Using Public Data to Generate Industrial Classification Codes*

John Cuffe[†] Bhattacharjee, Sudip [‡] Edudo, Ugochukwu [§]
Smith, Justin C [¶] Basdeo, Nevada [‖]

March 6, 2019

Draft. Do not cite or circulate without permission.

Abstract

The North American Industrial Classification System (NAICS) is the system by which multiple federal and international statistical agencies assign business establishments into industries. Generating these codes may be a costly enterprise, and the variety of data sources used across federal agencies leads to disagreement over the "true" classification of establishments. In this paper, we propose an improvement to the generation of these codes that could improve the quality of these codes and the efficiency of the generation process. The NAICS codes serve as a basis for survey frames and published economic statistics. In the current state, multiple statistical agencies and bureaus generate their own codes (e.g. Census Bureau, Bureau of Labor Statistics (BLS), and Social Security Administration) which can introduce inconsistencies across datasets housed at different agencies. For example, the business list comparison project undertaken by BLS and the Census Bureau found differences in classification even for single-unit establishments (Fairman et al., 2008; Foster et al., 2006). We propose that combining publicly available data and modern machine learning techniques can improve accuracy and timeliness of Census data products while also reducing costs. Using an initial sample of approximately 1.3 million businesses gathered from public APIs,

*Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed, DRB approval CBDRB-FY19-191. We thank ...s.

[†]Mojo Development Team, U.S. Census Bureau, (corresponding author) john.cuffe@census.gov
[‡]Center for Optimization and Data Science, U.S. Census Bureau
[§]University of Connecticut
[¶]Center for Optimization and Data Science, U.S. Census Bureau
[‖]Center for Optimization and Data Science, U.S. Census Bureau

we use user reviews and website information to accurately predict two-digit NAICS codes in approximately 59% of cases. Our approach may have some merit, however substantial methodological and possible privacy issues remain before statistical agencies can implement such a system.

1 Introduction

The North American Industrial Classification System (NAICS) is the system by which multiple federal and international statistical agencies assign business establishments into industrial sectors or classes. Both economic statistics such as the Business Dynamics Statistics (Haltiwanger et al., 2008) as well as survey sampling frames rely on timely and accurate industrial classification data. Despite the obvious need for this information, NAICS codes remain expensive to produce and often the same establishment will receive a different NAICS code from different statistical agencies. In this paper, we propose a solution to these issues that relies on publicly available data about businesses to generate NAICS codes. Specifically, we seek to combine web scraping with textual analysis, user reviews, and website text from approximately 120,000 single-unit employer establishments to generate NAICS sector classifications. Our approach shows that public data may be a useful tool for generating NAICS codes, but there are substantial challenges to any agency implementing such a system, and public data alone will not immediately replace current administrative and Census data collected through surveys. The paper proceeds as follows: first we will highlight the business issues with current methods, before discussing new methods being used to generate industrial and occupational classifications in statistical agencies in several countries. Then, we discuss our approach, combining web scraping with modern machine learning techniques to provide a low-cost alternative to current methods. Finally, we discuss our findings in the context of the Census Bureau's current capabilities and limitations.

2 Current Methods

Currently, NAICS codes are produced by multiple statistical agencies: The Census Bureau produces classifications through multiple surveys, most notably the Economic Census. The Bureau of Labor Statistics (BLS) generates and uses NAICS codes in its surveys, and the Social Security Administration (SSA) produces codes for newly established businesses via information on the SS4 Application for Employee Identification Number form. One would believe that the standardized list of industries would mean the agencies were able to coordinate their lists and ensure all establishments receive the same code, however this is not the case. Figure 1 below shows the percentage of agreement, on the two digit level, between NAICS codes produced by the 2012 Economic Census, the BLS, and the SSA for the same set of single-unit establishments active in 2012. It shows that the Census and BLS, when coding the same cases, only agree on the NAICS *sector* in approximately 86% of cases, whereas the BLS and SSA only concur in around 70% of cases. The central causes of this disagreement

are results from agencies using different data and different methods, even for the same sample. Initial discussions of the NAICS system opined for the production of standardized codes (Committe), and yet a quarter of a century later these issues remain. Further, NAICS codes require a substantial investment in both time and manpower: Census QA standards require a hand-review of 10,000 cases per quarter, and the SSA's currently methodology leaves over 10% of codes requiring hand-coding by subject matter experts. Even assuming 1 minute per case, this requires over 4 *weeks* of employee time to complete.

FIGURE 1 HERE

Thus, multiple federal agencies produce potentially drastically different "standard" codes, each at great expense. We propose easing these burdens by using publicly available data from businesses as a way to generate NAICS codes. Generally, an industrial classification for an establishment will need to take into account the "products it produces, processes, or sells....It may also be necessary to know how the products will be used and whether they are custom used for particular clients." (Jab, 1984) Gathering this information in a survey would be particularly burdensome to each respondent, however many businesses will have this information readily available in public forums. Specifically, public reviews of businesses will describe the products consumers purchased, and company websites will be tailored to highlight product offerings. Using these sources of data provides four specific advantages: First, much of this information is available through free or relatively low-cost APIs such as Google and Yelp, and website information exists in the public domain. This makes it easier for statistical agencies to share data used to generate NAICS codes, ensuring greater agreement. Secondly, this approach will allow the Census to provide more timely data, with searches and implementation of data modeling occurring far faster than traditional surveys. Thirdly, this approach will lower respondent burden on surveys by allowing the Census to forgo asking questions relating to industrial classification on survey forms. Finally, this approach allows for comparable industrial codes across the various statistical agencies. Currently statistical agencies may face legal barriers to sharing the source data for their industrial classifications (e.g. Title 13, Title 26 USC restrict how data may be shared across Bureaus). Using public data avoids these issues, allowing statistical agencies to make direct comparisons of competing models and assumptions, improving overall data quality.

Several statistical agencies, both in the U.S.A. and internationally, have attempted to use textual data as a means for classification. Much of the work has focused on the use of generating occupational classifications based on write-in survey responses (Gweon et al.,

2017; Jung et al., 2008; Fairman et al., 2012)[e.g.], however there are notable attempts to generate classifications of businesses. The British Office for National Statistics has attempted to use public textual information on companies to generate unsupervised classifications of industries (Office for National Statistics), with meaningful industrial clusters through a combination of Doc2Vec and Singular Value Decomposition models, however the data were fit on a "relatively small" (Office for National Statistics) number of observations, leaving the usefulness of the method at much more fine-grained levels unknown. These methods were similarly deployed by researchers at the Italian National Institute of Statistics to generate a classifier to distinguish non-profit institutions from the business universe. Researchers from the National Statistics Netherlands explored how to generate industrial classifications similar to NAICS codes using Term Frequency-Inverse Document Frequency (TF-IDF) and dictionary-based feature selections via Naive Bayes, Support Vector Machine, and Random Forest classifiers, finding three main complicating factors for classification: the size of the businesses, the source of the industrial code, and the complexity of the business' website (Roelands et al., 2017). Finally, the Australian Bureau of Statistics implemented a system that generates classifications based on short, free text responses into classification hierarchies based on a bag of words, one-hot encoding approach. This approach has the advantage of simplicity–for each word in the vocabulary, a record receives a "1" if its response contains that word, and a zero otherwise. However, this approach also ignores the context of words, a possible issue when seeking to distinguish closely related industries (Tarnow-Mordi). In the U.S. statistical system, Kearney and Kornbau (2005) produced the Social Security Administrations "Autocoder", a system that uses word dictionaries and predicted NAICS codes based on open-response text on IRS Form SS4, the application for an EIN. The Autocoder, first developed in the early 2000s, remains in service, and relies on a combination of logistic regression and subject-matter experts both for QA and for manual coding tasks.

We seek to build on this work in the context of 2-digit NAICS sectors for a sample of single-unit, employer businesses active in 2015 and/or 2016. Our approach builds on those above, by combining web-scraping of company websites, company names, and user reviews to generate Doc2Vec methods to reduce the dimensionality of the data in a similar manner to the previous attempts (Roelands et al., 2017; Tarnow-Mordi, e.g.). Finally, we use the outputs of this textual analysis as inputs into a Random Forest classifier, seeking to identify 2-digit NAICS By doing so, we hope to provide another example of uses of "Big Data" in statistical terms, and in particular demonstrate an avenue that the Census Bureau and other statistical agencies may provide more timely, more accurate industrial classifications, lower respondent burden, and generate substantial cost savings over current methods. The next section gives an overview of our methodological choices, before describing our dataset in

more details.

3 Data and Methods

Our approach combines publicly available data from company websites and user-generated reviews of businesses with Census Bureau protected information on individual business establishments. We first utilized public APIs to generate a target sample of approximately 1.3 million business establishments, matched those records to the Business Register by name and address, and then analyzed available text in user reviews, the company website, and company name to reduce the dimensionality of the data, and finally use these outputs as features (independent variables) in a Random Forest classifier to predict 2-digit NAICS codes. Here we give a brief overview of each stage of our approach before summarizing how our dataset matches up to the universe of single-unit employer businesses.

3.1 Web Scraping via APIs

An Application Program Interface, or API, is a web-based application that allows users to access the "building blocks" of information on websites. For example, the Google Places API used to gather our data allows access to business information such as name, address, rating, opening hours, user reviews, website links, and contact information. We leverage this information in two ways. Firstly, public user reviews provide a rich source of information about a business. For our purposes, we would be interested to know what kinds of products users describe in their reviews–multiple reviews on the quality of steak from an establishment increases the likelihood the business is a restaurant versus a manufacturing plant. Secondly, we use the linked website (when available) to gather the HTML text, with the logic that the homepage of the website is a place that the business is seeking to give a clear and directed message to potential customers on what products they offer. Next, we use Google Types Tags, a list of over 100 different classification tags assigned by Google. These tags vary in use, as they include words like 'establishment' or 'point of interest' which will not aid in classification, but also words such as "hotel", "bar", or even "Hindu Temple", which would greatly aid a model in classifying a business. Finally, we also use the name of the company, as company names often indicate the type of products on offer (e.g. Krusty Burger). Together, these four sources provide us with the exact type of information needed to describe a business, what products it may sell, and how its customers use or perceive those products (Jab, 1984).

To generate our sample of businesses, we conducted a grid search on both the Yelp and Google Places APIs, based on a combination of a lat/long coordinate and keyword. Geogra-

phies were sampled by first locating the centroid of each Core-Based Statistical Area (CBSA) and each county therein. This geographical search pattern will certainly mean that businesses not residing in a CBSA, or any industries that are more common in rural areas, may be under-sampled. This concern is not without merit–as discussed below, industries more common in rural areas (e.g. farming, mining) are heavily under-sampled when we match to the BR. Further research is seeking to rectify this bias. To identify keywords, we found all words contained in the titles of all two-digit NAICS sector (https://www.census.gov/eos/www/naics/). We then executed an API search for each keyword in 50 random locations around the centroids provided above, with a set search radius of 10km. This resulted in 1,272,000 records, with approximately 70% of those coming from the Yelp API. Next, we used the Google Places API to search for each business directly, retrieving any website text or user reviews, as well as other information such as Google Types tags. The sample we discuss below is the subset of these 1,272,000 million records–records contained at least one user review AND had a working URL linked through the API. These restrictions make our initial modeling simpler, but reduced the number of available establishments to approximately 290,000. For this initial exploratory study, we chose to eliminate records that did not have a website/user reviews to have the best sample to determine the overall utility of both sources of data, however further research will attempt to generate NAICS codes for establishments that lack either a website or user reviews.

3.2 Matching to the Business Register

The Business Register is the Census Bureau's comprehensive database of U.S. business establishments and companies, covering all employer and non-employer businesses (see:). In order to identify if these records appear in the Business Register, we utilized the Multiple Algorithm Matching for Better Analytics (MAMBA) software (Cuffe and Goldschlag, 2018). This software utilizes machine learning techniques to link records based on name and address. MAMBA provides high-quality matches, but also provides us with match metrics so we may identify quality matches over more tenuous linkages. In order to reduce the possibility of spurious linkages, we required that any matched pair must have either a 5-digit zip code, city name, or 3-digit zip code in common. We ran two particular matches–the first matching on both name and address, and then a residual match matching by only business name. The training data used for MAMBA may be prone to Type 2 errors, however given this initial study, we argue that accepting a small number of false-negatives in our matcher is a price worth paying to ensure our overall findings are based on high-quality matches.

An additional complexity arising during this matching process that we account for is the

possibility that a single BR record may be the "best" match for multiple input records. This would impact our model performance by potentially providing the same website/review data to two different NAICS codes. For example, a BR record with the name "Bob's Burgers" may match to Google records for "Bob's Burgers" but also "Bob Burgerson, Inc". To account for this possibility, we identified the best BR match for each Google record *that was not a better match for another Google record.* This ensures our final dataset is a set of unique pairs, however the impact of these assumptions on our final model requires further investigation. Finally, to ensure more stable estimates, we exclude several NAICS codes that returned fewer than 1000 matches, including the Agriculture, Mining, and Mangement sectors. This decision does impact the applicability of our findings, as these sectors provide valuable contributions to the economy. However, as we discuss below, our model has a distinct performance advantage in NAICS codes with larger numbers of observations, and future work will attempt to tune models for smaller observations.

In total, we identify 120,000 single-unit establishments that have both website and review text through the MAMBA process, accounting for only 43.44% of our businesses gathered from Google. While this match rate may seem poor, we argue this is not a symptom of poor matching or data but the result of four circumstances. First, the initial scraped occurred in December 2017/January 2018, whereas the Business Register data used to identify matches is from 2015/2016. Thus, in some cases the BR is almost two years behind Google. In some industries this is a substantial burden: approximately 19% of all server-providing (e.g. NAICS code 41 or higher) businesses fail within their first year of operation(Luo and Stark, 2014, p. 11) meaning many BR businesses may no longer exist in the Google database. Secondly, many Google records may not exist in the Business Register. The Census Bureau estimated that approximately 350,000 businesses would form after 2016Q3 (Bayard et al., 2018b,a)[1] and before we initiated our search, meaning any of these businesses may appear in Google but would not appear as an employer business in the Census data. Third, we only measure *single-unit* employer businesses. We chose to only analyze single-establishment firms in this paper due to complicated nature of assigned industrial codes to multi-unit firms: for example, large retail stores may have storefronts (NAICS 42), warehouses (48-49), and corporate headquarters (55), all pointing to the same website with similar user reviews, making identification using our methods problematic. However, given the Google API sorts results by importance and search popularity, it seems logical to assume many of the businesses we scrape actually belong to multi-unit firms, as these firms are more

[1]The Business Register defines a business as an employer business if it has payroll on March 12 of a given year. So by measuring from 2016Q3, we are account for any formations after this period. Figure sourced by taking the number of expected business formations for 2016Q3, 2016Q4, 2017Q1, and then multiplying 2017Qs 2-4 by the proportion of quarters remaining in the year.

likely to be searched (and thus be placed higher in subsequent results). In fact, we match approximately 84,000 records found in our Google search to *multi-unit* firms,[2] taking our match rate to employer businesses (either SU or MU) to approximately 70%. Fourth, we only account for businesses listed as employers in either the 2015 or 2016 Business Register, meaning non-employer businesses are not included in our sample. While larger, employer firms may appear more prominently in search results, our analysis shows that the Google data do contain non-employer businesses.

3.3 Matched Data Quality

Figure 2 below shows the percentage in our sample (upper bar) and the BR single-unit employer universe (lower bar) in each NAICS sector. The figure reveals that the scraped sample heavily over-samples NAICS 44/45 (Retail Trade) and 72 (Accommodation and Food Services). Approximately 12.28% of all BR single-unit employers fall into the Retail Trade sector, however this sector makes up almost 19% of the scraped sample. This heavy over-sample is expected; about 2/3rds of our sample was sourced from Yelp, which dominated by food services, and in general Google Places and Yelp both target these public-facing industries in their APIs. On the other hand, our approach badly *under-samples* NAICS code 54, Professional, Scientific, and Technical Services. This sector makes up about 12.6% of the universe of businesses, but only 4.36% in our sample. Additionally, our sample also badly under-samples the Construction and Agriculture, Forestry, and Mining sectors relative to their size in the broader economy. Discussed further below, the variation and severity of these sampling errors raise valid questions on the use of these methods to generate a universe of businesses at the NAICS sector level, let alone at more detailed industrial classifications.

FIGURE 2 HERE

Another question is whether our method gathers information more effectively in different geographical areas. Although our initial search pattern was devised to guarantee coverage of all CBSAs, some businesses are more common in some places than others, and this may impact Google's (and by extension, our) data coverage. To test this theory, we first create a simple index that is the weighted mean absolute percentage error between each BR NAICS sector in the BR single-unit employer universe and our matched sample for each state. This index gives a general sense of how far off our sample is from the BR universe, while also tak-

[2] a firm is defined as common ownership, based on BR Alpha. Figure shows number of valid Google Places IDs matched between Google and BR, but does not account for possible multi-match scenarios.

ing into account the heterogeneity between states. Lighter shading indicates the state is, accounting for size of NAICS codes within a state, *more* accurately represented by our scraped sample, while darker scores indicate worse accuracy. Although California, Texas, New York, Florida, and Ohio all appear in the top 12 most accurate states, we have little indication that population of businesses alone determines if a state is well-represented: our 12 most well-represented states account for approximately 23% of all US establishments (Haltiwanger et al., 2008), an almost exactly proportional figure[3]. While having a large state certainly does not necessarily indicate best fit, it does appear that *small* states without major metro areas are less represented: using our scale, North Dakota, Delaware, Vermont, Montana, and West Virginia are the least accurate 5 states. However, a lack of an urban area does not explain why Virginia and New Jersey also perform poorly–New Jersey is entirely covered by CBSAs, while the bulk of Virginia's population resides in three (Washington-Baltimore-Arlington, Richmond, and Virginia Beach-Norfolk). The actual differences between the states can be significant–West Virginia has a mean percentage error double that of the best state, Alabama. Combined with Figure 3 above, we can see evidence fo the difficulty of using the API approach for statistical purposes, as they appear to generate biased results that are also inconsistent in different areas of the country. However, with a more thorough search, or directly searching for records, the MAMBA results show it is feasible to match the Google API data in a reasonable manner to Census data.

FIGURE 3 HERE

3.4 Textual Data

Once we collected and matched our 120,000 records, we first analyzed how unique each NAICS sector was by the words used in the user reviews and websites. A model will have the easiest time identifying a NAICS sector if *all* of the words used in the reviews or website are unique to that sector. However, since the English language was not created to ease the classifications of businesses, we have to focus on how clear the signals in our data seem. Figure 4 below shows, the proportion of words found in website and review text *that are unique to that sector*. The larger the proportion of unique words, we argue, the simpler the classification decision for a model. Two clear trends emerge. Firstly, there is a great deal of heterogeneity between NAICS sectors. For example, the Information sector contains only

[3]12 states accounts for 23% of the 50 states plus DC

22% of words used on websites are unique to that sector, compared to almost 58% in Accommodation and Food Services. Secondly, website text always contains a greater proportion of words that are unique to the sector compared to user reviews across all sectors. This may provide early indications that website text, and not user reviews, provides a clearer way to identify NAICS codes, however more sophisticated Natural Language Processing techniques are required to validate this speculation.[4]

FIGURE 4 HERE

3.5 Natural Language Processing

Natural Language Processing (NLP) is a suite of analysis tools that gives mathematical meaning to words and phrases. For this research, we require this approach to convert website and review text into sensible dimensions, which we can then use to classify NAICS sectors. The most basic form of NLP appears as "one-hot encoding", demonstrated below in Equation 1. This method can be used for many classifiers (e.g. Naive Bayes), it has some major disadvantages, namely that it does not account for the context of words, an extremely problematic assumption in our work. For example, when identifying if the word "club" is associated with either a restaurant or a golf course, we would need to know if the word "club", when used in context, appears near to the words "sandwich" or "golf".

$$\begin{pmatrix} Do \\ Or \\ Do \\ Not \\ There \\ Is \\ No \\ Try \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{1}$$

As an alternative to context-less approaches, Word2Vec methods were first developed by Mikolov *et al* (2013) to more adequately capture the meanings behind words. Word2Vec models operate by calculating the likelihood that a word appears *given the words surround-*

[4]Another possibility here is insufficient HTML parsing. We used standardized software (BeautifulSoup4, (**?**)) for our parsing, however it is possible many words in the HTML text are insufficiently parsed fragments.

ing it. This is the 'skip-gram' model, which uses a Neural Network to identify a latent layer of relationships between words by assessing how likely different words are to appear near each other in sets of text. Figure 5 below shows a basic illustration of the skip-gram model. If this type of model would be useful in predicting NAICS sector codes, then words associated with each NAICS sector appear as more likely to appear near one another. For example, we should expect to see more mentions of the words 'burger', 'salad', 'pork belly', and 'pizza' near one another in reviews and websites belonging to businesses in the accommodation/food services NAICS code, whereas we may see words like 'oil', 'gas', and 'mine' from reviews from Construction and Mining industries. In Figure 5 below, the model seeks to identify the probability of any of the listed words appearing given the word 'burger' appears nearby. If our hypothesis that review and website text is accurate, we would see higher probabilities of the words 'beef' and 'fries' appearing near the word 'burger' compared to 'architect' and 'blinds', and thus a model will be able to identify these patterns and classify businesses based on the words used in our data. The key output of the Word2Vec model is not the output probabilities: it is the 'hidden layer', in effect a latent variable similar to factor loadings in a standard factor analysis, which reduces the dimensionality of the data, and can be used as predictors in a classification model.

FIGURE 5 HERE

The Word2Vec model, then, provides us with the ability to distinguish how *likely* words are to appear given their context, however it only provides the information for individual words, whereas our data has paragraphs of text for each observation. To solve this issue, we turn to Doc2Vec models (Mikolov et al., 2013), which function in the same way to Word2Vec, but return a hidden layer of factor loadings *for an entire document* of text. In a Doc2Vec model, a value on a hidden layer i for document k can be considered the average loading of document k on i. The Doc2Vec model, then, returns a series of values for each establishment in our data, accounting for the context of the words used, averaging across all the sentences in a document. If we are correct, then user reviews and websites for businesses in different NAICS sectors should have different contexts, and this method should allow us to evaluate how user reviews for restaurants and hotels differ from those for educational establishments.

3.6 Machine Learning

The vector outputs from Doc2Vec models lend themselves well to unsupervised classification techniques such as clustering, however they can also function as features (independent variables) in supervised machine learning algorithms. Given we have already matched our data to the Business Register, we have "true" NAICS sector codes for each establishment that we have matched, which we can use as our dependent variable. In this paper, we rely on Random Forest classifiers to predict the NAICS sector of each establishment given a set of generated vectors for business name, user reviews, and websites, as well as a series of binary variables indicating the "Type" tag for each establishment in the Google API data. Random Forests are a method of classification techniques derived from Decision Tree classifiers, but are relatively immune to issues such as over-fitting that often impact Decision Trees. Important for our analysis, Random Forests have been shown to out-perform (in terms of performance on test data) more common approaches such as logistic regression in class-imbalanced circumstances (Muchlinski et al., 2016).

However, as with all models the assumptions and features included play a huge role in determining outcomes in Random Forests. In order to ensure our model selection is both replicable and maximizes accuracy, we performed an analysis of 1000 different model configurations. For each configuration, we tasked the model to compute a Doc2Vec model for business name, user reviews, and website text with a random number of vectors. Then, in order to optimize those parameters, we executed a 50-iteration, 10-fold cross-validated grid search across those parameter configurations. Plainly, we randomly altered the number of vectors a Doc2Vec model was expecting, as well as how many, and how deep, trees the Random Forest model uses, and then tested how those different model configurations altered each model, and repeat this process to ensure we do not over-fit our data. This random process reduces the possibility of over-fitting–in and out-of sample tests show nearly identical quality, an encouraging sign for future use of this method for production. Our criteria to select our "best" model was to minimize log-loss of the model. Log loss is a penalizing function that allows us to weigh the trade-off between the prediction we make *and how certain we are about it.* Log loss penalizes incorrect predictions with high predicted probabilities, but does not penalize less certain incorrect assumptions. For our purposes, this is an ideal trade-off: the SSA Autocoder does not assign NAICS codes if the predicted probability is less than .638, so any system based on our model will need to be sensitive to the need to prevent assigning codes without high levels of certainty–this is especially relevant as more findings reveal that studies utilizing Machine Learning often over-fit their data, leading to disagreement when different data or methods are used to answer the same question (Allen).

4 Results

4.1 Model Evaluation

Figure 6 below shows the predicted log loss (bold line) and 95% confidence interval (shaded area) across the range of number of vectors used in our analysis. The goal of our grid search analysis was to *minimize* log-loss, however to aide interpretation higher scores on the y-axis indicate superior fit. The figure highlights one major outcome of this experimentation: in general, a relatively small number of vectors produces better results for user reviews and websites, with little different for business name. These findings are slightly counterintuitive: Doc2Vec models can be fit with up to 1000 vectors, and one would think that a complex task such as generating NAICS codes would require *more*, not less vectors. There are two possible explanations for this pattern. First, since a Random Forest takes a random subset of the features for each tree, providing a larger number of features only means the model is less likely to be making predictions based on the more important features. Secondly, given our data set is tiny compared to the original training data for Doc2Vec models, we may be simply unable to generate sufficiently predictive vectors with our current sample. The findings below discuss our best fitting model, which utilizes 119 trees in the Random Forest, with 20 vectors for business name, 8 vectors for user reviews, and 16 vectors for websites.
FIGURE 6 HERE

4.1.1 Predictive Accuracy

Overall, our model predicts approximately 59% of cases accurately. This figure places our model substantially below the current auto-coding methods used by the Social Security Administration, however it is a similar level to the initial match rates for the method, and shows comparable performance to similar exercises in other countries (Roelands et al., 2017). The model also exhibits considerable variation, with some NAICS codes (Information, Manufacturing) seeing fewer than 5% of observations correctly predicted, but

Accommodation and Food servers sees approximately 83% of establishments correctly assigned into the NAICS sector. However, given the unbalanced nature of our sample, evaluating strictly on accuracy may be misleading–it would encourage a model to over-fit to only large NAICS codes. The F1 score is the harmonic mean of the Precision and Sensitivity. For each NAICS code k, precision measures the total number of correctly identified cases in k divided by the total number of cases identified as k by the model. recall, or sensitivity, measures the proportion of cases in NAICS code k accurately predicted. Formally:

$$precision_k = \frac{TruePositive_k}{TruePositive_k + FalsePositive_k}$$
$$recall_k = \frac{TruePositive_k}{TruePositive_k + FalseNegative_k}$$

The F1 score is the harmonic mean of the precision and recall, and provides an advantage for us in that it shows a balanced way to evaluate model fit: in our case, a model may have a high recall merely by predicting all observations in larger NAICS codes, whereas a model fit to maximize precision may result in a high number of false negatives. Formally:

$$F1 = 2 \cdot \frac{precision \times recall}{precision + recall}$$

Figure **??** below shows a scatter plot of the average number of words *unique to the NAICS code* in our data, taken from figure 7 on the x-axis and the F1 Score for each NAICS sector on the y-axis. Clearly, our F1 scores for the Information, Wholesale Trade, and Manufacturing are exceedingly low, but *we also have the least number of words appearing only in that NAICS code.* The clear relationship shows why we are encouraged by the model. Simply put, words that are unique to a certain NAICS code represent a better signal for a model to use as a classifier. For example, a user discussing While our current model performs poorly in some sectors, it clearly shows potential for

better performance with more data. When our data collection efforts yielded a large corpora of words, our model was fairly efficient at predicted NAICS sectors. However, when our data collection efforts failed to find a sufficient corpora of text for a NAICS sector, our model performed poorly. We argue, then, our performance will improve with additional data from these under-sampled sectors–although the increase in number of unique words may not be linear compared to the number of obseravtions, our findings point directly to our model struggling to predict a (relatively) small number of businesses off of a relatively small number of unique words, which would be ameliorated with a broader search. However, it is also possilbe that a larger serach yields no improvement in terms of unique words for these sectors, which would call our search method into question.
FIGURE 7 HERE

One advantage of our multinomial classification is that we can evaluate how difficult our model finds distinguishing between two NAICS codes. Figure 8 shows the confusion matrix between actual NAICS codes (y-axis) and predicted NAICS codes (x-axis), excluding correctly predicted observations. This presentation enables us to evaluate *the kinds of errors* our model makes. Encouragingly, in every NAICS code, our model assigns the highest average predicted probability to correct predictions, however it also assigns Retail Trade (NAICS 44-45) as the *second* most likely NAICS code for each sector. This has a particularly large impact, we see, on Wholesale Trade (NAICS sector 42). Logically, this outcome is expected–the key difference between Wholesale and Retail trade may often not be the actual goods *but the customers.* Wholesale traders sell merchandise not directly to the public but to other businesses, but the types of words used on websites and in user reviews will often be similar, and this pattern may also appear across other NAICS sectors–for example, the term "golf clubs" may appear in the Manufacturing,

Wholesale Trade, Retail Trade, and Arts and Recreation sectors. In cases like this, when words give similar loadings, our model tends to select the NAICS code with the largest number of observations, as this reduces the impurity of the decision tree. This difficulty highlights the need for further investigation on methods and models to overcome these weaknesses.
FIGURE 8 HERE

5 Discussion

This paper has presented a new way for the Census Bureau and other statistical agencies to gather and generate industrial classification codes using publicly available data and modern Machine Learning techniques. Clearly, the current methods are not of sufficient accuracy to replace currently-used methods, however we have displayed some hope that a development program may be able to use the general framework to produce NAICS codes in a more timely and efficient manner. In this section, we will discuss two sets of issues that future work must grapple with before the Census can implement some of our techniques.

Clearly, a model only correctly predicting 60% of cases accurately cannot serve as a substitute for current methodologies. However our findings do indicate *how* these methods may eventually serve as the basis for a statistical product. Firstly, this paper has shown that using text as data to generate NAICS codes requires large numbers of establishments in order to attain a sufficiently large and diverse dictionary which allow vector reduction methods to identify any distinct signal from each NAICS code. However, even with large, diverse data, these methods may still struggle to disentangle NAICS codes with similar corpora of words such as Retail and Wholesale Trade. In these cases, the Census may have to look at alternative public or governmen-

tal data to supplement these efforts. Secondly, related research has shown improved fit for a smaller dataset using Naive Bayes classifiers and Term-Frequency-Inverse Document Frequency (TF-IDF) vectorization in place of Doc2Vec. We chose not to pursue these methods in this paper as the Naive Bayesian framework assumes no relationship between subsequent words, a patently false assumption in our case, however further investigation into alternative approaches may be able to use these methods more effectively.

The next set of issues focuses on how the Census Bureau may seek to put our ideas into production. Other than the obvious need to improve the modeling fit and performance, we see four possible challenges. Firstly, while substantially cheaper than survey collection, access to APIs is not free, and grid searches on the scale needed would require substantial computing and programming effort in order to effectively generate enough data. Secondly, APIs are specifically designed to *prevent* users from replicating databases, and only provide users information based on proprietary algorithms. Practically, this may necessitate enterprise-level agreements between the Census Bureau and data providers such as Google in order to gain access to the entirety of the data available. Next, it has been noted that the performance of textual-based classification models is very data specific, in that models designed to classify NAICS sectors may not perform well for sub-sectors or lower levels of NAICS classifications.

The Census Bureau maintains the highest standards in data privacy and confidentiality, but the prospect of web-scraping presents two possible ways the general public may view this activity as threatening this dedication. Firstly, the Census may wish to avoid seemingly gathering data without consent. While the data we analyze here is in the public domain, other research efforts at the Bureau (Dumbacher and Hanna, 2017) left a "calling card" on scraped websites to inform the host as to the activities and purposes of the data collected. Another possible alternative is that Economic surveys

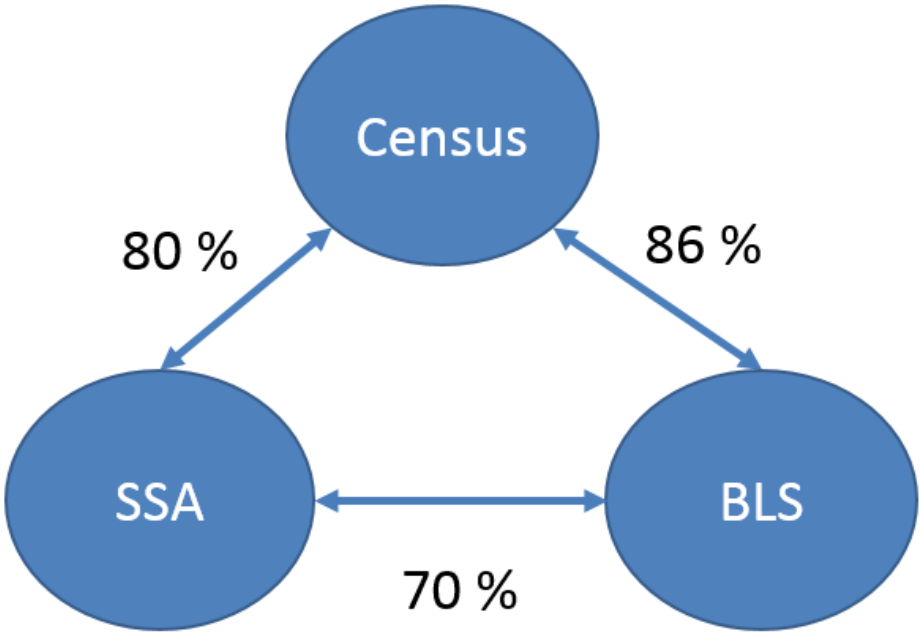
include an option for the respondent to give consent for the Census to gather information from their website to ensure data quality. Secondly, other data gathering efforts for this paper required that data already protected by Titles 13 and 26 received heavy "salting" before searching Google to avoid fact of filing disclosures. Such an approach, while ensuring data privacy and confidentiality, complicates the identification of larger samples of BR records, as there are fewer "salting" records available from external sources (i.e. other APIs).

While we are optimistic our approach may yield useful statistical products, we are realistic that, given the nature of APIs and the issues discussed above, the Census may need to consider alternative uses for text analysis. We can identify two immediate possibilities. Firstly, current auto-coding methods rely on dictionaries of words first gathered from EIN applications between 2002 and 2004. An approach that seeks to combine web-scraping, MAMBA matching, and NAICS classification could be used as a means to update these dictionaries in an efficient, cost-effective manner. This would provide immediate added value to the Census and the SSA, as it would not require new models to be developed, and could easily compared to the previous dictionaries for QA purposes. Secondly, our approach could be used for *targeted* searches of samples of BR data where current methods are unable to automatically assign a NAICS code. In this circumstance, Census staff could leverage these techniques as opposed to hand-review, reducing costs and the time investment required to produce accurate NAICS codes. In particular, the response rate for 'classification cards', which identify the NAICS sector for the business, for the Economic Census in 2017 declined over compared to previous marks, and compared to overall response rates. This produces substantial costs and delay for Census operations, and we argue provides a clear example of the utility of our approach of leveraging alternative data sources and modern machine learning techniques to help the Census accomplish its

mission.

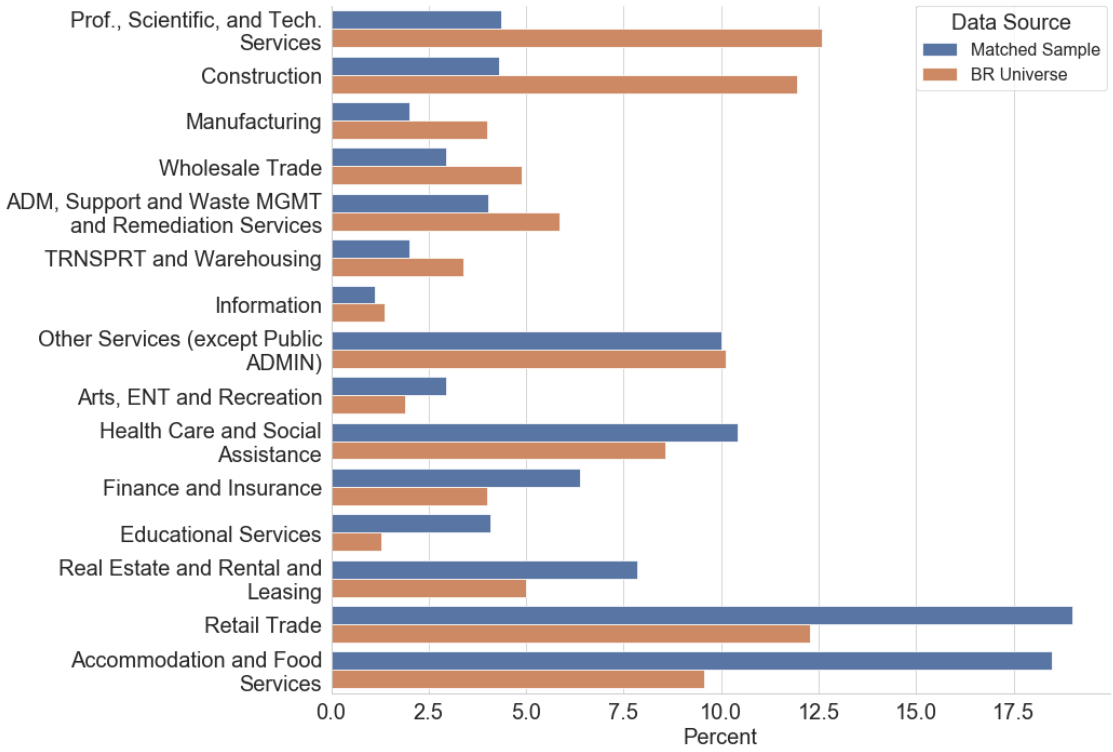
6 Figures

Figure 1: Agreement on NAICS Sectors between Census, BLS, and SSA.



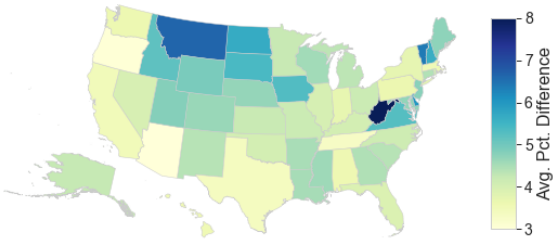
Note: Figure shows the Percentage of BR establishments that share a common 2-digit NAICS sector when present in each respective data source.

Figure 2: NAICS Code Representation.



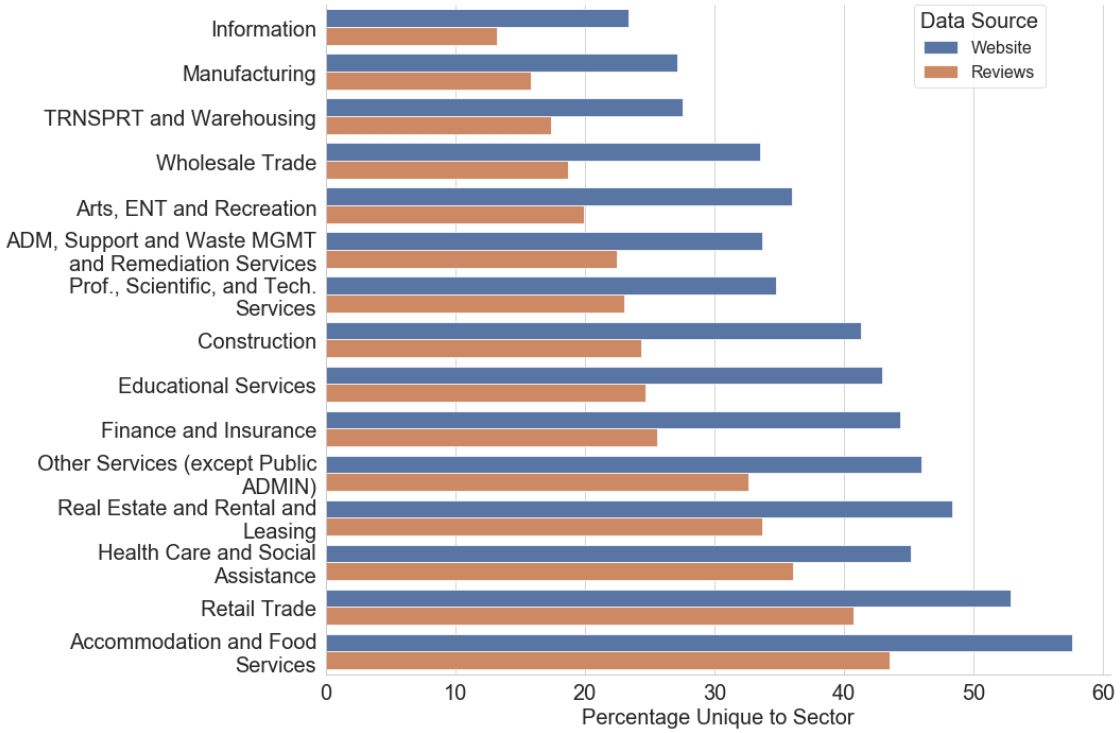
Note: Figure shows the Percentage of Single-Unit estabs in each sector on the 2015/2016 (pool) BR (blue, top)
and the percentage of establishments in our matched sample (orange, bottom).

Figure 3: Geographic Representation of Matched Data.



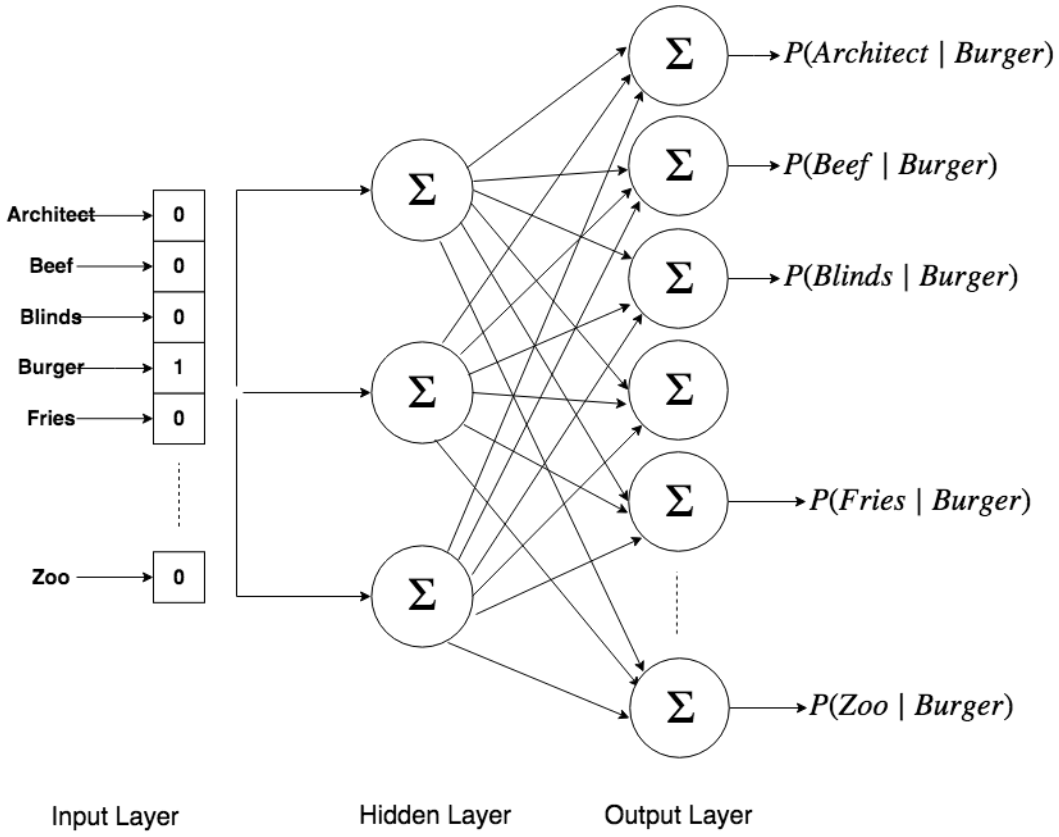
Note: Figure shows representation of each state by weighted average of absolute percentage error between BR SU universe and scraped data. Lighter color indicates more representative sample.

Figure 4: Uniqueness of Word Corpora by NAICS Code.



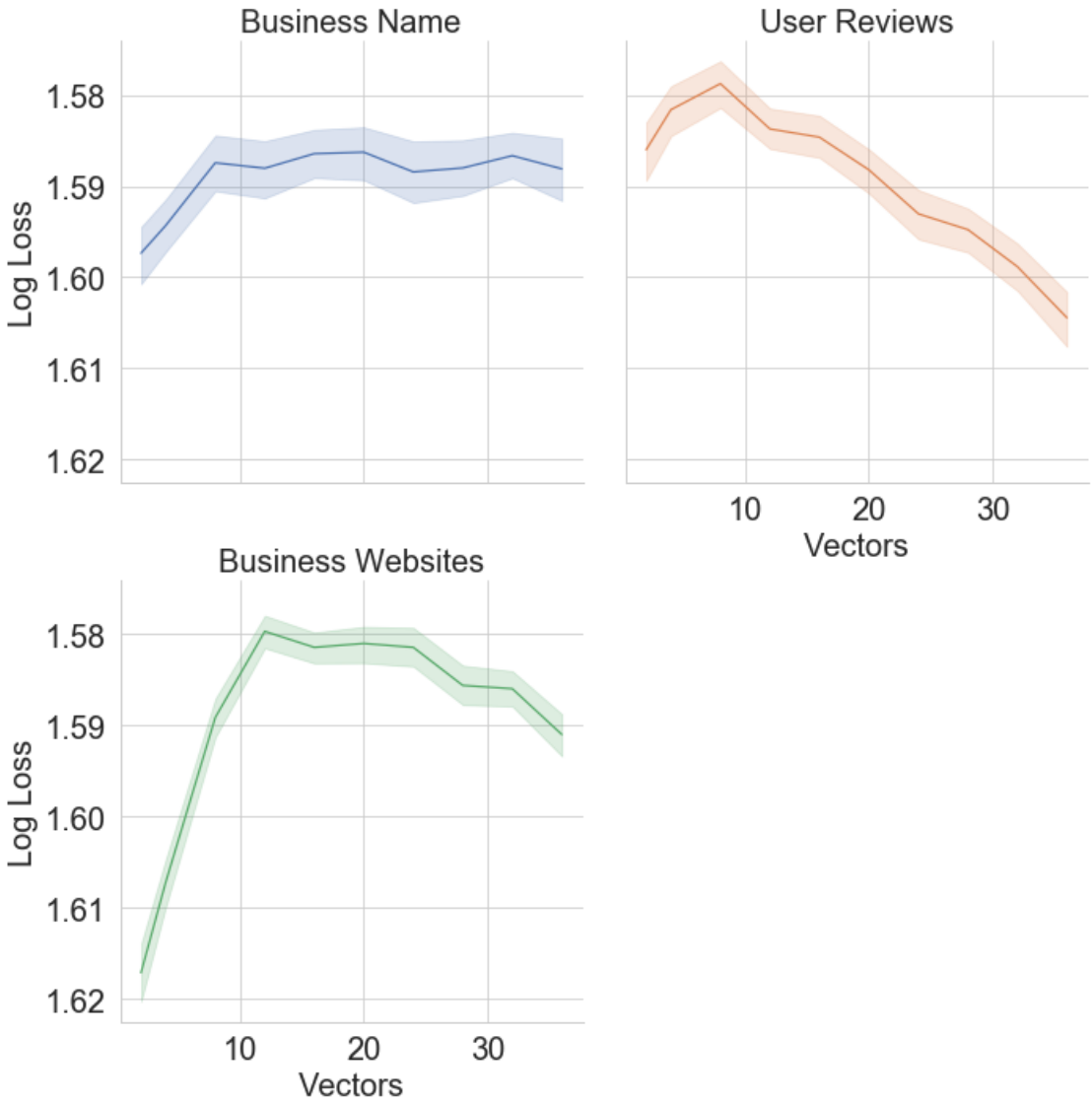
Note: Figure shows the percentage of words appearing in website (top, blue) and review (bottom, orange) that are unique to the particular NAICS sector.

Figure 5: Illustration of Word2Vec Model.



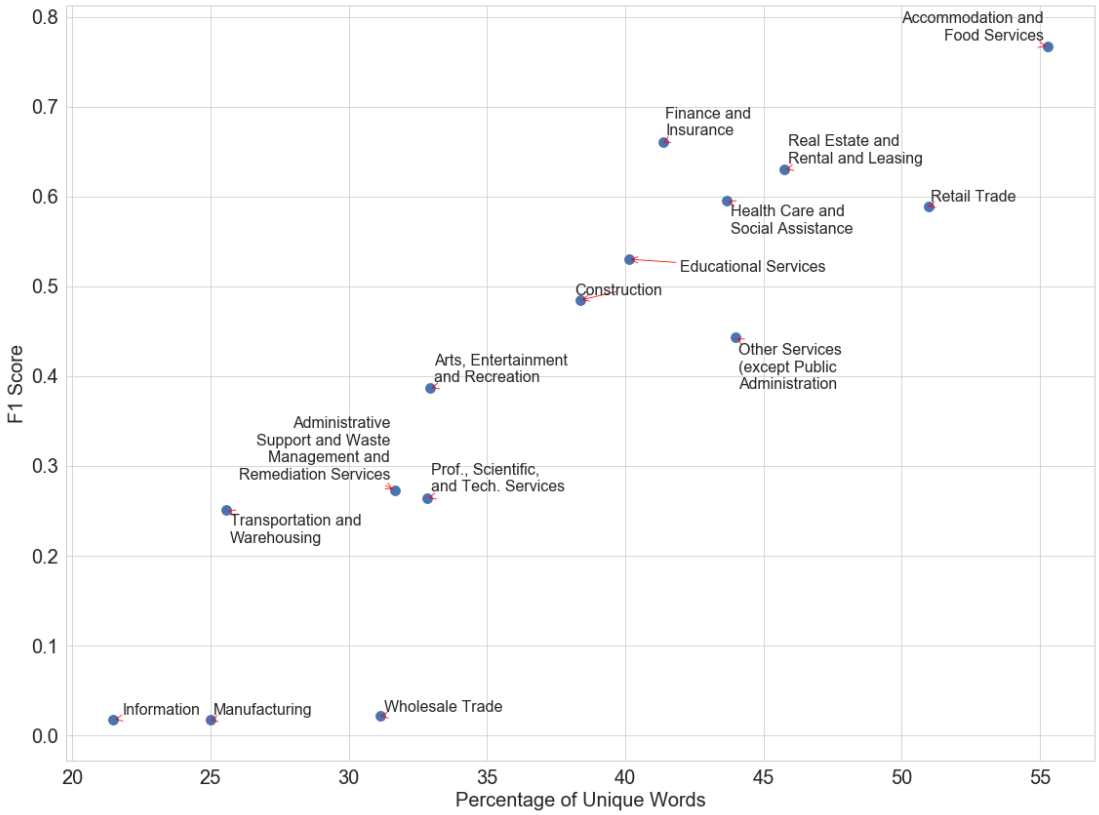
Adapted from: http://mccormickml.com/assets/word2vec

Figure 6: Model Performance Across Parameter Space.



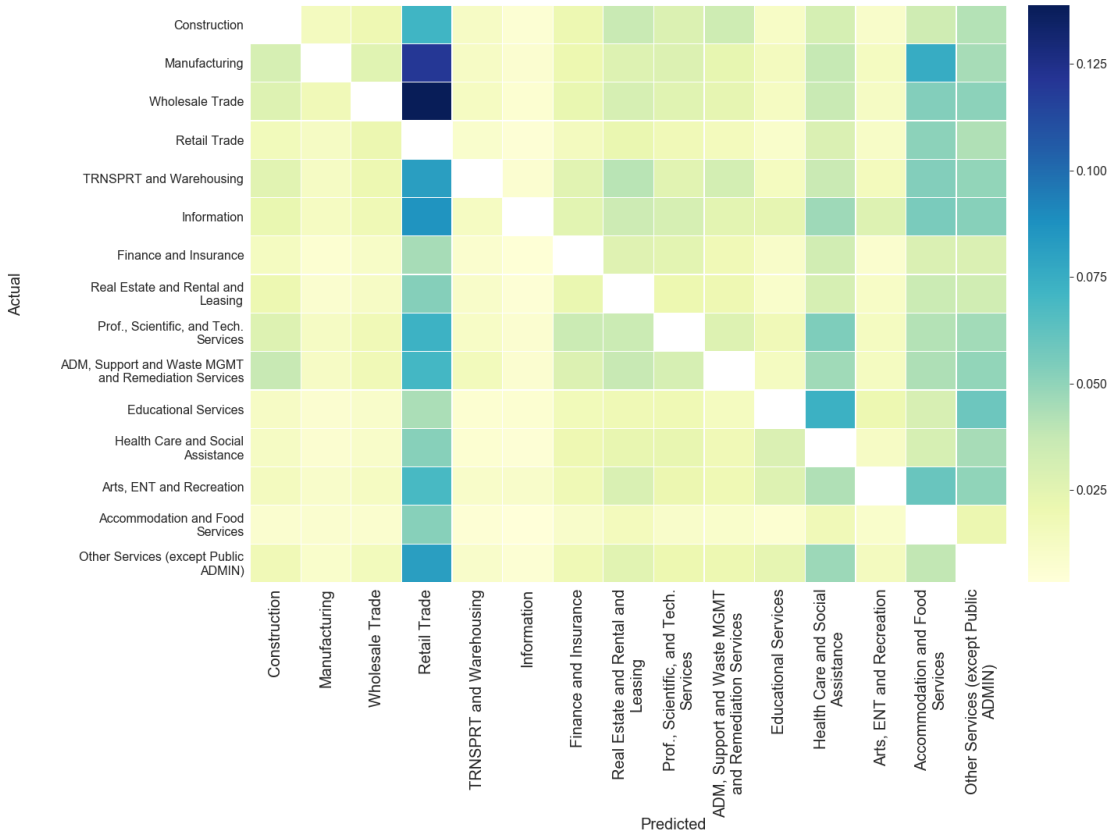
Note: Figure shows the mean and 95% confidence interval for a model using the number of vectors for the respective text source. Y-axis inverted to ease interpretation.

Figure 7: Model Performance by NAICS Sector.



Note: Figure shows the (averaged) percentage of words used in website and review text.
for each NAICS sector that are unique to that sector (x-axis) and F1 score from our model (y-axis).

Figure 8: Heatmap of Predicted Probability for Each NAICS Sector.



Note: Figure shows the proportion of unique words (averaged) from websites and reviews. (x-axis) and F1 score (y-axis) for each NAICS sector in our model.

References

(1984). *The Comparability and Accuracy of Industry Codes in Different Data Systems*. The National Academies Press, Washington, DC.

Allen, G. Machine learning: The view from statistics. In *American Academies of Arts and Sciences Annual Meetings*.

Bayard, K., Dinlersoz, E., Dunne, T., Haltiwanger, J., Miranda, J., and Stevens, J. (2018a). Business formation statistics. `https://www.census.gov/programs-surveys/bfs/data/datasets.html`.

Bayard, K., Dinlersoz, E., Dunne, T., Haltiwanger, J., Miranda, J., and Stevens, J. (2018b). Early-stage business formation: An analysis of applications for employer identification numbers. Working Paper 24364, National Bureau of Economic Research.

Committe, E. C. P. Issues paper no. 3: Collectibility of data. In *Issues Papers*.

Cuffe, J. and Goldschlag, N. (2018). Squeezing more out of your data: Business record linkage with python. In *Center for Economic Studies Working Paper Series*.

Dumbacher, B. and Hanna, D. (2017). Using passive data colection system-to-syystem data collection, and machine learning to improve economic surveys. In *Joint Statistical Meetings*.

Fairman, K., Foster, L., Krizan, C., and Rucker, I. (2008). An Analysis of Key Differences in Micro Data: Results from the Business List Comparison Project. Working Papers 08-28, Center for Economic Studies, U.S. Census Bureau.

Fairman, K., Foster, L., Krizan, C., and Rucker, I. (2012). An Analysis of Key Differences in Micro Data: Results from the Business List Comparison Project. Working papers, U.S. Census Bureau.

for National Statistics, O. Unsupervised document clustering with cluster

topic identification. In *Office for National Statistics Working Paper Series*.

Foster, L., Elvery, J., Becker, R., Krizan, C., Nguyen, S., and Talan, D. (2006). A comparison for the business registers used by the bureau of labor statistics and the u.s. census bureau. Working papers, Bureau of Labor Statistics Office of Survey Methods Research.

Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M., and Steiner, S. (2017). Three methods for occupation coding based on statistical learning. *Journal of Official Statistics*, 33(1).

Haltiwanger, J., Jarimin, R., and Miranda, J. (2008). Jobs created from business startups in the united states. Working papers, Center for Economic Studies, U.S. Census Bureau.

Jung, Y., Yoo, J., Myaeng, S.-H., and Han, D.-C. (2008). A web-based automated system for industry and occupation coding. In Bailey, J., Maier, D., Schewe, K.-D., Thalheim, B., and Wang, X. S., editors, *Web Information Systems Engineering - WISE 2008*, pages 443–457, Berlin, Heidelberg. Springer Berlin Heidelberg.

Luo, T. and Stark, P. (2014). Only the bad die young: Restaurant mortality in the western us.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Muchlinski, D., Siroky, D., He, J., and Kocher, M. (2016). Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis*, 24(1):87103.

Roelands, M., van Delden, A., and Windmeijer, D. (2017). Classifying businesses by economic activity using web-based text mining. In *Centraal Bureau voor de Statisteik*.

Tarnow-Mordi, R. The intelligent coder: Developing a machine learning classification system. In *Methological News 3*. Australian Bureau of Statistics.