WHAT CHARACTERISTICS OF A POST ON REDDIT ARE MOST PREDICTIVE OF THE OVERALL INTERACTION ON A THREAD (AS MEASURED BY NUMBER OF COMMENTS)?

BY

IZUNNA OHIA

Using Natural Language Processing as a tool to extract insight from a post text

This project focuses on the level of overall interactions on a thread by predicting the number of interactions and classifying such interaction as high or low with respect to the median interactions(As measured by number of comments). The quality of a post title, time post lasted, subreddit associated to a particular post title can often make or break the popularity of the submission

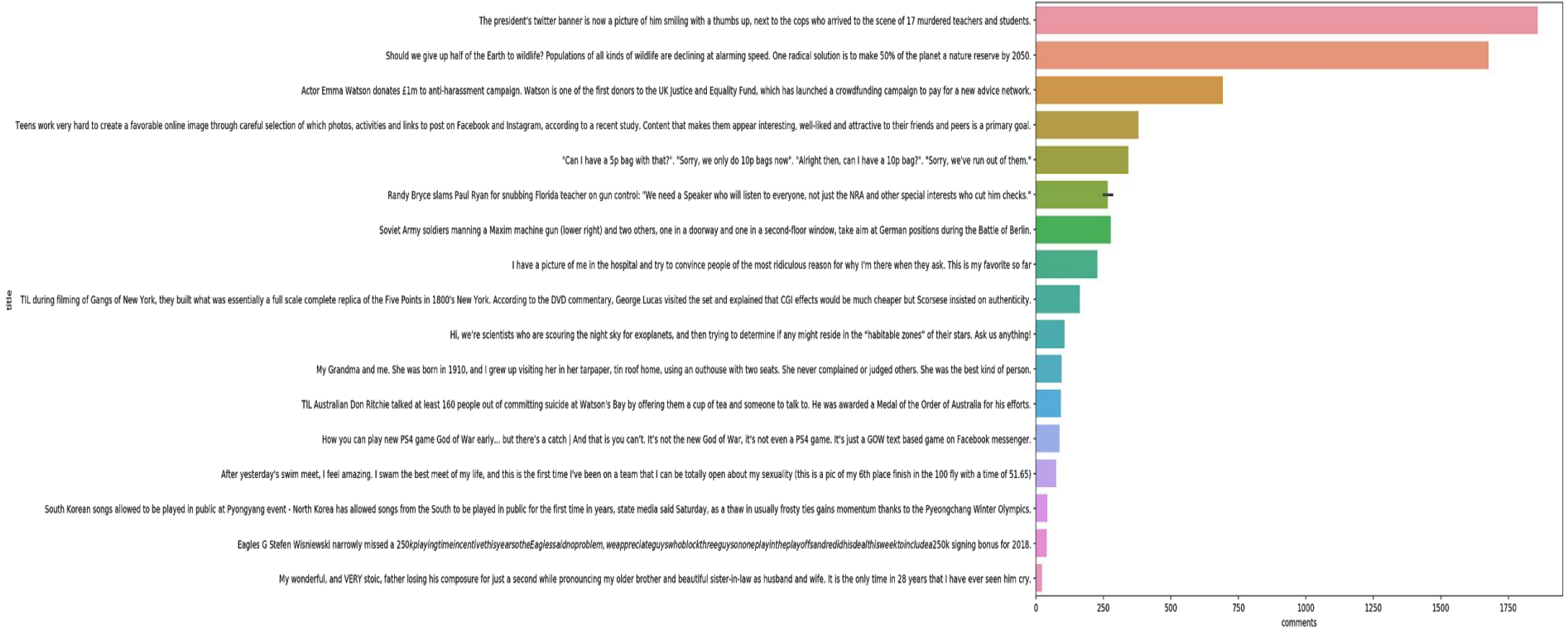
So the question becomes “What in a thread post attracts high or low interactions (Many or Low number of comments) on reddit.com?

1. Is it the length of the thread text?

So in order to accept or reject this hypothesis I sorted my data by number of comments. The thread title with most number of words had 28 comments, the second title had 11 comments, the earlier lasted for 7 hours while the later lasted for 3 hours

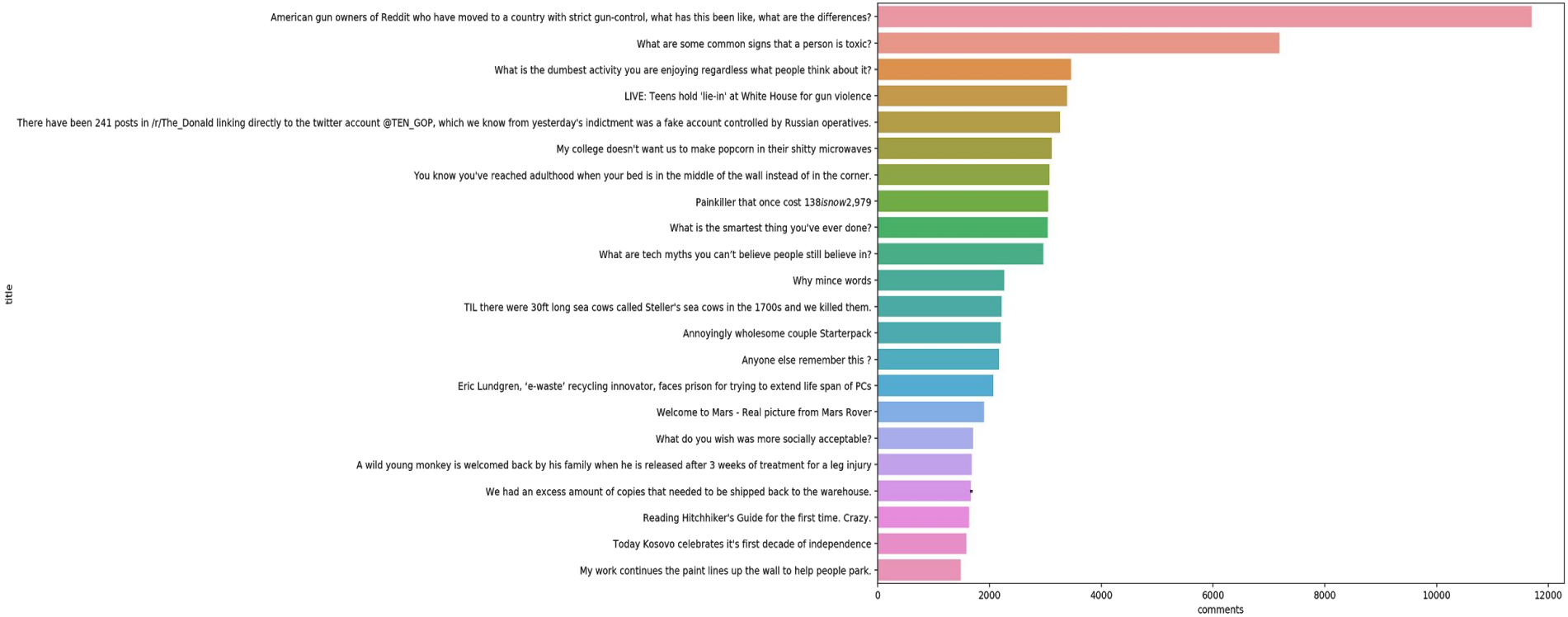
A bar plot of titles with greater than or equal to 34 words against their number of comments

1.a



A bar plot of titles with less than 34 words against their number of comments

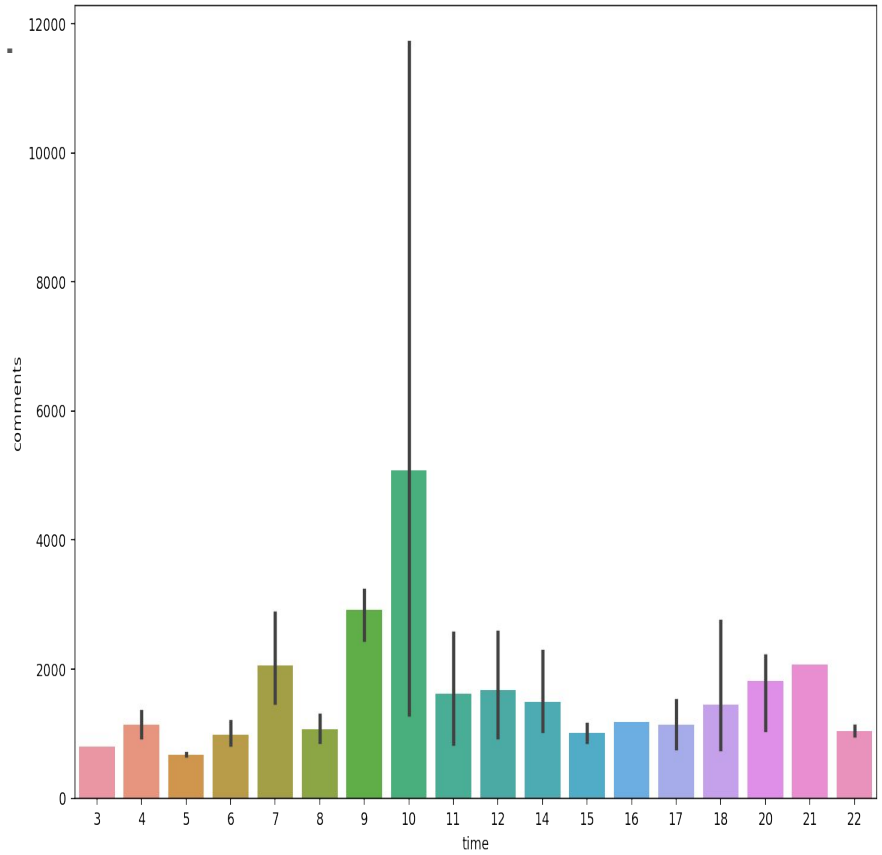
1.b

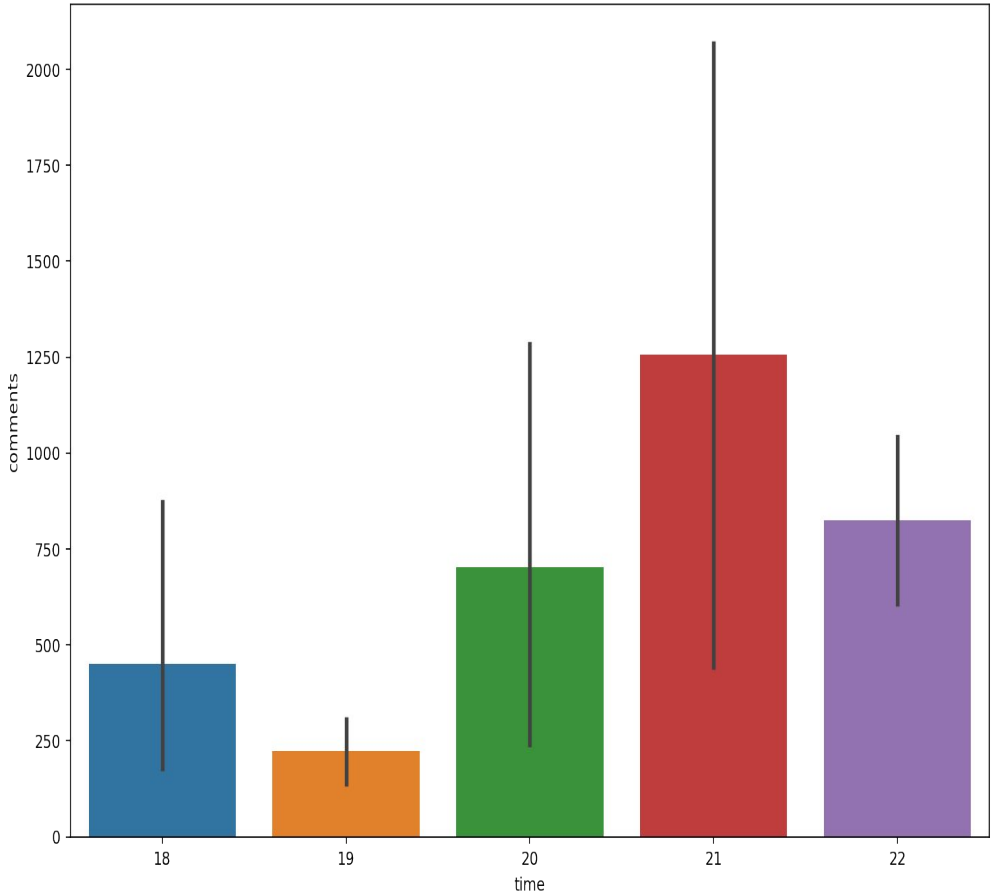


2. Could it be the entire time the post stayed up on reddit prior to scraping?

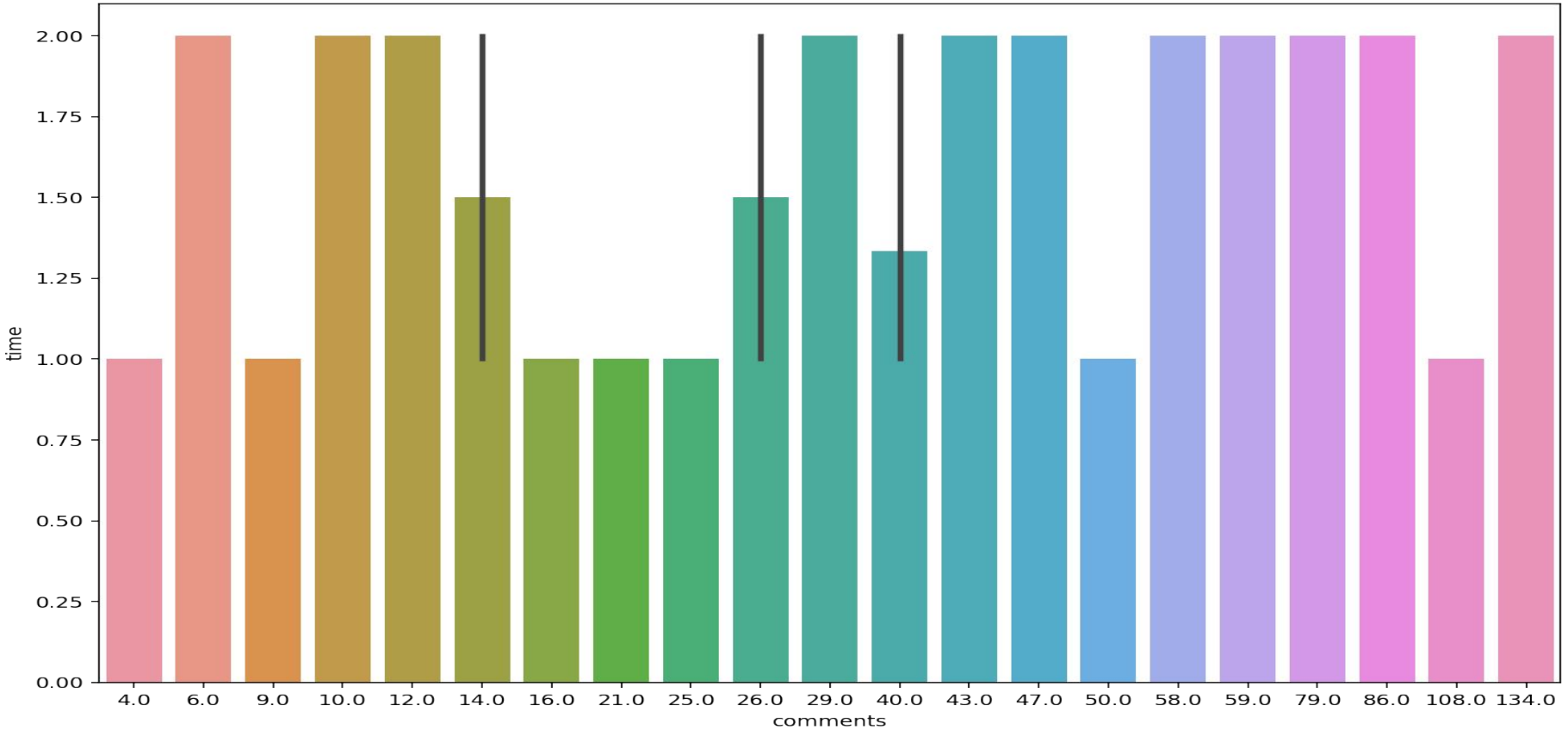
Titles with the most time prior to scraping all lasted for 22 hours with number of comments ranging from 608 to 1108 While titles with least time prior to scraping all lasted for 1 hour with number of comments ranging from 4 to 108

Graph showing the most time and their number of comments





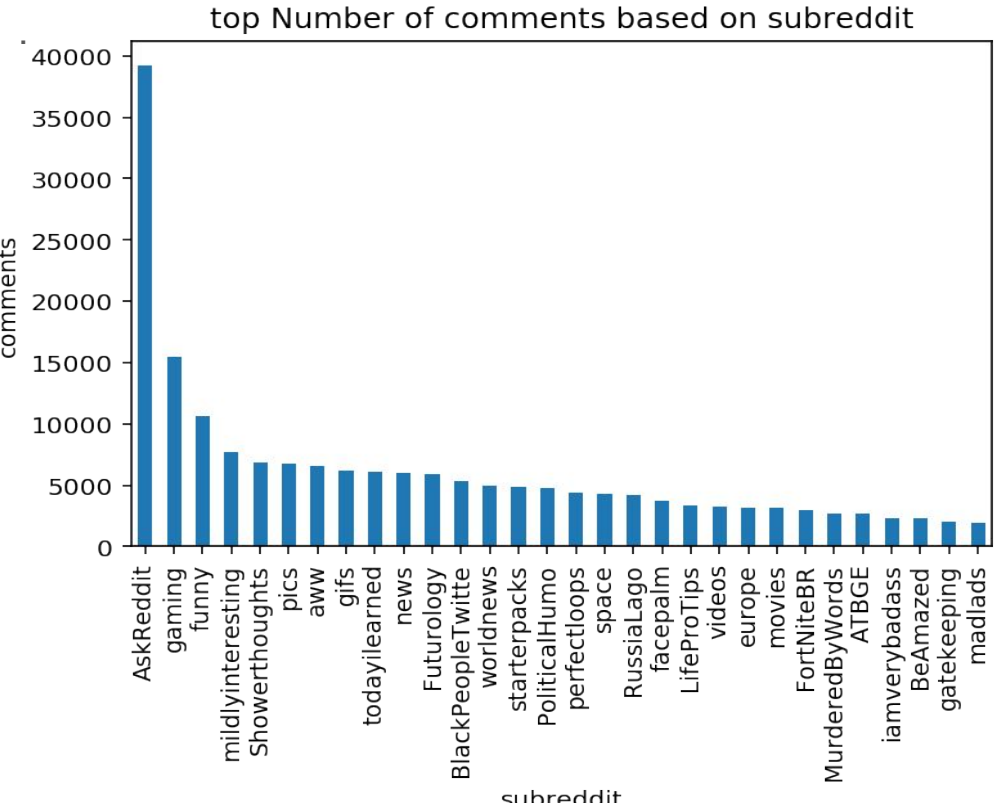
Graph showing the least time and their number of comments

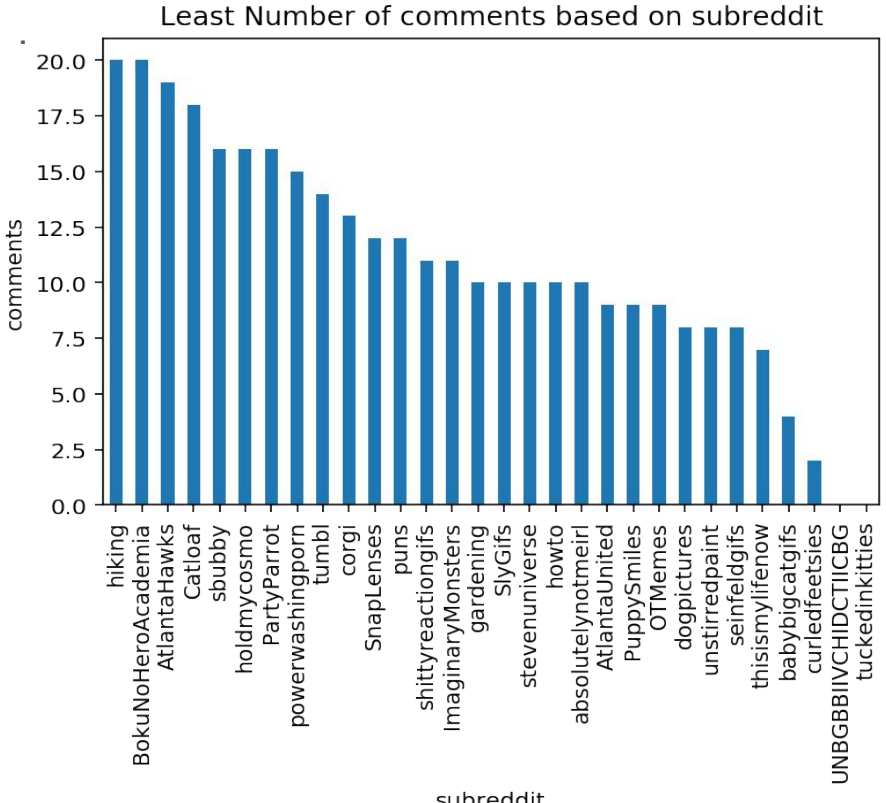


3. Could it be the subreddit associated with the particular thread title?

Here the top five subreddits associated to each thread title with the most comments on aggregate all have number of comments associated to their titles ranging from 6,862 to 39,201, While the least five have 0 to 7 on aggregate.

Graph showing the top and the least subreddit based on number of comments on aggregate associated to their respective thread titles

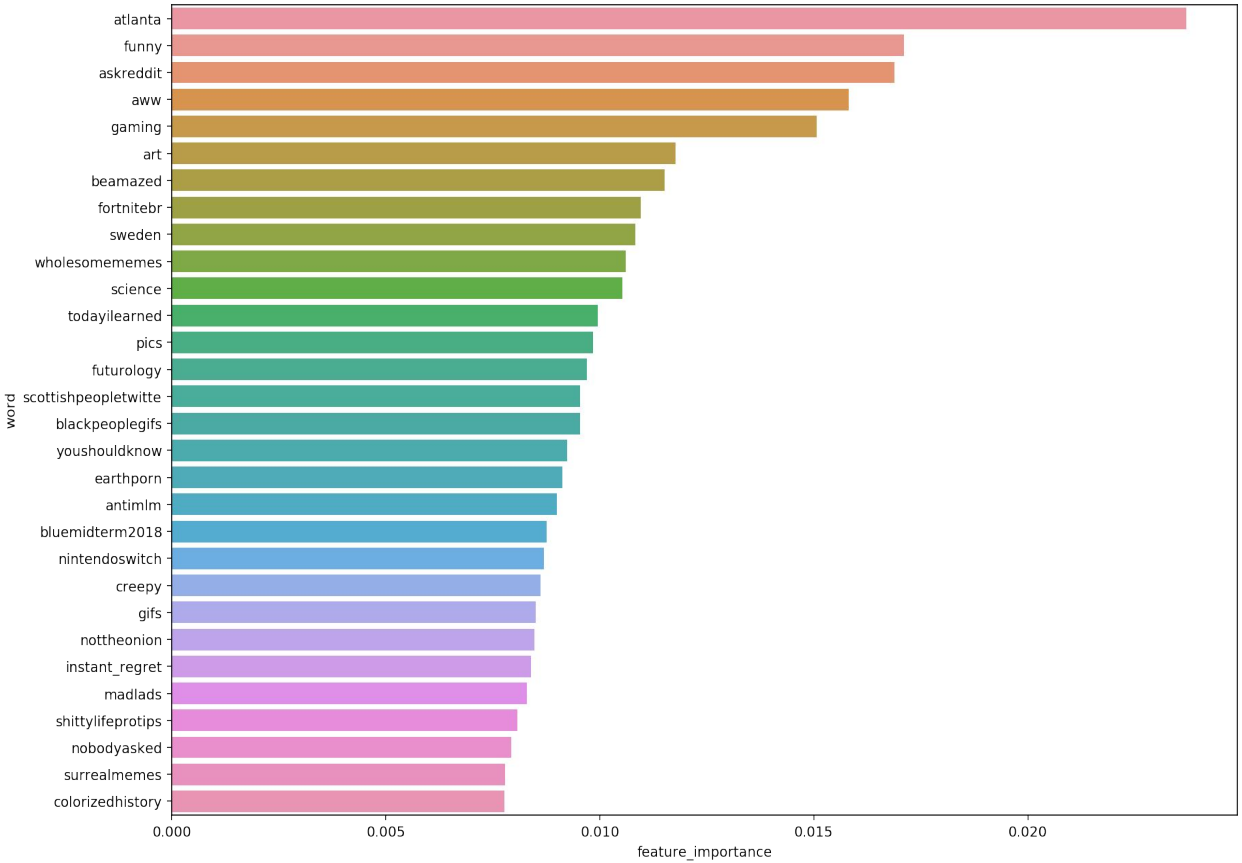




A bar graph showing predictor variables based on feature importance (subreddit).

Using Random Forest Classifier to predict and classify as high or low the median number of comments with a univariate predictor (Subreddit)

The model achieved a training accuracy of 85% and a test accuracy of 67%. This model is good at predicting “Low” than “High”.



Confusion matrix

This table shows **True Positive** of 89 which is the number of predictions that the Random Forest Classifier got right that were above the median number of comments While 35 was the number of prediction that it predicted to be above the median number of comments that were wrong (**False Positive**). Also it predicted correctly the number of **True Negative** as 93 Which is the number of comments that are below the median number of comments but also predicted 55 comments wrongly as the number of comments that are below the median number of comments when they are actually the number above the median number of comments **False Negative.**

Predicted Actual	0	1	ALL
0	93	35	128
1	55	89	144
ALL	148	124	272

4. Could it be the contextual structure of the thread title?

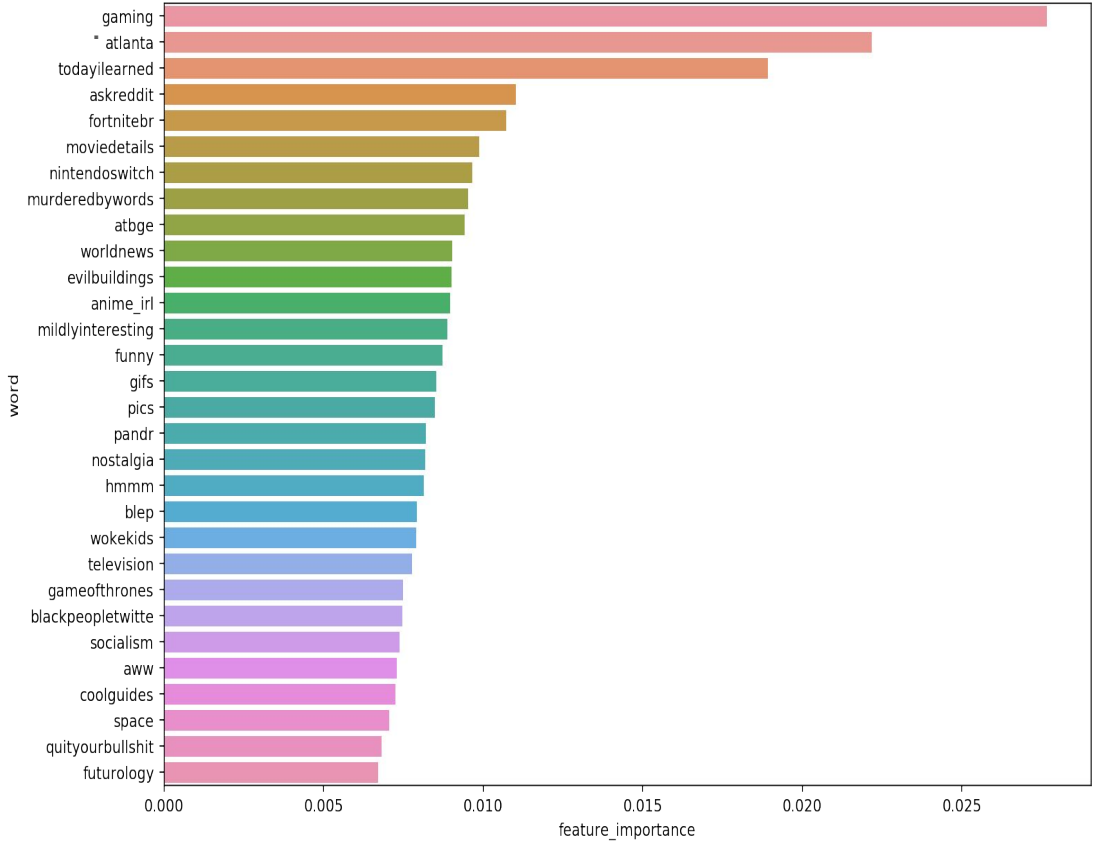
New variable based on if some certain word exist in the thread title

I wrote a function that return 0 if “twitter” or 1 if “Gun” or 2 if “Russia” or 3 if “Donald” or 4 if none was found in the thread title. Using a multivariate predictor resulted to the below outcome.

A bar graph showing predictor variable base on feature importance (subreddit, New_variable).

Using Random Forest Classifier to predict and classify as high or low the median number of comments with a multivariate predictor

The model achieved a test accuracy of 70%. This model is good at predicting “Low” than “High”.

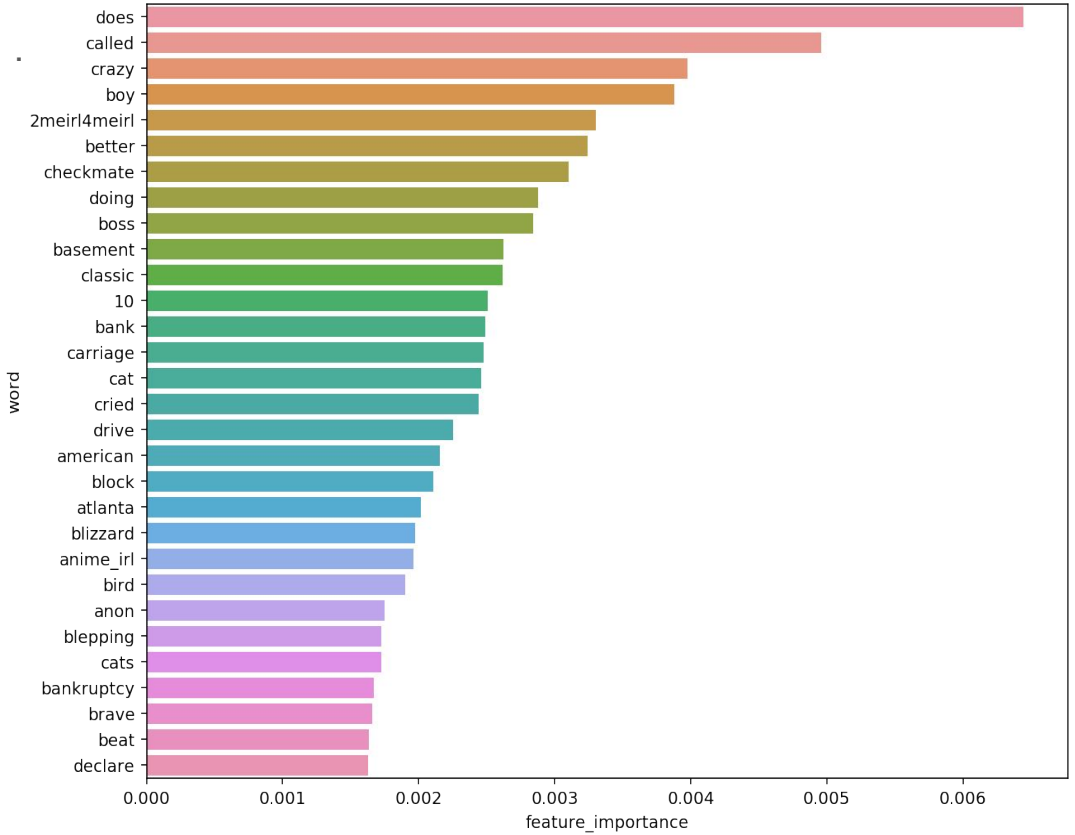


Having thread title, subreddit and the class variable as predictors

Using Random Forest Classifier to predict and classify as high or low the median number of comments with 3 predictors

The model achieved a test accuracy of 69%. This model is good at predicting “Low” than “High”

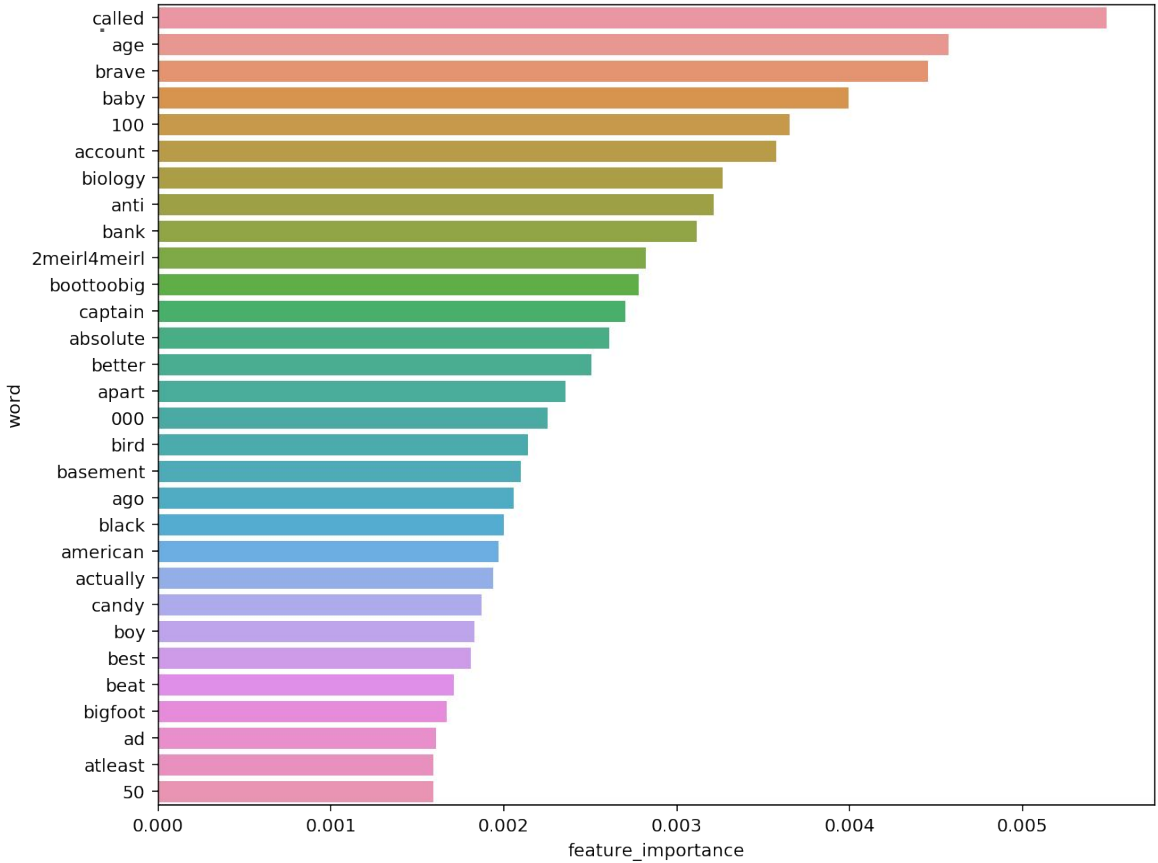
Using Cross-Validation on this model resulted in accuracy of 63% for 5 folds and 65% for 6 folds.



Using thread title as predictor for a 70/30 train/test_split

Using Random Forest Classifier to predict and classify as high or low the median number of comments with a univariate predictor (thread title)

The model achieved a test accuracy of 69%. This model is good at predicting “High” than “Low”



Conclusion

. I should have considered using the thread title with the most comments within a certain time frame and least comments within a certain time frame as control.

. The variables used do not totally reflect all possible factors that are most predictive of the overall interaction on a thread.

.Some subreddits with high feature importance are r/AskReddit, r/Atlanta, and r/aww. This is expected because these subreddits have tokens that are unique to their posts. r/AskReddit is mostly a place to ask and answer thought-provoking questions and often contains the token ”?” at the end of a post. Some like r/news, contains token ”LIVE” at the front of the post and it contains posts with all kinds of news. r/RussiaLago, Which is talking about Russia, that actually makes sense considering the political climate in the country now.

Future work should consider the incorporation of additional features such as;

1. Scraping more datas over a long period of time

2. More subreddits with the same number of thread title associated with them at random in a different posting time frame over a period of time can be scraped.

3. Using n-grams as inputs

4. Finally, hyperparameter tuning can also be optimized using Bayesian methods, which is significantly better than the Cross Validation method used.