Gen-Z Kampung Spirit: A Comparative Analysis of the Reddit Activity of Singaporean Students and their International Peers

Ezra Koh See Hwa
College of Computing and Data Science

Supervised by Asst Prof Farhan Ali
National Institute of Education

Abstract - Reddit is a fast-growing anonymous social media platform which houses over 100,000 communities, or subreddits. One such community is r/SGExams, which is dedicated to Singaporean students. Currently 236,000 members strong, it is the largest country-specific education subreddit. How do Reddit profiles of Singaporean adolescent students differ from their peers around the world?

To answer this question, 109 users who actively create posts and 112 users who actively comment on r/SGExams were first identified. Their posting habits within the community and their interests outside of it were analysed using mathematical, machine learning, and natural language processing methods. A random sample of students around the world was taken from r/teenagers and the same analysis methods were repeated.

Out of necessity, we developed a new method of collecting data from Reddit after the commonly used Pushift API was shut down in May 2023. Large-scale data from Reddit (2 million posts and 30 million comments across 11,000 subreddits and 1.6 million users) was collected using the Python Reddit API Wrapper (PRAW). We also developed a multi-class model to efficiently classify subreddits into 19 topics.

Some teachers have already turned to r/SGExams to learn about their students and are incorporating feedback and popular memes into their lessons. Given the large and increasing proportion of Singaporean students participating in this community, the findings of this study and subsequent studies will provide influential insights invaluable to policymakers, educators, and parents.

Keywords - Reddit, students, Singapore, machine learning

1 Introduction

We began this study interested in exploring the online behaviour of Singaporean adolescent students with respect to their interests in various topics and contrasting these observations with that of students around the world. For these purposes, we decided to analyse Reddit data as it is relatively easy to retrieve and differentiate between Singaporean and international users. The sample Reddit activity of Singaporean students was taken from the r/SGExams subreddit, which is “the largest community on reddit discussing education and student life in Singapore” [1]. It boasts 236,000 members, most of whom are from secondary schools and pre-tertiary institutions, and consistently ranks in the top 10 active subreddits in the Education category. For international students, we sampled the r/teenagers subreddit, which has over 3 million members, asserts itself as “the biggest community forum run by teenagers for teenagers” and primarily caters to users aged 13 to 19 years old [2].

In this paper, we question a remark commented on a post in r/SGExams [3] by Reddit user u/Eshuon, “This is basically r/teenagers”. Exactly how true is that statement?

2 Literature Review

[4] investigated the evolution of the interests of over 7 million Reddit users and summarized the general trend of user “lifetimes” and activity. They manually classified almost 1000 subreddits into 15 topics and quantified user movements across different subreddits and topics by geometrically encoding the activity of each user into a 15-dimensional space and applying linear algebraic methods to analyse the temporal changes. [4] observed that “users are more prone to a drifting behavior”, which is defined as a sudden change from one subreddit to another within the same topic.

While some algorithms and calculations were not outlined clearly in the paper, the definitions of important terms were specific and well-written. This made it easy for us to understand, recreate, and improve the processes performed by [4].

[5] made significant contributions to analysis of Reddit content and identification of user attributes on social media platforms in the form of a large annotated dataset (RedDust) that was carefully curated from posts and comments containing

statements the users made about themselves. They also applied a unique approach to data collection by including information unique to social media platforms such as “bracket patterns matching structured assertions of users’ ages and genders” and flairs, which are keywords used on user profiles and posts to assert personal traits particular to specific subreddits. For example, “ ‘male’, ‘female’, or ‘trans’ in r/AskWomen”.

Using Hidden Attribute Models (HAM) and Convolutional Neural Networks (CNN) as test cases, the authors demonstrated that RedDust is suitable to be used as training data for attribute predicting models.

3 Dataset Creation & Preparation

3.1 Initial Data Collection

In May 2023, just a few months before the start of this study, the popular Reddit API Pushift had its access to Reddit data revoked for failing to comply with updated terms of service. Thus, we turned to Python Reddit API Wrapper (PRAW) as an alternate means of collecting post and comment data from Reddit. The method that was developed involves creating a large list of keywords with varying degrees of commonality and using PRAW to iteratively search r/SGExams and r/teenagers and record the unique IDs of posts which contain these keywords. The search process was halted when a large majority of the keywords failed to return any new posts. After the search ended, the IDs were used to retrieve and download all posts that were found and all the comments on them. Posts and comments without author information (generally occurring when a Reddit user deletes their account) were removed.

This method yielded reasonably large datasets from both r/SGExams and r/teenagers; 38,920 posts and 666,978 comments and 1,676,845 posts and 29,196,257 comments respectively. All posts and comments retrieved were from 1 January 2019 to 12 January 2024. We used the Pandas and Numpy libraries for all data organisation and cleaning.

3.2 Identifying Active Users

For convenience, we define a “home subreddit” for each user in the activity dataset as being that subreddit (r/SGExams or r/teenagers) which they were identified as being active in.

Active posters on r/SGExams and r/teenagers were identified by counting the number of posts made by each user in their home subreddit and calculating the months between the first and last posts, thus obtaining the mean number of posts made per month by each user. Originally, we had planned to classify active posters as those who had at least 5 posts per month and 12 months between their first and last posts. However, all 1256 users who met the first requirement did not meet the second.

Therefore, we compromised the requirements and defined active posters as users who made at least 5 posts per month and a total of at least 50 posts and comments in their home subreddit . Users with the word “bot” or “automod” in their username were also removed. Finally, 114 users were classified as active posters.

A similar process was used to identify 112 active commenters, which we defined as users who made at least 10 comments per month in their home subreddit and had at least 12 months between their first and last comments.

Here it is necessary to include a defense of the defining criteria of active users. At first glance, the reader may have a knee-jerk reaction to the differences between the requirements for active posters and commenters, making the keen observation that a user only needs a total of 50 posts and comments in their home subreddit to be classified as an active poster, but needs a total of 120 comments in their home subreddit (posts are not counted) to be classified as an active commenter. Following from this observation, one might jump to the conclusion that it is obvious that active posters are less active on Reddit compared to active commenters. However, this assumption is shaky as there is no evidence or common sense to suggest that a user’s activity within their home subreddit is representative of all their Reddit activity, since the two home subreddits were selected by us, and these users may be more interested and active in other subreddits.

3.3 Activity Datasets

The posts and comments (hereafter referred to as activity) of all active r/SGExams posters and commenters were retrieved from their user profiles using PRAW. The activity dataset of active r/SGExams posters contains 109 unique users and 30628 activity (49.5% of which was in r/SGExams) spanning 1033 subreddits. The activity dataset of active r/SGExams commenters contains 112 unique users and 83946 activity (58.6% of which was in r/SGExams) spanning 1419 subreddits. The time range of the activity datasets for active users of r/SGExams is from 27 December 2015 to 13 February 2024.

To perform a consistent comparison with the r/SGExams active users, 109 active posters and 112 active commenters were randomly sampled from the r/teenagers activity datasets and their activity (199320 in total) was used for analysis. The time range of the activity datasets for active users of r/teenagers is from 8 June 2016 to 11 March 2024.

3 active users from r/SGExams and 1 active user from r/teenagers met the requirements to be classified as both an active poster and an active commenter. These users were retained in both groups.

3.4 Subreddit Classification

In preparation for analysis similar to that done by [4], subreddits had to be classified into topics. First, 3564 subreddits were manually classified into 19 topics (here) while 6781 subreddits remained to be classified by a machine learning model. Out of the manually classified subreddits, 3511 subreddit descriptions containing at least at least 10 characters were embedded using the Sentence-BERT all-MiniLM-L6-v2 model to be training data.

The compare_models function in Pycaret was applied to this training data and the linear discriminant analysis (LDA) model demonstrated the highest performance indices among other models such as logistic regression, K-nearest neighbours, LightGBM, naive Bayes, random forest, and ridge classifier. The hyperparameters of the LDA model were fine tuned using the tune_model function and the tuned model was used to classify 5543 subreddits which had at least 10 characters in their description. Most confidence scores for these classifications were high, with a first quartile of 0.839 and median of 0.986. For the remaining 1238 subreddits which did not have sufficiently long descriptions, their names were used as input data instead of the description. The confidence scores for these classifications were slightly lower, with a first quartile of 0.786 and median of 0.940.

We also attempted to train a model on subreddit descriptions at least 50 characters long, but the best performing model (ridge classifier), failed to predict some topics due to insufficient training data.

919 subreddits were classified as NSFW, the 20th topic, based on a native Reddit tag found using PRAW. In total, 11264 subreddits were classified into 20 topics.

4 Analysis & Results

4.1 Topic Attention Span

4.1.1 Calculations

To understand the varying “attention spans” that each active user has with respect to different topics, we followed the drift/shift analysis process performed by [4] closely. Drift refers to a sudden change of subreddit within a topic, while shift refers to a switch from topic to topic. The activity of active users on r/SGExams and r/teenagers was analysed separately using the same methods.

First we performed the shift analysis. Using the subreddit-topic tables, each activity is assigned a topic based on which subreddit it was posted in. Next, a table is created for each user and the user’s activity is sorted in chronological order from earliest to latest. Each column in the table represents a topic the user participated in and each row represents a bin of 10 activity which can also be formatted as a vector in n dimensions where n is the number of topics (columns). The last few activity of a user which is insufficient to make a new bin of 10 is excluded from calculations.

A new column “angle” is created and this column is filled up by calculating the cosine similarity of consecutive bins and taking the inverse of it to get the relative angle. For example, row 6 in the “angle” column records the relative angle between bin 6 and bin 7. Cosine similarity measures the similarity of two vectors, i.e. how close they are to each other. It is calculated by dividing the dot product of both vectors by the product of the lengths of each vector. Naturally then, a smaller angle between two bins indicates less attention shift and a larger angle indicates a larger attention shift. As defined by [4], a shift occurs when the relative shift angle between two consecutive topic bins is at least 45 degrees.

The drift angle calculation is similar to that of shift, but slightly more complicated. A similar table is created for each user and each row represents a bin of 10 activity. However, each column in the table represents a subreddit the user participated in. Next, between consecutive bins, the relative angle has to be calculated for each topic instead of just each bin. The relative drift angle between two bins is then taken to be the mean of all relative angles with respect to topic. Null relative angles (which occur when a user does not participate in a topic in both bins) are ignored in the mean calculation. Similarly, a drift occurs when the relative drift angle is at least 45 degrees.

4.1.2 Graphs

Figures 1 and 2 are samples of personalized user drift/shift graphs which are printed directly from the previously mentioned “angle” columns. The y-axis plots the angles between each consecutive bin and the x-axis is the user’s activity chronologically sorted into bins of 10. This sets up a pseudo time-series that is easy to understand and interpret. The green horizontal line is set at y = 45 to clearly show when a drift or shift occurs. Note that points directly on the green line constitute drifts/shifts. The two lists on the right of each graph show the top 4 topics/subreddits that the user has been most active in, in descending order. The graph has been deidentified by censoring the user’s Reddit username to preserve their privacy.

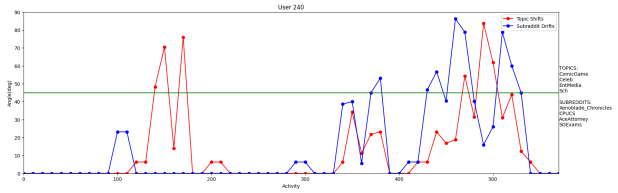


Figure 1: r/SGExams user

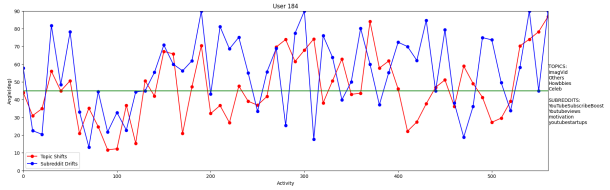


Figure 2: r/teenagers user

Figures 3 and 4 show on the y-axis the proportion of active users on r/SGExams and r/teenagers who have an average number of bins per shift which is at least x, where the x-axis represents number of bins. The blue lines represent posters and the red lines represent commenters.

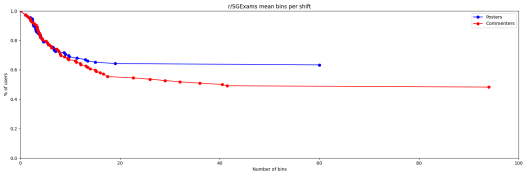


Figure 3

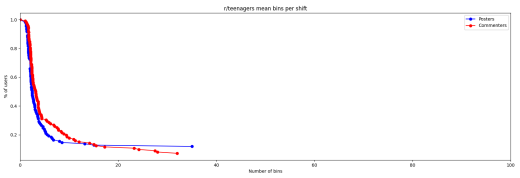


Figure 4

4.2 Topic Interest Trends

Figures 5 and 6 respectively show the amount and proportion of activity in each topic by active users from r/SGExams and r/teenagers. The respective home subreddits have been removed from Sch and Teens and reclassified into a new topic called HomeSub. The blue bars represent users from r/SGExams and the orange bars represent users from r/teenagers. The bars are sorted in descending order with respect to r/SGExams.

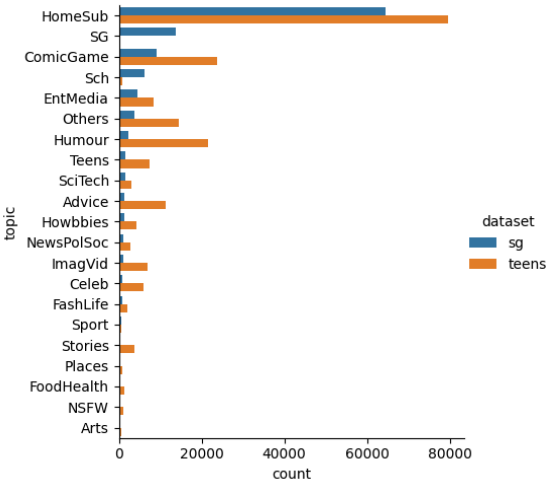


Figure 5

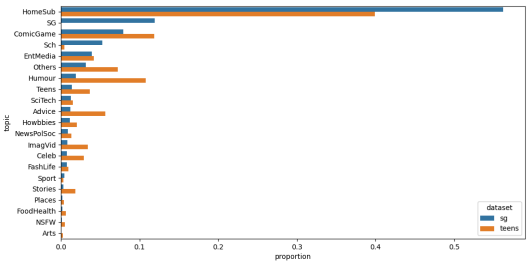


Figure 6

Figures 7 and 8 show the nominal/proportional changes in activity in popular topics over time. The top 3 topics from each home subreddit are displayed, namely HomeSub, SG, and ComicGame for r/SGExams, and HomeSub, ComicGame, and Humour for r/teenagers. The points marked by circles represent r/SGExams active users and the points marked by crosses represent active users from r/teenagers.

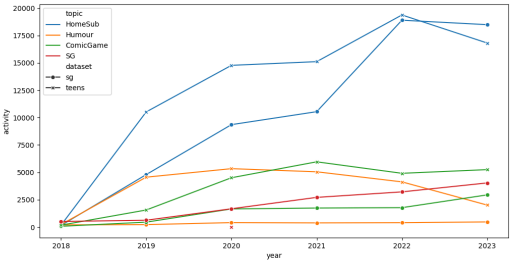


Figure 7

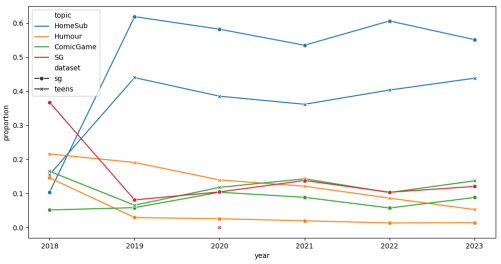


Figure 8

4.3 Other Statistics

Nominal statistics of the number of subreddits visited by active users from each group

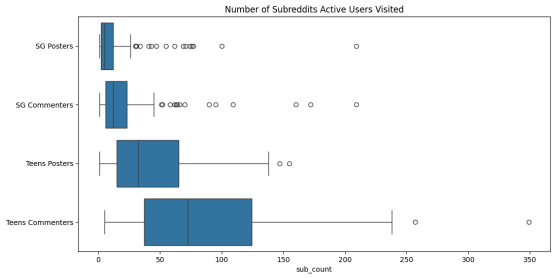


Figure 9

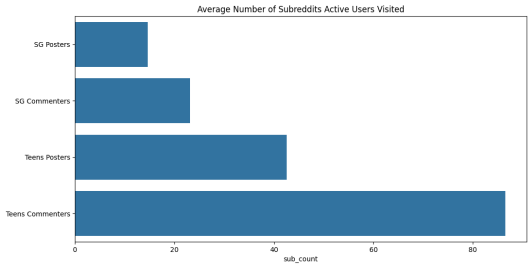


Figure 10

After adjusting the above statistics for activity levels by the formula sub_adjusted = sub_count/activity with respect to each individual user, we get the following plots.

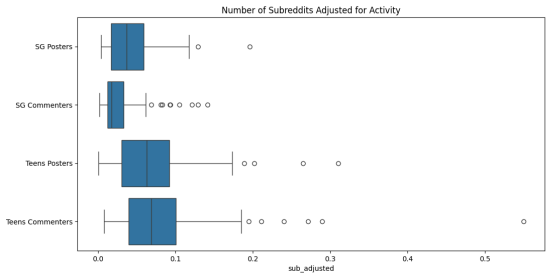


Figure 11

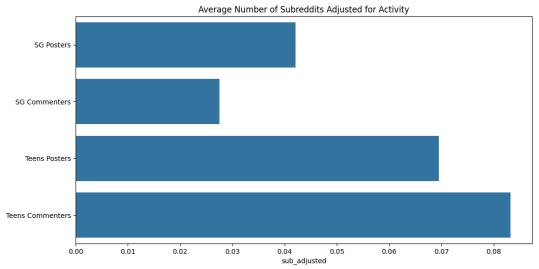


Figure 12

5 Discussion

5.1. Topic Attention Span

5.1.1 Comparing active posters and active commenters

We analysed the activity of a similar number of active posters and commenters in r/SGExams, 109 and 112 respectively. An equally sized sample was randomly drawn from active users on r/teenagers.

In r/SGExams, the proportion of commenters who never shifted topics (48.2%) is smaller than that of posters (63.3%). In r/teenagers, the proportion of commenters who never shifted topics (7.14%) is slightly smaller than that of posters (16.7%).

For both home subreddits, the red and blue lines in Figures 3 and 4 are very close, leading us to conclude that the interest shift patterns of posters and commenters are, for the most part, homogenous. This means that active users in r/SGExams and r/teenagers can be viewed as 2 different groups instead of 4.

5.1.2 Comparing active users in r/SGExams and r/teenagers

Figures 1 and 2 were randomly selected from pairs of r/SGExams and r/teenagers that have the same number of activity bins. A stark difference in the number of drifts and shifts is observed.

Comparing Figures 3 and 4, on average, active users in r/SGExams are less prone to topic shifts compared to their peers in r/teenagers. 55.7% of the Singaporean students never shifted topics, while the same can be said for only 10.7% of the international sample.

We attribute this drastic contrast in attention commitment to several factors. On average, Singaporean students spend less time on Reddit. Thus, it is expected that they would prefer a

one-stop-shop over switching between other subreddits and topics.

Secondly, Singapore's centralized, cookie-cutter school system and its accompanying struggles and stress are prime topics that students bond over, enabled by platforms such as Reddit.

Thirdly, r/SGExams is a village-like community with its own culture and trends which draw users in because of how familiar it feels.

5.2 Topic Interest Trends

From Figure 6, Singaporean students spend more time in r/SGExams than teens in r/teenagers. This is linked to the observation that r/SGExams is a community where users can get their daily doses of humour and advice, which accounts for the distinct differences between r/SGExams and r/teenagers at the Humour and Advice bars.

In line with expectations, ComicGame subreddits accounted for about 10% of all activity in the dataset.

5.3 Limitations and Improvements

With respect to the analysis of user interests, a key omission in the available data and consequently our resultant dataset is the inability to observe the subreddits, posts, and comments that each user has viewed.

Concerning the subreddit classifier, future studies may consider concatenating each subreddit name to its description for potentially increased accuracy.

Conclusion

In this study, we collected the largest available dataset of activity by Singaporean adolescents on Reddit. We developed a classification model to divide subreddits into 20 different topics and analysed the different topic interests and topic-shifting behaviours of various subpopulations.

As a pioneer study of Singaporeans on Reddit, this paper will certainly leave readers with more unanswered questions than they had. Given the increasing prevalence of digital social networks, we hope that this project will inspire future studies on the online behaviours of Singaporean students and that our datasets will contribute to these studies.

Acknowledgements

This project was supported by Nanyang Technological University under the URECA Undergraduate Research Programme.

I would like to express my gratitude to Asst Prof Farhan Ali for his consistent and contributive mentorship, without which this study would not have been possible.

References

[1] Retrieved from reddit.com/r/SGExams/about on 28 May 2024
[2] Retrieved from reddit.com/r/teenagers/about on 28 May 2024
[3] Retrieved from reddit.com/r/SGExams/comments/1bhs4lr/so_many_relationship_posts/ on 28 May 2024
[4] Characterizing the Evolution of Users Interests on Reddit, Valensise et al. (2019)
[5] RedDust: a Large Reusable Dataset of Reddit User Traits, Tigunova et al. (2020)

Appendix

Github link

All of the Jupyter notebooks written and used for this study can be found in this repository: https://github.com/OSA7JIMI/genz_kampung/tree/main.

Topic Definitions

15 topics defined by [4]
Some names were truncated and/or concatenated for ease of reference
Other changes made to the original definitions are styled in *italics;* formatted as *additions* or [*removals*]

Sport: Subreddits collecting discussions about sports or supporting teams or athletes.

FoodHealth: Subreddits collecting discussions about food and related issues as well as health and wellbeing related problems.

ComicGame: Subreddits collecting discussions about comics and games both online and offline.

NewsPolSoc: Subreddits collecting discussions about news, politics (regardless the party involved)

and societal issues such as migration or abortion and religion/belief systems

SciTech: Subreddits collecting discussions about hard sciences and technology from computers to cryptocurrencies

Advice: Subreddits collecting discussions about emotional issues and providing a forum to share experiences and receive advice

Humour: Subreddits collecting humoristic and funny objects including memes

EntMedia (originally: Books, Movies, Music): Subreddits collecting discussions about entertainment media, such as books, movies (including tv series and shows), *and pop music*

ImagVid: Subreddits dedicted to collect images and videos including GIFs without a specific topic

FashLife: Subreddit collecting discussions about fashion and lifestyle

Stories: Subreddits collecting stories from users regarding their daily experience and everyday life that are not related to existential problems or emotional issues or the request of support

Howbbies: Subreddits dedicated to guide and support users in technical problems or for discussing their hobbies/passions

Arts: Subreddits dedicated to discussion about soft sciences including History, Philosophy etc

Places: Subreddits dedicated to [*pictures and videos of*] specific places/locations

Others: Subreddits that cannot be classified in *all other* categories

5 topics added after looking over the datasets and considering the context of this study:

Celebrity (Celeb): Subreddits dedicated to specific people or groups; added because a significant number of celebrity-specific subreddits were observed

School/Educational (Sch): Subreddits dedicated to education systems, individual schools or the general pursuit of academic knowledge

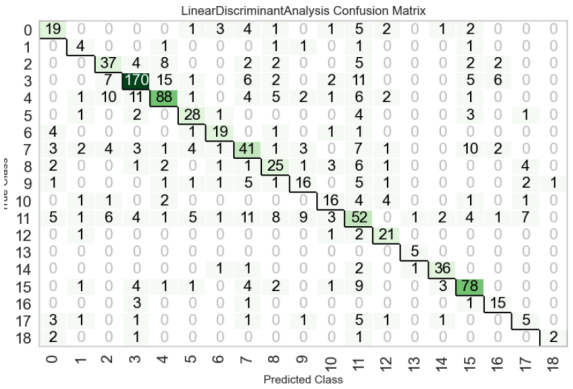
Teens: General purpose subreddits dedicated to teenagers

Singaporean (SG): Subreddits dedicated to Singaporeans and Singapore

NSFW: Subreddits marked as NSFW by Reddit, indicated by the is_nsfw property on PRAW

Subreddit Classifier

Confusion matrix of the initial LDA model: each number represents a topic (sorted in alphabetical order, excluding NSFW; for example, 0 = Advice, 1 = Arts etc)



Performance indices of the tuned LDA model

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.6480	0.0000	0.6480	0.6602	0.6462	0.6103	0.6113
1	0.6840	0.0000	0.6840	0.6917	0.6814	0.6504	0.6512
2	0.6400	0.0000	0.6400	0.6623	0.6429	0.6026	0.6039
3	0.6440	0.0000	0.6440	0.6452	0.6404	0.6036	0.6043
4	0.6426	0.0000	0.6426	0.6531	0.6407	0.6039	0.6049
5	0.6546	0.0000	0.6546	0.6742	0.6548	0.6180	0.6187
6	0.6104	0.0000	0.6104	0.6239	0.6082	0.5681	0.5691
7	0.6466	0.0000	0.6466	0.6655	0.6477	0.6090	0.6099
8	0.6225	0.0000	0.6225	0.6405	0.6248	0.5827	0.5837
9	0.5904	0.0000	0.5904	0.6151	0.5958	0.5474	0.5485
Mean	0.6383	0.0000	0.6383	0.6532	0.6383	0.5996	0.6005
Std	0.0243	0.0000	0.0243	0.0218	0.0228	0.0268	0.0267

Confidence score stats for LDA predictions

```
pred['prediction_score'].describe()
```

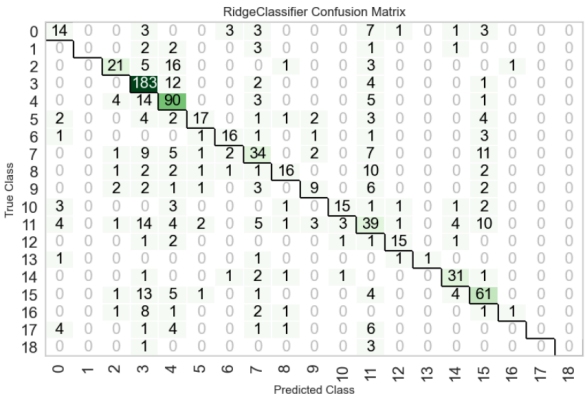
```
count    5543.000000
mean        0.894725
std         0.158562
min         0.262300
25%         0.839250
50%         0.986200
75%         0.999900
max         1.000000
Name: prediction_score, dtype: float64
```

```
names['confidence'].describe()
```

```
count    1238.000000
mean        0.867605
std         0.159073
min         0.330000
25%         0.786325
50%         0.940350
75%         0.994000
max         1.000000
```

- Include some examples from subreddit-topic table

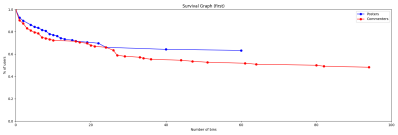
When only subreddits with at least 50 character long descriptions are used, the training data is missing some topics

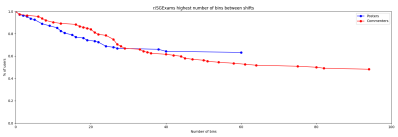


Drift / Shift

Other topic attention span survival graphs: max bins between shifts and number of bins before first shift

r/SGExams





r/teenagers