Improving the American Community Survey's Industry and Occupation Autocoding Process

2024 FCSM Research and Policy Conference

Collaborators: Alexander Zakrzeski (Presenter), Jackson Chen, Yezzi Angi Lee, Lynda Laughlin, Ana J. Montalvo, and Julia Beckhusen

United States Census Bureau

October 22, 2024

United States® Census Bureau

Agenda

1. Introduction
   - 1.1 American Community Survey
   - 1.2 Industry and Occupation (I&O) Autocoding
2. New Preprocessing Techniques
   - 2.1 Managing NA Values
   - 2.2 String Matching
   - 2.3 Lemmatization and Removing Stop Words
3. Improvements to Autocoding
   - 3.1 Utilizing Large Language Models
   - 3.2 Implementing Semantic Search
   - 3.3 Optimizing Through Fine-Tuning
   - 3.4 Evaluating Model Performance
4. Conclusion
   - 4.1 Future Work

1. Introduction

1.1 American Community Survey

Introduction

- The American Community Survey (ACS) is a continuous, nationwide survey conducted by the U.S. Census Bureau, collecting detailed demographic, social, economic, and housing data from U.S. households monthly.
- This research focuses on the two open-ended industry questions and the two open-ended occupation questions in the ACS.

Problem Statement

- How can the industry and occupation descriptions from the ACS be more accurately and efficiently autocoded with the official Census codes, further reducing manual effort and advancing the existing autocoding process?

Industry Questions

1. What is the name of your employer, business, agency, or branch of the Armed Forces?

   Example Response: United States Census Bureau

2. What kind of business or industry is this?

   Example Response: Federal Agency

Occupation Questions

1. What is your main occupation?

   Example Response: Data Scientist

2. Describe the most important activities or duties of your occupation.

   Example Response: Building Machine Learning Models

1.2 Industry and Occupation (I&O) Autocoding

Current Process

- The existing autocoding process uses two supervised machine learning models that incorporate logistic regression, n-gram frequencies, and various demographic explanatory variables.
- The industry model assigns the most appropriate of the 271 Census industry codes, while the occupation model assigns the most appropriate of the 570 Census occupation codes to the responses.
- Both machine learning models have a predicted probability cutoff of 88%.

Table I. Average Monthly Autocoder Rate

Occupation	Industry	Joint[1]
54%	45%	29%

[1] The remaining 71% of the descriptions go to the National Processing Center (NPC) for manual coding.

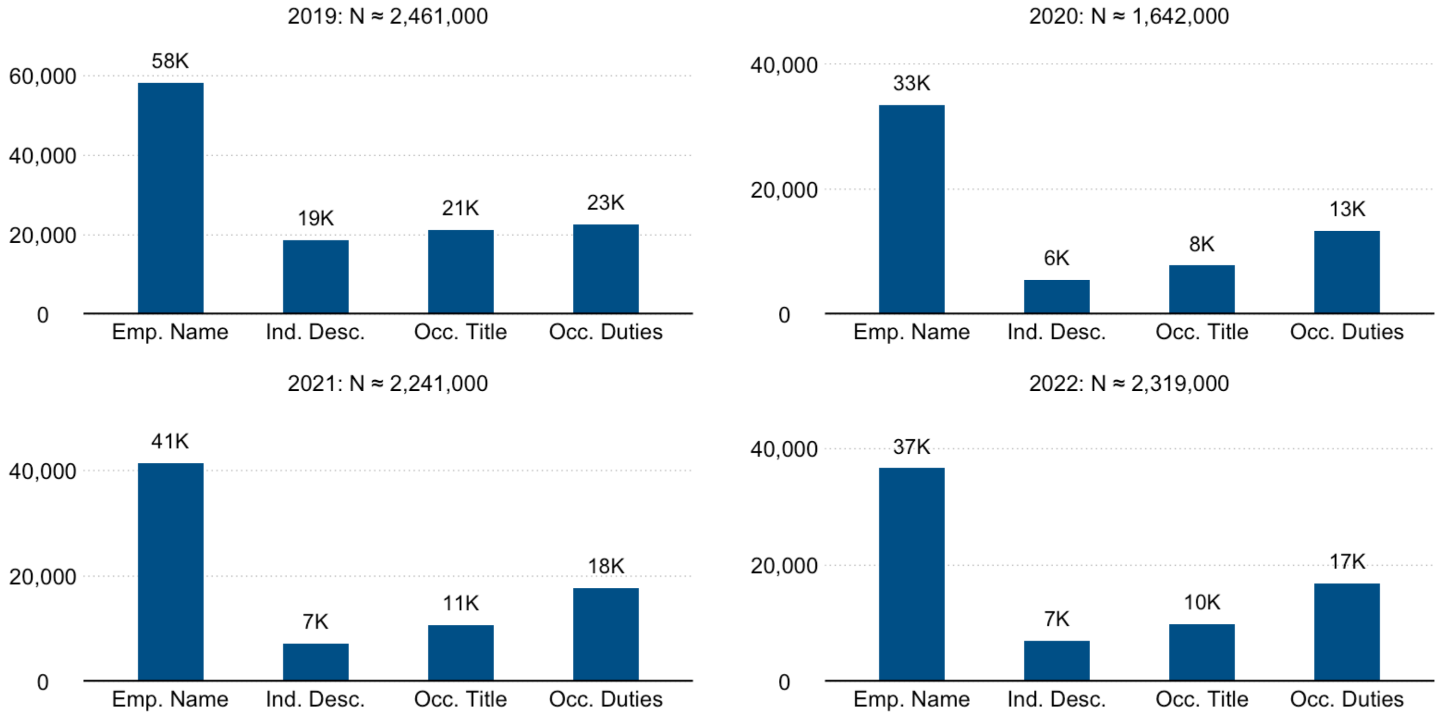
2. New Preprocessing Techniques

2.1 Managing NA Values

Improving Data Quality

- All NA values in the industry and occupation responses from the 2019–2022 ACS were identified and disregarded.
- Then, around 56 different non-useful values (e.g., "Unknown") were identified over four years, changed to NA, and disregarded.
- Regular expressions were used to locate many of these non-useful responses.
- Changing these values to NA and disregarding them reduces potential noise when moving to the modeling process.

Figure I. Counts of I&O Responses Changed to NA, 2019-2022

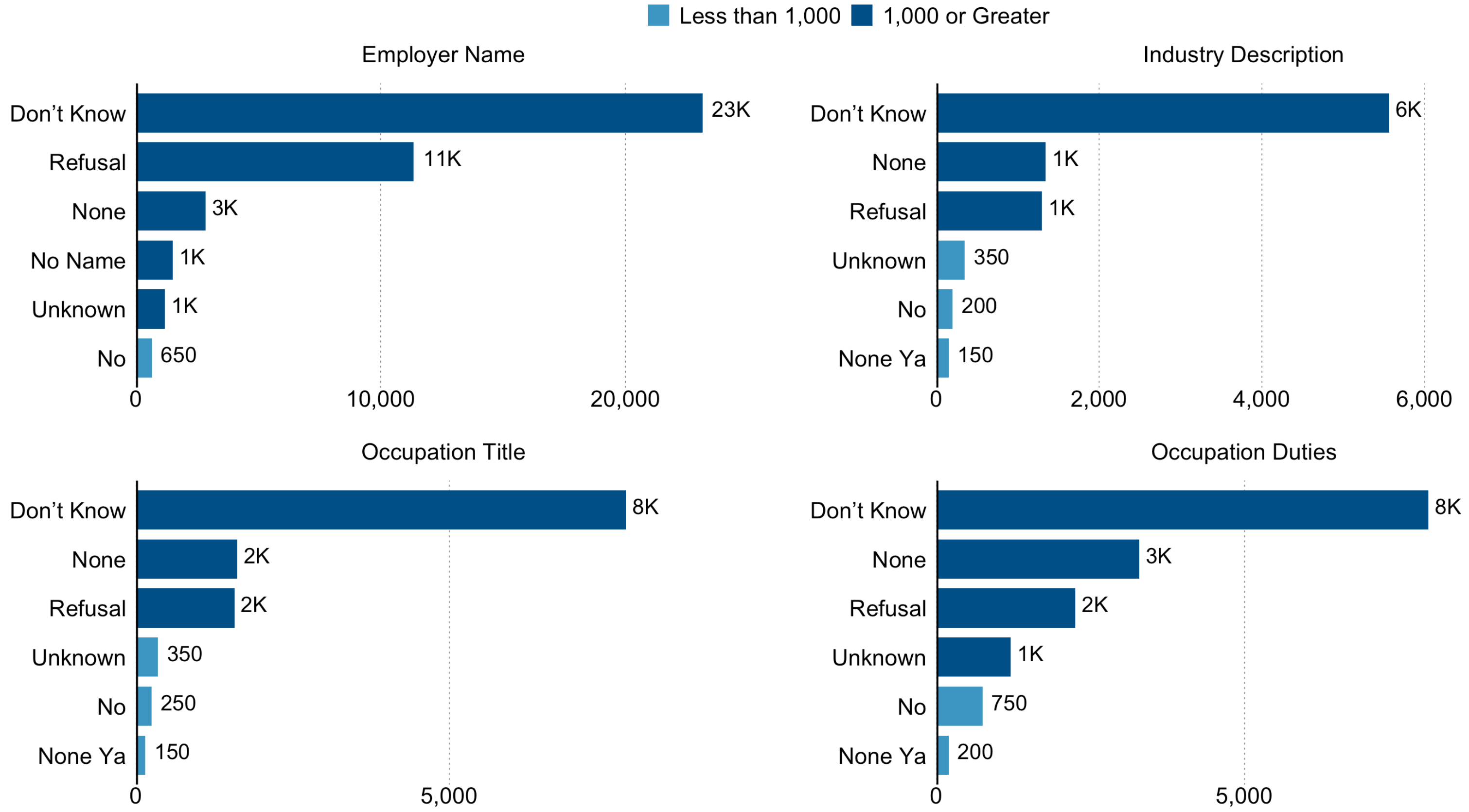


Abbreviations: Emp. = Employer, Ind. = Industry, Desc. = Description, and Occ. = Occupation

2.1 Managing NA Values

The most common value changed to NA in each of the industry and occupation responses each year is "Don't Know", with several others trailing behind and appearing thousands of times on average as well.

Figure II. Average Counts of Top I&O Responses Changed to NA, 2019–2022



2.2 String Matching

Correcting Spelling

- Fuzzy matching, using the Jaro-Winkler distance, and exact string matching were both used to correct responses for a respondent's employer or business name.
- The process described above corrected around 485,000 responses (about 6% of all non-NA values) from the 2019-2022 ACS responses.

Detecting Patterns

- Logic-based string matching with regular expressions was used to identify acronyms and other abbreviations in all the industry and occupation responses, such as detecting all values with four or fewer characters in all capital letters.

Table II. Examples of Employer Name Corrections Using Fuzzy Matching

Employer Type	Incorrect Employer Name	Corrected Employer Name	Similarity Score[1]
Government	Internal Revenue Services	Internal Revenue Service	0.99
Government	United State Census Bureau	United States Census Bureau	0.97
Government	United States Marine Corp	United States Marine Corps	0.99
Government	United States Postal Sevice	United States Postal Service	0.99
Educational	University of FL	University of Florida	0.95
Educational	University of Mich	University of Michigan	0.96
Educational	University of Pittsburg	University of Pittsburgh	0.99
Government	VA Hosptial	VA Hospital	0.98

[1] The Jaro-Winkler Distance's similarity score ranges between 0 and 1.

2.3 Lemmatization and Removing Stop Words

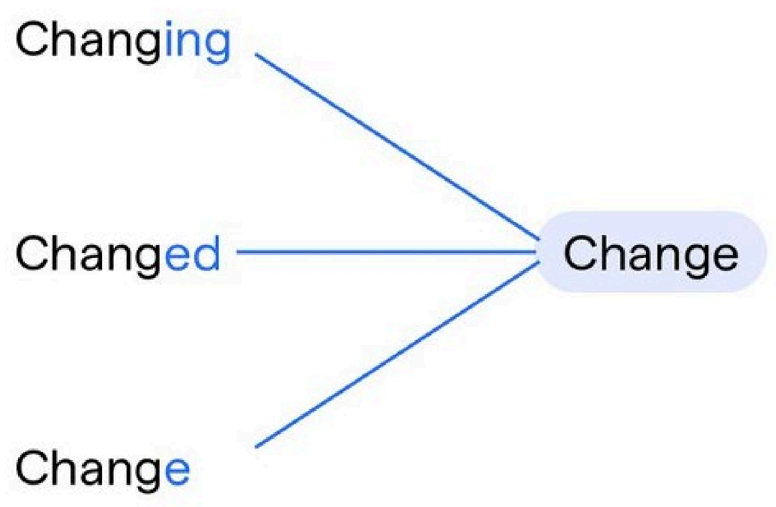
Reducing Words to Their Lemma

- After tokenizing industry and occupation responses to unigrams, lemmatization was implemented to reduce words to their lemma (root form).

Removing Noise

- Stop words were removed in the preprocessing steps to reduce noise from common words that are not meaningful in semantics.
- Certain stop words, such as "IT", which can stand for Information Technology, were retained and not removed after a careful examination of the list of stop words in the context of industries and occupations.

Figure III. Example of Lemmatization



3. Improvements to Autocoding

3.1 Utilizing Large Language Models

Understanding Responses

- Large language models (LLMs) represent sentences as vectors, capturing semantic information that helps them understand the meaning of sentences.
- Consequently, Census codes can be assigned without relying on historical data to train a machine learning model.
- With numerous models available, optimization is focused on both size and performance to maximize efficiency.

Retrieving Data

- The large language model "gte-large" was selected and augmented with the Census's Alphabetical Indexes, referenced in Table III, and the Census's Occupation Code List, referenced in Table IV, to ensure that only relevant codes and titles are retrieved.

Table III: Examples of Entries from the Alphabetical Index of Occupations

Occupation Title	Census Code	Industry Restriction
Biologist	1610	None
Biophysicist	1610	None
Biostatistician	1230	None
Bird Tender	4350	0180
Birth Attendant	3603	None
Biscuit Maker	7840	None

Table IV: Examples of Entries from the 2018 Census Occupation Code List

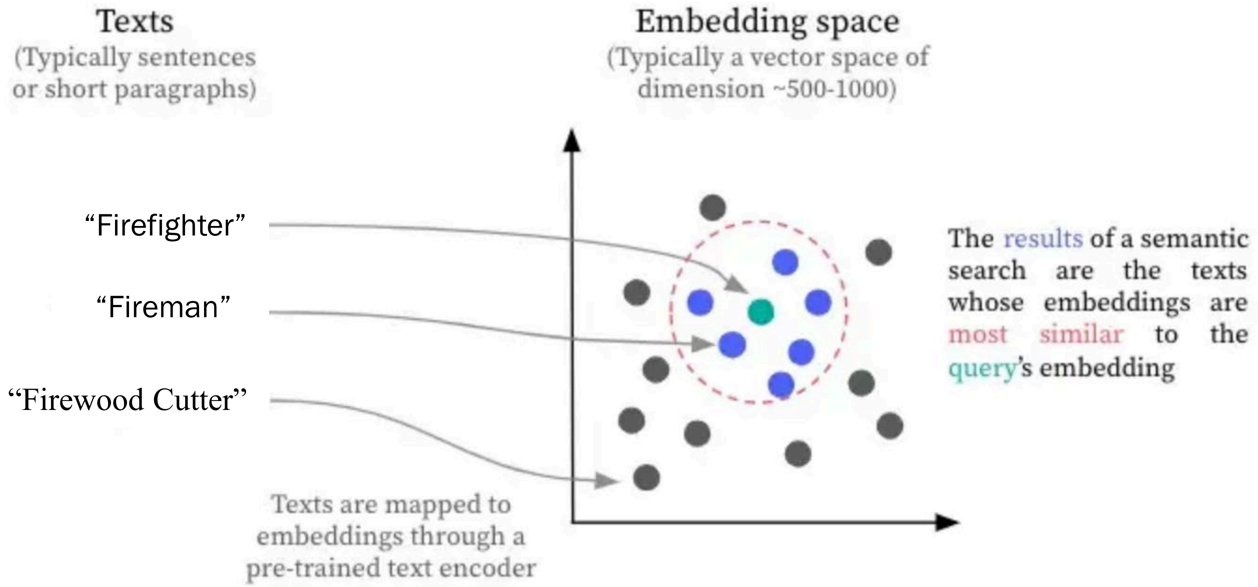
Census Occupation Title	Census Code
Life, Physical, and Social Science Occupations:	1600-1980
Medical Scientists	1650
Astronomers and Physicists	1700
Physical Scientists, All Other	1760
Economists	1800
Survey Researchers	1815

3.2 Implementing Semantic Search

Functionality

- Semantic Search utilizes vectors to capture the meaning of a sentence, using a large language model to generate vector embeddings.
- Cosine similarity is applied to compare different vector embeddings, calculating similarity and providing a similarity score. Results are then ranked by this score.
- Vector embeddings are stored in a vector database for efficient retrieval, reducing the time needed to generate embeddings.

Figure IV. Example of Semantic Search

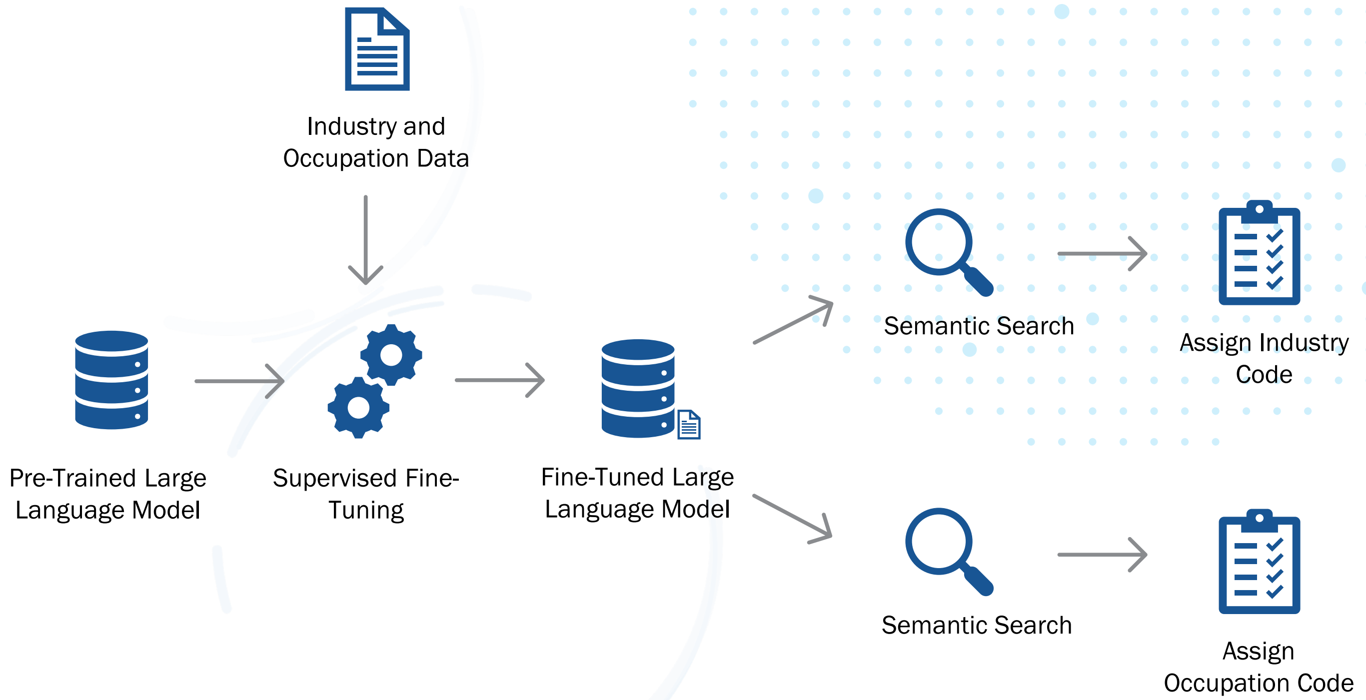


3.3 Optimizing Through Fine-Tuning

Improving Performance

- Although large language models are powerful, performance can still be improved by fine-tuning the chosen model to better understand the occupation and industry domain.
- Fine-tuning involves changing the weights in the chosen model to adapt to the provided data.
- The model was fine-tuned using the Census's Alphabetical Indexes of Industries and Occupations, and this technique provided the most significant improvement in performance compared to any other techniques tested.

Figure V. Autocoding Process



3.4 Evaluating Model Performance

Testing

- The model was tested on the 2019 ACS Public Use Sample of Occupation and Industry Write-ins dataset; an example of this is in Table V.
- This dataset contains 10,449 entries, with each entry including a full response and the assigned codes for industry and occupation.

Comparison

- The current autocoding process has a rate of 29% in assigning the best code for both industry and occupation, while the new autocoding process improved this rate to 51% (+22%).

Table V: Examples of Entries from the ACS Public Use Sample Dataset

Ind. Code	Industry Description	Occ. Code	Occupation Title	Occupation Duties
9480	Regional Office of Education	2016	Youth Outreach	Trainee to At-Risk Students
9470	Police	2016	Victim Services Unit	Servicing Crime Victims
8270	Nursing Home	2016	Social Services ASST/CNA	Spending Time with Residents
8270	Nursing Home	2016	Social Services Coordinator	Assist Residents with their Needs
0770	Military Engineering	2016	Small Business Deputy	Groups
9470	Law Enforcement	2016	Community Service Officer	Calls for Service

4. Conclusion

4.1 Future Work

Improvements

- A plan is in place to leverage previous data to enhance assignment accuracy and identify whether code assignments have been previously encountered.
- A quality control process will be established to evaluate the autocoder's performance.
- Additional fine-tuning will be incorporated, as it has proven effective in significantly improving assignment accuracy; fine-tuning specific to occupation and industry groups will also be explored to further enhance the model.

Questions