Westat

Improving Lives Through Research®

Can I have your name? Classification of names for case prioritization in household CAPI surveys

Xin (Rosalynn) Yang, Anil Battalahalli Sreenath, Ting Yan

Westat @ AAPOR 2022 — Take Survey Research to New Heights

Background

- In household CAPI surveys, interviewers are instructed to collect sampled respondents' names among other contact information as part of the screener.
- In the name fields, interviewers tend to enter any name info they can get in the fields…

Q. Can I have your name?

- *e.g., John Doe*
- *e.g., John*
- *e.g., Doe*
- *e.g., Husband, daughter (when no name is provided)*

Background

Q. Can I have your name? ➡ Type of name information

e.g., John Doe	Full name
e.g., John	First name only
e.g., Doe	Last name only
e.g., Husband, daughter	Generic labels

- In multi-stage household surveys, the type of name information given by respondents on the screener is highly predictive of interview response propensity
  - In a prior study, we found that respondents who gave full names and first names were significantly more likely to complete the interview.

Background

- Goal: use “name type” in the response propensity model to inform case prioritization
- Challenges:
  - The need to process a large amount of name information
  - Can’t have interviewers code on-site (may bias the data)
- Solution:
  - Real-time name classification via natural language processing and machine learning

Task and possible solutions

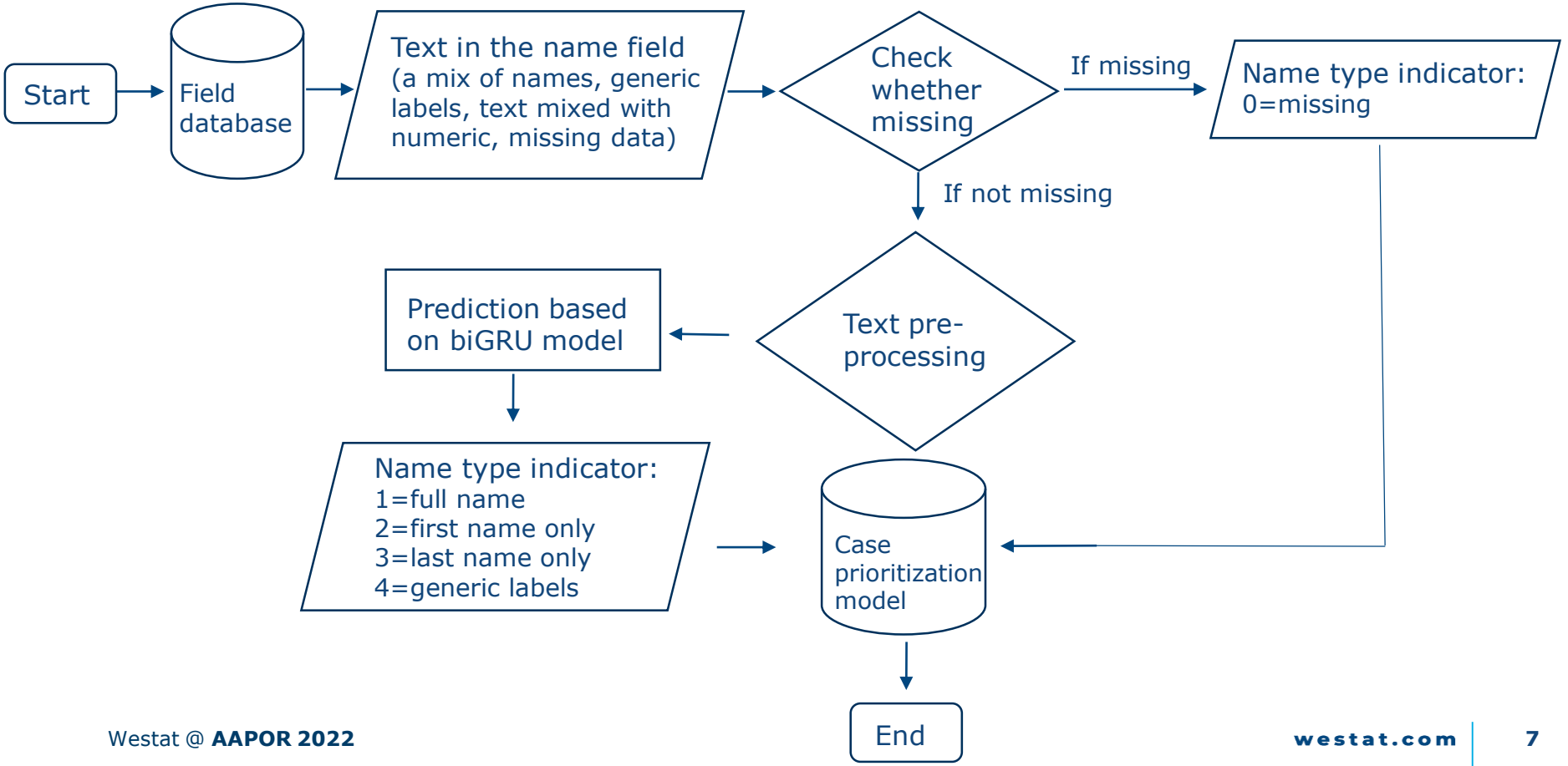
- Task: automatically classify the name information in text format into different classes of a name type indictor
- A multi-class text classification problem
- Possible solutions:
  - Text matching
  - Neural networks

Our method

- Our approach: natural language processing and machine learning model (BiDirectional Gated Recurrent Units (GRU))
- Model trained with publicly available name data from the Social Security Administration and FiveThirtyEight's GitHub data repo*
  - Cross-validation accuracy: 98%
- Model performance evaluated with prediction accuracy on real field data from a large national-scale household health survey

*Source of names: https://www.ssa.gov/OACT/babynames/limits.html
data/most-common-name at master · fivethirtyeight/data · GitHub

Method


Start
Field database
Text in the name field (a mix of names, generic labels, text mixed with numeric, missing data)
Check whether missing
If missing
Name type indicator: 0=missing
If not missing
Text pre-processing
Prediction based on biGRU model
Name type indicator: 1=full name 2=first name only 3=last name only 4=generic labels
Case prioritization model
End

Compare against traditional approaches

Approach	Prediction time	Prediction accuracy
Bi-directional GRU	0.375 **ms** for a single data point	0.9555
Fuzzy string match (via difflib)	0.371 **seconds** for a single data point	0.8089

~100% accuracy identifying full names and generic labels;
Lower accuracy for first names vs. last names

Discussion

- We trained a model (bi-directional GRU) that can classify name text information into whether it is a full name, a first name only, a last name only, or a generic label
- Deployed as a REST API, it can be incorporated into any propensity model/case prioritization pipelines
- Future research:
  - Real-time processing of interviewer notes to inform field operations

Thank You

Rosalynn Yang

rosalynnyang@Westat.com