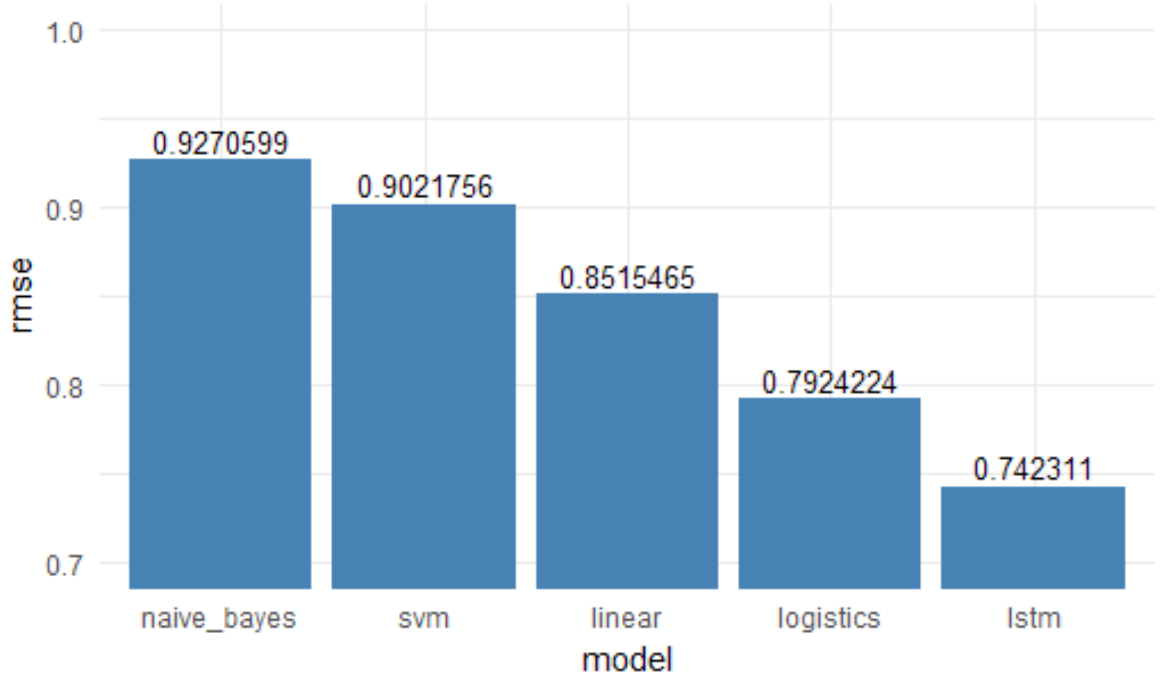
# Final analysis on yelp data

GROUP MEMBERS: CHENG LU, LINHAI ZHANG, LAN WANG

# 1. High Accuracy Model
# Deep Learning Model

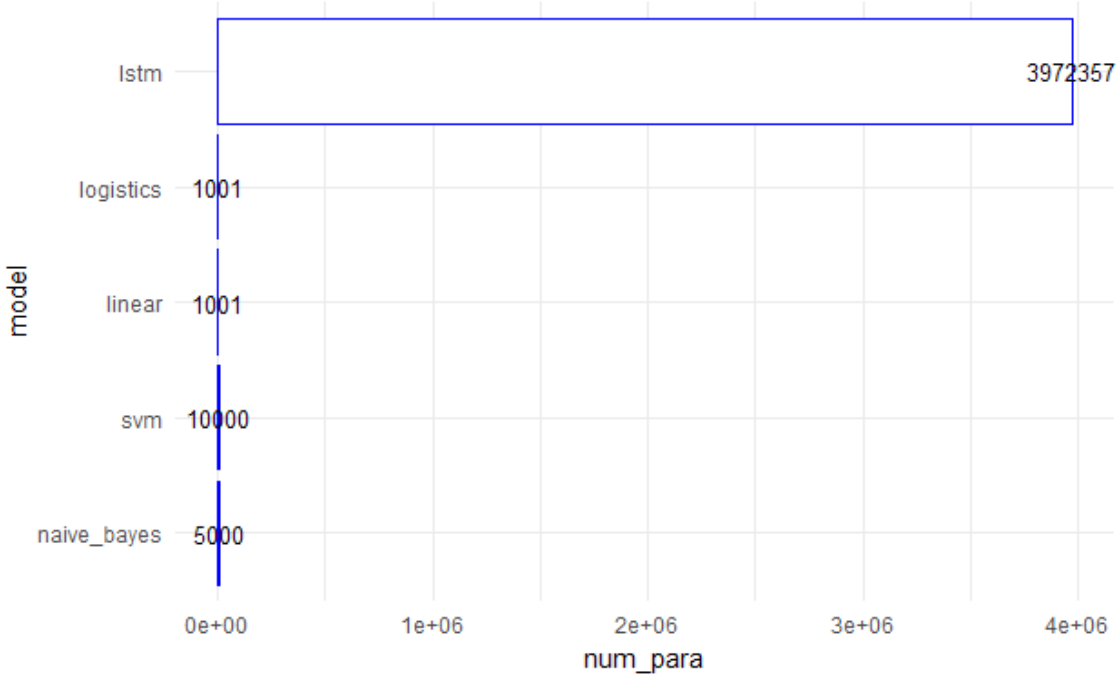
# 1.1 Why deep learning?

Deep learning method is generally better than others.



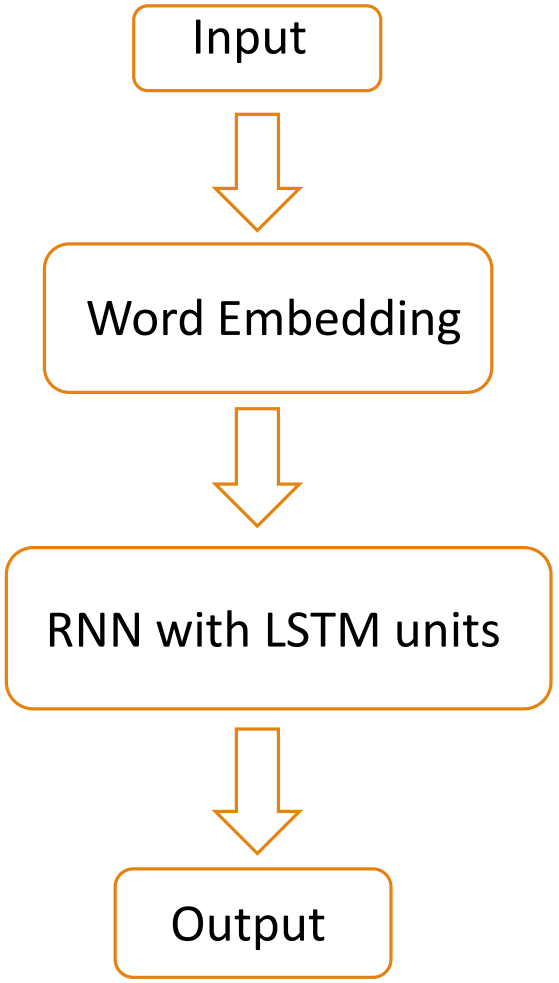

Model vs RMSE

One of the reasons may be here:




Model vs # parameters

# 1.2 Our framework




Word Embedding:

- An algorithm to project one word to n dimension real number vector
- Much more information compared with tradition way(i.e. 1 dimension)
- Commonly used in nature language processing

Long Short Time Memory units (LSTM)

- A building unit for layers of a recurrent neural network
- Allows the neural network uses the previous information

# 1.3 Parameter tuning

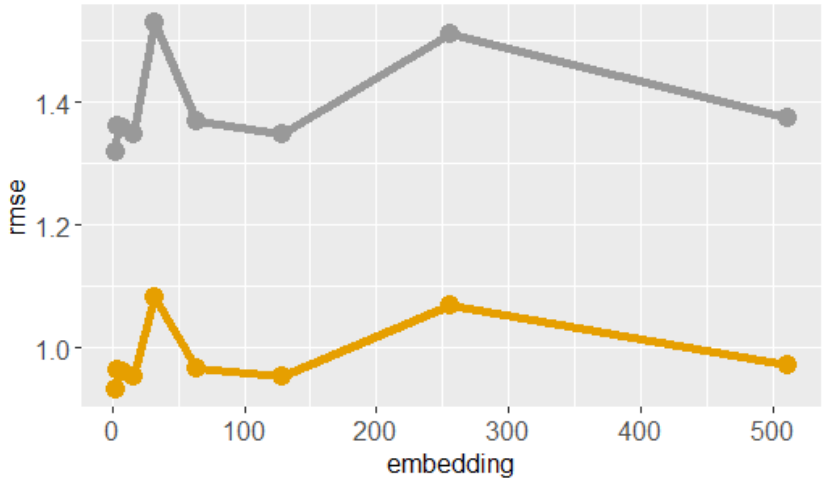
Important parameters:

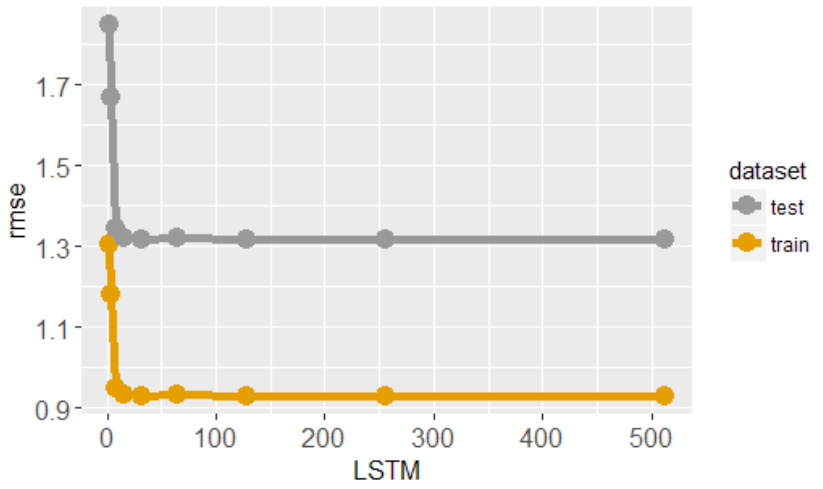
- Drop out rate
- Number of epoch
- Number of units in embedding
- Number of units in LSTM network

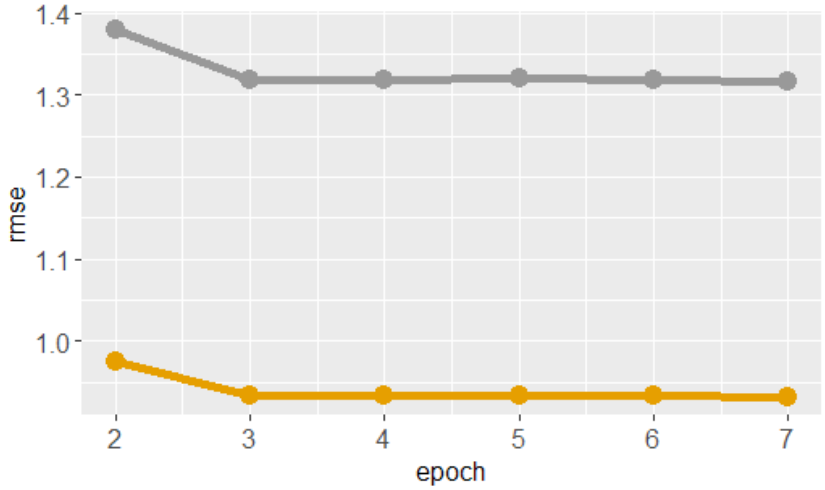
Parameter tuning set up:

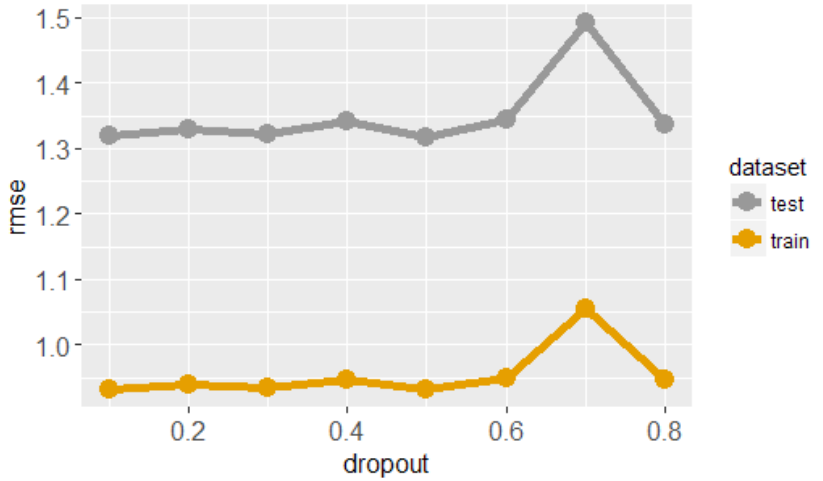
- Train on 10000 data
- Test on 5000 data
- Record the RMSE of train data and test data

# 1.3 Parameter tuning




Drop out rate: 0.2
# epoch: 4
# units in embedding: 128
# units in LSTM network: 128

# 1.4 Final Model

Final model set up:

- Train on: 1546379 (all the data)
- Drop out rate: 0.2
- Number of epoch: 4
- Number of units in embedding: 128
- Number of units in LSTM network: 128

**Final score on test data: 0.59889**

# 2. Positive/Negative ?

# Choose positive/negative words

**Use PCA based on this matrix**

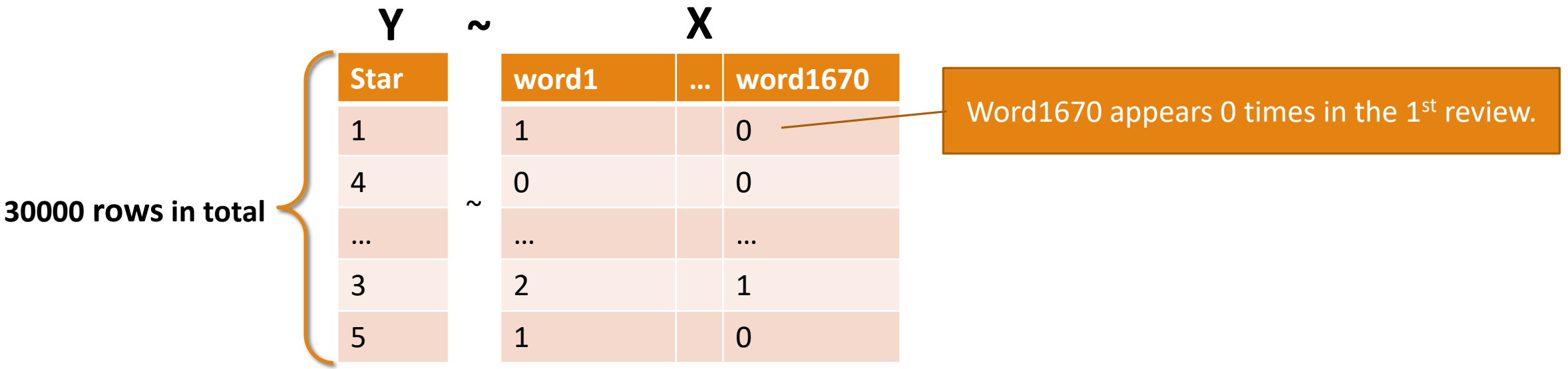
| Word | 1 star | 2 star | 3 star | 4 star | 5 star |
|---|---|---|---|---|---|
| impress | 3.41 | 4.57 | 3.48 | 0.82 | 0.73 |
| …. | … | … | … | … | … |
| miss | 4.71 | 5.44 | 5.25 | 4.93 | 4.76 |

| Top 10 Negative words | Top 10 Positive words |
|---|---|
| worst | perfect |
| horrible | delicious |
| awful | gem |
| rude | fantastic |
| disgust | not disappoint |
| never come | yum |
| never return | affordable |
| poor | incredible |

- **Taste**
  - delicious
  - yum
- **Price**
  - affordable
  - inexpensive
- **Specialty**
  - unique
  - homemade
- **Dessert**
  - cookie
  - pudding

- **Service**
  - rude
  - ignore
- **Taste**
  - tasteless
  - flavorless
- **Price**
  - not worth
  - overprice
- **Quanlity**
  - Not fresh
  - Hair

# Linear Model based on 1670 words' count




RMSE: 0.84;  Adjusted R-square: 0.55

# Laymen’s interpretation

Great place. Came in for lunch, Brodey, Ashley, and one other female server (We didn't catch her name) helped us out. Service was great! Very friendly, and inviting, and they were more than accommodating for us and our toddler. Food was delicious, love the sauce! I got the trifecta level 2, so good, nice and spicy! Very messy, but worth it.

We will be definitely back... With gloves and a raincoat. :)

True rating: 5

Prediction: 4.9

# Strength and weakness

Strength:

1. Deep Learning model: precise
2. Linear model: interpretable
3. PCA: effective

# Strength and weakness

Weakness:

1. Ignore bigrams
2. Ignore other languages
3. Small sample size

# Conclusion

1. Do the variable selection and text vectorizing.

2. Use PCA to create positive and negative wordlists, then fit a linear regression based on it, which is relatively precise (RMSE = 0.84).

3. Deep Learning Model gives precise prediction.

# Thank you!