Yelp Data Prediction

Preliminary Analysis

Yifan Li, Chenlai Shi, Jianmin Chen

Monday Group 1

1.1 Introduction

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

1.2 Data Cleaning

1. Modify Abbreviation and Special Symbol
2. Remove Non-English
3. Negative Sentences
4. Remove Punctuation

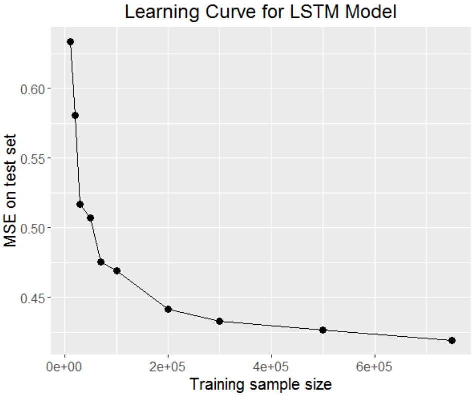
LSTM

- Networks with loops in them, allowing information to persist
- When users write reviews, their thoughts/scratches have persistence too

2 Model

- Input node: 100
- Output node: 50
- Dense layer: 2

2 Model

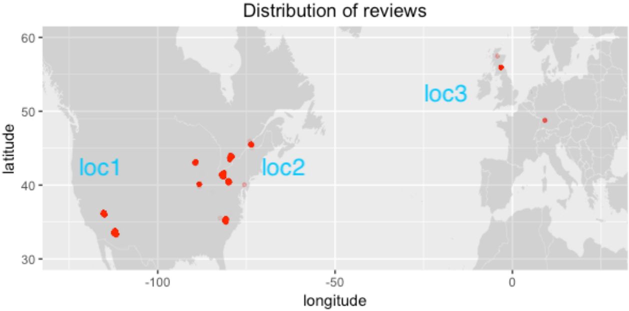

Learning Curve for LSTM Model
MSE on test set
0.60
0.55
0.50
0.45
0e+00
2e+05
4e+05
6e+05
Training sample size

Doc2vec

- Sentence embeddings
- Extension of word2vec

2.2 Additional Variable

- **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 estern United States, otherwise 0.
- **loc3**: 1 if the restaurant isn't in the United States, otherwise 0.



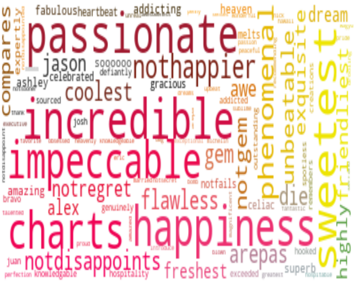
2.2 Additional Variable

- **S1** $\sim$ **S5**: $\text{S1[word]} = \frac{\text{P(this word is included in reviews with 1 star)}}{\text{P(this word is included in reviews with other stars)}}$

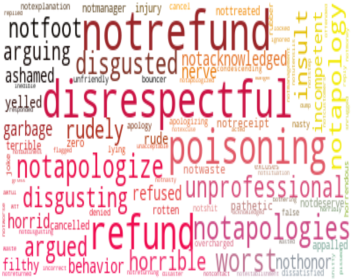
Word	Variable	1-star	2-star	3-star	4-star	5-star
refund	frequence	115	15	7	4	2
	probability	0.011	0.002	0	0	0
	S1 $\sim$ S5	34.200	1.080	0.300	0.072	0.025
notdisappoints	frequence	0	2	5	43	110
	probability	0	0	0	0.002	0.003
	S1 $\sim$ S5	0	0.116	0.188	0.917	3.870
and	frequence	9196	8691	12851	25604	32071
	probability	0.859	0.886	0.877	0.895	0.886
	S1 $\sim$ S5	0.968	1.000	0.991	1.020	1.000

2.2 Additional Variable

Positive



Negative



3 Compare RMSE with other method

RMSE

Feature\ Model	LM	NB	NN	LSTM	GLM	SVM
vector + ad	0.673	0.974	0.494	**0.493**	0.698	NA
vector	0.720	1.112	0.524	0.526	0.756	0.585
additional	**0.836**	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

4 Interpretable Model

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

5 Strengths and Weaknesses

Strengths

MSE 0.493 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!