Notes

Emily Maloney

February 10, 2019

Data Plan

1. score subreddits based on political ideology:
    a) VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text Pros - seems legit Cons - python
    b) https://maxcandocia.com/article/2017/Jan/05/analyzing-politics-of-reddit-part-1/ Pros - network stuff can be done in r Cons - not a score but just arrangement of political subreddits
    c) sentimentr Pros - R, not extremely difficult Cons - positive/negative only informaiton
    d) creating a dictionary ourselves based on entire corpus?
    e) doing a text net to find central users and get their sentiment score?
    f) probably should talk to Chris and/or Tom/Danielle
2. create dictionaries of conversion narrative terms
    a) previous lit?
    b) just using a few words and then tf-idf the "archetypical" ones that come up?
3. identify conversion narratives in subreddits
    a) topic model of these
    b) differences by party (specifically references to past self) or gender
4. Identify people who interacted with conversion narratives (commented, posted themselves?)
5. See if those people had a change in behavior (directly?) following first interaction with cn:
    a) sentiment (more/less positive and or liberal/conservative and or extreme)
    b) started posting/commenting more in more extreme subreddits (why scoring of subreddits is an especially important decision)
6. Try to classify race, gender, age, etc. on users from text as well?

Questions

1. What to do about missing data - e.g. deleted posts and people who quit reddit?
2. Want to collect super recent participation? (probably would need to use the PRAW python package to do so, which I feel semi-comfortable using but not 100%)
3. What time range? Just around election?
4. How to control for other media consumption? I.E. how will we attempt to account for omitted variables like offline behavior and behavior on other websites?
5. Probably need to come up with a good data storage and cleaning routine, not entirely sure how to do so tbh (soc server and dnac server are possibilities but I don't know how to use them yet, Aidan knows I think)
6. IRB application - should probably start working on so it's complete before we begin to systematically collect data