

User Review Classification

By Rachel Edwards

Presentation Outline

01

Business Proposition

02

Business Solution

03

Data

04

Yelp EDA

05

Final Model

06

Grubhub

07

Thank You

08

Appendix

Business Proposition

- Use existing data, reviews/ratings
- Fix three star review ambiguity
- Help small businesses make the most of their data



Business Solution

- An NLP based Binary Text Classification Model

- Takes in text based reviews
- Cleans up data
- Picks out key linguistic features
- Makes predictions based on those features
- Classifies a review as Positive or Negative

Yelp Open Dataset

An all-purpose dataset for learning



The Data :

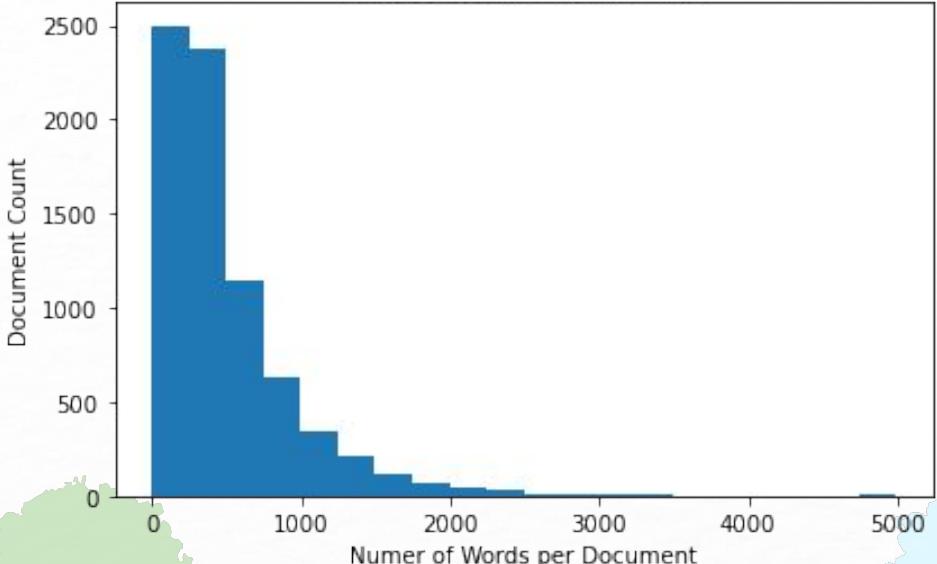


8,635,403 reviews

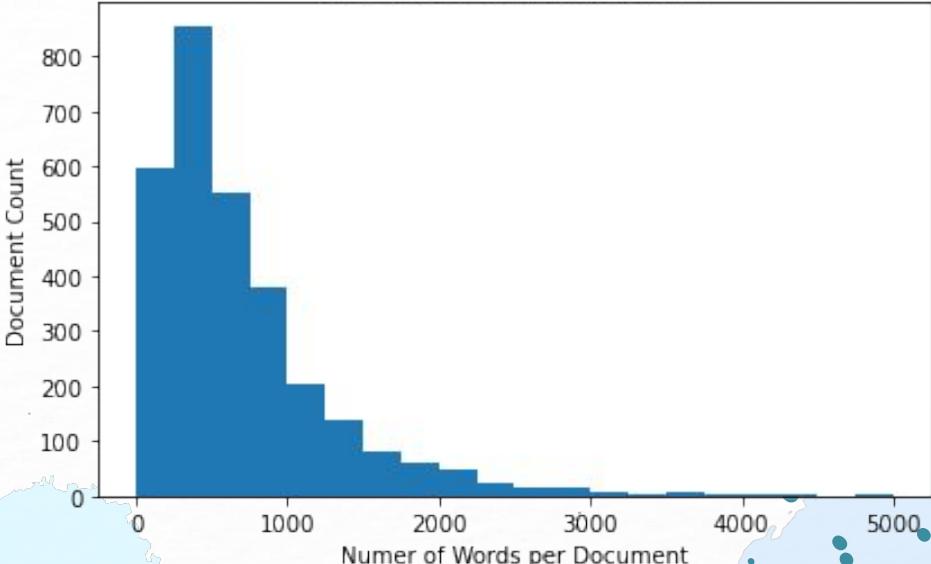
- Yelp OpenDataset
- Eight Million Reviews
- Ratio of 5 positive to 2 negative reviews

EDA

Positive Document Length



Negative Document Length



EDA

Positive WordCloud



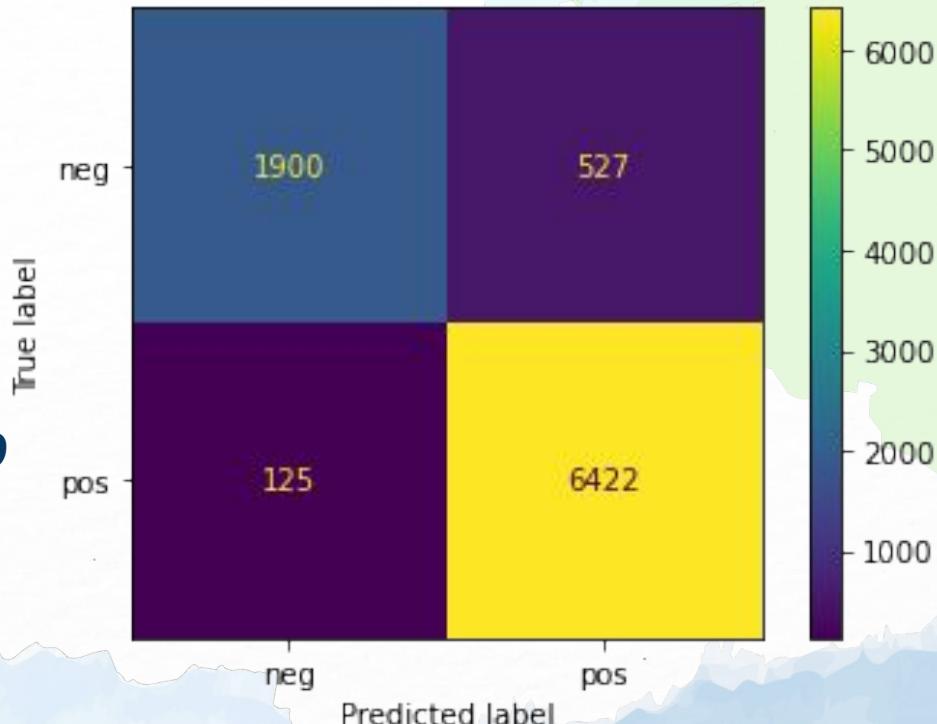
Negative WordCloud



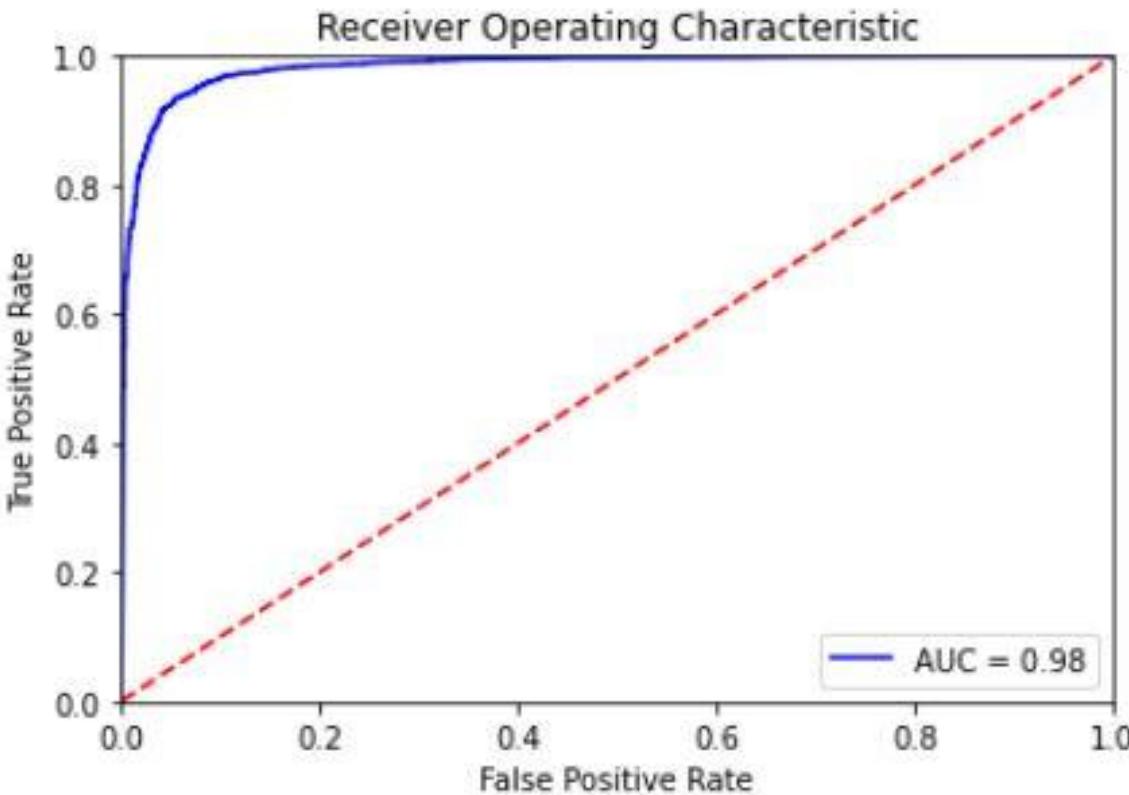
Final Model

- SGD Model
- 5000 features total
- Brigrams and Unigrams

-Mean Accuracy: **0.923111641093959**



ROC Curve



Key Features

Positive Features

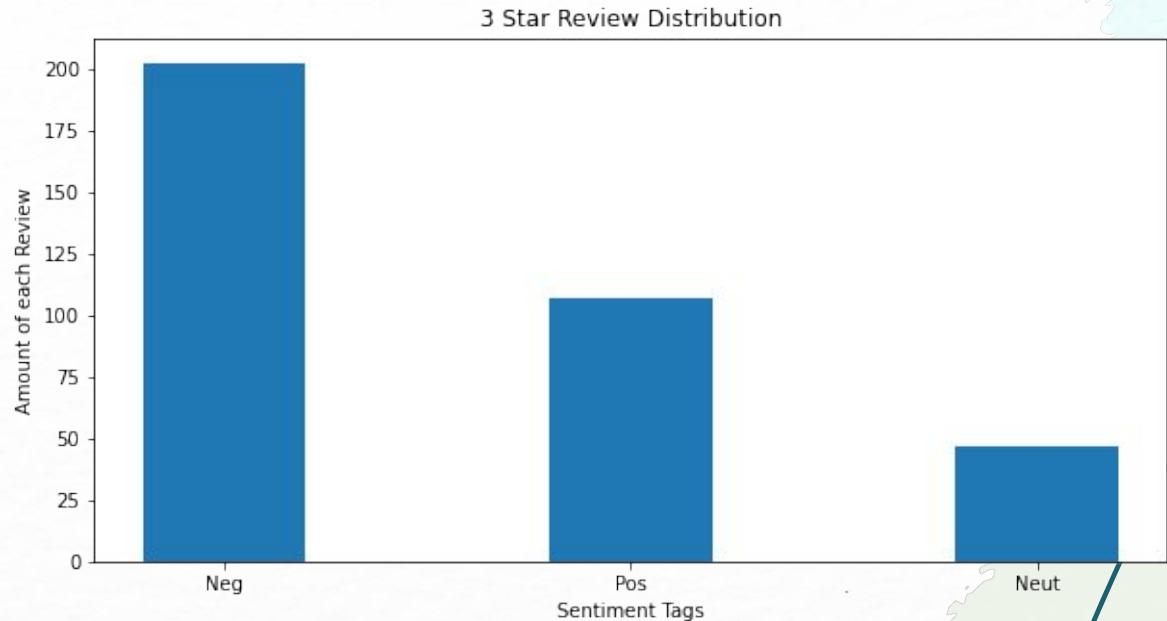
| | features | values |
|------|-----------|----------|
| 124 | amaze | 1.529868 |
| 383 | best | 1.737771 |
| 1047 | delicious | 1.758500 |
| 2559 | love | 1.874754 |
| 1954 | great | 2.269837 |

Negative Features

| | features | values |
|------|----------|-----------|
| 298 | bad | -1.706756 |
| 2140 | horrible | -1.498959 |
| 4378 | terrible | -1.489581 |
| 3699 | rude | -1.461884 |
| 3140 | pay | -1.284664 |

EDA : Grubhub

- Ratio of 3 negative reviews to 2 Positive reviews



EDA: Grubhub

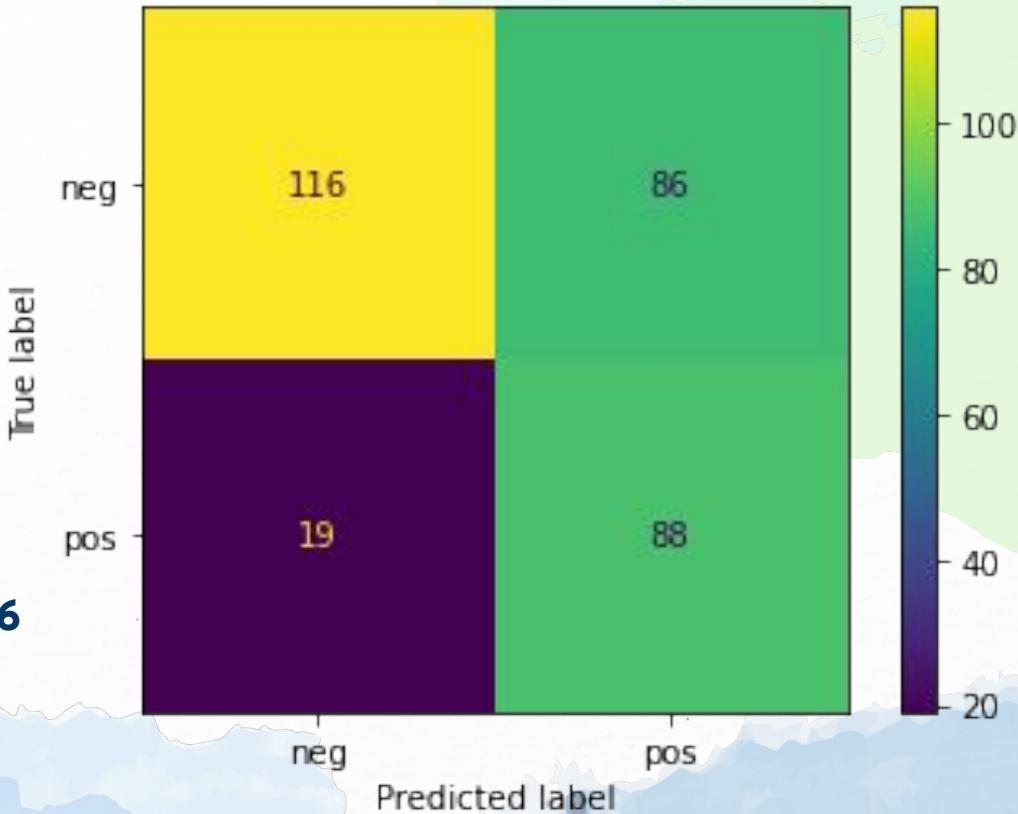
Negative Wordcloud



Grubhub Predictions

- SGD Yelp Model
- 5000 features total
- Brigrams and Unigrams

–Mean Accuracy: 0.7606028556319406



Thank You!

Rachel Edwards

Github

Linkedin

Appendix

Seattle Restaurant Alliance