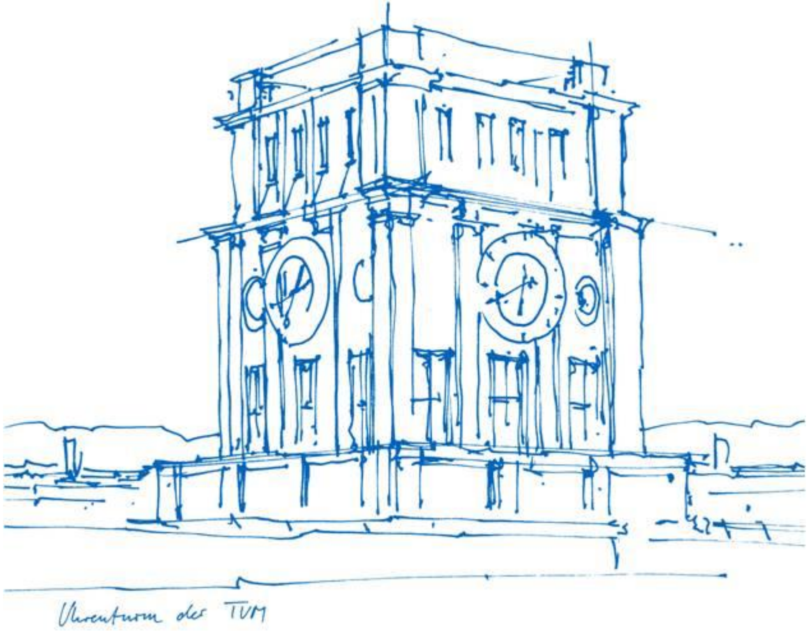
Multi-Lingual Theme Prediction of Customer Reviews Using Deep Pre-Trained Embeddings

Team 06

Michael Sorg

17.07.2019

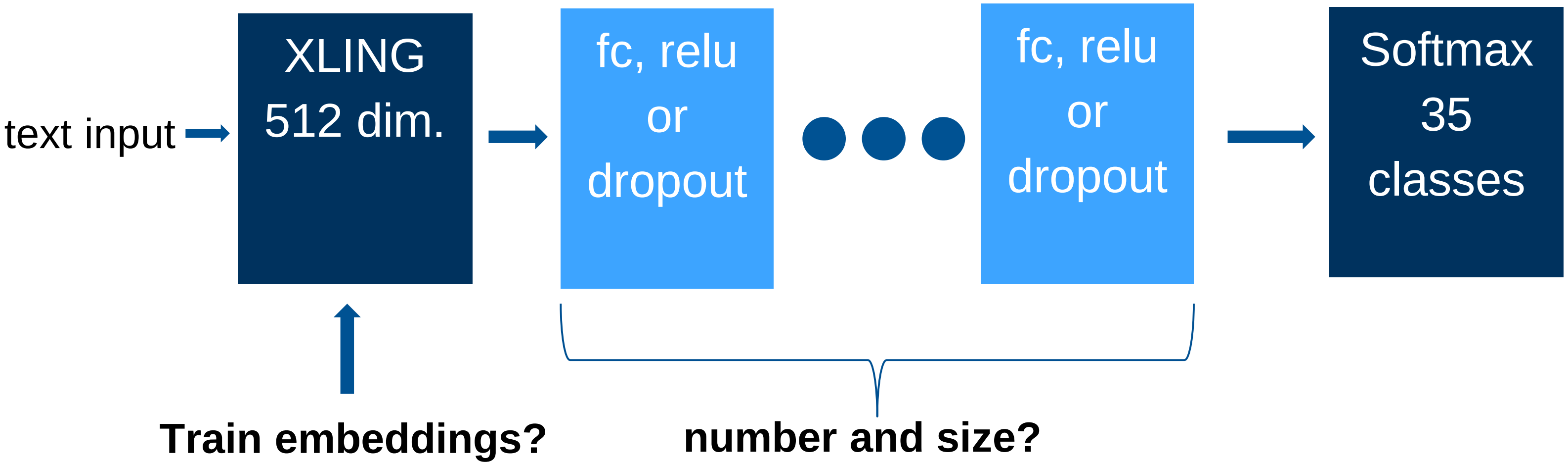


Outline

- Task description
- Amazon Reviews
  - Unbalanced training
  - Balanced training
  - Architecture search
- Organic Dataset
  - w/o fine-tuning

Task

- category prediction on amazon reviews based on XLING embeddings
- Evaluate on german reviews without training on german data
- Fine-tune on organic dataset for relevance, attribute and entity classification



text input

XLING 512 dim.

fc, relu or dropout

fc, relu or dropout

Softmax 35 classes

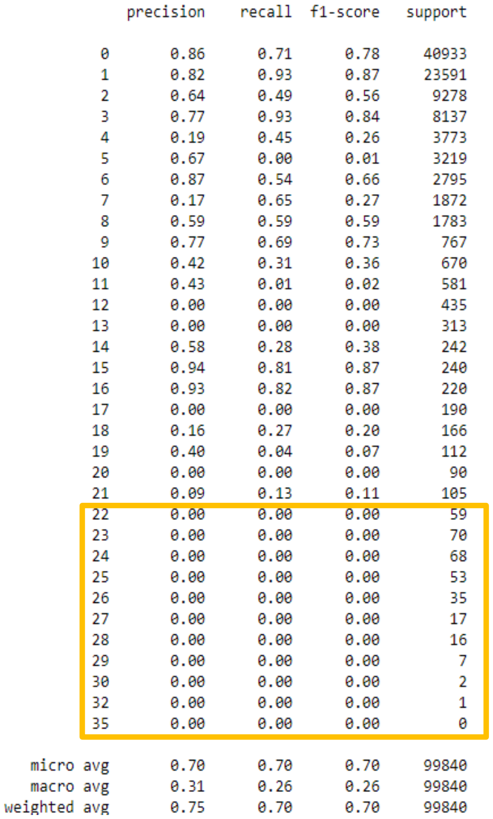
Train embeddings?

number and size?

Baseline experiment

- no hidden layer (only xling + softmax)
- Unbalanced training

- Some classes are never predicted
- Micro-f1 score on validation set: ~ 70%

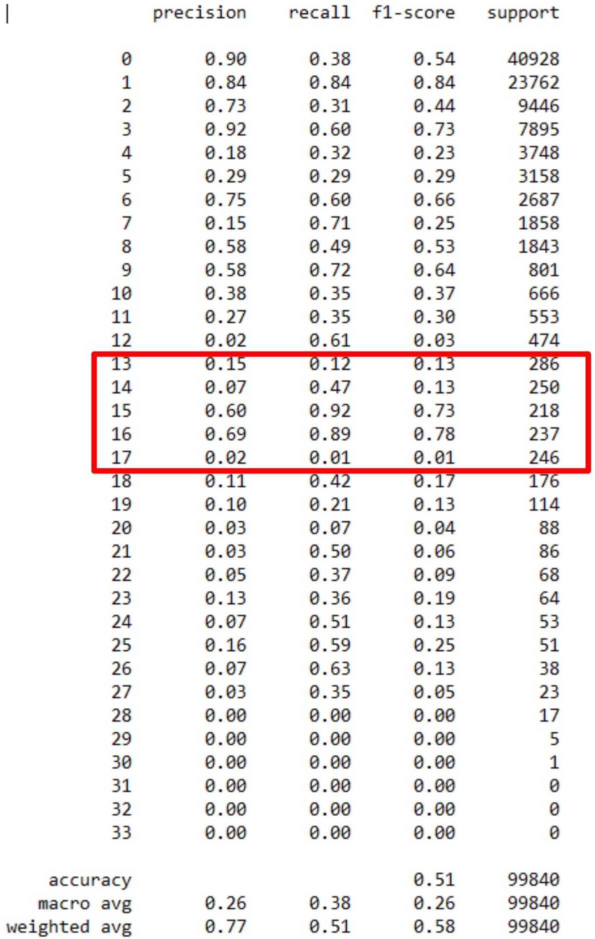


```
              precision    recall  f1-score   support

           0       0.86      0.71      0.78     40933
           1       0.82      0.93      0.87     23591
           2       0.64      0.49      0.56      9278
           3       0.77      0.93      0.84      8137
           4       0.19      0.45      0.26      3773
           5       0.67      0.00      0.01      3219
           6       0.87      0.54      0.66      2795
           7       0.17      0.65      0.27      1872
           8       0.59      0.59      0.59      1783
           9       0.77      0.69      0.73       767
          10       0.42      0.31      0.36       670
          11       0.43      0.01      0.02       581
          12       0.00      0.00      0.00       435
          13       0.00      0.00      0.00       313
          14       0.58      0.28      0.38       242
          15       0.94      0.81      0.87       240
          16       0.93      0.82      0.87       220
          17       0.00      0.00      0.00       190
          18       0.16      0.27      0.20       166
          19       0.40      0.04      0.07       112
          20       0.00      0.00      0.00        90
          21       0.09      0.13      0.11       105
          22       0.00      0.00      0.00        59
          23       0.00      0.00      0.00        70
          24       0.00      0.00      0.00        68
          25       0.00      0.00      0.00        53
          26       0.00      0.00      0.00        35
          27       0.00      0.00      0.00        17
          28       0.00      0.00      0.00        16
          29       0.00      0.00      0.00         7
          30       0.00      0.00      0.00         2
          32       0.00      0.00      0.00         1
          35       0.00      0.00      0.00         0

   micro avg       0.70      0.70      0.70     99840
   macro avg       0.31      0.26      0.26     99840
weighted avg       0.75      0.70      0.70     99840
```

Baseline experiment

- no hidden layer (only xling + softmax)
- **balanced training**

- Predictions for some classes are very bad
- Micro-f1 score on validation set: ~ 51%



```
|                 precision    recall  f1-score   support

               0       0.90      0.38      0.54     40928
               1       0.84      0.84      0.84     23762
               2       0.73      0.31      0.44      9446
               3       0.92      0.60      0.73      7895
               4       0.18      0.32      0.23      3748
               5       0.29      0.29      0.29      3158
               6       0.75      0.60      0.66      2687
               7       0.15      0.71      0.25      1858
               8       0.58      0.49      0.53      1843
               9       0.58      0.72      0.64       801
              10       0.38      0.35      0.37       666
              11       0.27      0.35      0.30       553
              12       0.02      0.61      0.03       474
              13       0.15      0.12      0.13       286
              14       0.07      0.47      0.13       250
              15       0.60      0.92      0.73       218
              16       0.69      0.89      0.78       237
              17       0.02      0.01      0.01       246
              18       0.11      0.42      0.17       176
              19       0.10      0.21      0.13       114
              20       0.03      0.07      0.04        88
              21       0.03      0.50      0.06        86
              22       0.05      0.37      0.09        68
              23       0.13      0.36      0.19        64
              24       0.07      0.51      0.13        53
              25       0.16      0.59      0.25        51
              26       0.07      0.63      0.13        38
              27       0.03      0.35      0.05        23
              28       0.00      0.00      0.00        17
              29       0.00      0.00      0.00         5
              30       0.00      0.00      0.00         1
              31       0.00      0.00      0.00         0
              32       0.00      0.00      0.00         0
              33       0.00      0.00      0.00         0

        accuracy                           0.51     99840
       macro avg       0.26      0.38      0.26     99840
    weighted avg       0.77      0.51      0.58     99840
```

Architecture search

- Training xling embeddings leads to overfitting

- Problems when going deeper:
  - f1 score doesn‘t improve much after 2-3 epochs
  - Overfitting when using deeper networks (more than 2-3 layers)

	micro-f1 on test set
Tf-idf + svm	44 %
Baseline model	51 %
Best model (xling,150,relu,dropout,softmax)	57 %

Results

- Increased data set size doesn‘t lead to better performance

- Including review headline has no effect
- Filtering out English reviews from the German test set has no effect

Organic dataset

- Fine-tuning increases performance

	f1 score	F1 score with fine-tuning
relevance	74 %	77 %
entity	51 %	57 %
attribute	44 %	50 %