Using Deep Learning to Classify Open-text Field Comments in a Panel Study

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

Government Advances in Statistical Programming Workshop, 11/06/2020

Background

- Field comments from Medical Expenditure Panel Study (MEPS): entered by interviewers into CAPI to request data updates or corrections to the data they collect earlier in the interview.
    - *“PID 102 also has AARP as a supplemental insurance.”*
    - *“Ibuprofen 800 mg prescribed for Andy on June 17, 2019.”*
    - *“For Catherine, the employer “Westad” should be spelled “Westat”.*
    - *“Nancy visited Dr. Grace Yang on 2/14/18 (not on 1/17/18)".*
- Field comments require costly (and mostly manually) data processing, but they can improve data quality.
- Category: designed to improve processing efficiency. Before entering a comment, interviewers are required to select a list of broad categories such as **Health Insurance**, **Employment**, or **Prescribed Medicines**

Problem

- Can Machine Learning help to determine the correct comment category?

Data

- ~8,000 data points
- 10 categories
- The proportion is unbalanced

```
Health Care Events          0.42
Other                       0.13
Prescribed Medicines        0.12
Health Insurance            0.11
RU / RU Member              0.10
Employment                  0.06
Condition                   0.02
RU Member Refusal           0.02
Other Medical Expenses      0.02
Glasses / Contact Lenses    0.01
Name: category, dtype: float64
```

Methodology - method 1: feature extraction, TF-IDF word embeddings and a linear model (in production)

feature extraction

- survey metadata: *question number* and *field name* where the comments entered
- **key information** presented in the sentences that could help predict the category: *dollar amounts*; *zipcode*; *phone number*; *states or cities*; *date*; **age of a person*; *keywords indicated respondent(s): "R(s)", "respondent(s)", "household(s)", "RU(s)", "PID(s)", "HH", "resp", "pt"*; *name of a person*; *insurer*; *drug*; *medical provider*
- Examples:
  - *"PID 102 also has **AARP** as a supplemental insurance"*. - features: **insurer** ; category: **Health Insurance**
  - *"**Ibuprofen 800 mg** prescribed for Andy on June 17, 2019."*. - features: **drug**, category: **Prescribed Medicines**
  - *"Nancy visited **Dr. Grace Yang** on **2/14/18** (not on **1/17/18**)*". - features: **medical provider**, **date**; category: **Health Care Events**

Methodology - method 1: feature extraction, TF-IDF word embeddings and a linear model (in production)

- model: ElasticNet
- performance: 76% accuracy

Can a deep model outperform the linear model in production?

- Deep learning-based Natural Language Processing models provide *pre-training contextual representations*, resulting in substantial accuracy improvements compared to training on a small set of datasets from scratch (Devlin, 2019).

Methodology - method 2: feature extraction, and a Feed Forward Neural Network (FFNN)

```
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
_________________________________________________________________
```

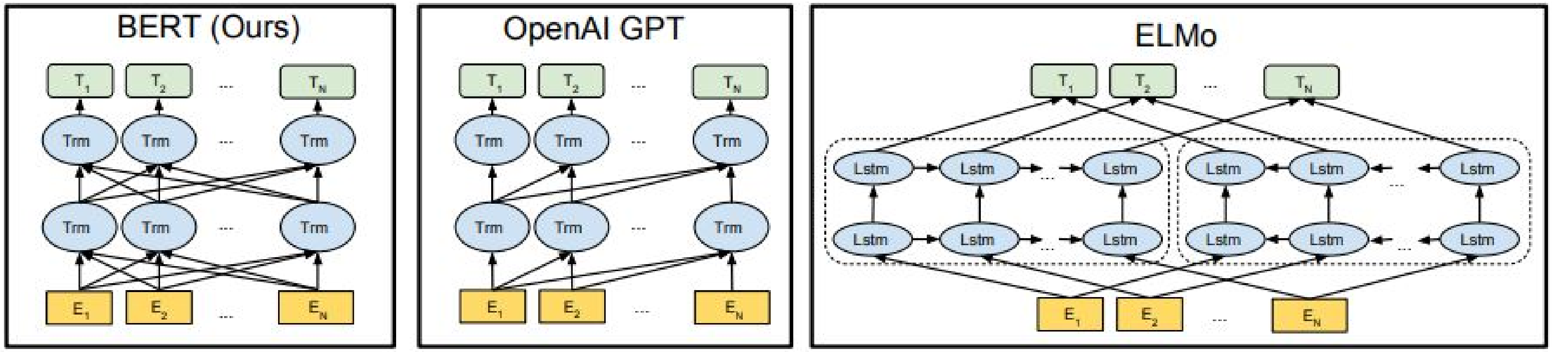
Methodology - method 2: feature extraction, and a FFNN

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

```
('                          precision    recall  f1-score   support\n'
 '\n'
 '          RU / RU Member       0.59      0.22      0.32       190\n'
 '       RU Member Refusal       0.17      0.66      0.27        32\n'
 '               Condition       0.29      0.57      0.38        40\n'
 '      Health Care Events       0.79      0.82      0.80       764\n'
 'Glasses / Contact Lenses       0.22      0.35      0.27        17\n'
 '  Other Medical Expenses       0.27      0.35      0.31        34\n'
 '    Prescribed Medicines       0.83      0.84      0.83       214\n'
 '              Employment       0.85      0.88      0.87       136\n'
 '        Health Insurance       0.80      0.82      0.81       213\n'
 '                   Other       0.54      0.37      0.44       241\n'
 '\n'
 '                accuracy                           0.69      1881\n'
 '               macro avg       0.53      0.59      0.53      1881\n'
 '            weighted avg       0.71      0.69      0.68      1881\n')
```

Methodology - method 3: comments and Bidirectional Transformers for Language Understanding (BERT)

- BERT (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 - method 3: comments and DistilBERT

```
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
_________________________________________________________________
```

Methodology - method 3: comments and DistilBERT

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

```
('                          precision    recall  f1-score   support\n'
 '\n'
 '            RU / RU Member       0.39      0.56      0.46       190\n'
 '         RU Member Refusal       0.55      0.84      0.67        32\n'
 '                 Condition       0.49      0.68      0.57        40\n'
 '        Health Care Events       0.79      0.68      0.73       764\n'
 'Glasses / Contact Lenses       0.46      0.71      0.56        17\n'
 '    Other Medical Expenses       0.33      0.50      0.40        34\n'
 '      Prescribed Medicines       0.83      0.83      0.83       214\n'
 '                Employment       0.74      0.77      0.76       136\n'
 '          Health Insurance       0.77      0.81      0.79       213\n'
 '                     Other       0.46      0.36      0.40       241\n'
 '\n'
 '                  accuracy                           0.66      1881\n'
 '                 macro avg       0.58      0.67      0.62      1881\n'
 '              weighted avg       0.68      0.66      0.67      1881\n')
```

Methodology - Model comparison

- Accuracy:
  - an ElasticNet model trained on extracted features and TF-IDF word embeddings: 76%
  - a FFNN trained on extracted features: 69%
  - DistilBERT for sentence classification trained on comments: 66%

- ElasticNet/FFNN > DistilBERT
  - The key to distinguishing classes among field comments mostly bases on the presence or absence of key information.
    - The knowledge derived better prediction exceeds the context of language itself.
  - Contextual representation doesn't add values in our classification.

- ElasticNet > FFNN
  - The linearity assumption of the extracted features applies better to the structure of our data.

Takeaways

- Superior deep learning models, i.e. BERT, don't always outperform in all kinds of Natural Language Understanding tasks.
- The value of rigorous research, data understanding, and feature engineering shouldn't be underestimated in Natural Language Understanding tasks. In some cases, a simpler model could solve the problem better.

Reference

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.

Thank You

KristinChen@westat.com