Using Public Data to Generate Industrial Classification Codes*

John Cuffe (john.cuffe@census.gov) † #
Sudip Bhattacharjee (sudip.bhattacharjee@uconn.edu) § #
Ugochukwu Etudo (ugochukwu.etudo@uconn.edu) §
Justin C. Smith (Justin.c.smith@census.gov) ‡
Nevada Basdeo (nevada.basdeo@census.gov) ‡

July 30, 2019

Submitted to CRIW 2019

Draft. Do not cite or circulate without permission.

†MOJO Development Team, U.S. Census Bureau
‡Center for Optimization and Data Science, U.S. Census Bureau
§University of Connecticut

Corresponding author

*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 Catherine Buffington, Lucia Foster, Javier Miranda, and the participants of the CRIW conference for their comments and suggestions.

1 Introduction

Statistical agencies face increasing costs, lower response rates, and increased demands for timely and accurate statistical data. These increased demands on agency resources reveal the need for alternative data sources, ideally data that is cheaper than current surveys and is available within a short time frame. Textual data available on public-facing websites present an ideal data source for certain US Census Bureau (henceforth Census) statistical products. In this paper, we identify such data sources and argue that these sources may be particularly well suited for classification tasks such as industrial or occupational coding. Using these sources of data provide five specific advantages:

i) 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 gather and share data used to generate statistics, ensuring greater transparency for federal statistics.
ii) This approach will allow the Census to provide more timely data, with searches and implementation of data modeling occurring far faster than traditional surveys.
iii) This approach could lower respondent burden by identifying answers to questions from the public data.
iv) This approach allows for clear comparisons between agencies producing similar statistics.
v) By utilizing new techniques such as textual analysis, public data may reveal avenues for new and innovative data products in addition to improving current offerings.

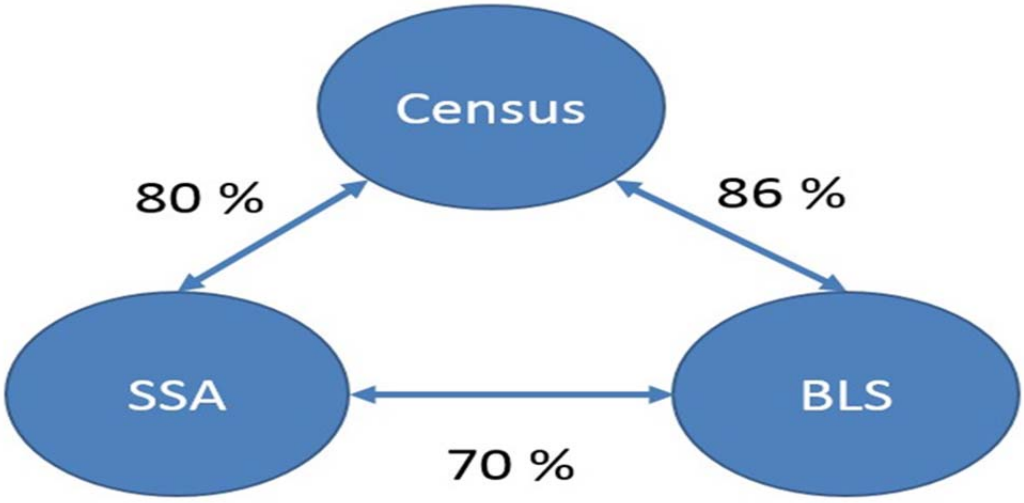
In this paper, we explore how public data can improve the production of federal statistics, using the specific case of using website text and user reviews, gathered from Google Places API, to generate North American Industrial Classification System (NAICS) codes for approximately 120,000 single-unit employer establishments. Our approach shows that public data is a useful tool for generating NAICS codes. We also find challenges, and provide suggestions for agencies implementing such a system for production purposes. The paper proceeds as follows: first, we 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.

1.1 The Case for NAICS Codes

The NAICS is the system by which multiple federal and international statistical agencies assign business establishments into industrial sectors or classes. Economic statistics, such as the Business Dynamics Statistics (Haltiwanger et al., 2008), and survey sampling frames rely on timely and accurate industrial classification data. Currently, NAICS codes are produced by multiple statistical agencies: The Census 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. NAICS classification provides an ideal testbed for use of public data – more accurate, timely, and consistent NAICS codes would save Census considerable effort, and improve statistical quality and timeliness. For example, the Economic Census uses 'classification cards', which are forms sent to a business prior to the Economic Census in an attempt to identify its correct NAICS code, which enables the correct Economic Census electronic survey path for that business. Filling out such an additional 'classification card' form adds substantial burden to respondents, increases survey costs, and may also suffer from lower response rates. Our proposed methodology has the potential to allow Census to avoid such costly classification procedures and deliver better data products at a faster rate. Another compelling reason to develop NAICS codes from public data sources is that laws that govern data sharing between agencies prevent reconciliation between agency codes. A standardized set of assigned classifications would allow agencies to coordinate their lists and ensure all establishments receive the same code. Figure 1 shows the percentage of agreement, at the 2-digit level, between NAICS codes produced by the 2012 Economic Census, BLS, and SSA for the same set of single-unit establishments active in 2012. It shows that the Census and BLS, when coding the same cases, agree on the NAICS *sector* in approximately 86% of cases, whereas the BLS and SSA concur in around 70% of cases.

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. Source: 2012 Business Register Single Unit Businesses.

Several statistical agencies have attempted to use textual data as a means for classification. Much of the work has focused on generating occupational classifications based on write-in survey responses (for example, Gweon et al., 2017; Jung et al., 2008; Fairman et al., 2012). There are also 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, 2018), identifying industrial clusters using a combination of Doc2Vec and Singular Value Decomposition (SVD) models. The data were fit on a "relatively small" number of observations, leaving the usefulness of the method at much more fine-grained levels unknown (Office for National Statistics, 2018). Researchers from 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 business, 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, 2017). In the U.S. statistical system, Kearney and Kornbau (2005) produced the SSA's "Autocoder", a system that uses word dictionaries and predicted NAICS codes based on open-response text on Internal Revenue Service (IRS) Form SS4, the application for a new Employer Identification Number (EIN). The Autocoder, developed in the early 2000s, remains in service, and relies on a combination of logistic regression and subject-matter experts for quality assurance and manual coding tasks. Other work has sought to apply similar methods as ours to coding occupational injuries and occupational types (Bertke et al., 2016; Measure, 2014; Gweon et al., 2017; Ikudo et al., 2018).

We seek to build on previous work by generating 2-digit NAICS sectors for a sample of single-unit, employer businesses active in 2015 to 2016. Our approach combines web-scraping of company websites, company names, and user reviews to generate a corpus of text associated with each business. We then apply Doc2Vec methods to reduce dimensionality of the data in a similar manner to the previous attempts (Roelands et al., 2017; Tarnow-Mordi, 2017). Finally, we use the outputs of this textual analysis as inputs into a Random Forest classifier, seeking to identify 2-digit NAICS codes.

2 Data and Methods

Our approach includes collecting publicly available data from company websites and user-generated reviews of businesses, and combining it with Census protected information on individual business establishments. We utilize public APIs to collect a target sample of approximately 1.3 million business establishments, match those records to the Business Register by name and address, perform textual pre-processing on available text in user reviews, company websites, and company name, and finally use these outputs as features (independent variables) in a Random Forest classifier to predict 2-digit NAICS codes. We first provide a brief overview of each stage of our approach, then compare our dataset sample to the universe of single-unit employer businesses.

2.1 Data from APIs and Web Scraping

An Application Program Interface (API) is a set of procedures that allows users to access information or other services from a provider. For example, Google Places API (used to collect our data) allows access to business information such as name, address, rating, user reviews,

website URL[1], contact information, and Google Types[2] tags. We leverage this information in two ways. First, public user reviews provide a rich source of contextual information about a business. For example, 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. Second, we visit the website (when available) and "scrape" its visible textual information. The working assumption is that a company website provides clear and direct information about products or services it offers. Next, we use Google Types, which vary in usefulness, with less useful words like "establishment" and "point of interest", but also words such as "hotel", "bar", or even "Hindu Temple", which would greatly aid a model in classifying a business. Finally, we use the name of the company, as company names often indicate the type of products on offer (e.g. Krusty Burger). Together, these four elements – all sourced from publicly gathered data – provide us with the type of information needed to describe a business, what products it may sell, and how its customers use or perceive those products (Jabine, 1984).

To generate our sample of businesses, we conducted a grid search on both Yelp and Google Places APIs, based on a combination of lat/long coordinate and keywords. We identified the geographic center of each Core-Based Statistical Area (CBSA) and each county, therein to serve as the starting point for our search.[3] To identify keywords, we found all words contained in the titles of all two-digit NAICS sectors.[4] We then executed an API search for each keyword in 50 random locations for each CBSA and county, 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 Yelp API. Next, we performed a search for each of those businesses on Google Places API , retrieving website URL, user reviews, and Google Types. The website URL was then visited and text was scraped using an internally developed procedure.

For this study, we eliminated records that did not have a website and user reviews, to have the best sample to determine the overall utility of both sources of data jointly. This restriction reduced the number of available records from 1,272,000 million to approximately 290,000. Future research can attempt to generate NAICS codes for establishments that lack either a

[1] URL: Uniform Resource Locator, or website address.
[2] A list of over 100 different classification tags assigned by Google to describe a place.
[3] 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. 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.
[4] https://www.census.gov/eos/www/naics/

website or user reviews.

2.2 Matching Collected Data to the Business Register

The Business Register (BR) is the Census Bureau's comprehensive database of U.S. business establishments and companies, covering all employer and non-employer businesses.[5] To identify if our 290,000 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, and provides high-quality matches. It 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, or city name, or 3-digit zip code in common – in order of importance. We ran two particular matches – the first matching on both name and address, and then a residual matching by only business name.

After matching the Google API data with the BR, we focus on the 120,000 *single-unit* (SU)[6] establishments that have both website and review text, and are employer-based businesses. This accounts for 43.44% of the records. This seemingly low match rate is the result of four circumstances:

i) We only use *single-unit* employer businesses for a cleaner analysis. *Multi-unit* (MU) firms sometimes have a complicated nature of assigned industrial codes. For example, large retail companies 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.

ii) Many Google records may not exist in the BR. The Census Bureau estimated that approximately 350,000 businesses would form after 2016Q3 (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 (Bayard et al., 2018b,a)[7]. Also, some Google

[5] https://www.census.gov/econ/overview/mu0600.html

[6] A single-unit (SU) establishment is a standalone business, where an "establishment" is defined as a particular location. A multi-unit (MU) establishment in a given location is part of a larger business which operates in many locations. Our sample includes only employer-based businesses.

[7] The Business Register defines a business as an employer business if it has payroll on March 12 of a given year. By measuring from 2016Q3, we 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.

records may be fake, and hence cannot be matched (Copeland and Bindley, 2019).

iii) The initial scraping occurred in December 2017/January 2018, whereas the BR data is from 2015/2016. Thus, in some cases the BR is almost two years older than the Google data. In some industries this is a substantial issue: studies have found that approximately 19% of all service-providing businesses (e.g. NAICS code 41 or higher) fail within their first year of operation (Luo and Stark, 2014, p. 11), meaning that many BR businesses may no longer exist, or appear as prominent search results, in the Google database.

iv) We only match for businesses listed as employer based, in either the 2015 or 2016 BR, meaning non-employer businesses are not included in our sample.

2.3 Matched Data Quality

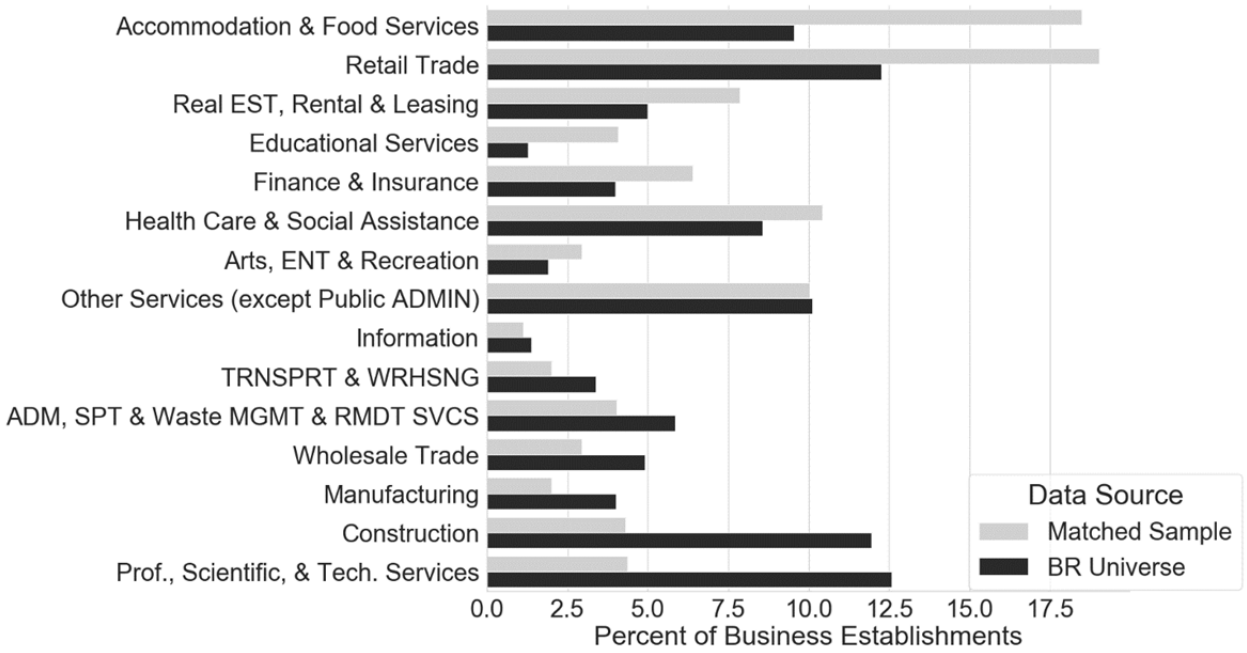
Figure 2 shows the percentage comparison for each NAICS sector between our sample (upper bar) and the BR single-unit employer universe (lower bar). It reveals that our sample heavily over-samples NAICS 44/45 (Retail Trade) and 72 (Accommodation and Food Services). Approximately 12.28% of all BR single-unit employers fall into the Retail Trade sector, however this sector makes up almost 19% of our sample. This is expected, as approximately two-thirds of our sample was sourced from Yelp, which is dominated by food services. In general, Google Places and Yelp both target public-facing industries in their APIs. On the other hand, our approach *under-samples* NAICS code 54, Professional, Scientific, and Technical Services, which is about 12.6% of all businesses, but only 4.36% in our sample. Our sample also under-samples Construction, and Agriculture and Forestry, and Mining sectors relative to their size in the Business Register.

2.4 Textual Data

We analyzed our sample (120,000 records) to see how many unique words were used within the user reviews and website text for each NAICS sector. This provides a measure of signal to noise (textual information) for a given sector, which helps in classification accuracy of that sector. 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. Figure 3 shows the proportion of words found in website and review text that are unique to that sector. The larger the proportion of unique words, the simpler the classification decision for a model should be. Two clear trends emerge. First, there is a great deal of heterogeneity between NAICS sectors. For example, the Information

sector contains only 22% of words used on websites are unique to that sector, compared to almost 58% in Accommodation and Food Services. Second, 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 may provide a clearer way to identify NAICS codes; however more sophisticated Natural Language Processing techniques are required for verification.[8]

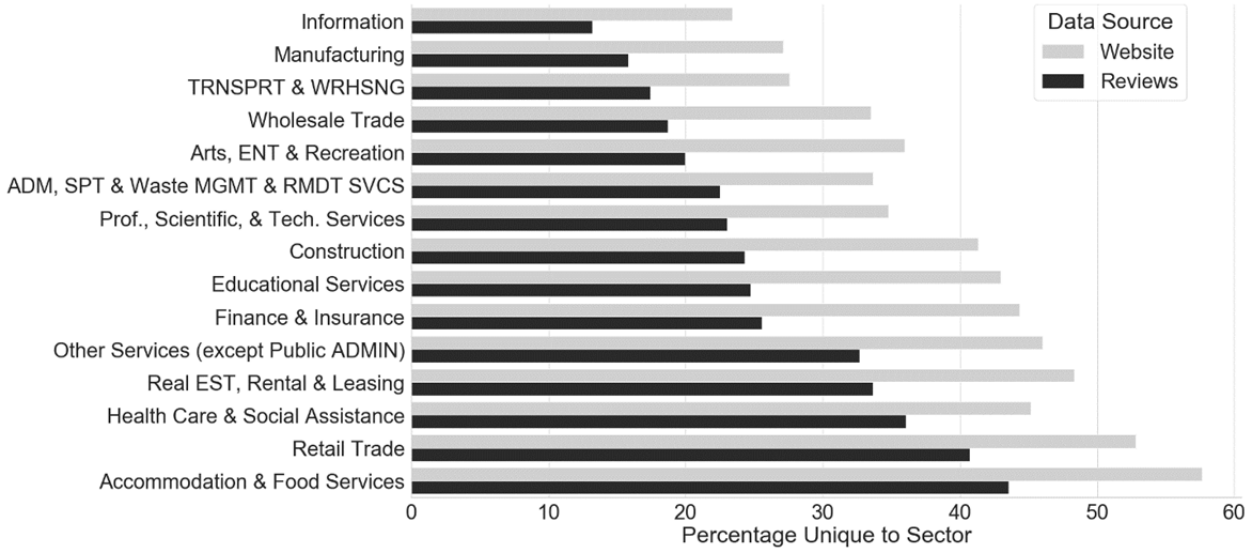
Figure 2: NAICS Code Sample Representation vs. Business Register



Note: Figure shows the Percentage of Single-Unit establishments in each sector on the 2015/2016 (pool) BR (black, bottom) and the percentage of establishments in our matched sample (gray, top). Source: Business Register, 2015-16. Google Places API.

[8] Another possibility here is insufficient HTML parsing. We used standardized software (BeautifulSoup4, https://www.crummy.com/software/BeautifulSoup/) for our parsing; however, it is possible many words in the HTML text are insufficiently parsed fragments.

Figure 3: Uniqueness of Word Corpora by NAICS Code



Note: Figure shows the percentage of words appearing in website (top, gray) and review (bottom, black) that are unique to the particular NAICS sector. Source: Business Register, 2015-6. Google Places API.

2.5 Natural Language Processing

Natural Language Processing (NLP) is a suite of analysis tools that gives mathematical meaning to words and phrases, converting words and phrases to a numerical format based on their semantic and contextual occurrence within a corpus of documents. For this research, we require this approach to convert website and review text into sensible dimensions, which we can then use in a model to classify companies into NAICS sectors. The most basic form of NLP appears as "one-hot encoding", demonstrated in Matrix 1. Although 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. 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".

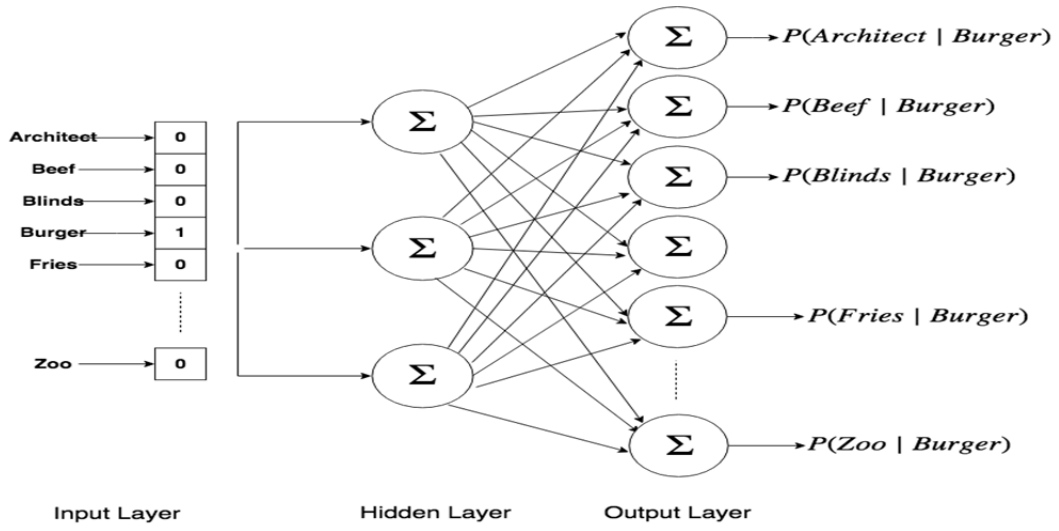
As an alternative to context-less approaches, Word2Vec methods were first developed by Mikolov et al (2013) to more adequately capture context in words. Word2Vec models operate by calculating the likelihood of a word appearing, *given the words surrounding it*. In this 'skip-gram' model, a neural network is used 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 4 shows a basic illustration, where the model seeks to identify the probability of any of the listed words appearing given the word 'burger' appears nearby. In our case, we should expect to see

more mentions of the words 'burger', 'salad', 'pork', and 'pizza' near one another in reviews and websites belonging to businesses in the Accommodation and Food services NAICS code, whereas we may see words like 'oil', 'gas', and 'mine' from reviews in Construction or Mining industries. Thus a model will be able to identify these patterns and classify businesses based on the words used in our dataset. 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 factor analysis, which reduces the dimensionality of the data and can be used as predictors in a classification model.

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

The Word2Vec model 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. On the other hand, our data has paragraphs of text for each observation. To solve this issue, we use 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 returns a series of values for each establishment, accounting for the context of the words used, averaged across all the sentences in a document. In this case, 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.

Figure 4: Illustration of the Word2Vec Model



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

2.6 Machine Learning

The vector outputs from Doc2Vec models lend themselves well to unsupervised classification techniques such as clustering. They can also function as features (independent variables) in supervised machine learning algorithms. After matching our data to the BR, we get the actual NAICS sector codes for each establishment matched, which we use as our dependent variable. We build a Random Forest model based classifier to predict the NAICS sector of each establishment, where the independent variables are the generated vectors for business name, user reviews, and websites, as well as a series of binary variables indicating the Google Type tag for each establishment. Random Forests are a method of classification techniques derived from Decision Tree classifiers, but are relatively immune to over-fitting that often impact Decision Trees. In some cases, Random Forests out-perform more common approaches such as logistic regression in class-imbalanced circumstances (Muchlinski et al., 2016). The 120,000 records are split into 80% training and 20% validation set for model training and evaluation.

In order to ensure our model selection is both replicable and maximizes accuracy, we performed an analysis of 1000 different model configurations. We randomly alter the number of vectors a Doc2Vec model produces, as well as how many, and how deep, the trees are in the Random Forest model. We then tested how those different model configurations altered the accuracy, and repeat this process. Minimum log-loss is chosen as the model comparison criteria, as log-loss is a penalizing function that allows us to weigh the trade-off between the prediction

and its certainty. 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, as the comparable SSA Autocoder does not assign NAICS codes if the predicted probability is less than 0.638 (Kearney and Kornbau, 2005). Hence any system based on our model will need to be sensitive to the need to prevent assigning codes without high levels of certainty.

3 Results

3.1 Model Evaluation

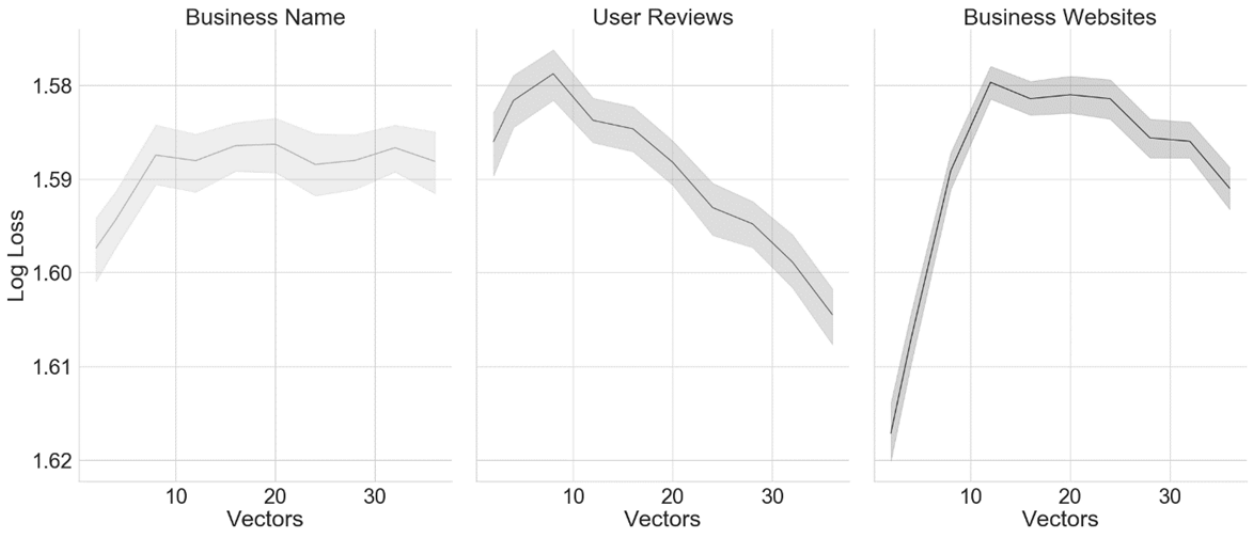
Figure 5 shows the predicted log loss (bold line) and 95% confidence interval (shaded area) across a range of number of vectors used in our analysis. The goal of our grid search analysis was to minimize log-loss. Lower scores on the y-axis indicate superior fit (y-axis is inverted in Figure 5 to ease interpretation). The figure highlights one major outcome of this experimentation: in general, a relatively small number of vectors (around 10) produce better results for user reviews and websites, while it takes approximately 20 vectors for business name. These findings are slightly counterintuitive: Doc2Vec models can be fit with up to 1000 vectors, and one would assume that a complex task such as generating NAICS codes would require more, not less vectors. It is possible that given our sample is tiny compared to the normal training data for Doc2Vec models, we may be simply unable to generate sufficiently predictive vectors with our current sample.

3.2 Predictive Accuracy

The findings here discuss our best fitting model, which utilizes 119 trees in the Random Forest, with 20 vectors for business name, 8 for user reviews, and 16 for websites. Overall, across all NAICS sectors, and for SU establishments only, our model predicts approximately 59% of cases accurately. This places our model substantially below the current auto-coding methods used by the SSA, however it is at a similar level to initial match rates for the SSA method, and shows comparable performance to similar exercises in other countries (Kearney and Kornbau, 2005; Roelands et al., 2017). The model also exhibits considerable variation, with some NAICS codes (Information, Manufacturing) seeing fewer than 5% of observations correctly predicted, while Accommodation and Food Services has approximately 83% of establishments correctly predicted into its NAICS sector. 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. Instead, we use the F1 score to evaluate our model.[9]

Figure 5: Model Performance Across Parameter Space

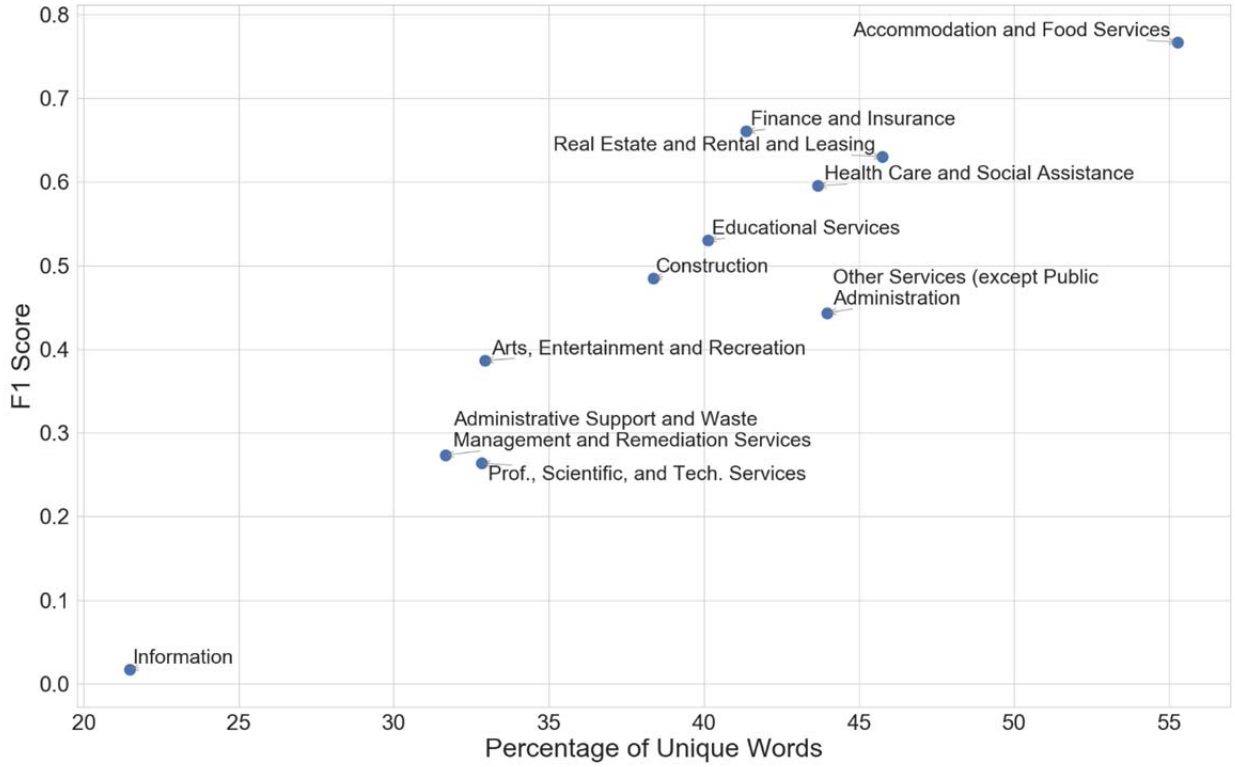


Note: Figure shows the mean and 95% confidence interval for a model using the number of vectors for the respective text source. Y-axis inverted to ease interpretation.

Figure 6 shows a scatter plot of the average number of words *unique to the NAICS sector* in our data (from Figure 3) on the x-axis, and the F1 Score for each NAICS sector on the y-axis. Clearly, Accommodation and Food Services, and Retail Trade have the highest F1 scores, and corresponding *highest percentage of unique words*. Similarly, F1 scores for Information, Wholesale Trade, and Manufacturing sectors are exceedingly low, and also have the *least percentage of unique words* appearing in those NAICS codes, This clear relationship demonstrates encouraging signs of this modeling and approach – words that are unique to a certain NAICS code represent a better signal for a model to use as a classifier. Therefore, we argue that our model performance will improve with additional data from under-sampled sectors. Although the increase in number of unique words may not be linear compared to the number of observations, our findings point directly to our model not able to correctly predict businesses in a sector from a relatively small number of unique words, which may be ameliorated with a broader search.

[9] 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.

Figure 6: 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).

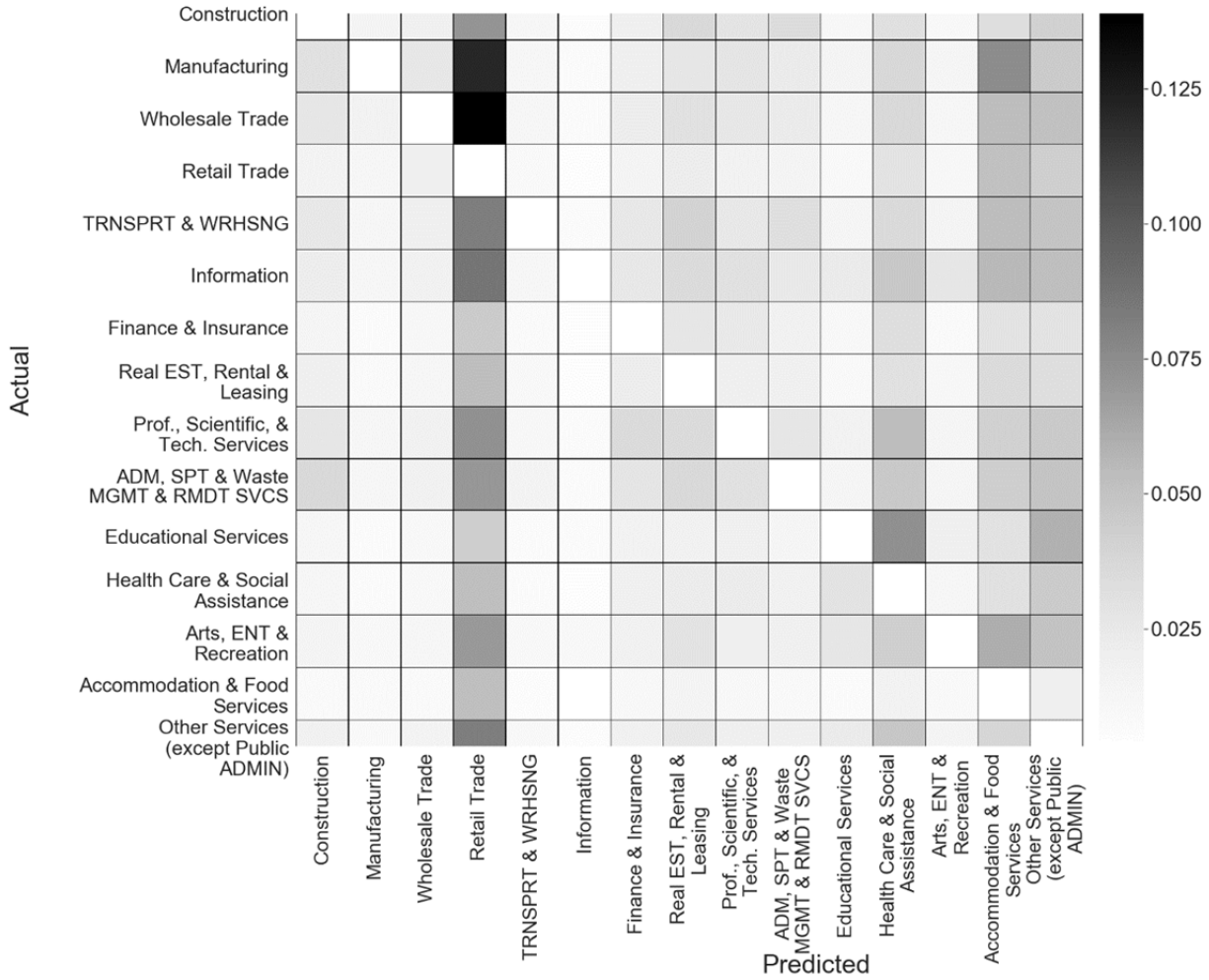
3.3 Misclassification Analysis

One advantage of our multinomial classification approach is that we can evaluate the difficulty in distinguishing between two NAICS codes, one of which is the correct one. Figure 7 shows the confusion matrix between actual (y-axis) and predicted NAICS codes (x-axis), excluding correctly predicted observations. This enables us to evaluate model errors and biases.

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 on Wholesale Trade (NAICS sector 42). Logically, this outcome is expected – the key difference between Wholesale and Retail Trade may not often be the actual goods, but the customers. Wholesale traders sell merchandise to other businesses and not directly to the public, but the types of words used on websites and in user reviews will often be similar. This pattern may also appear across other NAICS sectors – for example, the term "golf clubs" may appear in

Manufacturing, Wholesale Trade, Retail Trade, and "Arts, Entertainment and Recreation" sectors. In such cases, when words have 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 7: Heatmap of Incorrect Classifications in 2-digit NAICS Sectors



Note: Figure shows the proportion of incorrectly predicted businesses in each NAICS sector.

4 Discussion

This paper presented a new approach for Census and other statistical agencies to gather and generate industrial classification codes, using publicly available textual data and machine learning techniques. The approach shows significant promise – in NAICS sectors where more data is available (with high signal-to-noise ratio) to train classification models, the accuracy goes

up to 83%, with negligible fine-tuning of models. On the other hand, in sectors where little data is available, or where there are less unique words describing a sector, accuracy lowers to 5%. Subsequent research has demonstrated that a larger datasets, and alternative modelling approaches, do indeed increase accuracy. Hence, further development of this approach and framework promises to improve NAICS coding at the 2, 4, and 6-digit levels, using publicly available data sources, in a timely and efficient manner.

Our findings indicate that these methods may eventually serve as the basis for a statistical product, once accuracy, bias, reliability and replicability of the techniques are further researched and proven. This paper has shown that using text as data to generate NAICS codes requires data from a sufficiently large number of establishments in each NAICS sector to identify distinct signals from each NAICS code. Further, other types of predictive algorithms, e.g. logistic regression, gradient boosting, and deep learning, should be tested to find their efficacy in solving this problem. In addition, well-known methods of feature engineering, which adds derived variables from the data, have also been shown to improve model accuracy (Chakraborty et al, 2019; Forman, 2003; Liao et al, 2017; Xu et al 2012). Even with advanced methods, it is possible to still struggle to disentangle NAICS codes with similar corpora of words, such as for Retail and Wholesale Trade. This may need clerical or other approaches for a coordinated solution.

We can also identify additional possibilities where our approach can enhance current products. First, current auto-coding methods rely on dictionaries of words first gathered from EIN applications between 2002 and 2004, and updated periodically. The new textual corpus could be used to update these dictionaries in an efficient, cost-effective manner. This would provide immediate added value to the Census and the SSA, and could be compared to previous dictionaries for QA purposes. Second, 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 our approach as opposed to hand-review, reducing cost and time investment required to produce accurate NAICS codes.

Statistical production processes require steady access to source data, and related constraints on budget and computation resources. Web scraping of company websites is substantially cheaper than survey collection, even considering the computation resources needed. However, surveys may gather additional information not be available on websites. In addition, access to APIs for data collection is not free, and grid searches across geographies on the scale needed

would require substantial computing effort in order to effectively generate enough data. Also, 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 Census and data providers (e.g. Google) in order to gain access to the entirety of the data available. If the data is sourced from a single provider, it introduces risk as the data format, availability or even the underlying measurement criteria in the data might change. The provider may even discontinue the data collection, or show monopolistic behavior. These factors need to be carefully addressed for production purposes of statistical products from public or restricted data sources.

The prospect of web-scraping public sources of data may present two risks. First, a *perceptual risk* may be that data is being gathered without consent, although the data is in the public domain. The US Census Bureau should be transparent and announce its intent to publicly gather such information to improve national statistics, reduce respondent burden, save organizations time and resources, and reduce cost.[10] Second, large-scale directed search efforts using data that is protected by Titles 13 and 26 is complicated, and *risks* not being scalable and repeatable. Such protected data need to be mixed with heavy "salting" with external data before searching can occur, 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).

We are excited that our approach can yield useful statistical products. We also suggest that policies be developed to reduce the risk and enhance the usability of such approaches for production purposes. This would provide a clear advantage if Census operations can utilize our approach of alternative data sources and modern machine learning techniques to help Census accomplish its mission more effectively.

References

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

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.

[10] See, for example Statistics Canada web scraping explanation at: https://www.statcan.gc.ca/eng/our-data/where/web-scraping

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.

Bertke, S., Meyers, A., Wurzelbacher, S., Measure, A., Lampl, M., and Robins, D. (2016). Comparison of methods for auto-coding causation of injury narratives. *Accident Analysis Prevention*, 88:117 – 123.

Chakraborty, A., Bhattacharjee, S., Raghu, T.S. (2019). "Data Privacy and Security for US Consumers: Assessing Legislative Success of Data Protection Through Feature Engineering and Prediction". Working paper.

Copeland, R. and Bindley, K. (2019). Millions of business listings on google maps are fake—and google profits. *The Wall Street Journal*, June 20, 2019.

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*.

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.

Office for National Statistics. (2018) Unsupervised document clustering with cluster topic identification. In *Office for National Statistics Working Paper Series number 14*. https://cy.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsworkingpaperseriesnumber14unsupervisedocumentclusteringwithclustertopicidentification (accessed on July 30, 2019).

Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research*, 3(Mar), pp.1289-1305.

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., Jarmin, R., and Miranda, J. (2008). Jobs created from business startups in the United States. Working papers, ADRM, U.S. Census Bureau, December 2008.

Ikudo, A., Lane, J., Staudt, J., and Weinberg, B. (2018) Occupational classifications: A machine learning approach. *NBER Working Paper Series*, number 24951, August 2018. https://www.nber.org/papers/w24951.pdf (accessed July 30, 2019).

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.

Kearney, A. T., & Kornbau, M. E. (2005). An automated industry coding application for new US business establishments. In *Proceedings of the American Statistical Association*.

Liao, W., Al-Kofahi, K., and Moulinier, I. 2017. Thomson Reuters Global Resources ULC, 2017. *Feature engineering and user behavior analysis*. U.S. Patent 9,552,420.

Luo, T. and Stark, P. (2014). Only the bad die young: Restaurant mortality in the Western US. https://arxiv.org/abs/1410.8603 (accessed July 30, 2019)

Measure, A. (2014). Automated coding of worker injury narratives. JSM.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *https://arxiv.org/abs/1301.3781* (accessed July 30, 2019)

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):87–103.

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

Tarnow-Mordi, R. (2017) The intelligent coder: Developing a machine learning classification system. In *Methological News 3*. Australian Bureau of Statistics.

Xu, Y., Hong, K., Tsujii, J., and Chang, E.I.C. 2012. Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries. *Journal of the American Medical Informatics Association*, 19(5), pp.824-832.