r/TalesFromYourDataScientist:

Brandon Bergeron, Data Scientist

Data sources:

r/TalesFromRetail
r/TalesFromYourServer

The Process

Acquiring Data → Cleaning/preprocessing → Modeling

Web-scraping

Acquired roughly 10,000 posts from each subreddit spanning back over 2 years

Addressed issues such as:

- Duplicate posts
- Deleted posts
- Removing page names/etc. from posts
- Looking at keywords in each subreddit

Balancing performance with ____:

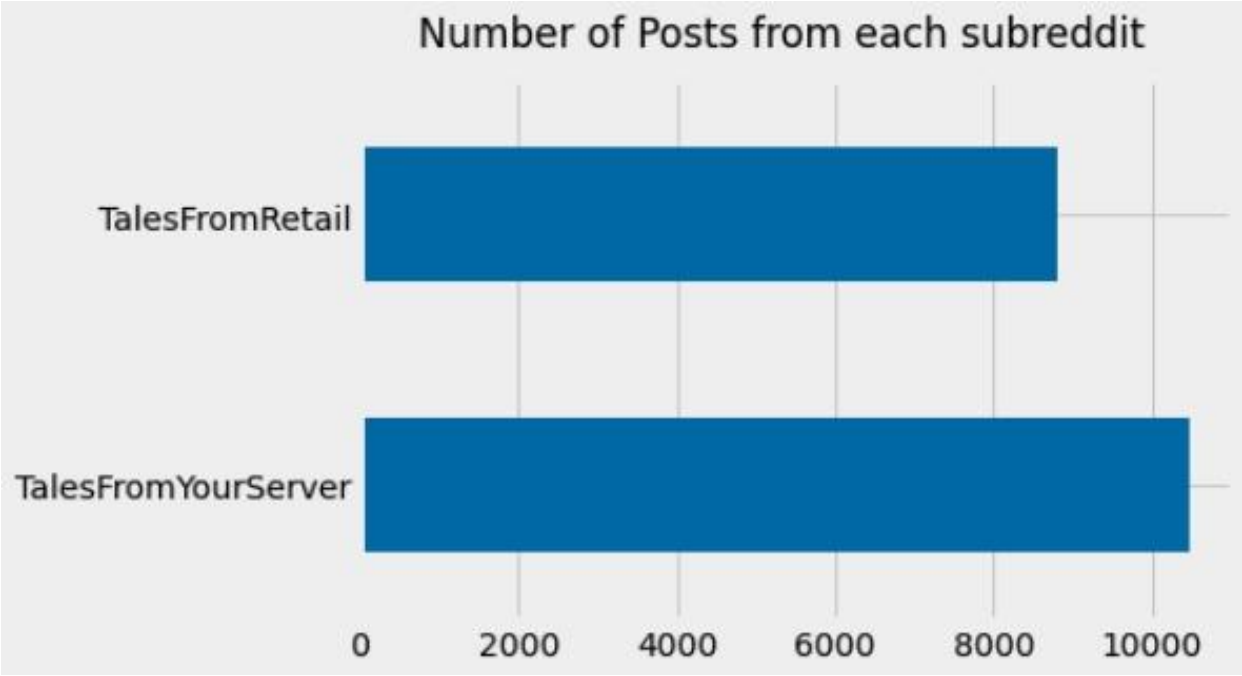
Building a model that _____

Number of posts gathered

Working with just under 20,000 posts after deleting removed posts and duplicates.

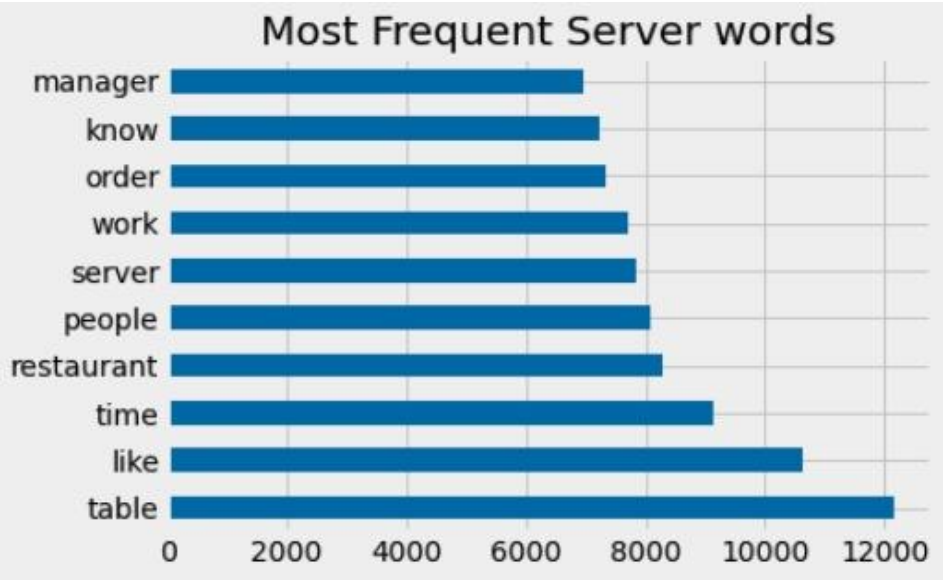
~9000 Retail

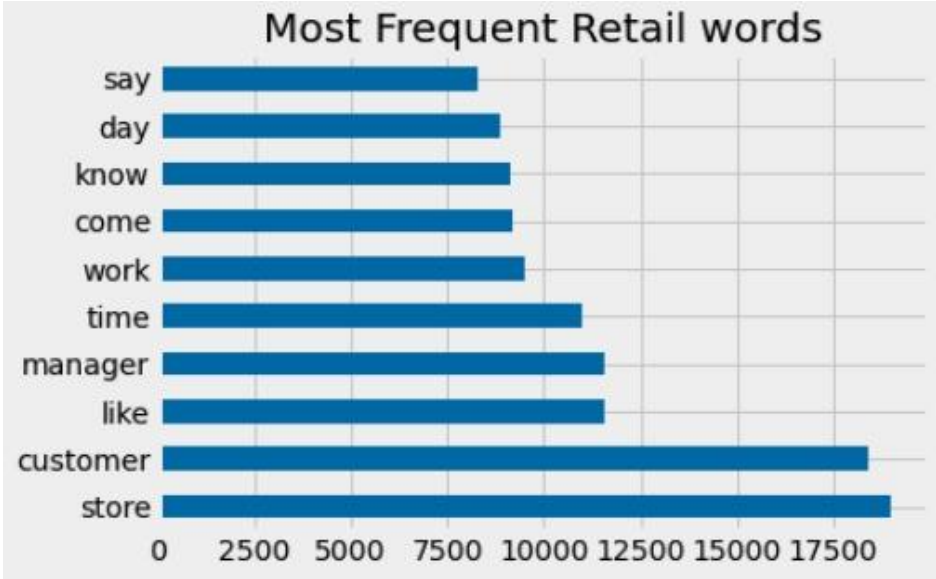
~10500 Server



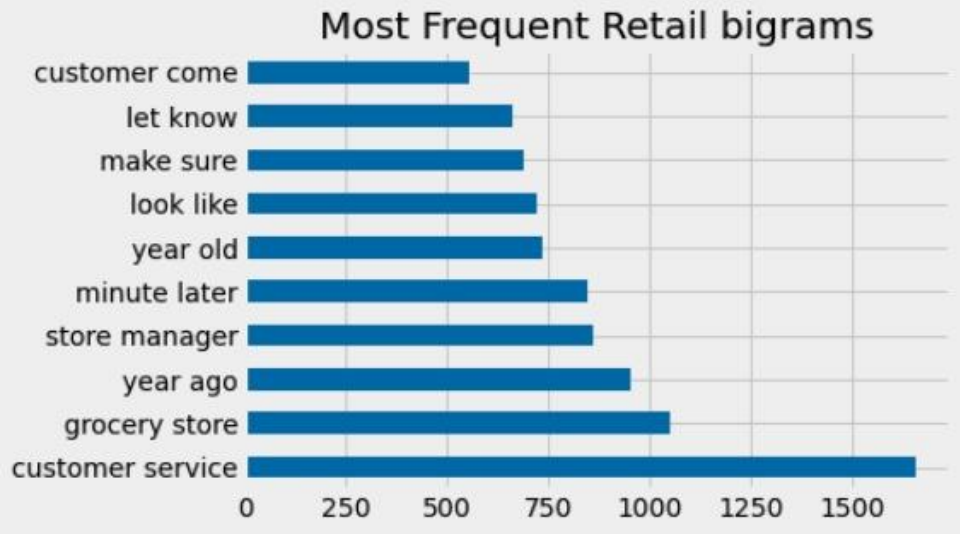
What’s in the two subreddits?

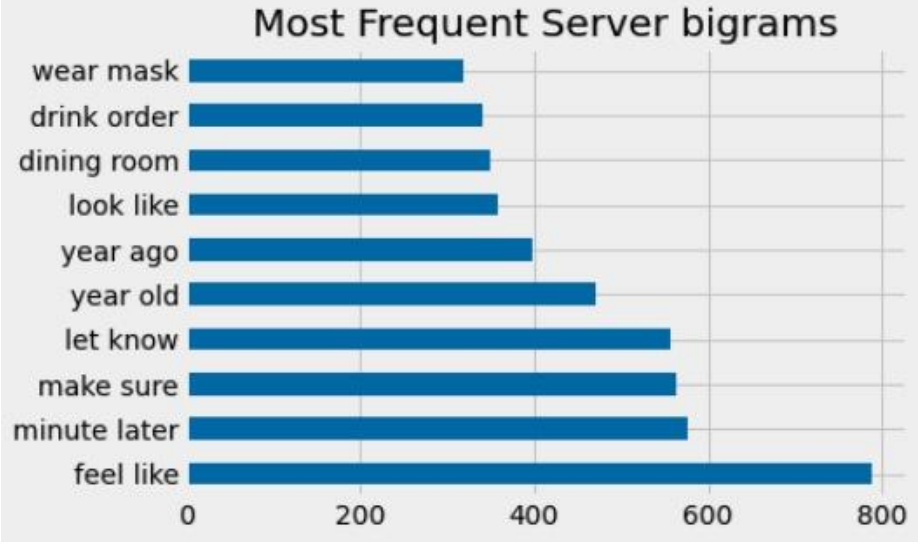
Highest occuring words



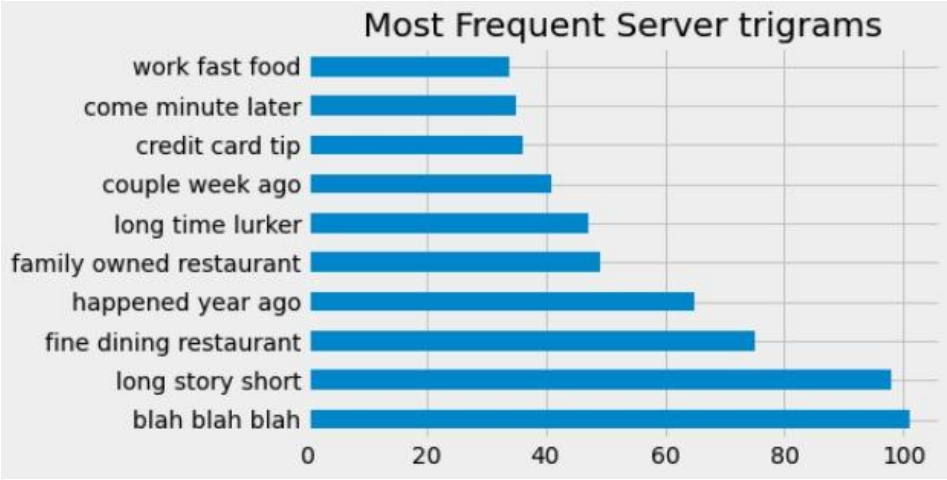


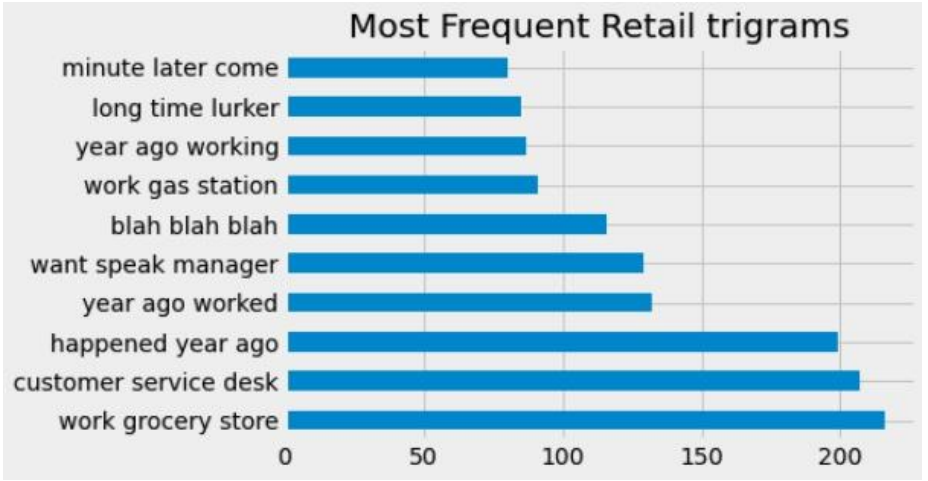
Highest Occuring Bigrams





Highest Occuring Trigrams





Modeling

After initial preprocessing, all models performed with between 90-95% accuracy in their basic states

Models tested:

- Logistic Regression
- Naive Bayes
- RandomForestClassifier
- AdaBoostClassifier
- GradientBoostingClassifier
- SVC

Best Performing Models:

Logistic Regression

- 94.5% accuracy on new data
- Fast to train
- Easy to interpret

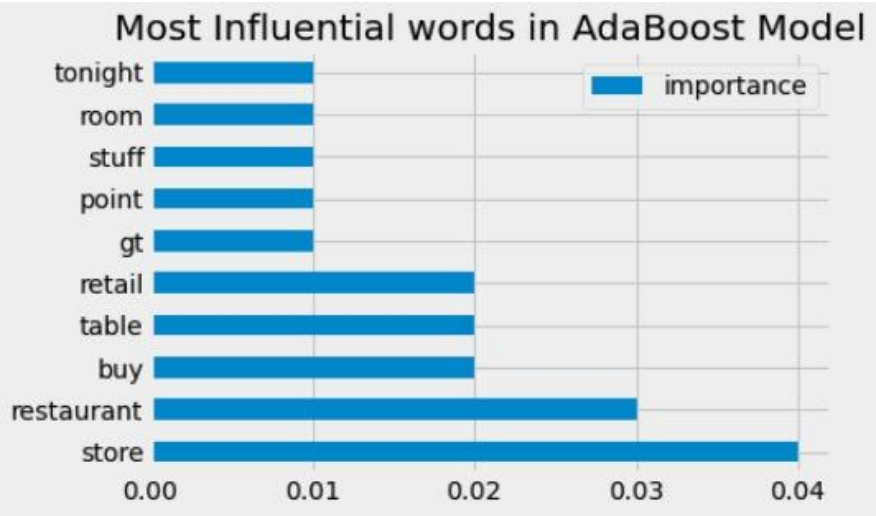
SVC

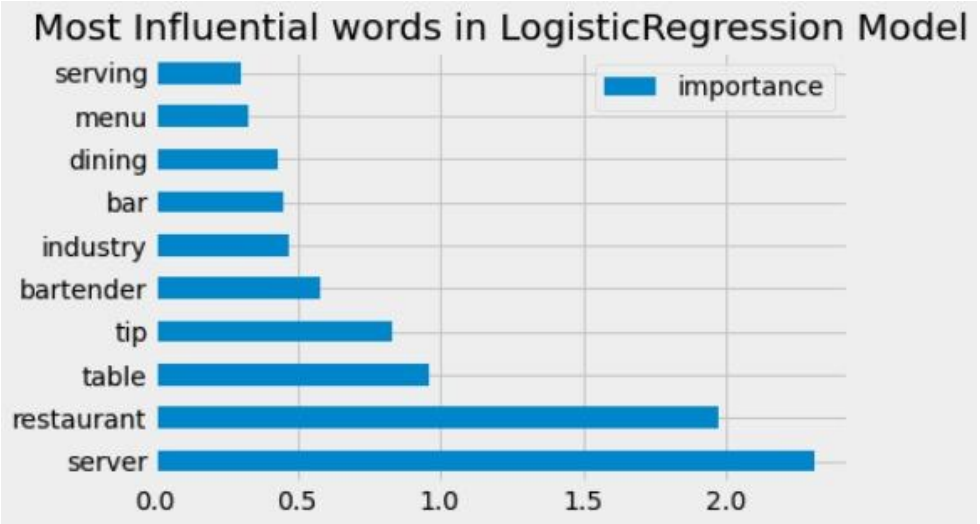
- 94.5% accuracy on new data
- least overfit with no parameter tuning
- CON: Slow...

AdaBoostClassifier

- 94.3% accuracy on new data
- Moderately quick to train
- Flexible

Model Defining Features:





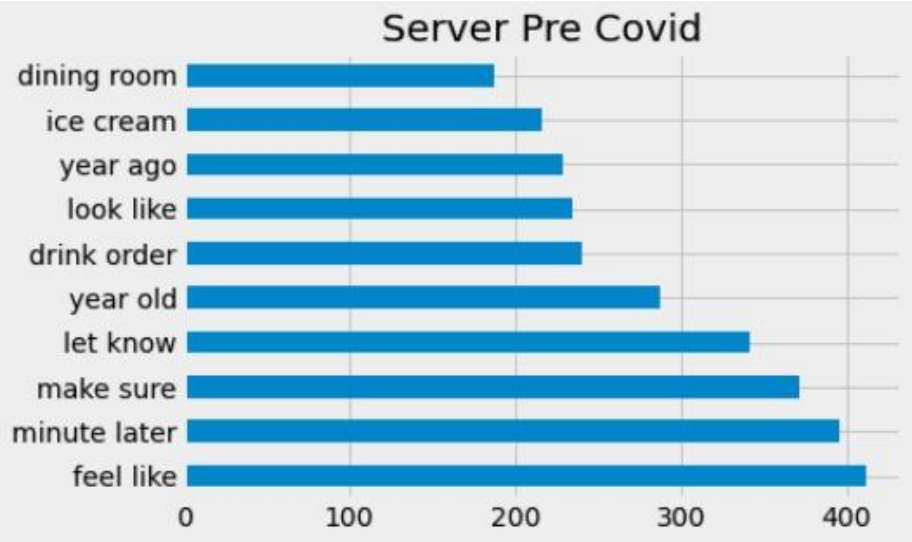
After Searching Parameters:

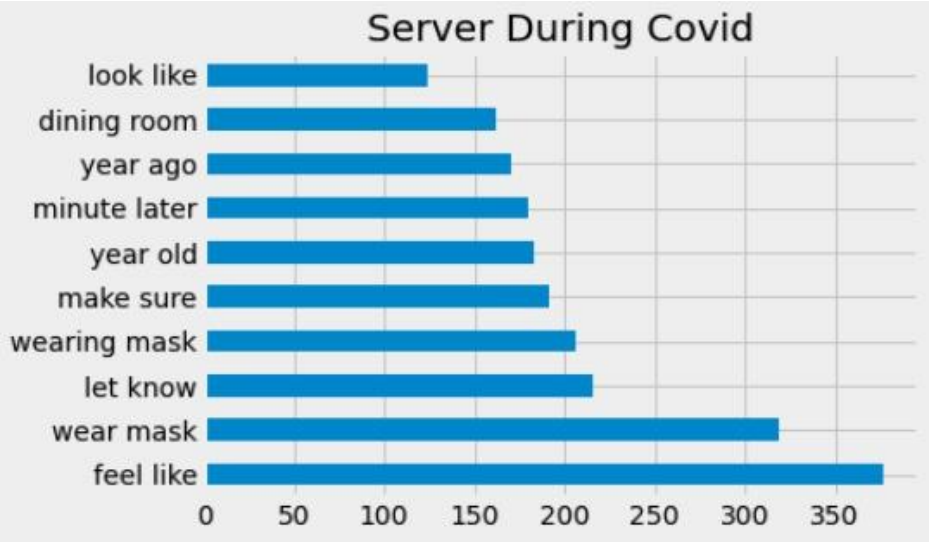
Logistic Regression	SVC	AdaBoostClassifier
- Accuracy on new data up 0.3% to 94.7%	- Too slow to search over - Any parameter tuning had bigger issues with overfitting	- Accuracy on new data up 0.5% to 94.8%

Not able to improve performance of any of the models substantially.

Impact of Covid on Results

/TalesFromYourServer:





Effect on Modeling

- The posts from today’s pandemic time were slightly more prone to overfitting

- Models trained on pre-covid data performed slightly worse on data from the past year

Can this be utilized elsewhere?

Questions!