NAICS Coding Approaches

In early 2018, Statistics Canada decided to explore new NAICS coding approaches to not only improve the quality of NAICS coding on the BR, but also to assign a NAICS to the uncoded BR entities (see section 5 for more information). The following approaches will be discussed in further details in the following sub-sections: (4.1) Text Mining, and (4.2) Machine Learning.

4.1 Text Mining

Text mining is the process of deriving high-quality information from text. Text mining usually involves information retrieval, preprocessing to convert raw input textual data to structured format, uncovering patterns within the resulting derived structured data, lexical analysis to study word frequency distributions, pattern recognition and predictive analytics. A typical application is to scan a set of documents (commonly referred to as the "corpus" in text mining terminology) written in a natural language and either model the corpus for predictive classification purposes or populate a certain search index with the information extracted.

4.1.1 Text Coding

For each NAICS-specific word in a business name, an industry description field or the main activity field can be more prevalent (or even unique). For example, the word "BRDCST" appears in the industry descriptions of over 6,000 businesses, all of which belong to the NAICS classification 519130 (Internet broadcasting and web search portals). In contrast, the more common word "EGGS" appears in the industry descriptions of more than 20,000 businesses spanning across several NAICS. Among the businesses whose industry descriptions contain the word "EGGS", the most common NAICS is 112310 (Chicken egg production), which occurs 30.4% of the time, while the NAICS 413130 (Poultry and egg merchant) occurs around 20% of the time among these businesses. Table 4.1.1-1 shows a number of examples of words in industry descriptions, the most common NAICS among the businesses whose industry descriptions contain these words:

$$P(NAICS|W_{NAICS}) = \frac{Freq(W_{NAICS})}{\sum_{k \in All\ NAICS} Freq(W_k)}$$

where $Freq(W_{NAICS})$ is the number of businesses whose industry descriptions contain the word W and whose NAICS classification is NAICS.

Table 4.1.1-1
Word in the description field and the associated NAICS

Word (W)	$Freq(W)$)	$P(NAICS\|W_{NAICS})$	NAICS	NAICS Title
BRDCST	6,000	100.0%	519130	Internet publishing and broadcasting and web search portals
POLLING	4,000	100.0%	541910	Marketing research and public opinion polling
APICULTURE	2,000	100.0%	112910	Apiculture
UNIFAMILIALES	1,000	100.0%	236110	Residential building construction
DOUGHNUT	400	100.0%	722512	Limited-service eating place
BAPTIST	200	100.0%	813110	Religious organizations
CAMPINGS	150	100.0%	721211	Recreational vehicle (RV) parks and campgrounds
JANITORIAL	9,000	88.9%	561722	Janitorial services (except window cleaning)
CHURCH	15,000	84.1%	813110	Religious organizations
ADJUSTERS	200	75.6%	524291	Claims adjusters
EGGS	20,000+	30.4%	112310	Chicken egg production

4.1.2 Predicting NAICS Using Text Mining

As shown is sub-section 4.1.1, it is possible to predict a business' NAICS based on words and names using text mining algorithms. Therefore, multiple words or names can be used to predict a specific NAICS. For example, words like "Church", "Nativity", "Baptists" and many others can help code businesses to the 813110 NAICS (Religious organizations), see figure 4.1.2-1.

Figure 4.1.2-1
Example of words associated to NAICS 813110 (accuracy rate is presented in %)

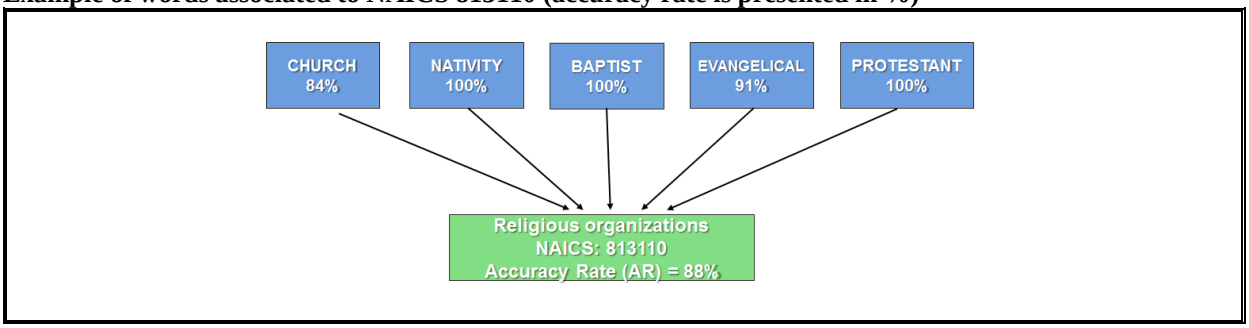

Table 4.1.2-1 presents the accuracy rate when using text mining with business name (NM), description field (DM), and the main activity field (AM). It was found that coding NAICS is of higher quality when using the description field. This means that words in DM have a stronger association to the NAICS than words found in NM or AM. The table also shows that the accuracy rate is much lower for the manufacturing (NAICS-2: 31, 32 and 33), wholesale trade (NAICS-2: 41) and management (NAICS-2: 55) sectors, regardless of whether NM, DM or AM was used. This means that the words found in businesses in these sectors are not good predictors of NAICS.

Table 4.1.2-1
Text mining accuracy rate (in %) using business name (NM), description field (DM) and the main activity field (AM)

NAICS	11	21	22	23	31-33	41	44-45	48-49	51	52	53	54	55	56	61	62	71	72	81	91	Total
NM	63	49	54	90	36	44	85	83	65	80	91	80	2	76	29	91	58	95	78	82	83
DM	98	80	86	89	66	24	90	97	91	97	99	93	75	93	94	98	97	96	90	88	94
AM	95	39	24	94	31	13	79	78	71	90	39	92	2	82	60	98	77	97	85	93	87

4.2 Machine Learning

Following the text mining NAICS coding, the BR decided to explore the use of a multivariate Bernoulli naïve Bayes classifier to predict NAICS, based on features derived from the following input textual variables:

- Business Name,
- Description Field, and
- Main Activity Field.

The name, description, and main activity of each business were concatenated and treated as a single document. A vocabulary was constructed based on the words that occur in the resulting corpus of documents. Multivariate one-hot feature vector was then generated for each document, based on the word occurrences in the given document. The results were of higher quality than the text mining method presented in section 4.1; see table 4.2.1-2. The following are the steps used in our preprocessing, feature engineering and machine learning workflow:

- Standardize words: Up case, remove accents, stemming, etc.
- Make a document frequency matrix (DFM): Possible use of n-grams;
- Pick ideal parameters in DFM: Highest accuracy with testing/training data;
- Build model;
- Predict; and
- Evaluate/Decision.

4.2.1 Naïve Bayes Classifiers

The naïve Bayes (NB) classifiers are a collection of classification techniques whose underlying models are probabilistic and make the naïve Bayes assumption. The naïve Bayes assumption is that the features are conditionally independent given the class label. For certain classification problems, naïve Bayes classifiers can be trained very efficiently. For this reason, naïve Bayes classifiers remain popular methods for text categorization (see table 4.2.1-1 for examples). For the purpose of the NAICS machine learning coding, the R package *Quanteda*, which provides an implementation of the multivariate Bernoulli naïve Bayes classifier was used. Table 4.2.1-2 presents this method accuracy rate by industrial sectors. The naïve Bayes classifier method uses the following formula:

$$\boldsymbol{P(C|X_i)} = \frac{\prod_i \big(\boldsymbol{P(X_i|C)}\big)\, \boldsymbol{P(C)}}{\boldsymbol{P(X_i)}}$$

Where C is the class label (NAICS) and X_i are the features (words).
$P(C)$ is be the probability of businesses belonging to a specific NAICS.
$P(X_i|C)$ is the probability of businesses with a combination of specific features knowing that it belongs to a specific NAICS.

Table 4.2.1-1
Naïve Bayesian machine learning NAICS coding based on words

Word	NAICS 1	Title 1	Prob 1	NAICS 2	Title 2	Prob 2	NAICS 3	Title 3	Prob 3
ADJUSTERS	524291	Claims adjusters	75.6%	524210	Insurance agencies and brokerages	12.3%	524299	All other insurance related activities	3.4%
CLAIM ADJUSTERS	524291	Claims adjusters	88.0%	524210	Insurance agencies and brokerages	6.8%	524299	All other insurance related activities	3.0%
"H&S BRAND" CLAIM ADJUSTERS	524291	Claims adjusters	98.2%	524210	Insurance agencies and brokerages	1.4%	524299	All other insurance related activities	0.2%

Table 4.2.1-2
Machine Learning NAICS coded accuracy rate (in %)

NAICS	11	21	22	23	31-33	41	44-45	48-49	51	52	53	54	55	56	61	62	71	72	81	91	Total
NB (1)	99	94	96	99	88	96	95	98	94	98	100	98	80	97	98	98	99	95	96	99	97
NB (2)	93	83	78	85	72	75	91	88	89	88	96	85	77	78	88	92	93	90	84	86	89

Notes: (1) Naïve Bayes Classifier with a posterior probability threshold of greater than 80%. (2) Naïve Bayes Classifier with no probability threshold.

5. Business Register's Backlog

The Business Register's backlog refers to "the active" population present on the BR that current NAICS coding approaches are unable to code. These units therefore cannot be in-scope for NAICS-driven economic surveys (a majority of Statistics Canada's business surveys), which may leads to the underestimation of economic parameters. According to the following key size measures, the backlog represents:

- 500,000 active establishments (5%-8% of the BR);
- $200 billion in revenue (3%-5% of the BR); and
- 400,000 in employment (2%-3% of the BR).

As shown in section 4, text mining and machine learning techniques could be used to code a significant number of backlog units. Tables 4.1.2-1 and 4.2.1-2 demonstrate the accuracy rate of the various text mining and machine learning coding techniques (Business Name Mining (NM), Description Mining (DM) and tax activity mining (AM) and naïve Bayes Classifier (NB)). Table 5-1 presents the number of units that can be coded using the previously mentioned methods.

Table 5-1
Backlog coded using text mining and machine learning

NAICS	NM	DM	AM	NB
11	441	3,790	312	4,997
21	100	141	27	2,757
22	14	498	1	224
23	4,035	10,847	2,145	21,327
31-33	195	2,002	42	1,020
41	335	505	14	7,347
44-45	2,083	4,164	1,079	8,150
48-49	636	5,754	268	6,787
51	533	4,038	778	3,200
52	1,762	11,980	782	32,073
53	5,705	5,892	138	28,650
54	2,959	19,411	5,605	42,877
55	9	362	0	30,096
56	1,827	8,596	603	10,288
61	204	1,504	463	3,321
62	3,742	5,408	2,278	8,095
71	337	4,447	603	4,569
72	2,730	3,539	1,341	6,188
81	11,857	10,439	2,865	33,255
91	51	119	15	78
Total	39,555	103,436	19,359	255,299

6. Conclusion

The accuracy of the Business Register has a direct impact on the efficiency of the business survey process, the reliability of data produced by business statistics programs and the coherence of the national accounting system. The text mining and machine learning methodologies presented in this article provide Statistics Canada with new high-quality NAICS coding approaches. These new approaches can be used to code a large portion of the BR's NAICS backlog with positive impact on Statistics Canada economic program's estimates.

Acknowledgements

The author would like to thank the following contributors who helped make this project possible: Sonja Simic, Aaron McBride, Shuai Zhang, Yi Li, Laura Wile, Kenneth Chu, Christian Wolfe, Anthony Yeung, Danielle Lebrasseur, Alain Therrien, Jamie Brunet, Jeff Mondoux, Linda Scantland, Amanda Maddicks and Stan Hatko.

References

Benoit, K., K. Watanabe, H. Wang, P. Nulty, A. Obeng, S. Müller, and A. Matsuo. (2018), "Quanteda: An R package for the Quantitative Analysis of Textual Data", *Journal of Open Source Software*, 3(30), p. 774.

Horwood, J. (2018), "Machine Learning for Scanner Data Classification", unpublished document, Ottawa, Canada: Statistics Canada.

Oyarzun, J. (2018), “NAICS coding: How can we do better?” presented at Statistics Canada’s Business Survey Methods Division Technical Committee, May 25, 2018.

Simic, S., and J. Oyarzun (2018), “OK Computer: Using Machine Learning to Code NAICS on the Business Register” presented at Statistics Canada’s Business Survey Methods Division Seminars, June 14, 2018.

Statistics Canada (2018), *NAICS Canada 2017 Version 3.0*.