Classification models on subreddits [r/Capitalism and r/communism]

By Genaro Garcia Pereyra

Problem Statement

This projects seeks to demonstrate a classification model that is capable of classifying the subreddits of **r/Capitalism** and **r/communism** better than the baseline model given by the majority class in the dataset collected. In addition, the goal for this data analysis is to provide more information of the sentiment that users display on each of the subreddits and how sentiment and influence in some degree to the accuracy of the classifier.

The classification model uses ensemble methods in order to reduce the risk of overfitting and have a good perform in unseen data by combining predictions from several base models.

Accuracy is used as the metric for evaluation for all the models, and depending the model's performance on this metric the final model is determined.

Data Collection

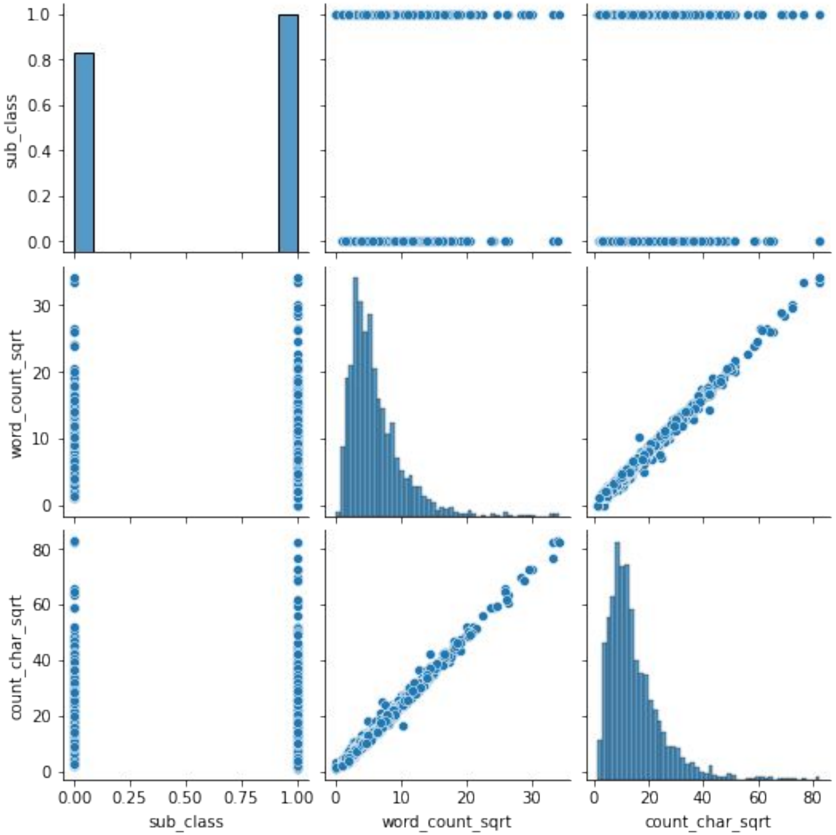
Data collection was performed with the help of the **pushshift API** which allows us to search all publicly available comments and submissions on Reddit. For this particular case the text to be classified only considers the **comments section** of each post, this decision was made with the idea of having more available text that could allow a better prediction of the models implemented. In addition, the interactions between users in the comments section is a critical component on how to determine the general sentiment of the users that use that particular subreddit.

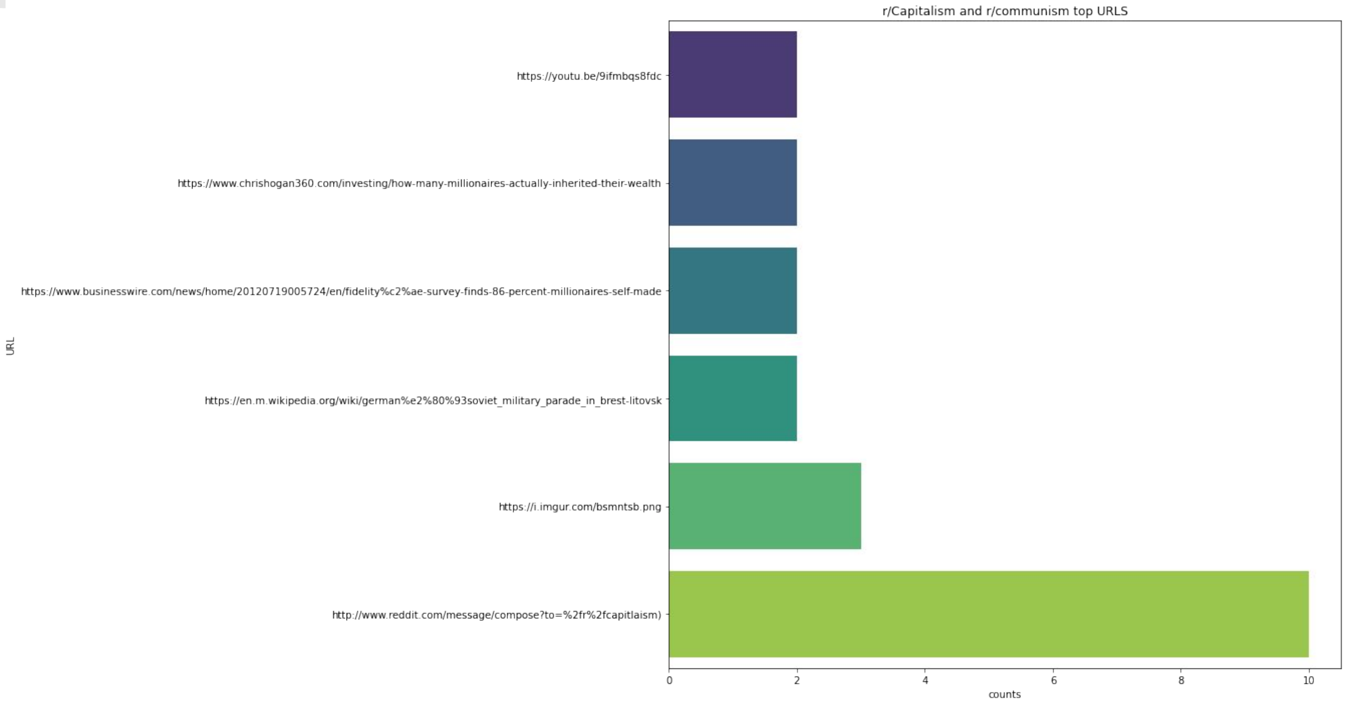
A size of 100 most recent comments was extracted for each of the last 11 months starting form (10/19/2020) for both the r/Capitalism and r/communism subreddits. This data was then used to created a single balanced dataset containing 2200 observations.

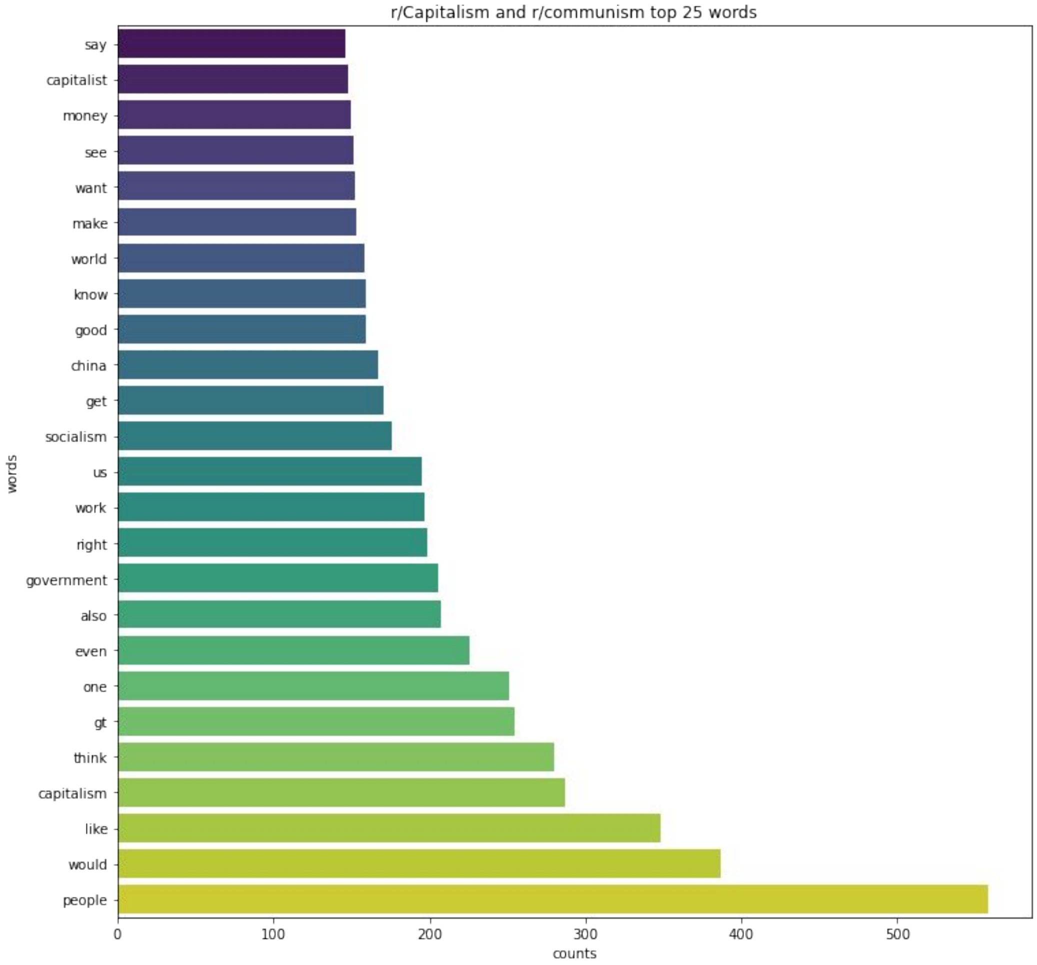
Data Cleaning + EDA

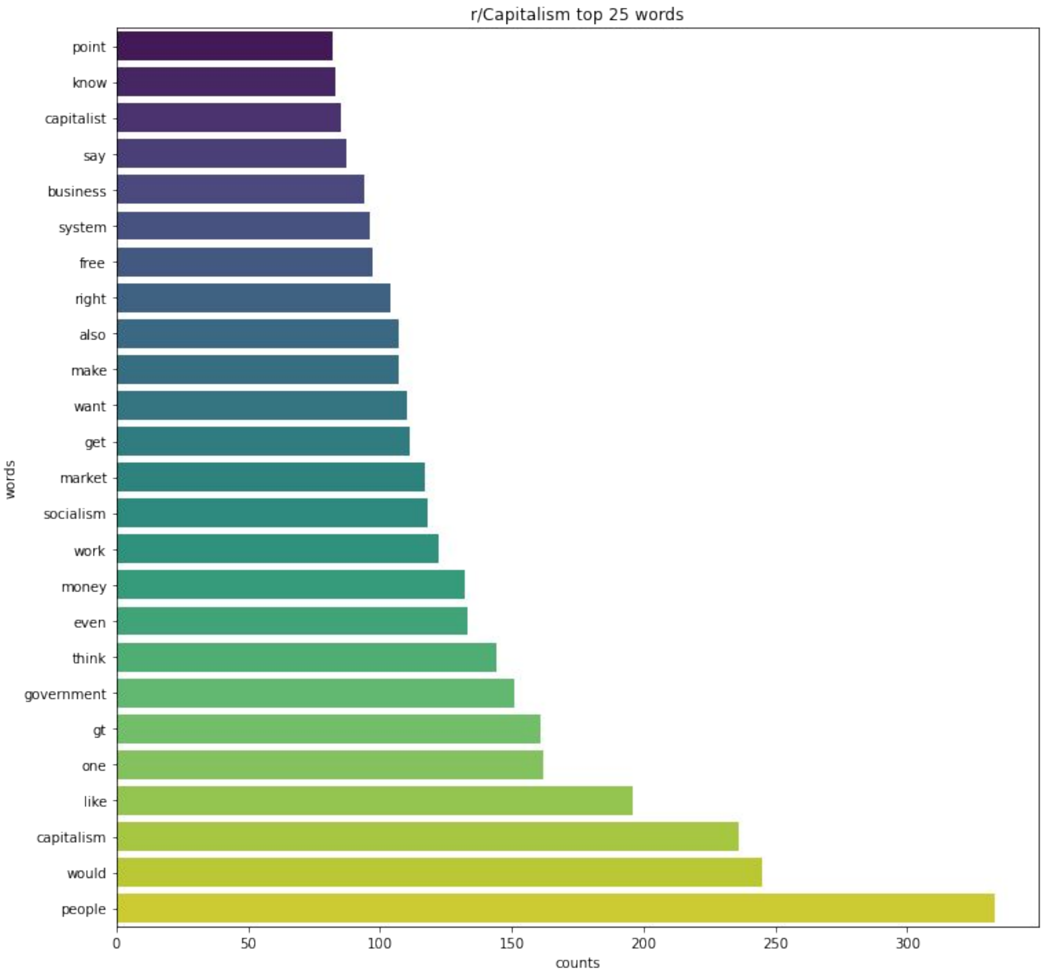
Goals after performing data cleaning on 2200 comments extracted:

1. Remove presence of URLs
2. Remove presence of Russian words
3. Remove presence of Chinese words
4. Remove english stop words
5. Remove numerical values
6. Take care of all removed comments by users
7. Binarize class

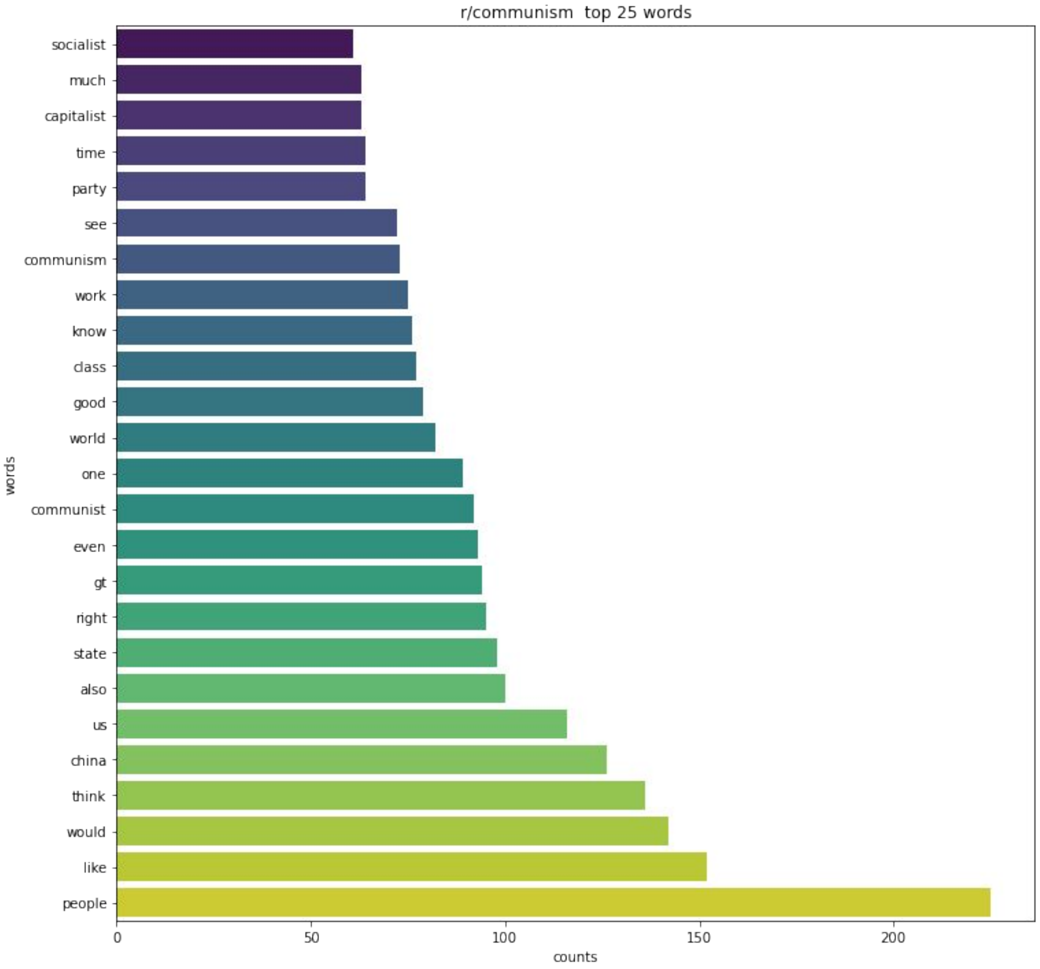

sub_class
word_count_sqrt
count_char_sqrt
1.0
0.8
0.6
0.4
0.2
0.0
30
20
10
0
80
60
40
20
0
0.00
0.25
0.50
0.75
1.00
0
10
20
30
0
20
40
60
80
sub_class
word_count_sqrt
count_char_sqrt


r/Capitalism and r/communism top URLS
URL
https://youtu.be/9ifmbqs8fdc
https://www.chrishogan360.com/investing/how-many-millionaires-actually-inherited-their-wealth
https://www.businesswire.com/news/home/20120719005724/en/fidelity%c2%ae-survey-finds-86-percent-millionaires-self-made
https://en.m.wikipedia.org/wiki/german%e2%80%93soviet_military_parade_in_brest-litovsk
https://i.imgur.com/bsmntsb.png
http://www.reddit.com/message/compose?to=%2fr%2fcapitlaism)
0
2
4
6
8
10
counts

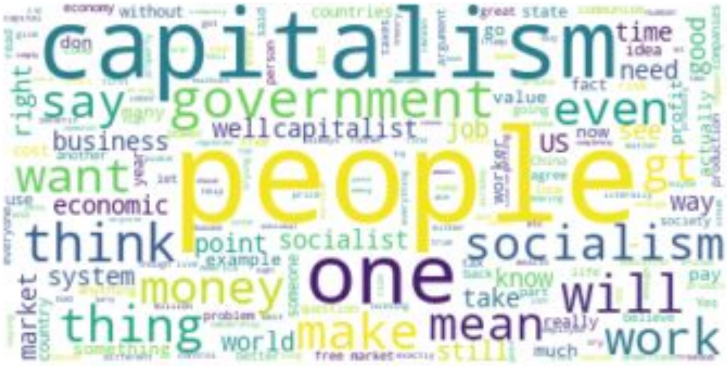

r/Capitalism and r/communism top 25 words
words
say
capitalist
money
see
want
make
world
know
good
china
get
socialism
us
work
right
government
also
even
one
gt
think
capitalism
like
would
people
0
100
200
300
400
500
counts



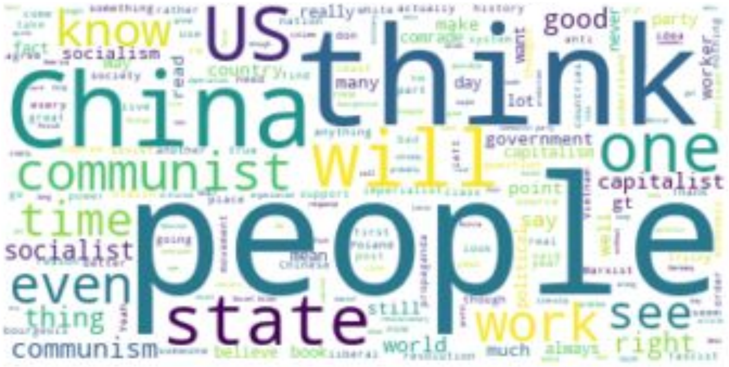
r/Capitalism top 25 words
words
point
know
capitalist
say
business
system
free
right
also
make
want
get
market
socialism
work
money
even
think
government
gt
one
like
capitalism
would
people
0
50
100
150
200
250
300
counts

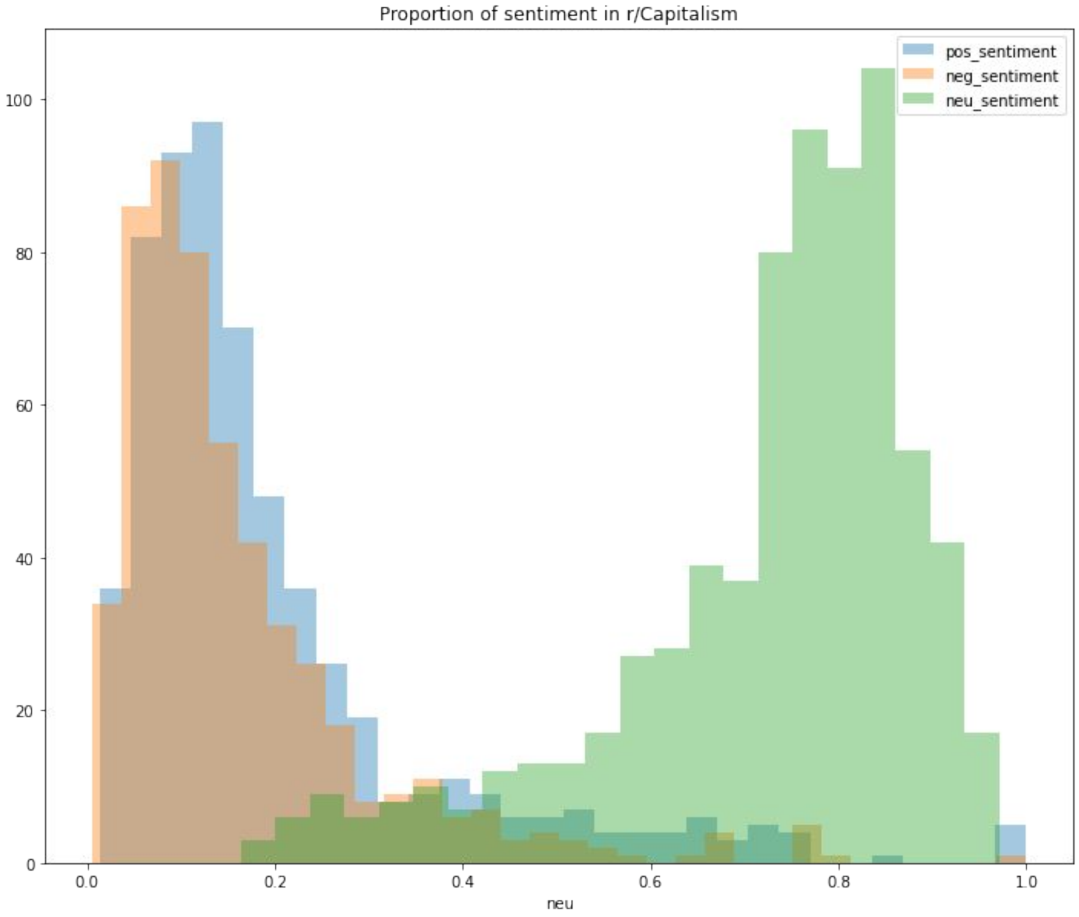

r/communism top 25 words
words
socialist
much
capitalist
time
party
see
communism
work
know
class
good
world
one
communist
even
gt
right
state
also
us
china
think
would
like
people
0
50
100
150
200
counts

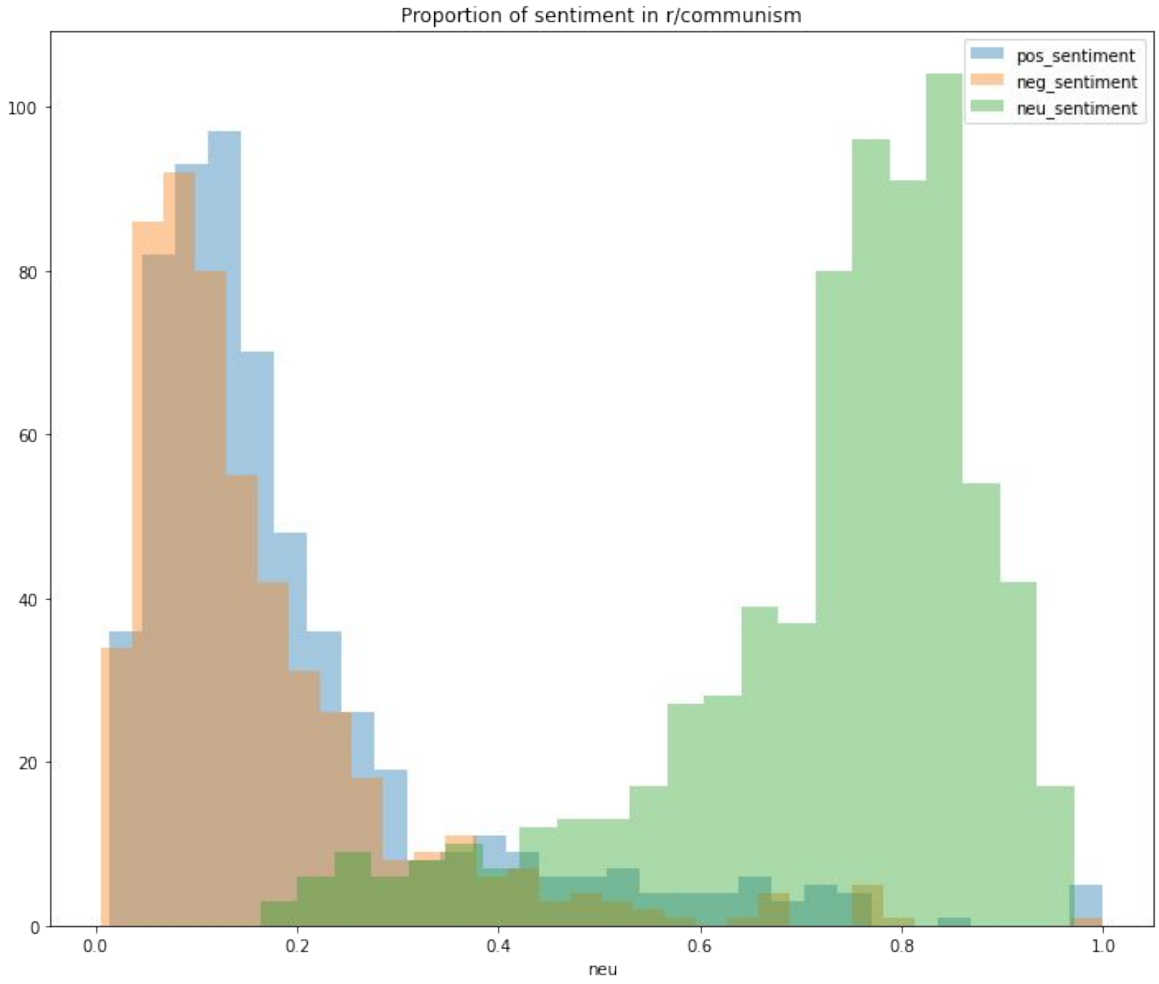
r/Capitalism top 25 words

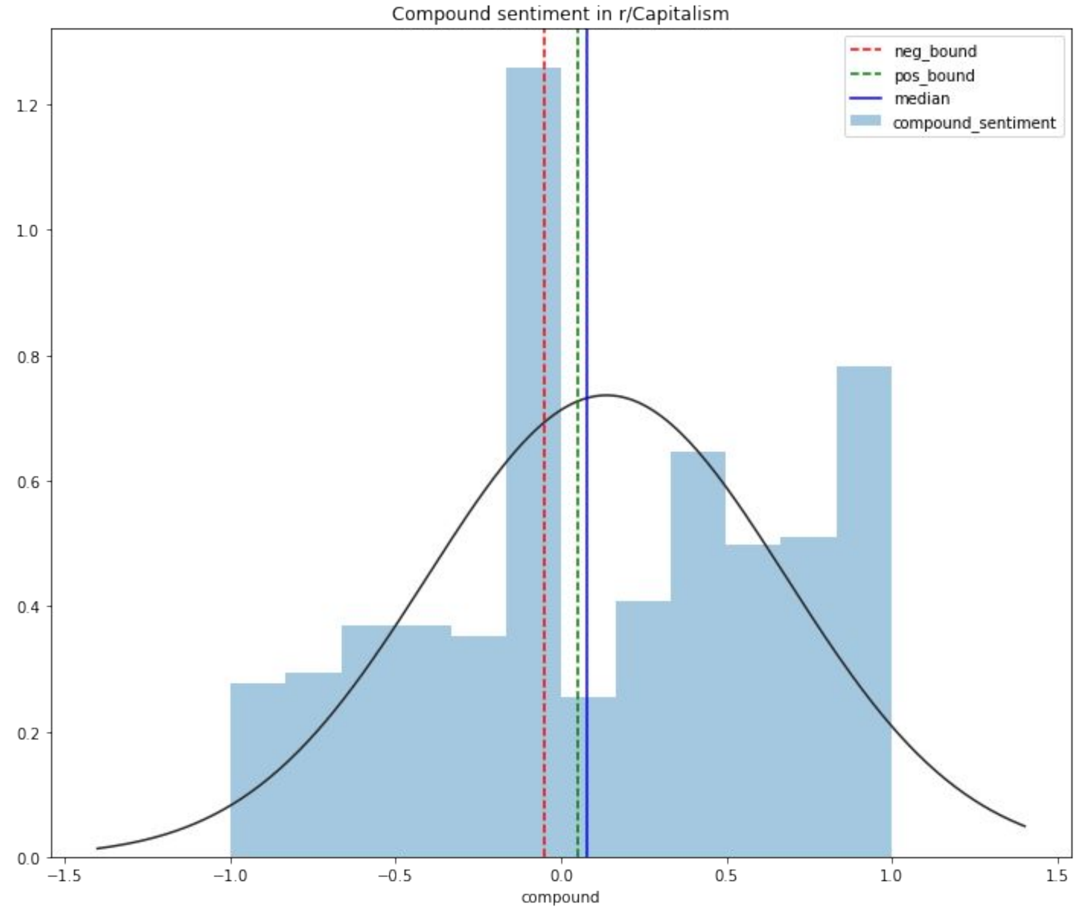


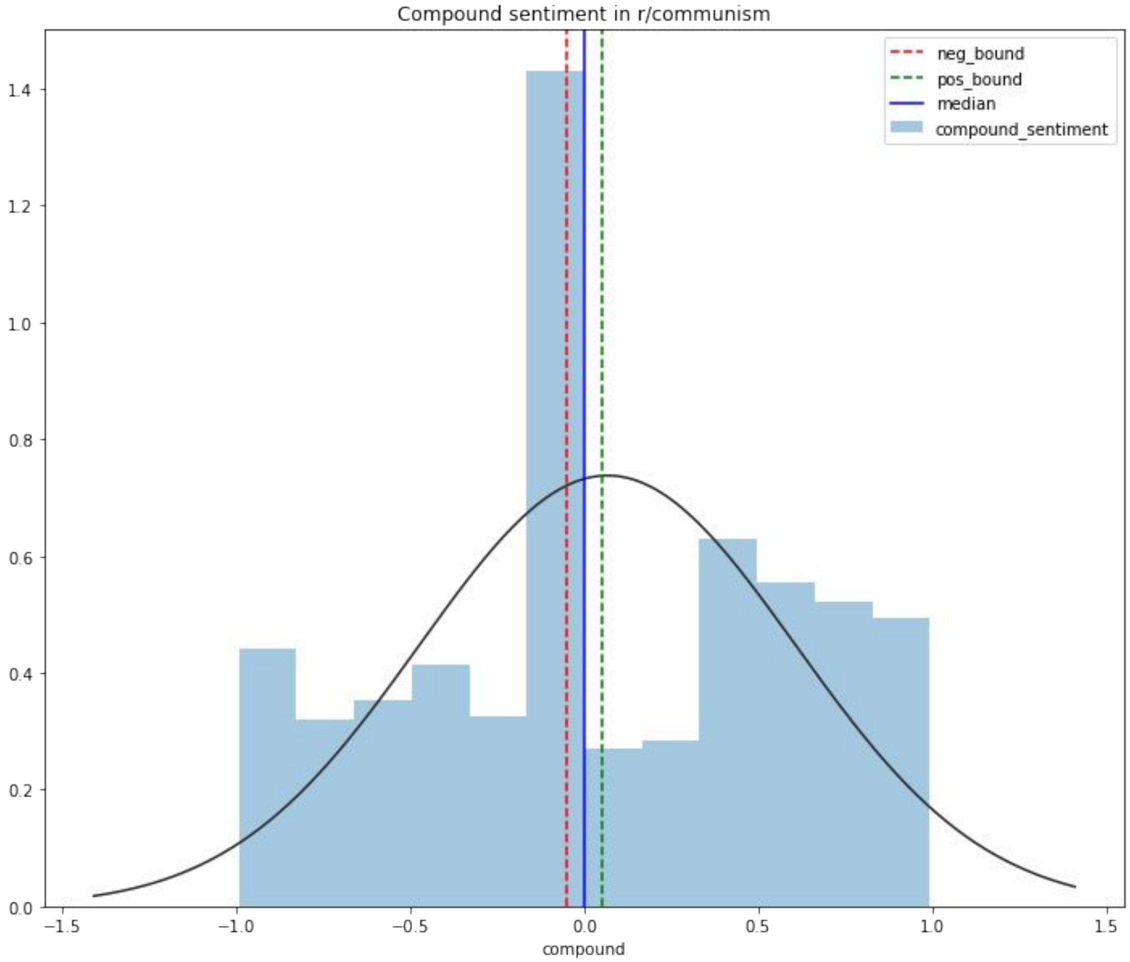
r/communism top 25 words




Proportion of sentiment in r/Capitalism
pos_sentiment
neg_sentiment
neu_sentiment
100
80
60
40
20
0
0.0
0.2
0.4
0.6
0.8
1.0
neu


Proportion of sentiment in r/communism
pos_sentiment
neg_sentiment
neu_sentiment
100
80
60
40
20
0
0.0
0.2
0.4
0.6
0.8
1.0
neu


Compound sentiment in r/Capitalism
neg_bound
pos_bound
median
compound_sentiment
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
compound


Compound sentiment in r/communism
neg_bound
pos_bound
median
compound_sentiment
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
compound

Baseline score: 54%

Modeling

Model	Score on Training set	Score on Testing set	Vectorized text	VADER sentiment analysis
RFC + GS	~ 0.9849	~ 0.7116	Yes	No
Logistic Reg + GS	~ 0.9071	~ 0.7382	Yes	No
RFC + GS	~ 0.9979	~ 0.7668	Yes	Yes
Logistic Reg + GS	~ 0.9098	~ 0.7525	Yes	Yes
Ensemble [LogReg + RFC]	~ 0.9938	~ 0.7689	Yes	Yes

Conclusion and Recommendations

The use of sentiment analysis metrics as part of the features used in the model to classify the holdout test set of subreddits of c/Capitalism and c/communism was essential for the creation of a model that could generalize better on unseen data. As seen in the word clouds for both subreddits there is high similarity of words used in both comment sections, and because of this it makes it harder for the model without VADER to obtain an acceptable accuracy score in the test set.

It is recommended to use tools such as VADER to have a better classification model. Furthemore, because of the presence of other languages it might be useful to incorporate corpus from Russian and Chinese languages.