Contents

Abstract

Focusing on the good faith debate subreddit r/ChangeMyView, this paper utilizes historical information on users who submit their opinion in this debate space, in order to see whether this information can be used to predict whether said users will end up changing their opinion in the face of persuasion. Some user

history is evaluated under the psychological theory of self-affirmation, in an attempt to further test the theory's viability in an online context. Additional user information, not directly applicable to reciprocity or self-affirmation, such as notions of subreddit diversity and equality of participation is utilized as a means of classing similar users. This paper finds some evidence of the applicability of self-affirmation theory to online debate spaces, the relevance of other Reddit-specific indicators, and other differences between those with malleable and static opinions.

Introduction / Literature Review

Persuasion

Persuasion, a goal in a number of settings, from political and marketing campaigns to friendly or professional conversations, has been the subject of significant research efforts (Tan et al. 2016; Dillard and Pfau 2002). Before the advent of social media websites like Facebook or Reddit, these research efforts were mostly confined to laboratory settings, but thanks to the increasing number of social interactions online, interpersonal persuasion has become observable on a massive scale (Fogg 2008).

Tan et al. explored persuasion in /r/changemyview (CMV), an area of the popular social media website Reddit (2016). CMV is particularly conducive to the study of mass interpersonal persuasion, as posters must state the reasoning behind their views, and successful arguments must be awarded with explicit confirmation. Thus, the outcome of the persuasion efforts, reasoning behind people's views, and the full interactions are accessible. Additionally, with the emphasis Reddit as a whole places on long-form content (Wood, Molly and Ainpour, Shaheen 2018), these interactions provide substantial data for training natural language processing models.

With access to this information, Tan et al. focused primarily on how interaction dynamics and choice of language within arguments were associated with a successful change in someone's opinion. A third focus of the study was an attempt to determine the malleability of an opinion, i.e. the likelihood that the holder of that opinion would award successful arguments to change it. Assuming that at least 10 unique challengers to the opinion were present, and that the holder of the opinion responded at least once, Tan et al. analyzed the way in which the opinion was presented and attempted to predict whether or not it could be changed.

This last task, attempting to determine the malleability of an opinion without respect to any of the arguments attempting to change it, was difficult indeed, and Tan et al. only achieved an ROC AUC of .54. Still,

using weighted logistic regression, they found some significant features consistent with self-affirmation theory (Cohen, Aronson, and Steele 2000; Correll, Spencer, and Zanna 2004).

Self-Affirmation Theory

> "[S]elf-affirmation theory . . . suggests that every person strives for positive self-regard and, to achieve it, draws on successes in important domains in her or his life. These domains constitute aspects of individual identity, including important social roles, abilities, and beliefs. Because the individual depends on a constellation of domains for feelings of adequacy, a threat to one of the domains can prompt a defensive reaction. A crucial tenet of self-affirmation theory, though, is that the ultimate goal of a defensive reaction is the security of the global sense of self-worth, no the security of the domain, per se. The individual should defend a given domain only to the degree that the more general sense of self-worth is compromised by its loss."

(Correll, Spencer, and Zanna 2004)

In psychology, self-affirmation, which can be thought to reinforce one's global sense of self-worth, has been found to indicate open-mindedness and make beliefs more likely to yield (Correll, Spencer, and Zanna 2004; Cohen, Aronson, and Steele 2000).

Tan et al. found that within the text of an opinion, the use of first person pronouns were strong indicators of malleability, but first person plural pronouns correlated with resistance.

> "[I]ndividualizing one's relationship with a belief using the first person pronouns affirms the self, while first person plurals can indicate a dilluted sense of group responsibility for the view."

(Tan et al. 2016)

While Tan et al. attempted to derive the level of self-affirmation present within the stating of an opinion, the user stating that opinion can have other sources of self-affirmation. Returning to Correll:

> "[I]f global self-worth is temporarily bolstered by success in a second, unrelated domain, the individual should be more willing to tolerate a threat to the domain of interest."

Looking at the wording of the opinion itself is a related domain, but it is reasonable to assume that if a Redditor has previous submission history, that some of that history is unrelated to the opinion they are presenting for change in CMV. Additionally, self-affirmation theory does not restrict the source of bolstering one's global self-worth; it can be self-affirmation or affirmation from third parties.

Thus, within past submissions one can look at the same features as Tan et al., first person singular and plural pronouns for self-affirmation, but also for features that are indicative of third party affirmation, like the score, given by other users, of the submission in question.

Nearly all of a Reddit user's past submission history is available for perusal. Exploring the affirming nature of this history allows a deeper testing of self-affirmation theory, as a lab experiment can only really test the history created within the lab settings itself. Cohen et. al (2000), for example. . .

> [The researchers] asked half of their participants to write a paragraph about an important value (to affirm their sense of self-worth) before exposing them to arguments that challenged their views on capital punishment or abortion. Compared with control participants who wrote about less important values, those who wrote about a central value were more willing to recognize the strengths of the challenging argument.

Utilizing a Reddit user's past submission history, on the other hand, not only allows a more extended look into instances in which self and third party affirmation may have occurred, but also provides NLP models data to quantify the "similarity" or "relatedness" between the past submission history and opinions made thereafter. Tan et al. extended self-affirmation theory as it was onto an online debate space. Utilizing an opinion author's past submissions could extend the relevance of self-affirmation theory beyond the most recent past, by looking further back in time for potential affirmation.

Data

Overview of Data Source: r/ChangeMyView

The popular social media site, Reddit, is composed of a variety of subcommunities, or "subreddits". One such subreddit, r/ChangeMyView (CMV):

> . . . is a subreddit dedicated to the civil discourse of opinions, and is built around the idea that in order to resolve our differences, we must first understand them. We believe that productive conversation requires respect and openness, and that certitude is the enemy of understanding.

Redditors can participate in CMV in one of two main ways. They can post their *opinions*, along with supporting reasoning, as *submissions*, or they can *comment* directly on a submission in an *attempt to change the opinion*. If a user posts a view, the subreddit's rules require them to respond to any comments, attempts to change the view, within three hours otherwise the post is removed. The rules of CMV also forbid low

effort comments, and an active team of moderators stringently enforce all of CMV's rules. If a respondent mangages to change the original poster's (OP) opinion, then the OP can award the comment that changed his or her mind a "delta", by replying directly to the worthy comment with a signal that a delta has been awarded (typically "!delta"), along with a brief explanation of why the comment changed their view. While uncommon, it is also possible for another user that is not the OP, who one might presume holds a similar view, to award deltas as well.

Before the delta is officially awarded to a respondent, a bot (*DeltaBot*) must confirm it. Assuming the OP's

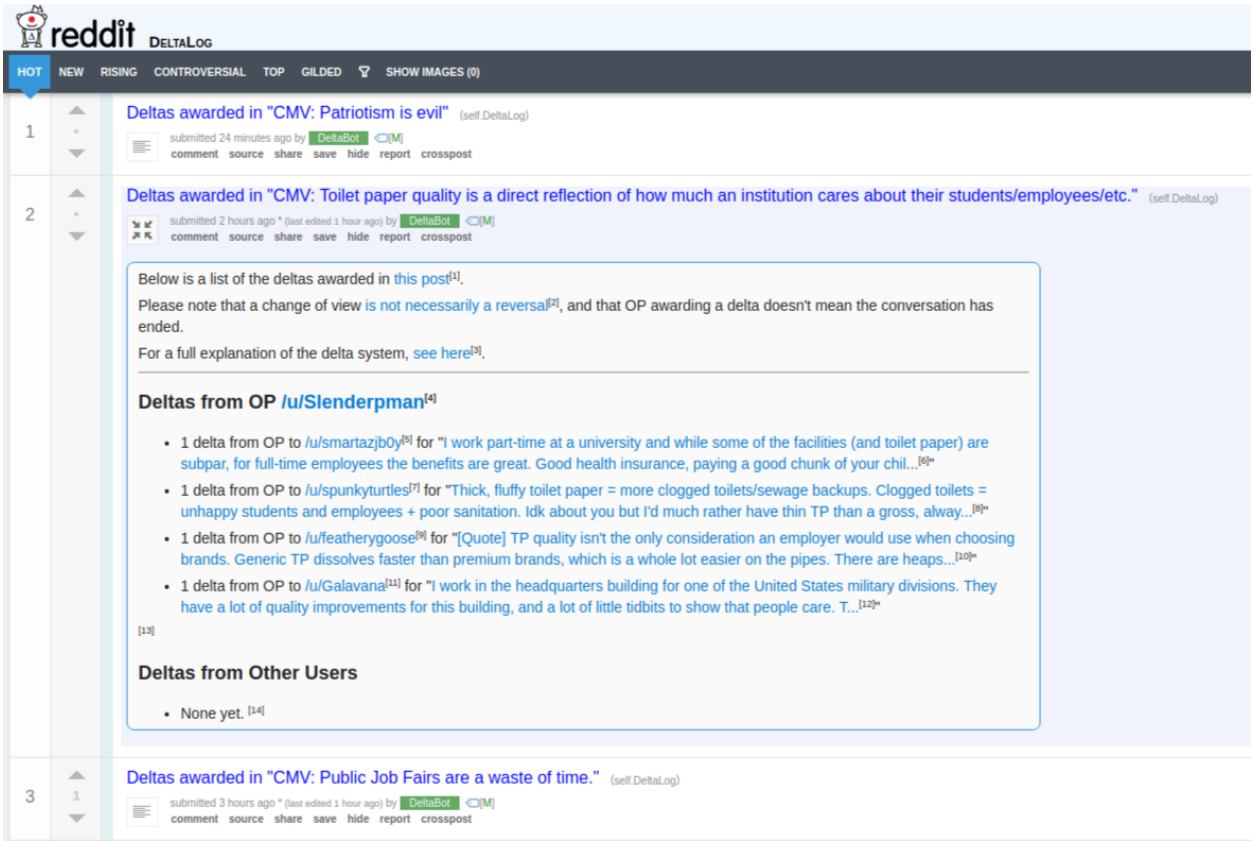


Figure 1: The Front Page of r/DeltaLog. Delta winning comments are sorted by submission.

explanation is long enough (a sentence or two), DeltaBot responds to the OP's delta signalling comment to confirm the award. After this confirmation, any comment that the user who received the delta makes, will have an updated total of deltas awarded to the right of their name. This confirmation also contributes to a record of deltas on a separate subreddit, r/DeltaLog, a browseable archive of comments that received deltas in r/ChangeMyView. Additionally, the wiki of r/ChangeMyView also features a Deltaboards section which tracks the users with the most deltas awarded daily, weekly, and monthly.

Thus, the CMV subreddit allows for access to the reasoning behind a person's views, the debate that takes place for each view, and an easily extractable outcome of the debate: either the opinion is stable and no delta is awarded, or the opinion changes, and at least one delta from the OP is awarded. CMV is an ideal setting for the study of persuasion (Tan et al. (2016)). There are many questions to be explored in CMV, but one that leverages the open nature of Reddit user histories is:

> "Using what we know about a user who submits a view to CMV, can we predict if they will change their opinion?"

Weekly

Rank	Username	Deltas
1	**SaintBio**	49
2	poundfoolishhh	48
3	TheManWhoWasNotShort	48
4	kublahkoala	7
5	callmesixone	6
6	Ansuz07	6
7	Iswallowedafly	5
8	GoIdfinch	5
9	Huntingmoa	5
10	MyUsernameIsJudge	5
	As of 4/13/18 19:56 UTC	

Monthly

Rank	Username	Deltas
1	**SaintBio**	50
2	poundfoolishhh	48
3	TheManWhoWasNotShort	48
4	AnythingApplied	19
5	kublahkoala	12
6	MyUsernameIsJudge	11
7	bguy74	9
8	Polychrist	8
9	boundbythecurve	7

Figure 2: Weekly and Monthly Sections of the Deltaboards

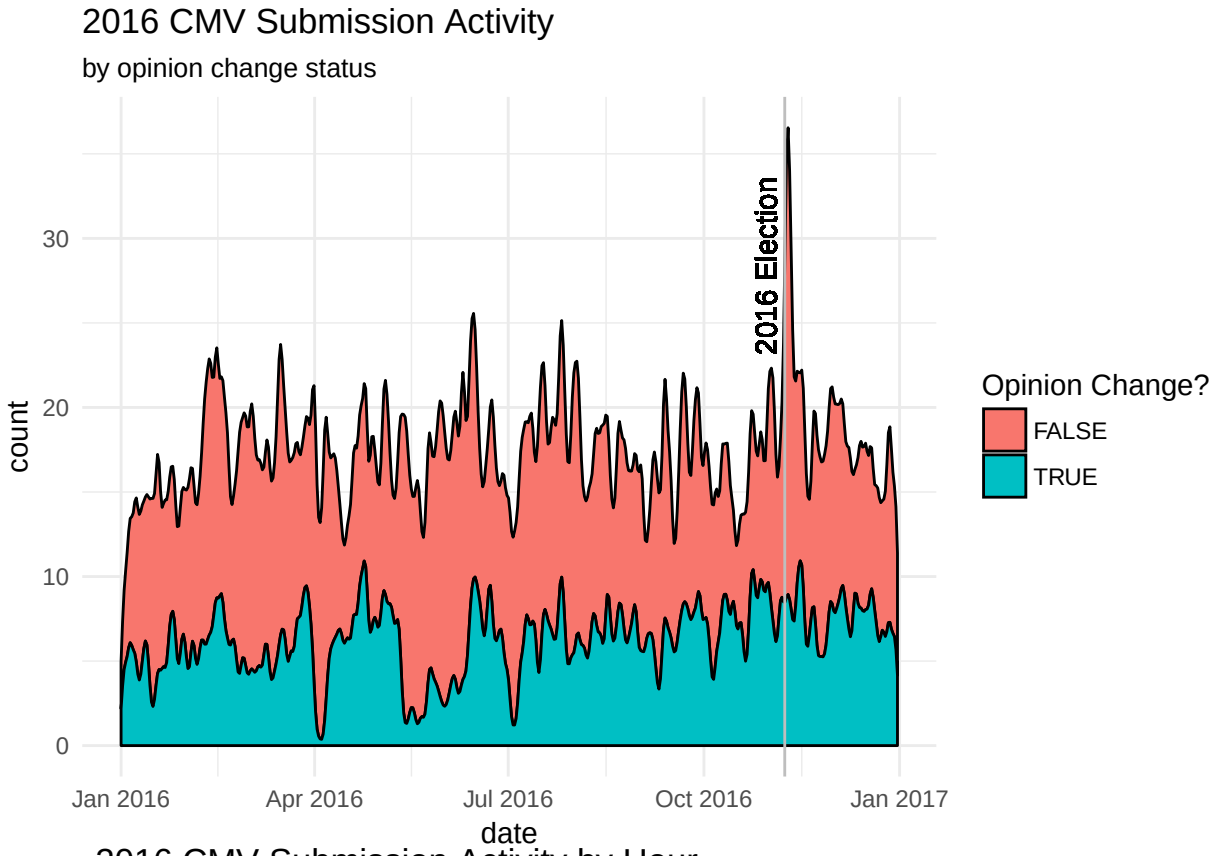
Exploring the Data Used

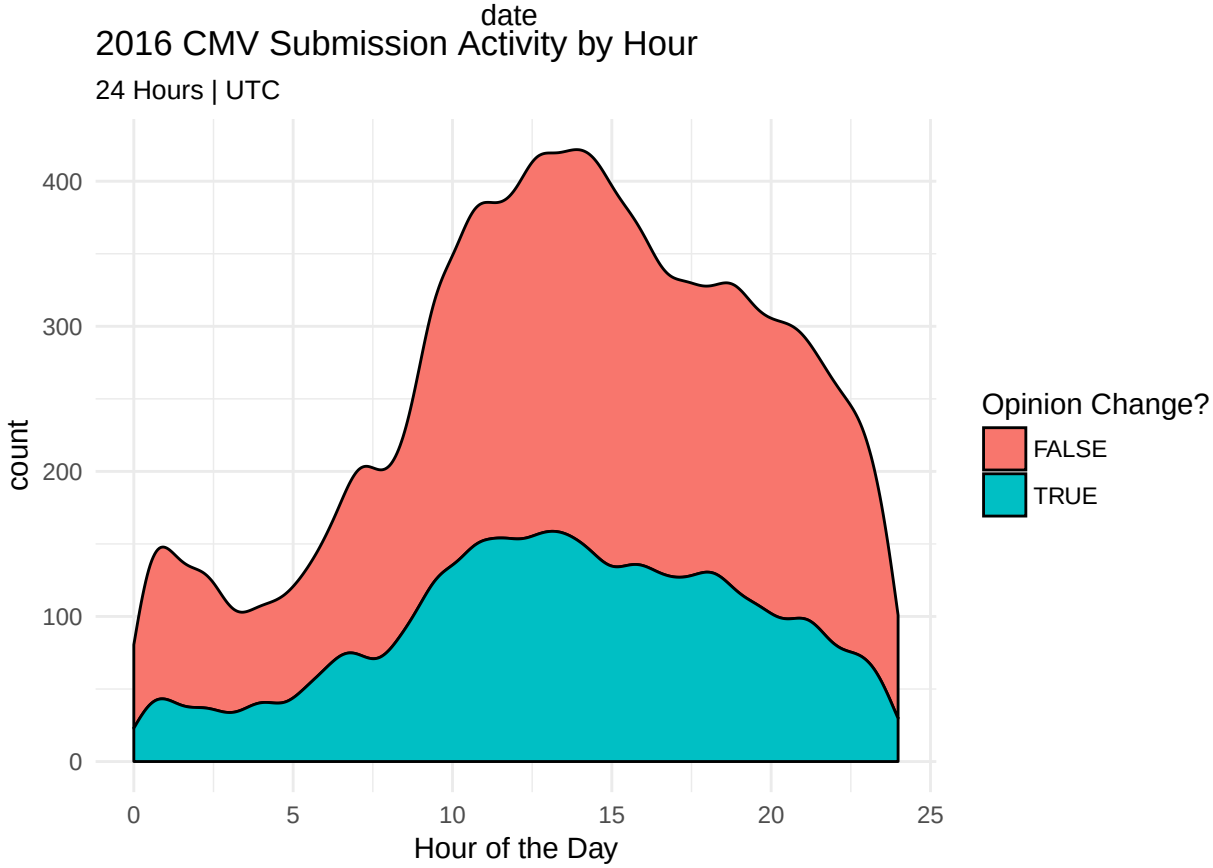
Firstly, in order to answer this question, opinions must be gathered. Because Reddit submissions are "archived", i.e. made immutable, once they are 6 months old I look at the opinion submissions posted to CMV in 2016, and the authors of these submissions. While there is no doubt that 2016 opinion submissions fit into and contain distinct time series patterns for many of the covariates I analyze, I make a tradeoff between "all" archived CMV submisions (2013 to today - 6 months) and too few. In choosing a year's worth of *relatively* recent data, I aim to average over fewer distinct times trends, maintain statistical power, and to counteract the remaining possibility, for older archived submissions, that a Reddit users deletes or edits their submissions. Because this study aims to gather information on Reddit user's partially through their submission history, deletions and edits create introduce anomalies in the data that are difficult to counteract. Deletions are a missing data problem with no practical imputation solutions, and while one can tell whether a submission is edited, there is no way to reliably count the number of edits or discern the changes made. Additionally, a number of older significant CMV submissions are authored by "[deleted]", i.e. an unknown author. A Reddit user can delete there account, and henceforth all of their submissions will appear to be authored by "[deleted]".

Due to the risk of account deletion teogether with submission alteration and/or deletion, the time at which data is important to note, as later versions of the same Reddit content may very well be different from their earlier versions. Starting February 19th, 2018, I utilized PRAW, The Python Reddit API Wrapper, with a MySQL database to begin scraping data from CMV opinion submissions from 2016. All the content of these opinion submissions were gathered. In the vast majority of cases in which an opinion submission author could be identified, of the submissions the author made before the time of the scraping were also gathered. On February 28th the data was backed up and ready to process. To reiterate generally, 2 major sources of data were gathered:

- All CMV opinion submissions and their content
- All available CMV authors' submission histories
  - Comments to these submissions were not tracked.

In 2016, 6429 opinion submissions were recorded in CMV. Of these, 0.361% of these opinions are recorded as having changed by their OP. Below are figures detailing day-by-day and hour-to-hour opinion submission activity by their resulting change in opinion, or lack thereof.

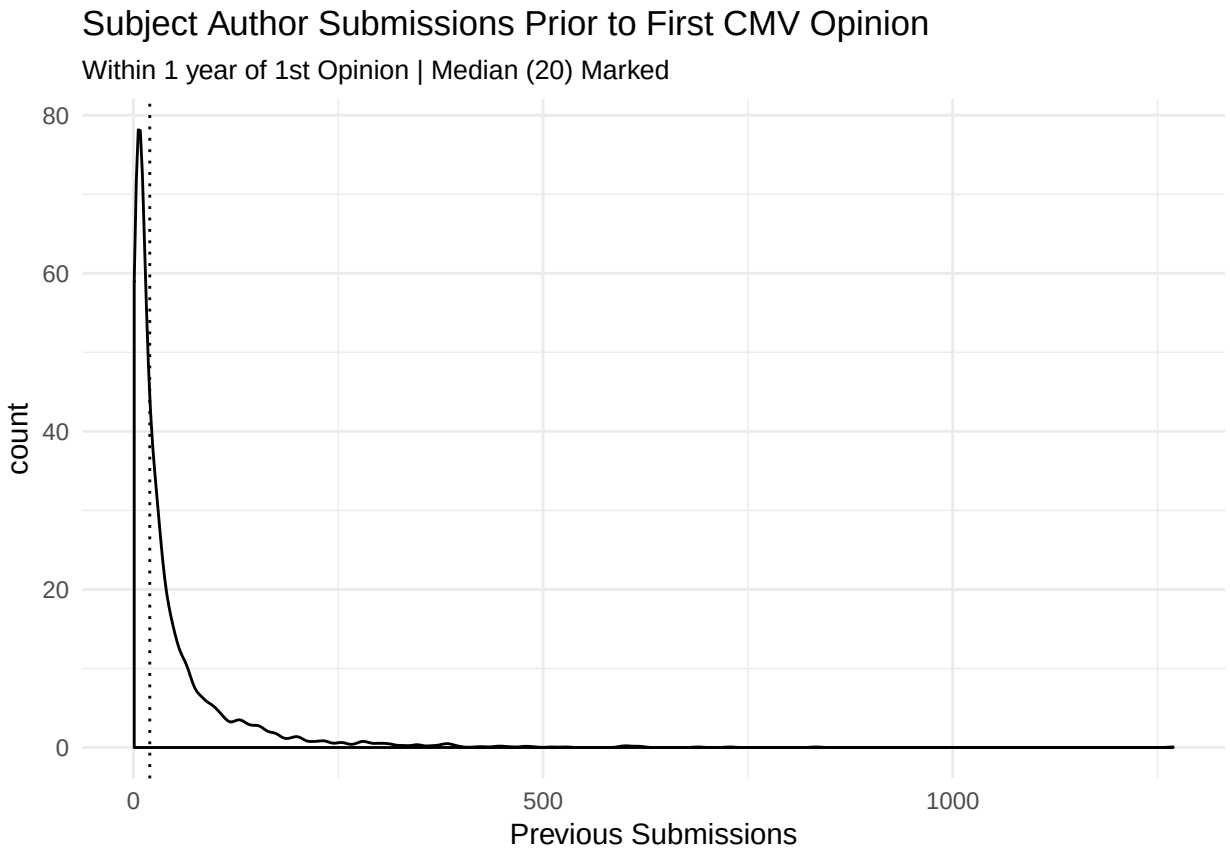

2016 CMV Submission Activity
by opinion change status
2016 Election
count
30
20
10
0
Jan 2016
Apr 2016
Jul 2016
Oct 2016
Jan 2017
date
Opinion Change?
FALSE
TRUE


2016 CMV Submission Activity by Hour
24 Hours | UTC
count
400
300
200
100
0
0
5
10
15
20
25
Hour of the Day
Opinion Change?
FALSE
TRUE

3882 2016 opinion submission were made by first time authors. Indeed, first time authors with submission history are the subjects of interest in this study (hereafter: subject authors). After all, without some sort of prior Reddit participation, there is very little information to discern from subject authors. There are many that have at least 1 prior submission in a subreddit other than r/ChangeMyView, and some even have hundreds of prior submissions. Of course, some submissions are older, relative to that author's first CMV opinion, than others. Should a submission from 2013 be taken into account for it's potential effect on a CMV opinion given in 2016? Deciding on a cutoff, if to have one at all introduces another tradeoff, one the one hand, a shorter cutoff, say only including the last 30 days worth of submissions before a subject author's first CMV opinion allows only the recent, and perhaps most relevant information of that author to be taken into account. However, a short cutoff also precludes more subject authors and submission from study, given the higher bar on activity demanded from the cutoff. A longer cutoff, or not having a cutoff at all includes more subject authors and submission in the study, with all possible information, but information gleaned from older submissions might be "noise" to our model that attempts to predict whether the further off opinion changes or not.

In light of these tradeoffs (which are heuristics at best), I begin by including only submission histories that are no more than 1 year older than a CMV author's first CMV opinion. If an author does not have submission history within one year of their first CMV opinion, they are not counted as subject authors. Because this pre-modelling parameter has such a large effect on how many subject authors there are, as well as the sort of information that is gathered about them, I will later relax my constraint on the submission history cutoff. More on this in the methodology.

3165 (0.815%), viable subject authors are available for study. A density figure of the number of submissions prior to a subject author's first CMV submission is given below.



Methodology

Features

I construct three distinct groups of features to utilize in my models. . .

- Author History (AH)
- Pre Debate
- Post Debate

For a complete overview of every feature, please see the appendix. A table of summary statistics for these variables is given below.

Model Selection, Comparison, and Performance Evaluation

Given the quixotic aim of this study, that is, the prediction of the opinion malleability of individuals on the internet, the choice of model(s) to use, different iterations of these models to compare, and how to evaluate

Table 1: Summary Statistics

Statistic	Mean	St. Dev.	Min	Max
(AH) # Submissions within 1 years before 1st CMV Post	45.440	75.442	1	1,269
(AH) # Unique Subreddits Posted In	16.905	18.844	1	163
(AH) Edits Per Previous Submission	0.094	0.147	0.000	1.000
(AH) Fraction Plural First Person Pronouns	0.004	0.006	0.000	0.105
(AH) Fraction Singular First Person Pronouns	0.041	0.021	0.000	0.250
(AH) Mean # of Words	87.217	103.384	0.000	1,585.833
(AH) Mean Plural First Person Pronouns	12.783	41.260	0	1,328
(AH) Mean Previous Submission Sentiment	0.036	0.066	−0.429	0.613
(AH) Mean Singular First Person Pronouns	109.307	199.928	0	3,158
(AH) Mean Submission Score	60.717	296.901	0.000	13,102.000
(AH) Removed CMV Subs 1 years before 1st CMV Post	0.476	1.275	0	19
(AH) Submissions with Content 1 years before 1st CMV Post	23.487	37.765	0	413
(AH) Subreddit Gini Index	0.295	0.209	0.000	0.886
(Post Debate) # Direct Comments	13.315	14.505	0	221
(Post Debate) # OP Comments	13.122	15.322	0	274
(Post Debate) # Total Comments	132.022	240.862	0	4,885
(Pre Debate) # Words	238.584	232.389	1	3,379
(Pre Debate) CMV Submission Hour	13.088	5.858	0	23
(Pre Debate) Fraction Plural FP Pronouns	0.006	0.007	0.000	0.118
(Pre Debate) Fraction Singular FP Pronouns	0.012	0.017	0.000	0.121
(Pre Debate) Sentiment	0.069	0.128	−0.638	0.635
(Pre Debate) Similarity Score	−0.003	0.065	−0.270	0.379
(Pre Debate) Topic 1 Score	0.084	0.233	0.00002	0.999
(Pre Debate) Topic 2 Score	0.092	0.246	0.00002	1.000
(Pre Debate) Topic 3 Score	0.101	0.245	0.00002	1.000
(Pre Debate) Topic 4 Score	0.022	0.095	0.00002	0.998
(Pre Debate) Topic 5 Score	0.072	0.212	0.00002	0.999
(Pre Debate) Topic 6 Score	0.072	0.213	0.00002	1.000
(Pre Debate) Topic 7 Score	0.073	0.213	0.00002	1.000
(Pre Debate) Topic 8 Score	0.385	0.482	0.00002	0.997
(Pre Debate) Topic 9 Score	0.101	0.255	0.00002	1.000
(Pre Debate) Total Plural FP Pronouns	1.298	2.227	0	25
(Pre Debate) Total Singular FP Pronouns	3.791	7.918	0	240

their importance are issues that must be addressed. The 3 groups of features I utilize are divisible into two categories. Author history and pre debate features, if utilized as a model's only features, makes for a true *prediction* task because it aims to utilize information that only is available in the intermittent period between when a subject author posts their opinion and when other Reddit users begin attempting to change the view; these are "pre persuasion features". While I do not utilize such a model in any sort of causal framework, by focusing on pre persuasion features attempting to predict a post persuasion outcome, the temporality of cause and effect, at least, is respected, placing confidence in the model perhaps a hair above the oft-quoted notion of "correlation does not imply causation". Utilizing the two groups of pre persuasion features and the post debate (post persuasion features), which utilize features which potentially capture information available after the outcome variable is fixed, does not respect the temporality of cause and effect, but is useful as a benchmark to compare against the pre persuasion based model.

That being said, I plan to use models that:

1. (Base) Utilize a single nominal feature as a base of comparison.
2. (Past) Utilize soley pre persuasion features (Pre Debate and Author History).
3. (Full) Utilize pre and post persuasion features.

My model of interest, however, will be the pre persuasion model, and comparing it's performance and the roles of its component features against the base and full model. As one may expect however, using only pre persuasion features disregards a great deal of useful intra and post debate information, and even information pertaining to the subject author's pre persuasion characteristics.

Thankfully, my main goal is not eking out model performance in a straightforward classification task (a model where the *prediction* aspect does not respect temporality). As such, I do not utilize less interpretable boosted or bagging models, but opt for a simple logistic regression model. With this model, I can measure the roles of my features within the easily understandable log odds, where I can at least have the notions of significance and log odds, which for this study are preferable to something like "variable importance", where concepts of power are almost impossible to perceive. Differentiating significant effects between less influential variables is practically impossible with variable importance.

Besides comparing 95% CIs for log odds between the three models, I also utilize standard model performance metrics of the receiver operating characteristic curve (ROC), as well as sensitivity (true positive rate) and specificity (true negative rate) with these latter two statistics based on a `p=.5` decision boundary. 95% bootstrap CIs for each of these three statistics are calculated using 20 reptitions of 10-fold cross-validation, for each of the models, all of which used centered and scaled data.

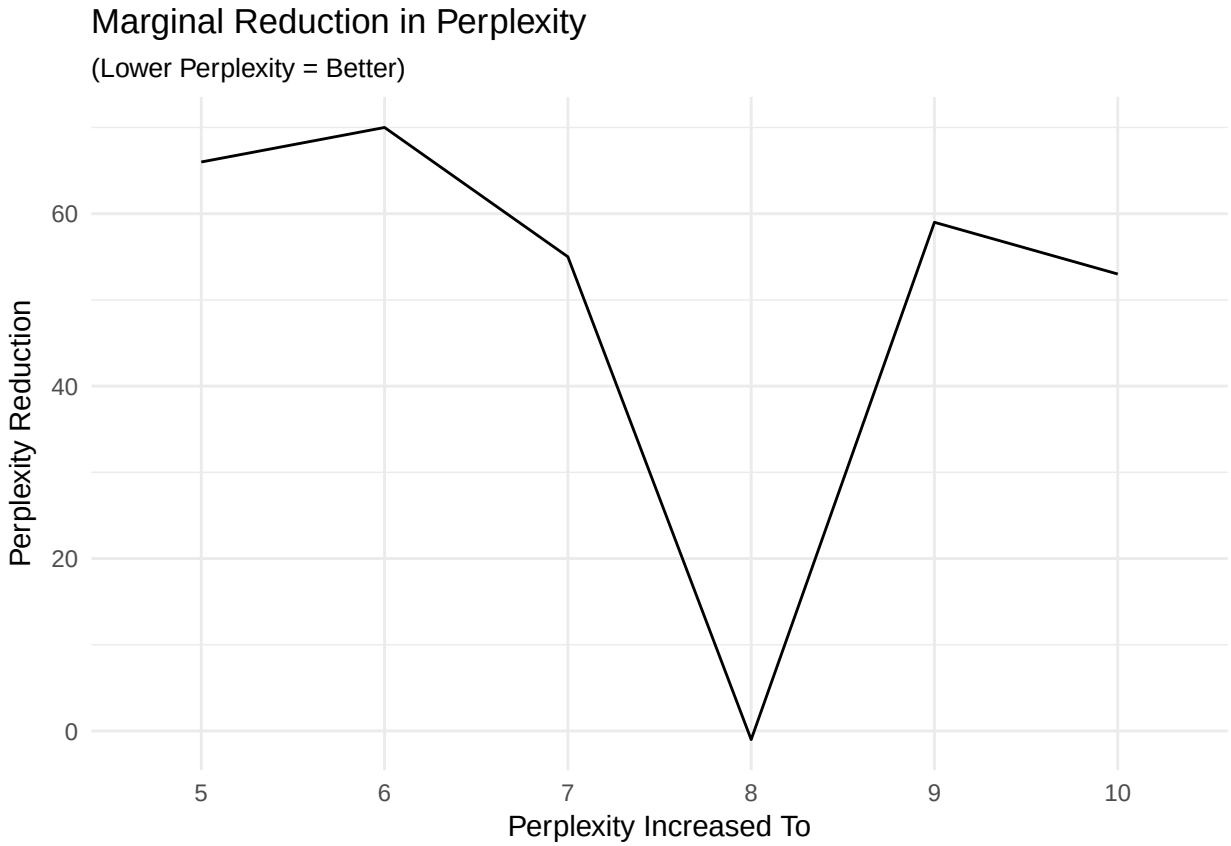
Feature Engineering Practices

I mentioned that I have constrained the scope of the subject authors' submission history. Facing a number of tradeoffs surrounding relevancy and abundance, I have elected to only include subject author submissions within a year before their first CMV post. This choice not only has effect on the number of viable subject authors, but also on any variable that utilizes the subject author's submission history. Conceivably, this choice of cutoff has effects on the performance and statistical significance of the features in the model. There are also other features that rely on certain parameters, resulting in either numerically distinct features, or even different numbers of related features. Thus, these choices guarantee a thousands of combinations of potential model-ready data, and some judgement must be exercised in selecting these parameters.

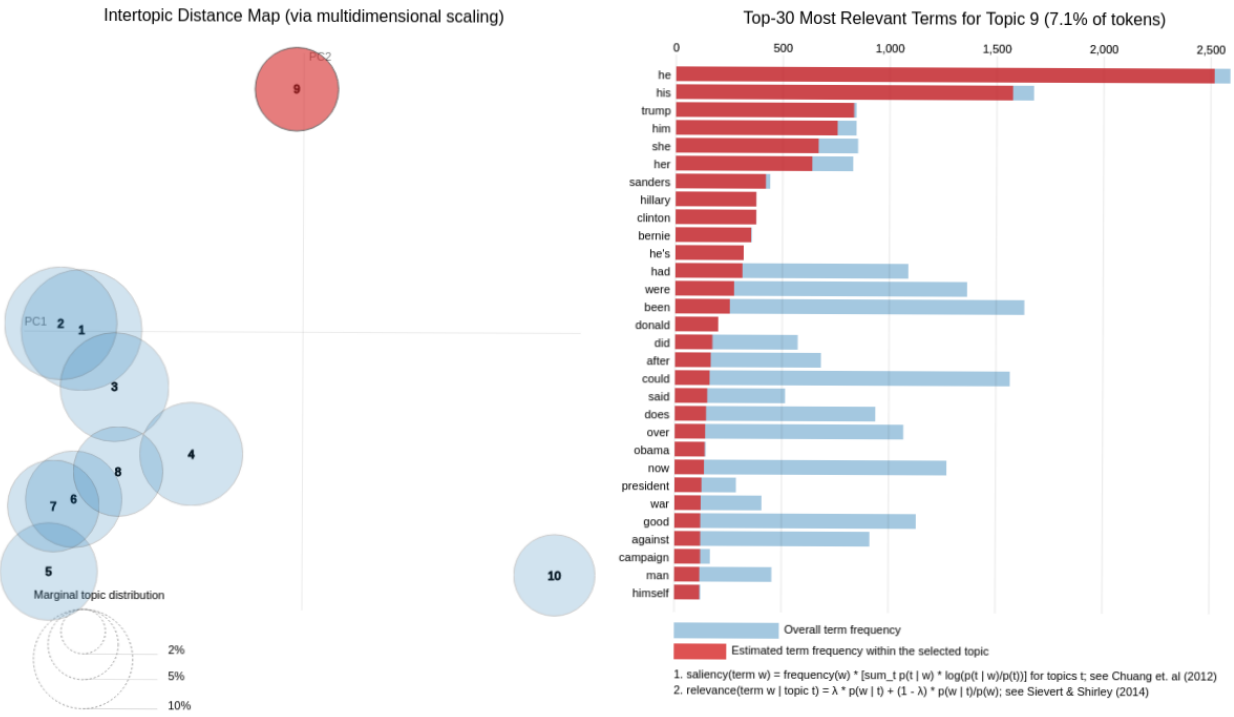
Choosing # of Topics for CMV Submission Topic Model

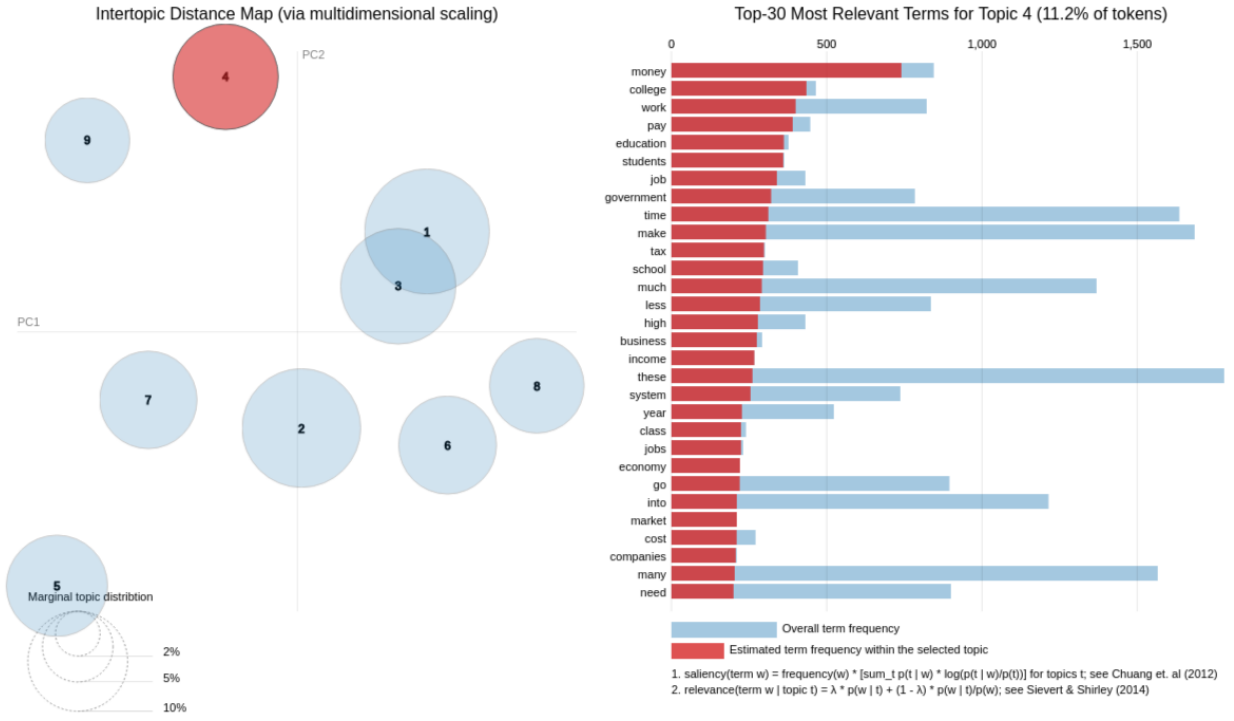
Besides submission history cutoff, choosing K topics for a Latent Direchlet Allocation (LDA) topic model is the 2nd such parameter that must be dealt with. This LDA topic model is applied to all CMV opinions submitted in 2016. With this model fitted, the subject author's first CMV opinions are given `K` topic scores summing to 1, in order to see, generally, how their subject matter fits within 2016. Determining a suitable `K` is important to effectively differentiate subject author's first opinions on the basis of subject matter.

In that case, how should one determine `K`? A number of heuristic methods exist for this task. Some involve using certain information criterion measures to find an "elbow", i.e. a place where the marginal information gained with an increase in `K` drops steeply. Other on visual comparison using PCA and judgement of most common words associated with the topics modeled. In this study, I look at both the reduction in "perplexity", a measure of how well a probability distribution or model predicts a sample, as well as a visual and topical inspection of the data.



Often, when using soemthing like Akaike information criterion (AIC) or Bayesian information criterion (BIC), graphs like the one above will be monotonically decreasing. Using perplexity, this is not the case, as one can clearly see marginal reductions exhibiting non-monotonic behavior. This non-monotonic behavior, while making it impossible to have an identifiable "elbow" in the graph, does point towards `K=8` or `K=9` topics as potentially viable, but increasing `K` to 10 topics still results in relatively large perplexity reductions. Turning to a visual inspection helps to differentiate `K=9` and `K=10`.





The above two figures are taken from interactive plots available to LDA topic models made with the "text2vec" package. In this case, I focus on the left section, the intertopic distance map, which projects the topics using PCA onto two dimensions, with size indicating marginal topic distribution. Immediately noticeable is the relative lack of overlap between different topics under `K=9` compared to `K=10`. Even with the caveat that dimension reduction/projection cannot completely represent topic overlap/separation, comparing these two

visualizations makes `K=9` more amenable to my purpose of topic differentiation. Even compared to $K = 4...8$, `K=9` displays better visual separation between topics. A table of the top 10 words comprising each topic is given below. Note that topic 3 seems to be a very generalized topic, as its top words are common URL prefixes, nubmers, and very common words. The rest of the topics seem to fall into distinguishiable categories.

1	2	3	4	5	6	7	8	9
white	money	https	gun	believe	he	want	vote	women
really	college	1	law	these	his	life	trump	person
why	work	2	police	many	him	could	party	she
black	pay	use	rights	world	game	way	sanders	her
make	education	3	laws	am	been	make	hillary	sex
know	students	had	government	such	war	really	clinton	men
been	job	were	abortion	those	were	why	bernie	someone
good	government	new	right	human	now	know	election	man
i've	time	may	crime	society	he's	now	voting	gender
am	make	watch	against	view	could	going	political	he

Choosing # Topics for Submission History and CMV Opinion Similarity

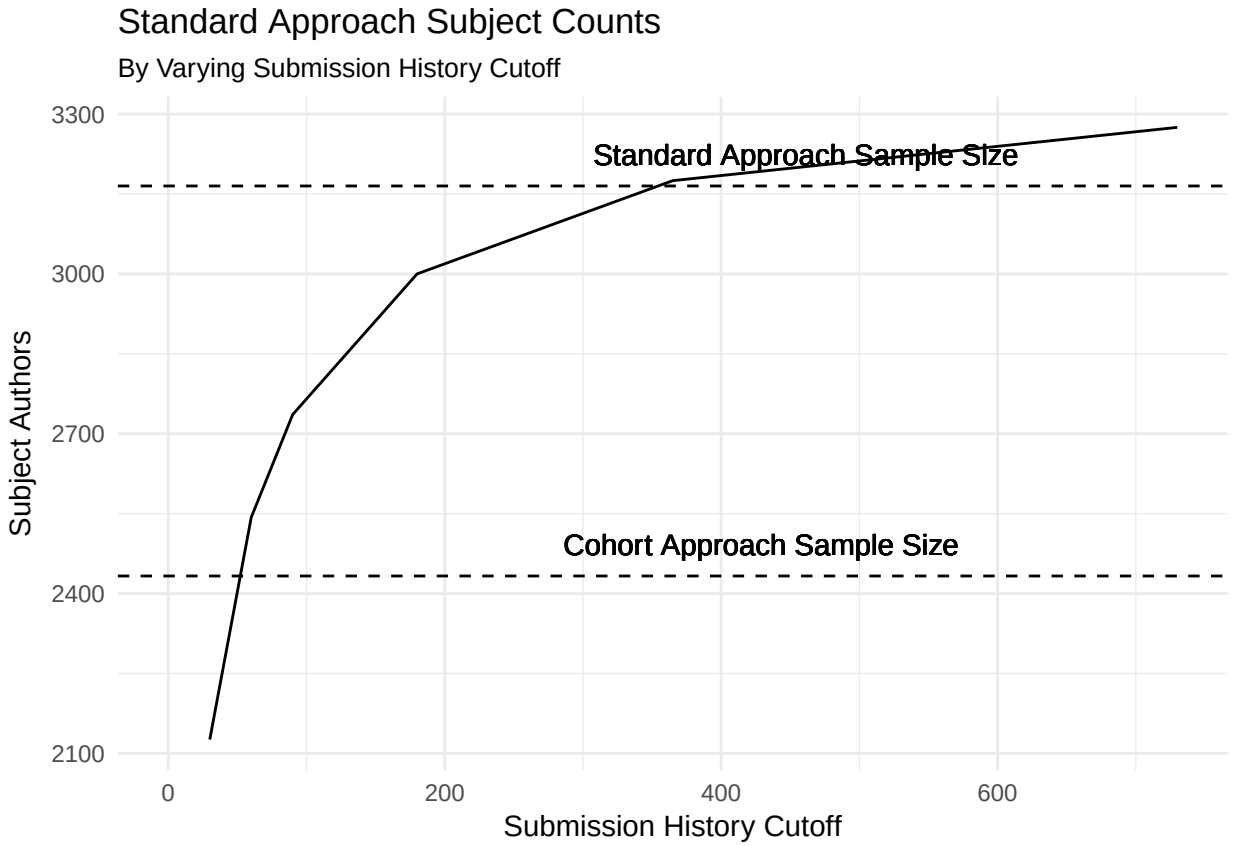
One of the Pre Debate features is a similarity score between the subject author's first CMV opinion and their included submission history. I choose to model similarity using cosine similarity with a Latent Semantic Analysis (LSA) topic model. The LSA model is first trained upon all subject authors' included submission history with `J` topics. Then, cosine similarity between each individual subject author's body of prior work and their first CMV opinion is calculated, ranging from -1 (least similar), to 1 (most similar). However, in this case, unable to find heuristics to choose `J`, I opt for a performance metric based approach to see if there are "better" values of `J`.

In light of the Base, Pre, and Full models discussed earlier, I utilize varying values of `J`, as well as `K`, and submission history cutoff for the three models. Utilizing a grid search over these values, I compare 95% CIs of area under the ROC curve to see if specific combinations of these parameters that create my training data result in significantly better model performance *within* either the Base, Pre, or Full models. I found that varying these three parameters within a reasonable set of options does not result in significantly better models, which justifies the heuristics I previously described for the selection of submission history cutoff and `K`. With `J` then I simply opt for `J=150`.

Introducing A "Cohort" Approach

All of the author history variables (AH) are highly dependent on the aforementioned submission history cutoff, or more specifically, for a given subject author, the number of submissions added or dropped due to an adjustment of submission cutoff can significantly affect author history variables. In the case of a subject author with only two submissions, this effect is quite substantial should the cutoff change halve the history the models look at. Also, for those subject authors whose submissions span a long period of time, what plays a larger role? Their earlier, or later submissions?

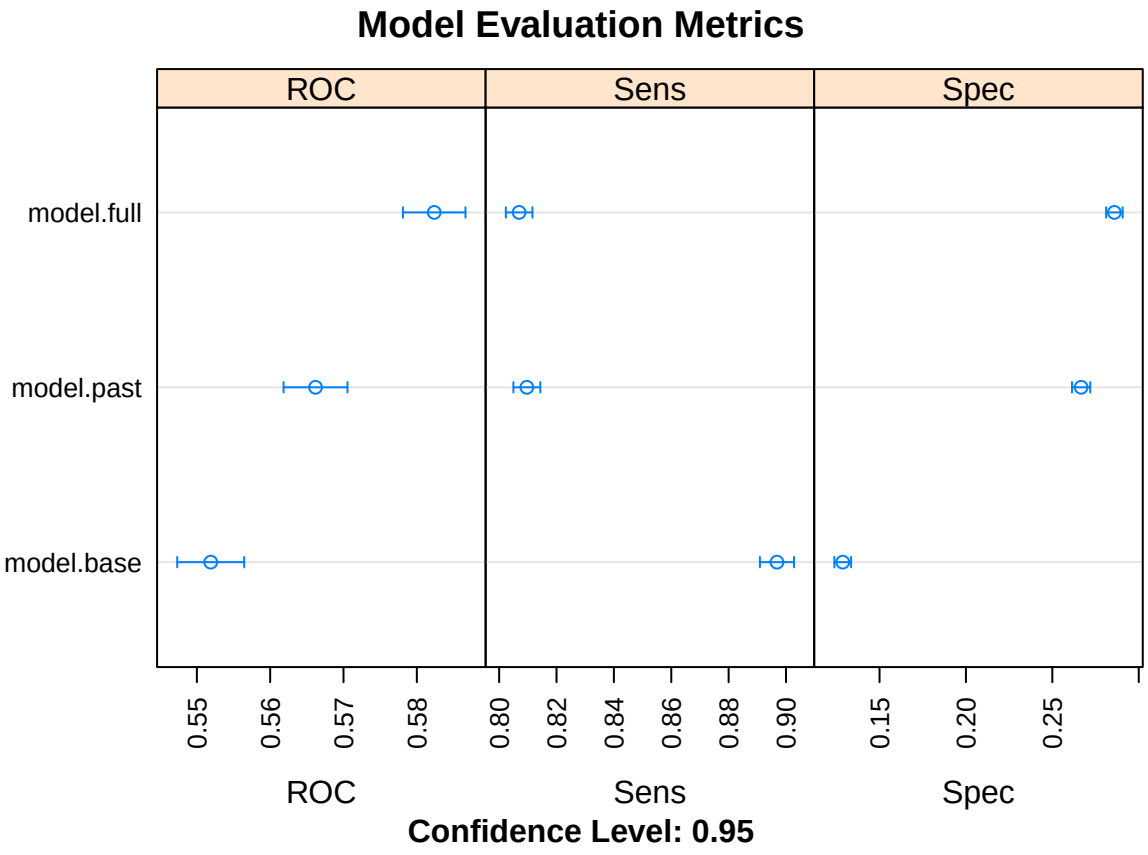
Thus, there must be some effort taken to adjust the history cutoff in a meaningful way. My previous criteria for selecting subject authors, given a fixed history cutoff was simply to include them should they have submission history within the window alotted by the cutoff. However, this criteria is not condusive to a method which calls for a variable submission cutoff. Adjusting the cutoff, depending on the direction, would lead to including or rejecting subject authors for modelling, making comparison of models run with different cutoffs more difficult. To more directly address the issues of time and submission history, I use a "Cohort" approach, which fixes the subject author's initially in the shortest submission history cutoff (30 days), and keeps these authors while adjusting that cutoff to 60 days, 90 days, 180 days, and beyond. Under the Cohort approach, the sample size remains constant at 2433, as opposed to the varying sample size under the standard approach (see figure below). In effect, this approach acts as a sort of brute force, weighting scheme with $0, 1$ weights for a subject author's submissions.



In utilizing a "Cohort" approach, the signifiance of many of the author history (AH) features can be more thoroughly analyzed. The standard approach that I have previously described, with fixed values of `K`, `J`, and cutoff, can still be analyzed, and variable significance can still be determined, but one can have more confidence of the link between variables should they pass through the gauntlet of signifiance in the standard past model, the standard full model, and finally, under the evaluation of the "cohort" approach.

Results

Model Performance (ROC, Sens, Spec)



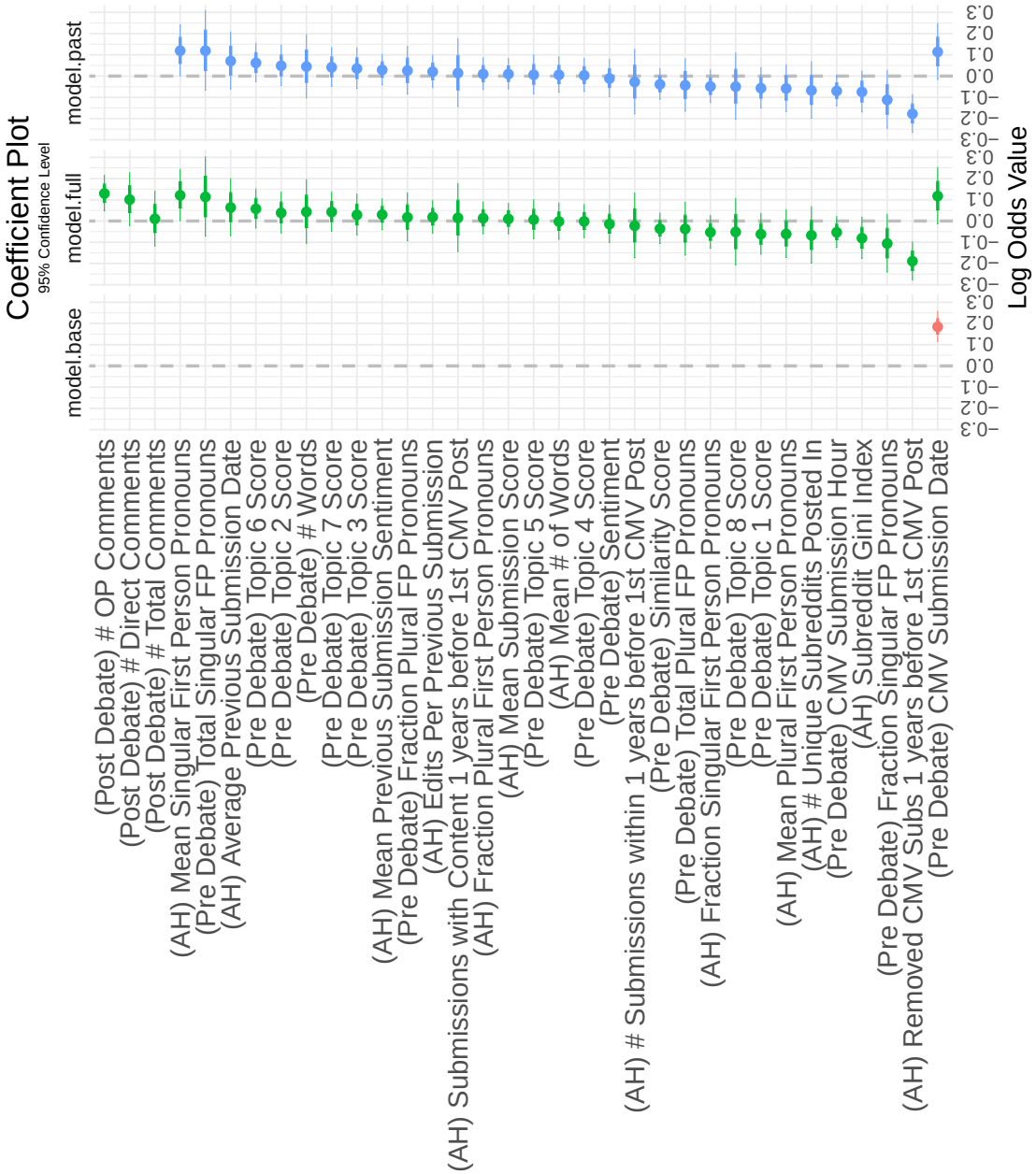
The three models shown in the figure above are significantly different in terms of ROC, the general performance measure seen here, with even the base model, which only uses the number of words in the subject author's CMV submission, performing better than randomly guessing. Of course, opinion malleability, the subject authors' susceptibility to persuasion stil proves to be a rather difficult modelling task, as even with the highest ROC (token) full model is still shy of just .6 ROC.

In comparing the full model to the past model, in terms of sensitivity, the two are not distinguishable under the CIs constructed. This speaks somewhat well for the past model, as it is able to correctly classify ~81% of the subject authors changed opinions, and this true positive rate performance is not improved by utilizing the full model with variables taking into account aspects of the debate; intuitively these should be important. Additionally, specificity, the true identification of no opinion change, does differ between the past and full models, with the two models sitting just shy of 26% specificity. I have admitted that my version of the "full" model does not include a full gambit of debate interaction variables (Tan et al. (2016)). Still, the

lack of difference between the past and full models sensitivity, and their difference in specificity points to an interesting difference in the effectiveness of the past model between those subject author's whose first opinions were changed and those who remained fixed in their views.

Coefficient Analysis

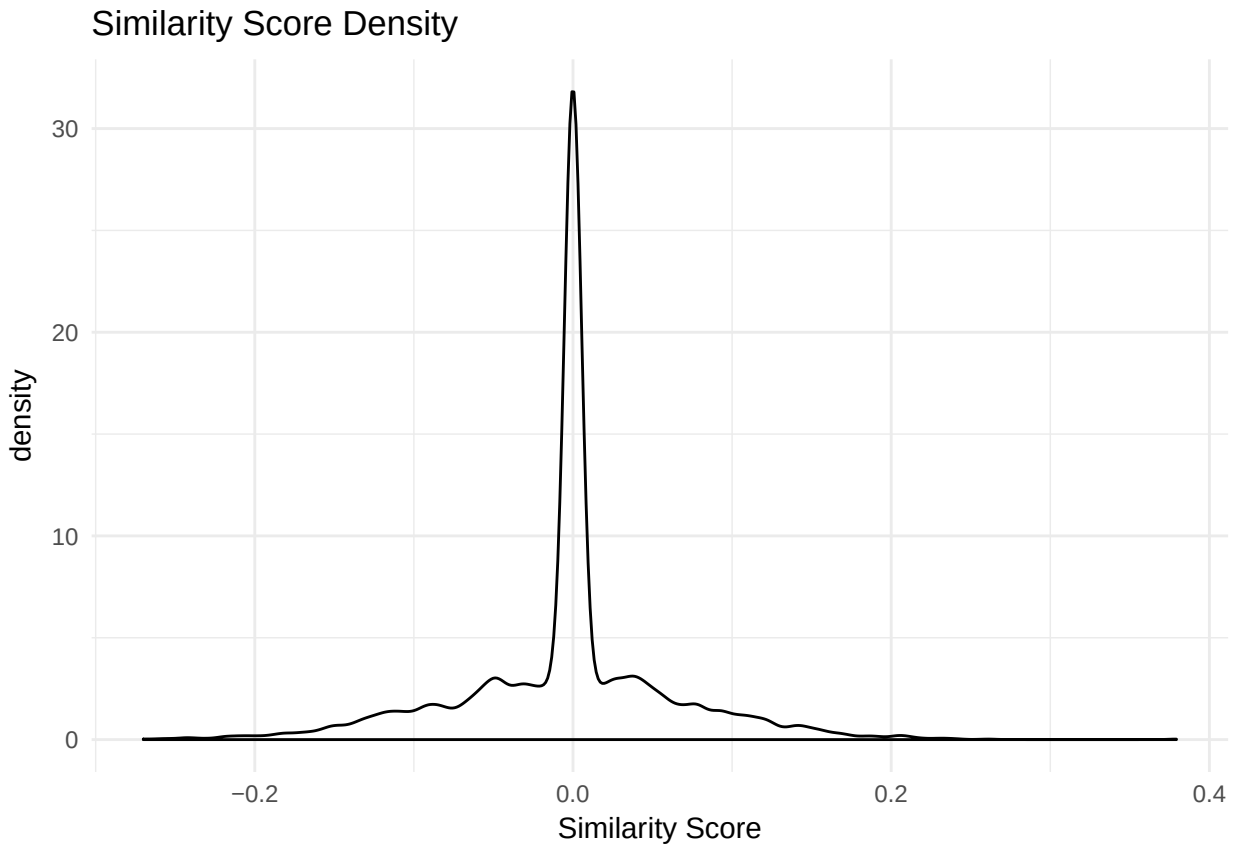
Below I give a coefficient plot of the three models, sorted by magnitude of their (log odds) effect. Because the model is trained on centered and scaled data, the effects displayed below pertain to a one standard deviation increase of the variable in question.



As alluded to before, the past model is the one most pertinent to analysis, but by having the full model included in the coefficient plot, one can verify the (in)stability of the coefficient's effects. Starting from the top, as is expected, most of the post debate variables (# OP Comments, # Direct Comments), are significant. After all, deltas cannot occur unless the subject author, as the OP, posts a comment that responds to a direct comment. There are however, more interesting coefficients to look at further down.

Self-Affirmation

One of the main mechanisms which I have used to test self-affirmation is the calculation of a "similarity score" between the subject author's first CMV opinion reasoning and their submission history within 1 year. Self-affirmation theory notes that if global self-worth is bolstered by an unrelated domain, the subject author should tolerate a threat to the domain of interest. In the context of this study, I then expect similarity score to be have a negative effect on the subject author's chance of changing their views, and this is exactly the effect seen. Reassuringly, the similarity scores across all of the subject authors is distributed evenly about 0, so anomolies in this covariate do not seem to be the cause. Again, while I make no causal claims here there does seem to be evidence of self-affirmation theory's tenents extending beyond the immediate past.



First Person Pronouns

Within the coefficient plot, a number of first person pronoun coefficients prove to be significant, even in the full model, yet there are some slightly curious results to parse as well.

Singular

For author history (AH) and Pre Debate, the mean and total number of singular first person pronouns have a significantly positive effect on first opinion malleability. However, the *fraction* version of both of these variables have significantly negative effects. This stands in constrast the findings of Tan et al. (2016), who found, at least for CMV opinions, that *both* fraction and total singular FP pronouns had positive effects. The findings on first person pronouns and "openness", or in this case malleability, are not universal. Pennebaker and King (1999) found negative correlations between openness and first person singular pronouns.

The full and past models both utilize variables accounting for the number of words in a subject author's submission history and first opinion post, so the most cogent explanation of these effects is that increased useage of singular FP pronouns, but coupled with longer submission or opinion reasonings in general, as to keep the fraction relatively constant, are correlated with increased first opinion malleability.

Plural

For plural FP pronouns, the only coefficient that survives the transition from past to full model is the negative significant effect of mean number of plural first person pronouns in the author history (AH). While Tan et al. (2016) found more signifiance with the fraction of plural FP pronouns in the CMV opinion as well, the directionality of this finding makes sense, given:

> [F]irst person plurals can indicate a diluted sense of group responsibility for the view.

CMV Topics

Naturally, one assumes that certain subject matter influences first opinion malleability, and this is reinforced by the coefficient plot. Topic 6 and Topic 1 have significant effects, but in opposite directions, positive and negative, respectively. Looking again at the top ten words for these topics, Topic 1 can perhaps be interpreted as black-white race related questions, while Topic 6 is a little more difficult to interpret, consisting of many male pronouns.

Given the self-selecting nature of r/ChangeMyViews users, it is interesting to see increased malleability associated with black-white race relations, but decreased when it comes to (perhaps) "male" issues. However, given the 18-29 male skew of the Reddit userbase as a whole, maybe this is not so surprising (Duggan and Smith (2013)).

1	6
white	he
really	his
why	him
black	game
make	been
know	war
been	were
good	now
i've	he's
am	could

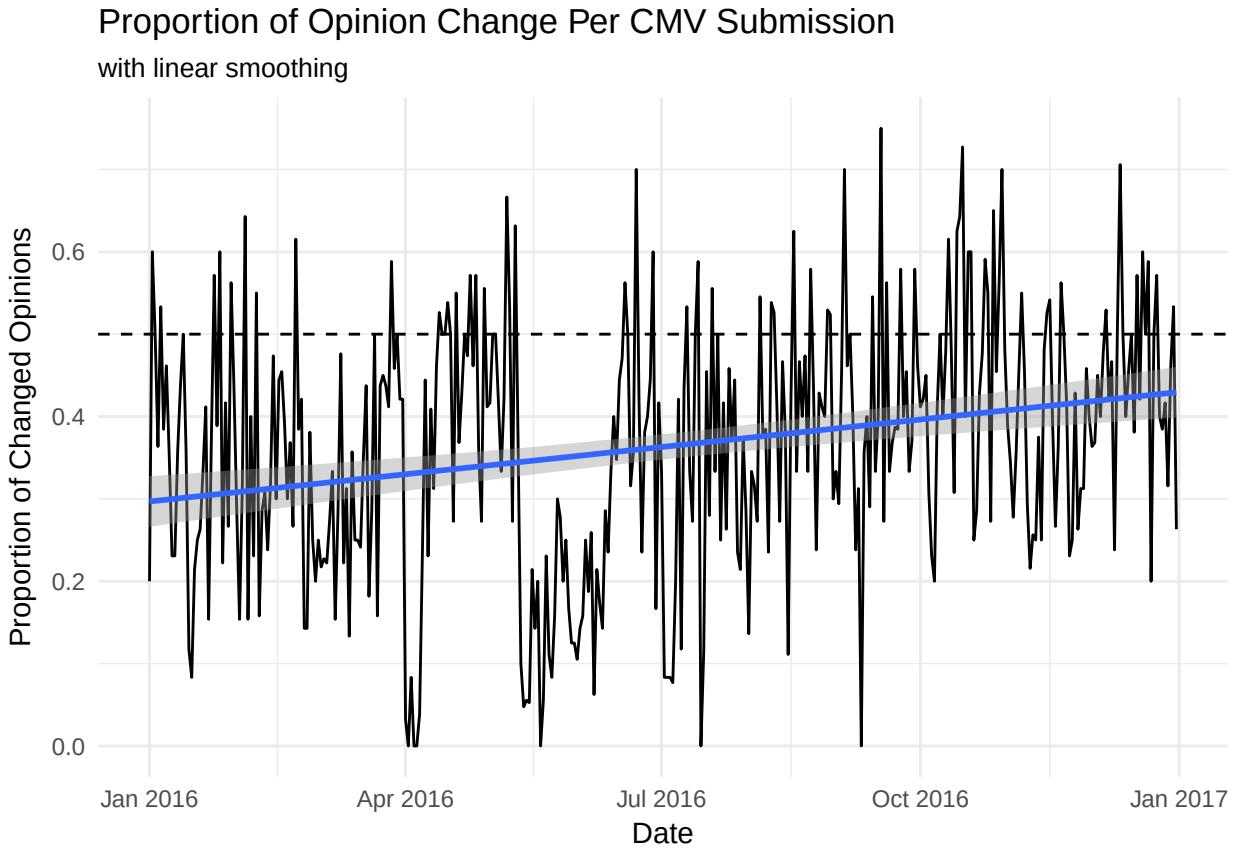
Submission History Diversity and Equality

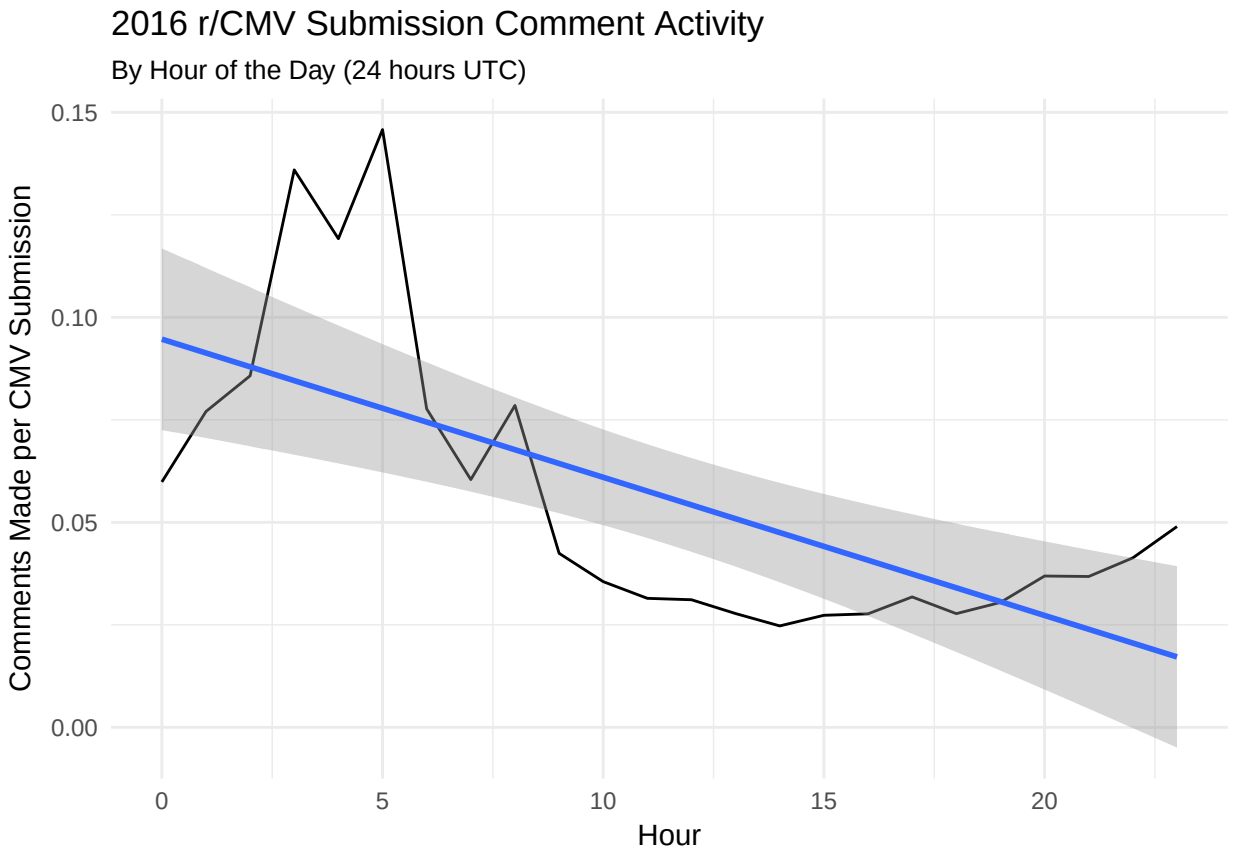
The author history variables "(AH) # Unique Subreddits Posted In" and "(AH) Subreddit Gini Index" together capture the diversity and equality of communities for a subject author. For both the Gini Index and number of unique subreddits, one sees a negative significant effect. More diversity then, (more unique subreddits), has decreased odds of opinion malleability, while more equality (lower Gini) between a fixed number of communities has increased odds of opinion malleability. The model suggests that increased equality of participation can offset, up to a certain point (given the 0, 1 bounds of the gini coefficient), the effects of increased diversity.

Time

The coefficient plot also indicates that the timing of a subject author's first opinion also effects malleability. The variable for the date and hour of the day, both indicate significant postive and negative effects, respectively, on the odds of opinion malleability. Though 2016, opinion submissions became more likely, overall to change, and within each day, opinion submissions posted later in the day (UTC), were less likely to

change. The date effect result seems quite in line with information that can be gleaned from unconditional, exploratory data analysis (example below), suggesting time trends, and other information not in the model, like current news, likely influences malleability. However, the time of day/hour effect has a less obvious cause. Comments per CMV submission tend to decrease for later hours, but in the full model, this variable is already taken into account. It might be possible then for the time of day trend to indicate differing malleability for subject author's from different time zones, but under the current framework this is difficult to determine.





Community Norms

Like many subreddits, r/ChangeMyView has a variety of rules when it comes to submitting or commenting. In some cases, subject authors may have attempted to post an opinion submission on CMV, only to have it removed due to rule violations. Because these removed opinions cannot be changed, they are not counted as the subject author's "first" opinion submission. However, they do provide useful information as to how well the subject author can conform to CMV's rules. In the coefficient plot, the number of removed CMV submissions before the subject author's first "successful" one has a significant negative effect on malleability. This effect has the largest magnitude under the full model.

Because submissions can be removed for a variety of reasons, one cannot say with certainty what the significance of this variable points to. Submissions can be removed if a user responds and the subject author, (the OP) does not respond within 3 hours. If submission were removed for this reason, perhaps the subject author's innate level of engagement could be the state measured by this variable. Posts can also be removed for being too short, not expressing a neutral stance, or playing devil's advocate. In a number of cases, community moderators cite specific rules broken when removing a post, and the classification of

removed CMV posts based on such information provides an obvious but likely fruitful direction for further study in this area.

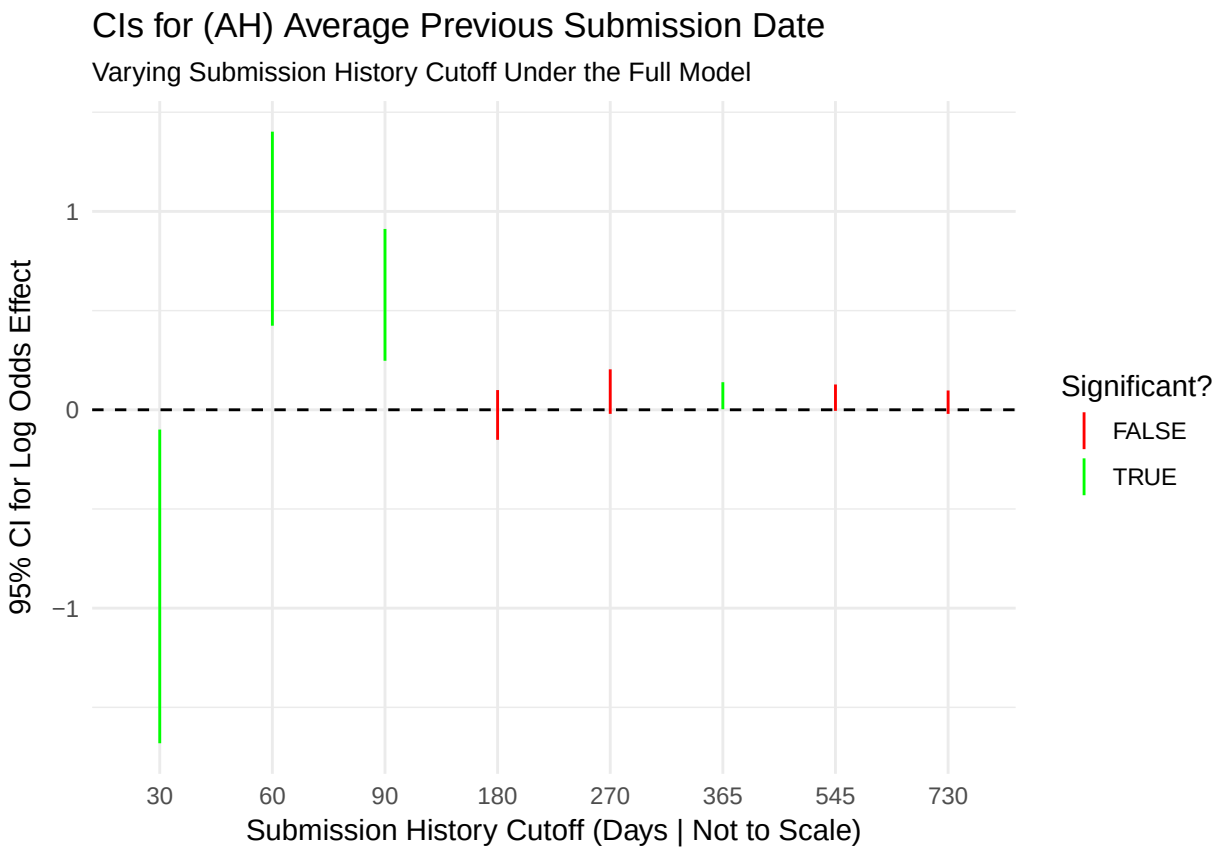
Submission History Timing/Concentration

Much of this paper has treated subject author submission history as occurring simply "within" a blockoff time, with the first CMV opinion and cutoff dates serving as bounds. However, between subject authors, submission histories are concentrated further or closer to their first CMV opinion. The (AH) Average Previous Submission Date introduces some of this information into the model.

For the past model, this variable has a positive significant effect on opinion malleability, i.e. *closer* histories to the first CMV opinion increase the odds of an opinion changing. However, this effect becomes just insignificant in full model, lowering the confidence one can have in such an effect. However, the average previous submission date, like many author history variables, is highly susceptible to changes in history cutoff. For this variable especially, because means are susceptible to outliers, large swings in its value are possible as submissions nearby the cutoff are included or ignored.

"Cohort" Approach

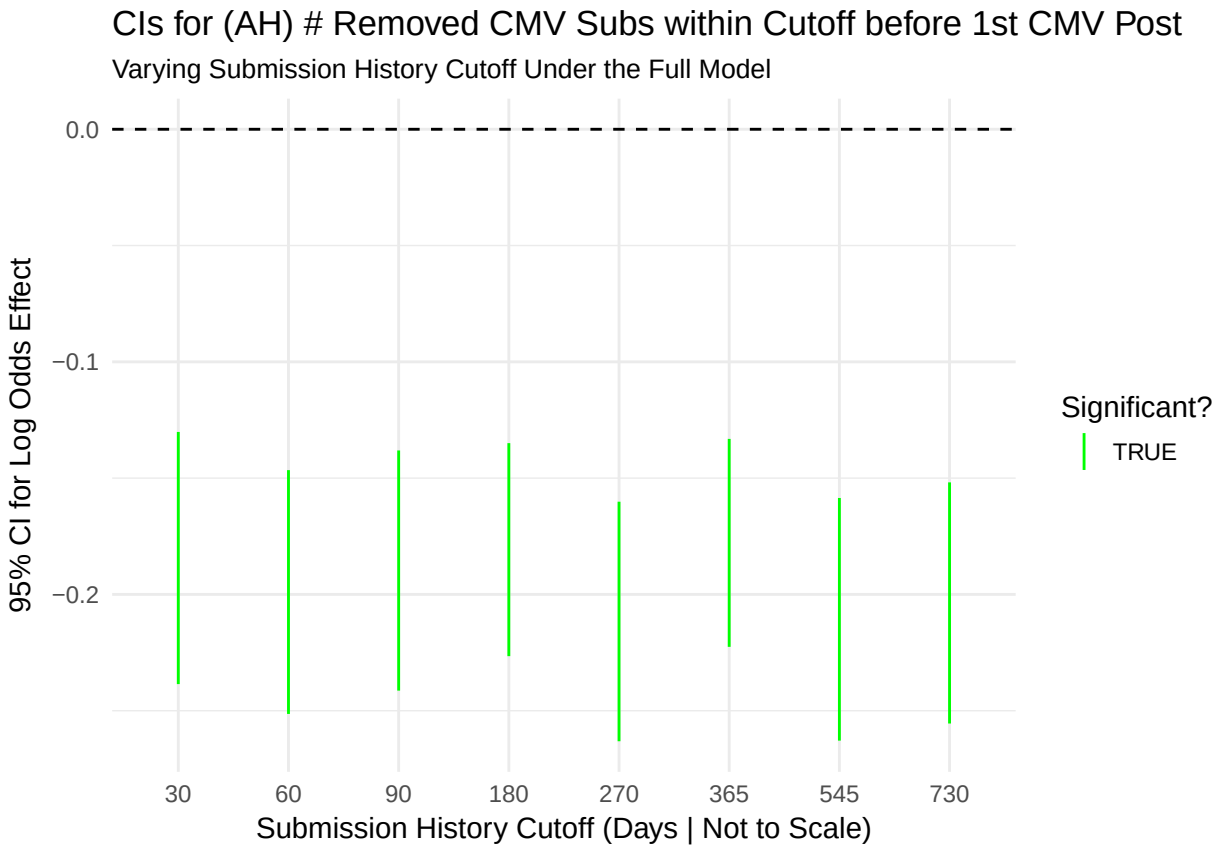
Recall that with the cohort approach, instead of varying the subject author sample size with the submission history cutoff, I instead fix the subject authors as those who had at least 1 submission within 30 days of their first CMV submission, and then adjust the amount of their submission history to include, altering the author history variables. While the results of cohort approach aren't *directly* comparable to that of the standard approach with its larger sample of subject authors and fixed submission history, the cohort approach offers a more nuanced evaluation to author history (AH) variables. For instance, although (AH) Average Previous Submission Date, was found to be just insignificant in the full model in the standard approach, the cohort approach tells provides some new insight, as seen below.

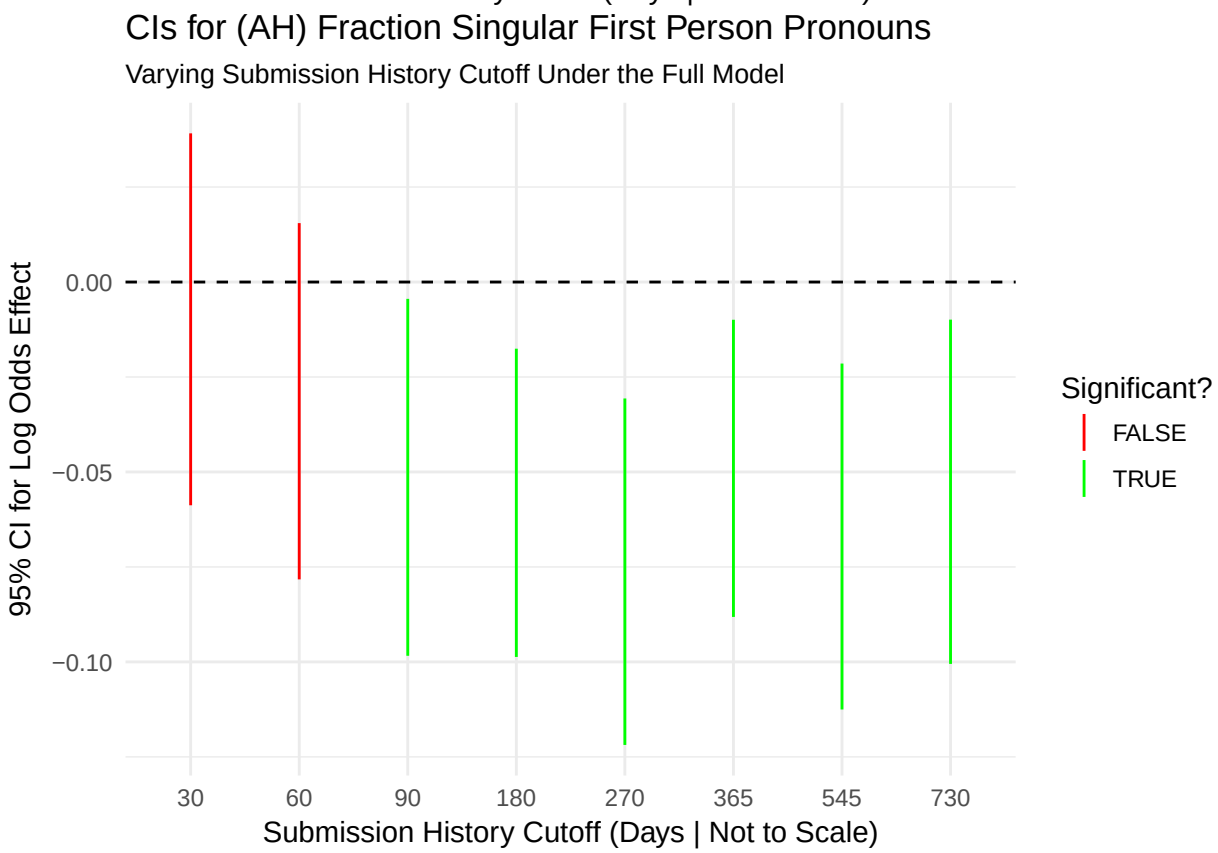


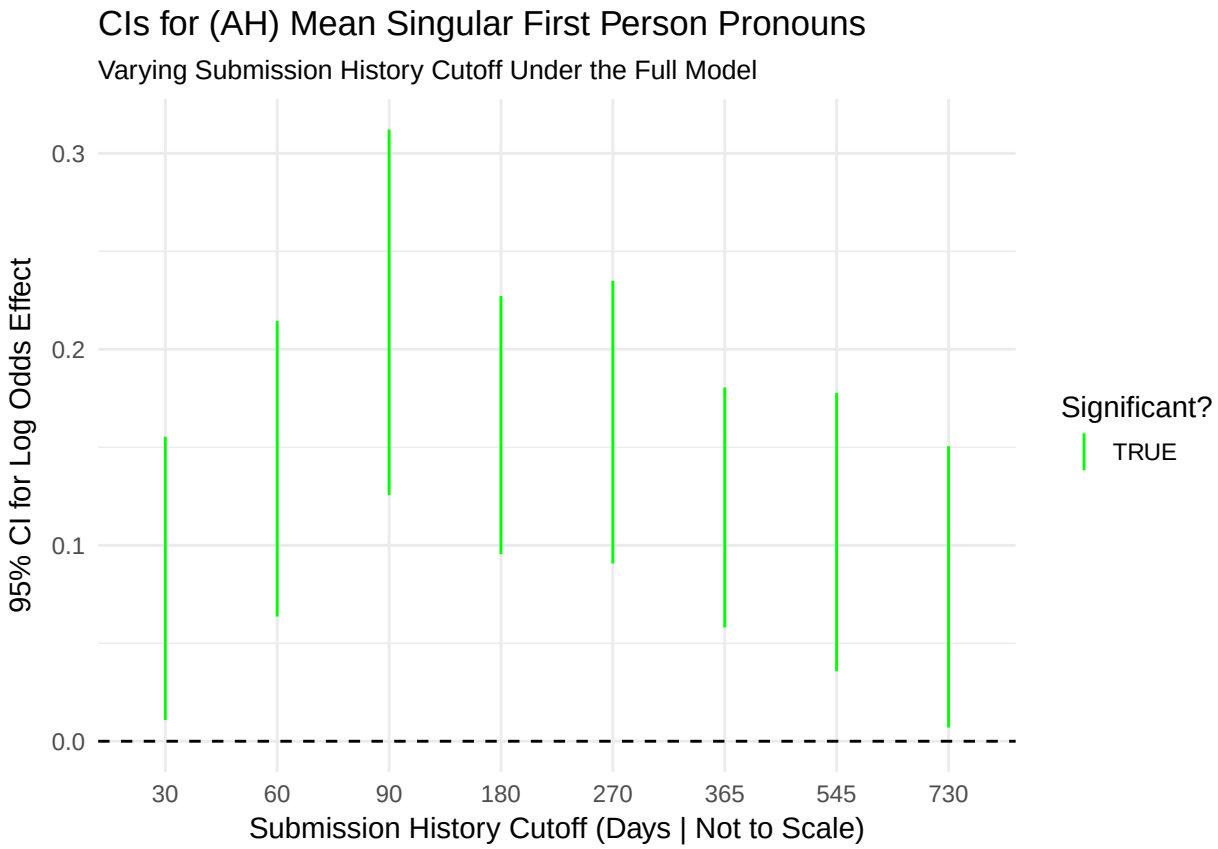
With this figure it's apparent that under the full model, lower cutoffs for submission history result in significant effects for this variable, though, there is even a curious change in direction when adjusting the cutoff from 30 to 60 days. While the results of this figure aren't conclusive evidence, one can at least see that the average submission date (i.e. distance from the subject author's first opinion) is more relevant for the given cohort until the cutoff expands.

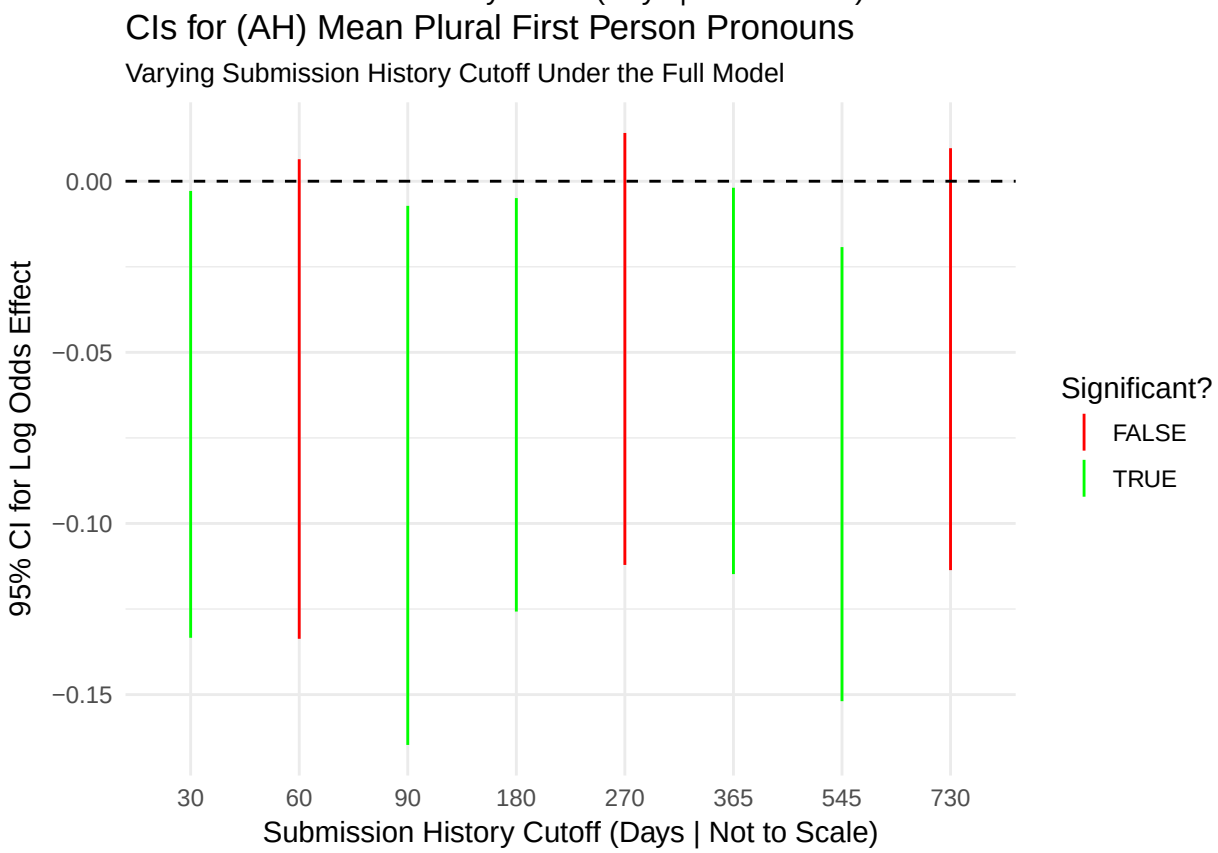
Reinforcing Previous Findings

Utilizing the cohort approach, I find that some author history variables have robust significance across changing definitions of submission history cutoffs.


CIs for (AH) # Removed CMV Subs within Cutoff before 1st CMV Post
Varying Submission History Cutoff Under the Full Model
0.0
−0.1
−0.2
95% CI for Log Odds Effect
30
60
90
180
270
365
545
730
Submission History Cutoff (Days | Not to Scale)
Significant?
TRUE


CIs for (AH) Fraction Singular First Person Pronouns
Varying Submission History Cutoff Under the Full Model
0.00
−0.05
−0.10
95% CI for Log Odds Effect
30
60
90
180
270
365
545
730
Submission History Cutoff (Days | Not to Scale)
Significant?
FALSE
TRUE

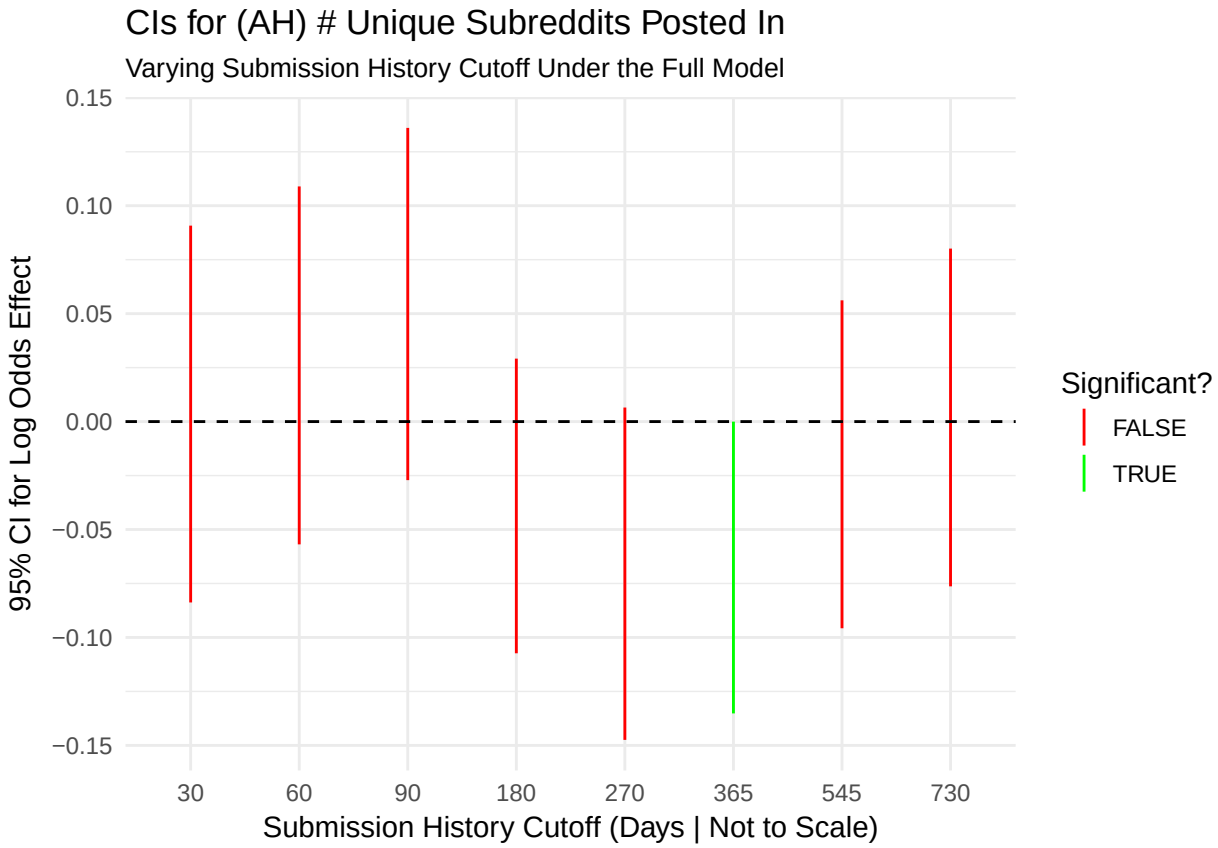

CIs for (AH) Mean Singular First Person Pronouns
Varying Submission History Cutoff Under the Full Model
95% CI for Log Odds Effect
0.3
0.2
0.1
0.0
30
60
90
180
270
365
545
730
Submission History Cutoff (Days | Not to Scale)
Significant?
TRUE


CIs for (AH) Mean Plural First Person Pronouns
Varying Submission History Cutoff Under the Full Model
95% CI for Log Odds Effect
0.00
−0.05
−0.10
−0.15
30
60
90
180
270
365
545
730
Submission History Cutoff (Days | Not to Scale)
Significant?
FALSE
TRUE

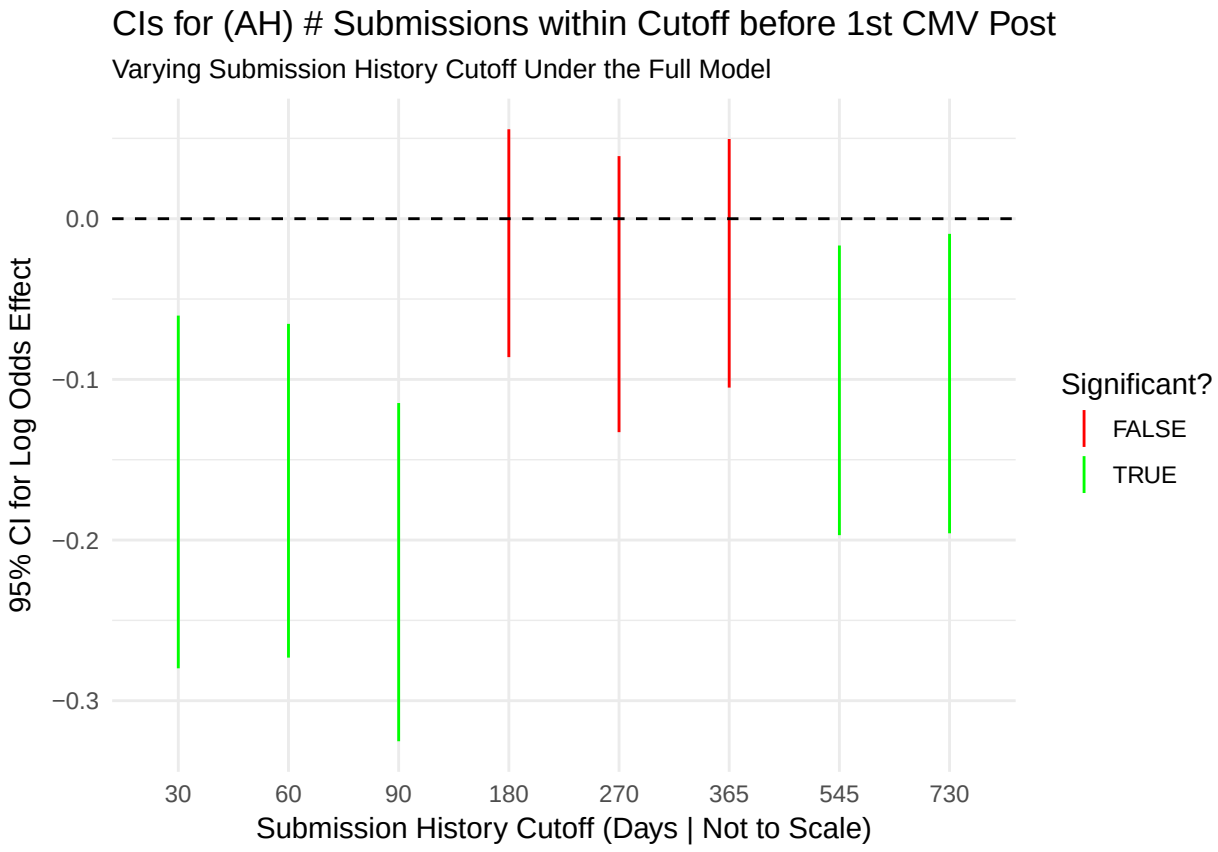
In the grahps above, for instance, we see that the community norms findings, and those of singular first person pronouns and plural first person pronouns are mostly stable through differing submission history cutoffs.

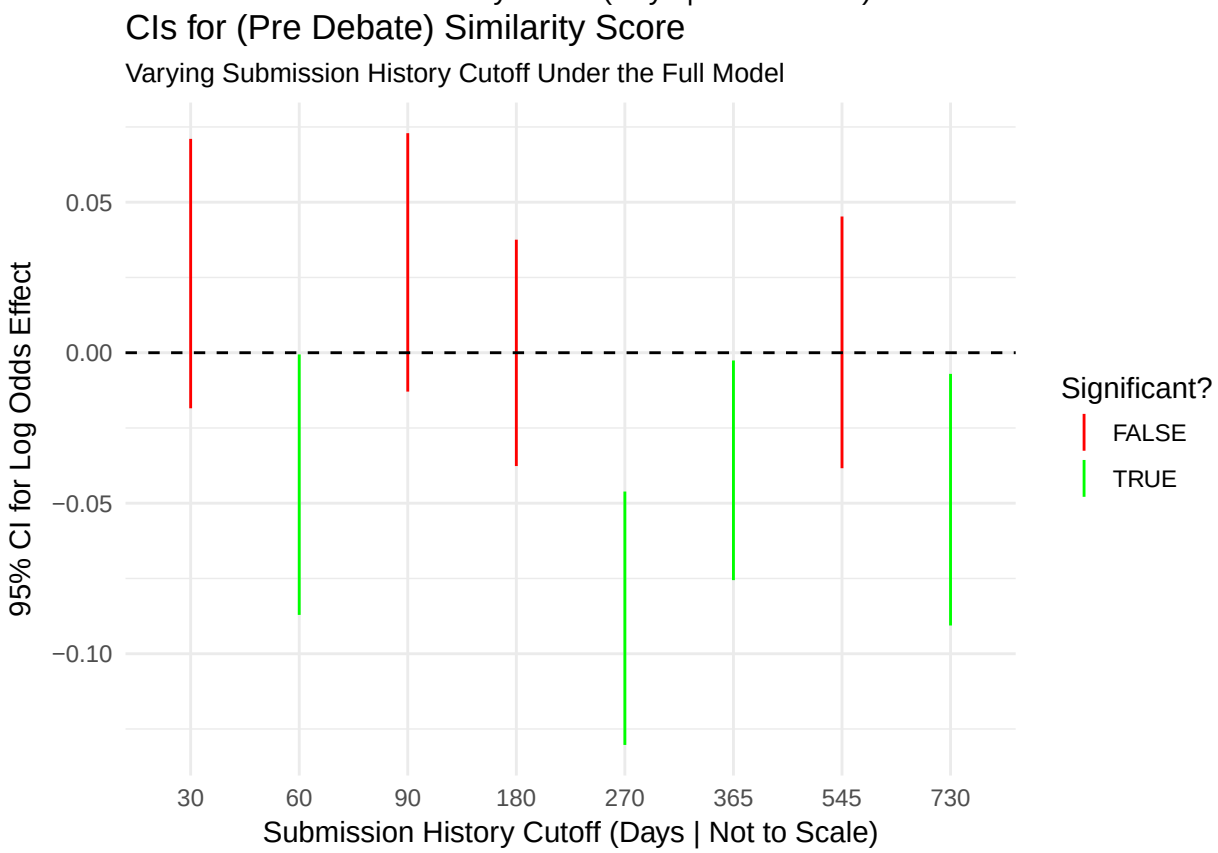
Refuting or Fleshing Out Previous Findings

In some cases, the specific cutoff chosen under the standard approach resulted in variable significance and effect direction that, in light of the cohort approach, seem to be incorrect or part of a larger pattern.

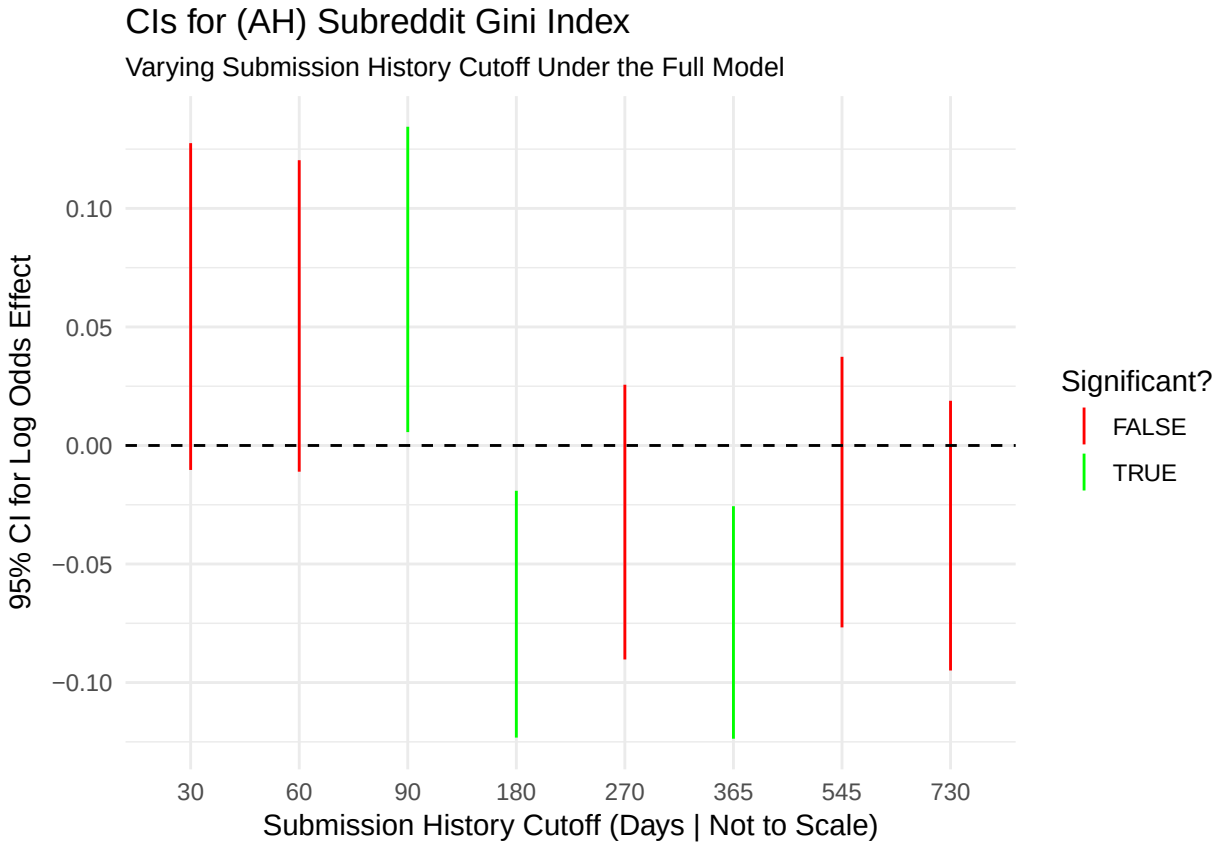


For instance, for the # Unique subreddits posted in, the standard approach showed a negative significant effect, but the cohort approach shows that this significance and direction instead appear to be an anomaly of the data.


CIs for (AH) # Submissions within Cutoff before 1st CMV Post
Varying Submission History Cutoff Under the Full Model
95% CI for Log Odds Effect
0.0
−0.1
−0.2
−0.3
30
60
90
180
270
365
545
730
Submission History Cutoff (Days | Not to Scale)
Significant?
FALSE
TRUE


CIs for (Pre Debate) Similarity Score
Varying Submission History Cutoff Under the Full Model
95% CI for Log Odds Effect
0.05
0.00
−0.05
−0.10
30
60
90
180
270
365
545
730
Submission History Cutoff (Days | Not to Scale)
Significant?
FALSE
TRUE

For # Submissions within the history cutoff, the opposite is true. That is, within the confines of the standard approach, this variable seemed insignificant, but the cohort approach bears out that under most other cutoffs, this variable is significant, although in a slightly counterintuitive direction. Under the self-affirmation framework, one might expect the opposite effect, that the "self-affirmation" that occurs via a subject author making submissions to make it more likely for the first opinion to change. However, even adjusting for the similarity between these submissions and the first opinion, the effect is negative. The similarity score however behaves mostly the way it did under the standard approach.



The behavior of the subreddit gini index under the cohort approach is perhaps the most interesting. Increasing the cutoff from 90 to 180 days (and beyond) results in an apparent effect flip. If only taking earlier history into account, it seems that more concentrated submission activity within a few subreddits (higher Gini) increases the odds for a first opinion change, but if that behavior is born out past half a year, then the odds actually decrease. Clearly, any blanket notions about subreddit submission "equality" must be evaluated in the context of the time period when the submissions were made.

Limitations

Much of the effort of this thesis lies in the infrastructure I created to gather the raw data from Reddit, through its API. Thus, the relevant limitations of that API undercut all of the data provided in this study. In the data section, I have already discussed the issue with not being able to access certain data due to author or moderator removal, but the Reddit API also provides limits on access to user submissions. Through the API, one can only access 1000 of a Redditor's most recent, "hottest", "top", or "controversial" comments. Thus, if a user has over 1000 submissions, there is no guarantee of being able to retrieve *all* of the rest of that user's submissions (by finding union of submissions found under the other 3 sorting criteria). Fortunately, most users have under 1000 submissions, but the same cannot be said for comments, which face the same issue. Since users in general post more comments than submissions, more users have over 1000 comments, and thus utilizing comments becomes fraught with a missing data problem on an semi-known scale. This is why this thesis eschews the use of comments as part of a subject author's history.

For many of these issues, I suspect whatever features internal Reddit teams must have allowing them to provide targeted advertising would be an effective pancea. An insider with a more solid assurance of "data completeness" would be very well positioned to thoroughly explore r/ChangeMyView, it's opinions, norms, and participants.

Extensions

r/ChangeMyView has existed since 2013, and thus the most obvious extension would be to include *all* possible data for review. Though, as previously discussed, increased complexity of time trends as well and the increased likelihood for missing data would loom even larger over the endeavour. Assuming that access to a version of "insider" data is impossible for researchers, the Reddit API could be utilized to ensure access to higher quality future data. In realtime, authors of CMV submissions could be identified, and their prior submission and comment history gathered *then*, to minimize missing data through API limitations or moderator/user removal. Comments of these submissions, and alterations to these comments could also be tracked, ensuring a more complete history of activity on r/ChangeMyview.

Additionally, my thesis, with the exception of a few features, almost completely ignored the *context* in which a subject author's prior submissions took place. For example, the language of other Reddit users commenting on a subject author's prior submission may have had some influence as the direct feedback to that author's participation. While including such information would contribute to the already sizeable parameter space of this thesis, it would also provide a more detailed means to differentiate between submission history.

More attention could be paid to CMV's propensity for topics to be influenced by news cycles, additional NLP and time series tools could be utilized in an attempt to characterize the relationship between CMV, news, and opinion malleability. Thorough analysis of CMV's self-documenting structures like the r/Deltalog and the Deltaboards would prove useful in this respect.

Conclusion

This study attempts the nigh insourmountable task of predicting opionion malleability for a self-selecting group of Reddit users, all while effectively eschewing a detailed analysis of the relevant debate and opting to rely on the user histories. Coupled with the inherently difficult task is the ephemeral nature of the data being collected and a complicated pre model-training parameter space to determine what history to analyze, and nuances of natural language processing topic models to use. Unsurprisingly, traditional performance metrics like ROC, sensitivty, and specificty reinforce the difficulty of such a task, but also display that traction for the problem is indeed possible. Opting for a more interpretable model allows for an analysis based on self-affirmation theory, pronoun usage, and relevant Reddit-specific heuristics.

The analysis has reiterated the relevance of self-affirmation, pronoun useage, and opinion timing for r/ChangeMyView opinion malleability, while also drawing attention to notions of participation "diversity" by utilizing features like unique subreddits submitted to, and subreddit gini index. Additionally, this study has provided some evidence that subjects on CMV with more malleable opinions are easier to detect with pre-debate information and author history compared to change resistant subjects. Hopefully the effort undertaken here demonstrates the novelty of the data available at r/ChangeMyView, and while that novelty doesn't engender the findings here to be generalized in other important debate spaces, it is an intriguing place to explore the relationship between opinion malleability, persuasion related theory and the unique aspects of Reddit and CMV itself.

Appendix

Full Feature Description

Author History

As the title suggests, this group of features focuses on the subject author's submission history.

Submissions within 1 year before 1st CMV Post

As alluded to above, this feature is a simple count of the subject author's submissions within a year before his or her first CMV Post. Since all the submissions included in this feature are subreddits that are *not* CMV, it fits nicely within self-affirmation's conception of success in an unrelated field, especially when paired with other author history features like "Mean Submission Score".

Unique Subreddits Posted In

For the subject author's submission history, this counts how many distinct subreddits the author participated in before posting to CMV.

Average Previous Submission Date

This feature seeks to establish the chronological "center" of the subject author's submission history.

Edits Per Previous Submission

If a subject author edits one of their previous post, any number of times, this can be tracked with the Reddit API. Edits occur for a variety of reasons, and may help indicate a user's responsiveness to comments on their submissions. For CMV submission this is usually the case, so I track previous, potentially similarly motivated behavior with this feature.

Fraction Plural First Person Pronouns

Fraction Singular First Person Pronouns

Mean Plural First Person Pronouns

Mean Singular First Person Pronouns

Tan et al. (2016) found:

> First person pronouns are strong indicators of malleability, but first person plural pronouns correlate with resistance. . . individualizing one's relationship with a belief using first person pronouns affirms the self, while first person plurals can indicate a diluted sense of group responsibility for the view

In light of this, I have included a number of features concerning plural and singular first person pronouns, with the mean based ones being more sensitive to submission length (and thus, in many cases, quality), and the fraction based features being agnostic to the number of submissions and their length.

Mean # of Words

This feature serves as a proxy for mean prior submission quality, and is also included to calculate first person pronoun fractions.

Mean Previous Submission Sentiment

Utilizing the R package "sentimentr" I calculate a negative-positive (-1 to 1) sentiment score and take the average across the subject author's previous submissions.

Mean Submission Score

This is simply the average of all the subject author's previous submissions. While Reddit upvote/score manipulation is impossible to rule out, this helps to delineate the "success" of self-affirmation provided by the submissions.

Removed CMV Subs 1 Year before 1st CMV Post

This feature counts the subject author's previous "unsuccessful" earlier attempts at posting to r/ChangeMyView. CMV posts counted in this feature are those that are removed for some violation of the rules of the subreddit. Given the self-selective nature of CMV participants, unsuccessful earlier integration into the community as measured through removed CMV posts could be an important feature to keep track of.

Submissions with Content 1 Year before 1st CMV Post

Not all submissions are created equal, at least compared to the higher quality discussion and opinion explanations found on CMV. This feature counts the number of subject author submissions that actually have textual content beyond, say, a hyperlink.

Subreddit Gini Index

This feature is a measure of subreddit submission inequality. This feature helps to differentiate between subject author's who focused a lot of their submissions in a few subreddits versus those who diversified the subreddits they posted submissions in.

Pre Debate

This features are based on the subject author's first opinion submitted to CMV, but utilizing information that could only be gleaned at the time of the submission, before any debate initiated by other user's could take place. Together with the Author History features, these groups comprise the main thrust of the study, to test the predictive capability of subject author information from their history and from their opinion on opinion malleability.

Words

Counts the number of words contained within the CMV Submission.

CMV Submission Date

Tracks the date the subject author submitted their first opinion to CMV. (UTC Epoch)

CMV Submission Hour

Tracks the hour of the day the opinion was posted. (UTC 24 hours)

Fraction Plural First Person Pronouns

Fraction Singular First Person Pronouns

Total Plural First Person Pronouns

Total Singular First Person Pronouns

These features track the fraction and totals of first person plural and singular pronouns of the subject author's first CMV Post.

Sentiment

Utilizing the R package "sentimentr" I calculate a negative-positive (-1 to 1) sentiment score of the subject author's first CMV Post.

Similarity Score

This features represents the similarity between two documents: (1) The subject author's previous submissions within a year before their first opinion post. (2) The subject author's first opinion post.

Utilizing the R package "text2vec", I calculate a score between -1 to 1 (least to most similar). This feature aims to serve as a proxy for how "unrelated" the subject author's previous submission content compared to their first opinion on CMV. Given the generally open nature of Reddit user submission histories, this feature helps exploit the opportunity to study self-affirmation in an online context.

Topic {K} Score

Before the debate surrounding a CMV submission occurs, the reasoning supporting a given opinion is available to browse. The subject matter of the reasoning and opinions varies widely. Accordingly a Latent Dirichlet Allocation (LDA) topic model is applied to *all* CMV opinion submissions in 2016, excluding those submissions whose content is unavailable (removed or deleted posts), and meta-subreddit announcement and discussion posts. Using the LDA topic model, the subject author's first opinion submissions are scored on K given topics (

$$sum_{i=1}^{K} topic\_score_i = 1$$

). These score features are used to model the differing subject matter within the CMV 2016 submissions and how the subject author's first opinions fit into such a model.

Post Debate

After the debate surrounding a CMV post has taken place, a variety of further information can be gleaned from the contents comprising the debate, and in some cases, moderation. In this study, the Post Debate variables only include the outcome of interest and a few other variables that are included in a model that serves to benchmark the effectiveness of Pre Debate + Author History model. For a more thorough examination of "Post Debate" and debate dynamics in general in CMV, please see (Tan et al. (2016)).

However, simply because these variables are gathered after the debate and archiving process have taken place, this does not mean that the information gathered has no bearing of characteristics that may have affected the subject author's decision making or the debate itself.

For example, measuring how many times the subject author commented on their first opinion submission, while only identifiable after the debate is finished and the post is archived, one could think of this variable as partially measuring the propensity of the author to engage and debate with other Redditors. Surely, some of this characteristic is determined by the quality and subject matter of the comments that "draw out" the subject author's engagement, but some of it must be an aspect of the author that he or she had before the debate, as well, as during, and after it.

Opinion Change?

This is the outcome of interest for the study. This feature tracks whether the subject author awarded at least 1 delta in their submission, indicating that the subject author changed their opinion.

As alluded to previously, it is possible for comments on an CMV opinion to be awarded deltas from Reddit users besides the subject author. Presumably, Redditors who awarded such deltas share at the general opinion that the subject author has, but because these deltas are relatively rare and there is no mechanism to glean these Redditors' reasons why they hold the opinion, such deltas are not taken into account.

 # Direct Comments

Direct comments are those Reddit comments whose "parent" is the subject author opinion submission. In other words, these are the comments directly attempting to change the opinion. In general, the number of direct comments to a CMV Post is a reliable indicator of the level of debate that occurred. Including such a feature in a model helps it to differentiate posts that may have not received a delta due to low debate activity.

 # OP Comments

This feature simply tracks how many comments the subject author, as the original poster (OP) of their opinion submission, submitted during the discourse. Inclusion in a model helps to determine the role subject author participation played in the malleability of an opinion.

 # Total Comments

This includes all the comments on the subject author's submission. These comments may reply to other comments attempting to change the subject author's, besides direct comments and OP comments, this feature also captures comments whose parent is not the opinion submission itself. These comments reply to other comments, whether they be direct comments, OP comments, or arbitrarily nested comments in a comment tree. This feature helps to flesh out the "debates within the debates" that take place in a CMV post.

Citations

Cohen, Geoffrey L, Joshua Aronson, and Claude M Steele. 2000. "When Beliefs Yield to Evidence: Reducing Biased Evaluation by Affirming the Self." *Personality and Social Psychology Bulletin* 26 (9). Sage Publications Sage CA: Thousand Oaks, CA:1151–64.

Correll, Joshua, Steven J Spencer, and Mark P Zanna. 2004. "An Affirmed Self and an Open Mind: Self-Affirmation and Sensitivity to Argument Strength." *Journal of Experimental Social Psychology* 40 (3). Elsevier:350–56.

Dillard, James Price, and Michael Pfau. 2002. *The Persuasion Handbook: Developments in Theory and Practice*. Sage Publications.

Duggan, Maeve, and Aaron Smith. 2013. "6% of Online Adults Are Reddit Users." *Pew Internet & American Life Project* 3:1–10.

Fogg, BJ. 2008. "Mass Interpersonal Persuasion: An Early View of a New Phenomenon." In *International Conference on Persuasive Technology*, 23–34. Springer.

Pennebaker, James W, and Laura A King. 1999. "Linguistic Styles: Language Use as an Individual Difference." *Journal of Personality and Social Psychology* 77 (6). American Psychological Association:1296.

Tan, Chenhao, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. "Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-Faith Online Discussions." In *Proceedings of the 25th International Conference on World Wide Web*, 613–24. International World Wide Web Conferences Steering Committee.

Wood, Molly, Hughes, Stephanie, and Ainpour, Shaheen. 2018. "Reddit's Cto on Shedding Its "Dystopian Craigslist" Vibe." https://www.marketplace.org/2018/05/14/tech/reddit-cto.