r/interestingproject

Nate Cox | Data Scientist

Can you predict what subreddit a comment comes from?

Outline

- Info about the Data
- Subreddits
- Cleaning the Data
- Feature Engineering
- EDA
- Modeling
- Conclusion

Info About the Data

- Collected data using the Pushshift Reddit API
- 115,000 comments
- Text, author, and the subreddit

Subreddits

mildlyinteresting

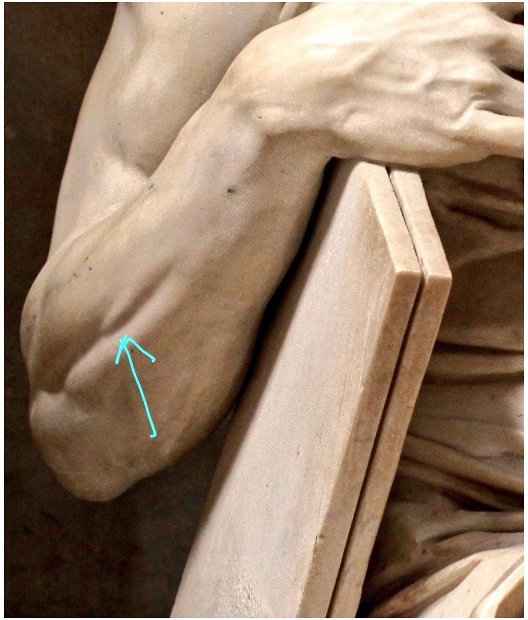
interestingasfuck

Mildlyinteresting

Source: https://www.reddit.com/r/mildlyinteresting/comments/9ykoe4/the_corrosion_on_this_water_tap_looks_like_a_map/

Source:
https://www.reddit.com/r/mildlyinteresting/comments/83ichd/found_the_cliff_this_clif_bar_came_from
/

Interestingasfuck

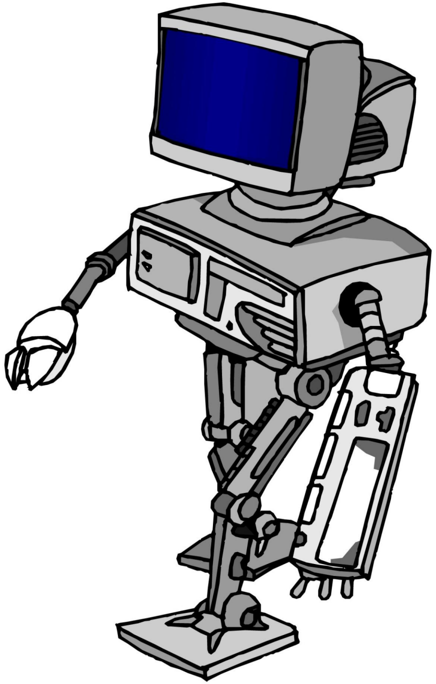


Source:
https://www.reddit.com/r/interestingasfuck/comments/k6ej0c/the_small_details_in_the_forearms_there_is_one

Source:
https://www.reddit.com/r/interestingasfuck/comments/n39zq5/i_created_a_photorealistic_image_of_george/

Cleaning the Data

- Deleted comments
- Bots
- URL’s
- Extra white space
- Contractions
- And much, much, more...



Feature Engineering

Number of characters

Number of words

Number of capital letters

Is there a question mark?

Is there an exclamation point?

Are there trailing sentences...

Is there text in quotes?

Is there text in italics?

Is there text in bold?

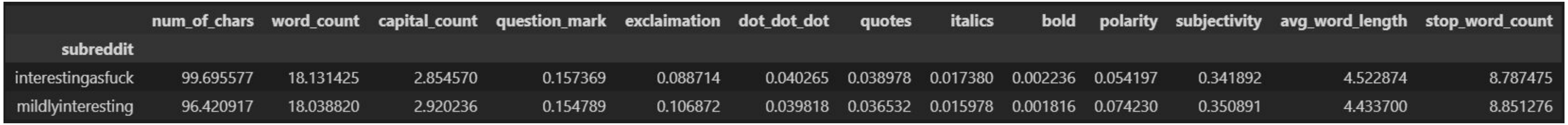
What is the polarity of the comment?

What is the subjectivity of the comment?

Average word length

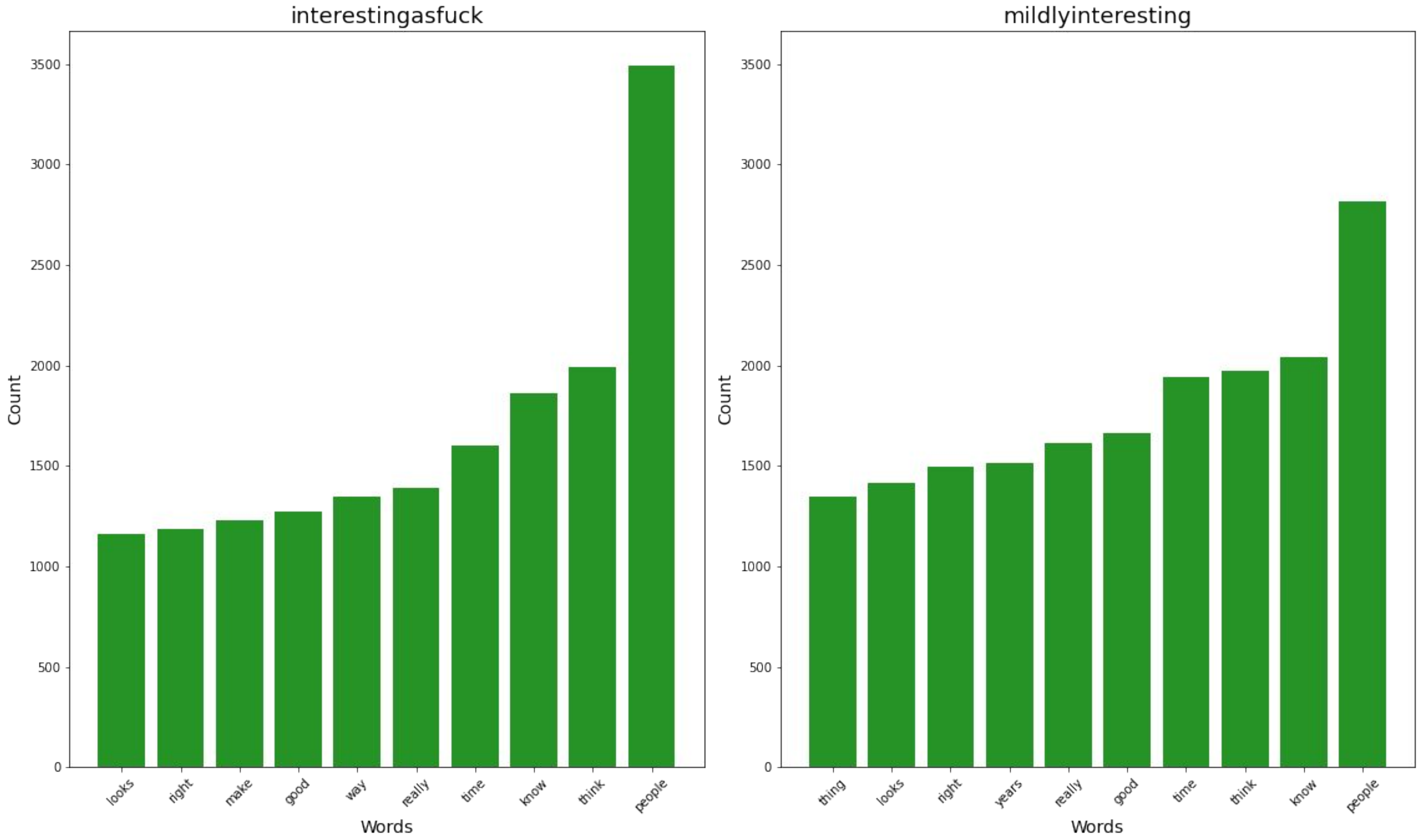
Number of stop words

Averages of Engineered Features



	num_of_chars	word_count	capital_count	question_mark	exclaimation	dot_dot_dot	quotes	italics	bold	polarity	subjectivity	avg_word_length	stop_word_count
subreddit													
interestingasfuck	99.695577	18.131425	2.854570	0.157369	0.088714	0.040265	0.038978	0.017380	0.002236	0.054197	0.341892	4.522874	8.787475
mildlyinteresting	96.420917	18.038820	2.920236	0.154789	0.106872	0.039818	0.036532	0.015978	0.001816	0.074230	0.350891	4.433700	8.851276

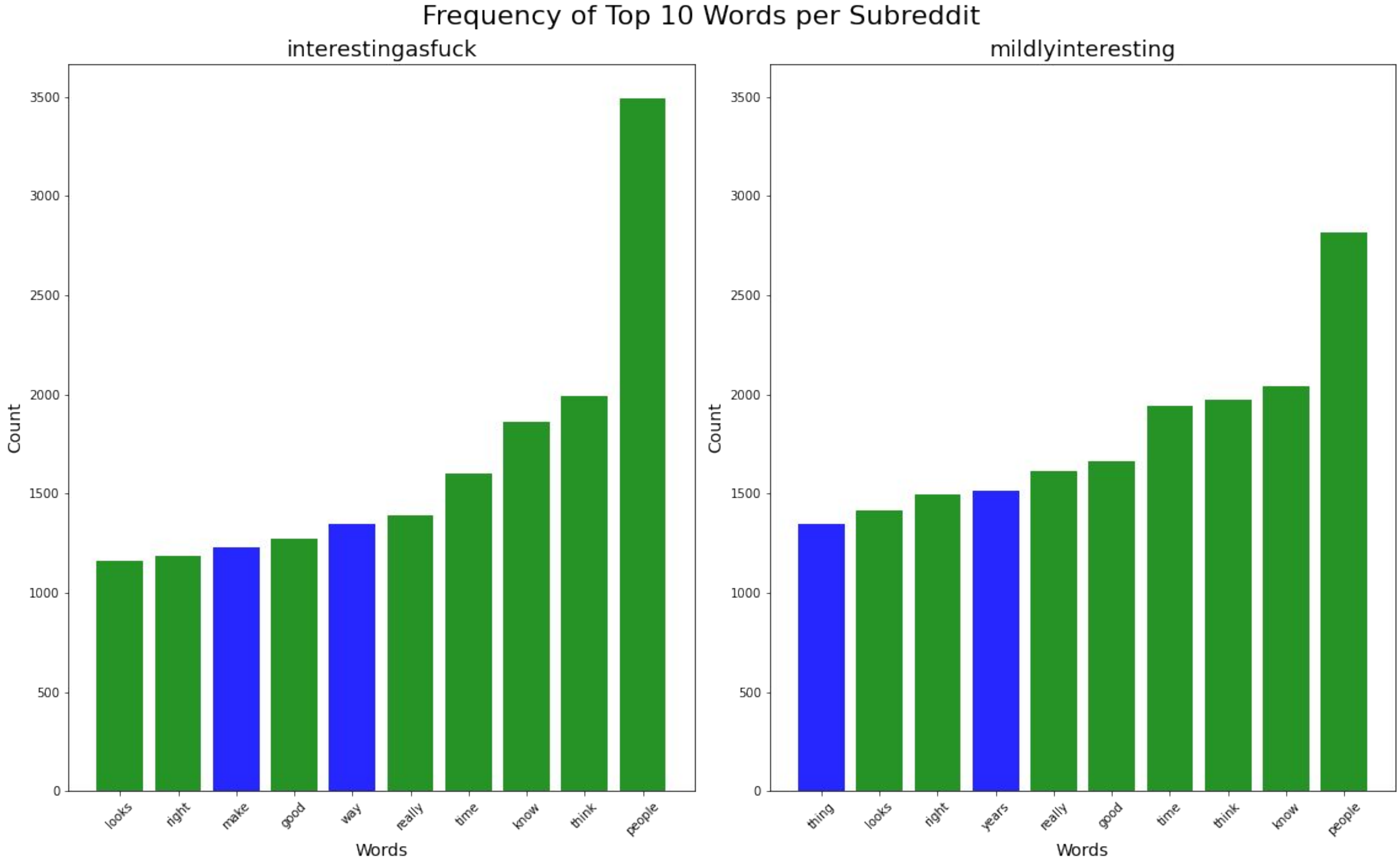
Summary: The averages for all features in both subreddits are very close together.



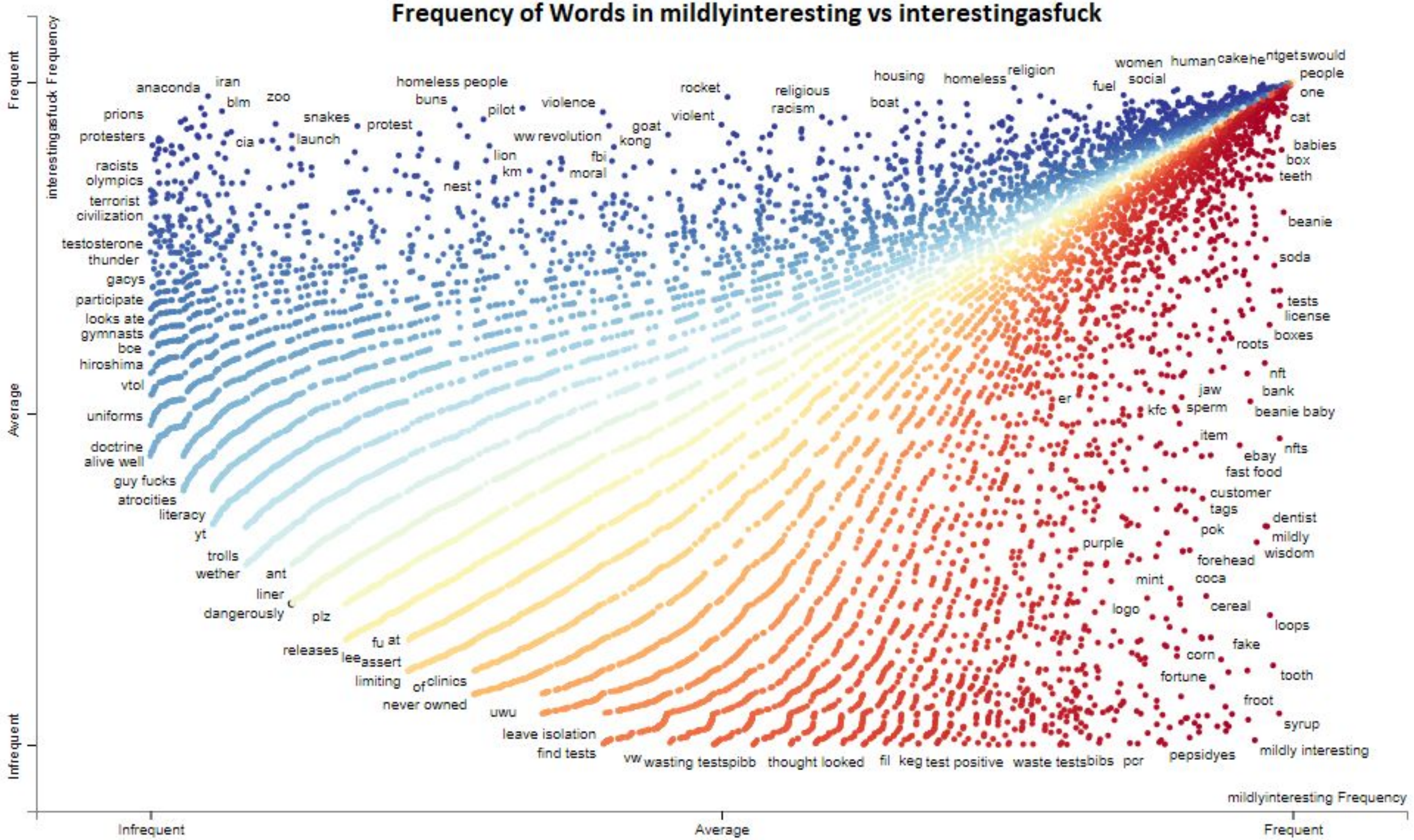
Frequency of Top 10 Words per Subreddit

interestingasfuck

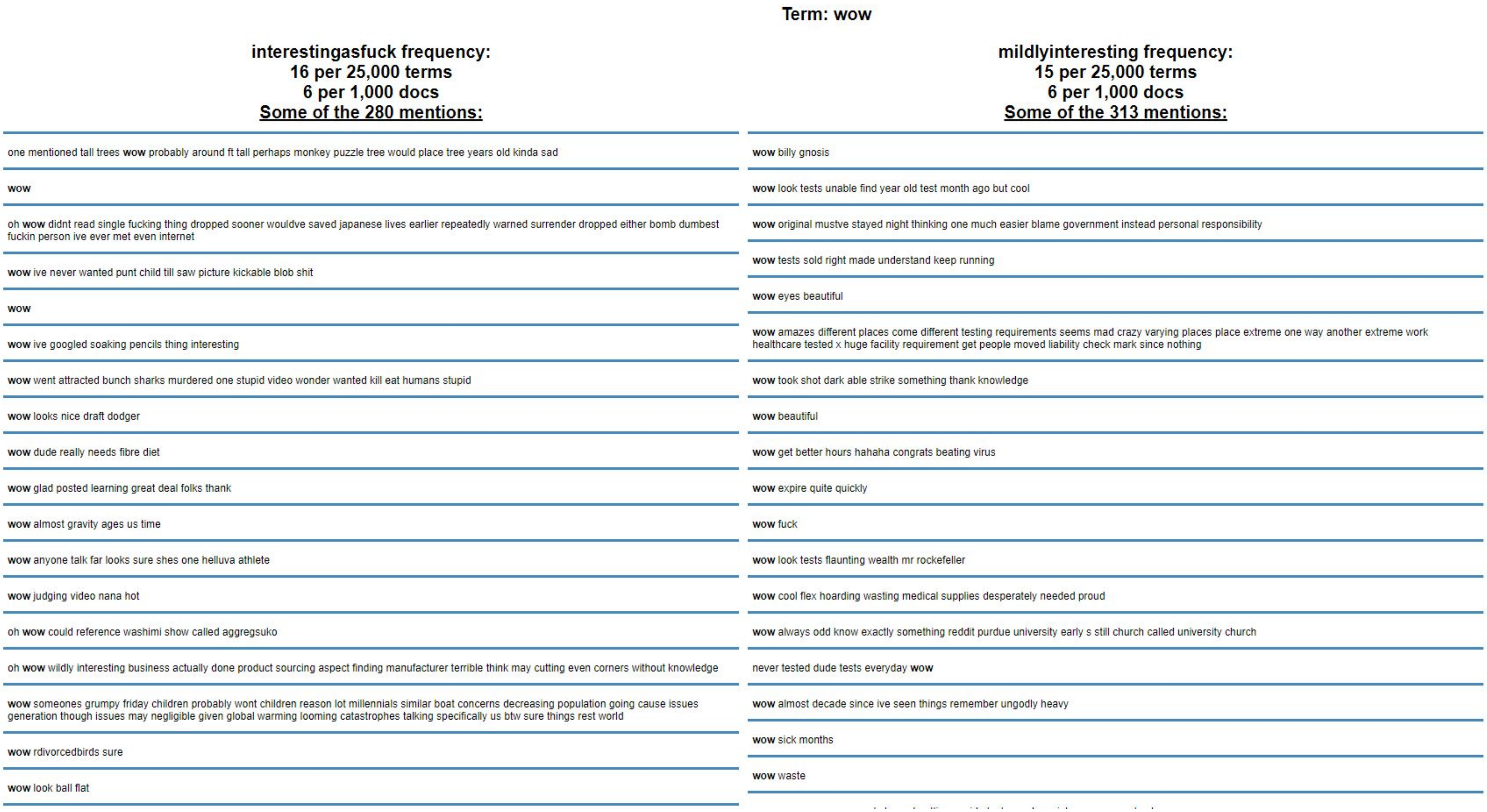
mildlyinteresting


Frequency of Top 10 Words per Subreddit
interestingasfuck
mildlyinteresting
Count
Count
3500
3000
2500
2000
1500
1000
500
0
looks
right
make
good
way
really
time
know
think
people
thing
looks
right
years
really
good
time
think
know
people
Words
Words

Frequency of Words in mildlyinteresting vs interestingasfuck



interestingasfuck document count: 47,411; word count: 434,605
mildlyinteresting document count: 55,075; word count: 494,438



Term: **wow**

interestingasfuck frequency:
16 per 25,000 terms
6 per 1,000 docs
Some of the 280 mentions:

one mentioned tall trees **wow** probably around ft tall perhaps monkey puzzle tree would place tree years old kinda sad

wow

oh **wow** didnt read single fucking thing dropped sooner wouldve saved japanese lives earlier repeatedly warned surrender dropped either bomb dumbest fuckin person ive ever met even internet

wow ive never wanted punt child till saw picture kickable blob shit

wow

wow ive googled soaking pencils thing interesting

wow went attracted bunch sharks murdered one stupid video wonder wanted kill eat humans stupid

wow looks nice draft dodger

wow dude really needs fibre diet

wow glad posted learning great deal folks thank

wow almost gravity ages us time

wow anyone talk far looks sure shes one helluva athlete

wow judging video nana hot

oh **wow** could reference washimi show called aggregsuko

oh **wow** wildly interesting business actually done product sourcing aspect finding manufacturer terrible think may cutting even corners without knowledge

wow someones grumpy friday children probably wont children reason lot millennials similar boat concerns decreasing population going cause issues generation though issues may negligible given global warming looming catastrophes talking specifically us btw sure things rest world

wow rdivorcedbirds sure

wow look ball flat

mildlyinteresting frequency:
15 per 25,000 terms
6 per 1,000 docs
Some of the 313 mentions:

wow billy gnosis

wow look tests unable find year old test month ago but cool

wow original mustve stayed night thinking one much easier blame government instead personal responsibility

wow tests sold right made understand keep running

wow eyes beautiful

wow amazes different places come different testing requirements seems mad crazy varying places place extreme one way another extreme work healthcare tested x huge facility requirement get people moved liability check mark since nothing

wow took shot dark able strike something thank knowledge

wow beautiful

wow get better hours hahaha congrats beating virus

wow expire quite quickly

wow fuck

wow look tests flaunting wealth mr rockefeller

wow cool flex hoarding wasting medical supplies desperately needed proud

wow always odd know exactly something reddit purdue university early s still church called university church

never tested dude tests everyday **wow**

wow almost decade since ive seen things remember ungodly heavy

wow sick months

wow waste

Models

- Tested six models
- Cut down to three
  - Multinomial Naive Bayes
  - Logistic Regression
  - Voting Classifier (Logreg + NB)
- Only used the text of the comments

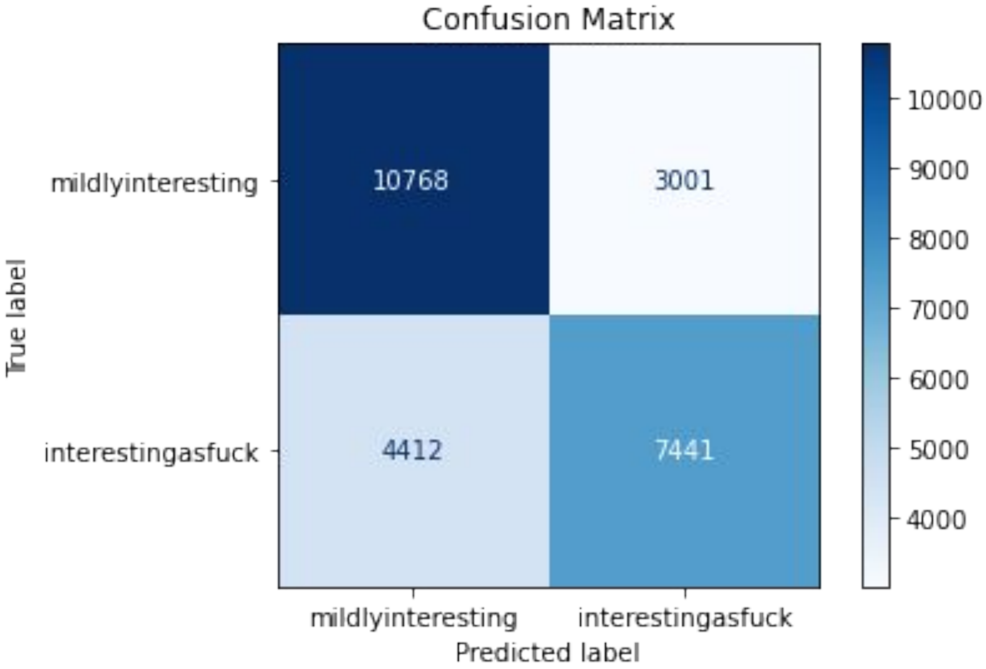
Results

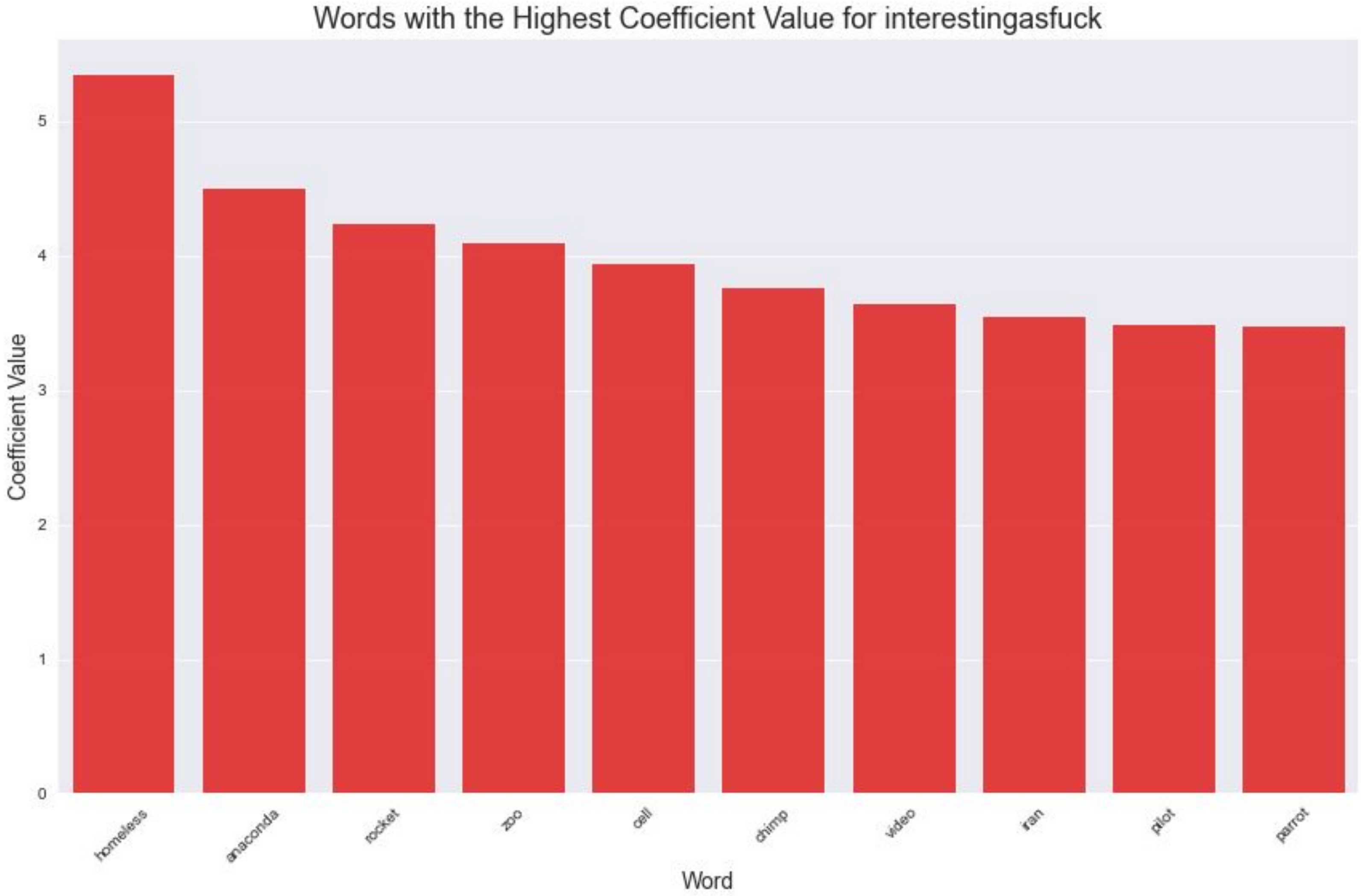
Model	Accuracy
Baseline	53.7%
Multinomial Naive Bayes	70.5%
Logistic Regression	70.8%
Voting Classifier (Logreg + NB)	71.3%

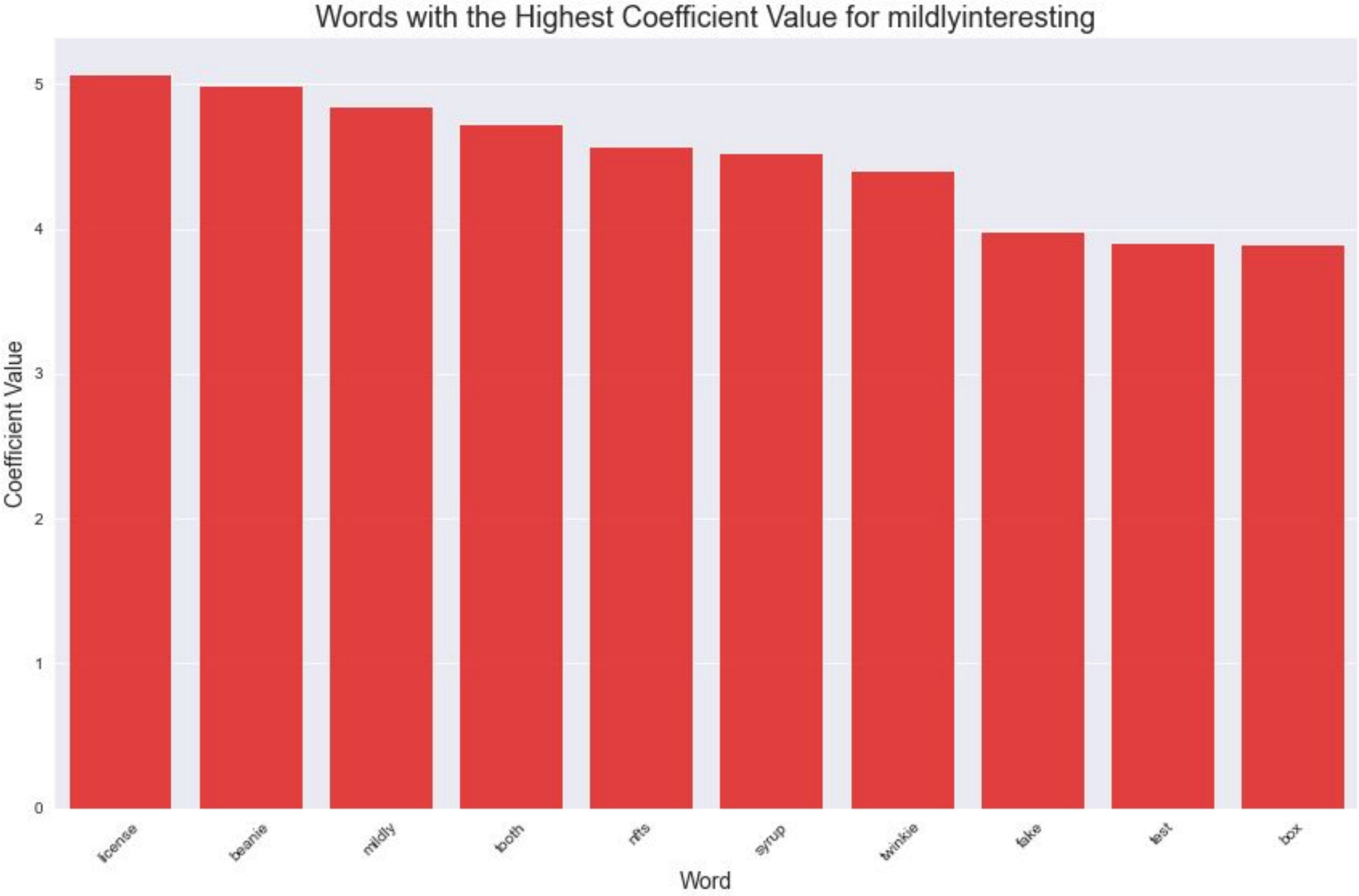
Breaking it down

Correctly predicted mildlyinteresting = 10,768

Correctly predicted interestingasfuck = 7,441




Words with the Highest Coefficient Value for interestingasfuck
Coefficient Value
0
1
2
3
4
5
homeless
anaconda
rocket
zoo
cell
chimp
video
iran
pilot
parrot
Word


Words with the Highest Coefficient Value for mildlyinteresting
Coefficient Value
0
1
2
3
4
5
license
beanie
mildly
tooth
nfts
syrup
twinkie
fake
test
box
Word

Conclusion

- Can you predict what subreddit a comment comes from?
  - **Yes.**
- The model performed **17.6%** better than the baseline
  - **Baseline: 53.7%**
  - **Voting Classifier: 71.3%**

Questions?