Using artificial intelligence tools to automate manual data processing and generate alternative data sources in the U.S. Census Bureau

Sara Alaoui[1], Louis Avenilla[1], Patrick Campanello[1], Yi-Tan Chang[1,2], Ugochukwu Etudo[1,2,3], Nisarahmad Hundewale[1,2], Haley Hunter-Zinck[1], Jennifer Hutnick[1], Yathish Kolli[1], Arezou Koohi[1], Anup Mathur[1], Lydia Shia[1]

[1]Center for Optimization and Data Science, U.S. Census Bureau [2]Brite Group [3]Virginia Commonwealth University | Authors listed in alphabetical order by last name

Introduction

The Center for Optimization and Data Science (CODS) is a research center within the U.S. Census Bureau that aims to incorporate aspects of data science, including artificial intelligence (AI), into survey operations and data processing.

The center's main objectives in the U.S Census include:
- Automating manual data processing procedures
- Generating alternative data sources to augment traditional data collection streams

We introduce several projects that cover areas in AI:
1. Supervised and unsupervised record linkage
2. Text classification of survey responses
3. Topic modeling of survey open responses fields
4. Web scraping with AI-driven information extraction
5. Named entity recognition in Census field notes

Future goals of the center include
- Evaluating the potential of generative AI tools
- Instituting continuous monitoring frameworks for AI models
- Streamlining AI prototypes from development to production

Our goal is to integrate high performing, efficient, and low bias AI models into survey data processing to continue to improve processing efficiency and data quality for U.S. Census data products.

(1) Record linkage

Research area : 2020 Post-Enumeration Survey (PES) [2]

Problem: The 2020 PES used clerical review to manually match PES to Decennial Census enumerated person records after initial computerized record matching, necessitating a large clerical review workload.

Objective: Enhance the computerized record matching procedure with the use of AI models to reduce the clerical review workload.

Methods: Use supervised and unsupervised record linkage models using comparative feature vectors derived from person record names, demographics, and other data elements to predict matches.

Evaluation: Calculate recall and precision against a held out clerically reviewed dataset and error rate for matched sample identifiers.

(2) Text classification

Research area: Current Population Survey (CPS) [1]

Problem: Name entry values need to be quickly screened for input to survey mode selection decisions.

Objective: Develop a model to automatically classify survey name entries as name, description, or invalid responses with high performance and efficiency.

Methods: Develop guidelines for categorization and programmatically encode rules to generate labeled training sets. Fine tune transformer models for the text classification task [3].

Evaluation: Calculate precision and recall against a manually labeled set of name entries. Simulate implementation to estimate automation and error rates.

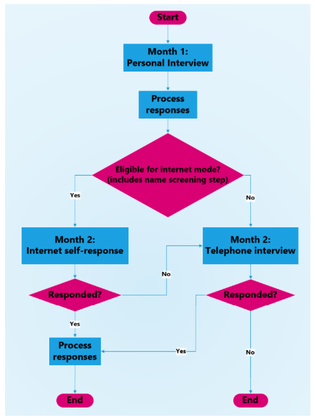

Figure 1: Integration of the AI model for name screening in the Current Population Survey workflow.

(3) Topic modeling

Research area : Teacher Follow-up Survey (TFS) [4]

Problem: Free text responses are labor intensive to analyze and thematically categorize manually.

Objective: Partially automate discovery of themes in the TFS open response question asking about the effects of the COVID-19 pandemic on staffs' teaching experience.

Methods: Use embedding based topic modeling methods implemented in Top2Vec [5] and BERTopic [6].

Evaluation: Clustering metrics and manual interpretation of topic themes along with cross tabulations with other collected response data variables.

(4) Web scraping

Research area : 2020 Decennial Census [7]

Problem: Use public websites as an alternative data source for Group Quarter (GQ) frame augmentation and validation.

Objective: Develop generalized extraction methods for data elements such as addresses and capacity from a GQ website's content.

Methods: Generate labeled datasets and fine-tune question-answer and named entity recognition transformer models [3].

Evaluation: Calculate precision and recall against a held out curated dataset.

```
Question: How many beds are in the shelter?
Context: The shelter has 100 beds.
Answer: {'score': 0.9967477321624756, 'start': 16, 'end': 24, 'answer': 100}

Question: How many beds are in the shelter?
Context: The shelter holds 20 women.
Answer: {'score': 0.9925102591514587, 'start': 18, 'end': 26, 'answer': 20}

Question: How many beds are in the shelter?
Context: The shelter Can serve up to 12 men and 8 women
Answer: {'score': 0.9411935806274414, 'start': 28, 'end': 46, 'answer': 20}

Question: How many beds are in the shelter?
Context: This shelter does not state the number of beds.
Answer: {'score': 0.37632274627685547, 'start': 32, 'end': 46, 'answer': None}

Question: How many beds are in the shelter?
Context: The shelter has eighty beds.
Answer: {'score': 0.9956754446029663, 'start': 16, 'end': 27, 'answer': 80}

Question: How many beds are in the shelter?
Context: The shelter is open for 24 hours.
Answer: {'score': 0.08937286585569382, 'start': 4, 'end': 11, 'answer': None}
```

Figure 2: Output of the fine-tuned question-answer transformer model for extracting a facility's capacity on synthetic example documents (contexts).

(5) Named entity recognition

Research area : 2020 Decennial Census [7] and 2020 Post-Enumeration Survey (PES) [2]

Problem: Field notes contain valuable information, but this information is labor intensive to extract manually.

Objective: Automate the extraction of target data elements such as person names, physical addresses, and dates to incorporate in downstream processes.

Methods: Use pretrained, transformer-based named entity recognition (NER) models and train custom NER models for application to specific data elements using packages like spaCy [8].

Evaluation: Calculate precision and recall against a held out curated dataset.

Future directions

- **Evaluating the potential of generative AI**: The release of generative large language models provides a new suite of tools for generalized natural language processing. We plan to benchmark these models against currently used techniques for our applications. Due to data use restrictions, we will only evaluate open source and downloadable models.
- **Streamlining models from development to production**: The operationalization of our models start with the beginning of the data science research project. We aim to further streamline the transition for modeling frameworks from research into production by continuing to invest in infrastructure and best practices for software development, user interface design, dependency management, documentation, and validation.
- **Implementing continuous monitoring frameworks for model bias and model drift**: For models in production, tracking metrics that calculate and detect reduced performance across sensitive subgroups and over time will be essential for maintaining performance in current operations.

References

- [1] "Current Population Survey (CPS)." [Online]. Available: https://www.census.gov/programs-surveys/cps.html
- [2] "Post-Enumeration Surveys." [Online]. Available: https://www.census.gov/programs-surveys/decennial-census/about/coverage-measurement/pes.html
- [3] T. Wolf et al., "HuggingFace's Transformers: State-of-the-art Natural Language Processing," arXiv, Jul. 13, 2020. Accessed: Apr. 24, 2024. [Online]. Available: http://arxiv.org/abs/1910.03771
- [4] "The 2021-22 Teacher Follow-up (TFS) and Principal Follow-up (PFS) Surveys." [Online]. Available: https://nces.ed.gov/surveys/ntps/participants_2022.asp
- [5] D. Angelov, "Top2Vec: Distributed Representations of Topics," 2020.
- [6] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," arXiv preprint arXiv:2203.05794, 2022.
- [7] "2020 Census." [Online]. Available: https://www.census.gov/programs-surveys/decennial-census/decade/2020/2020-census-main.html
- [8] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spaCy: Industrial-strength Natural Language Processing in Python." [Online]. Available: 10.5281/zenodo.1212303

Disclosure statement: Any opinions and conclusions expressed herein are those of the authors and do not reflect the views of the U.S. Census Bureau. This presentation does not contain sensitive information including Title 13, Title 26, Title 5, other controlled unclassified information, or administratively restricted information. All described models are for intended for internal research use only.

Acknowledgements

We would like to thank our collaborators at the U.S. Census, National Center for Education Statistics, and Bureau of Labor Statistics for expertise and funding.

Center for Optimization and Data Science (CODS)

United States® Census Bureau