

# Yelp Data Prediction

---

Yifan Li, Chenlai Shi, Jianmin Chen

Monday Group 1

# 1 Introduction and Data Cleaning

## Introduction

- Small set of informative features
- Accurate predictive model
- Based on about 1.5 million Yelp reviews

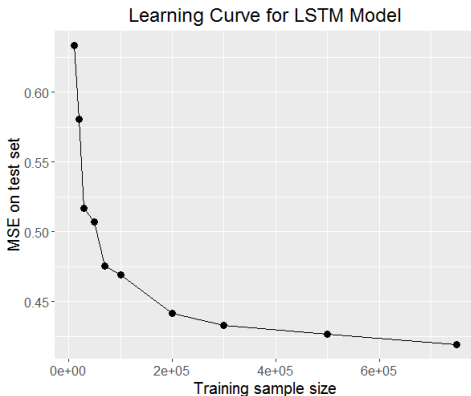
## Data Cleaning

- Modify Abbreviation and Special Symbol
- Remove Non-English
- Negative Sentences
- Remove Punctuation

## 2 Model: LSTM

**Model:** Neural Network with 3 layers

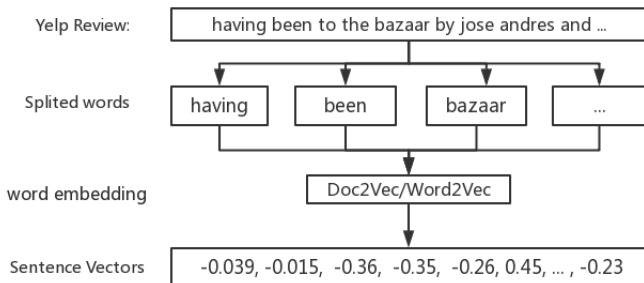
- layer1: LSTM layer with 50 output units
- layer2: Dense layer with 3 output units
- layer3: Dense layer with 1 output unit



## 2.1 Pretrained Sentence Vectors

### Model Features

- Pre-trained Sentence Vectors: Capture word counts and order
- Additional Variables: Capture sentiment, review date and location



## 2.2 Additional Variables

- **year**: scaled year variable.
- **loc1**: 1 if the restaurant is in the western United States, otherwise 0.
- **loc2**: 1 if the restaurant is in the eastern United States, otherwise 0.



## 2.2 Additional Variables

- **Score1~5**:  $\text{Score1}[\text{word}] = \frac{P(\text{this word is included in reviews with 1-star})}{P(\text{this word is included in reviews with other stars})}$
- **S1 ~ S5**:  $S1[\text{review}] = \# \text{ of words with high Score1 in the review.}$

Word	Variable	1-star	2-star	3-star	4-star	5-star
<b>refund</b>	frequency	115	15	7	4	2
	probability	0.011	0.002	0	0	0
	Score	34.200	1.080	0.300	0.072	0.025
<b>notdisappoints</b>	frequency	0	2	5	43	110
	probability	0	0	0	0.002	0.003
	Score	0	0.116	0.188	0.917	3.870
<b>and</b>	frequency	9196	8691	12851	25604	32071
	probability	0.859	0.886	0.877	0.895	0.886
	Score	0.968	1.000	0.991	1.020	1.000



### 3 Compare MSE with other method

## MSE

Feature\ Model	LM	NB	NN	LSTM	GLM	SVM
vector + ad	0.673	0.974	0.494	<b>0.493</b>	0.698	NA
vector	0.720	1.112	0.524	0.526	0.756	0.585
additional	<b>0.836</b>	1.459	0.614	0.612	0.894	NA
frequence	NA	1.126	1.210	NA	0.864	0.790
tf-idf	0.889	1.114	0.804	NA	0.836	0.770

tested on 100000 data



## 4 Interpretable Model

$$\begin{aligned}\hat{y} = & 3.65 + 0.04 * scale(year) + 0.04 * loc1 + 0.06 * loc2 \\ & - 0.11 * S1 - 0.17 * S2 - 0.03 * S3 + 0.03 * S4 + 0.14 * S5\end{aligned}$$

## 5 Strengths and Weaknesses

### **Strengths**

RMSE 0.635 for best model feature combination prediction

Inclusion of additional informative variables contributes to the reduction of MSE by 0.033

### **Weaknesses**

Grid search over various model parameters

Thank You!