Using Machine Learning to Categorize Person Name Entry Responses in the Current Population Survey

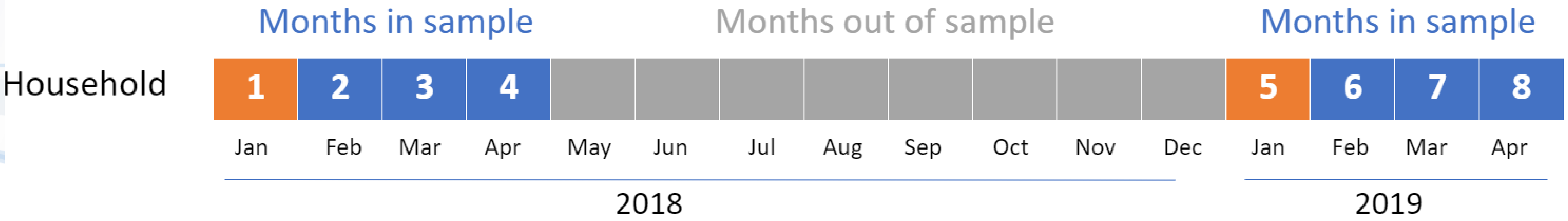
Haley Hunter-Zinck
Center for Optimization and Data Science
U.S. Census Bureau

AAPOR
May 17, 2024

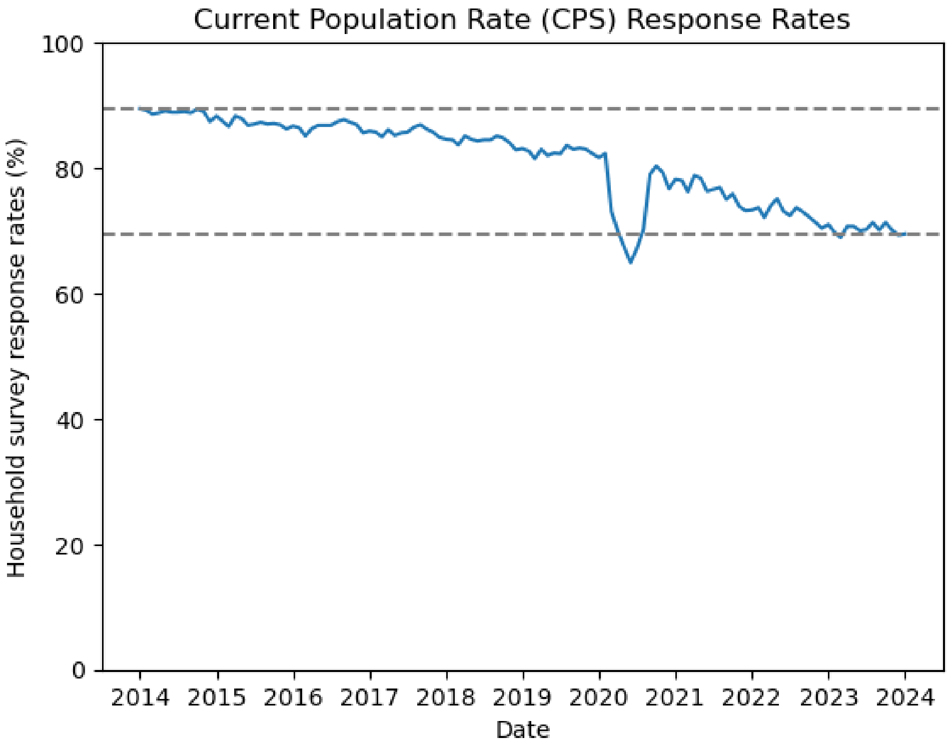
United States® Census Bureau

The Current Population Survey (CPS)

- Sponsored by the U.S. Census and the Bureau of Labor Statistics
- Collects household employment and income information
- Monthly survey
- Households are in survey for 8 months with an 8-month gap



CPS modernization efforts are underway to tackle declining response rates



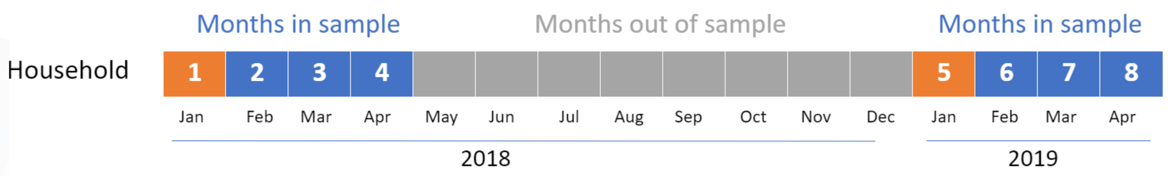
- CPS is currently conducted via telephone or personal interviews
- CPS will provide an Internet self-response option by 2027

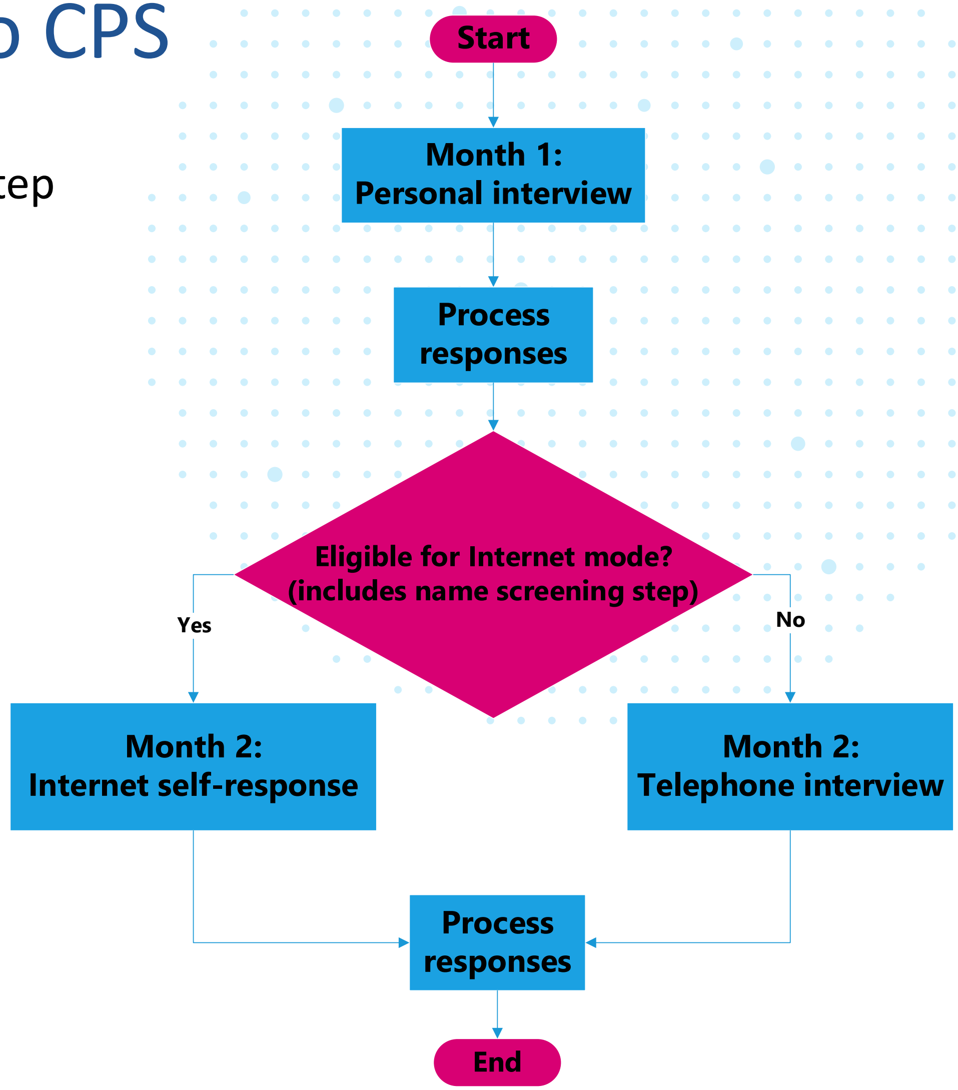
https://www.bls.gov/osmr/response-rates/household-survey-response-rates.htm

https://www.census.gov/programs-surveys/cps/about/modernization.html

Here, we present a simplified workflow for incorporating Internet self-response into CPS

Eligibility for Internet self-response includes a name entry screening step





United States® Census Bureau

Household respondents must have a valid name entry to be eligible for Internet mode

- The survey will display the respondents’ name to them to verify their identify
- Name entry must be …
  - Appropriate
  - Uniquely identifiable
- Respondents may refuse to give the interviewer their name
- The interviewer will enter a description or refusal in the name entry field
  - Resident
  - Jane Doe
  - Son
  - Refused
  - …

CPS needs an efficient and high performing name screening tool to categorize name entries

Options

- Manual curation
- Automated rules
- Machine learning (ML) model

Why ML?

Desired attribute	Manual	Rules	ML
Measure of certainty	✗	✗	✓
Flexibility	✓	✗	✓
Consistency	✗	✓	✓
Efficiency	✗	✓	✓

Three name entry categories

Label	Description	Examples
name	An actual person's name or initials	Debbie Chang Haley Hunter-Zinck D. C.
description	A word or phrase that is not a name but describes a person's role, profession, or familial relationship.	Head of household Sister Son-in-law
invalid	Any inappropriate words or phrases, generic placeholders, typos or completely non-alphabetic entries found in one or more words in names	Anonymous Jane Doe 000

We resolved entries adhering to more than one category via the following precedence rules:

invalid > **description > name**

We start with an unlabeled dataset of first and last names from the CPS

Gold
- Perform manual curation of name entries (gold standard dataset)
- Develop categorization guidelines by consensus

Silver
- Encode guidelines programmatically as rules
- Automatically categorize name entries by rules (silver standard dataset)

Training
- Construct features (derived data elements) from each name entry
- Train a supervised machine learning model based on features and silver labels

Validation
- Use the gold labels to validate the results of the trained machine learning model
- Perform error analysis on the trained machine learning models

We benchmarked supervised machine learning models against rules-based annotations

1. Feature engineering and classical machine learning (ML) classifier
2. Sentence transformer and classical ML classifier
3. Fine-tuned transformer model for text classification

We calculated 6 classes of features to represent the name entries as input to the classical ML model

Feature sets

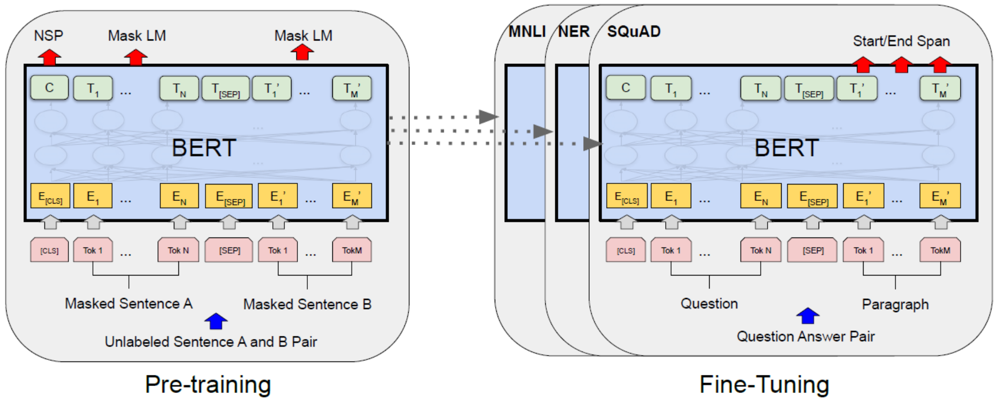
1. Character and word counts
2. Gazetteer (word list) based similarity
3. Typos check
4. Profanity score check
5. Named entity recognition and part of speech
6. Document level embeddings

Model training

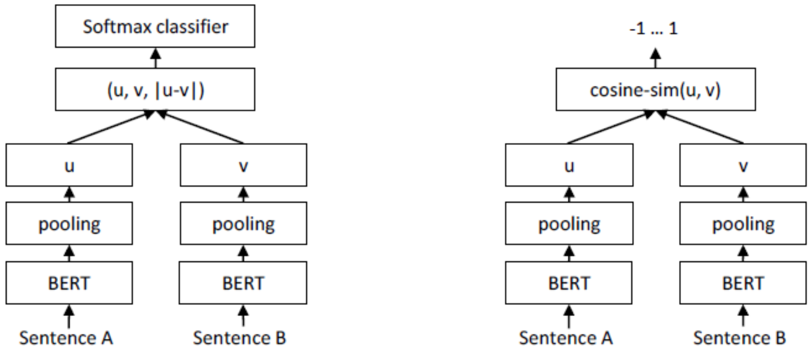
- XGBoost
- Hyperparameter tuning

Transformer-based text classification methods represent name entries as semantically meaningful vectors

- Use pretrained models for representing responses as vectors (embeddings)
  - Transformer such as BERT
  - Sentence transformer
- Fine-tune for classification task
  - Train final layers for classifying embedded responses
  - Input to classical ML model

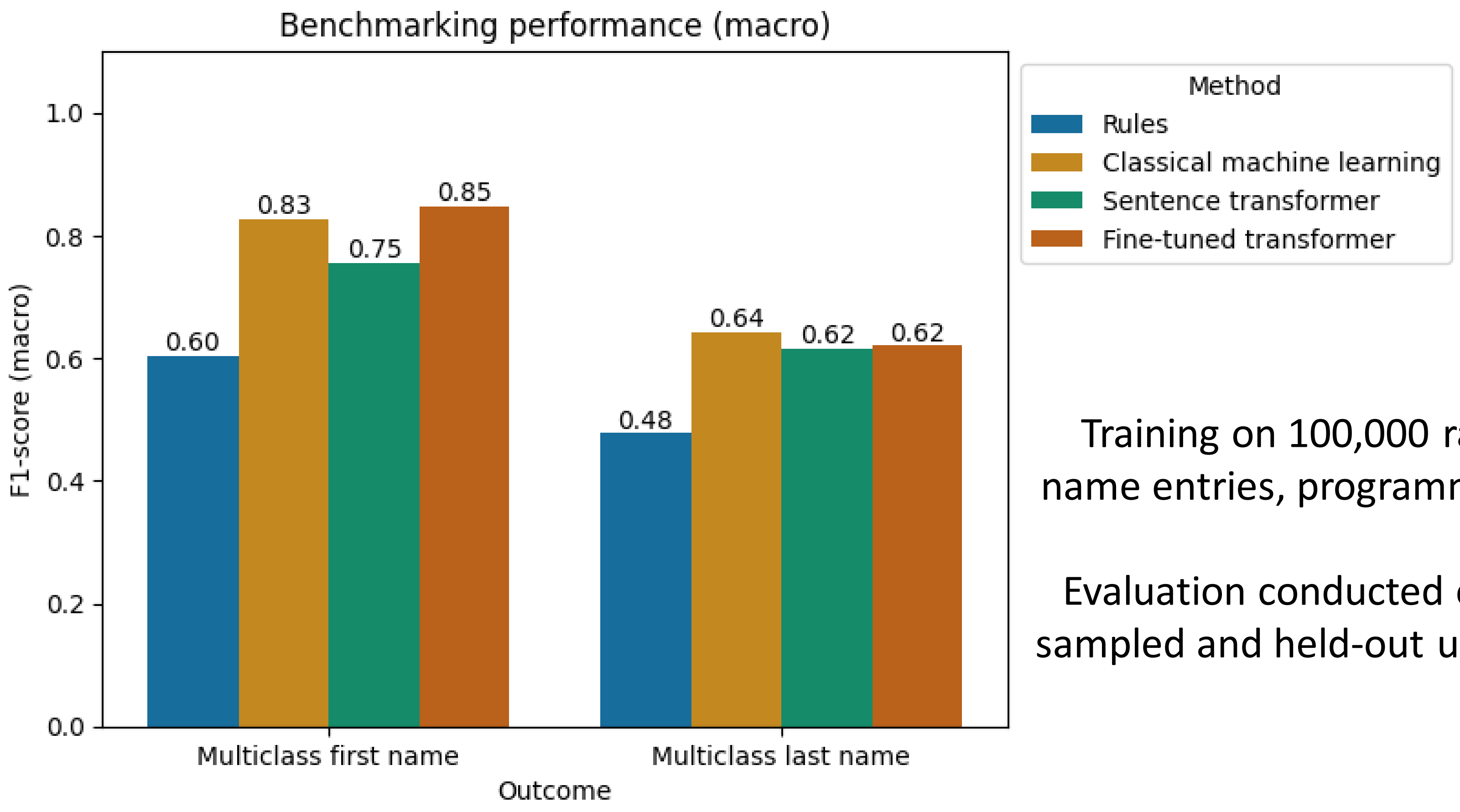


J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” arXiv, May 24, 2019. Accessed: Mar. 27, 2024. [Online]. Available: http://arxiv.org/abs/1810.04805



N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.” arXiv, Aug. 27, 2019. Accessed: Mar. 04, 2024. [Online]. Available: http://arxiv.org/abs/1908.10084

Machine learning models outperform rules-based annotation



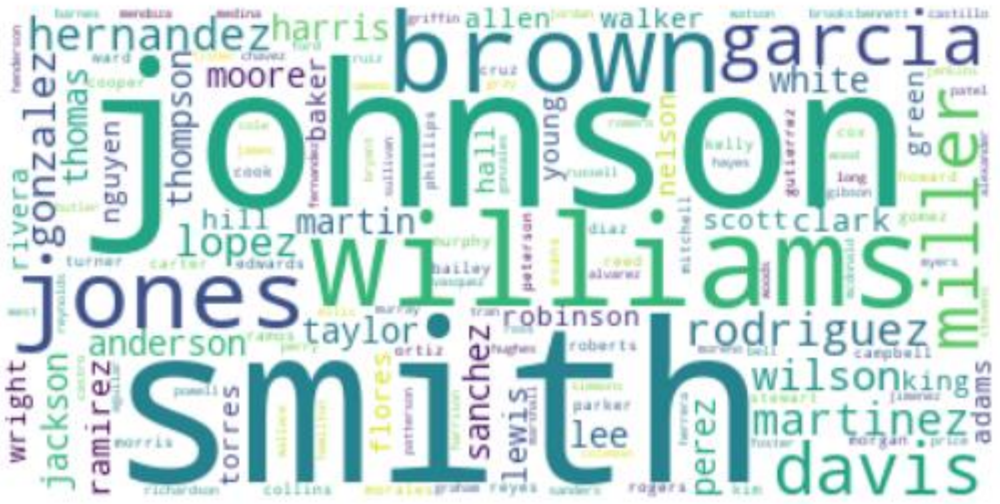
Training on 100,000 randomly sampled name entries, programmatically annotated.

Evaluation conducted on 5,000 randomly sampled and held-out unique name entries.

The most common prediction error for fine-tuned transformers occurs when names are predicted as descriptions

Surnames that are also dictionary words are often predicted as descriptions.

Common surnames in the 2010 U.S. Census

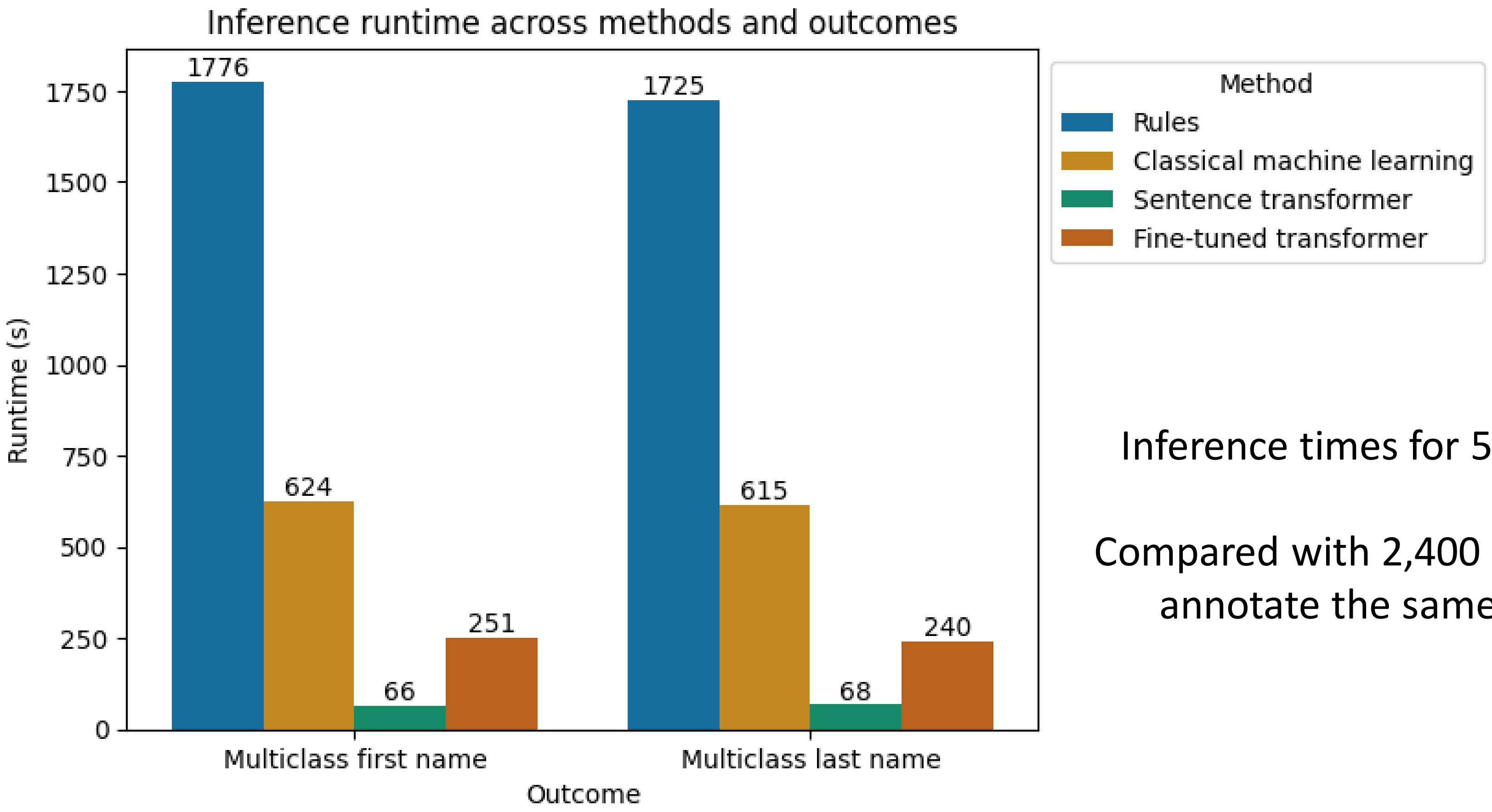


Common surnames that are also English or Spanish words

Word clouds generated from 2010 U.S. Census surname data publicly available at
https://www.census.gov/topics/population/genealogy/data/2010_surnames.html

United States® Census Bureau

Model inference, especially with sentence-transformer encoding, is faster than rules-based inference



Inference times for 5,000 name entries.

Compared with 2,400 seconds to manually annotate the same name entry set.

Conclusions

- Machine learning models provide increased performance and efficiency over rules-based strategies for name screening
- We train a high performing name screening model with programmatically labeled data
- Fine-tuned transformers provide a balance between performance and efficiency

Acknowledgements

Thank you

- **Yi-Tan Chang**
- Louis Avenilla
- Anup Mathur
- **Kyra Linse**
- James Back

Support

- Current population survey

Contact

- haley.s.hunter-zinck@census.gov

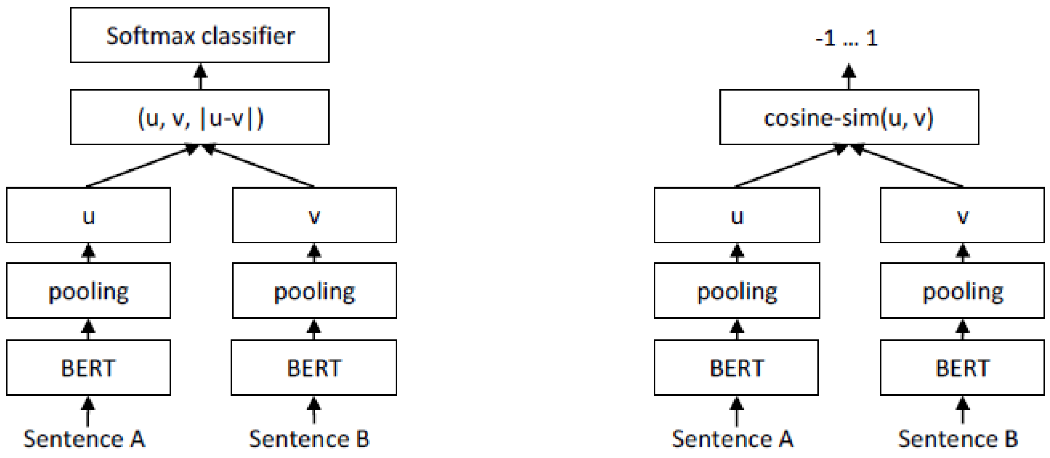
United States® Census Bureau

Sentence transformer

- Model: all-mpnet-base-v2

  https://huggingface.co/sentence-transformers/all-mpnet-base-v2

- Encodes each text as a 768-dimension vector
- Fine-tuned for clustering and semantic search
- Use XGBoost model to predict name entry categorization



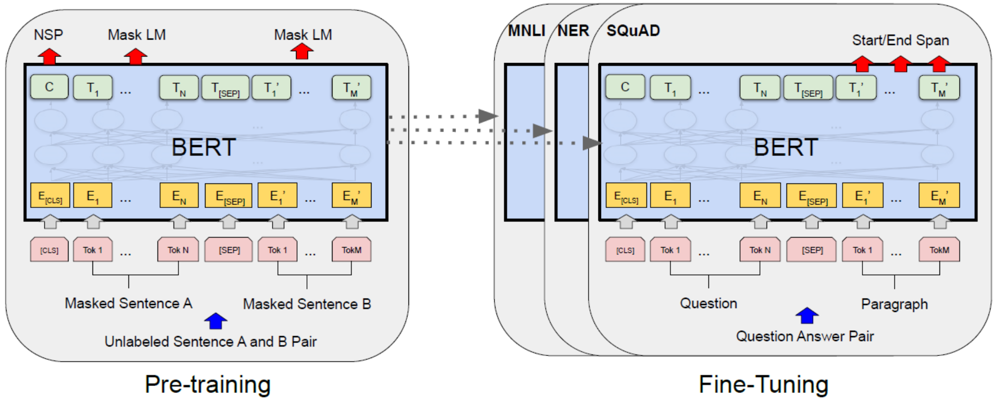
N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.” arXiv, Aug. 27, 2019. Accessed: Mar. 04, 2024. [Online]. Available: http://arxiv.org/abs/1908.10084

Fine-tuned transformer

- Model: distilled RoBERTa

  https://huggingface.co/distilbert/distilroberta-base

- Encodes each text as a 768-dimension vector
- Fine-tune for name entry classification task



J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv, May 24, 2019. Accessed: Mar. 27, 2024. [Online]. Available: http://arxiv.org/abs/1810.04805