r/Seattle or r/SeattleWA ?

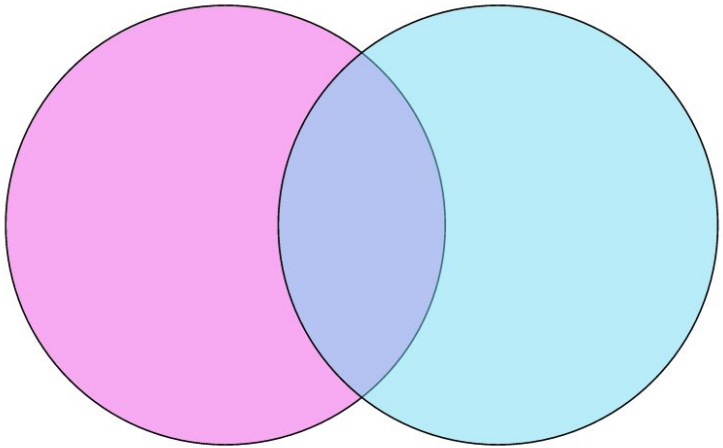
John Guo

What is reddit?

- Reddit is a forum website
- Partitioned into many communities called subreddits
- space for people to share things related to the specific subreddit

But these sound the same...

- Many subreddits overlap
- still have their own differences that distinguish them from one another
- r/Seattle and r/SeattleWA



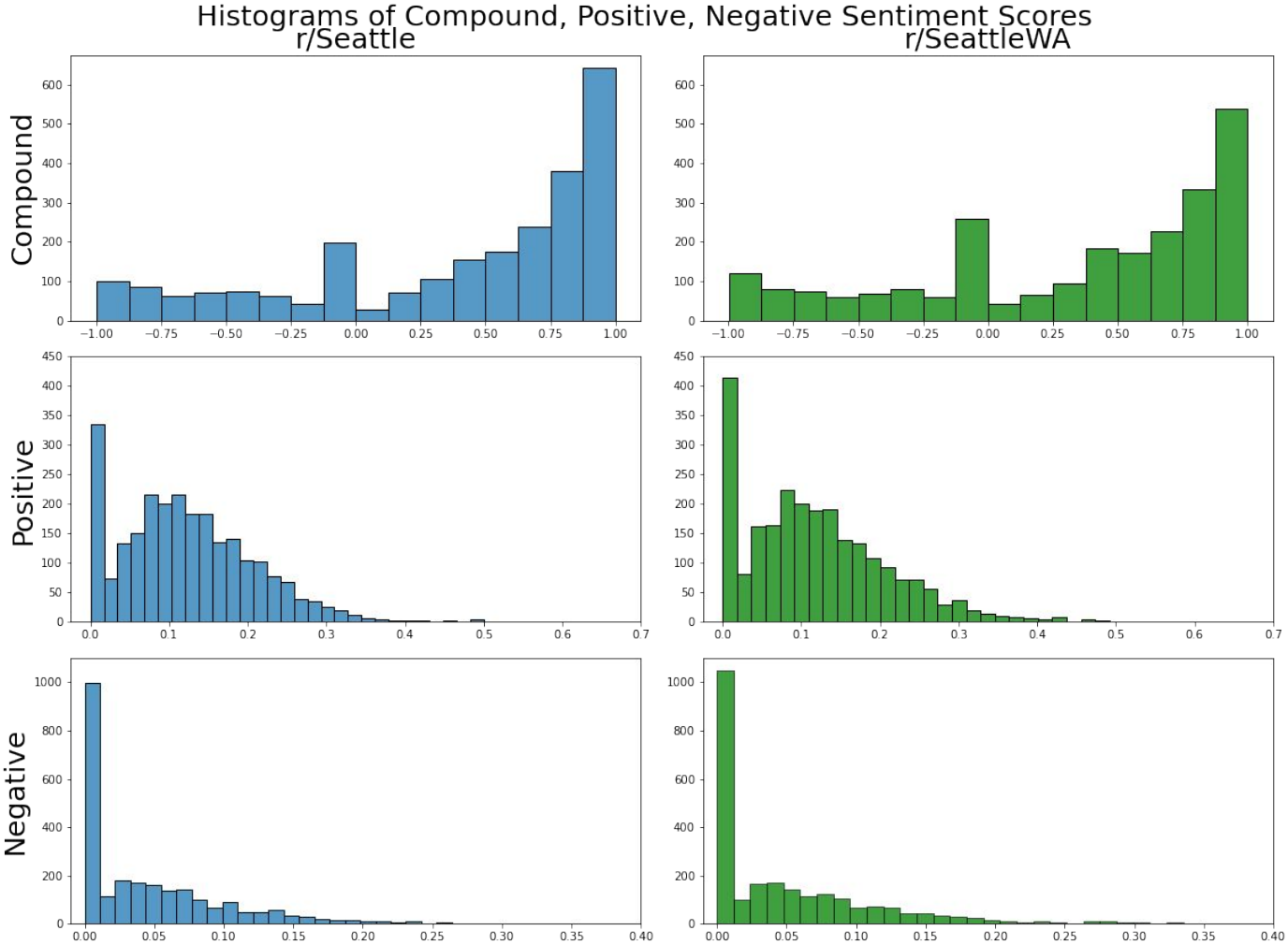
r/Seattle vs r/SeattleWA

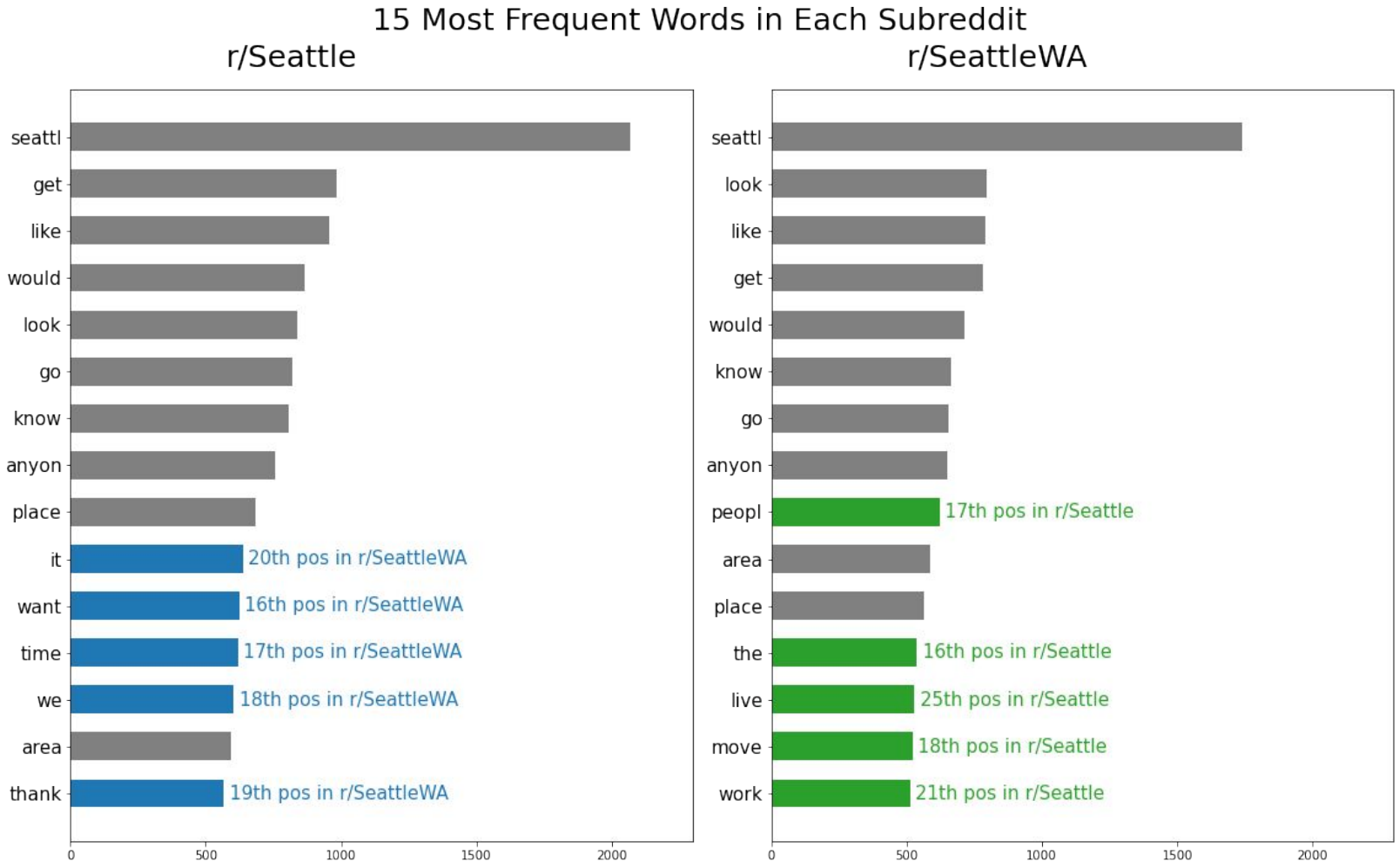
In this project we:

- explore text data from posts on r/Seattle and r/SeattleWA
- try to identify notable ways in how these two subreddits differ
- and build classification models to try and predict if a post is more likely to come from r/Seattle or r/SeattleWA

r/Seattle vs r/SeattleWA

- r/Seattle and r/SeattleWA likely to have many similar users and posts
- there are certainly people who participate in both subreddits
- consider our model successful if it can correctly classify which subreddit a post belongs to more than 60% of the time.


Histograms of Compound, Positive, Negative Sentiment Scores
r/Seattle
r/SeattleWA
Compound
Positive
Negative


15 Most Frequent Words in Each Subreddit
r/Seattle
r/SeattleWA
seattl
get
like
would
look
go
know
anyon
place
it
want
time
we
area
thank
20th pos in r/SeattleWA
16th pos in r/SeattleWA
17th pos in r/SeattleWA
18th pos in r/SeattleWA
19th pos in r/SeattleWA
0
500
1000
1500
2000
seattl
look
like
get
would
know
go
anyon
peopl
area
place
the
live
move
work
17th pos in r/Seattle
16th pos in r/Seattle
25th pos in r/Seattle
18th pos in r/Seattle
21th pos in r/Seattle
0
500
1000
1500
2000

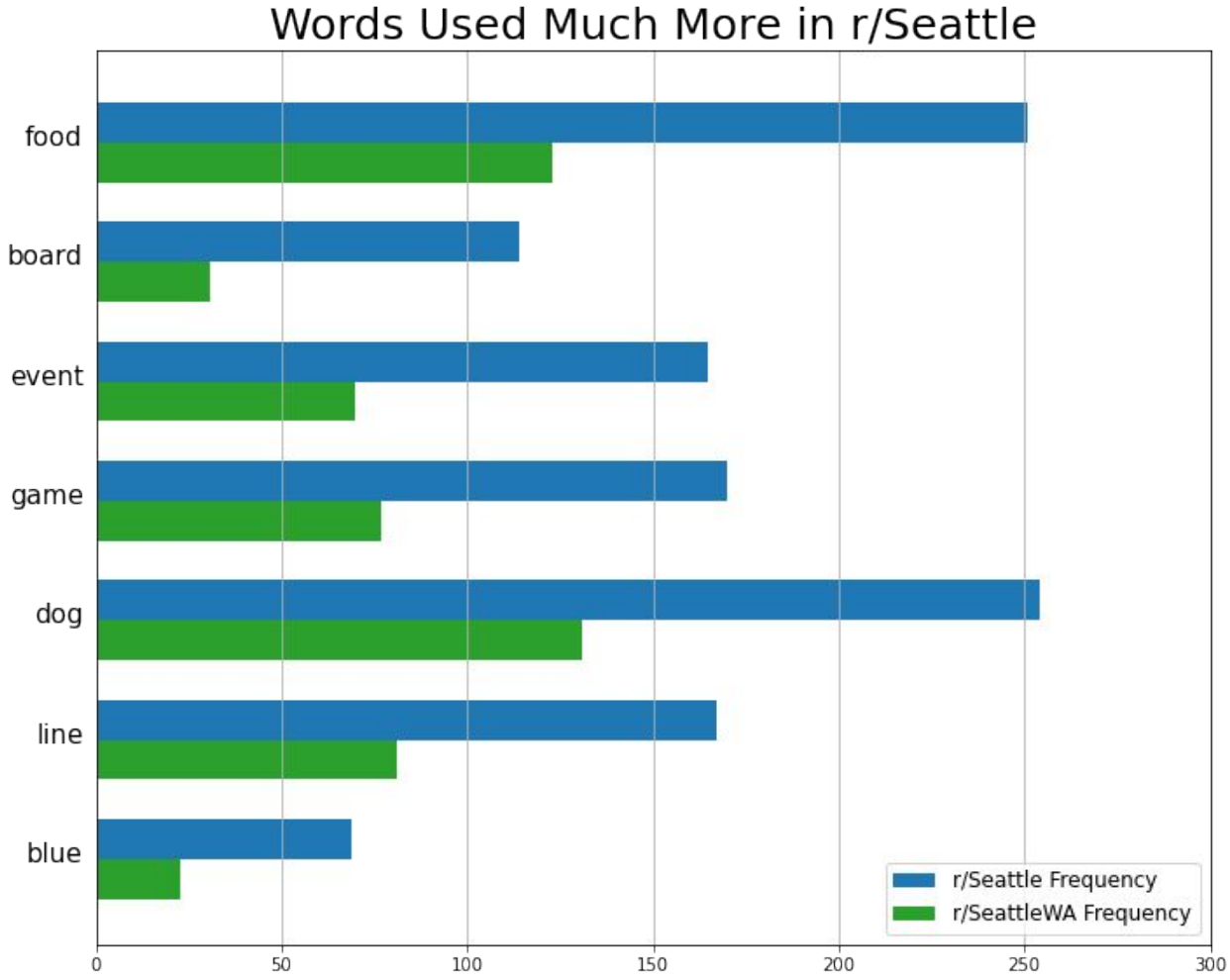
Why is word count not enough?

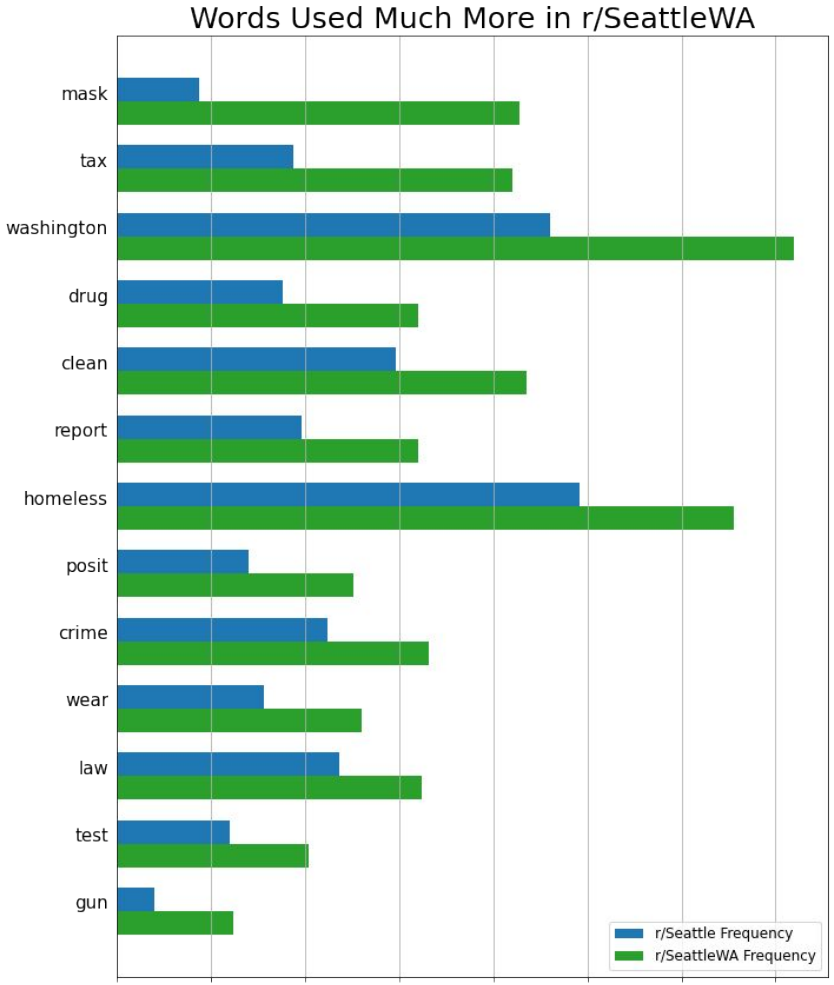
- Seattl was the most common word in both subreddits
- But Seattl appeared roughly 300 more times in r/Seattle than r/SeattleWA
- Just using word count would say this word is important
- In reality, it is the least important

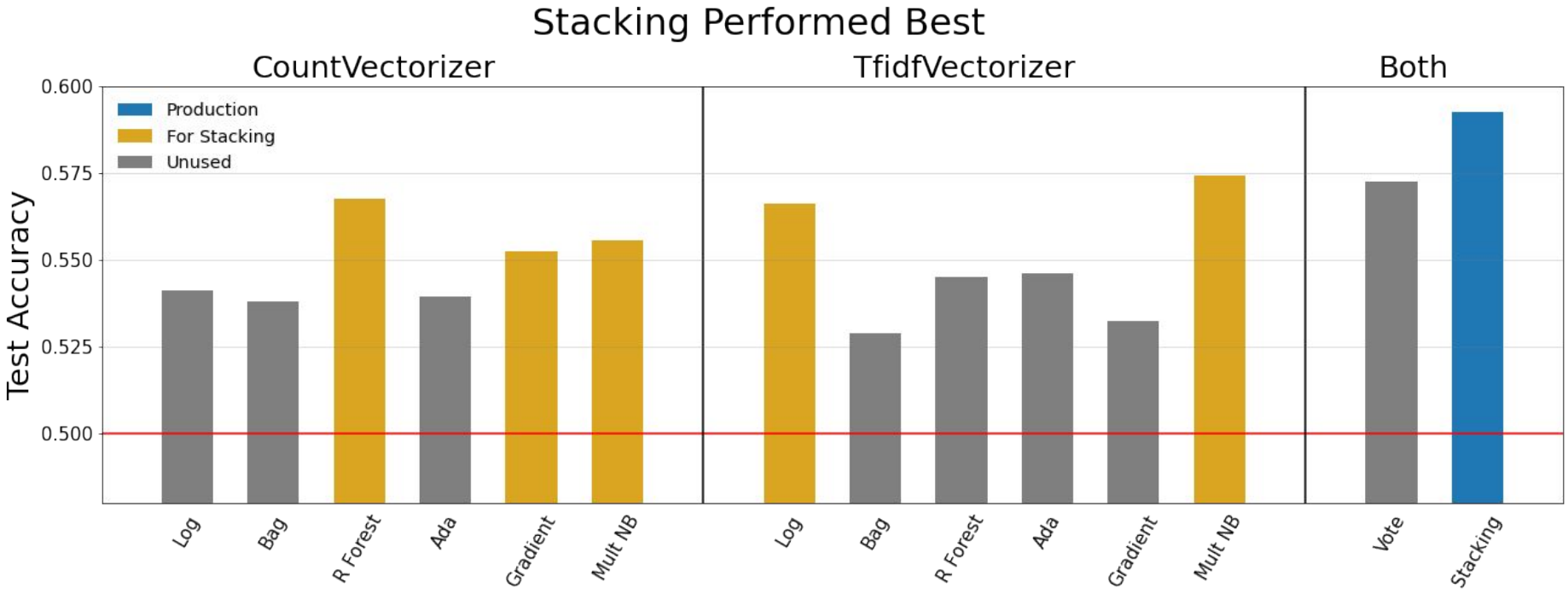
Ranking is Our Solution!

For each subreddit

- Rank each word within the subreddit
- Subtract word's rank in one subreddit from rank in the order
- Use difference in rankings


Words Used Much More in r/Seattle
food
board
event
game
dog
line
blue
0
50
100
150
200
250
300
r/Seattle Frequency
r/SeattleWA Frequency


Words Used Much More in r/SeattleWA
mask
tax
washington
drug
clean
report
homeless
posit
crime
wear
law
test
gun
0
25
50
75
100
125
150
175
r/Seattle Frequency
r/SeattleWA Frequency


Stacking Performed Best
CountVectorizer
TfidfVectorizer
Both
Test Accuracy
0.600
0.575
0.550
0.525
0.500
Production
For Stacking
Unused
Log
Bag
R Forest
Ada
Gradient
Mult NB
Log
Bag
R Forest
Ada
Gradient
Mult NB
Vote
Stacking

Conclusions

- Users on r/SeattleWA are likely more Conservative than the users on r/Seattle
  - mask, tax, drug, clean, homeless, report, crime, law, wear, gun

- Words that appeared more in r/Seattle were generic common
  - board, food, game, event, dog, line, blue

- Our production model slightly missed the target with an accuracy score of 59%

Future Work

Explore

- Words not in common between the two subreddits
- Bi-grams and tri-grams
- Beyond the bag-of-words approach