Finding Efficiencies for Open Text Review Using Natural Language Processing on a Panel Study

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

2021 Federal Computer Assisted Survey Information Collection Workshops

Agenda

Background & Business Challenge

Data

Methodology 1 – in production

Methodology 2 – for research

Takeaways

Survey Background

- During in person interviews, respondents may offer additional information too late for the interviewer to back up and incorporate the correction
- E.g., MEPS: Respondents are asked recall details about a broad range of health-related events from an extended time period
  - Respondents often offer additional or updated information at much later parts of the interview
  - Backing up through the system to correct the data increases interview length, increases risk of error, and affects the experience of the interviews
  - Instead, interviewers leave comments with details for post-hoc edits on specific questions or variables

Business Challenge

- Processing interviewer comments is time-consuming, labor-intensive and costly
- Interviewers must select a broad grouping category for each comment they enter
- Human coders use these categories to standardize data editing, but comment categories are sometimes incorrect
  - 80% of them are assigned to the catch-all "other" category
- **Can we use Machine Learning to predict the corrected categories?**

Training data

Category	Interviewer Comment	Proportion
Health Care Events	Nancy, PID 103, visited Dr. Grace Yang on 1/16/18 (not on 1/17/18)" for allergies at 1600 Research Blvd, Rockville MD 20850. 301-251-1500. Copay $50.	48.66
Health Insurance	PID 102 also has AARP as a supplemental insurance.	12.36
Prescribed Medicines	Ibuprofen 800 mg prescribed for Andy on June 17, 2019.	12.04
Other		8.09
RU/ RU Member		6.24
Employment	For Catherine, the employer “Westad” should be spelled “Westat”.	4.98
RU Member Refusal		2.39
Condition		2.31
Glasses/Contact Lenses		1.51
Other Medical Expenses		1.42

Methodology 1: Hand-crafted feature extraction, TF-IDF word embeddings and a ML model (in production)

- Feature extraction
  - Section and question number: comments about a given topic are more likely to appear on or after the section in which that information was collected
  - 11 non-text features: the feature captures attributes that are potentially relevant for identifying comment categories
    - A comment indicating the respondent used a specific pharmacy to fill a prescription may include the pharmacy’s phone number or address

Methodology 1: Hand-crafted feature extraction

- Comment: "Nancy, PID 103, visited Dr. Grace Yang on 1/16/18 (not on 1/17/18) for allergies at 1600 Research Blvd, Rockville MD 20850. 301-251-1500. Copay $50."
- Category: Health Care Events

Date	$	Zip code	Telephone #	A respondent	Age of a person	Any city or state	Any person name	drug	Insurer	Medical provider	Section #	Question #
1	1	1	1	1	0	1	1	0	0	1	PV	PV

Regular Expression (re)

Named Entity Recognition (SpaCy)

Syntactic Dependency Parser & Fuzzy String Matching (SpaCy, FuzzyWuzzy, Elastic Search)

Methodology 1: Hand-crafted feature extraction

- Comment: “Nancy, PID 103, visited Dr. Grace Yang on 1/16/18 (not on 1/17/18) for allergies at 1600 Research Blvd, Rockville MD 20850. 301-251-1500. Copay $50.”
- Category: Health Care Events

Extract noun chunks to lookup against reference database	Search results based on the string distance
'dr. grace yang' 'pid' 'rockvillemd' 'grace yang' 'nancy' '1600 research blvd' 'copay'	'grace cuihong yang'

Syntactic Dependency Parser

Fuzzy String Matching

Methodology 1: TF-IDF word embeddings

- Comment: “Nancy, PID 103, visited Dr. Grace Yang on 1/16/18 (not on 1/17/18) for allergies at 1600 Research Blvd, Rockville MD 20850. 301-251-1500. Copay $50.”
- Category: Health Care Events

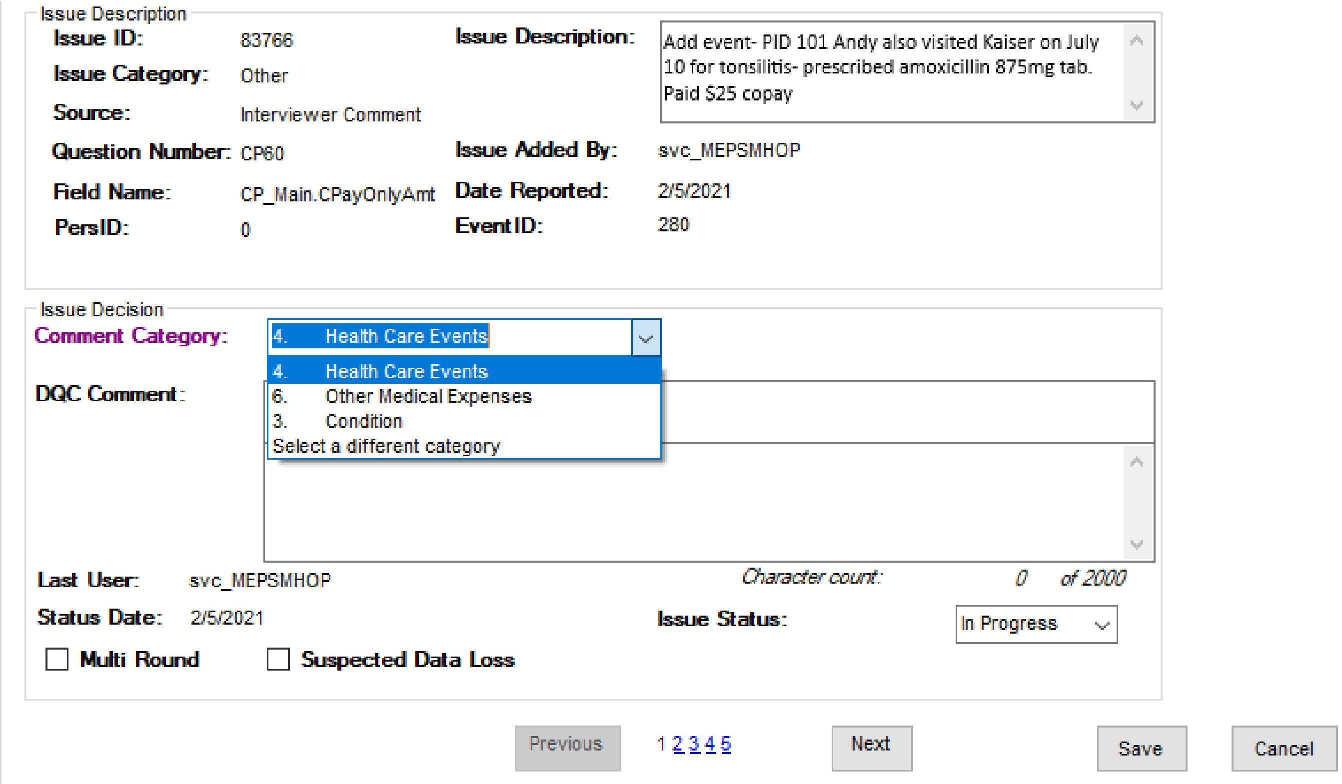
visit	allergy	research	blvd	rockville	copay
0.23	0.44	0.90	0.90	0.90	0.56

Methodology 1: a ML model

- 80% for training, 20% for testing
- 10-fold cross validation
- Explored from ElasticNet to XGBoost
- ElasticNet is selected as best option
  - Top 1 accuracy: 88.36%
  - Top 3 accuracy: 97.1%

Category	Precision	Recall	Testing Size
Health Care Events	0.897	0.963	507
Health Insurance	0.905	0.884	129
Prescribed Medicines	0.881	0.887	148
Other	0.882	0.788	85
RU/ RU Member	0.717	0.729	59
Employment	0.902	0.899	62
RU Member Refusal	0.941	0.842	19
Condition	0.929	0.433	30
Glasses/Contact Lenses	0.846	0.478	23
Other Medical Expenses	0.750	0.750	12

Methodology 1: Hand-crafted feature extraction, TF-IDF word embeddings and a ML model (in production)



Methodology 2: Deep Learning Neural Network (for research)

- **Can a deep learning model outperform the linear model in production?**
  - Deep learning enables multi-level automatic feature representation learning. In contrast, traditional machine learning liaises heavily on hand-crafted features. (Young 2018)

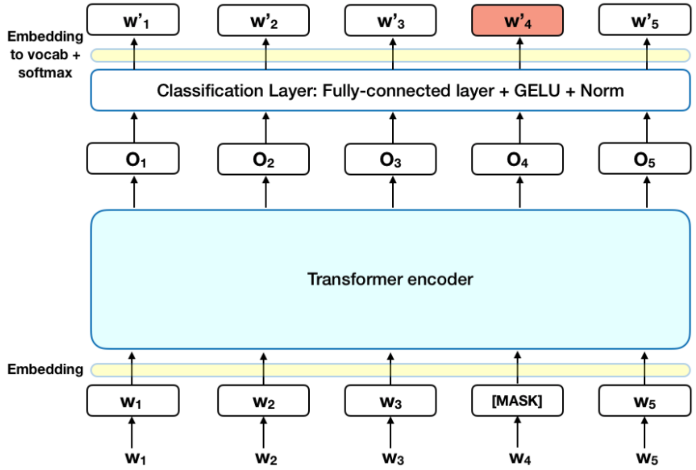
Methodology 2: Feature extraction, and a shallow Neural Network (for research)

- LR = 0.001; BATCH_SIZE = 32; EPOCHS = 30; weighted classes
- Accuracy: 69%

```
Model: "sequential_7"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_14 (Dense)             multiple                  11904
_________________________________________________________________
dense_15 (Dense)             multiple                  1290
=================================================================
Total params: 13,194
Trainable params: 13,194
Non-trainable params: 0
_________________________________________________________________
```

Methodology 2: Comments and DistilBERT (for research)

- BERT: Bidirectional Encoder Representations from Transformers (Devlin 2019)
  - DistilBERT: reduces the size of a BERT model by 40%, while retaining 97% of its language understanding capabilities and being 60% faster. (Sanh 2019)



Methodology 2: Comments and DistilBERT (for research)

- DBERT_MODEL = 'distilbert-base-uncased'; LR = 1e-5; EPOCHS = 10; BATCH_SIZE = 32; weighted classes
- Accuracy: 66%

```
Model: "tf_distil_bert_for_sequence_classification"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
distilbert (TFDistilBertMain multiple                  66362880
_________________________________________________________________
pre_classifier (Dense)       multiple                  590592
_________________________________________________________________
classifier (Dense)           multiple                  7690
_________________________________________________________________
dropout_19 (Dropout)         multiple                  0
=================================================================
Total params: 66,961,162
Trainable params: 66,961,162
Non-trainable params: 0
_________________________________________________________________
```

Summary

- Model comparison – Accuracy
  - An ElasticNet model trained on extracted features and TF-IDF word embeddings: 88%
  - A shallow NN trained on extracted features: 69%
  - DistilBERT for sentence classification trained on comments: 66%
- ElasticNet/shallow NN based on hand-crafted features > DistilBERT
  - The key to distinguishing classes among comments mostly based on the presence or absence of key information, which are captured by the hand-crafted features.
  - This knowledge exceeds the contextual representation captured by BERT.

Takeaways

- Superior deep learning models, i.e. BERT, doesn’t apply to this problem.

References

- Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT.
- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv, abs/1910.01108.
- Young, Tom & Hazarika, Devamanyu & Poria, Soujanya & Cambria, Erik. (2018). Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. IEEE Computational Intelligence Magazine. 13. 55-75. 10.1109/MCI.2018.2840738.

Thank You

kristinchen@westat.com