Journal of Data Science 23 (2), 429–448
April 2025

DOI: 10.6339/25-JDS1180
Data Science in Action

BEACON: A Tool for Industry Self-Classification in the Economic Census

Brian Dumbacher[1,*], Daniel Whitehead[1], Jiseok Jeong[1], and Sarah Pfeiff[1]

[1]*U.S. Census Bureau, Washington, DC 20233, United States*

Abstract

Business Establishment Automated Classification of NAICS (BEACON) is a text classification tool that helps respondents to the U.S. Census Bureau's economic surveys self-classify their business activity in real time. The tool is based on rich training data, natural language processing, machine learning, and information retrieval. It is implemented using Python and an application programming interface. This paper describes BEACON's methodology and successful application to the 2022 Economic Census, during which the tool was used over half a million times. BEACON has demonstrated that it recognizes a large vocabulary, quickly returns relevant results to respondents, and reduces clerical work associated with industry code assignment.

Keywords *Economic Census; machine learning; NAICS; ranked text classification; short text*

1 Introduction

1.1 NAICS and the Economic Census

Implemented in 1997, the North American Industry Classification System (NAICS) was developed jointly by the United States, Canada, and Mexico to facilitate economic analyses of these three North American countries (U.S. Census Bureau, 2024d). NAICS classifies establishments—physical locations where business is conducted—according to their production processes and primary business activity. A key use of NAICS is to provide a consistent and uniform way to present summary statistics about the U.S. economy. The U.S. Census Bureau and other statistical agencies use NAICS throughout the economic survey life cycle including sample selection, data collection, editing, and publication of establishment data (Kirkendall et al., 2018). The proper NAICS classification of establishments is therefore important for the accuracy of official economic statistics.

NAICS uses a hierarchical six-digit coding scheme to identify business activity at different levels of detail. The first two digits of the NAICS code represent the broad economic sector. Some sectors are represented by multiple two-digit codes. For example, Manufacturing consists of 31–33. There are 20 sectors such as Construction, Manufacturing, and Retail Trade. For a complete list, see Table 8 in Appendix A. Additional non-zero digits add industry detail. NAICS is revised every five years to reflect the changing economy. The 2022 vintage of NAICS identifies 1,012 codes at the 6-digit level. Table 1 breaks down the structure of an example NAICS code in the Accommodation and Food Services sector.

*Corresponding author. Email: brian.dumbacher@census.gov.

Received July 19, 2024; Accepted March 20, 2025

Table 1: Structure of an example NAICS code.

Level of detail	**NAICS code**	**Description**
Sector	72	Accommodation and Food Services
Subsector	721	Accommodation
Industry Group	7211	Traveler Accommodation
NAICS Industry	72119	Other Traveler Accommodation
National Industry	721191	Bed-and-Breakfast Inns

Source: U.S. Census Bureau (2024d). Disclosure Review Board (DRB) approval number: CBDRB-FY24-ESMD001-007.

Revisions to NAICS coincide with the Economic Census. Conducted by the U.S. Census Bureau for years ending in "2" and "7," the Economic Census is an extensive survey covering approximately eight million establishments with paid employees, most industries, and all geographic areas of the United States (U.S. Census Bureau, 2024a). About half of these establishments are asked to complete an electronic questionnaire. The other half is accounted for through administrative records (U.S. Census Bureau, 2024b). The Economic Census provides a wealth of information to help policymakers, businesses, trade associations, and other federal agencies understand economic activity at a granular level. Key statistics include total number of establishments; total number of employees; value of sales, shipments, receipts, and revenue; and total annual payroll. Data products from the Economic Census regarding establishments are broken down by geography and industry, as classified by NAICS. For technical details about the Economic Census design and methodology, see U.S. Census Bureau (2024b).

1.2 Problem Description

The Primary Business or Activity (PBA) question in the Economic Census asks respondents to describe their business. Answers help keep NAICS code assignments up to date. The PBA question displays pre-listed descriptions based on the establishment's current classification, and the respondent is asked to select one. If none of the pre-listed descriptions seems accurate, the respondent can select an "Other" option and provide a short, open-ended description known as a write-in. The term write-in is still used even though the descriptions are now typed in, as opposed to written in. The respondent is also able to indicate the economic sector using a drop-down menu. To illustrate, Figure 1 is a screenshot of the PBA question from the 2022 Economic Census Drinking Places (Alcoholic Beverages) questionnaire. This questionnaire is intended for bars, taverns, and similar establishments. Example write-ins that respondents might provide include "liquor distributor" and "brewpub."

Every Economic Census, the U.S. Census Bureau receives hundreds of thousands of write-in responses to the PBA question. For the most part, clerks process and assign NAICS codes to these cases manually, which is resource intensive. According to Snijkers et al. (2013), manual coding has the three key disadvantages of being expensive, time-consuming, and subjective. Given the same information about a write-in case, different clerks may assign different NAICS codes, and any resulting errors may be difficult to diagnose.

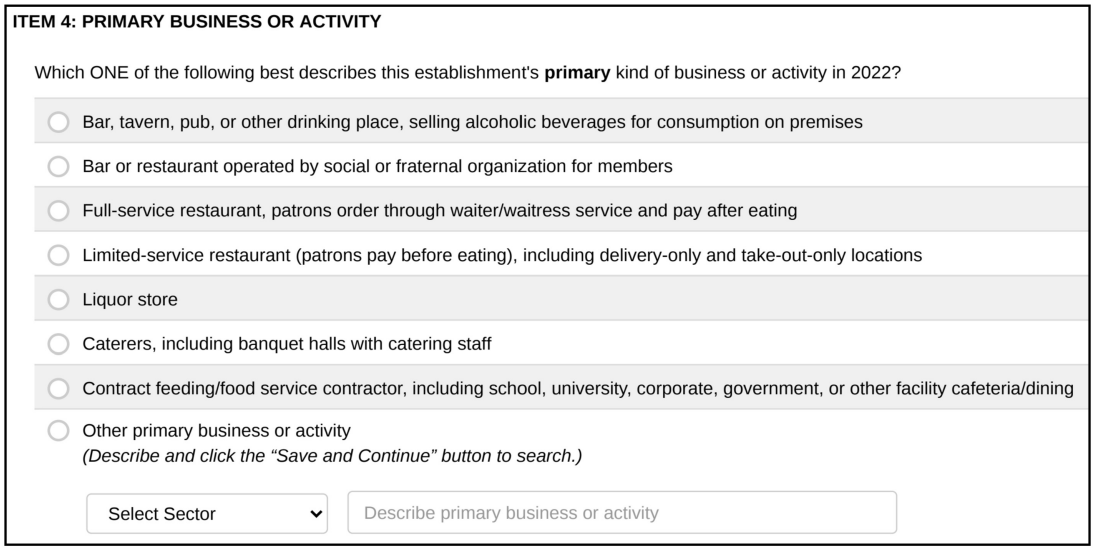

Figure 1: Primary Business or Activity question from the 2022 Economic Census Drinking Places (Alcoholic Beverages) questionnaire. Example write-ins include "liquor distributor" and "brew-pub." Source: https://bhs.econ.census.gov/ombpdfs2022/export/2022_AF-72240_su.pdf.

A more automated approach to NAICS code assignment can help address these disadvantages and improve efficiency (Baumgartner et al., 2021). To this end, the U.S. Census Bureau developed a tool called Business Establishment Automated Classification of NAICS (BEACON) to help respondents self-classify their NAICS code in real time. With BEACON, the "Other" option on the Economic Census pre-list screen essentially became a NAICS code search. The respondent provides a short business description as normal, and then BEACON returns a ranked list of candidate 6-digit NAICS codes for the respondent to choose from. Returning relevant NAICS codes in this manner is a supervised learning problem, specifically ranked text classification (Aggarwal, 2018).

1.3 Related Work

There are various examples of U.S. government agencies using supervised learning with textual features to assign NAICS codes, albeit not interactively with respondents. Kornbau (2016) and Kearney and Kornbau (2005) describe how the U.S. Census Bureau, Internal Revenue Service (IRS), and Social Security Administration developed a NAICS Autocoder for new businesses based on the IRS's SS-4 form (Internal Revenue Service, 2023), which is used for tax filing and reporting purposes. The Autocoder employs logistic regression and creates dictionaries of words, bigrams, and complete write-in text from the SS-4 business name and description text fields. In 2015, 79 percent of 3.6 million new business records were autocoded according to this methodology. In turn, about 69 percent of the autocoded records were classified to a full 6-digit NAICS code.

Dumbacher and Russell (2019) use 2012 and 2017 Economic Census data and traditional machine learning algorithms to build and evaluate classification models at the NAICS sector level. The authors consider the business name, write-in text, and industry-specific survey prompt as sources for word and bigram features. Logistic regression is found to outperform Naïve Bayes

for all but one feature combination. Some of the highest misclassification rates involve the Wholesale Trade, Retail Trade, and Other Services (except Public Administration) sectors.

More recently, Cuffe et al. (2022) explore NAICS classification models based on publicly available user reviews from the Google Places application programming interface (API) and text scraped from business websites. These data are matched to confidential U.S. Census Bureau economic records. The authors consider random forests with features derived from a Doc2Vec (Mikolov et al., 2013) representation of text. Prediction is at the NAICS sector level. The authors find model accuracy to be comparable to that of Kearney and Kornbau (2005) and note large differences in performance across sectors because of varying sample and dictionary sizes.

In the context of NAICS classification for IRS purposes, Oehlert et al. (2022) describe a supervised learning approach to validate or predict taxpayers' NAICS sector. The authors use random forests with a combination of (dollar value) tax return line-items and frequently occurring business description write-in tokens as features. In several of the applications considered, the random forest model provides appreciably better results than either the baseline method or the taxpayer self-reported sector.

Other National Statistical Organizations such as Statistics Canada (Evans and Oyarzun, 2021; Oyarzun, 2018), Statistics Netherlands (Roelands et al., 2018), and the Australian Bureau of Statistics (Tarnow-Mordi, 2017) are researching supervised learning methods to automate industry code assignment. Applications outside of the public sector are also relevant to the Economic Census write-in problem and illustrate alternative approaches. In the context of finance, insurance, and business analytics, Rizinski et al. (2024) review industry classification methods that combine natural language processing and deep learning. These applications involve various industry classification schemes, including NAICS.

2 Methodology

Drawing on the related work, BEACON is a complex model with many methodological components. It uses rich training data and adopts a new approach based on a combination of natural language processing (Jurafsky and Martin, 2009), machine learning, and information retrieval (Aggarwal, 2018). BEACON is designed to classify short text (typically fewer than ten words) and handle large numbers of observations, classes (6-digit NAICS codes), and model features.

2.1 Training Data

BEACON's training data consists of over 4.3 million observations across all NAICS sectors and industries. These observations were assembled from five sources: past Economic Census write-ins, IRS SS-4 forms, the U.S. Census Bureau's Classification Assistance Tool, autocoded write-ins from the 2017 Economic Census, and the Harmonized System (Dumbacher and Whitehead, 2022). The following are detailed descriptions of the data sources.

- For the Economic Census data source, BEACON uses write-in descriptions from the 2002, 2007, 2012, 2017, and 2022 Economic Census provided mainly by single-unit (one physical location) business establishments (Dumbacher and Whitehead, 2022).
- The IRS SS-4 data source consists of write-in descriptions from the IRS SS-4 form. The relevant open-ended question on the form asks for the "principal line of merchandise sold, specific construction work done, products produced, or services provided" (Internal Revenue Service, 2023). This data source covers 2002–2016.

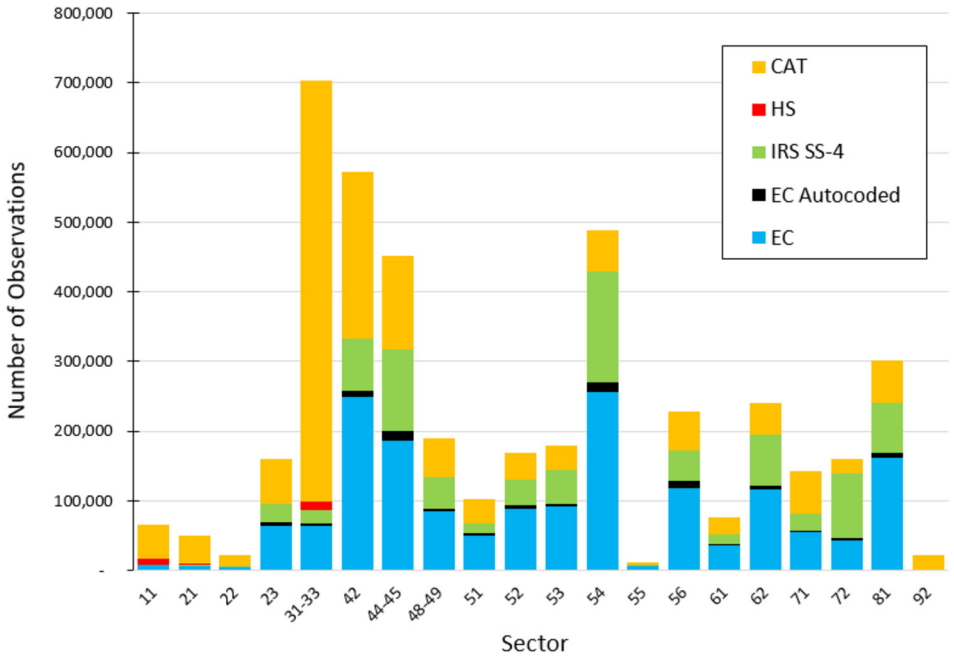

Figure 2: BEACON training data breakdown by sector and source. The five data sources are 2002, 2007, 2012, 2017, 2022 Economic Census (EC); 2017 EC Autocoded; 2002–2016 Internal Revenue Service (IRS) SS-4 forms; Classification Assistance Tool (CAT); and Harmonized System (HS).

- The Classification Assistance Tool is used internally by U.S. Census Bureau analysts in their industry classification work. The corresponding data source is the underlying catalog of business descriptions. These descriptions include those found in the publicly available NAICS manual (U.S. Census Bureau, 2024d).
- During the 2017 Economic Census, an exact-match autocoder was used to assign NAICS codes to frequently occurring write-ins. The autocoded data source consists of these write-ins and the corresponding NAICS codes that were assigned automatically.
- The Harmonized System is an internationally standardized system for classifying traded products. This publicly available data source consists of commodity descriptions and their associated NAICS codes (U.S. Census Bureau, 2024c).

Figure 2 is a breakdown of BEACON's training data by sector and source. These five sources complement one another well and combine to form a rich dataset. The Economic Census data source is the most representative of the target population for the problem BEACON is trying to solve. The other sources supplement it by providing coverage for seldom reported industries (Dumbacher and Whitehead, 2022). For example, the Classification Assistance Tool data source provides a large, albeit technical, vocabulary covering all industries. It includes duplicates and variations of descriptions to give certain industries more representation (Dumbacher and Whitehead, 2022). Similarly, the inclusion of the Harmonized System data source increases sample sizes for sectors not represented well elsewhere. The diversity in BEACON's training data allows for accurate predictions of both seldom reported industries and frequently occurring ones.

Table 2: Text cleaning algorithm examples. Misspellings are intentional.

Write-in text	**Clean text**
This is a convenence store.	conveni store
automobile MFG	car manufactur
We rapair watches & jewelry.	repair watch jewelri
New and ussed car dealer-ship	new used car dealership
long dist trckng	long distanc trucking
we do liq dist	liquor distribut
3PL	thirdparti logist

Disclosure Review Board (DRB) approval number: CBDRB-FY24-ESMD001-007.

2.2 Natural Language Processing

BEACON utilizes a detailed text cleaning algorithm to prepare the write-in business descriptions for modeling. Implemented in Python using regular expressions and lookup lists, this algorithm applies many natural language processing techniques (Jurafsky and Martin, 2009). For example, it converts text to lowercase, removes extraneous whitespace, and addresses numbers and common abbreviations in various ways. The algorithm includes many rules for handling compound or hyphenated words. Some compounds such as "fast-food" are concatenated to represent a single concept, whereas others are separated. Filler phrases, non-letters, and "stop" words (Jurafsky and Martin, 2009) are also removed.

To strip suffixes and reduce the number of word variations, the text cleaning algorithm applies a modified version of the Natural Language Toolkit's Porter 2/Snowball stemmer (Bird, 2006; Porter, 2001). Modifications take the form of additional rules for addressing over-stemming errors, plural nouns, and common technical suffixes in BEACON's training data. The cleaning algorithm applies final mapping rules to associate stems with other stems. These rules address under-stemming errors, lemmatize the text further (Jurafsky and Martin, 2009), and correct common misspellings.

The output of the text cleaning process is a standardized string of words (stems) separated by spaces. This clean text, in turn, serves as the input to the NAICS classification model. To illustrate different aspects of the text cleaning algorithm, Table 2 displays hypothetical write-ins and the resulting clean text. Misspellings are intentional. Note that the stemmer often converts the last letter "y" of words to "i."

2.3 Machine Learning

There are many available definitions of machine learning that BEACON satisfies. Chu and Poirier (2015, p. 1) define machine learning "as an application of artificial intelligence where available information is used through algorithms to process or assist the processing of statistical data." Roberson and Nguyen (2018, p. 1) define it as "a type of artificial intelligence that enables software applications to more precisely predict outcomes, without being programmed explicitly." BEACON clearly fits the first definition as it uses several algorithms to digest write-in business descriptions into a usable form. Likewise, BEACON is not "programmed explicitly" to map words to specific NAICS codes. Rather, BEACON learns how pieces of text in the cleaned training data are distributed across NAICS codes.

Table 3: BEACON dictionary counts.

Feature type	**Count**
Word	11,052
2-word combination	158,200
3-word combination	275,800
Exact/full-length description	62,173
Total	507,225

Disclosure Review Board (DRB) approval number: CBDRB-FY24-ESMD001-007.

Underlying BEACON is a dictionary of frequently occurring words, 2-word combinations, 3-word combinations, and exact/full-length business descriptions. The word combinations and exact/full-length descriptions are based on the concept of word co-occurrence (Figueiredo et al., 2011) and do not take word order or distance into account. All of these pieces of text serve as the model features. The dictionary is essentially a data structure that stores the features' NAICS distributions. Table 3 breaks down BEACON's dictionary by feature type. Currently, BEACON recognizes over 500,000 features.

As a machine learning application, BEACON translates short and often messy write-in business descriptions into a usable form for predictions. According to Mullainathan and Spiess (2017, p. 88), "the appeal of machine learning is that it manages to uncover generalizable patterns." While the pieces of text are the model features, BEACON must put these features into context. BEACON accomplishes this task with purity weights. A purity weight is associated with each feature and measures how concentrated, or pure, the feature's NAICS distribution is. This concept is equivalent to leaf node purity in decision trees (Tan et al., 2019). The purity weight w_f for feature f is defined by

$$w_f = \left(\frac{N_{naics}}{N_{naics} - 1}\right)\left(maxProp_f - \frac{1}{N_{naics}}\right), \tag{1}$$

where N_{naics} is the number of NAICS codes, and $maxProp_f$ is the maximum proportion in feature f's NAICS distribution. Values of w_f range from 0 (evenly distributed across NAICS codes in the training data) to 1 (occurring in only one NAICS code). Therefore, features with more concentrated distributions have more influence. Other purity weight definitions exist, but w_f defined in terms of the maximum proportion was found to work well for this problem (Dumbacher and Whitehead, 2024).

To illustrate BEACON's dictionary, Figure 3 shows the sector distributions ($N_{naics} = 20$) of four features: the word {"retail"}, the word {"bakeri"}, the 2-word combination {"retail", "bakeri"}, and the exact/full-length description exact{"retail", "bakeri"}. Only the sectors with the highest proportions are displayed: Manufacturing (sector 31–33), Wholesale Trade (sector 42), Retail Trade (sector 44–45), and Accommodation and Food Services (sector 72). For these four features, the maximum proportion $maxProp_f = 0.76$, 0.47, 0.61, and 0.64, and the purity weight $w_f = 0.75$, 0.44, 0.59, and 0.62. Observe that {"retail"} is highly associated with Retail Trade (sector 44–45). The other features are more strongly associated with Manufacturing (sector 31–33), which contains NAICS code 311811 (Retail Bakeries). In particular, the two features that take word co-occurrence into account show the greatest association with Manufacturing.

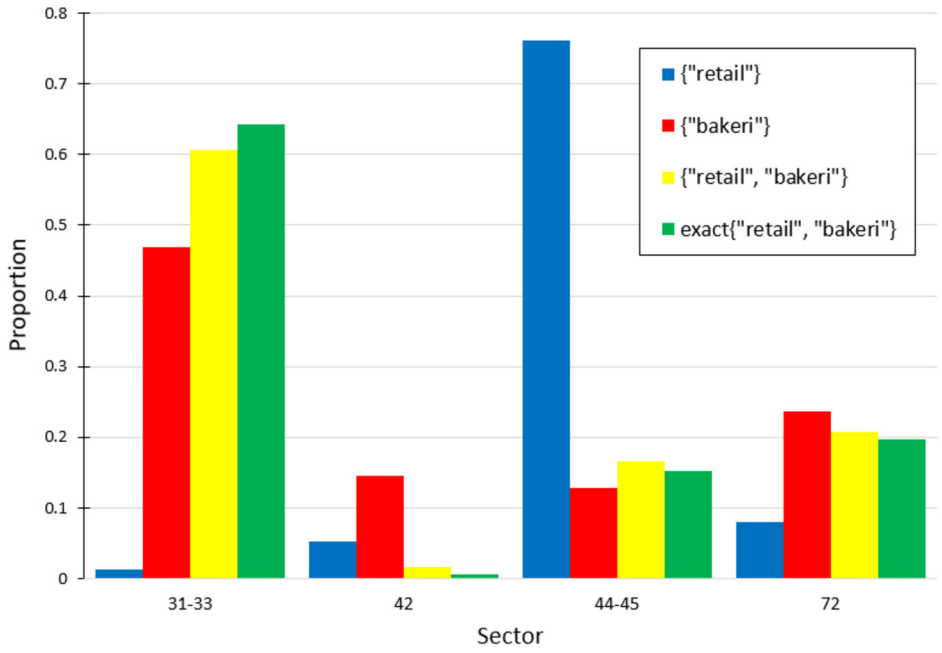

Figure 3: Sector distributions of the word {“retail”}, the word {“bakeri”}, the 2-word combination {“retail”, “bakeri”}, and the exact/full-length description exact{“retail”, “bakeri”}. Source: 2002, 2007, 2012, 2017, 2022 Economic Census; 2002–2016 Internal Revenue Service SS-4 forms; Classification Assistance Tool; and Harmonized System.

2.4 Information Retrieval

BEACON utilizes methods from information retrieval to perform ranked text classification (Aggarwal, 2018, ch. 9). This process is similar to how internet search engines operate. Given a respondent-provided business description, BEACON assigns a relevance score to each 6-digit NAICS code. The relevance score is a measure of confidence that the NAICS code is the correct industry. Generally speaking, the more highly associated the words in the business description are with the NAICS code in the training data, the higher the score. The scores are on the scale from 0 to 1, so they resemble probabilities. BEACON ranks the NAICS codes by relevance score and then returns up to ten of the highest-scoring NAICS codes.

One advantage of the information retrieval approach is that the information needed for each feature is pre-computed and stored in BEACON’s dictionary. Thus, BEACON simply retrieves the information needed to generate predictions, as will be detailed in the next section. This process occurs nearly instantaneously, allowing BEACON to be employed in real time. On a related note, in the early stages of research, the team found more traditional approaches such as logistic regression and random forests to be computationally infeasible. They could not handle the large numbers of observations, classes, and features involved. Another advantage of having a foundation based on information retrieval principles is that the resultant relevance scores are not overly sensitive to updates to the training data. As BEACON incorporates the latest available data into its training data, the ensuing rankings for common words and phrases should be relatively stable compared to procedures with more potential for unstable predictions such as k-nearest neighbors or decision trees (Hastie et al., 2009).

2.5 Model Ensemble

While the specific procedures used often vary, machine learning allows researchers to make the best use of multiple methods without being tied to a single choice (Jordan and Mitchell, 2015). This flexibility is very beneficial for most forms of data used in practice. To gain flexibility that a more traditional method may lack and to handle better the many relationships in the training data, BEACON applies multiple models to produce predictions. Such flexibility is valued by researchers from all fields as they often turn to machine learning for solutions to real-life problems (Bishop, 2013).

BEACON employs an ensemble methodology consisting of three information retrieval sub-models known as "standard," "umbrella," and "exact." These three sub-models complement one another by taking different sets of features into account. The standard sub-model considers all words and word combinations in the respondent-provided business description. This conservative approach allows every word and word combination in the description to contribute to the prediction. The umbrella sub-model, on the other hand, considers only the words and word combinations in the business description that are not subsets of other combinations. This sub-model focuses on the most detailed word co-occurrences through the umbrella-like features that cover other features. The third sub-model, exact, considers only the exact/full-length business description feature. Thus, it bases its prediction on the observations in the training data whose clean text consists of and only of the words in the description. These observations can be thought of as exactly matching the respondent's description.

The three sub-models assign relevance scores in a "term-at-a-time" manner (Aggarwal, 2018, ch. 9). Each sub-model computes a purity-weighted average of the NAICS distributions of the appropriate features, resulting in three sets of relevance scores. The ensemble score for a particular NAICS code equals the weighted average of the scores from the three sub-models, where the ensemble weights have been determined using the holdout method (Tan et al., 2019). A small fraction of the training data was randomly selected and set aside. The remaining observations were used to fit ensembles for various combinations of ensemble weights (multiples of 0.1 constrained to sum to 1). The different ensembles were then applied to and evaluated on the held-out fraction of the data. The combination of weights yielding the best results was the following: 0.1 for standard, 0.6 for umbrella, and 0.3 for exact. The ensemble places most weight on the umbrella sub-model, which, as described previously, considers the most detailed word co-occurrences.

Table 4 reports sub-model and ensemble performance on a held-out dataset that contains over 50,000 observations from the 2002, 2007, 2012, and 2017 Economic Census and covers all 20 NAICS sectors. The main metric is top-k accuracy ($k = 1$, 3, and 10), where success is defined as the true NAICS code appearing among the k highest scoring codes. Also given are the average rank of the true NAICS code (when it appears in the top ten) and the average relevance score of the true NAICS code (when it appears in the top ten). The model ensemble (with weights 0.1, 0.6, and 0.3) yields a modest improvement in accuracy over the sub-models. Although the exact sub-model underperforms the two other sub-models with regard to top-k accuracy, its inclusion does boost the performance of the ensemble. In terms of the average score of the true NAICS code, the ensemble is slightly more conservative than the umbrella and exact sub-models. However, this does not negatively affect the average rank of the true NAICS code. The exact sub-model is based on exact matches to the full-length description, so it makes sense that when it returns the true NAICS code as one of its top ten predictions, it does so with high confidence.

Table 4: Sub-model and ensemble performance on held-out dataset.

	Accuracy			Confidence	
Model	**Top-1**	**Top-3**	**Top-10**	**Avg. Rank**	**Avg. Score**
Standard	0.401	0.601	0.772	2.49	0.252
Umbrella	0.406	0.607	0.773	2.44	0.343
Exact	0.350	0.516	0.642	2.33	0.398
Ensemble	0.415	0.614	0.781	2.43	0.323

Source: 2002, 2007, 2012, 2017 Economic Census. Disclosure Review Board (DRB) approval number: CBDRB-FY24-ESMD001-007.

The following is an example that outlines how the model ensemble works at the 2-digit (sector) level. Suppose a respondent provides the write-in "This is a retail bakery." This description gets cleaned to "retail bakeri," where the features {"retail"}, {"bakeri"}, {"retail", "bakeri"}, and exact{"retail", "bakeri"} are all in BEACON's dictionary. In the next steps, note the subtle difference between the umbrella and exact sub-models.

- For the standard sub-model, retrieve the NAICS distributions of {"retail"}, {"bakeri"}, and {"retail", "bakeri"}. Then calculate a weighted average of the NAICS distributions using the features' purity weights.
- For the umbrella sub-model, the features {"retail"} and {"bakeri"} are subsets of {"retail", "bakeri"}, so they are excluded. To determine the relevance scores, retrieve the NAICS distribution of the 2-word combination {"retail", "bakeri"}. This is the distribution of observations in the training data whose clean text contains the words "retail" and "bakeri." Word order and distance do not matter.
- For the exact sub-model, retrieve the NAICS distribution of the feature exact{"retail", "bakeri"}, which corresponds to the exact/full-length description. This is the distribution of observations in the training data whose clean text consists of and only of the words "retail" and "bakeri." Word order and distance do not matter.
- To determine the ensemble relevance scores, calculate a weighted average of the relevance scores from the standard, umbrella, and exact sub-models using the ensemble weights 0.1, 0.6, and 0.3.

Figure 4 shows sector-level relevance scores produced by the standard, umbrella, exact, and ensemble models for this example. Note that only sectors with the highest scores are displayed. Influenced by {"retail"}, the standard sub-model assigns the highest score to Retail Trade (sector 44–45). The umbrella and exact sub-models, on the other hand, pick up on the co-occurrence of the words "retail" and "bakeri" and assign higher scores to Manufacturing (sector 31–33). The ensemble is swayed more by the umbrella and exact sub-models, which have a combined weight of 0.9 ($= 0.6 + 0.3$).

2.6 Hierarchical Model Structure

Assigning relevance scores directly at the 6-digit level is challenging. The approach used by BEACON takes advantage of the hierarchical structure of NAICS (see Table 1). First, BEACON uses the model ensemble to assign scores at the 2-digit (sector) level. Then, for each of the 20 sectors, BEACON uses the model ensemble again to assign sector-conditional scores to the constituent 6-digit NAICS codes. The conditional score for NAICS code *SS*#### can be interpreted as this

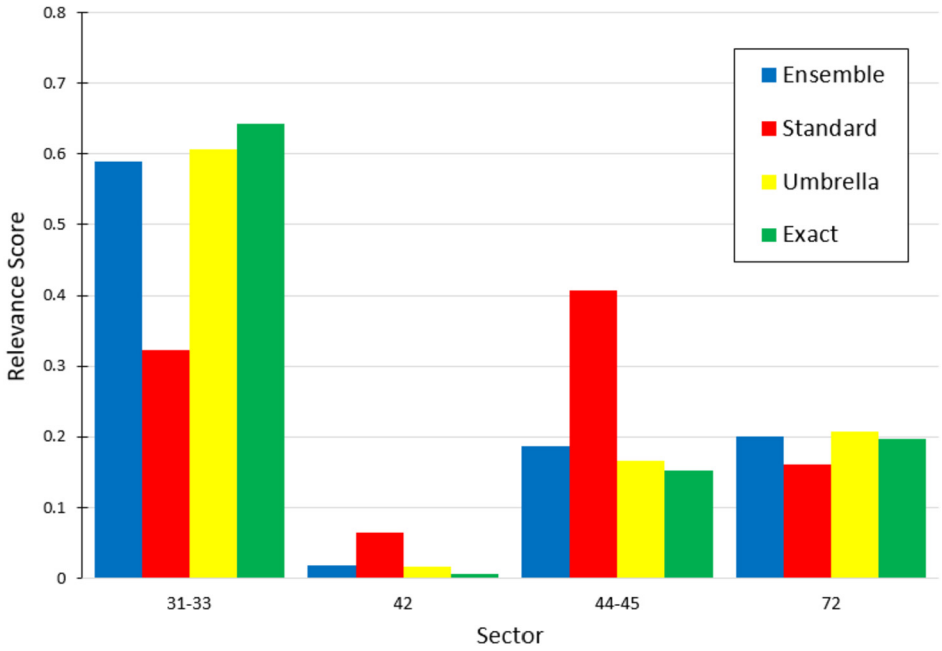

Figure 4: Sector-level relevance scores produced by the standard, umbrella, exact, and ensemble models for the write-in "This is a retail bakery." Source: 2002, 2007, 2012, 2017, 2022 Economic Census; 2002–2016 Internal Revenue Service SS-4 forms; Classification Assistance Tool; and Harmonized System.

industry's relevance score, given or assuming that the correct two-digit sector is SS.

To calculate the unconditional 6-digit scores, BEACON combines the 2-digit score and 6-digit sector-conditional scores using the conditional probability formula:

$$score\,(SS\#\#\#\#) = score\,(SS) \times score\,(SS\#\#\#\# \mid SS)\,. \tag{2}$$

This step essentially allocates the 2-digit score among the constituent 6-digit NAICS codes. In summary, the model ensemble is used 21 times in the hierarchy—once to assign scores at the 2-digit level, and 20 more times to assign sector-conditional scores at the 6-digit level.

3 Application and Results

3.1 Implementation

BEACON is programmed primarily in Python and implemented as an API. The U.S. Census Bureau's electronic survey instrument calls the API from the PBA question and displays results on the following screen. The respondent sees the NAICS description, NAICS code, and sector description for at most ten results. At this point, the respondent can select a NAICS code returned by BEACON, select "Not listed," conduct a new search, or return to the pre-list screen. Figure 5 shows what the results screen looks like for example input. The sector selected from the drop-down menu is "Retail Trade," and the write-in is "retail bakery." The highest-scoring NAICS codes from the selected sector are displayed first followed by the highest-scoring NAICS codes from outside that sector.

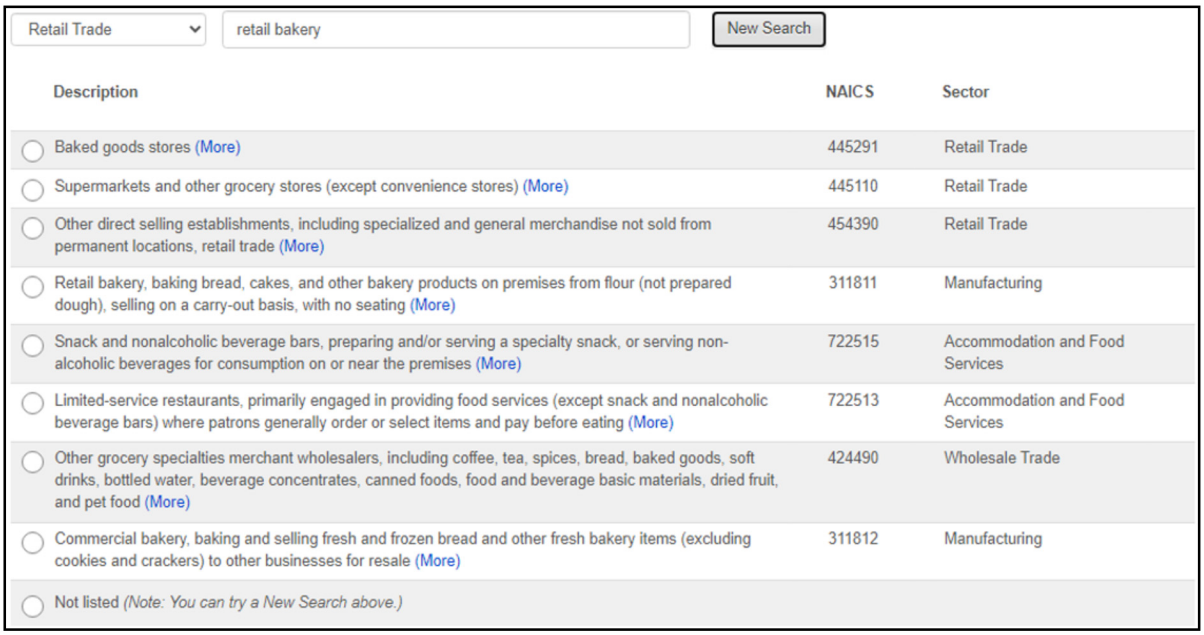

Figure 5: BEACON results screen for example input. The selected sector is "Retail Trade," and the write-in is "retail bakery." Source: 2022 Economic Census.

3.2 2021 Industry Classification Report

From October 2021 through February 2022, the U.S. Census Bureau conducted a survey called the 2021 Economic Census Industry Classification Report to obtain an updated NAICS classification for certain cases. As part of the 2022 Economic Census Pre-test, this survey also allowed the U.S. Census Bureau to assess new questionnaire features, including BEACON, with live respondents in a production environment. The sample consisted of 37,000 business establishments. By design, a third of the sample had a reliable NAICS code of record. These establishments comprised the "truth deck" and were used to evaluate BEACON's accuracy. The probability of a successful NAICS self-classification ($p_{success}$) equals the product of the probability of BEACON returning the correct NAICS code as one of its results (p_{return}) and the conditional probability of the respondent selecting the correct NAICS code ($p_{select|return}$):

$$p_{success} = p_{return} \times p_{select|return}. \tag{3}$$

Estimates of these three components of BEACON's accuracy are displayed in Table 5. They are based on the $n = 7{,}050$ respondents in the truth deck who used BEACON and selected a NAICS code from the returned results. Estimates are presented both for the overall sample and by the respondent's NAICS sector of record. Overall, BEACON returned the correct 6-digit NAICS code as one of its results with probability 0.901 and achieved a successful self-classification rate of 0.755. For the Utilities (sector 22), Finance and Insurance (sector 52), Health Care and Social Assistance (sector 62), and Accommodation and Food Services (sector 72) sectors, $\hat{p}_{success} > 0.9$. The Manufacturing (sector 31–33) and Wholesale Trade (sector 42) sectors had the lowest success rates of 0.577 and 0.549, respectively. These two sectors are historically difficult to classify (Whitehead and Dumbacher, 2023; Cuffe et al., 2022; Dumbacher and Russell, 2019). Manufacturing is particularly challenging because it represents approximately

Table 5: Estimates of different components of BEACON's accuracy, overall and by NAICS sector of record. Estimated probabilities greater than 0.9 are shaded in gray.

Sector	n	$\hat{p}_{return}$	$\hat{p}_{select\|return}$	$\hat{p}_{success}$
21	200	0.847	0.729	0.617
22	350	0.961	0.969	0.931
23	450	0.876	0.851	0.746
31–33	500	0.735	0.785	0.577
42	400	0.778	0.706	0.549
44–45	500	0.934	0.834	0.779
48–49	250	0.883	0.795	0.702
51	350	0.823	0.772	0.635
52	500	0.983	0.959	0.943
53	400	0.890	0.674	0.600
54	500	0.954	0.883	0.842
56	400	0.864	0.750	0.648
61	450	0.918	0.800	0.735
62	550	0.980	0.928	0.909
71	400	0.922	0.804	0.742
72	350	0.983	0.958	0.942
81	500	0.935	0.853	0.798
Overall	7,050	0.901	0.837	0.755

Source: 2021 Economic Census Industry Classification Report. Disclosure Review Board (DRB) approval number: CBDRB-FY24-ESMD001-007.

one third of all 6-digit NAICS codes. Interestingly, it is also one of two sectors in this analysis for which $\hat{p}_{return} < \hat{p}_{select|return}$.

Altogether, these estimates provided confidence in BEACON's ability to return the correct NAICS code as one of its highest-scoring results. In terms of improving the electronic survey instrument and helping respondents understand what the various industries represent, some NAICS descriptions were rewritten with examples and exclusions. Links to detailed entries from the online NAICS manual were also added to the BEACON results screen (see the "More" links in Figure 5).

3.3 2022 Economic Census

From October 2022 through November 2023, the U.S. Census Bureau conducted the 2022 Economic Census. BEACON was used 526,000 times by respondents. With an average response time of 0.17 seconds, the BEACON API performed very well, even during peak usage periods. The model was updated four times during data collection to learn from recently received write-ins. The statistics reported in this section take into account the version of BEACON used by the respondent.

In terms of length and detail, the write-ins are typical of those from previous Economic Censuses. The mean, median, and mode number of write-in words (sequences of non-whitespace characters) equal 3.78, 3, and 2, respectively. Owing to its large dictionary, BEACON recognized

Table 6: Most common write-ins provided by single-unit establishments.

Rank	Write-in	Rank	Write-in
1	CONSULTING	16	SOFTWARE
2	RESTAURANT	17	CONSULTING SERVICES
3	SALES	18	MANAGEMENT CONSULTING
4	MARKETING	19	ENTERTAINMENT
5	MANAGEMENT	20	CONSULTANT
6	GENERAL CONTRACTOR	21	HEALTHCARE
7	CHURCH	22	CONSTRUCTION
8	PROPERTY MANAGEMENT	23	PAINTING
9	REAL ESTATE	24	541990
10	MANAGEMENT COMPANY	25	REAL ESTATE SALES
11	BUSINESS CONSULTING	26	MEDICAL SPA
12	HOLDING COMPANY	27	FULL SERVICE RESTAURANT
13	TRUCKING	28	IT CONSULTING
14	SOFTWARE DEVELOPMENT	29	ECOMMERCE
15	EDUCATION	30	MANAGEMENT SERVICES

Source: 2022 Economic Census. Disclosure Review Board (DRB) approval number: CBDRB-FY24-ESMD001-007.

98.39 percent of the 1,597,000 word instances in the aggregate clean text. Furthermore, 96.24 percent of write-ins contained at least one letter. The remaining 3.76 percent of write-ins were predominantly NAICS codes.

Table 6 lists the most common write-ins provided by single-unit establishments. Some write-ins such as "CHURCH" and "FULL SERVICE RESTAURANT" are highly associated with one industry, but many others on the list such as "SALES" and "MANAGEMENT" are vague. In particular, there are several variations of the general description "CONSULTING." The most common NAICS code write-in provided by single units was "541990" (All Other Professional, Scientific, and Technical Services).

Figure 6 shows the distribution of the rank of the NAICS code selected from the BEACON results screen. The option "Not listed" is also included. A large majority of respondents, 82.01 percent, selected a 6-digit NAICS code returned by BEACON. When respondents did so, they tended to select from the top of the results. This part of the analysis excludes respondents who may have used BEACON, returned to the pre-list screen, and selected from among the pre-listed NAICS codes.

3.4 BEACON and Autocoder Comparative Analysis

Described in Section 1.3, the Autocoder (Kornbau, 2016; Kearney and Kornbau, 2005) is another NAICS classification tool used by the U.S. Census Bureau. Whereas BEACON helps respondents choose a NAICS code from among a ranked list of options, the Autocoder assigns a single NAICS code to new establishments based on textual administrative data. A comparative analysis between the two tools was performed to better understand their strengths and weaknesses and to discover potential areas of improvement. The analysis was based on a test dataset of over 100,000 recent IRS SS-4 write-ins for which there was a reliable NAICS code of record. Because

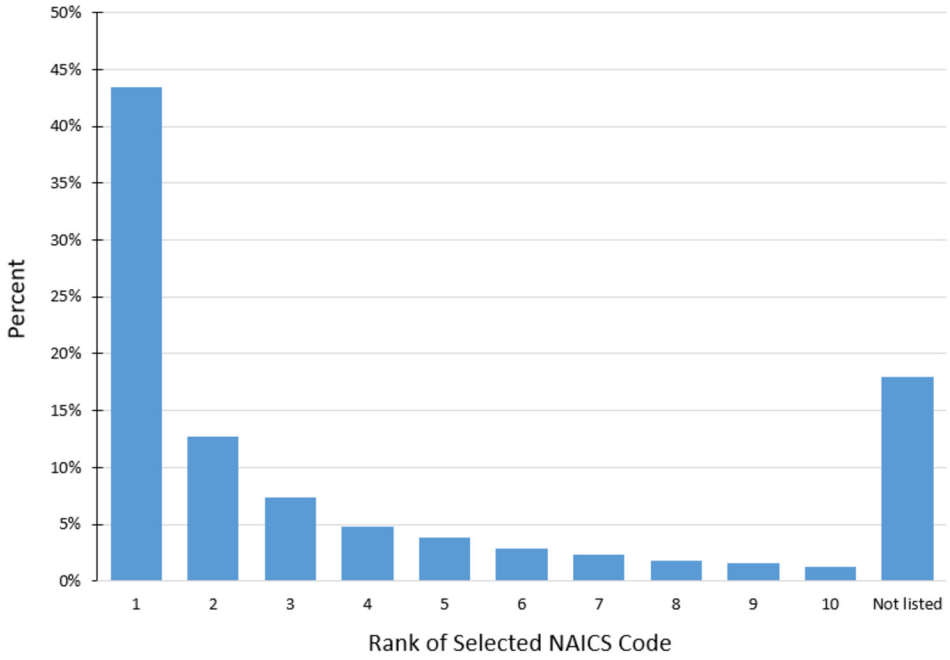

Figure 6: Distribution of the rank of the selected NAICS code. Of the respondents who used BEACON and made a selection from the results screen, 82.01 percent selected a NAICS code. The other 17.99 percent of respondents selected "Not listed." Source: 2022 Economic Census.

BEACON and the Autocoder were designed for different contexts, attempts were made for a fair comparison. For example, the Autocoder may return a less detailed NAICS code at the 2-, 3-, 4-, or 5-digit level, so some comparisons were restricted to the approximately 75,000 observations for which both methods returned a full 6-digit code. The Autocoder also uses the business name as input, so the team considered an experimental version of BEACON that incorporates the business name into the write-in. Metrics used to evaluate the methods included top-k accuracy, precision, recall, and F1 score (Tan et al., 2019).

For all methods, predictions were more reliable when write-ins contained specific words such as "law" or "dentist" versus broader terms such as "asset" or "company." Table 7 reports the top-k accuracy ($k = 1$, 3, and 10) for the Autocoder, BEACON, and the experimental version of BEACON that takes the business name into account. Overall, the Autocoder outperformed both versions of BEACON by a small margin when comparisons were made using just the top prediction from BEACON. As the Autocoder returns a single NAICS code, the team could not make valid comparisons against the full set of predictions from BEACON. Naturally, BEACON's performance improved when its additional predictions were accounted for. Including the business name did not improve BEACON's performance when only the top prediction was considered but did improve performance when all returned predictions were considered. Altogether, the analysis suggested some benefit from BEACON using the business name as an additional input and provided evidence that BEACON can assist the Autocoder with more detailed NAICS predictions at the 6-digit level.

Table 7: Model performance on recent SS-4 write-ins.

	Accuracy		
Model	**Top-1**	**Top-3**	**Top-10**
Autocoder	0.803	–	–
BEACON	0.768	0.880	0.920
BEACON (+ business name)	0.754	0.884	0.931

Source: 2017–2023 Internal Revenue Service SS-4 forms. Disclosure Review Board (DRB) approval number: CBDRB-FY24-ESMD001-007.

4 Discussion

BEACON has demonstrated that it recognizes a large vocabulary and quickly returns relevant NAICS codes to respondents. BEACON contributed to a 60 percent reduction in the number of write-ins needing to be processed post–data collection, compared to the previous 2017 Economic Census. This is a substantial decrease that has helped save clerical resources. Some analysts have even used BEACON post–data collection to aid in the assignment of more detailed NAICS codes. This application takes into account the relevance score of the 6-digit NAICS code selected by the respondent. Thorough analysis by subject matter experts regarding the accuracy of self-classified codes during the 2022 Economic Census is ongoing.

BEACON has various methodological components, and there are many ways to extend the research. In terms of improving the training data, the team is considering adding a large number of recent IRS SS-4 observations. This would increase sample and dictionary sizes for difficult sectors such as Manufacturing and Wholesale Trade (Whitehead and Dumbacher, 2023; Cuffe et al., 2022; Dumbacher and Russell, 2019). Another area of improvement is BEACON's ability to recognize Spanish write-ins. BEACON's text cleaning algorithm currently includes Spanish-to-English mapping rules to handle commonly provided Spanish words, but other rules and approaches could be researched.

In terms of improving the model, one idea is to use advanced ensemble methods such as model stacking (Whitehead and Dumbacher, 2023; Džeroski and Ženko, 2004). The current ensemble weights (0.1, 0.6, and 0.3) allow BEACON's predictions to be informed by three different sets of features but are static. Model stacking involves a "meta model" that more dynamically learns from the constituent models' relevance scores (Džeroski and Ženko, 2004; Todorovski and Džeroski, 2003). A related idea is to incorporate additional approaches into BEACON's model ensemble such as fastText (Bojanowski et al., 2017; Evans and Oyarzun, 2021). FastText is an open-source package for word embeddings and text classification based on character n-grams (Bojanowski et al., 2017). The benefits of this approach include comprehending word context better and making predictions using words not in BEACON's dictionary.

A Appendix

Table 8 lists the 20 NAICS sectors.

Table 8: NAICS sectors.

Sector	Description
11	Agriculture, Forestry, Fishing and Hunting
21	Mining, Quarrying, and Oil and Gas Extraction
22	Utilities
23	Construction
31–33	Manufacturing
42	Wholesale Trade
44–45	Retail Trade
48–49	Transportation and Warehousing
51	Information
52	Finance and Insurance
53	Real Estate and Rental and Leasing
54	Professional, Scientific, and Professional Services
55	Management of Companies and Enterprises
56	Administrative and Support and 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)
92	Public Administration

Source: U.S. Census Bureau (2024d). Disclosure Review Board (DRB) approval number: CBDRB-FY24-ESMD001-007.

Supplementary Material

The supplementary material consists of a Python program that implements a simplified version of BEACON. All of the methodological components are present, but the full text cleaning algorithm cannot be shared for confidentiality reasons. Likewise, the confidential data sources used by BEACON cannot be shared. The public data sources that are part of BEACON's training data are available at the references cited. See https://github.com/uscensusbureau/BEACON for additional files and documentation.

Acknowledgments

The authors would like to acknowledge and thank colleagues at the U.S. Census Bureau for reviewing drafts of this article and providing helpful comments. Thanks also to two anonymous reviewers for their constructive feedback, which strengthened the manuscript.

Disclaimer

Any opinions and conclusions expressed herein are those of the authors and do not reflect the views of the U.S. Census Bureau. The Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data used to produce this product [Data Management System (DMS) number: P-7504847, subproject P-7514952; Disclosure Review Board (DRB) approval number: CBDRB-FY24-ESMD001-007].

References

Aggarwal CC (2018). *Machine Learning for Text*. Springer International Publishing, Cham.

Baumgartner P, Smith A, Olmsted M, Ohse D (2021). A framework for using machine learning to support qualitative data coding. OSF Preprints. https://doi.org/10.31219/osf.io/fueyj

Bird S (2006). NLTK: The natural language toolkit. In: *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, 69–72.

Bishop CM (2013). Model-based machine learning. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984): 1–17. https://doi.org/10.1098/rsta.2012.0222

Bojanowski P, Grave E, Joulin A, Mikolov T (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5: 135–146. https://doi.org/10.1162/tacl_a_00051

Chu K, Poirier C (2015). Machine learning documentation initiative. United Nations Economic Commission for Europe. In: *Conference of European Statisticians: Workshop on the Modernisation of Statistical Production Meeting*. 15–17 April 2015, https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2015/Topic3_Canada_paper.pdf. [Online; accessed 15 March 2024].

Cuffe J, Bhattacharjee S, Etudo U, Smith JC, Basdeo N, Burbank N, et al. (2022). Using public data to generate industrial classification codes. In: *Big Data for 21st Century Economic Statistics* (K Abraham, R Jarmin, B Moyer, M Shapiro, eds.), volume 79 of *National Bureau of Economic Research: Studies in Income and Wealth*, chapter 8, 229–246. University of Chicago Press.

Dumbacher B, Russell A (2019). Using machine learning to assign North American industry classification system codes to establishments based on business description write-ins. In: *2019 Proceedings of the American Statistical Association*, 1497–1514.

Dumbacher B, Whitehead D (2022). Industry self-classification in the Economic Census. In: *2022 Proceedings of the American Statistical Association*, 1049–1064.

Dumbacher B, Whitehead D (2024). Ranked short text classification using co-occurrence features and score functions. U.S. Census Bureau ADEP Working Paper Series, (ADEP-WP-2024-06).

Džeroski S, Ženko B (2004). Is combining classifiers with stacking better than selecting the best one? *Machine Learning*, 54: 255–273. https://doi.org/10.1023/B:MACH.0000015881.36452.6e

Evans J, Oyarzun J (2021). Need for speed: Using fastText (machine learning) to code the Labour Force Survey. In: *2021 Proceedings of the Statistics Canada Symposium*.

Figueiredo F, Rocha L, Couto T, Salles T, Gonçalves MA, Meira W Jr (2011). Word co-occurrence features for text classification. *Information Systems*, 36(5): 843–858. https://doi.org/10.1016/j.is.2011.02.002

Hastie T, Tibshirani R, Friedman J (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Internal Revenue Service (2023). Form SS-4 application for Employer Identification Number. https://www.irs.gov/pub/irs-pdf/fss4.pdf. [Online; accessed 7 March 2024].

Jordan MI, Mitchell TM (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245): 255–260. https://doi.org/10.1126/science.aaa8415

Jurafsky D, Martin JH (2009). *Speech and Language Processing*. Pearson Education, Inc., Upper Saddle River.

Kearney AT, Kornbau ME (2005). An automated industry coding application for new U.S. business establishments. In: *2005 Proceedings of the American Statistical Association*, 867–874.

Kirkendall NK, White Jr GD, Citro CF, Abraham KG (2018). *Reengineering the Census Bureau's Annual Economic Surveys*. National Academies Press, Washington, DC.

Kornbau ME (2016). Automating processes for the U.S. Census Bureau register. *25th Meeting of the Wiesbaden Group on Business Registers*.

Mikolov T, Chen K, Corrado G, Dean J (2013). Efficient estimation of word representations in vector space. arXiv preprint: https://arxiv.org/abs/1301.3781

Mullainathan S, Spiess J (2017). Machine learning: An applied econometric approach. *The Journal of Economic Perspectives*, 31(2): 87–106. https://doi.org/10.1257/jep.31.2.87

Oehlert C, Schulz E, Parker A (2022). NAICS code prediction using supervised methods. *Statistics and Public Policy*, 9(1): 58–66. https://doi.org/10.1080/2330443X.2022.2033654

Oyarzun J (2018). The imitation game: An overview of a machine learning approach to code the industrial classification. In: *2018 Proceedings of the Statistics Canada Symposium*.

Porter MF (2001). Snowball: A language for stemming algorithms. http://snowball.tartarus.org/texts/introduction.html. [Online; accessed 11 March 2024].

Rizinski M, Jankov A, Sankaradas V, Pinsky E, Mishkovski I, Trajanov D (2024). Comparative analysis of NLP-based models for company classification. *Information*, 15(77): 1–32. https://doi.org/10.3390/info15020077

Roberson A, Nguyen J (2018). Comparison of machine learning algorithms to build a predictive model for classification of survey write-in responses. In: *Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research and Policy Conference*.

Roelands M, van Delden A, Windmeijer D (2018). *Classifying Businesses by Economic Activity using Web-based Text Mining*. Statistics Netherlands, Technical report.

Snijkers G, Haraldsen G, Jones J, Willimack DK (2013). *Designing and Conducting Business Surveys*. John Wiley & Sons, Inc., Hoboken.

Tan PN, Steinbach M, Karpatne A, Kumar V (2019). *Introduction to Data Mining*. Pearson Education, Inc., New York.

Tarnow-Mordi R (2017). The intelligent coder: Developing a machine-learning classification system. Methodological News. Australian Bureau of Statistics. https://www.abs.gov.au/ausstats/abs@.nsf/Previousproducts/1504.0Main%20Features5Sep%202017. [Online; accessed 8 March 2024].

Todorovski L, Džeroski S (2003). Combining classifiers with meta decision trees. *Machine Learning*, 50: 223–249. https://doi.org/10.1023/A:1021709817809

U.S. Census Bureau (2024a). Economic Census. https://www.census.gov/programs-surveys/economic-census.html. [Online; accessed 4 March 2024].

U.S. Census Bureau (2024b). Economic Census technical documentation. https://www.census.gov/programs-surveys/economic-census/technical-documentation.html. [Online; accessed 4

March 2024].

U.S. Census Bureau (2024c). Foreign trade reference codes. https://www.census.gov/foreign-trade/reference/codes/index.html. [Online; accessed 8 April 2024].

U.S. Census Bureau (2024d). North American Industry Classification System. https://www.census.gov/naics/. [Online; accessed 4 March 2024].

Whitehead D, Dumbacher B (2023). Ensemble modeling techniques for NAICS classification in the Economic Census. In: *Proceedings of the 2023 Federal Committee on Statistical Methodology (FCSM) Research and Policy Conference*.