Data Story

By MD Wasim Zaman, 1007007640

Introduction

In this data blog post, we will be looking into how generalists and specialists correlate with creators online. What do we mean by creators? In our data analysis we are using reddit pushshift.io dataset, so creators in this context are the post creators. The communities on reddit are called subreddits, where a user can post a thread and others can comment on it.

This study builds on top of "Generalists and Specialists: Using Community Embeddings to Quantify Activity Diversity in Online Platforms", a paper which quantifies specialization into a score known as GS-score. In our processes this score varies from 0.5 to 1.0, with the closer it is to 1.0, the more specialized the user is in which subreddits they interact with.

We want to use this GS-score to investigate how specialization of creators vary on social media. We shall look into it via 4 avenues: key post statistics (reddit upvote scores, number of comments, number of posts, etc), look in-terms of communities as a whole, sentiment analysis of comments and post titles, and finally use social dimensions to see how they vary with specialization.

First we load and clean our two data files which are reddit post statistics and comment statistics from 2019 to 2021 of the top 5k subreddits.

Cleaning pushshift.io data and computing GS-scores

Since the data we are loading is very large we have to preprocess some critical files:

a) GS-scores calculated via posts by users
b) Sentiment analysis on each post title
c) Sentiment analysis on each post's comment

For sentiment analysis we are using Vader with NLTK package.
For the GS-scores we are loading in a pre-existing word embedding (mentioned in README.md).

Calculating GS-scores:
For the post gs-scores we calculate the center of mass of each poster which is the average of the vector of each subreddit they post on.
Then we calculate the weighted average of the dot products of the center of mass against the vector of each subreddit they post on.

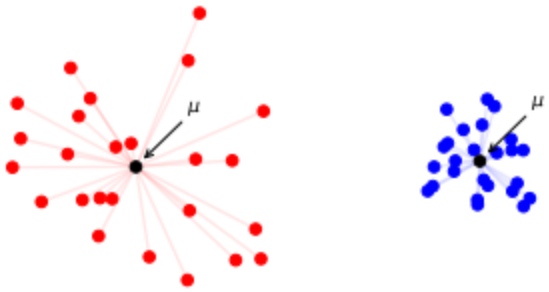
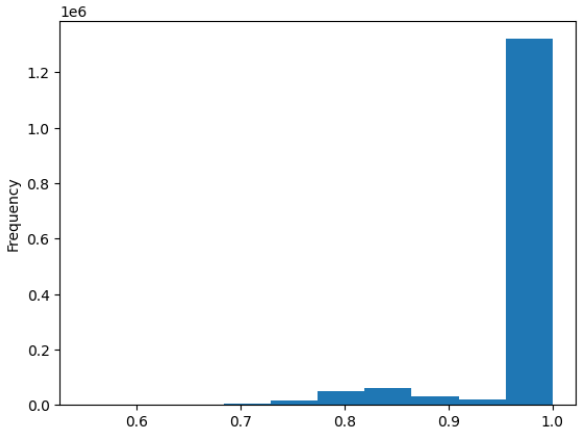


Figure 1: A schematic depicting the vector representations of communities contributed to by a generalist (left) and a specialist (right). The generalist's communities are spread out, and the specialist's communities are clustered together.

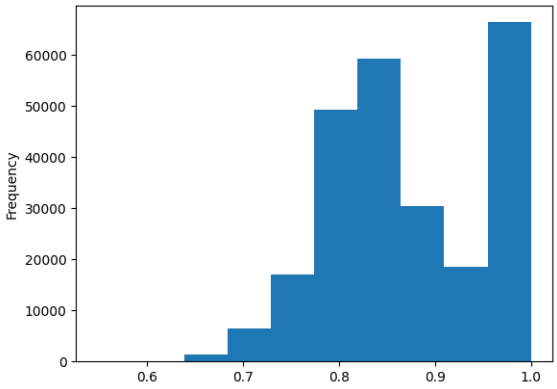
(Wallet et al, 2019)

We store each of the GS-score of each poster (or author in the actual csv).
This is our independent variable for our analysis.

We can see the distribution of our GS-scores:



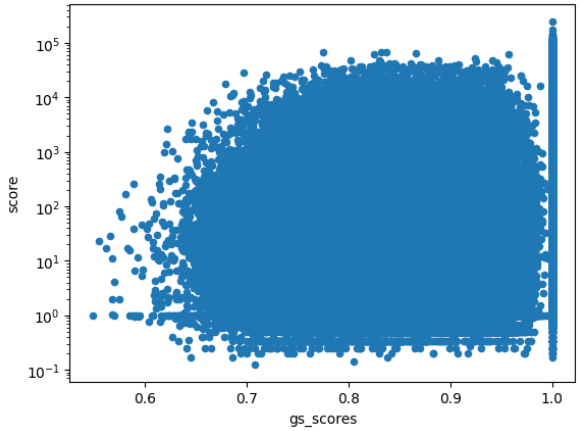
A large number of people have number_of_engagements, the total number of posts the user has made, at 1 which makes them a specialist by default. We can filter these users out to see the a distribution of our gs-scores of users who aren't one time.

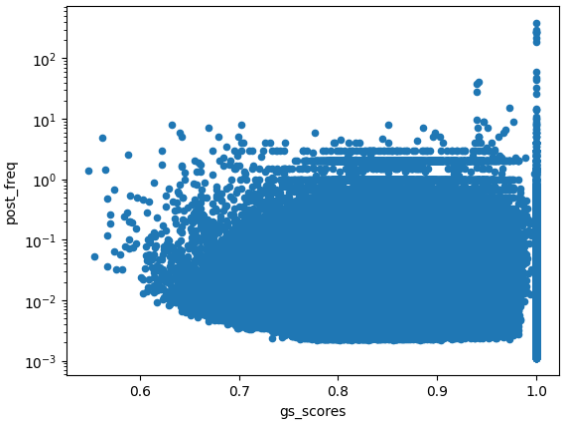


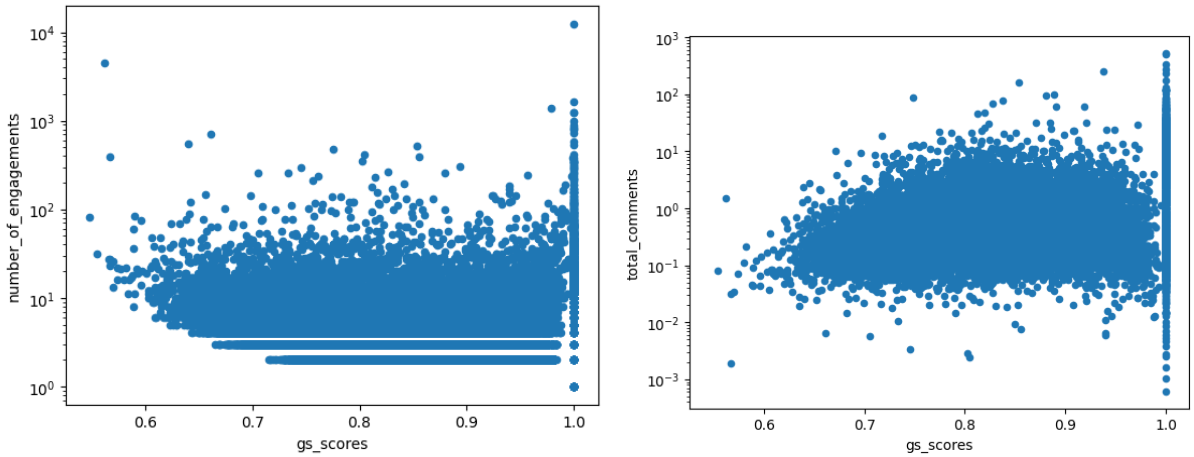
Note: the number of pure specialists (GS = 1) are still high, meaning creators on reddit are already very specialized by default.

Comparison with post statistics

After loading in the GS-scores from a pre-processed csv we can do scatter plots against some key statistics which measures the success of a post.
Some key statistics we can immediately get from the pre-existing data are the:
Average scores(upvotes - downvotes) per post
Total number of posts by user (number_of_engagements)
Total comments per post of a user
Post frequency (total number of posts / (duration between first and last post in days))





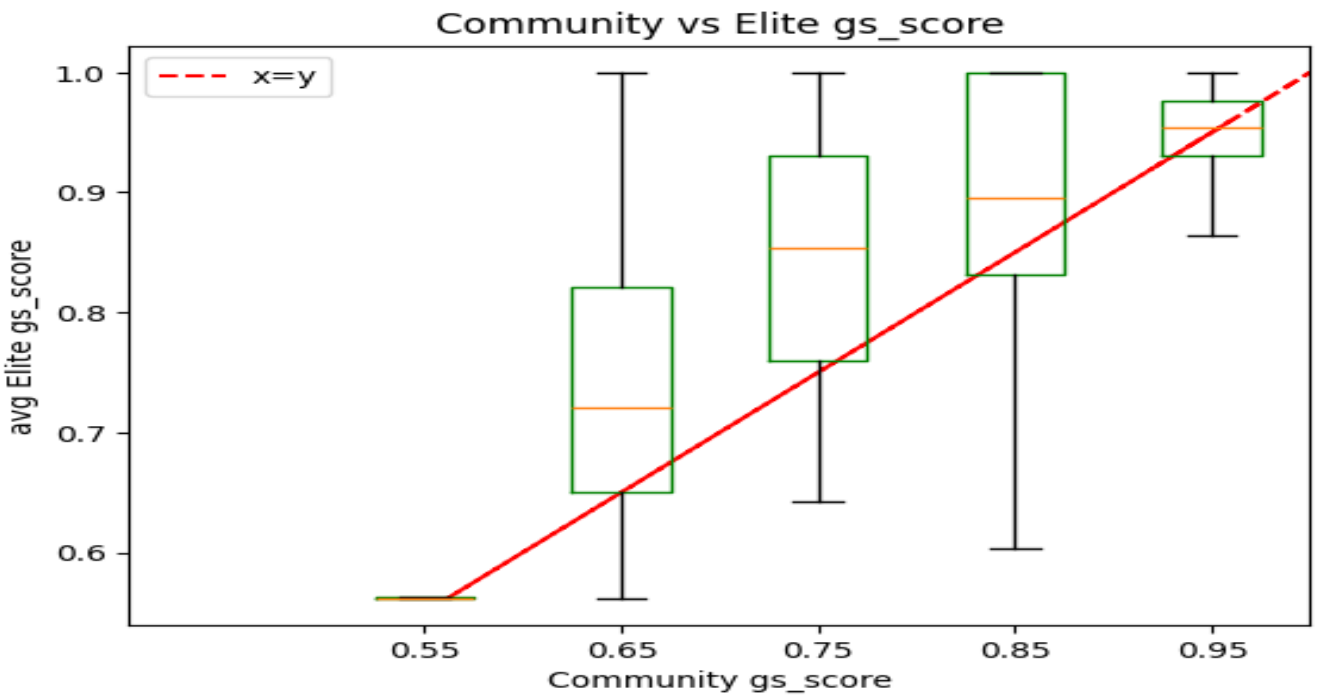


We have two key take aways from this preliminary analysis:
Looking interms of posters may not be too wise as there is a very large variance in data, and most key statistics don't seem to vary with specialization (this may be due to the communities themselves being a better determining factor).
Thus we can now look interms of the communities themselves

Who are the elite posters in a community

In the original paper they looked at the community gs-scores and the elites in the community (top 5% user ranked by average post score).
We can draw a line of x=y and based on whether average elite posters are above or below that line, we can see if the elite posters are generalists or specialists

The community gs score is the average weighted gs scores of active posters weighted by number of posts made on that subreddit.



Each box plot is made using the communities with gs_scores within 0.025 of its center.
For example the box plot at 0.65 is made using 0.625 to 0.675.

Interpreting this graph we can see that the median is always >= the y=x line, however there are some outliers which increases the range.

Using the same methodology from the original paper we can see that the elite posters are more specialized than the community they post in. This is interesting since in the original study (using commenter) they found the opposite. This makes sense since creating a post is a much more involved activity, and more specialized users would be more skilled and hence they make better posts.

Using sentimental analysis to view post success

Other than posts we can utilize the actual comments on the post. Total comments may not be a very good indicator of post success and it may not correlate well with gs scores (as seen with the graph).

Thus we can use sentimental analysis on the text inside the post and view how positive they are. Vader gives 4 sentiment values: neu (neutrality of a text), pos (how positive a text is), neg(negativity of a text), compound (one score which shows the overall sentiment)

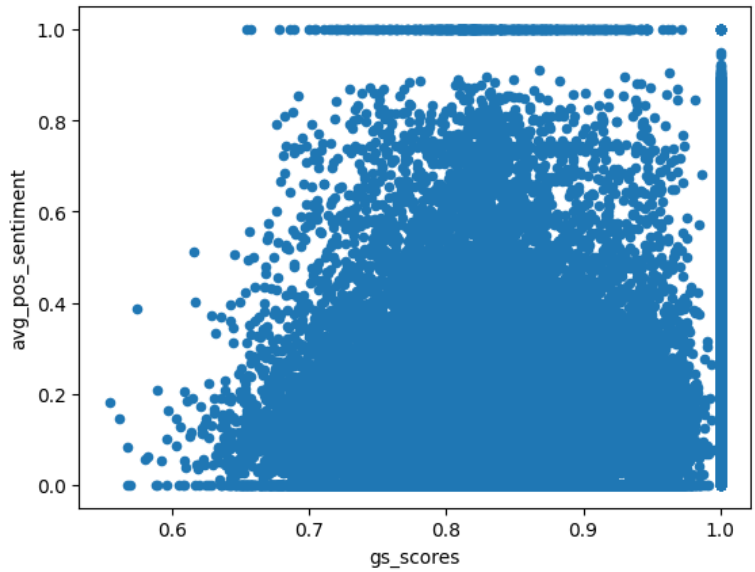


Fig a: average positivity of comments against gs_score of user on a post level

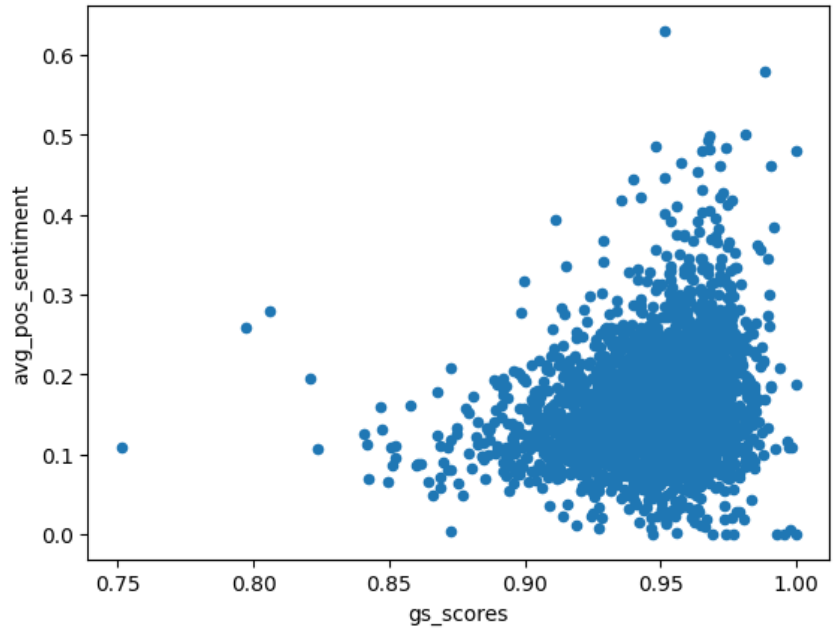


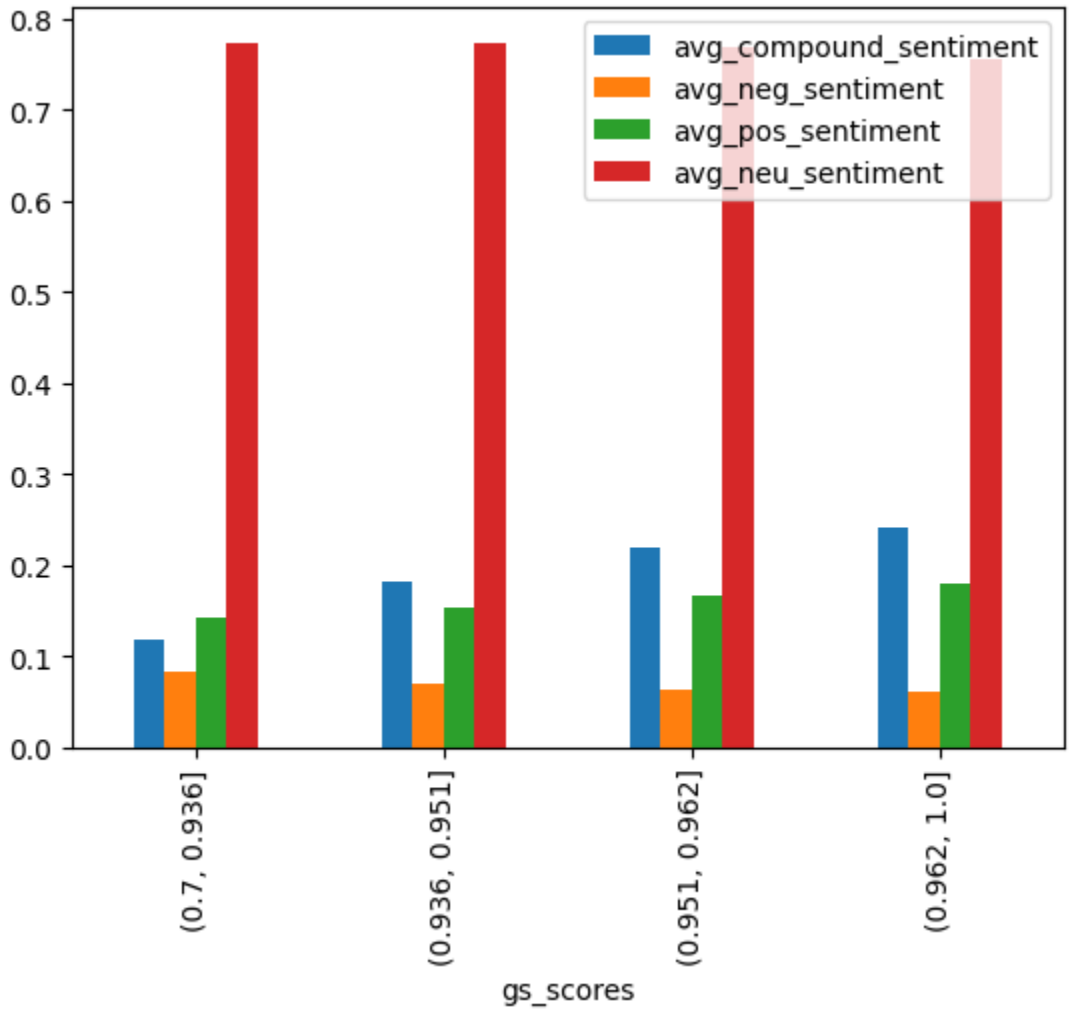
Fig b: positivity vs community gs scores, this is on a community level

We can see that the data is less chaotic on a community level, so we shall conduct out analysis using community gs scores.

We can break the sentiments along the quartiles (0 to 25th percentile, 25th to 50th, 50th to 75th, 75th onwards), and see how each sentiment varies.

```
25%          0.936279
50%          0.951048
75%          0.962732
max          1.000000
```

Our quartile values



Community gs scores vs average of each sentiment

gs_scores	avg_compound_sentiment	avg_neg_sentiment	avg_pos_sentiment	avg_neu_sentiment	gs_scores
(0.7, 0.936]	0.118537	0.082475	0.141126	0.773485	0.917273
(0.936, 0.951]	0.181046	0.070245	0.153998	0.772151	0.944124
(0.951, 0.962]	0.219658	0.062838	0.167056	0.767909	0.956529
(0.962, 1.0]	0.240096	0.061953	0.180432	0.755036	0.970843

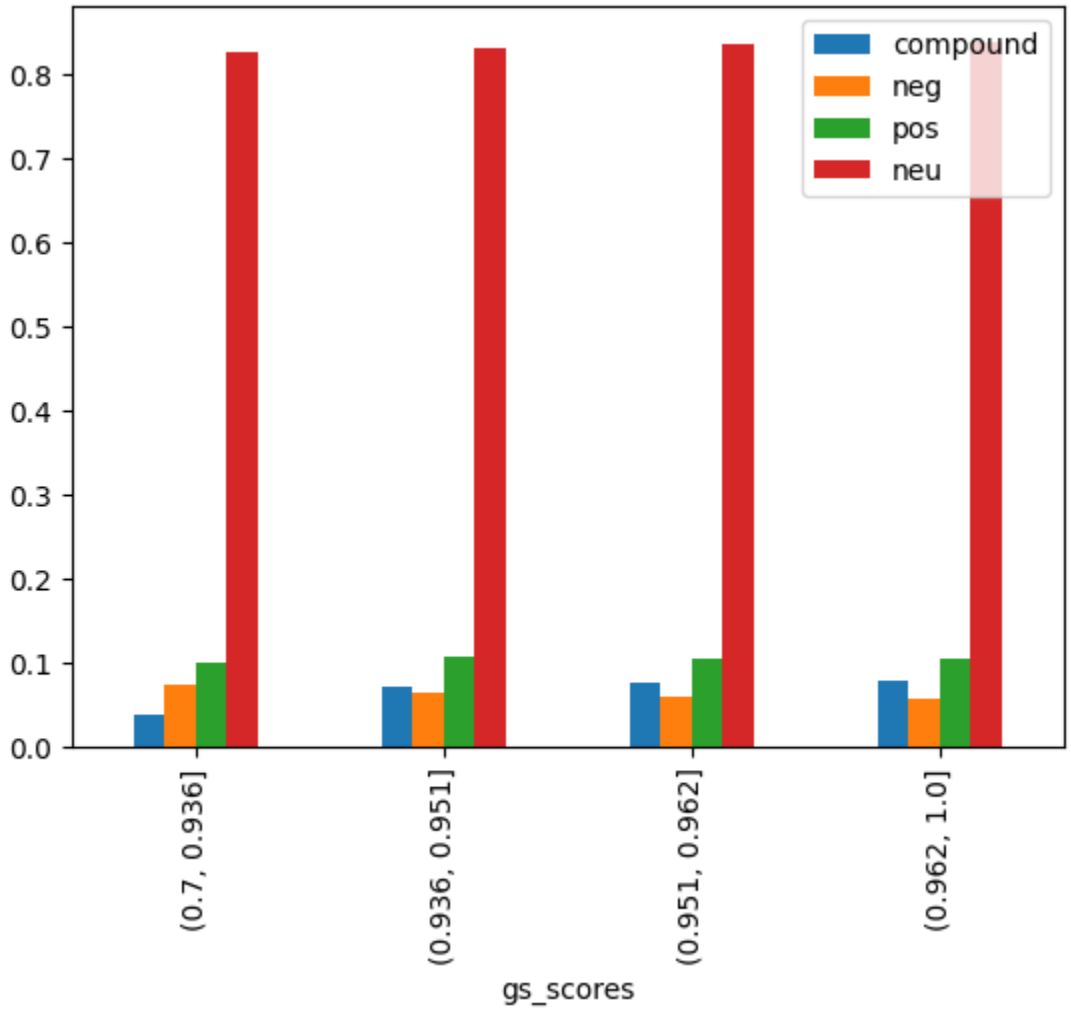
Actual values

We can see that most of the comments are neutral in all posts.
However, the positivity increases alongside gs scores, and negativity decreases slightly.

	avg_pos_sentiment	gs_scores
avg_pos_sentiment	1.000000	0.220888
gs_scores	0.220888	1.000000

We have some weak correlation between positivity and gs scores.

We can look at the post's title's sentiment and how they vary along side the quartiles (0 to 25th percentile, 25th to 50th, 50th to 75th, 75th onwards) too.



	pos	neu	neg	compound	gs_scores
gs_scores					
(0.7, 0.936]	0.099821	0.825551	0.073322	0.038405	0.917273
(0.936, 0.951]	0.107465	0.828770	0.062425	0.069783	0.944124
(0.951, 0.962]	0.104506	0.835552	0.058826	0.074576	0.956529
(0.962, 1.0]	0.104659	0.837571	0.056644	0.079033	0.970843

The values don't vary much along the quartiles and there isn't much of a trend. The post's title sentiment may not be a very worthwhile indicator of the post.

Thus, we can conclude that more specialized communities have slightly more positive discussions.

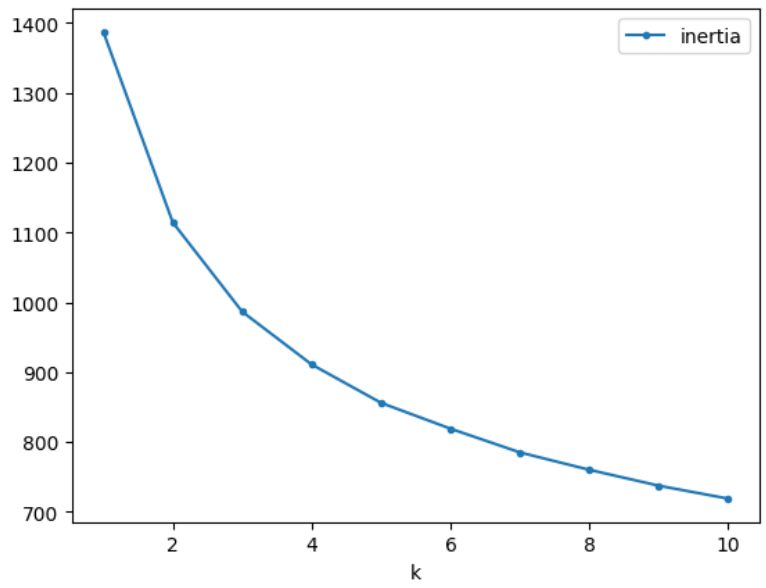
Viewing specialization in terms of social dimensions

The same data source from which we got our word embeddings also built social dimensions for each subreddit embedding.

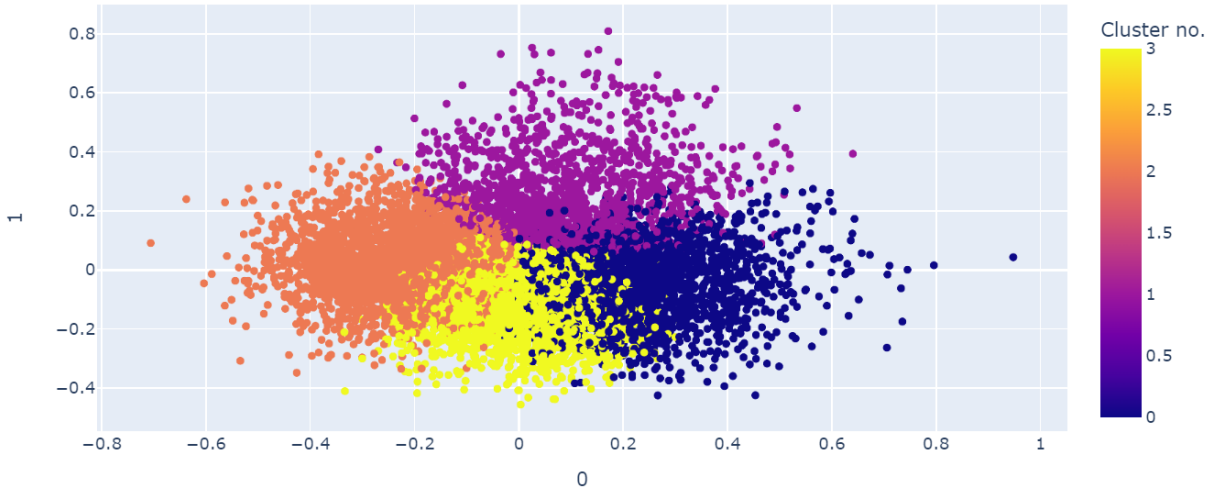
For example the 'sociality" dimension is made up of some similar subreddit differences, such as the vector direction of nyc to nycmeetup and law to LSAT, this vector is believed to be sociality. The sociality dimension is then calculated by projecting the word2vec embedding of a subreddit upon that direction.

We can try and cluster the subreddits based on the social embeddings and then visualized them via PCA.

First use elbow method to find optimum number of centers.



We use 4 centers.



We can then cluster using kmeans and then visualize via PCA.
Note there isnt any clear boundaries.

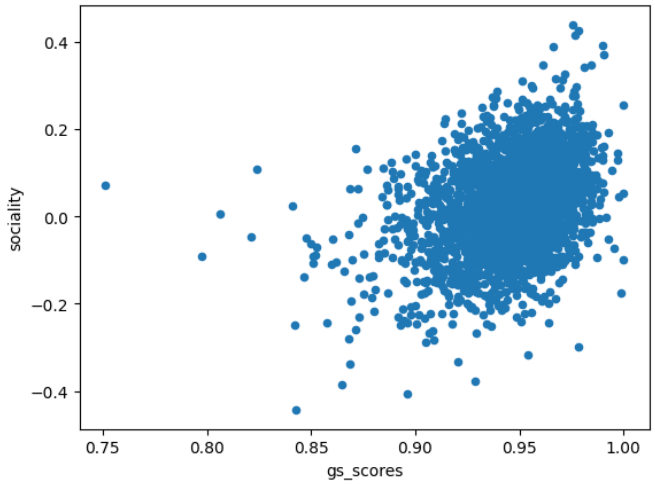
We can then find the average gs_scores of each cluster and compare them.

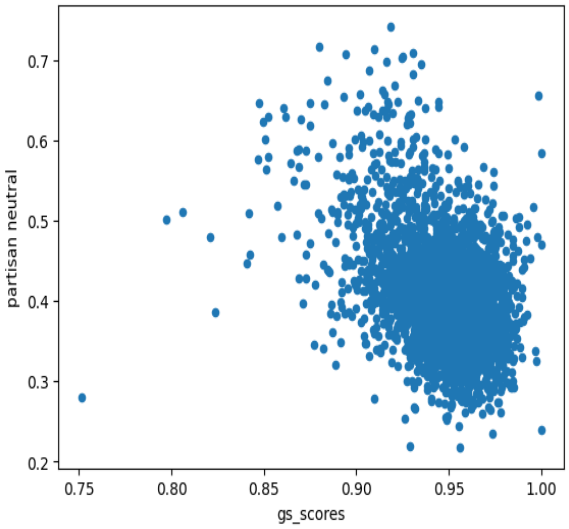
cluster	gs_scores
0	0.937511
1	0.954034
2	0.952540
3	0.945133

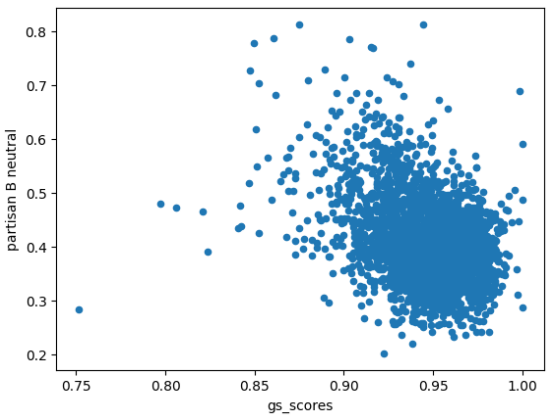
There seems to be some variance in the gs_scores however not by much.
Thus we can look into individual social dimensions to see how community gs scores vary.

In the analysis we plot community gs along each social dimension ('age', 'age B', 'age neutral', 'affluence', 'gender', 'gender B', 'gender neutral', 'partisan B', 'partisan B neutral', 'partisan', 'partisan neutral', 'edginess', 'sociality', 'time').

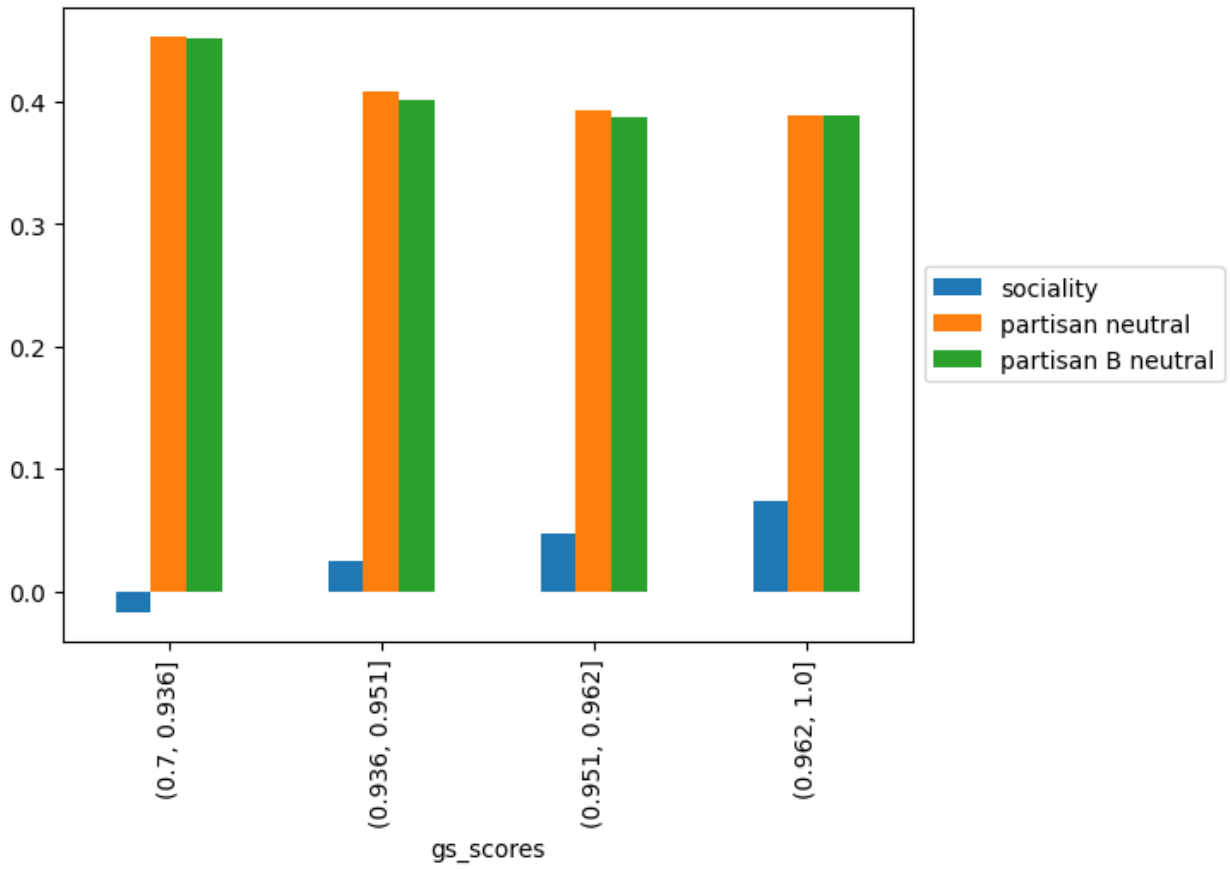
3 dimensions have scatter plots which look like theres some correlation:
Sociality, partisan B, partisan B neutral.







We can similarly break them up into quartiles and then see how each dimension varies:



gs_scores	sociality	partisan neutral	partisan B neutral
(0.7, 0.936]	-0.017214	0.452758	0.450972
(0.936, 0.951]	0.024542	0.408416	0.400206
(0.951, 0.962]	0.047196	0.392273	0.386893
(0.962, 1.0]	0.073651	0.388882	0.387812

We can clearly see that sociality positively correlates with gs_score quartiles.

	sociality	gs_scores
sociality	1.000000	0.360378
gs_scores	0.360378	1.000000

Weak pearson correlation between gs_scores and sociality

Conclusions:

We can make 3 conclusions:

The elite creators in a community are the specialists. This makes sense as more specialized creators would be able to make posts which has more niche appeal hence higher score.

More specialized communities have more positive discussions. We can think of as more tightknit communities and hence the discussion is more likely to be better.

The “sociality” dimension correlates with more specialized communities.

However, due to lower frequency of posts our results may be skewed. The gs-score is not a good indicator of specialization when the number of datapoints per user is low, as low engagements would make the average user more specialized.