Using Natural Language Processing and Machine Learning to Quickly Classify Open Text Field Comments in a Longitudinal Study

Catherine Billington, Jiating (Kristin) Chen, Andrew Jannett, Gonzalo Rivero, John Riddles

Government Advances in Statistical Programming (GASP!) Workshop, Washington D.C.
September 22, 2019

Agenda

Background

Problem

Data

Methodology

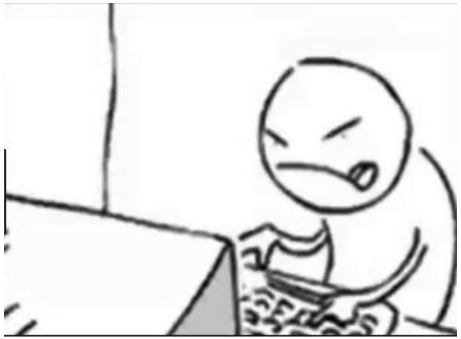
Discussion

Background

- Field comments from Medical Expenditure Panel Study (MEPS): Interviewers type comments to request updates or corrections to the data they collect earlier in the interview
- Why interviewers make comments:
  - MEPS asks respondents to recall specific information about medical events over the last few months. It is usual for them to correct or add to the information collected earlier in the interview.
  - MEPS is large: 1267 questions, 90 minutes to administer. Backing-up to edit earlier responses is time-consuming and can cause errors.
- To avoid backing up, field interviewers can leave comments on the case file. Interviewers must select a category from a drop-down each time they enter a comment, to facilitate data processing.

Background (Cont.)

- Challenges for comment processing under the current approach:
  - Costly to address: text; large amount
  - Quick turnaround
- What if we could determine the correct comment category more quickly?



Technicians process comments:

- *Correct category?*
- *Update responses?*

Problem

- Can we use natural language processing and machine learning to **quickly** *classify* open text field comments **with acceptable accuracy**?

Computer:
- *Suggest category*

Technicians :
- *Choose correct one*

Data

› 4400 comments, 16 categories

Category	Sample Comments	%
Edit/Delete health care event	It was not a hospital stay. Nancy visited to a private doctor clinic once a week on Mondays for allergies.	22.3
Edit Charge/Payment details		16.8
Edit health insurance information		15.36
Edit/Delete prescribed medicine(s)	Correct spelling of following prescriptions: OXYCODONE & PREDNISONE.	8.61
Edit health care utilization details		7.7
Edit employment information		7.3
Edit/Delete provider information	Nancy, PID 103, visited Dr. Grace Yang on 1/16/18 and 2/14/18 at 1600 Research Blvd, Rockville MD 20850. 301-251-1500. Copay $50.	7.3
Change in household roster		2.84
Add prescribed medicine(s)		2.59
Edit/Delete condition(s)		2.3
Other comment		2.14
Person level refusal within RU		2.09
Add Purchase of Eyeglasses/Contact Lenses		0.89
Add Purchase of Other Medical Expenses		0.89
Edit RU Member Name		0.64
Edit RU member Date of Birth		0.27

Methodology

- Feature Engineering (Python: spaCy, NLTK, re)
  - 11 features
    - 9 key information items in text
    - 2 Metadata from the abbreviations of field name and question number
  - Text features: Term Frequency – Inverse Document Frequency matrix
- Machine learning (Python: scikit-learn)
- Model Deployment for production (Python: Flask, Nginx, Angular JS, Docker)

Methodology – Feature Engineering (9 key information items in text)

› Different combinations of features are associated with different categories.

Category	Date	$	Zip Code	Phone	Name	Verbs and synonyms	Provider	Drug	Insurer
Edit/Delete provider information			✓	✓			✓		
Edit health insurance information									✓
Edit/Delete prescribed medicine(s)								✓	
Edit Charge/Payment details		✓							
Add Purchase of Eyeglasses/Contact Lenses		✓							
Add Purchase of Other Medical Expenses		✓			✓				
Edit RU Member Name									
Edit RU member Date of Birth	✓								

Methodology – Feature Engineering

Example comment #1:

Nancy, PID 103, visited Dr. Grace Yang on 1/16/18 and 2/14/18 at 1600 Research Blvd, Rockville MD 20850. 301-251-1500. Copay $50.

Category: *Edit/Delete provider information*

Features:

Date	$	Zip Code	Phone	Name	Verbs and synonyms	Provider	Drug	Insurer	Question num	Field name
1	1	1	1	1	0	1	0	0	PV	PV

Regular expression (re)

Named entity recognition (spaCy)

Synset (NLTK)

Match against three reference DB (spaCy, Levenshtein, Elastic Search)

Methodology – Feature Engineering (verbs "correct", "change", "edit", "delete" and their synonyms)

Example comment #2:

Correct spelling of following prescriptions: OXYCODONE & PREDNISONE.

Category*: Edit/Delete prescribed medicine(s)*

Features*:*

Date	$	Zip Code	Phone	Name	Verbs and synonyms	Provider	Drug	Insurer	Question num	Field name
0	0	0	0	0	1	0	1	0	PM	PM

Add new information

Add Purchase of Eyeglasses/Contact Lenses

Add Purchase of Other Medical Expenses

Add prescribed medicine(s)

Edit, delete, change existing information

Edit/Delete health care event

Edit/Delete provider information

Edit/Delete prescribed medicine(s)

Methodology – Features Engineering (match against provider, drug, insurer reference databases)

- Lookup
  - Huge database: 1GB+
  - 2 minutes for 1 lookup against text file
    - Elastic Search (ES): a few milliseconds for 1 lookup
- Match
  - Which part in comments to lookup against ES?
    - Noun chunks
  - How to verify the matching between the part in comments and the results from ES?
    - Fuzzy string matching

Methodology – Features Engineering (match against provider, drug, insurer reference databases)

- Comment: *Nancy, PID 103, visited Dr. Grace Yang on 1/16/18 and 2/14/18 at 1600 Research Blvd, Rockville MD 20850. 301-251-1500. Copay $50.*

The part in comments to lookup against databases (noun chunks)		The results from ES
'dr. grace yang' 'pid' 'rockville md' 'grace yang' 'nancy' '1600 research blvd' 'copay'	=	'grace cuihong yang'
Syntactic dependency parser (spaCy)		Fuzzy string matching (Levenshtein)

Methodology – Text Features (tf-idf matrix)

- Comment: *It was not a hospital stay. Nancy visited a private clinic once a week on Mondays for allergies.*
- Category*: edit/delete health event details*

Date	$	Zip Code	Phone	Name	Verbs	Provider	Drug	Insurer	Question num	Field name
0	0	0	0	1	0	0	0	0	HS	HS

+

One Hot Encoder
(scikit-learn)

hospital	stay	visit	private	clinic	allergy
0.44	0.55	0.22	0.22	0.22	0.76

• • •

Methodology – Machine Learning

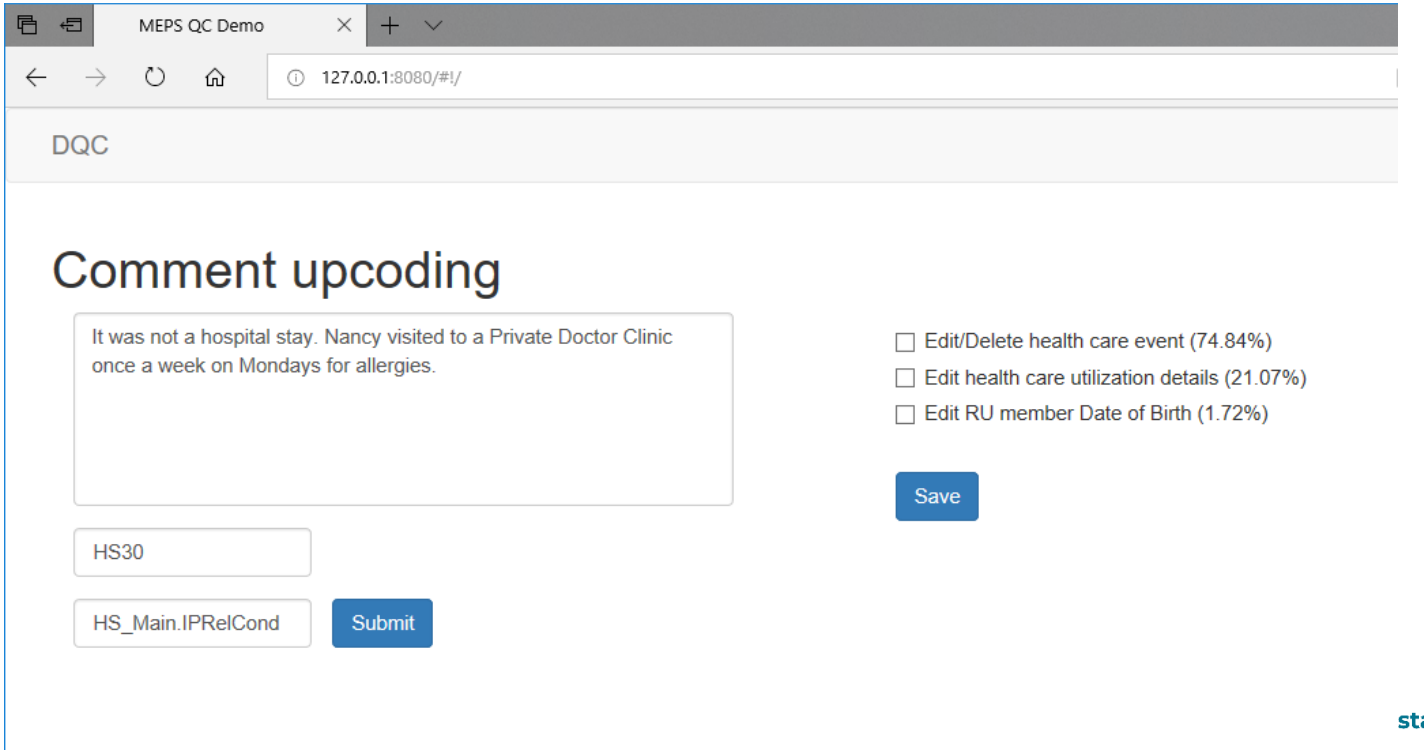
- 80% for training, 20% for testing
- explored from LASSO to XGBoost
- LASSO is selected as best option

Discussion – Model Performance

- 76.14% classification accuracy across 16 categories
- 94.2% accuracy with top 3 predicted categories ranked by probability

Discussion – Conclusion

- We use NLP and ML to **suggest 3 categories** for technicians to further process open text field comments. Our end product allows **real-time responses with 94.2% accuracy.**



Thank You

kristinchen@westat.com