Project 3

Quantifying Reddit

By Evgeny Didenko

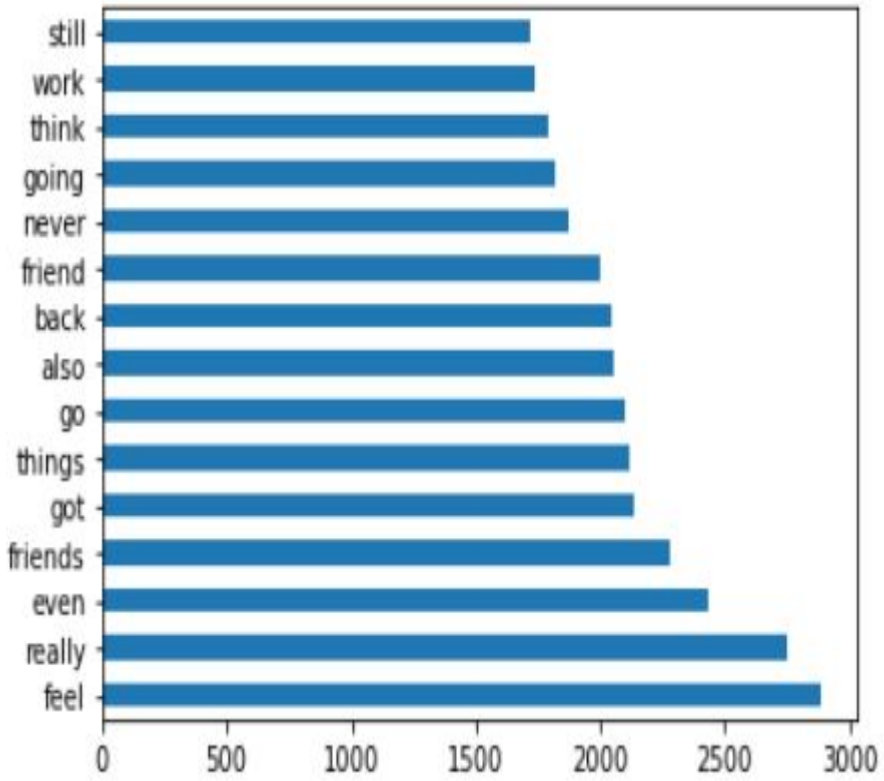
Overview

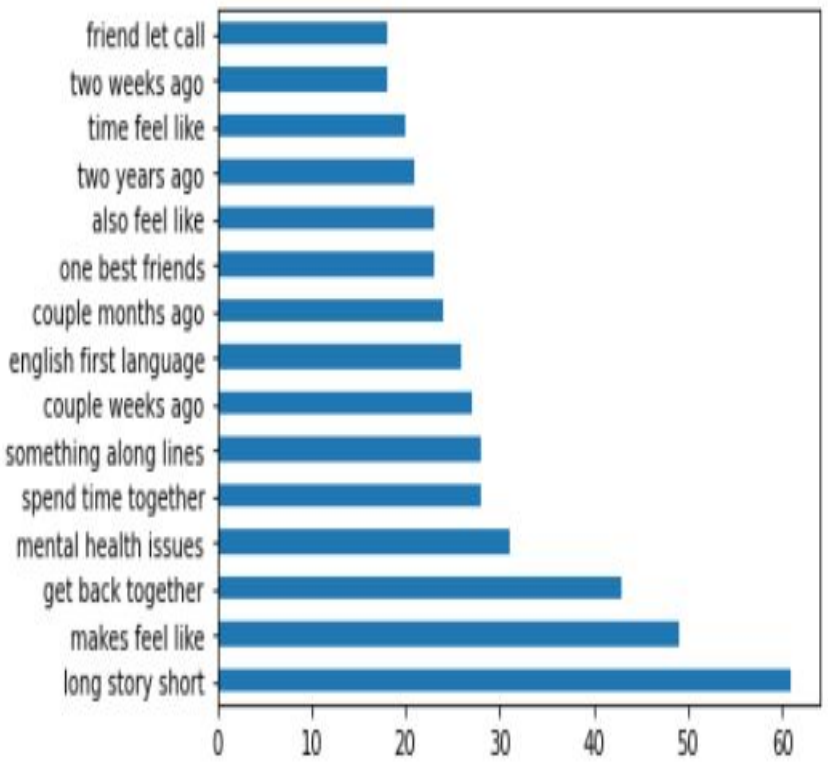
- The goal is to investigate how capable are the basic classification functions at distinguishing text data from two subreddits of similar general topic.
- The subreddits are "Relationship" and "Am I the Asshole", they both cover stories of conflict in interpersonal relationships and they are both very text heavy.
- Generally very large portion of Reddit is links, pictures or video clips, I don’t think we have the tools to analyze these yet.

Procuring the Data

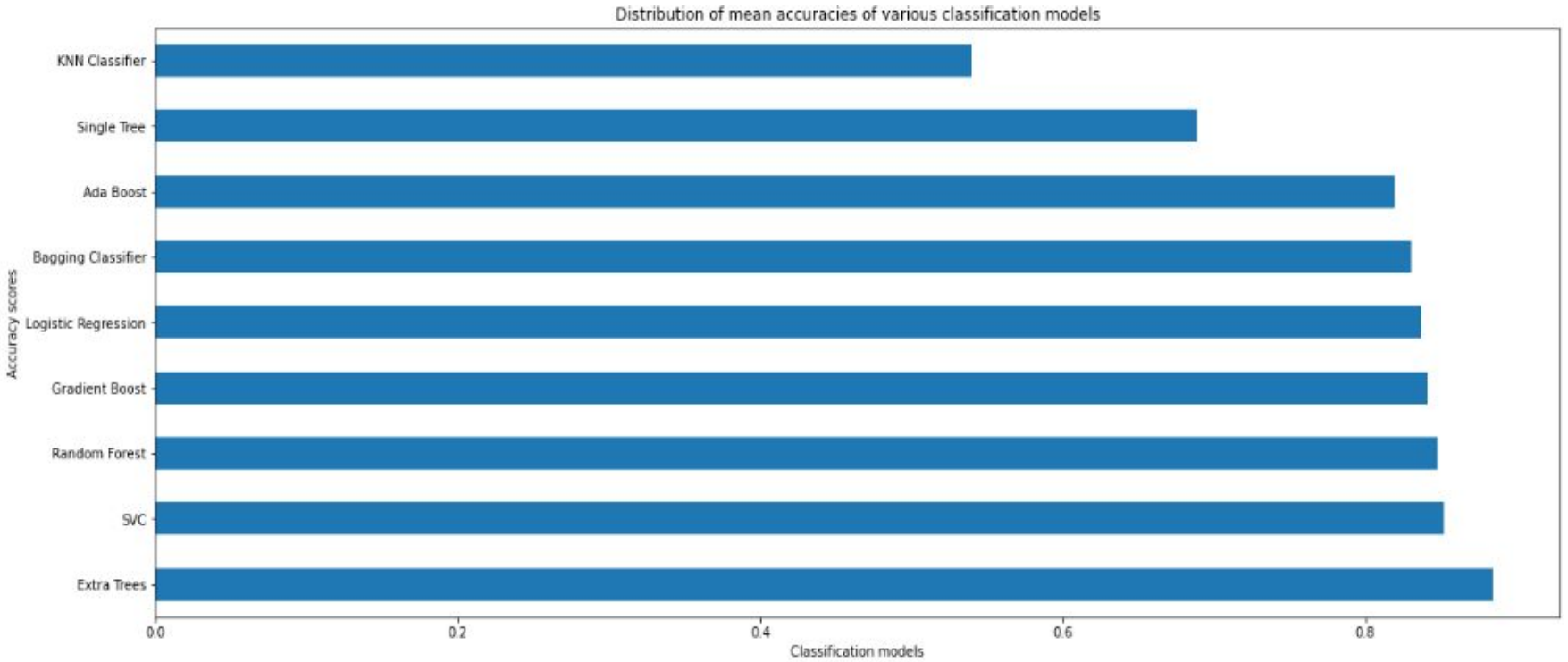
- We were given a link to API capable of downloading data from reddit as well as detailed instructions one how to use it.
- This API does not like handling too many posts or too many requests at once, so a loop with 2 second delay between cycles was implemented.
- Each loop after first used the date of the first (chronologically) post as the starting point, so although there were some duplicated, vast majority of posts was unique.
- Relevant (just the subreddit name, post title and post text) data from each subreddit was fed to a separate dataframes that were merged afterwards.

EDA


still
work
think
going
never
friend
back
also
go
things
got
friends
even
really
feel
0
500
1000
1500
2000
2500
3000


friend let call
two weeks ago
time feel like
two years ago
also feel like
one best friends
couple months ago
english first language
couple weeks ago
something along lines
spend time together
mental health issues
get back together
makes feel like
long story short
0
10
20
30
40
50
60

Models


Distribution of mean accuracies of various classification models
KNN Classifier
Single Tree
Ada Boost
Bagging Classifier
Logistic Regression
Gradient Boost
Random Forest
SVC
Extra Trees
Accuracy scores
0.0
0.2
0.4
0.6
0.8
Classification models

Questions