A text mining and machine learning platform to classify businesses in NAICS codes

Sudip Bhattacharjee (Presenter), Sudip.Bhattacharjee@uconn.edu
Senior Research Fellow, US Census Bureau; Professor, University of Connecticut, USA

Ugochukwu Etudo, University of Connecticut, USA; US Census Bureau
Justin C Smith, U.S. Census Bureau

1

1

Why is NAICS important?

- **Foundation to measure $19 trillion US economy**
- Standard used by federal agencies
- Adopted in 1997 (replaces SIC)

North American Industry Classification System

Sector	Description
11	Agriculture, Forestry, Fishing and Hunting
21	Mining, Quarrying, and Oil and Gas Extraction
22	Utilities
23	Construction
31-33	Manufacturing
42	Wholesale Trade
44-45	Retail Trade
48-49	Transportation and Warehousing
51	Information
52	Finance and Insurance
53	Real Estate and Rental and Leasing
54	Professional, Scientific, and Technical Services
55	Management of Companies and Enterprises
56	Administrative and Support and Waste Management and Remediation Services
61	Educational Services
62	Health Care and Social Assistance
71	Arts, Entertainment, and Recreation
72	Accommodation and Food Services
81	Other Services (except Public Administration)
92	Public Administration

2

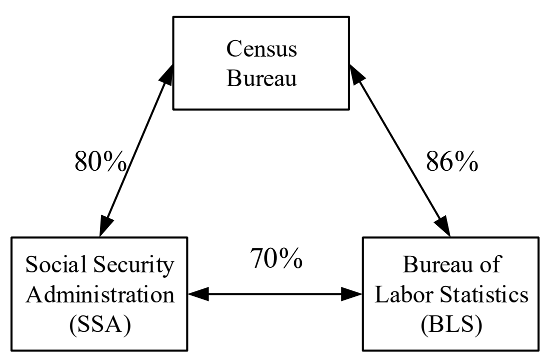
2

NAICS codes received from multiple sources

- **Surveys/interactive updates:**
  - ASM = Annual Survey of Manufactures
  - BSR = Business Sample Revision
  - CCC = Census Classification Card
  - **CEN = Economic Census (every 5 years)**
  - IPG = Global Correction
  - IPS = Interactive correction from D-IPSE
  - RFL = REFILE
  - UPD= Interactive correction from D-IPSE
- **Administrative sources:**
  - **BLS = Bureau of Labor Statistics**
  - BMF = Business Master File
  - **SSA = Social Security Administration – generated from EIN application**
  - TAX = Derived from IRS tax return

NAICS Codes don’t match across govt agencies

- Important element of Census operations, and economic measurement, BUT….
  - Potentially outdated information.
  - Burdensome to respondents.
  - Not all sources agree.
  - No known ground truth.
  - Unable to share data between agencies.



2-digit level match for NAICS codes for single-unit establishments (2012)

Literature review

- Textual analysis to generate occupational classification (Gweon et al., 2017; Jung et al., 2008; Fairman et al., 2012)
- Unsupervised classification of industries (British Office for National Statistics, 2018, Shi et al 2016)
- National Statistics Netherlands (Roelands et al., 2017)
  - challenges: size of the business, the source of the industrial code, and the complexity of the business website
- Australian Bureau of Statistics (Tarnow-Mordi, 2017).
  - based on short, free text responses into classification hierarchies
- US statistical system – "Autocoder" (Kearney and Kornbau, 2005)
  - combination of logistic regression and subject-matter experts for quality assurance and manual coding tasks
- Multiple statistical agencies worldwide attempting similar research

United States Census Bureau | U.S. Department of Commerce Economics and Statistics Administration U.S. CENSUS BUREAU census.gov

5

5

Research hypotheses

- Declarative bias:
  - Adding feature sets will improve upon current "Autocoder" accuracy
    - Publicly available data (Company name, Website text)
    - Commercial data (Google Place type, Yelp tag)
    - Customer-sourced data (customer reviews)
  - Each new feature set, and their combination, will improve classification accuracy
- Procedural (model) bias:
  - Sophisticated modeling approach will improve accuracy over simpler models
  - Ensemble/stacked model will improve accuracy over individual models

United States Census Bureau | U.S. Department of Commerce Economics and Statistics Administration U.S. CENSUS BUREAU census.gov

6

6

Data

- Collected 500,000 business places from Google Places API
  - business name, website URL (publicly available data)
  - Google types tags (commercial data)
  - User reviews (customer sourced data)
- Scraped homepage of businesses using URL
- Match to the NAICS code in Business Register (Single-Units (SU), Multiple-Units (MU))
  - official data
- Record linkage using business name and address → **hard problem**
- Matched at establishment level, not company level → get NAICS code

Data linkage and cleaning

- Match to the Business Register (Single-Units (SU), Multiple-Units (MU))
- Drop records without any of:
  - Business name, machine readable website (publicly available)
  - Google type, reviews (commercial)
  - 6-digit NAICS code (official)
- Also drop records if less than 10 in a 6-digit NAICS code (model stability)
- Final record set:
  - Single units: 79,500
  - Multi units: 49,000

Hierarchical models

All input is textual

NAICS code = f(business name, homepage, Google types, reviews)

Level	Classes (SU/MU)
2-digit	20/20
4-digit	200/150
6-digit	450/300

Records:
SU = 79,500
MU = 49,000

3 levels x 2 business types

Methods Overview

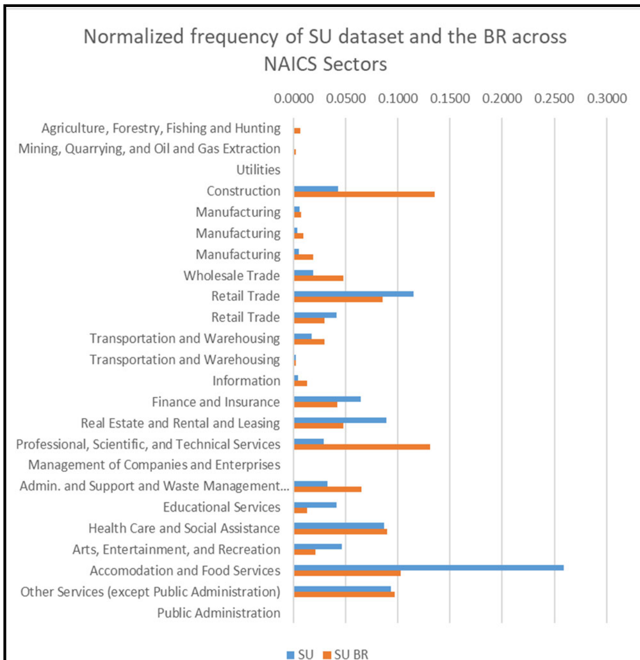
- Natural language processing:
  - TF-IDF (Term frequency-inverse document frequency) — NLP model
- Machine learning models: — 4 ML models
  - Logistic regression
  - Random forest
  - Support vector machines (SVM)
  - Gradient boosting (XGBoost)
- **Stacked model (linear combination of individual models)**
- Generate predictions at 2, 4 and 6 digit NAICS levels
- **3 levels x 2 business types x 4 variables x 1 NLP model x 5 ML models**

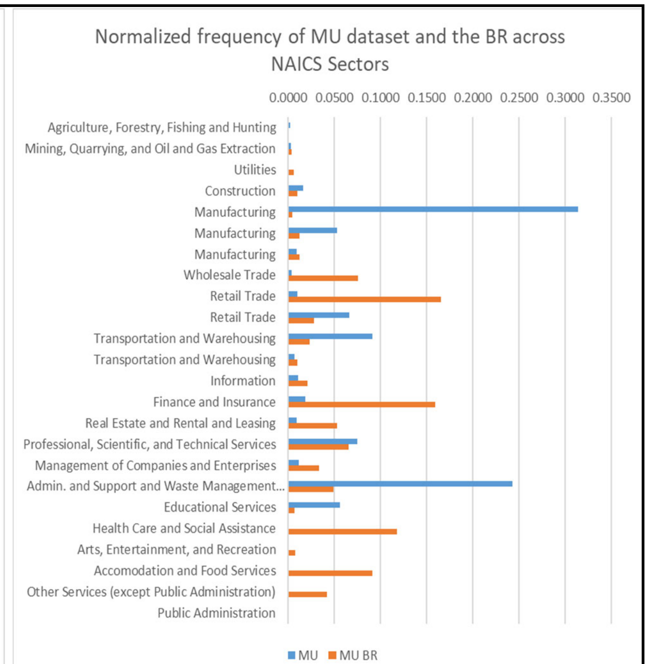
Data coverage at 2, 4, 6-digit NAICS

Code	Data		Business Register		
	SU	**MU**	**BR_{SU}**	**BR_{MU}**	**Overall**
Sector (2 digit)	20	20	20	20	20
Industry Group (4 digit)	200	150	300	300	311
National Industry (6 digit)	450	300	1000	1000	1057

"Overall" column is from publicly available data at https://www.census.gov/eos/www/naics/
All other columns are rounded as per disclosure rules.

11





12

RESULTS

Model stability: 2, 4, and 6-digit accuracy

	SU			MU		
Model	**2 digit**	**4 digit**	**6 digit**	**2 digit**	**4 digit**	**6 digit**
Logistic regression	**0.850**	**0.789**	**0.731**	**0.909**	**0.864**	0.846
Random forest	0.814	0.760	0.709	0.896	0.863	**0.851**
Support vector machines	0.845	0.776	0.717	0.909	0.860	0.840
XGBoost (gradient boosting)	0.793	0.759	0.705	0.878	0.858	0.844
Stack	**0.851**	**0.796**	**0.746**	**0.919**	**0.879**	**0.865**

Accuracy comparison across 4 machine learning models (6-digit)

Testing Procedural Bias

	Accuracy without any minimum probability threshold		Accuracy considering only sub-sample where prob. >= 0.60 accepted for prediction		Dataset coverage with prob. >= 0.60 sub-sample (percent)	
Model	**SU**	**MU**	**SU**	**MU**	**SU**	**MU**
Logistic regression	**0.731**	0.846	0.890	0.950	59.500	74.900
Random forest	0.709	**0.851**	**0.913**	**0.963**	51.000	73.700
Support vector machines	0.717	0.840	0.880	0.944	49.800	68.100
XGBoost (gradient boosting)	0.705	0.844	0.838	0.925	**69.300**	**84.500**
Stack	**0.746**	**0.865**	0.876	0.941	**71.600**	**84.700**

Model comparison with current benchmarks

Model	2 Digit		4 Digit		6 Digit	
	SU	MU	SU	MU	SU	MU
Stack	0.676	**0.919**	**0.657**	**0.805**	**0.627**	**0.798**
SSA	**0.865**	0.724	0.556	0.387	0.452	0.326
BLS	0.872	0.712	0.810	0.666	0.780	0.570

Impact of Information source on overall accuracy (SU establishments)

Testing Declarative Bias

Model variable	Machine Learning Model Used				
	Stack	Logistic regression	Random Forest	Support Vector Machine	XG Boost
Name (N)	**0.702**	0.618	0.700	0.690	0.587
Types (T)	**0.443#**	0.438	0.443#	0.441	0.441
Homepage (H)	**0.679**	0.618	0.659	0.668	0.642
Reviews (R)	**0.614#**	0.533	0.589	0.614#	0.552
N+T	**0.695**	0.664	0.690	0.679	0.634
N+H	**0.742**	0.704	0.705	0.724	0.688
N+R	**0.732**	0.688	0.683	0.716	0.655
N+T+H	**0.747**	0.716	0.719	0.714	0.701
N+T+R	**0.721**	0.700	0.689	0.696	0.663
N+H+R	**0.749**	0.727	0.711	0.729	0.694
N+T+H+R	**0.746**	0.731	0.709	0.717	0.705

Note: Stacked model results bolded where accuracies are statistically better than other machine learning models used (as per Mcnemar test)
Statistically not different

United States Census Bureau | U.S. Department of Commerce Economics and Statistics Administration U.S. CENSUS BUREAU census.gov

17

17

Research Contributions

- Supervised classification of business establishments
- Highly accurate at 2, 4 and 6 digits
  - 2-digit: 20 classes; 6-digit: ~500 classes
- Supports hypotheses that publicly available information can accurately classify business establishments
  - Can share model and results more easily across statistical agencies
- Benchmarks 2 different NLP vectorization methods
- Benchmarks 4 different ML models
- Models are stable

United States Census Bureau | U.S. Department of Commerce Economics and Statistics Administration U.S. CENSUS BUREAU census.gov

18

18

Conclusions and next steps

- Respondent burden reduced
  - Do not have to rely on businesses for survey response and business description
- Easy process for classifying a new case given a trained model
- Can be purposed for production