Rating Prediction of Yelp Reviews

Shaodong Wang, Wenjin Li, Jia Liu, Jiayin Wang
University of Wisconsin-Madison – Department of Statistics

Outline

- Feature Extraction
- Model Selection
- Model Evaluation
- Conclusion
- Further Proposals

Feature Extraction

Unigram:

- Select the first 10,000 words with the largest frequencies
- Delete 5,000 words with high variance

$$variance(t) = \frac{\sum_i [\#\text{ of } t \text{ in } i \text{ star reviews} * i - mean(t)]^2}{\#\text{ of } t \text{ in all of the reviews}}$$

Bigram:

- Select the first 20,000 words with the largest frequencies
- Delete 10,000 words with high variance

Feature Extraction

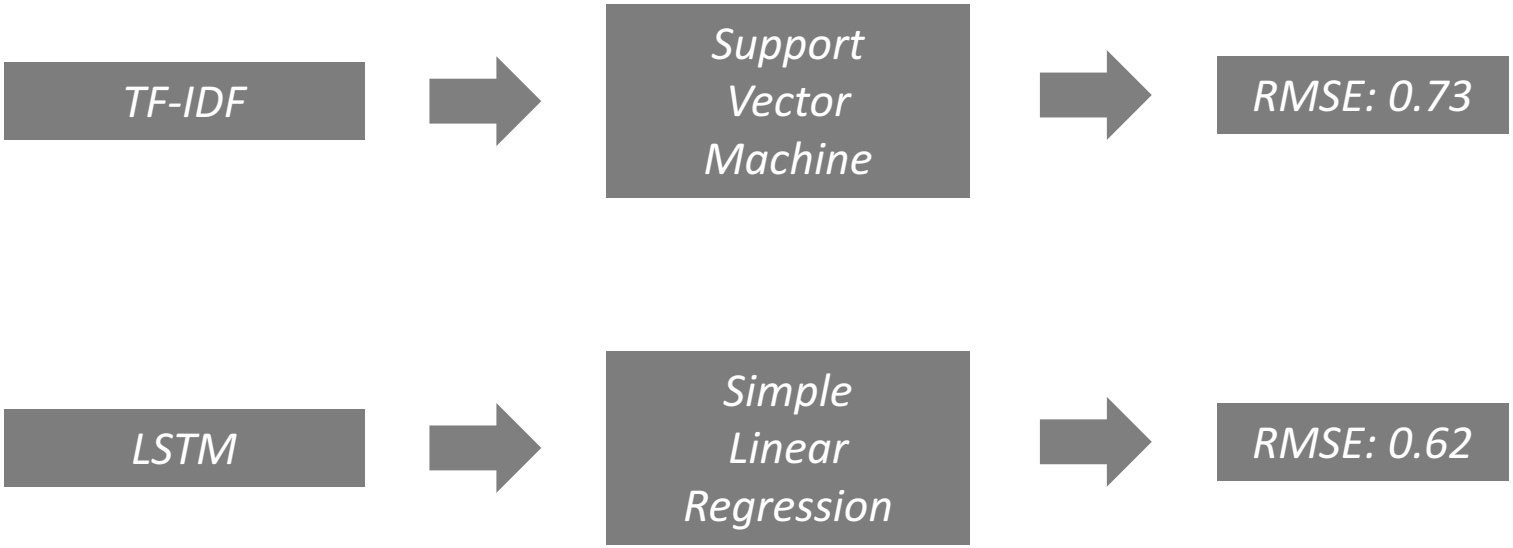
Predictors

- Unigram and bigram features from review texts
- Categories
- City
- Business name
- …

Model Selection

- Support Vector Machine with TF-IDF
- LSTM + Linear Regression
- Model Comparison

Model Selection



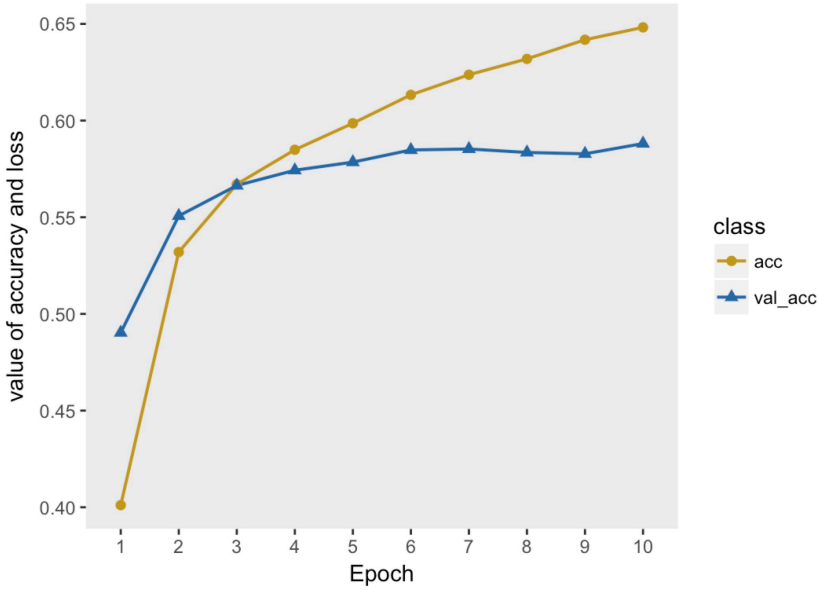
Model Selection

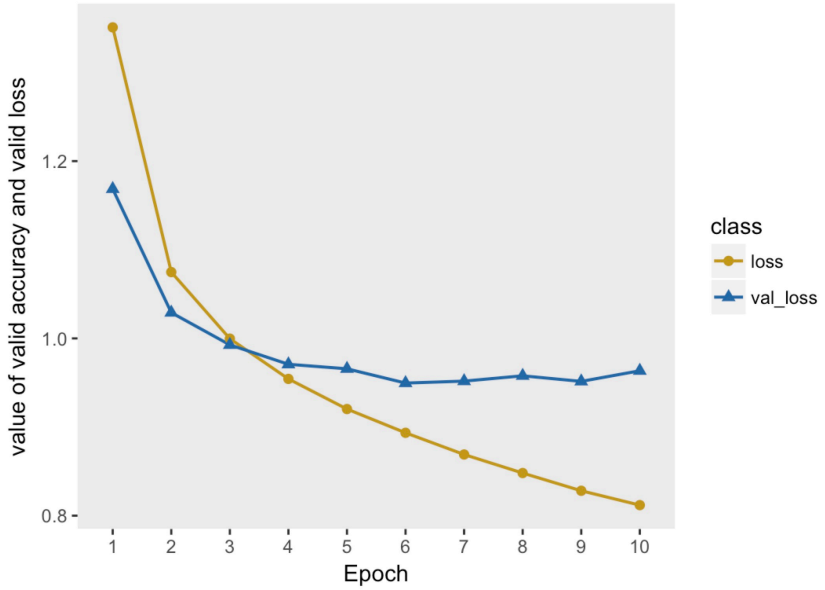
LSTM Model

- Vocabulary size: 5,000
- Maximum review length: 400
- Training set: 150k random samples for each of the 5 rating brackets
- Embedding: 128-dimension vectors
- Drop out rate: 0.2

Model Selection

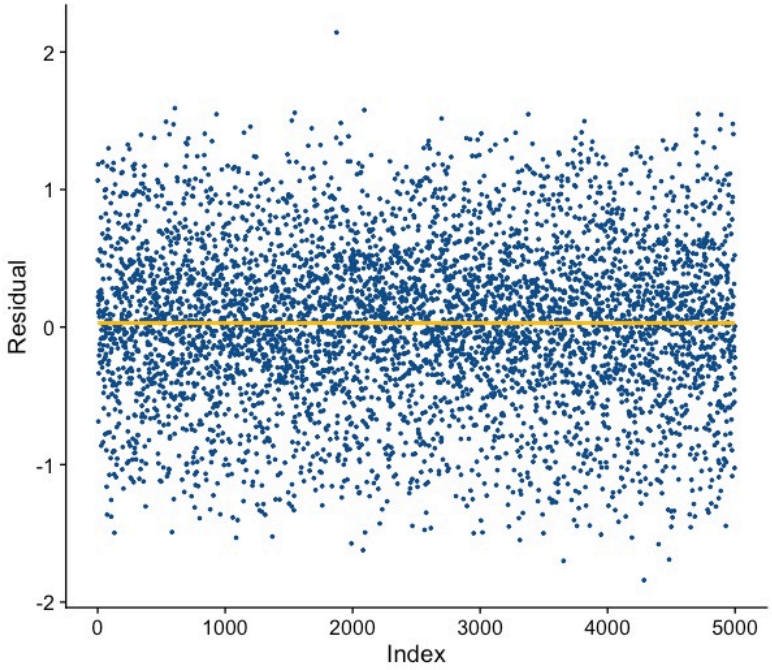
Accuracy and Loss of LSTM Model

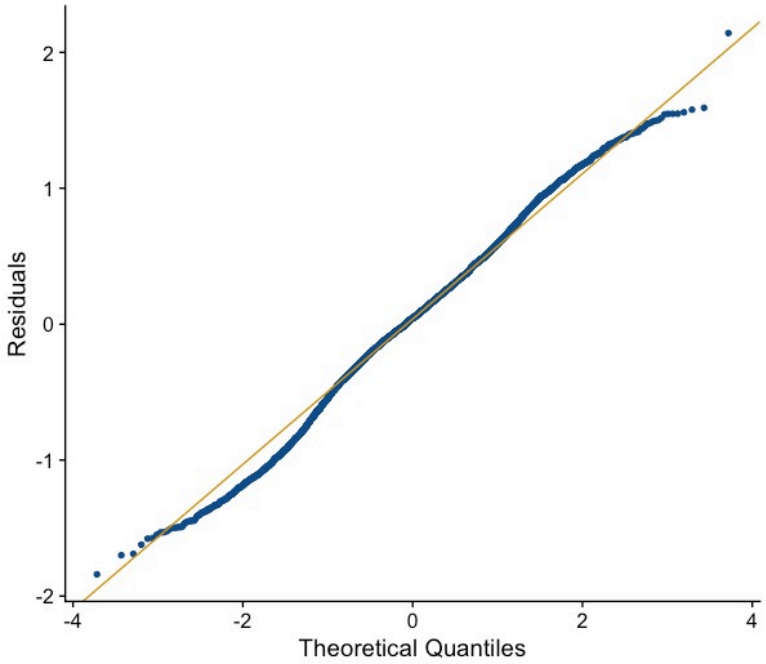




Model Selection

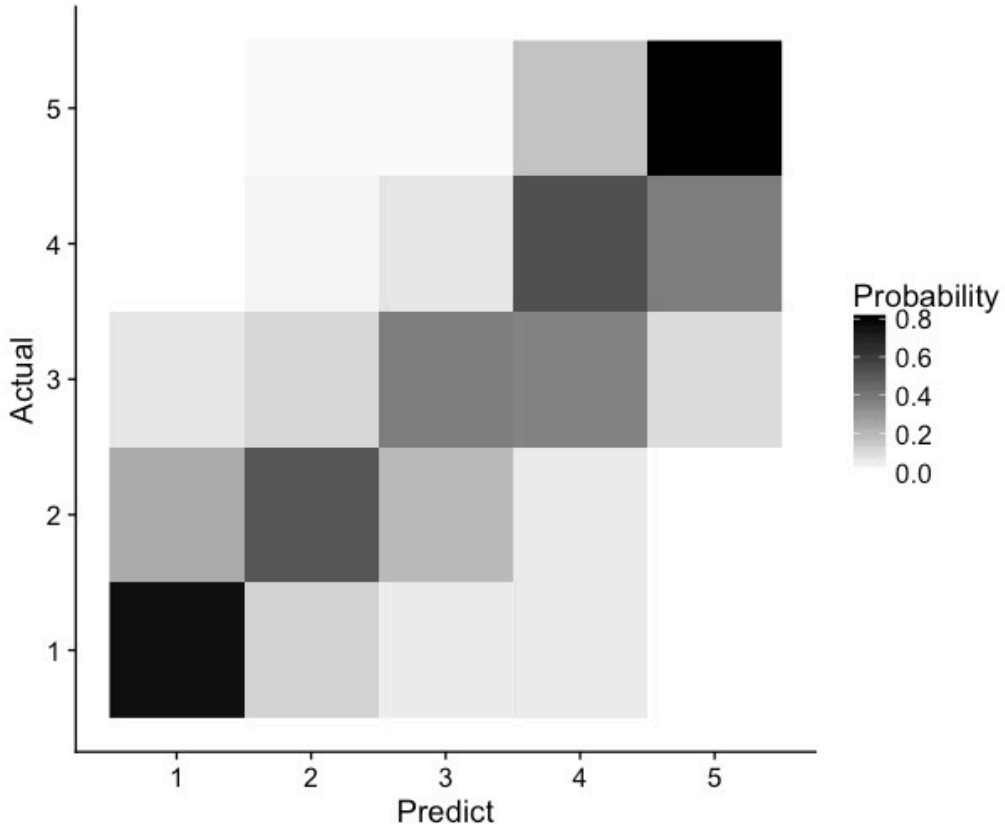
Residual Diagnostic





Model Selection

Confusion Matrix of Prediction Values



Performs better at both extremes of the rating scale: 1 and 5-star ratings.

Model Selection

Model	Ridge	K-nn	Random Forest	SLR	SVR	SVM	LSTM	LSTM+ Linear
Training Size	ALL	100K	200K	ALL	100K	ALL	600K	1,000K
Feature	UNI	UNI	UNI	UNI	UNI&BI	UNI&BI	UNI	UNI
RMSE	1.71	1.49	0.81	0.79	0.75	0.73	0.68	0.62

Model Evaluation

Strengths

- Accuracy
- Robustness

Weaknesses

- Time Consuming
- Feature Selection
- Typos
- Dimension Reduction

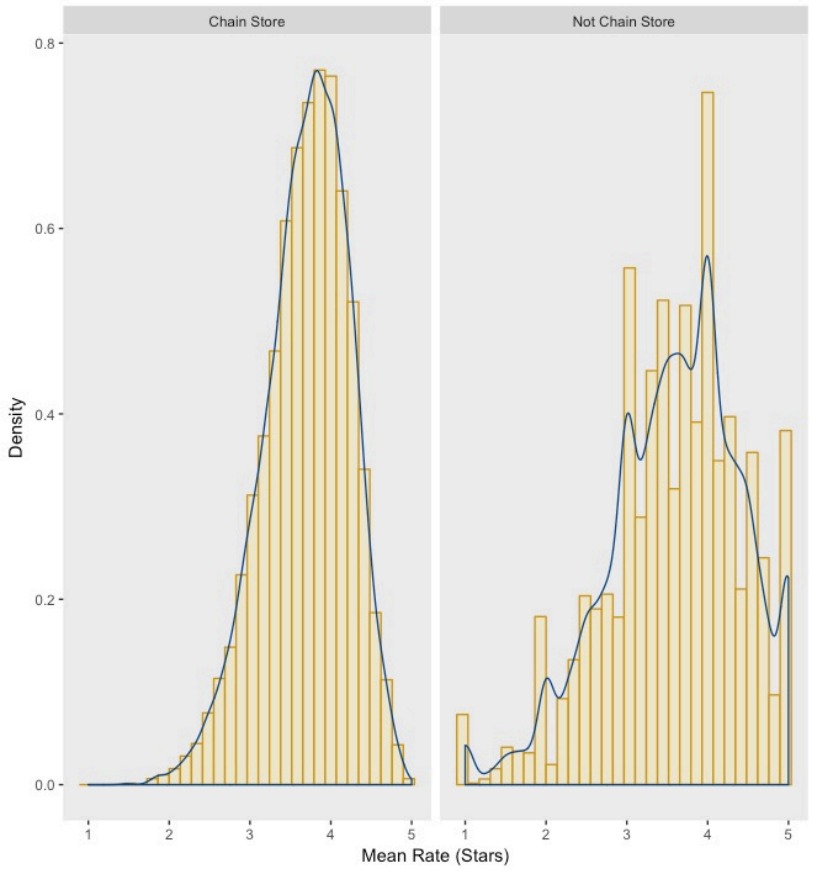
Conclusion

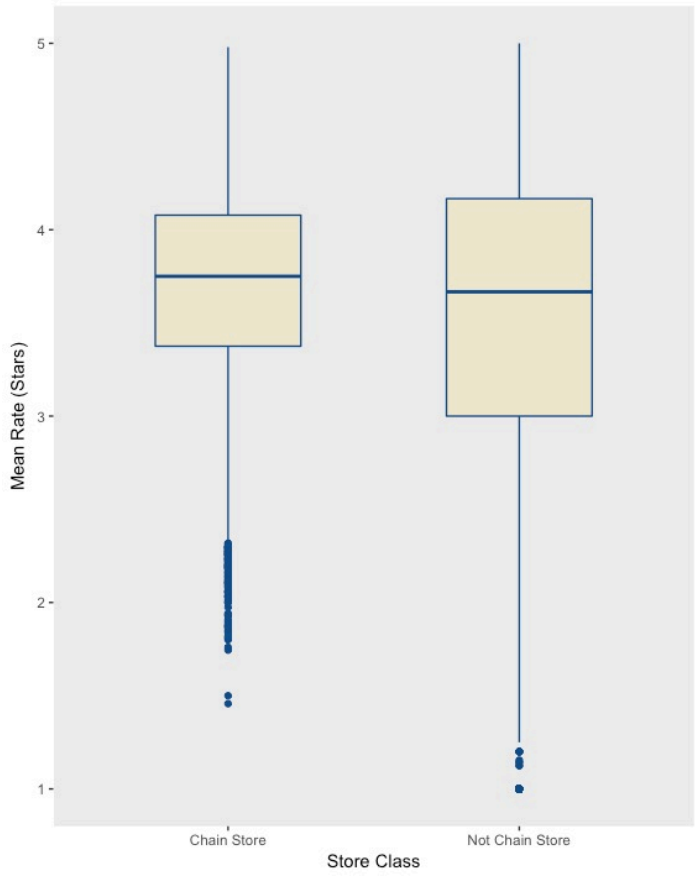
- What makes a review positive or negative?

- Propose a prediction model to predict the ratings of reviews:
  - Remove the meaningless part in the text reviews.
  - Create sparse matrices based on unigram and bigram features by TF-IDF.
  - Achieve an accurate prediction model through the combination of LSTM using 100-dimension vectors and simple linear model.
  - Add features into model such as categories to improve accuracy.

Further Proposals

Average Stars by Business Type





Questions?