Using NLP to predict Yelp Stars Given

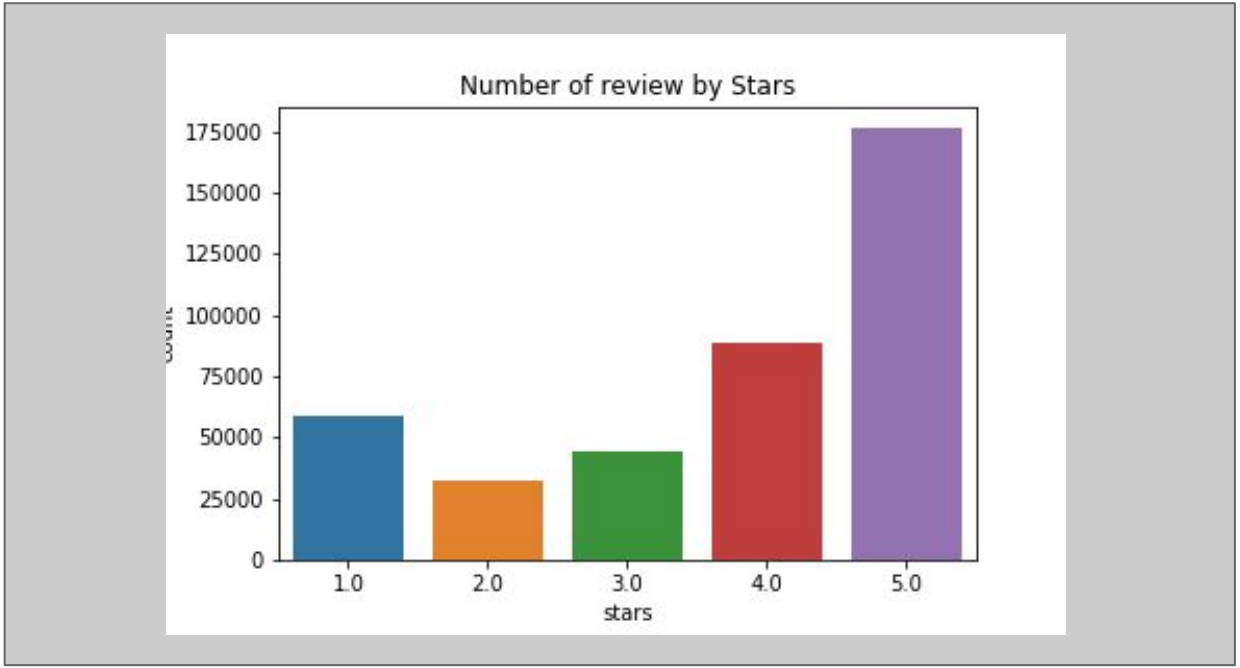
FlatIron School- Capstone Project

Purpose and Process

Can we predict number of stars given through Natural Language Processing?

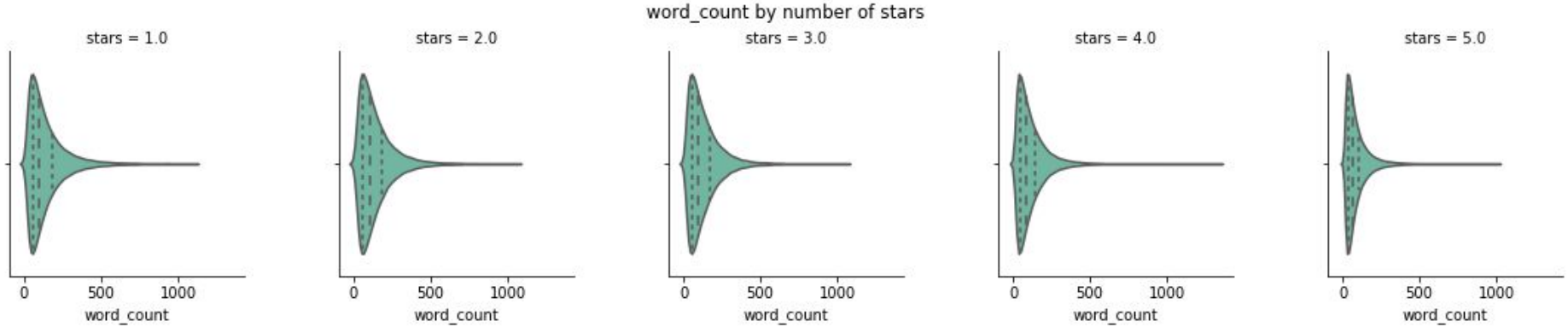
EDA Questions

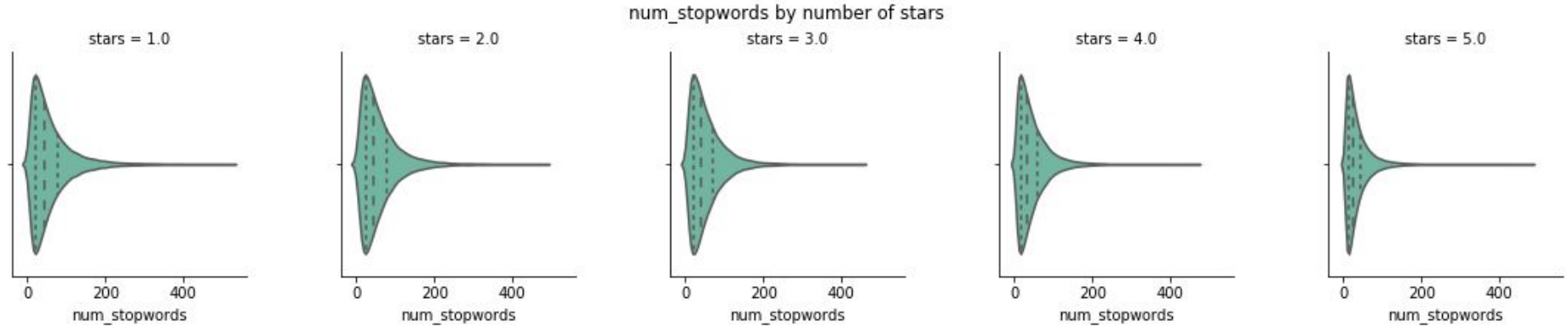
- **What is the breakdown of reviews given by customers?**
- Is there a difference in language used based on the number of stars given?
- Are there any important word associations and how do we capture those?
- Are there any themes that emerge from the data and do those themes vary by number of stars?

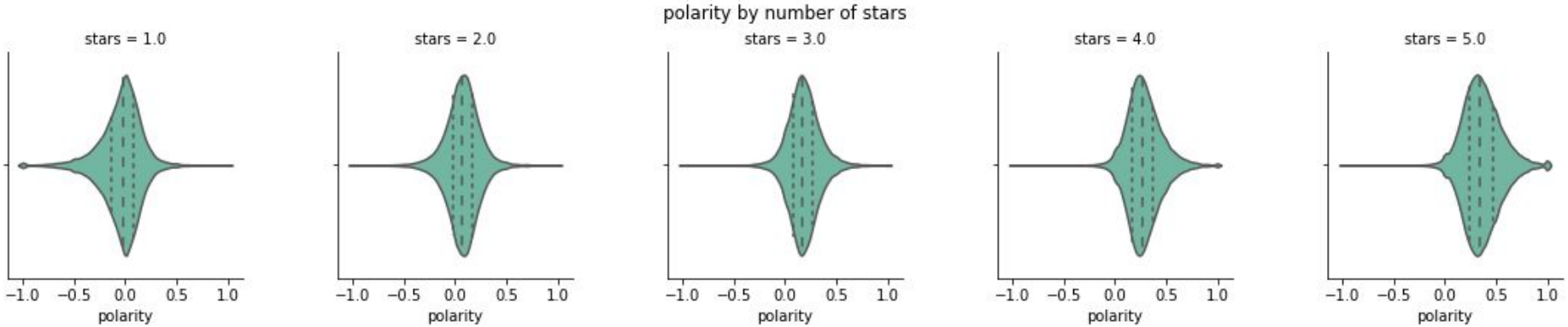

Number of review by Stars
175000
150000
125000
100000
75000
50000
25000
0
1.0
2.0
3.0
4.0
5.0
stars

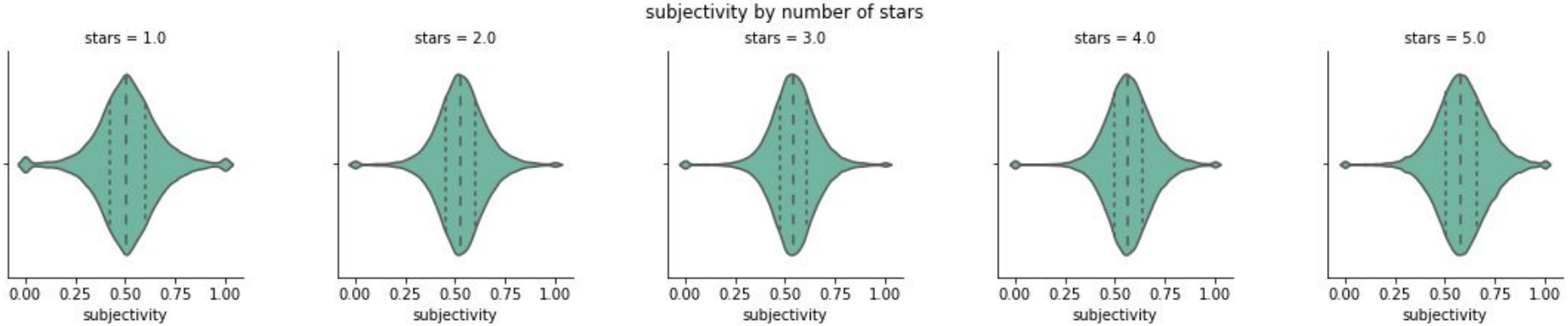
EDA Questions

- What is the breakdown of reviews given by customers?
- **Is there a difference in language used based on the number of stars given?**
- Are there any important word associations and how do we capture those?
- Are there any themes that emerge from the data and do those themes vary by number of stars?


word_count by number of stars
stars = 1.0
stars = 2.0
stars = 3.0
stars = 4.0
stars = 5.0
0
500
1000
word_count


num_stopwords by number of stars
stars = 1.0
stars = 2.0
stars = 3.0
stars = 4.0
stars = 5.0
0
200
400
num_stopwords


polarity by number of stars
stars = 1.0
stars = 2.0
stars = 3.0
stars = 4.0
stars = 5.0
−1.0 −0.5 0.0 0.5 1.0
polarity


subjectivity by number of stars
stars = 1.0
stars = 2.0
stars = 3.0
stars = 4.0
stars = 5.0
0.00 0.25 0.50 0.75 1.00
subjectivity

EDA Questions

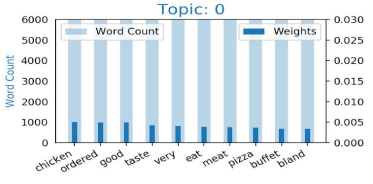
- What is the breakdown of reviews given by customers?
- Is there a difference in language used based on the number of stars given?
- **Are there any important word associations and how do we capture those?**
- Are there any themes that emerge from the data and do those themes vary by number of stars?

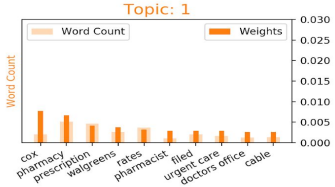
```
Most Similar Words to sandwhich ignoring None
------
panini: 0.8566
club sandwich: 0.8416
turkey sandwich: 0.8385
tuna salad: 0.8136
cobb salad: 0.8114
roast beef sandwich: 0.8080
blt: 0.8065
pastrami sandwich: 0.8028
turkey club: 0.7944
hoagie: 0.7884
```

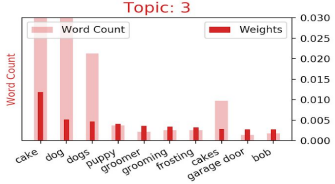
EDA Questions

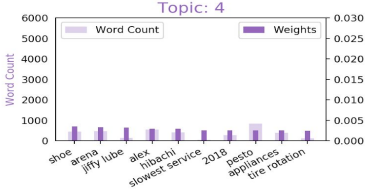
- What is the breakdown of reviews given by customers?
- Is there a difference in language used based on the number of stars given?
- Are there any important word associations and how do we capture those?
- **Are there any themes that emerge from the data and do those themes vary by number of stars?**

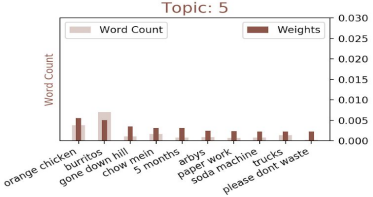
One Star Reviews

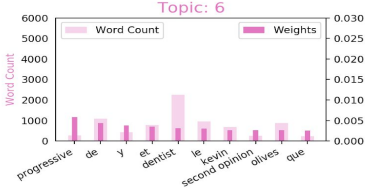


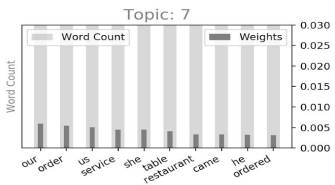


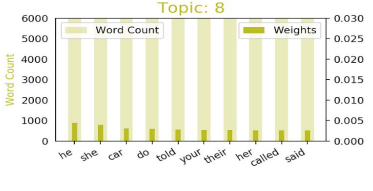


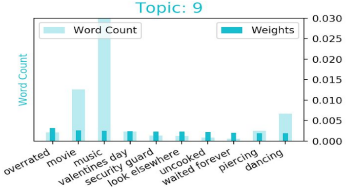




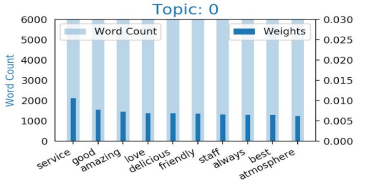


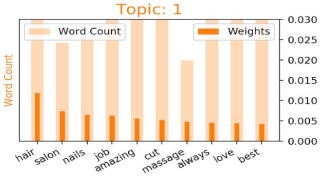


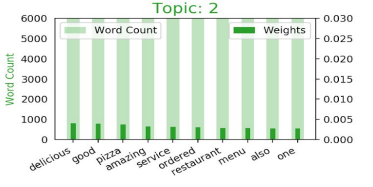


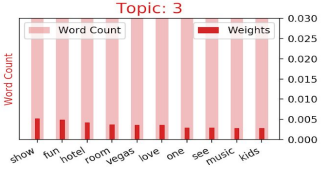


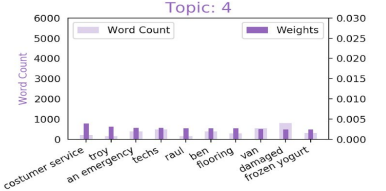
Five Star Reviews

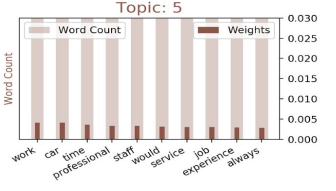


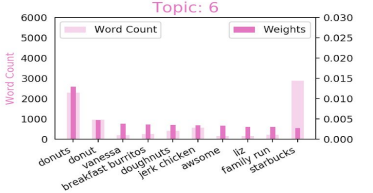


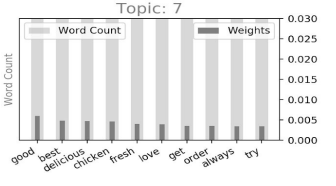


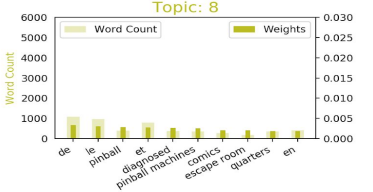


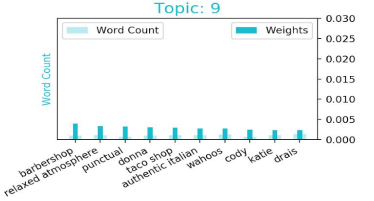












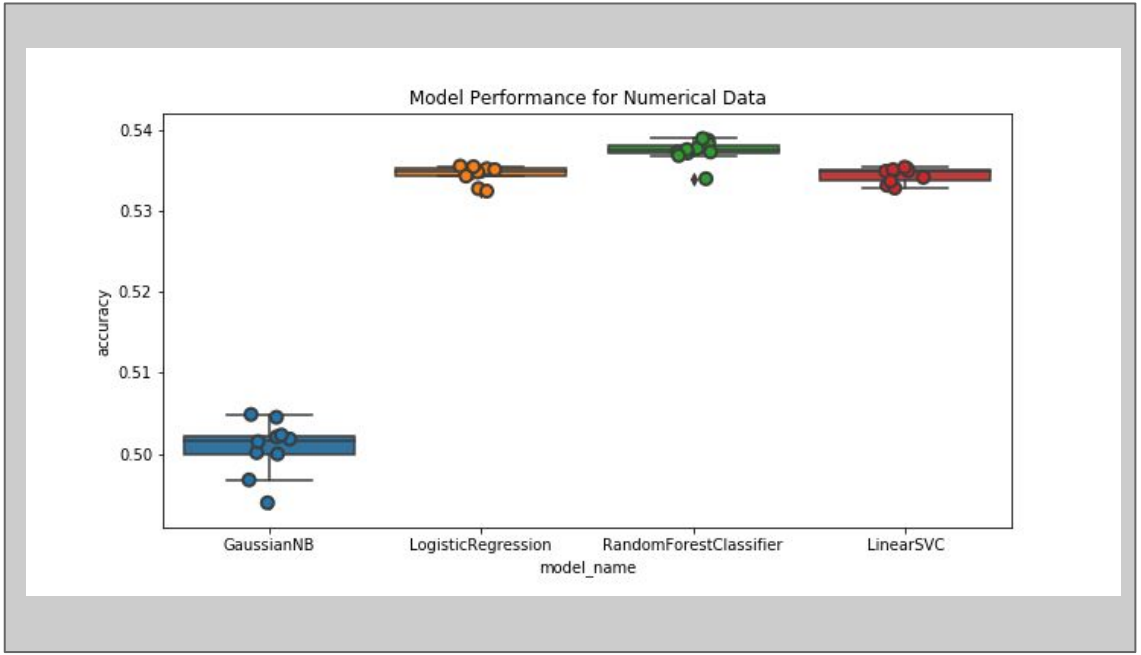
Model

Naive Bayes Model:
50.08% (+/- 0.3167%)

Logistical Regression Model:
53.46% (+/- 0.1023%)

Random Forest Model:
53.74% (+/- 0.1313%)

Linear SVC Model:
53.45% (+/- 0.08733%)



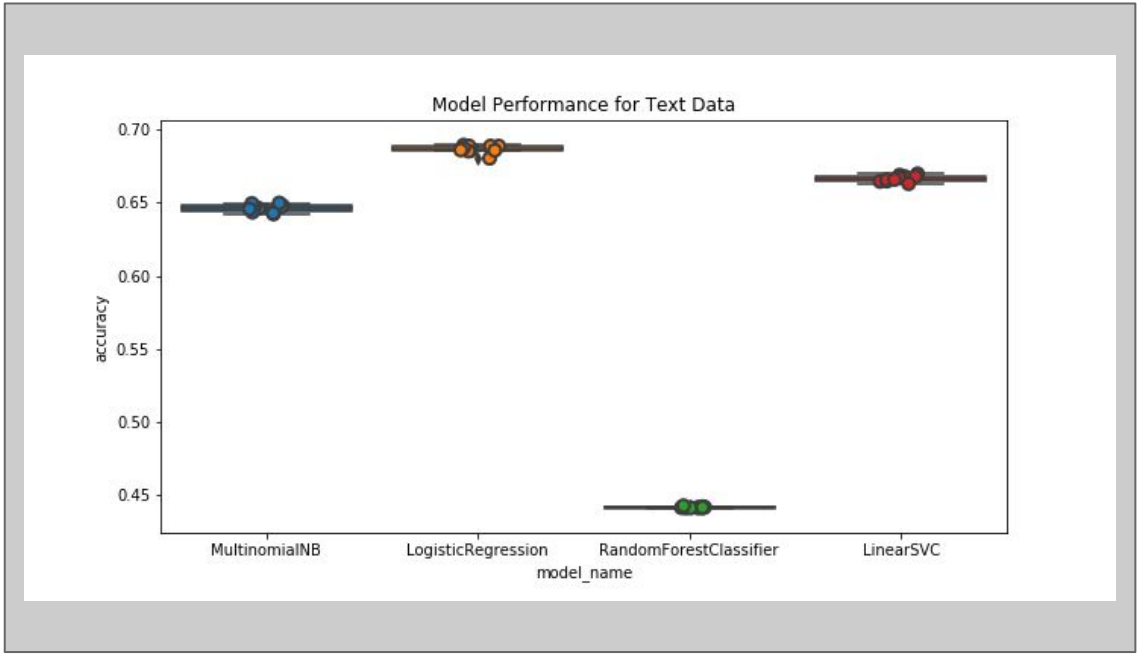
Model

Naive Bayes Model:
64.62% (+/- 0.2514%)

Logistical Regression Model:
68.65% (+/- 0.2543%)

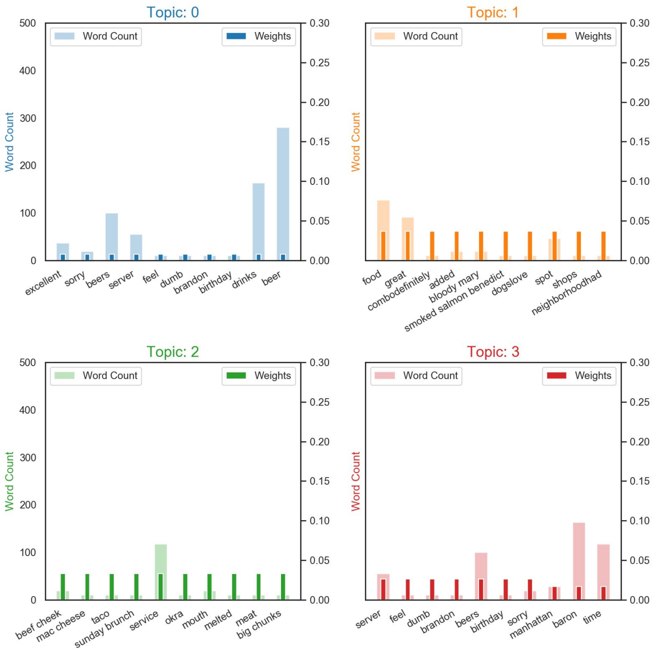
Random Forest Model:
44.19% (+/- 0.04548%)

Linear SVC Model:
66.64% (+/- 0.1989%)



Testing with New Data

5 Star Review Themes



Next Steps

1. Work to improve our cleaner function to improve our data.
2. Test an unsupervised model that combines both numerical and text data

Thank You!

Any questions?