Subreddit Classification

By: Royce Phillips Jr

Problem Statement

Reddit is a large platform with millions of submissions and hundreds if not thousands of subreddits. Each subreddit has its own topics and personality.

Is it possible to create a classification model that can decipher between two separate subreddits.

Which subreddits?

- 'CasualConversation'
- 'SeriousConversation'.

Why?

- Sister subreddits?
- Any similar topics?
- Are there any major differences

Image: https://www.flickr.com/photos/djandywdotcom/31437486496

What data??

Used Pushshift's API to pull data

Image: https://freesvg.org/shrug-gesture

Initial cleaning and EDA

- Many null values
- Many submissions had been ‘removed’
  - Mostly in ‘CasualConversation’
- Unbalanced subreddits

Initial thoughts

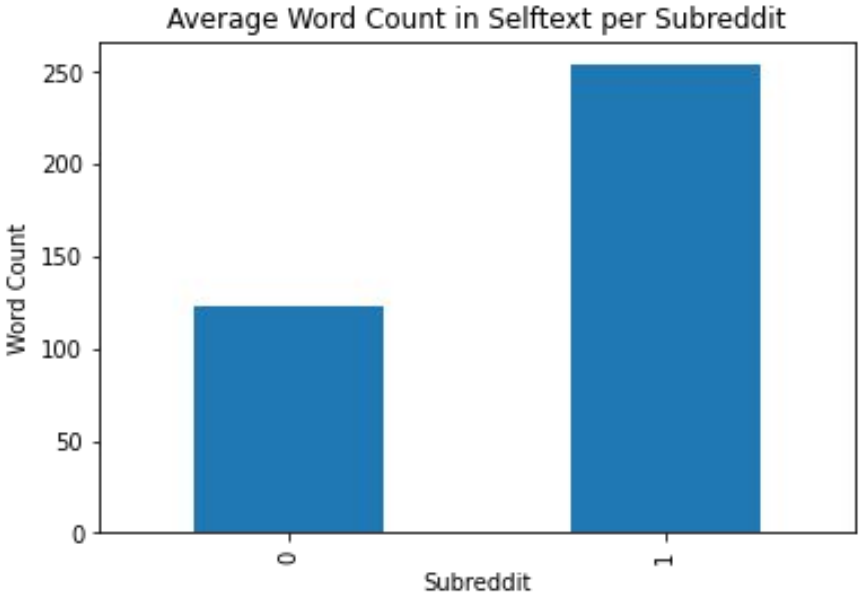
- ‘SeriousConversation’ more quality?
- ‘CasualConversation’ too many people posting?

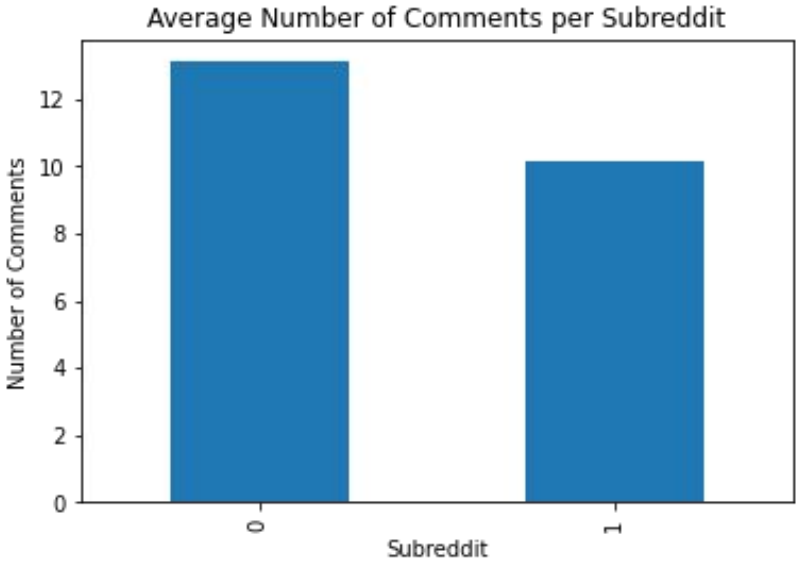
Some quick facts

	Number of Comments	Score	Title Word Count	SelfText Word Count
Casual Conversation	13	14	10	123
Serious Conversation	10	9	12	254

Comparing Subreddits

0 = ‘CasualConversation’
1 = ‘SeriousConversation’





Which gets more interaction?

	More than 100 Comments	More than 250 Comments
CasualConversation	135 ~ 76.3%	42 ~ 23.7%
SeriousConversation	61 ~ 98.4%	1 ~ 1.6

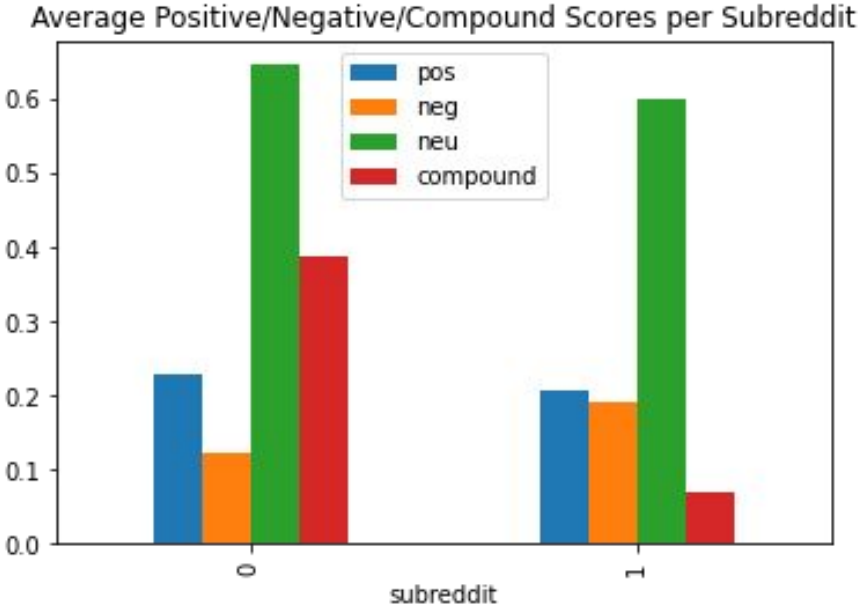
Most Commented Post

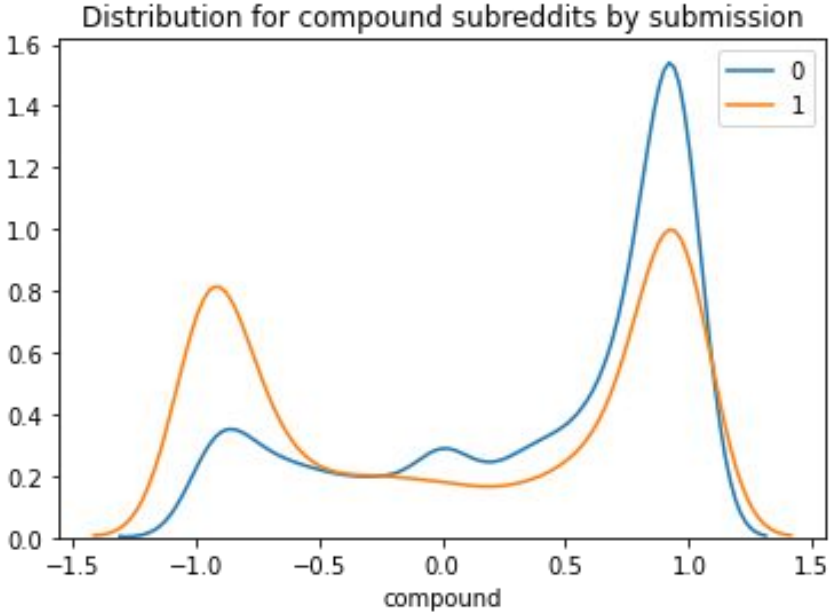
“CasualConversation”

- 3001 Comments

'This coronavirus things has made me realize people would be a lot happier and explore their passions and interests if they didn’t have to work so much.'

Sentiment Analysis





How about models?

- Gridsearch
- Pipelines
- Linear Regression
- Naive Bayes Multinomial

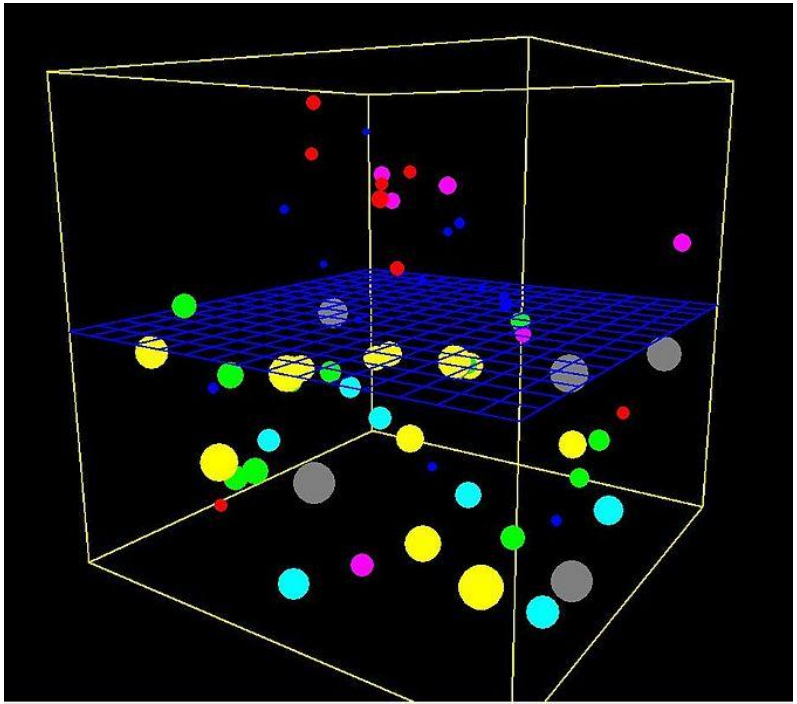


Image: https://commons.wikimedia.org/wiki/File:Osmosis_computer_simulation.jpg

Top 20 words

SeriousConversation

	0
religious	2.371477
sexual	2.197139
death	1.962149
died	1.853095
sexually	1.686206
politics	1.667395
trauma	1.660829
cancer	1.633291
drugs	1.562543
dying	1.441792
political	1.379804
sex	1.366498
dead	1.270756
religion	1.240917
discuss	1.192002
much time	1.135842
climate	1.121131
illness	1.077652
passed away	1.058376
porn	1.032302

CasualConversation

	0
quarantine	-2.026719
wanted share	-1.709266
corona	-1.599827
isolation	-1.434727
lockdown	-1.287376
caring	-1.267222
cake	-1.224584
stay home	-1.188175
social distancing	-1.171571
bought	-1.133384
haha	-1.109965
applied	-1.078960
tldr	-1.061328
series	-1.052097
working home	-1.051667
chatting	-1.022924
coronavirus	-0.991696
virus	-0.977520
holding	-0.971848
covid	-0.961733
thankful	-0.950105

Ideas to improve model

- Gather more data
- Try different models
  - SVM, KNN, etc.
- See how TFIDF could work

Image:
https://commons.wikimedia.org/wiki/File:Performance-Evaluation-Process-z.jpg

Conclusion

- 80% success rate.
- Amount of words in selftext
- Sentiment analysis
- Number of comments
- Topics
    - 'SeriousConversation' seemed to be more personal.
    - 'CasualConversations' were more revolved around the topic of COVID-19.

2nd Most commented : 1814 Comments

‘I just ate rich people pasta for the first time'

"I'm about to graduate and get my first full time job in a couple of months. Meanwhile, I'm still living the poor plebejian life. I always get the cheapest 39 cents no name pasta. However, due to all the panic buying, almost all pasta was sold out. Except for the expensive brand pasta, which usually costs 1.43€, but it was reduced to 77 cents for some reason. It's in a fancy carton package with a little see through window. The cheap pasta I get is just a plastic bag. So I got the expensive pasta, but I still bought the cheapest pasta sauce for 79 cents. I cooked the pasta 15 minutes ago. What the actual f@%$. I cooked it, and it's still thin. But it's not hard. It's stretchy, but it doesn't rip. You don't have to drown it in sauce until it's edible, no, it tastes good even without the sauce. I mixed it with the sauce. What is this. I didn't know pasta could have physical property like that. I can't go back to 39 cents pasta. Now imagine you're rich and this x100 is your average meal."