NLP & binary text classification

Capstone Project # 2

Dmitri Kochubei
15.08.2020

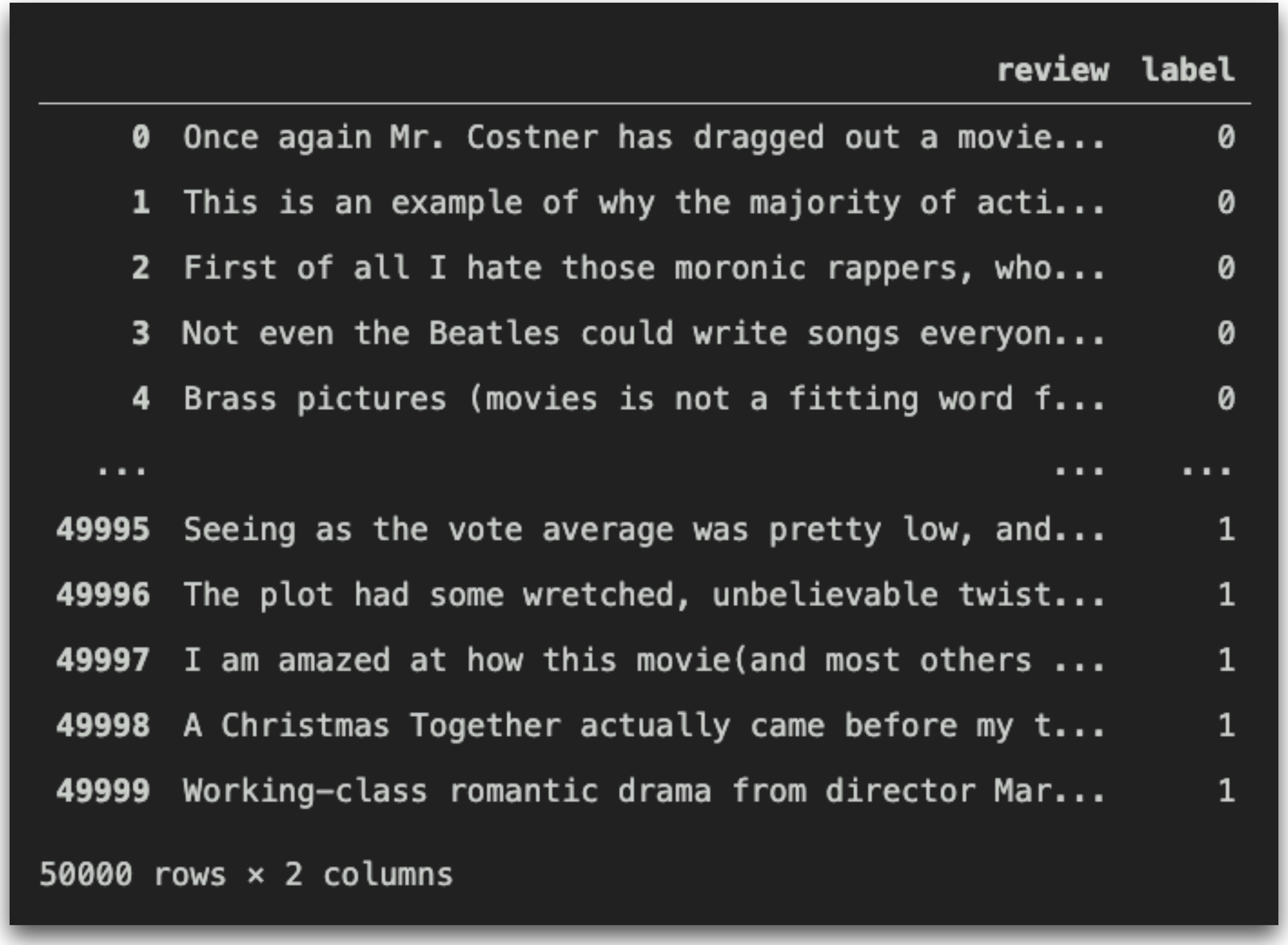
Motivation

Lex Fridman: So, what advice do you have for someone who wants to get started in deep learning?

Jeremy Howard: Train lots of models...

double angle
2sin²(θ) = 1 − cos(2θ)
(Check)
t if we move in opposite ...tion?
Sign ch
t happens if we chang ...zation? Same result.

Problem overview



```
                                                  review  label
    0  Once again Mr. Costner has dragged out a movie...      0
    1  This is an example of why the majority of acti...      0
    2  First of all I hate those moronic rappers, who...      0
    3  Not even the Beatles could write songs everyon...      0
    4  Brass pictures (movies is not a fitting word f...      0
  ...                                                ...    ...
49995  Seeing as the vote average was pretty low, and...      1
49996  The plot had some wretched, unbelievable twist...      1
49997  I am amazed at how this movie(and most others ...      1
49998  A Christmas Together actually came before my t...      1
49999  Working-class romantic drama from director Mar...      1

50000 rows × 2 columns
```

Dataset

Size - 135.5 mb

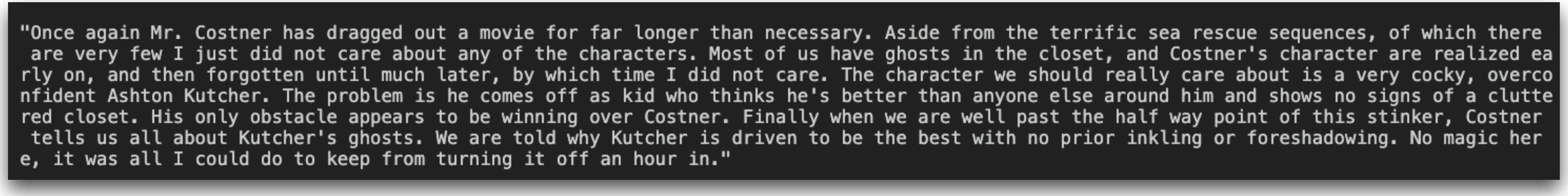
# of observations - 50000

Review - textual data

Class balance - 50/50

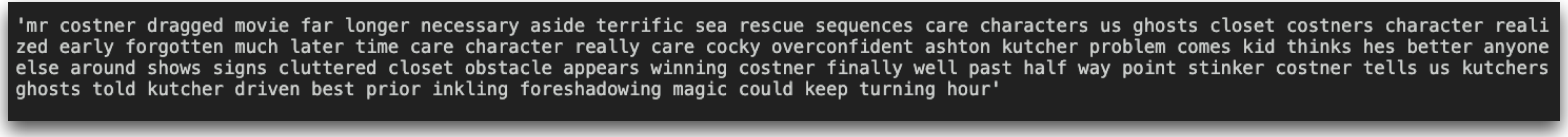
Goal - build a binary classification model to correctly classify reviews' sentiment

Data pre-processing



```
"Once again Mr. Costner has dragged out a movie for far longer than necessary. Aside from the terrific sea rescue sequences, of which there
 are very few I just did not care about any of the characters. Most of us have ghosts in the closet, and Costner's character are realized ea
rly on, and then forgotten until much later, by which time I did not care. The character we should really care about is a very cocky, overco
nfident Ashton Kutcher. The problem is he comes off as kid who thinks he's better than anyone else around him and shows no signs of a clutte
red closet. His only obstacle appears to be winning over Costner. Finally when we are well past the half way point of this stinker, Costner
 tells us all about Kutcher's ghosts. We are told why Kutcher is driven to be the best with no prior inkling or foreshadowing. No magic her
e, it was all I could do to keep from turning it off an hour in."
```

1. **“Lowercase” every word**
2. **Remove punctuations**
3. **Remove stop words**



```
'mr costner dragged movie far longer necessary aside terrific sea rescue sequences care characters us ghosts closet costners character reali
zed early forgotten much later time care character really care cocky overconfident ashton kutcher problem comes kid thinks hes better anyone
else around shows signs cluttered closet obstacle appears winning costner finally well past half way point stinker costner tells us kutchers
ghosts told kutcher driven best prior inkling foreshadowing magic could keep turning hour'
```

Data pre-processing

ML

tf-idf

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

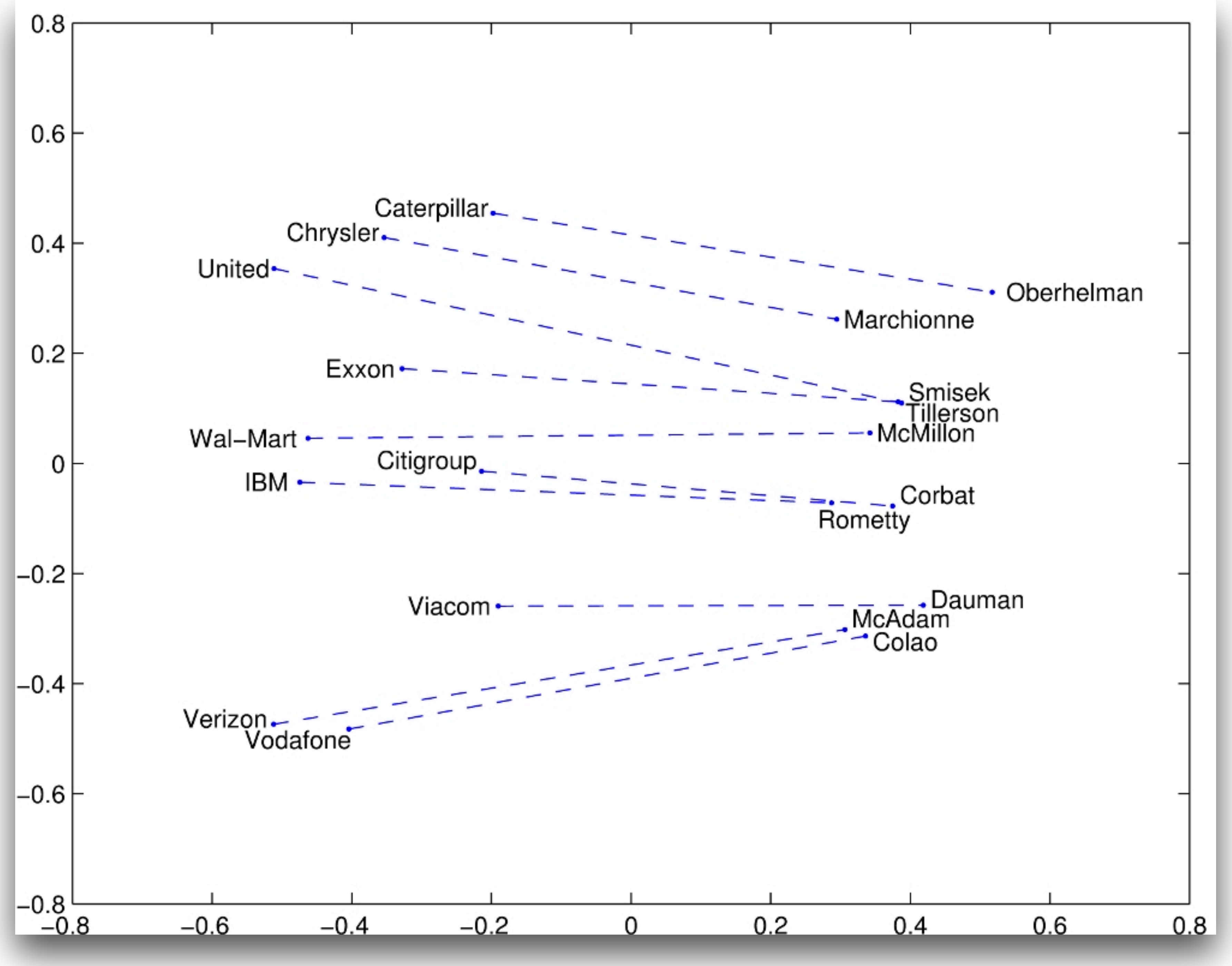
$tf_{ij} =$ number of occurrences of i in j

$df_i =$ number of documents containing i

$N =$ total number of documents

DL

Pre-trained word embedding (GloVe)



GloVe pre-processing

1. Word tokenization
2. Padding
3. Parsing 'glove.42B.300d.txt'
4. Creating embedding matrix

We need a look-up table where every word has a number that corresponds to a GloVe embedding.

DL

1. LSTM
2. Convolutional NN
3. Dense NN

In every network the first layer is the embedding layer.

ML

1. Logistic regression

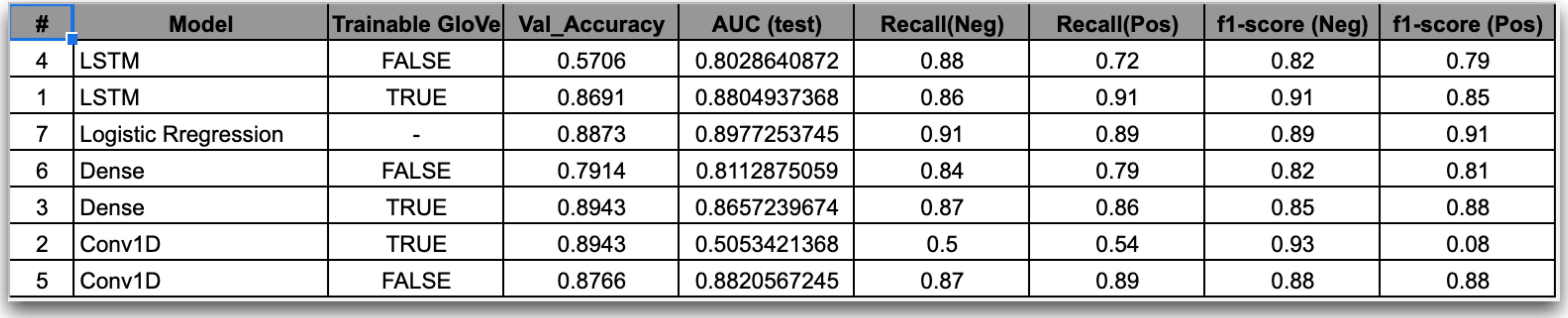
Simple tf-idf matrix is used as X for the model.

Should GloVe weights be updated while training?

- It depends, some models benefited from updatable weights while other didn’t.

- Slower convergence is observed in cases when GloVe vectors are not updatable.

Results



#	Model	Trainable GloVe	Val_Accuracy	AUC (test)	Recall(Neg)	Recall(Pos)	f1-score (Neg)	f1-score (Pos)
4	LSTM	FALSE	0.5706	0.8028640872	0.88	0.72	0.82	0.79
1	LSTM	TRUE	0.8691	0.8804937368	0.86	0.91	0.91	0.85
7	Logistic Rregression	-	0.8873	0.8977253745	0.91	0.89	0.89	0.91
6	Dense	FALSE	0.7914	0.8112875059	0.84	0.79	0.82	0.81
3	Dense	TRUE	0.8943	0.8657239674	0.87	0.86	0.85	0.88
2	Conv1D	TRUE	0.8943	0.5053421368	0.5	0.54	0.93	0.08
5	Conv1D	FALSE	0.8766	0.8820567245	0.87	0.89	0.88	0.88

FRESH GRAD
SENIOR DEV
NOOOOOOO!!!!! YOU SHOULD USE DEEP NEURAL NETWORKS IN THIS PROBLEM!! ONLY 200+ HOURS OF TRAINING ON BIG CLUSTER OF TPU'S CAN GRANT YOU DECENT PERFORMANCE!!!! OMG YOUR MODEL GIVES 3% LESS ACCURACY NOOOO!!!!!!
haha TF-IDF + LOGISTIC REGRESSION GO BRRRRRRRRRR
scikit learn

Conclusion

1. The desired hands-on NLP experience has been obtained both with Keras and PyTorch.
2. “Simple” logistic regression has outperformed every other model in this particular task.
3. There is a ready-to-use project code that can be applied in other NLP projects.