Smelling the Covfefe: A Computational Analysis of User Responses to the Banning of r/The_Donald

Introduction:

On 29th June 2020, social-sharing site Reddit took the unprecedented step of blocking access to the alt-right subreddit r/The_Donald in response to consistent violations of the platform's code of conduct. This project seeks to fill a gap in the literature for empirical study of the effects of such deplatforming measures on user activity in the post-2016 digital information ecology; it uses computational methods of causal inference and temporal analysis to address the following questions:

- RQ1: What effect did Reddit's ban have on former contributors to r/The_Donald?
  - RQ1a: (How) were their activity levels affected?
  - RQ1b: (How) did their hate speech usage change?

- RQ2: What effect did the ban have on subreddits which experienced an influx of former r/The_Donald users?
  - RQ2a: To which subreddits did r/The_Donald contributors migrate following the ban?
  - RQ2b: Did hate speech usage by these migrants change?
  - RQ2c: Did hate speech usage by pre-existing contributors to these subreddits change?

Context:

Previous analysis of hate-speech regulation has primarily focussed on sites such as Facebook and Twitter which operate 'industrial' moderation strategies; Reddit's preference for an idiosyncratic 'community-led' approach restricts the extent to which findings from such studies can be considered relevant.[1] Where Reddit-related analysis *has* been published, this has usually followed small communities such as r/FatPeopleHate;[2] moreover, given the rapid pace of change in digital and political cultures, we might question such studies' currency.[3]

Datasets, Methodology and Timelines:

I anticipate that each of the following phases will take one-and-a-half weeks to complete.

Phase 1: Constructing a dataset and lexicon of hate-speech

- This project will draw on raw data obtained from Baumgartner's open-source repository of all 3bn+ historic posts on the Reddit site.[4] Google's SQL-based BigQuery platform will be used to extract from this all text submitted to r/The_Donald in 2020, which will serve as a corpus for the construction of a hate-speech lexicon.

---

[1] Caplan, Robyn. Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches. New York: Data and Society, 2018. Accessed 17th January 2021. https://datasociety.net/library/content-or-context-moderation/

[2] Squirrell, Tim, "Platform dialectics: The relationships between volunteer moderators and end users on reddit," New Media and Society 23, no. 9 (2019): 1910-1927.

[3] Chandrasekharan, Eshwar, Pavalanathan,Umashanthi, Srinivasan, Anirudh, Glynn, Adam, Eisenstein, Jacob and Gilbert, Eric, "You Can't Stay Here," Proceedings of the ACM on Human-Computer Interaction 2, no.31 (2017): 1-22.

[4] Baumgartner, Jason. Directory Contents. pushift.io, 2020. Accessed: 17th January 2021. https://files.pushshift.io/reddit/

- The corpus will be run through an automated keyword identification tool in accordance with the BigQuery protocol book.[5] A baseline comparison will be conducted against a random sample of Reddit posts by using Python SAGE implementation to identify the 100 words with the highest SAGE coefficients in the random sample and comparing these to the r/The_Donald lexicon to identify words with especially high normalised incidence. The automatically generated lexicon will then be manually inspected to remove any non-hate-oriented terms.

Phase 2: RQ1

- The r/The_Donald timeline will be used to generate a treatment group of accounts by mining the handles of a systematic sample- ordered by post volume- of 20 users from those who have submitted to the subreddit eight times or more.
- A control set of users will be constructed by selecting 50 alternative subreddits where treatment users contributed the highest portion of posts prior to 29/06/2020. R-based Mahalanobis Distance Matching will be used to identify contributors to these subreddits closest to the treatment users.[6]
- RQ1a) Using 29/06/2020 as the origin, relevant user-timelines will be segmented into pre- and post-ban periods divided into 15-day intervals. The number of posts made by control and treatment users per interval will be aggregated and plotted on a line graph.
- RQ1b) The frequency with which words from the hate lexicon were posted by treatment and control users each 15-day period will be observed, then normalised by the total number of words submitted. Mean values for each group per interval will be counted and plotted on a line graph.

Phase 3: RQ2

- RQ2a) Using interrupted time-series analysis, all posts in the timelines of the treatment users will be analysed, and subreddits where their activity increased by >100% following the ban will be tabulated.
- RQ2b) Changes in the submission frequency of words in the hate-speech lexicon from both groups, normalised by the total number of words in all the individual's posts, will be calculated. Mean values will be plotted as a line-graph.

Phase 4: Presenting findings in a report and poster.

- If time permits, data visualisation tools such as Halfviz and RAWGraphs could be used to provide a richer cartography of the subreddit's ecology.

<u>Ethics:</u>

The usual ethical approval process will suffice for this project, though precautions should be taken to anonymise and user information (user-handles, for example) in the final project products.
Content warnings will be provided on all final products and at any oral presentations since sensitive and/or offensive language is likely to arise.

[5] Brautmartner, Jason. Using BigQuery with Reddit Data, pushift.io, 2020. Accessed 17th January 2021. https://pushshift.io/using-bigquery-with-reddit-data/

[6] Rubin, Donald, and Stuart, Elizabeth, "Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions," The Annals of Statistics 34, no. 4 (2006): 1814-1826.