Using Machine Learning to Assign North American Industry Classification System Codes to Establishments Based on Business Description Write-Ins

Brian Dumbacher[1], Anne Russell[1]
[1]U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233
Brian.Dumbacher@census.gov, Anne.Sigda.Russell@census.gov

Abstract
The U.S. Census Bureau classifies business establishments according to the North American Industry Classification System (NAICS). NAICS groups establishments into industries based on the activities in which they are primarily engaged. The Census Bureau uses NAICS for many purposes such as stratifying establishments for sample selection and tailoring survey questionnaires to respondents. To assign NAICS codes to establishments, the Census Bureau uses information from different sources such as the Economic Census, the Internal Revenue Service, and the Social Security Administration. Aspects of NAICS coding can be manually intensive, expensive, and time consuming and can introduce systematic errors that are difficult to diagnose. Assigning codes in a more automated way using models can address these disadvantages. In this paper, we review NAICS autocoding efforts and explore machine learning and text classification methods for assigning NAICS codes using business description write-in responses to the Economic Census. Models are trained on write-ins from the 2012 Economic Census and applied to write-ins from the 2017 Economic Census. We also discuss associated concerns and challenges.

Key Words: U.S. Census Bureau, Economic Census, North American Industry Classification System, business establishments, machine learning, text classification

1. Introduction

1.1 North American Industry Classification System (NAICS)

The North American Industry Classification System (NAICS) was implemented in 1997 and replaced the Standard Industrial Classification system, which was the Federal Government's official industry classification system since the 1930's. NAICS was developed in conjunction with Canada and Mexico to facilitate economic analyses of the three North American countries. A key use of NAICS is to provide a consistent and uniform way to present summary statistics about the U.S. economy. Also, the U.S. Census Bureau and other statistical agencies use NAICS throughout the survey life cycle including sample selection, data collection, editing, publication, and analysis of establishment data.

Disclaimer: Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau. The Census Bureau has reviewed this data product for unauthorized disclosure of confidential information and has approved the disclosure avoidance practices applied. (Approval ID: CBDRB-FY19-B00004)

NAICS classifies business establishments based on their production processes. As such, all establishments with the same or similar production process will be grouped together in NAICS. Establishments are classified based on their primary business activity. In theory, NAICS would be assigned based on the primary goods or services share of production costs and capital investment. However, in practice, other variables such as revenue or shipments are used frequently.

NAICS uses a six-digit coding scheme to identify industry classes that are organized in a hierarchical structure. The first two digits of the code represent the economic sector. There are 20 sectors across the entire economy. The third digit represents the subsector, the fourth digit represents the industry group, the fifth digit represents the NAICS industry, and the sixth digit represents the national industry. If the sixth digit is a zero, it typically means that the NAICS industry (three-country industry) and the U.S. industry are the same. NAICS is typically revised once every five years. According to the 2017 NAICS Manual (U.S. Census Bureau, 2017), the 2017 NAICS identifies 1,057 U.S. industry codes. For additional information about NAICS, see https://www.census.gov/eos/www/naics/.

1.2 Economic Census

Every five years, for years ending in "2" and "7", the Census Bureau conducts the Economic Census, an extensive survey of approximately eight million establishments with paid employees that covers most industries[1] and all geographic areas of the United States, including U.S. territories. The Economic Census asks about half of the eight million establishments to complete questionnaires whereas the other half is accounted for through data from administrative records. The Economic Census provides a wealth of information to help policymakers, trade and business associations, individual businesses, and other federal agencies understand U.S. economic activity at a granular level. Some 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 industry, as classified by NAICS. For technical details about the Economic Census design and methodology, see https://www.census.gov/programs-surveys/economic-census/technical-documentation.html.

1.3 Business Description Write-Ins

The Census Bureau sends Economic Census questionnaires to business establishments based on the most recent estimate of the establishment's NAICS code at the time of mail-out. One question on the questionnaire, the self-designated kind of business (SDKB) question, asks respondents to describe their business. This question consists of a list of checkboxes with business descriptions, and the respondent is asked to mark one box. The respondent has the option to write or type in a business description if none of the check

[1] NAICS 11 and 92 along with a few industries in other sectors are out of scope of the Economic Census. Out-of-scope codes do show up in the data for this research because of classification changes and misclassifications. For more information on industry coverage, see the "Industry Classification of Establishments" section of the Economic Census methodology.

box descriptions is accurate. To illustrate, Figure 1 is a screenshot of the SDKB question from the 2012 Economic Census pipelines questionnaire (TW-48601).

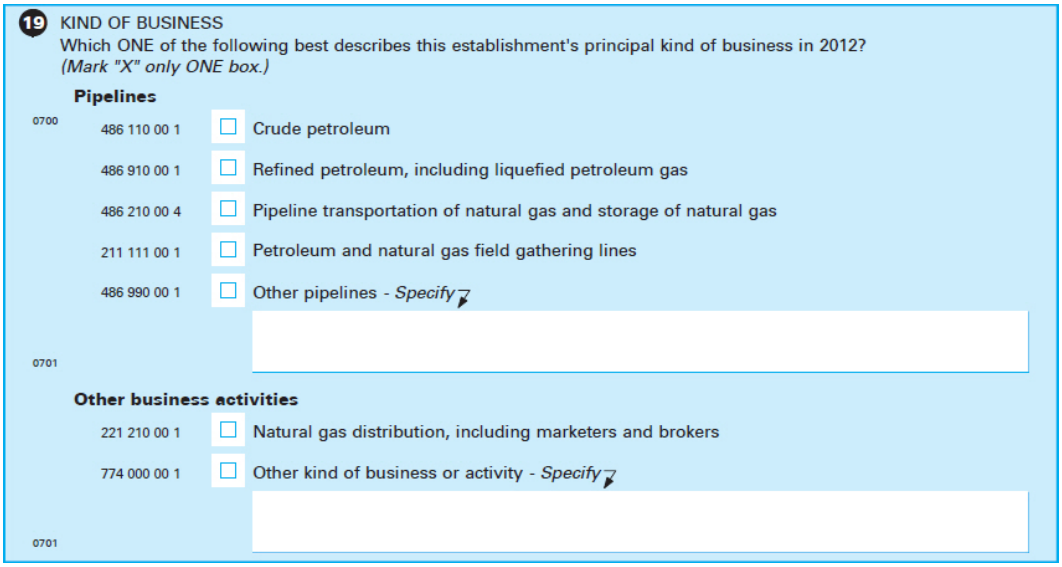

Figure 1. Self-designated kind of business question from the 2012 Economic Census pipelines questionnaire (TW-48601). Respondents can write or type in their own description. Example write-in text for this questionnaire includes "pipeline terminal" and "field office." Source: 2012 Economic Census.

For the 2012 Economic Census, there were hundreds of thousands of these so-called "write-in" responses. For the most part, clerks process and assign NAICS codes manually to these cases, which is a very resource-intensive activity. According to Snijkers *et al.* (2013, p. 478), manual coding has three key disadvantages: (1) it is expensive, (2) it is time-consuming, and (3) it can introduce systematic errors. Using an autocoder based on models or predetermined rules to assign NAICS codes can help address these disadvantages and make it easier to diagnose errors. The goal of this research is to develop a text classification model using machine learning that assigns NAICS codes to establishments based on the SDKB write-in text and other text variables from the Economic Census. We focus on predicting NAICS at the 2-digit, or economic sector, level. For context regarding how this project fits into the Census Bureau's larger efforts to use machine learning in support of its economic programs, see Dumbacher and Hanna (2017).

1.4 Outline of Paper

The rest of the paper is organized as follows. Section 2 reviews other NAICS autocoding efforts including an autocoder currently used in production to assign NAICS codes to newly identified business establishments. In Section 3, we introduce the 2012 Economic Census write-in data, which serve as the basis for model building and evaluation. Section 4 describes our machine learning and text classification methodology. In Sections 5 and 6, we evaluate models on 2012 and 2017 Economic Census data, respectively. Section 7 states conclusions, and Section 8 outlines future work.

2. NAICS Autocoding Efforts

2.1 Autocoder for New Establishments

To assign NAICS codes to business establishments, the Census Bureau uses information from different sources such as the Economic Census, Bureau of Labor Statistics, Internal Revenue Service (IRS), and Social Security Administration (SSA). Kornbau (2016, sec. 2) and Kearney and Kornbau (2005) describe how the Census Bureau, IRS, and SSA developed a NAICS autocoder for new businesses. The autocoder assigns a NAICS code using write-in text and other variables from the IRS's SS-4 form that businesses use to apply for an Employer Identification Number. The methodology uses dictionaries of words, two-word sequences, and complete write-in text from the SS-4 business name and description fields that (1) occur frequently and that (2) map a large percentage of the time to a particular NAICS code. A logistic regression model with dictionary mapping percentages as the main predictors is used to assign the NAICS code. In 2015, 79 percent of 3.6 million new business records were autocoded using this methodology, and about 69 percent of these coded records were classified to a complete 6-digit NAICS level (Kornbau, 2016, p. 3). Continual improvements and a robust quality control process have helped ensure quality autocoding over time. In this research, we borrow many elements from this successful approach.

2.2 Other Autocoding Efforts

Two other autocoding efforts were implemented more recently for the 2017 Economic Census. Both involve predetermined look-up lists. Write-in text was compared to a look-up list of 5,000 descriptions and their associated NAICS code. If there was an exact match, then the NAICS code was assigned. Approximately 69,000 write-in observations were assigned a NAICS code using this method. Similarly, write-in text was compared to another look-up list to identify text not expected to be predictive of NAICS. Many of these observations are associated with out-of-scope establishments. Examples of these so-called throw-away write-ins are "business closed," "NA," and "unknown." If there was an exact match, then the observation was flagged as unusable, and the response was treated similar to a missing response.

3. Data

3.1 Description

For this research project, we have access to a dataset of 634,473 write-in responses to the SDKB question on the 2012 Economic Census. The text-based variables in this dataset consist of the SDKB write-in, the business name, and the line label associated with the write-in text field. For example, the line labels associated with the two write-in text fields in Figure 1 are "Other pipelines – Specify" and "Other kind of business or activity – Specify." Line labels with industry-specific text such as "Other pipelines – Specify" are expected to be helpful in determining the NAICS classification. In this case, the line label serves as a proxy for the questionnaire, which, in turn, represents the establishment's estimated NAICS at the time of questionnaire mail-out. On the other hand, the generic line

label[2] "Other kind of business or activity – Specify" is not expected to be predictive of NAICS. The write-in text is the focus of this research, but we would like to research the usefulness of the two other text variables, business name and line label. The dataset also contains the NAICS code that was later assigned to the business establishment. Although this code has varying degrees of reliability, it is regarded as the true industry when evaluating the text classification models.

3.2 Preprocessing

We preprocess and pare down the 2012 Economic Census write-in dataset before applying machine learning. First of all, to avoid a large number of repeat write-ins from multiple establishments belonging to the same firm, we focus attention on single-unit establishments[3]. Observations are removed that have an invalid or missing NAICS code or that are associated with Puerto Rico (in order to avoid Spanish text). We also remove write-ins with the previously described throw-away text. Of the 634,473 write-in responses on the original dataset, 35,493 have throw-away text. Finally, we remove any duplicate observations from the same establishment.

3.3 Data Summary

After preprocessing the 2012 Economic Census data, the final dataset consists of 377,708 observations. Figure 2 breaks down this dataset by 2-digit NAICS code, which represents economic sector. The four most frequently occurring 2-digit NAICS codes are 42 (wholesale trade), 44-45 (retail trade), 54 (professional, scientific, and technical services), and 81 [other services (except public administration)]. For a complete list of 2-digit NAICS code descriptions, see Appendix A.

[2] All questionnaires with the SDKB question have this generic line label. Questionnaires for sectors 21, 23, and 31-33 did not have the SDKB question for the 2012 Economic Census. The SDKB question was added to the questionnaires for these sectors for the 2017 Economic Census.

[3] A single-unit establishment is a business establishment that makes up the entirety of its firm. In this case, the firm and establishment represent the same physical location, and these two terms can be used interchangeably.

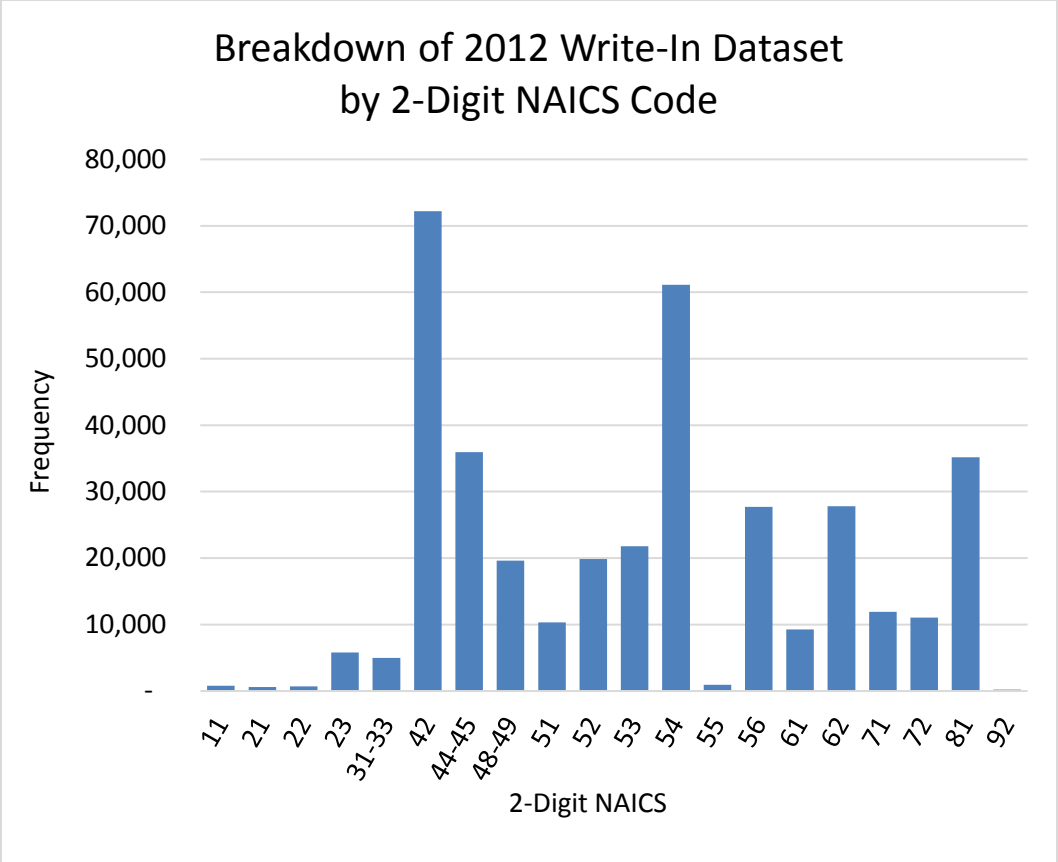

Figure 2. Breakdown of the 2012 Economic Census write-in dataset by 2-digit NAICS code (economic sector). Source: 2012 Economic Census.

4. Machine Learning Methodology

4.1 Features

The goal of this research is to develop a text classification model using machine learning that takes as input the write-in, business name, and line label text and outputs a predicted 2-digit NAICS code. To this end, we adopt a bag-of-words approach (Jurafsky and Martin, 2009, p. 641), which means the models are based on the occurrences of individual words and word sequences in the text. We consider only words and two-word sequences, or bigrams, because we have not found longer word sequences to provide appreciable predictive power given their added complexity. For each observation in the 2012 Economic Census write-in dataset, we construct a model predictor, or feature, for each word and bigram. This feature equals 1 if the text contains the word or bigram and 0 otherwise. We do this separately for the three text variables. For example, the feature for the word "retailer" in the write-in differs from the feature for the word "retailer" in the business name. Also, we consider four sets of features based on combinations of the three text variables: write-in alone; write-in and business name; write-in and line label; and write-in, business name, and line label.

Before constructing the features, we first clean, or standardize, the text variables. Standardization involves the following steps:

1. Convert the text to lowercase
2. Deal with punctuation either by deleting or converting to white space
3. Remove extraneous white space
4. Remove common English "stop" words such as articles and pronouns that are not expected to be predictive of industry

We do not deal with misspellings or employ stemming techniques[4]. In the end, every standardized write-in, business name, and line label text has the simple format of a string of words separated by spaces. Below is an example of standardizing a fictional SDKB write-in. Features are based on the individual words "waverunner", "sea", "doo", "jet", "ski", "sales", "parts", and "service" and the bigrams "waverunner sea", "sea doo", "doo jet", "jet ski", "ski sales", "sales parts", and "parts service."

Original write-in text:

```
Waverunner,Sea-Doo, and Jet  Ski sales, PARTS & service.
```

Standardized write-in text:

```
waverunner sea doo jet ski sales parts service
```

After standardizing the text variables, there remain hundreds of thousands of features for machine learning algorithms to learn from and increase overall predictive abilities. For example, the standardized SDKB write-ins alone contain 42,906 unique words and 390,065 unique bigrams. Table 1 presents the ten words and bigrams that appear in the most write-ins. Some of these words and bigrams appear to have potential predictive power such as "retail", "wholesale", "real estate", and "oil gas", which correspond directly to some economic sectors. On the other hand, general terms such as "equipment" and "business" are not expected to be as predictive. An important point is that we do not provide any input into which words and bigrams we think are predictive. The machine learning algorithms themselves determine which words and bigrams are highly associated with certain industries.

[4] Stemming involves mapping variations of words such as "manufactures" and "manufacturer" to a common concept such as "manufacturing."

Table 1. Frequently Occurring Words and Bigrams

Rank	Word	Number of Write-Ins	Bigram	Number of Write-Ins
1	services	28,345	real estate	6,593
2	sales	21,748	consulting services	3,857
3	consulting	21,545	non profit	3,239
4	service	17,708	sales service	2,625
5	management	14,416	oil gas	1,766
6	repair	11,352	management consulting	1,584
7	equipment	10,888	management services	1,304
8	retail	9,358	property management	1,292
9	wholesale	7,650	management company	1,260
10	business	7,507	retail sales	1,213

Source: 2012 Economic Census.

4.2 Learning Algorithms

In this research, we compare two commonly used learning algorithms for text classification: Bernoulli naïve Bayes (Scikit-learn, 2019a) with assumed uniform class priors to mitigate effects of class imbalance and logistic regression (Scikit-learn, 2019b) with a one-versus-rest approach to multiclass classification. Based on our experience with a bag-of-words approach to text classification, we expect the logistic regression model to perform better in terms of accuracy. However, naïve Bayes has a much faster runtime, which is important to consider given the size of the write-in dataset and the availability of computing resources.

For naïve Bayes, we optimize the smoothness parameter α, which relates to how previously unseen words and bigrams are treated when classifying new text. For logistic regression, we consider the L2 penalty and optimize the "inverse of regularization strength" parameter C, which governs model complexity. To determine more optimal values for these two parameters, we employ stratified 5-fold cross-validation with a grid search (Raschka, 2016, p. 177). The list of candidate values for α and C are (0.05, 0.1, 0.2, 0.33, 0.5, default = 1) and (0.5, 0.66, default = 1, 1.5, 2), respectively. These lists are based on results from preliminary models and cross-validation runtime considerations.

4.3 Summary of Models

In summary, we consider combinations of two learning algorithms (naïve Bayes and logistic regression), four sets of text features (based on combinations of the three text variables SDKB write-in, business name, and line label), and two parameter methods (default values or cross-validation). In total, there are 16 ($= 2 \times 4 \times 2$) models. The lists of candidate parameter values for cross-validation contain the default value, so the purpose of considering only the default value is to understand the importance of and any computational issues regarding parameter optimization. We implement the models in Python using two key modules: the Natural Language Toolkit (NLTK) for working with

text (Bird, 2006; Bird, Klein, and Loper, 2009) and scikit-learn for applying machine learning (Pedregosa *et al.*, 2011). NLTK and scikit-learn have complementary features that facilitate fitting text classification models.

5. Model Evaluation – 2012 Economic Census Data

5.1 Setup

From the 377,708 observations in the 2012 Economic Census dataset, we select a stratified simple random sample with strata defined by 2-digit NAICS code and sampling fraction equal to 0.9. The selected observations comprise the training set, and the remaining observations comprise the test set. Each model is fit using the training set and then applied to and evaluated on the test set. Other commonly used splitting proportions are 70/30 and 80/20, but we opted to include more data in the training set with a 90/10 split because of the large number of text-based features and relationships from which to learn. As stated in Hastie *et al.* (2009, p. 222), it is difficult to come up with a general rule for how large the training set should be as it depends on the complexity of the models and data. In the end, the training set has 339,936 observations, and the test set has 37,772 observations.

5.2 Results

Table 2 presents cross-validated parameter values and test set accuracies for the 16 models. The test set accuracy is simply the percentage of observations in the test set whose predicted 2-digit NAICS code agrees with the true NAICS code. As expected, given the same features, logistic regression achieves a higher accuracy than naïve Bayes. It is also apparent that cross-validation benefits naïve Bayes greatly. For example, for the naïve Bayes model with text features based on write-in, business name, and line label, the accuracy increases twenty percentage points from 0.5334 to 0.7336 when α is set to its cross-validated value of 0.1. On the other hand, cross-validation is less helpful for logistic regression. For most logistic regression models, cross-validation determines the default value of 1 to be optimal for C or results in a slightly lower test set accuracy, which is possible because the training and test sets are independent. Regarding the choice of text features, line label is more predictive than business name when used in combination with the write-in. However, for both naïve Bayes and logistic regression, the models that achieve the highest test set accuracy use features based on all three text variables – write-in, business name, and line label.

Table 2. Cross-Validated Parameter Values and 2012 Test Set Accuracies

Learning Algorithm	Text Features	Parameter and Value Method		Test Set Accuracy
Naïve Bayes	WI	Default	$\alpha = 1$	0.4657
		CV	$\alpha = 0.1$	0.6424
	WI, BN	Default	$\alpha = 1$	0.3658
		CV	$\alpha = 0.1$	0.6593
	WI, LL	Default	$\alpha = 1$	0.6340
		CV	$\alpha = 0.2$	0.7147
	WI, BN, LL	Default	$\alpha = 1$	0.5334
		CV	$\alpha = 0.1$	0.7336
Logistic Regression	WI	Default	$C = 1$	0.6457
		CV	$C = 1.5$	0.6454
	WI, BN	Default	$C = 1$	0.6866
		CV	$C = 1$	0.6866
	WI, LL	Default	$C = 1$	0.7490
		CV	$C = 1.5$	0.7483
	WI, BN, LL	Default	$C = 1$	0.7695
		CV	$C = 1.5$	0.7697

Notes: WI – write-in; BN – business name; LL – line label; CV – cross-validation. Test set accuracies greater than 70 percent are highlighted. Source: 2012 Economic Census.

The test set accuracy provides a broad measure of model performance. To understand better which observations the model is misclassifying in the test set, we organize the predicted 2-digit NAICS codes in a confusion matrix. Figure B-1 in Appendix B displays the confusion matrix for the logistic regression model with $C = 1.5$ and text features based on write-in, business name, and line label. The most misclassified observations occur between wholesale trade (42) and retail trade (44-45). This makes sense because when products are described in the write-in text, it is not always clear whether they are being sold to consumers. There is also substantial misclassification among the service sectors (in particular, 51, 52, 53, 54, 56, and 81). This indicates the model can determine that the establishment is associated with a service industry but not the details of the service.

A similar analysis for the naïve Bayes model with $\alpha = 0.1$ and text features based on write-in, business name, and line label reveals that this model does not predict any 2-digit NAICS codes to be 11, 21, 22, 55, or 92. These five sectors have the fewest observations in the 2012 write-in dataset. Even though we assumed uniform class priors to mitigate the effects of class imbalance, it appears naïve Bayes is too sensitive to imbalance in this setting. On an added note, the naïve Bayes models do fit much more quickly (less than three minutes with cross-validation) than the logistic regression models (up to 30 minutes without cross-validation and on the order of hours with cross-validation). We do prefer logistic regression but note that naïve Bayes performs decently given its speed. Future research could involve using the NAICS code prediction from a naïve Bayes model as a feature in other models.

6. Model Evaluation – 2017 Economic Census Data

6.1 Setup

We would like to see how the models perform on more recent 2017 Economic Census data. To this end, we fit the models from Section 5 using all 377,708 observations from the 2012 Economic Census dataset and then apply them to single-unit write-in observations pulled from the 2017 Economic Census database. The same criteria for paring down the 2012 dataset are applied to the 2017 dataset. At the time of this research, data collection and processing for the 2017 Economic Census were ongoing. On May 30, 2019, we accessed the database and created a dataset consisting of 226,124 write-in observations. This dataset contains the three text variables – SDKB write-in, business name, and line label – and a NAICS code that can be regarded as the truth with the caveat that the 2017 data are mid-review. As with the 2012 data, this NAICS code has varying degrees of reliability. Figure 3 breaks down the 2017 dataset by 2-digit NAICS code. Note that there are no observations for NAICS 2-digit codes 11 (agriculture, forestry, fishing and hunting) and 92 (public administration). The two dominant 2-digit NAICS codes are again 42 (wholesale trade) and 54 (professional, scientific, and technical services). However, other 2-digit NAICS codes such as 23 (construction) and 31-33 (manufacturing) occur more frequently, in relative terms, than in the 2012 dataset.

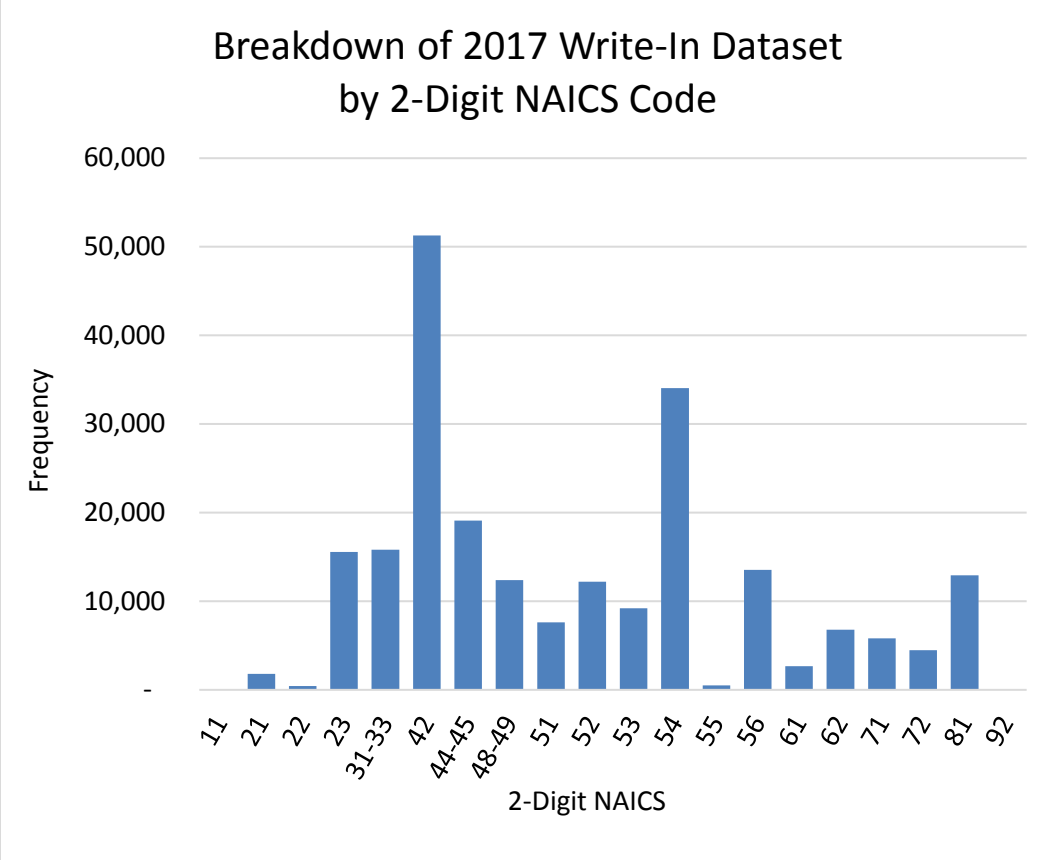

Figure 3. Breakdown of the 2017 Economic Census write-in dataset by 2-digit NAICS code (economic sector). Source: 2017 Economic Census.

6.2 Results

Table 3 presents accuracies on the 2017 dataset for eight combinations of learning algorithm and text features. The parameter values are set to the same cross-validated values from Section 5. The logistic regression model with C = 1.5 and text features based on write-in, business name, and line label achieves a very low accuracy of 0.4387 on the 2017 Economic Census dataset. Figure B-2 in Appendix B displays the confusion matrix for this model. There is a high rate of misclassification among observations in the wholesale trade sector (42). In general, many observations are being misclassified as other services (except public administration) (81). A key problem with using line label as a text variable for these mid-review 2017 data is the high use of the write-in text field with generic line label "Other principal business or activity – Describe," which is not predictive of NAICS. Other things to explore are differences in line label wording between the 2012 and 2017 questionnaires and the effect of NAICS reliability on results. Considering just write-in and business name as text features, the logistic regression model with C = 1 achieves a higher accuracy of 0.6118. This is still lower than the accuracy of 0.6866 achieved by the corresponding model that was fit using and applied to 2012 Economic Census data.

Table 3. Model Accuracies on the 2017 Economic Census Dataset

Learning Algorithm	Text Features	Parameter Method and Value		Accuracy
Naïve Bayes	WI	CV	α = 0.1	0.5355
	WI, BN	CV	α = 0.1	0.5686
	WI, LL	CV	α = 0.2	0.4030
	WI, BN, LL	CV	α = 0.1	0.4774
Logistic Regression	WI	CV	C = 1.5	0.5620
	WI, BN	CV	C = 1	0.6118
	WI, LL	CV	C = 1.5	0.4059
	WI, BN, LL	CV	C = 1.5	0.4387

Notes: WI – write-in; BN – business name; LL – line label; CV – cross-validation. Accuracies greater than 50 percent are highlighted. Source: 2012 and 2017 Economic Census.

7. Conclusions

We consider two commonly used machine learning algorithms for text classification – naïve Bayes and logistic regression – in order to predict a business establishment's industry at the 2-digit NAICS, or economic sector, level. Using 2012 Economic Census data and restricting attention to single-unit establishments, we find that the best performing models use features based on all three text variables – write-in, business name, and line label. Business name appears to be the least predictive of the three. Unlike naïve Bayes, logistic regression yields predicted 2-digit NAICS codes that represent all 20 sectors of the economy. The best performing logistic regression model achieves a decent test set accuracy above 76 percent. Upon inspection of misclassified observations, it is seen that this model has difficulty distinguishing between retail and wholesale and among the service sectors. Applying these models to mid-review 2017 Economic Census data, which are five

years removed from the 2012 training data, results in lower accuracies. Line label is not as predictive for 2017 possibly because of differences in questionnaire wording between 2012 and 2017 and fewer labels having industry-specific text.

8. Future Work

There are many directions in which to continue this research. First of all, there are more advanced machine learning algorithms that can be brought to bear on the write-in data such as decision trees, random forests, neural networks, and ensemble methods. These methods represent very different approaches to text classification, and it would be interesting to compare them with naïve Bayes and logistic regression. We would also like to research how best to predict industry at a more detailed NAICS level. One approach might be to build a separate model for each economic sector. Another approach could involve hierarchical modeling (Silla and Freitas, 2011). It would be useful to see how model performance changes as the level of detail of the predictions increases.

Furthermore, we want to explore more fully the impact of using the line label in the model. Specifically, a key question is why the line label improved model performance in 2012 but decreases model performance significantly in 2017. One thought is that in 2012 the line label served more as a proxy for the questionnaire but because of questionnaire and wording changes, the same relationships do not exist in 2017. We will explore using other variables such as mail-out NAICS sector to see if they result in a more consistent and accurate model for predicting NAICS for 2012 and 2017.

We also have plans to combine data from the 2012 and 2017 Economic Census (after more 2017 Economic Census observations have been processed) to create a larger and richer training set. This will help for write-ins associated with the three sectors that did not have the SDKB question in 2012 and that contributed limited data to the training set. In addition, this is likely to help with predicting NAICS accurately at a greater level of detail. An area of concern is the decline in predictive power from one Economic Census to the next even after removing line label. One solution we want to pursue is to augment the training set with more frequent and recent data sources such as descriptions from the SS-4 form to capture changes in the economy and emerging industries.

This research focuses on using text variables from the Economic Census, but there exist non-text variables that could serve as key predictors. As mentioned in Section 5.2 and in Dumbacher and Hanna (2017), one challenge is to figure out how best to incorporate other NAICS predictions. Two examples are the predicted 2-digit NAICS code from a naïve Bayes model, which can be obtained relatively fast, and the estimate of the establishment's 2-digit NAICS code at the time of questionnaire mail-out. These codes could serve as model predictors or stratification variables for fitting separate models by economic sector. Lastly, on the Economic Census questionnaire, there exist multiple "Class of Customer" questions that ask respondents about the customers to whom the establishment sells its goods (for example, consumers, retailers, wholesalers, and distributors) and the

corresponding percentage breakdown of the establishment's sales. This information could help models distinguish between retail and wholesale, a problem area that we identified for the best performing logistic regression model.

Acknowledgments

The authors would like to thank Matthew Thompson, William Samples, Justin Nguyen, Javier Miranda, and William Davie Jr. of the U.S. Census Bureau for reviewing drafts of this paper and providing helpful comments. Special thanks to Michael Kornbau for his input as well and for creating the 2012 Economic Census write-in dataset.

References

Bird, S. (2006). NLTK: The Natural Language Toolkit. *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*. Sydney, Australia: Association for Computational Linguistics, 69–72.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Beijing, China: O'Reilly Media, Inc.

Dumbacher, B. and Hanna, D. (2017). Using Passive Data Collection, System-to-System Data Collection, and Machine Learning to Improve Economic Surveys. *2017 Proceedings of the American Statistical Association, Business and Economic Statistics Section*. Alexandria, VA: American Statistical Association, 772–785.

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

Jurafsky, D. and Martin, J.H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (Second Edition). Upper Saddle River, NJ: Pearson Education, Inc.

Kearney, A.T. and Kornbau, M.E. (2005). An Automated Industry Coding Application for New U.S. Business Establishments. *2005 Proceedings of the American Statistical Association, Business and Economic Statistics Section*. Alexandria, VA: American Statistical Association, 867–874.

Kornbau, M.E. (2016). Automating Processes for the U.S. Census Business Register. *25th Meeting of the Wiesbaden Group on Business Registers*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Raschka, S. (2016). *Python Machine Learning*. Birmingham, United Kingdom: Packt Publishing.

Scikit-learn. (2019a). Naive Bayes: Bernoulli Naive Bayes. <https://scikit-learn.org/stable/modules/naive_bayes.html#bernoulli-naive-bayes>. Accessed April 10, 2019.

Scikit-learn. (2019b). Generalized Linear Models: Logistic regression. <https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression>. Accessed April 10, 2019.

Silla, C.N. and Freitas, A.A. (2011). A Survey of Hierarchical Classification across Different Application Domains. *Data Mining and Knowledge Discovery*, *22*(1-2), 31–72.

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

U.S. Census Bureau. (2017). 2017 North American Industry Classification System Manual. <https://www.census.gov/eos/www/naics/2017NAICS/2017_NAICS_Manual.pdf>. Accessed April 24, 2019.

Appendix A: Descriptions of 2-Digit NAICS Codes

The first two digits of the NAICS code represent the economic sector. There are 20 sectors across the entire economy. Table A-1 describes these 2-digit NAICS codes.

Table A-1. Descriptions of 2-Digit NAICS Codes

2-Digit NAICS	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 Technical 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 (2017).

Appendix B: Confusion Matrices

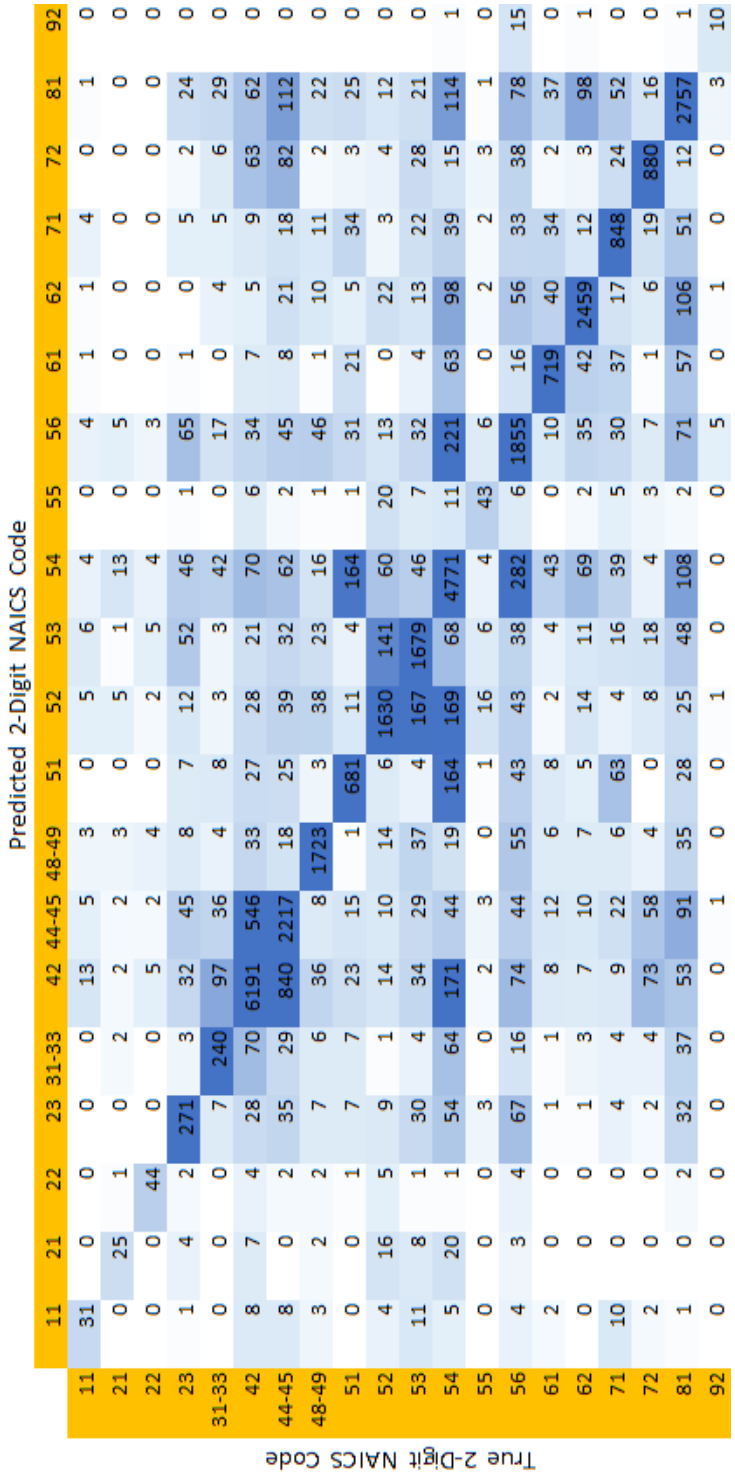

Figure B-1. Confusion matrix for the logistic regression model with C = 1.5 and text features based on write-in, business name, and line label fit using the 2012 Economic Census training set and applied to the 2012 test set. Source: 2012 Economic Census.

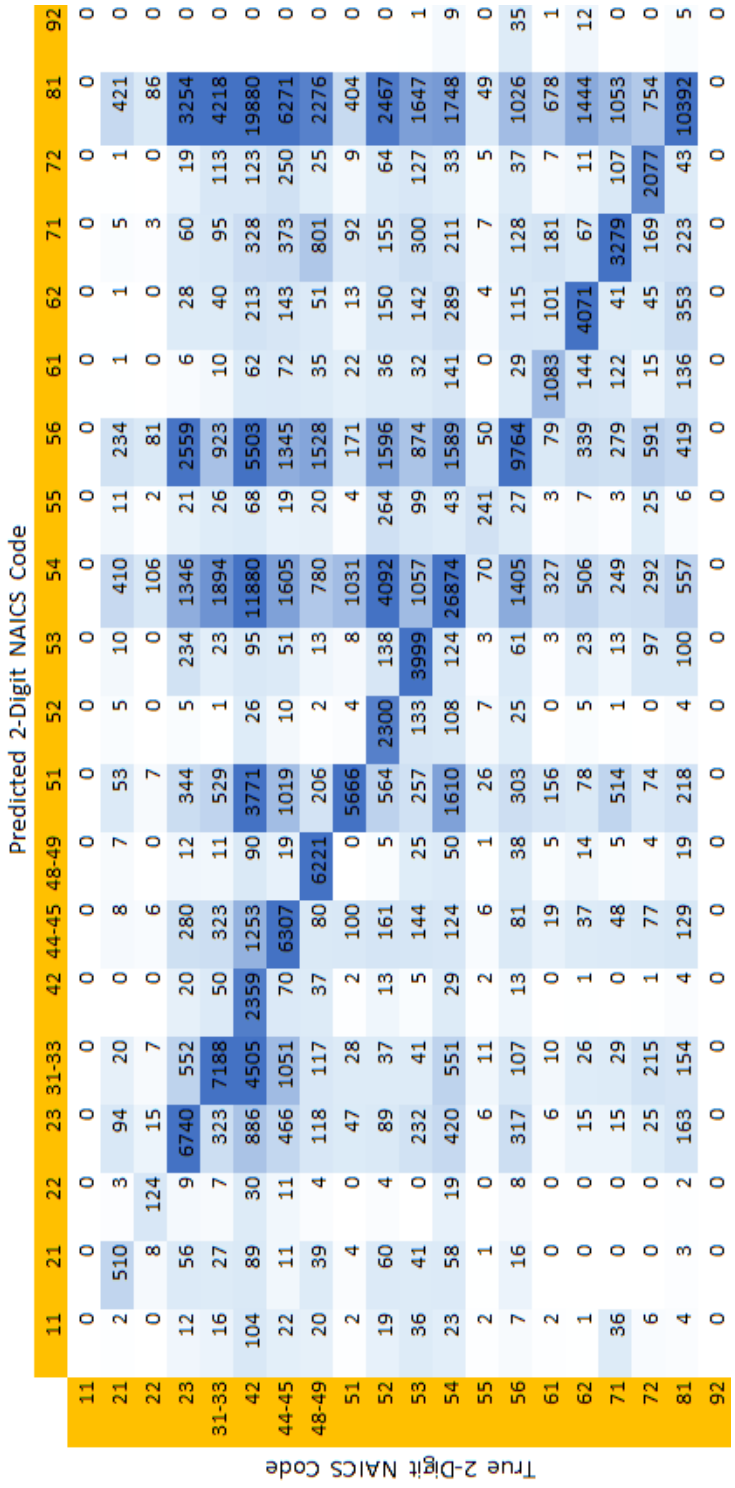

Figure B-2. Confusion matrix for the logistic regression model with C = 1.5 and text features based on write-in, business name, and line label fit using all of the 2012 Economic Census data and applied to the 2017 Economic Census dataset. Source: 2012 and 2017 Economic Census.