Classifying NLP Data

Comparing Language Differences between Retail Workers and Servers Based on Reddit Comments

Terri John, Data Scientist

Problem Statement

During the past year, the service industry has seen an unprecedented rate of resignations. I wanted to develop a classification model that could predict if written language originates from a service worker in the restaurant industry or in the retail industry. I wanted to see which language was common to both industries, and which language differentiates the two, as well as the sentiment of that language. By examining these similarities and differences, businesses can better understand what steps they need to take in order to retain workers, as well as to attract new employees.

Background

Reasons cited for the uptick in resignations include:

- Stagnant wages.
- Fluctuations in working hours (made worse by the COVID-19 pandemic).
- Jobless benefits? Maybe not.

Source: https://www.nytimes.com/2021/10/14/opinion/workers-quitting-wages.html?searchResultPosition=3

Background

Source: https://fred.stlouisfed.org/series/JTSQUR

The Data

I used the Pushshift API to collect the most recent 40,000 comments posted to two subreddits:

TalesFromRetail: https://www.reddit.com/r/TalesFromRetail/

- Members 645,000
- Created November 9, 2011
- “A place to exchange stories about your daily experiences in brick & mortar retail.”

TalesFromYourServer: https://www.reddit.com/r/TalesFromYourServer/

- Members: 444,000
- Created September 24, 2012
- A subreddit where servers share stories and advice.

Source: https://github.com/pushshift/api

The Methodology

1. **Data Collection.**
   a. Used Pushshift API to collect comments from subreddits.
2. **Data Cleaning and Exploratory Data Analysis.**
   a. Deleted ‘non-entries’: those labeled ‘deleted’ or ‘removed’
   b. Deleted obvious ‘noise:’ Automoderator posts and ‘Happy Cake Day!’
   c. Deleted duplicates.
   d. Tokenized, lemmatized, removed stop words=english.
3. **Modeling.**
   a. Vectorized and tried out various models, including RandomForest, Logistic Regression, and MultinomialNB.
4. **Analyzed the modeling results.**
5. **Performed Sentiment Analysis.**

Source: https://github.com/pushshift/api

The Model: Logistic Regression

Baseline model

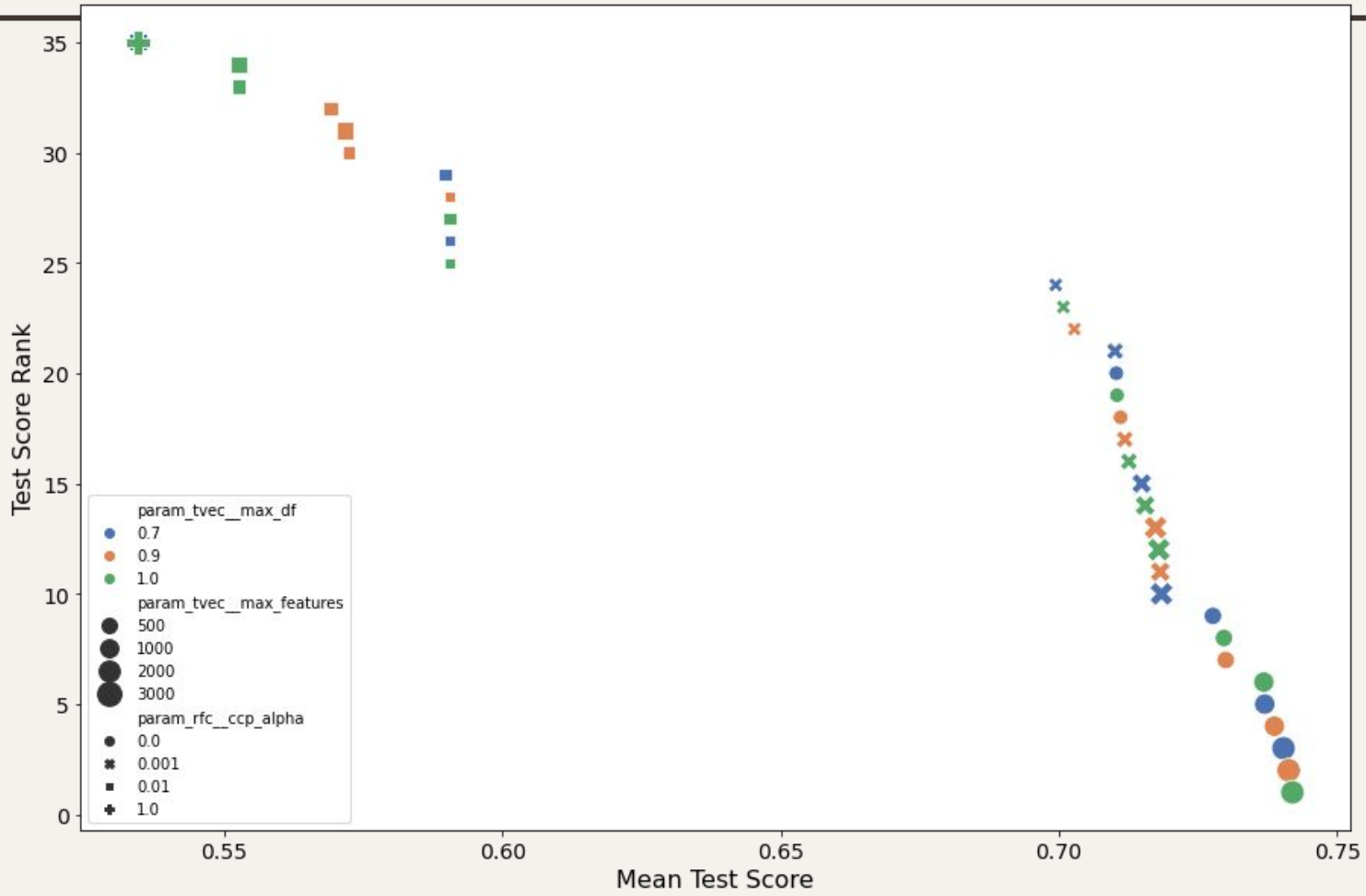
Accuracy score: **53%**
Based on the majority class.

RandomForest

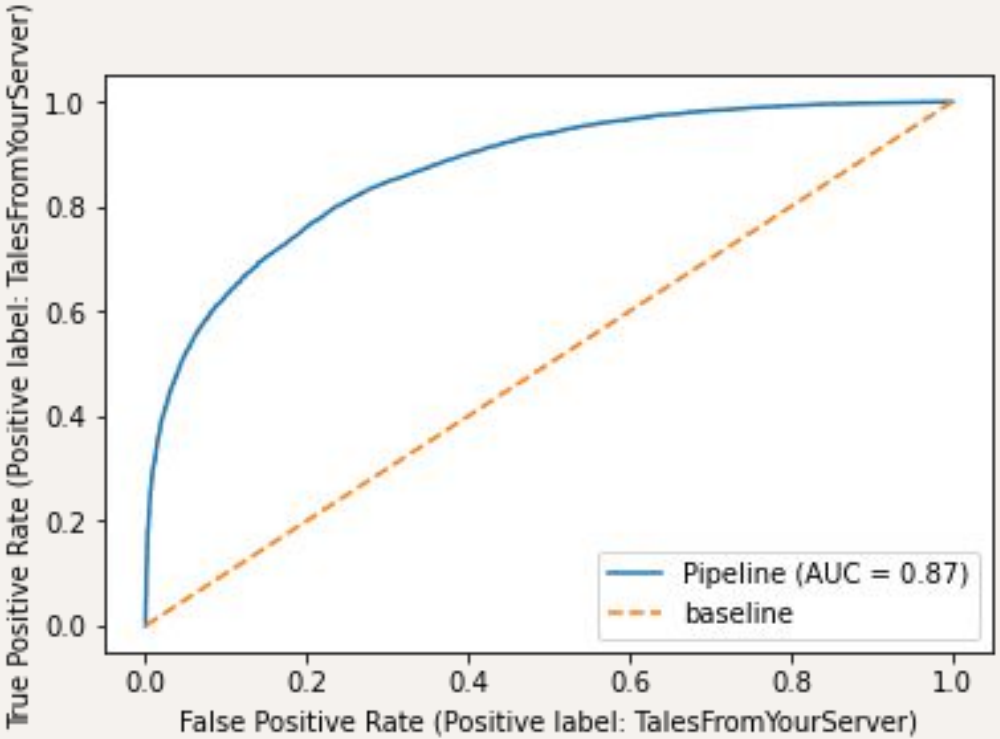
Training Accuracy Score: **.99**
Testing Accuracy Score: **.75**

LogisticRegression

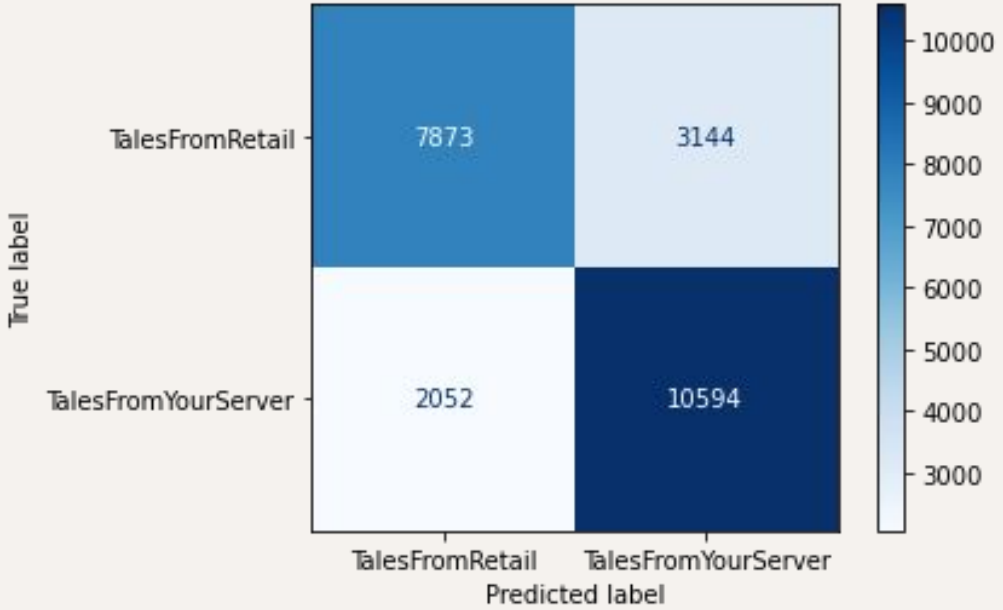
Training Accuracy Score: **.81**
Testing Accuracy Score: **.78**


Cross Validation Scores for Random Forest Gridline Search
Test Score Rank
Mean Test Score
0
5
10
15
20
25
30
35
0.55
0.60
0.65
0.70
0.75
param_tvec__max_df
0.7
0.9
1.0
param_tvec__max_features
500
1000
2000
3000
param_rfc__ccp_alpha
0.0
0.001
0.01
1.0

The Model: Logistic Regression

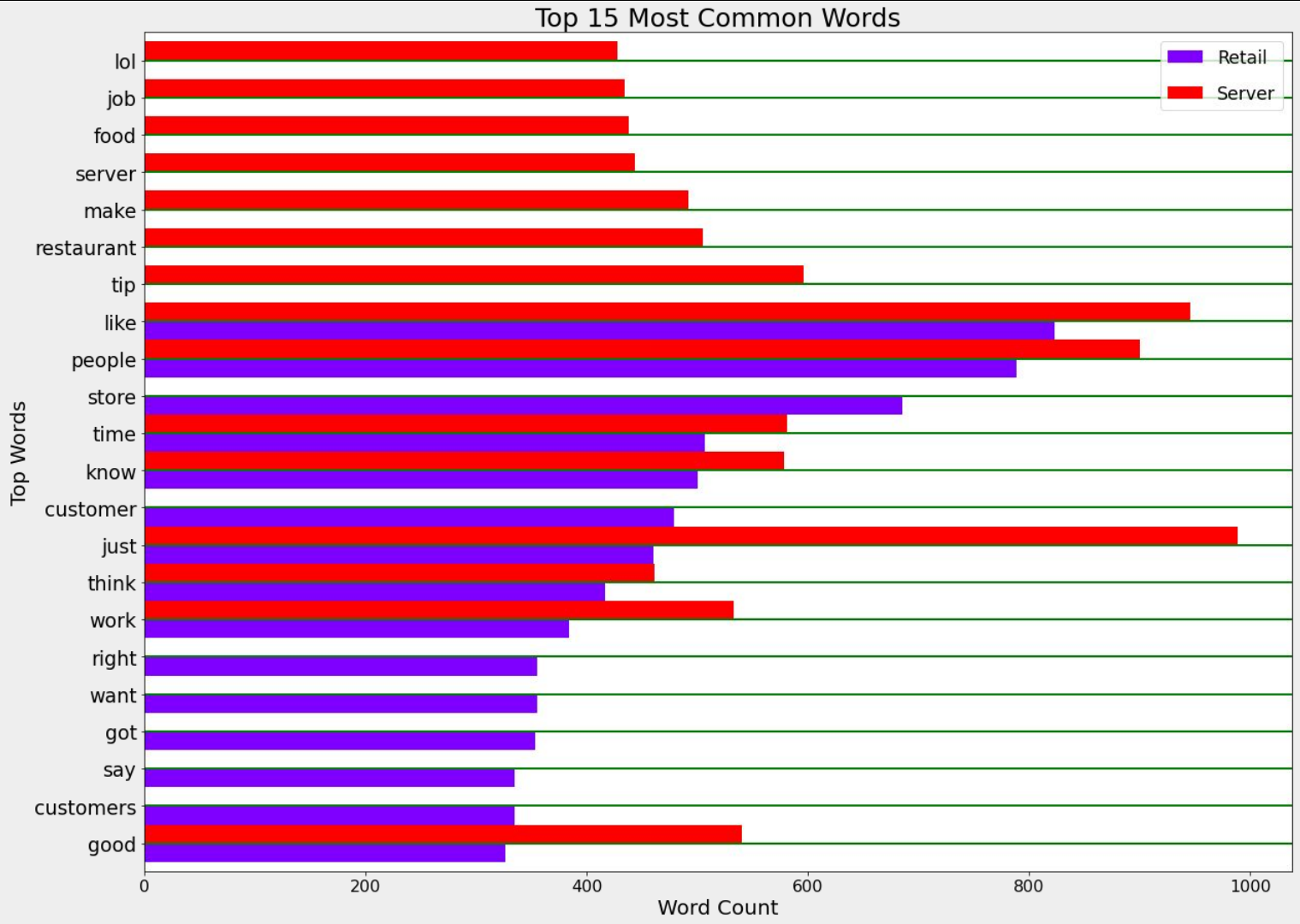


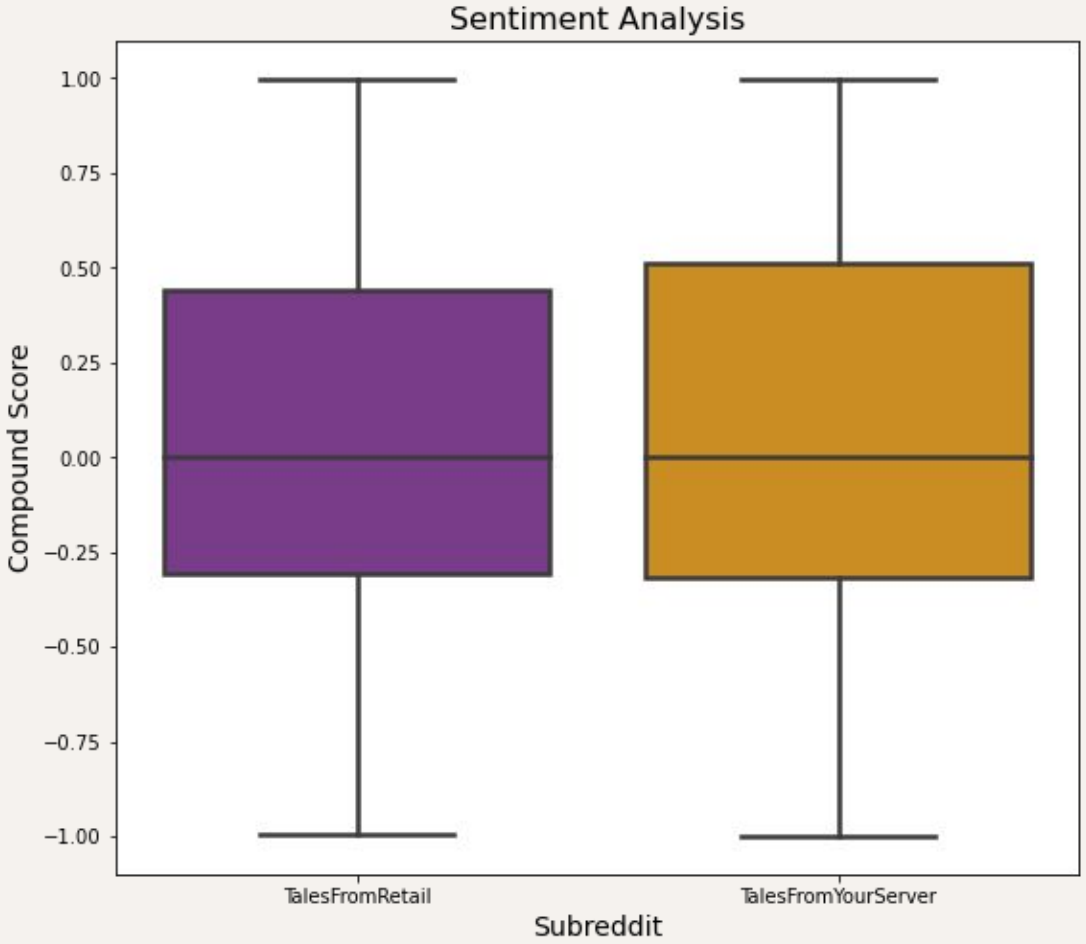
The Model: Logistic Regression

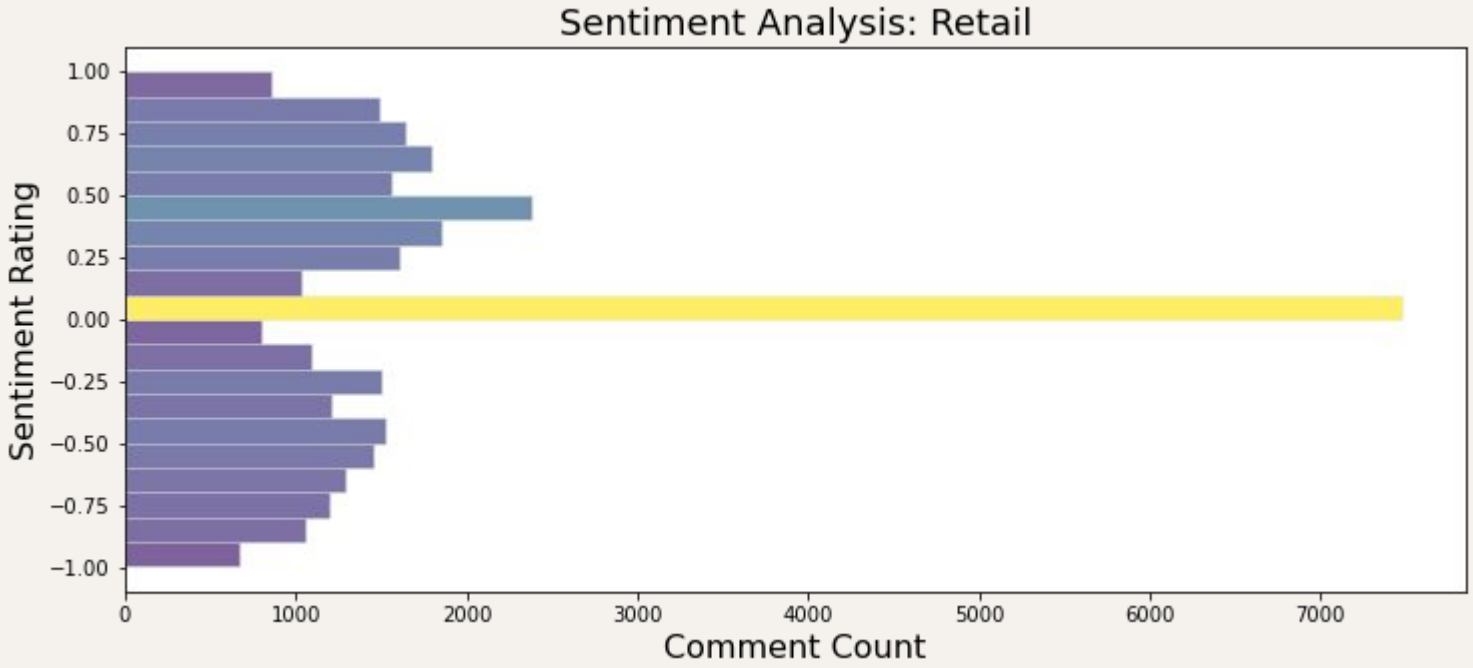


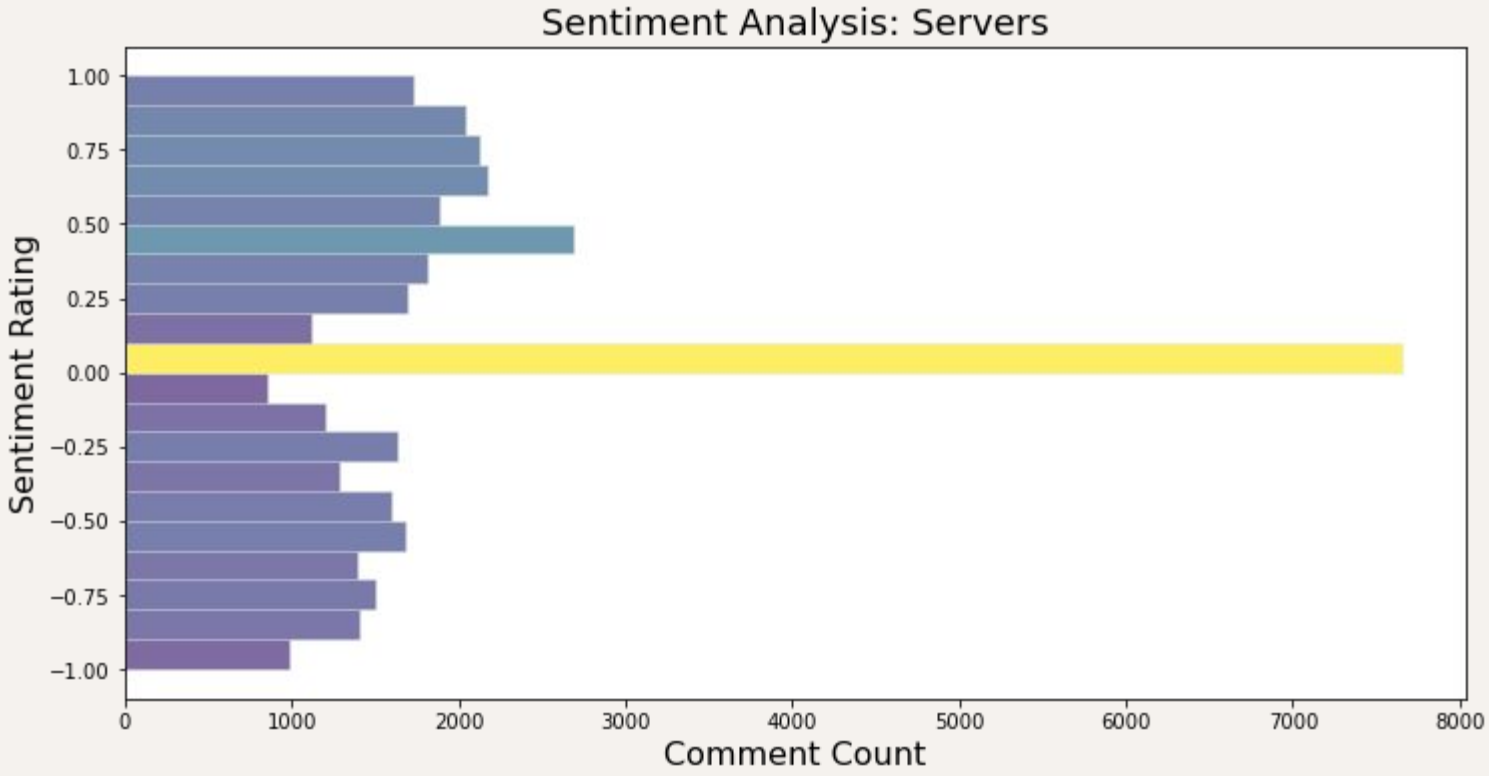
Subreddit	Precision	Recall	F1
TalesFrom Retail	.79	.71	.75
TalesFrom YourServer	.77	.84	.80

Training Accuracy Score: **.81**
Testing Accuracy Score: **.78**


Top 15 Most Common Words
Retail
Server
lol
job
food
server
make
restaurant
tip
like
people
store
time
know
customer
just
think
work
right
want
got
say
customers
good
Top Words
0
200
400
600
800
1000
Word Count


Sentiment Analysis
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Compound Score
TalesFromRetail
TalesFromYourServer
Subreddit


Sentiment Analysis: Retail
Sentiment Rating
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
1000
2000
3000
4000
5000
6000
7000
Comment Count


Sentiment Analysis: Servers
Sentiment Rating
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
1000
2000
3000
4000
5000
6000
7000
8000
Comment Count

Conclusion

- The model is able to differentiate between the TalesFromYourServer and the TalesFromRetail subreddits with about 80% accuracy.
- Generally, the sentiment from both groups is fairly neutral.
- Based on top word usage, it would appear that things like tips, food, and tables are important to servers, while things like customers are important to both servers and retail workers.

Recommendation for Next Steps

- In order to better understand how to best serve workers in order to retain them, I would recommend conducting a survey.
- Try more models!

Resources

- Pushift API: https://github.com/pushshift/api
- The New York Times: https://www.nytimes.com/2021/10/14/opinion/workers-quitting-wages.html?searchResultPosition=3
- TalesFromRetail: https://www.reddit.com/r/TalesFromRetail/
- TalesFromYourServer: https://www.reddit.com/r/TalesFromYourServer/
-
-

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**,and infographics & images by **Freepik**

Please keep this slide for attribution