
# Yelp Data Analysis

Group 3 : Yiqiao Zhang Xiawei Wang Shuyi Qu

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

# Data Pretreatment

## Negation transformation

- Find common 2-word and 3-word idioms contains negation words by tokenizing text into bigrams and trigrams
- Whenever a negation word occurs in a sentence,

  - if it occurs as a part of idioms, leave it alone;

  - else, attach “not” in front of every word after that in the same sentence and delete the negation word itself

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

# Data Pretreatment


No matter how the food is, it's ==not worth== being treated like that. Will ==never== be going ==back==.

No matter how the food is, it’s ==NOTworth== being treated like that. Will be going ==NOTback==.

# Data Pretreatment

- Lowercase first letter in each word
- Remove common stop words which did not show significant trend among different stars, like “I”, “me”
- Negation word transformation

UNIVERSITY OF WISCONSIN–MADISON

# Feature selection

Feature unigrams and bigrams are supposed to

- be highly used in text
- show different trends among star rates which are quantified by formulas below：

$$\sum_{i=1}^{5} occur(i)$$

$$\left(\sum_{i=1}^{5}\sum_{j=i+1}^{5} |occur(i) - occur(j)|\right) / \sum_{i=1}^{5} occur(i)$$

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

# Feature selection

Features selected include：

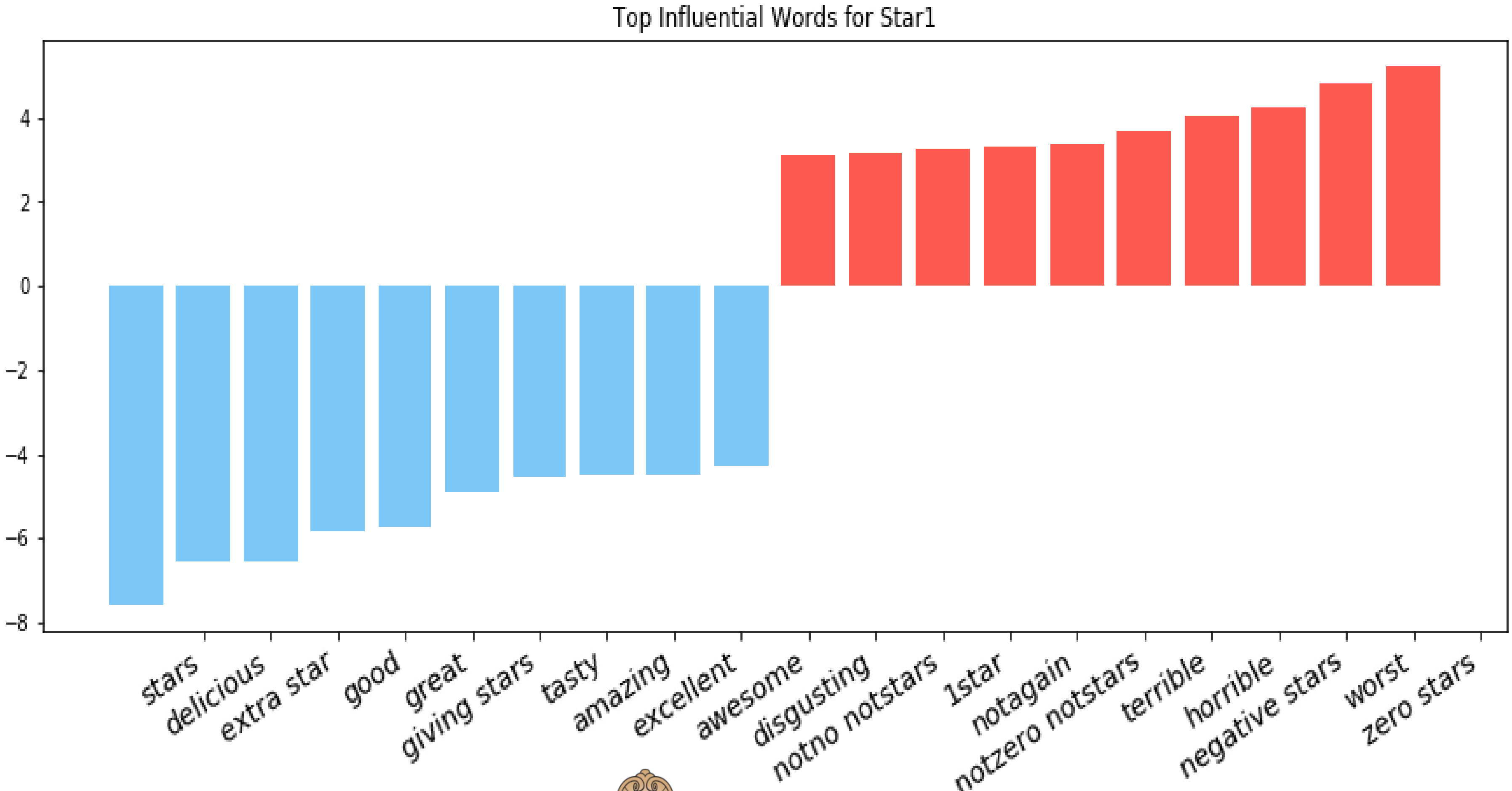
- Intersections of highly used unigrams and unigrams with large scaled L1-norm among star rates
- Intersections of highly used bigrams and bigrams with large scaled L1-norm among star rates
- Punctuations that convey attitude like "!", "?", "*" and "$"

# Design Matrix

- Each column represents a feature we selected
- Each row represent one piece of reformatted text
- Entry in this matrix is either 0 or 1
- Another design matrix is generated by features' frequency in each reformatted text

# Interpretable Model

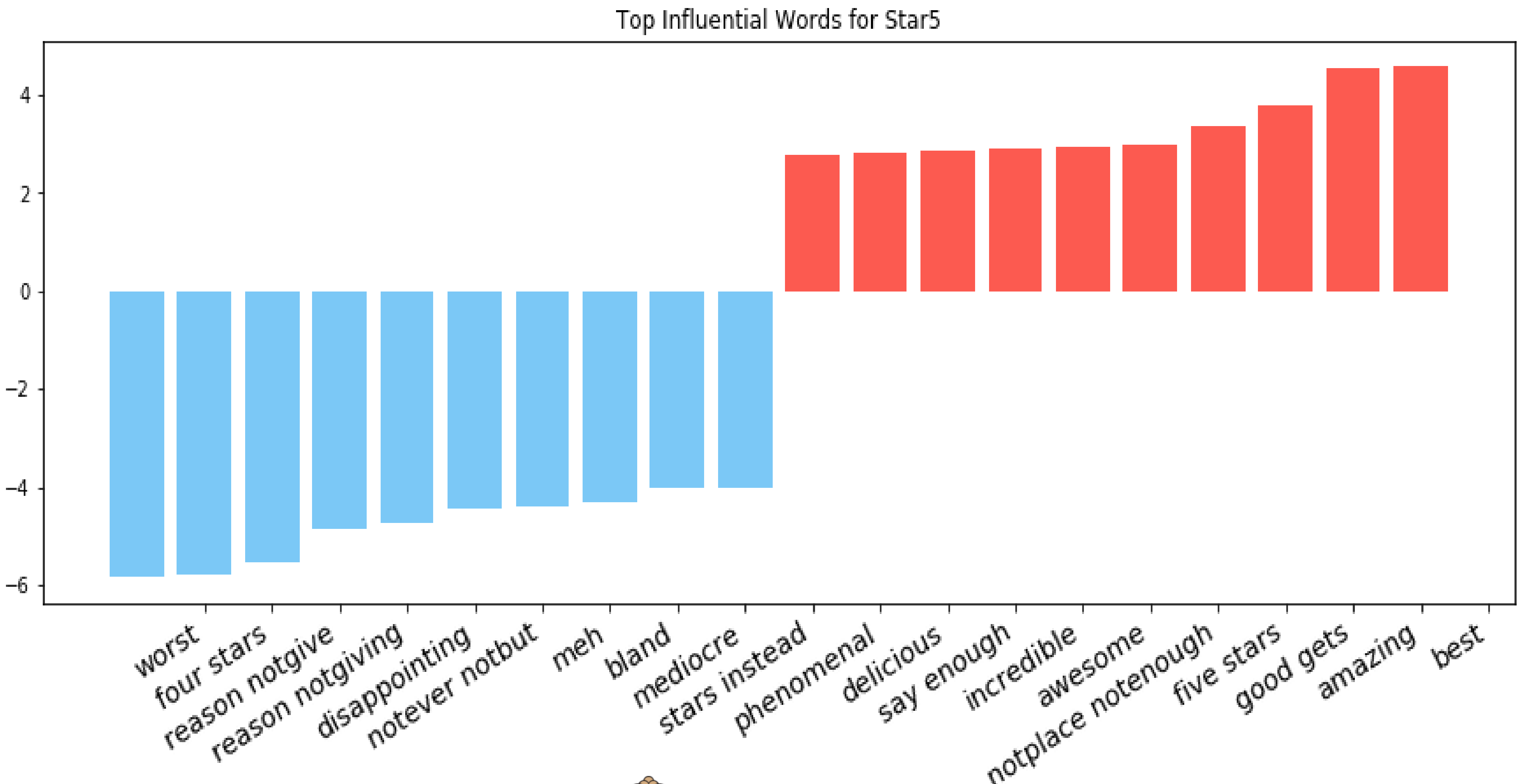
## Model: Linear Support Vector Machine




WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

# Interpretable Model

## Model: Linear Support Vector Machine




WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

# Interpretable Model

## Model: Linear Support Vector Machine

| Actual<br>Prediction | star 1 | star 2 | star 3 | star 4 | star 5 | |
|---|---|---|---|---|---|---|
| star 1 | 97586 | 2436 | 823 | 204 | 95 | 101144 |
| star 2 | 690 | 86221 | 1043 | 233 | 41 | 88228 |
| star 3 | 321 | 1887 | 123610 | 2573 | 492 | 128883 |
| star 4 | 112 | 735 | 7637 | 235508 | 12933 | 256925 |
| star 5 | 72 | 210 | 2025 | 27540 | 322792 | 352639 |
| | 98781 | 91489 | 135138 | 266058 | 336353 | 927819 |

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

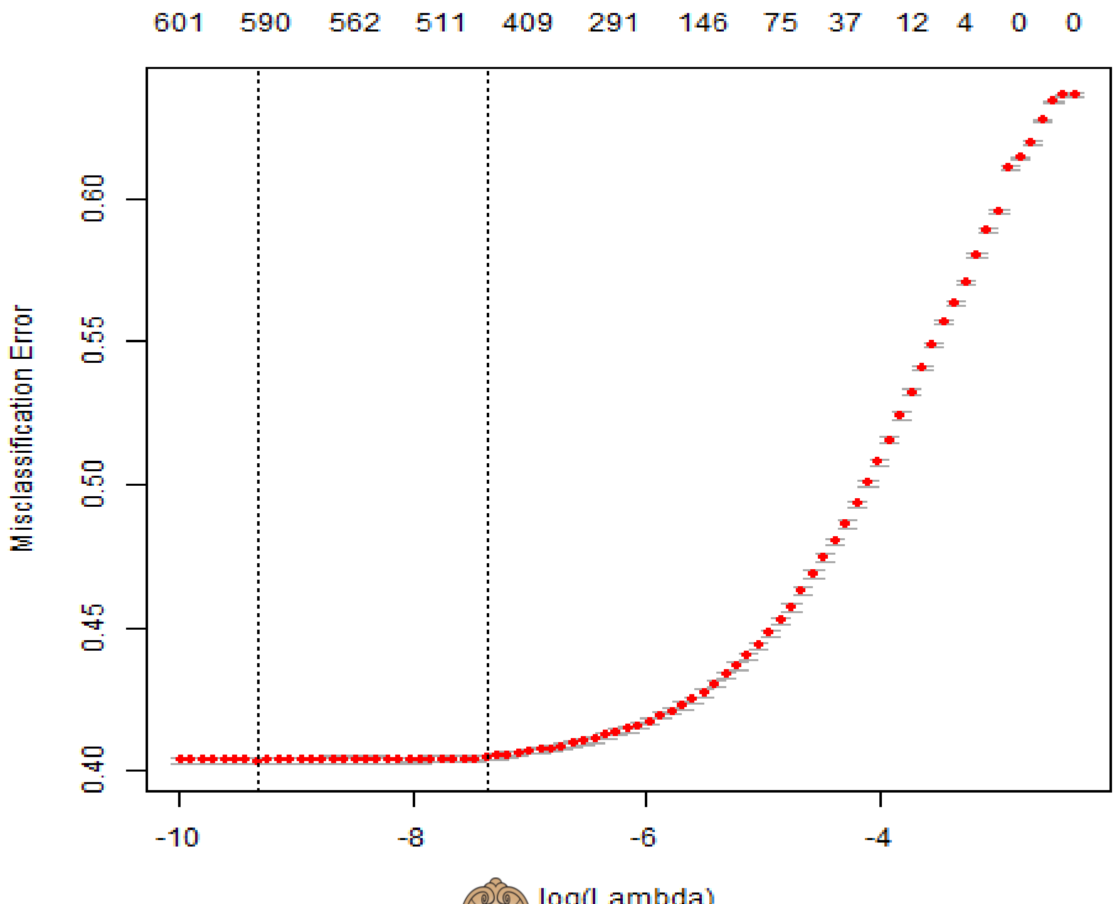
# Interpretable Model

## Model: Linear Support Vector Machine

| Statistics | Star 1 | Star 2 | Star 3 | Star 4 | Star 5 |
|---|---|---|---|---|---|
| Sensitivity | 0.99 | 0.94 | 0.91 | 0.89 | 0.96 |
| Specificity | 1 | 1 | 0.99 | 0.97 | 0.95 |
| Precision | 0.96 | 0.98 | 0.96 | 0.92 | 0.92 |
| Recall | 0.99 | 0.94 | 0.91 | 0.89 | 0.96 |
| Balanced Accuracy | 0.99 | 0.97 | 0.95 | 0.93 | 0.95 |
| Overall Accuracy | | | 0.9331 | | |

# Weakness

1. Model: Multinomial Regression with Lasso Penalty




WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

# Weakness

1. Model: Multinomial Regression with Lasso Penalty

| Actual<br>Prediction | Star 1 | Star 2 | Star 3 | Star 4 | Star 5 | |
|---|---|---|---|---|---|---|
| Star 1 | 11970 | 3916 | 1176 | 487 | 375 | 17924 |
| Star 2 | 2058 | 5438 | 2528 | 527 | 202 | 10753 |
| Star 3 | 679 | 3564 | 8718 | 3554 | 672 | 17187 |
| Star 4 | 660 | 1597 | 7576 | 21597 | 8943 | 40373 |
| Star 5 | 1060 | 943 | 2311 | 17977 | 46091 | 68382 |
| | 16427 | 15458 | 22309 | 44142 | 56283 | 154619 |

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

# Weakness

## 1. Model: Multinomial Regression with Lasso Penalty

| Statistics | Star 1 | Star 2 | Star 3 | Star 4 | Star 5 |
|---|---|---|---|---|---|
| Sensitivity | 0.73 | 0.35 | 0.39 | 0.49 | 0.82 |
| Specificity | 0.96 | 0.96 | 0.94 | 0.83 | 0.77 |
| Precision | 0.67 | 0.51 | 0.51 | 0.53 | 0.67 |
| Recall | 0.73 | 0.35 | 0.39 | 0.49 | 0.82 |
| Balanced Accuracy | 0.84 | 0.66 | 0.66 | 0.66 | 0.8 |
| Overall Accuracy | | | 0.6067 | | |

# Weakness

## 2. Model: Long-Short Term Memory Network

1) Input Layer:  new text
2) Embedding Layer: mapping to 32-dim space
3) LSTM layer: batch size of 32
4) Output Layer: classification
5) Configure learning process: MSE loss

# Weakness

2. Model: Long-Short Term Memory Network

   1. Input Layer
   2. Embedding Layer
   3. Output Layer
   4. Configure learning process
   5. Parameter Tuning

# Weakness

## 2. Model: Long-Short Term Memory Network

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, None, 32)          160000
_________________________________________________________________
lstm_1 (LSTM)                (None, 32)                8320
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 33
=================================================================
Total params: 168,353
Trainable params: 168,353
Non-trainable params: 0
_________________________________________________________________
None
Train on 742254 samples, validate on 185564 samples
```

# Weakness

## 2. Model: Long-Short Term Memory Network

```
Train on 742254 samples, validate on 185564 samples
Epoch 1/5
742254/742254  - loss: 0.5342 - val_loss: 0.4289
Epoch 2/5
742254/742254  - loss: 0.4273 - val_loss: 0.4135
Epoch 3/5
742254/742254  - loss: 0.4111 - val_loss: 0.4094
Epoch 4/5
742254/742254  - loss: 0.4013 - val_loss: 0.4076
Epoch 5/5
742254/742254  - loss: 0.3931 - val_loss: 0.4042
```

# Strength

1. Subtle data cleaning
2. Flexible feature selection
3. Diverse model application
4. Always keep trying

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON