Opinion Malleability In r/ChangeMyView :

Self-affirmation Theory and Submission Behavior

Julian McClellan | jmcclellan@uchicago.edu

Due 4/22/17

Abstract

How can you tell if someone will change their opinion? Using data from the popular debate subreddit r/ChangeMyView in 2016 and weighted logistic regression I utilize the wording of opinions, and the history of their authors, to answer this. The inclusion of author history in this model results in improved malleability predictions compared to earlier attempts. The model results provide evidence that self-affirmation's effects on open-mindedness extend past a person's immediate past. Other significant effects include previous opinion submission experience, diversity of subreddit participation, and sentiment of opinion and submission history.

Introduction

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 in a submission in an attempt to change the opinion presented in the submission. If a user posts an opinion, they are encouraged to conduct a dialogue with the respondents. The rules of CMV explicitly forbid low effort comments, and an active team of moderators stringently enforce all 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", along with a brief explanation.

Figure 1: Example of an Opinion

Figure 2: OP Awards a Delta and Interacts with Respondents

Thus, the CMV subreddit allows for access to the reasoning behind a person's views, the full 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 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 is: "Using what we know about a CMV participant, can we predict if they will change their opinion?"

Using submissions and author histroy scraped from CMV activity in 2016, I construct several relevant features to help predict an opinion change/delta awarding. Like Tan et al. I utilize the wording of the opinion itself, but I also go beyond and construct features from the OP's prior submission history. Like Tan et al. I create some features linked to self-affirmation theory in psychology, but also utilizes features specific to the place of submissions in Reddit overall.

I utilize weighted logistic regression models and outperform earlier attempts to predict the opinion change on CMV. Whereas previous psychological studies have found the effects of self-affirmation in the immediate past to be statistically significant in predicting opinion malleability (Cohen, Aronson, and Steele 2000), the models here find that self-affirmation features utilizing the entirety of an OP's submission history, dating back months and sometimes years, also have signifiant effects in the same domain. Thus, this paper provides the first evidence to suggest that self-affirmation has such far-reaching effects.

Literature Review

Persuasion and Malleability

Before the advent of social media websites like Facebook or Reddit, research efforts into persuasion (and by extension, malleability) 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, (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.

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 out of sample ROC of `.54` with their best model. 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

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 submission in question.

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. . .

> . . . 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.

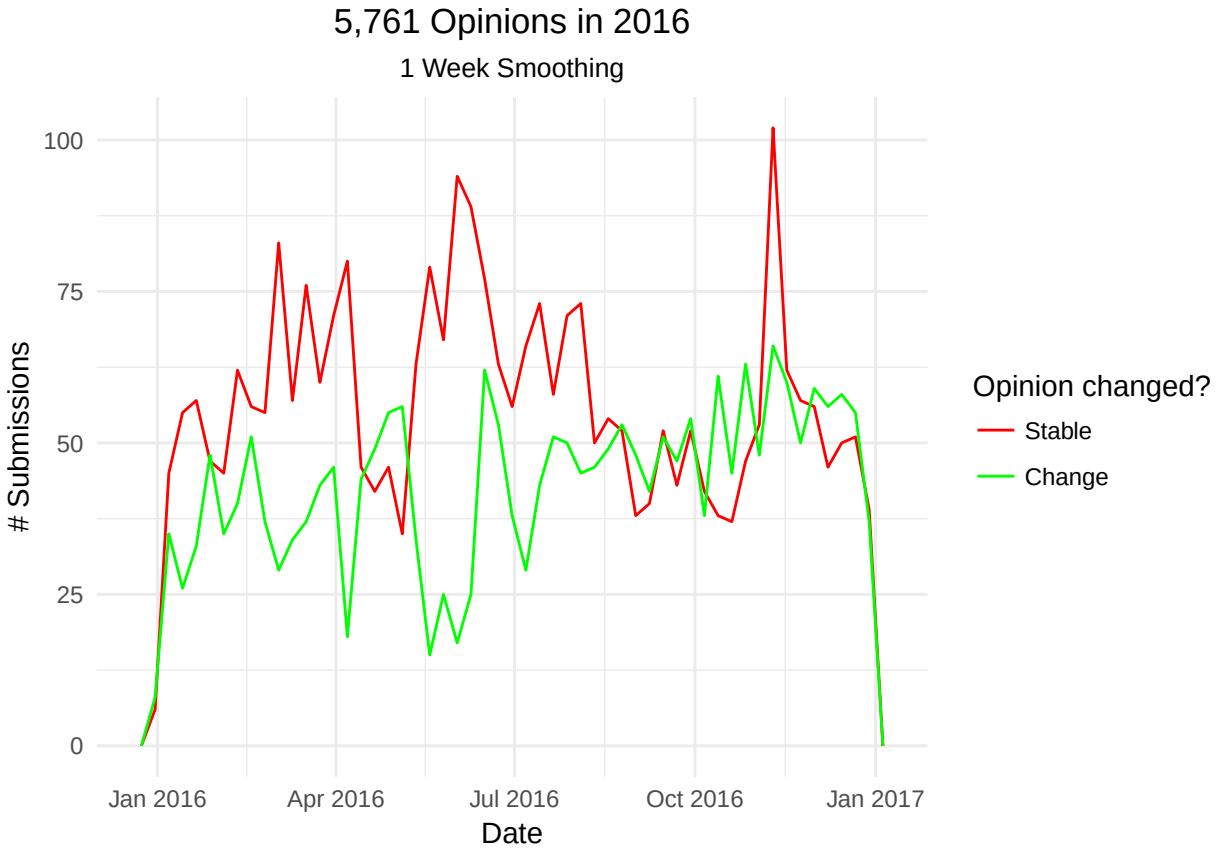
Utilizing a Reddit user's past submission history, on the other hand, not only allows a more extended look into instances in which self and external affirmation may have occurred, but it also provides instances of participation with the community that the view is being exposed to. If a Reddit user posited their opinion to CMV, but also had previous submissions within the CMV community, then that participation would be relevant, but previous participation in other subreddits can also be more or less relevant as well. CMV does not establish any barriers (besides following the subreddit rules), against first-time participation from other Reddit users new to CMV. If a user posts an opinion on CMV, but has also attempted to change an opinion on CMV, then it's possible that these attempts may or may not have changed the view (receiving or not receiving explicit recognition).

Tan et al. extended the study of self-affirmation theory into an online debate space. Utilizing an opinion author's past submission in Reddit can possibly extend evidence of self-affirmation theory's effects, by looking further back in time to potential affirmation, but also help determine the properties of malleable opinions and open minded authors.

Data

Using the Python Reddit API Wrapper (PRAW) the data was collected from the subreddit, r/ChangeMyView, and specific Reddit user histories. All `5,761` views posted in 2016 were gathered, and the complete submission histories of the `4,142` unique authors in 2016 were also gathered. For each submission data was collected on "Pre Debate" features, i.e. those

features that can be gathered as soon as a view was posted to the subreddit. These include the creation time of the submission, the number of words it contains, excluding the title, the sentiment of the submission, as measured by Valence Aware Dictionary and Sentiment Reasoner (VADER), and the raw number of singular and plural first person pronouns were collected, and the fraction of the total words they comprised. A few "Post Debate" features were also gathered. These are the number of comments the OP made on their own submission, the number of direct replies to the OP's opinion, the total number of user comments in the post, and to provide an accurate classification of a stable or changed opinion, whether or not the OP gave a delta. A timeline of opinion stability and change for 2016 is given below and an web application for exploring the dataset is available here.



However, the majority of features concern an OP's "Author History", concerning their submission history prior to posting on r/ChangeMyView. In general, these features are either aggregate statistics, like the total number of singular/plural first person pronouns, or mean statistics, and often there are aggregate and mean versions of similar features, like the mean number/fraction of first person pronouns used in prior submissions. Many author history features, such as first person pronouns and submission score are related to self or external affirmation, whereas other features are statistics that capture Reddit specific submission

behavior or other statistics.

The number of empty submissions an author has is tracked since an empty submission obfuscates the collection of pronoun or sentiment features (there is no content to analyze). A related feature is the number of submissions with available content. This feature mostly captures submissions with simple hyperlinks instead of explicit textual content. The average level of textual content an author produces is tracked by the mean number of words in a submission. There is also a gini index of subreddit participation. One might expect that a user who submits to subreddits unequally, i.e. a user that favors participation in a select few subreddits, may be less open minded, as their participation to the Reddit community lacks a certain diversity.

Data Filtering

As seen above, there were more submission made to r/ChangeMyView (`5,761`) than there are unique authors (`4,142`) in 2016. Thus, if attempting to utilize all `5,761` CMV submissions, very similar histories will be used between two or more submissions by the same author introducing collinearity in the model. To eliminate this, the data is filtered so that only the first submission an author made in 2016 is included with the data. Additionaly, the majority of the features pertain to an author's history, so if the author of a submission does not have a submission history prior to their first r/ChangeMyView post in 2016, then that submission is dropped. Lastly, submission are also dropped if there are no responding comments, and hence no oppoprtunites at all for the opinion to change. Unlike Tan et al., I do not implement stricter restriction on the dataset (10 unique responders and at least 1 OP response).

Utilizing these filtering procedures, `3,618` opinions and authors remain in 2016, each with unique submission histories. A summary data table for the most important features of these filtered opinions is given below. A number of features were created by interacting the total number of submissions and mean statistics and utilized in the model, and their statistics are not featured for brevity.

Table 1: 2016 r/ChangeMyView Unique Author Opinion Features

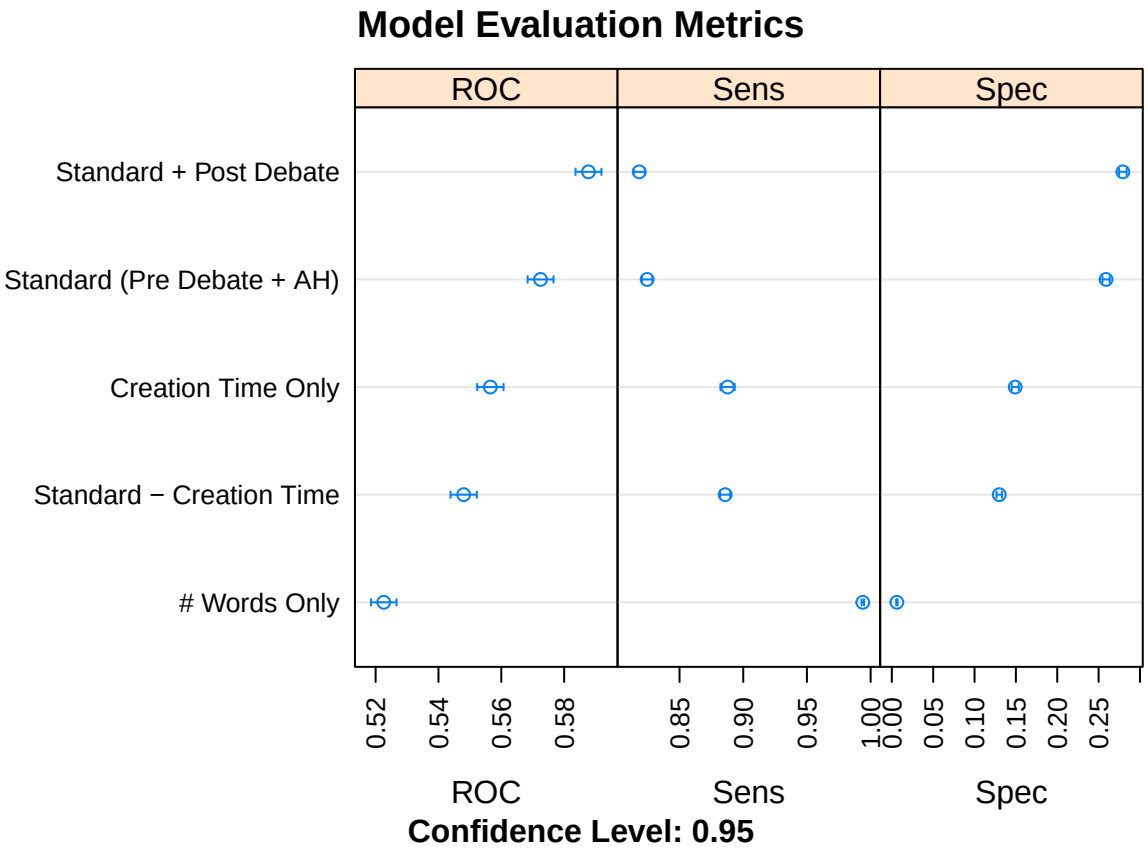
Statistic	Mean	St. Dev.	Min	Max
(Post Debate) # OP Comments	13.1	15.2	0	274
(Post Debate) # Direct Comments	13.2	14.2	1	224
(Post Debate) # Total Comments	130.1	233.1	1	4,903
(Pre Debate) Plural First Person Pronouns	1.8	3.3	0	43
(Pre Debate) Fraction Plural First Person Pronouns	0.005	0.01	0.0	0.1
(Pre Debate) Singular First Person Pronouns	1.1	2.0	0	26
(Pre Debate) Fraction Singular First Person Pronouns	0.003	0.01	0.0	0.1
(Pre Debate) # Words	368.1	294.4	1	3,785
(Pre Debate) Sentiment	0.2	0.8	−1.0	1.0
(AH) # All Prior Submissions	107.9	171.0	1	1,804
(AH) Mean Submission Score	52.4	169.9	0.0	5,271.2
(AH) Subreddit Gini Index	0.4	0.2	0.0	0.9
(AH) Fraction Removed Submissions	0.1	0.1	0.0	0.8
(AH) Fraction Empty Submissions	0.4	0.2	0.0	1.0
(AH) Mean Submission Sentiment	0.1	0.2	−1.0	1.0
(AH) Daily Submission Frequency	0.03	0.2	0.0	7.7
(AH) Number of CMV Submissions	1.1	4.0	0	91
(AH) Fraction of CMV Submissions	0.03	0.1	0.0	1.0
(AH) Mean Singular First Person Pronouns	0.9	1.9	0.0	80.0
(AH) Fraction Singular First Person Pronouns	0.01	0.01	0.0	0.1
(AH) Mean Plural First Person Pronouns	0.6	1.3	0.0	24.0
(AH) Fraction Plural First Person Pronouns	0.003	0.004	0.0	0.1
(AH) # Submissions with Available Content	50.9	84.9	1	1,096
(AH) Mean Number of Words	159.9	152.4	1.0	3,464.0

Models

Much like Tan et al. and psychologists who have studied self-affirmation theory, I want to establish the direction of the effect of the self-affirmation features included in the data. Since the variable of interest is binary using iteratively reweighted least squares logistic regression seems an ideal choice.

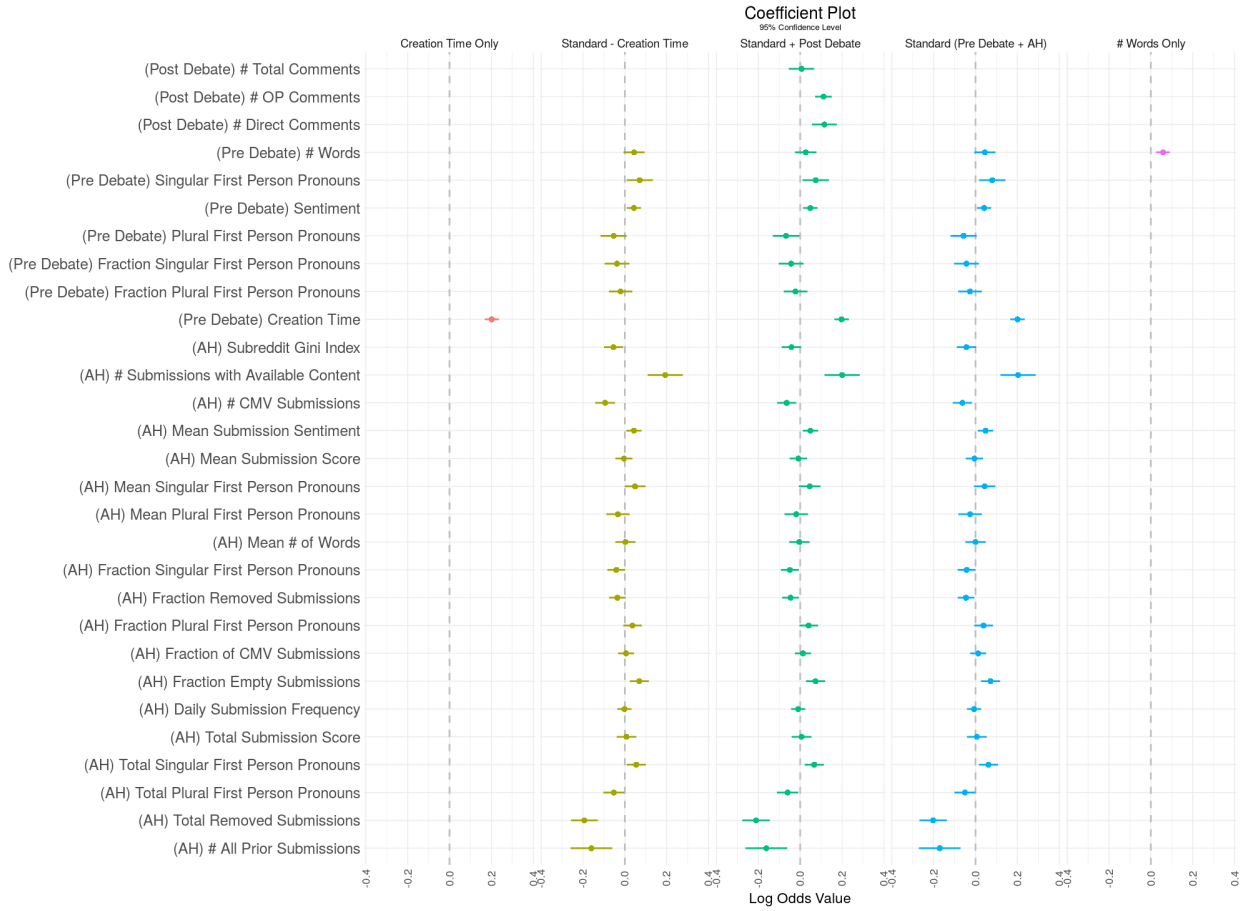
5 separate logistic regression models are tested using 20 repetitions of 10 fold cross validation. The results of these 200 different repititons for the models are used to construct 95% confidence intervals for model performance metrics: ROC, Sensitivity, and Specificity. There are two baseline models that only utilize the number of words in an opinion and the creation time of that opinion, respectively, and there are three "full" models: a "standard model" utilizing all the pre debate and author history variables, the standard model without creation time, and the standard model with all the features, including post debate features. All models are trained on centered and scaled data.

Results



Comparing the ROC's between the various models, the creation time of a submission is a surprisingly large contributor to ROC. In all models except "# Word Only", Tan et al.'s best AUC (.54) is outperformed with the standard model achieving ~.57 mean AUC, even without as strict of a pre-filtering procedure they utilized. However, complete superiority of the model's proposed here can't necessarily be established, as Tan et al. had access to more submissions, but in previous years, and they also utilized a single held-out test set versus the repeated 10-fold cross-validation used here.

Despite the low ROC across the board, with a 0.5 cutoff, the standard model is able to achieve a mean sensitivity (true detection rate of opinion changes) of 82%, and a specificity (true detection rate of opinion stability) of 26%. The model fit metrics show that detecting opinion malleability is a difficult task indeed.



The coefficient plot displays a number of significant features. Focusing on those that are specifically testing self or outside affirmation, the "(Pre Debate) Singular First Person Pronouns" has a statistically significant positive effect on the log odds of an opinion changing, which conforms with Tan et. al's findings. The total number of singular and plural first person pronouns in an author's history have statistically significant positive and negative effects on the log odds. The mean number of singular first person pronouns in an author's history also have a significant positive effect. These findings are also consistent with Tan et. al, but notably, it provides evidence that the effects of self-affirmation extend beyond the immidiate past.

Two subreddit participation features have interesting significant effects. The number of previous CMV submissions by an author actually has a significant negative effect on opinion change probability. This warrants more investigation to say exactly why, but it might be that initial views submitted to CMV are the "low hanging fruit", and are easier to change, but the user eventually runs out of these sorts of views, and continued participation involves the submission of more resistant views. The subreddit gini index also has a (just barely)

significant negative effect. Suggesting that equality in submission between subreddits might be a feature of more open minded view holders.

There are other significant features as well, and they mostly capture the effects of Reddit specific behavior. For instance, the total number of removed submissions in a user's history have significant negative effects on the odds of an opinion changing. A submission can be removed by the moderator of a subreddit, and this typically happens if a moderator determines that the submission is against the rules of the subreddit. This result is rather intuitive, as one would expect that users who have trouble posting in accordance with Reddit community rules to be less likely to credit the CMV community for changing their opinion. Of course however, these models cannot prove causality. The total number of prior submissions also has a significant effect, though in the negative direction. Here, an intuitive explanation is a little harder to come by, one might posit that more submissions make users more "jaded" to interaction with Reddit, and less susceptible to persuasion by the community as a whole. More positive sentiment, both in the wording of the opinion (Pre Debate), and the mean positive sentiment across previous author submissions (AH), both have significant positive effects for the model's prediction of an opinion change.

Returning to the creation time feature, the coefficient plot shows that it has a significant positive effect, in all models it appears in, on the log odds of an opinion change. This suggests an increasing trend in opinion changes during 2016. However, looking at the initial timeline presented in the data section, a polynomial trend might be more appropriate. Similarly, there are other variables where we might expect a polynomial trend to be more appropriate. For many of the self-affirmation features, we might expect an *overuse* of the first person, or really high submission scores to perhaps some sort of "ego" effect.

However, in utilizing a variety of modelling techinques with non-linear predictor effects baked in (bagged/boosted decision trees, random forest), performance measures did not differ much from the models above. Perhaps, if more data were collected across a longer timespan, these methods might have better performance than the models utilized here.

Conclusion

The data contained in this analysis only pertain to CMV activity in 2016, and even then only `3,618` (`62.8%`) of the submissions that year are included. Although I utilized less stringent filtering procedures than Tan et. al, with data ranging from 2013-2015, they still were able to utilize `10,473` submissions as training data in thier study of opinion malleability.

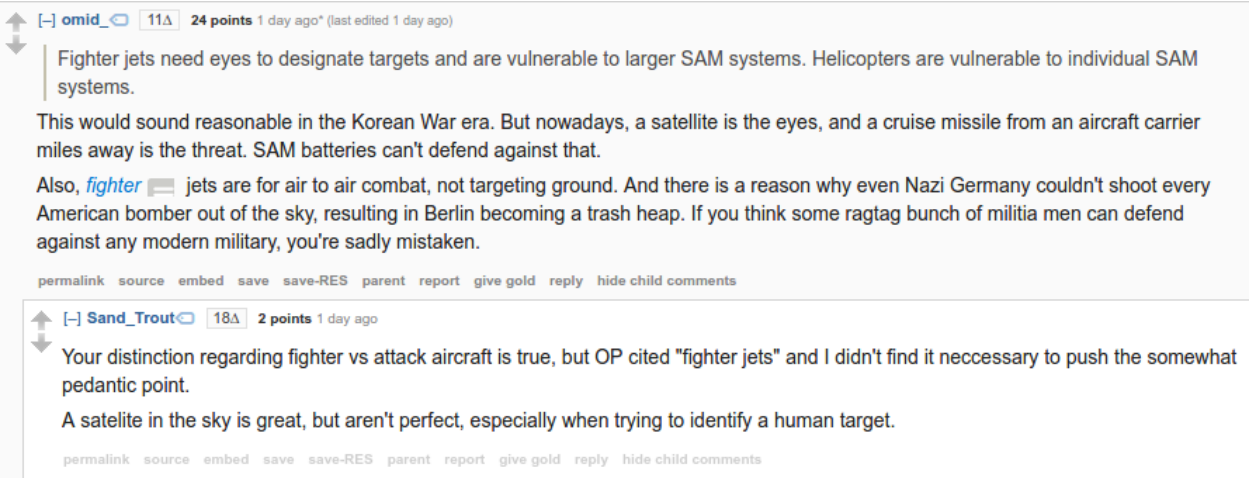


Figure 3: Delta 'reputation' displayed next to user handles

Still, the results motivate utilizing more data and a more extensive use of CMV author history in opinion prediction. More features can be extracted from an author's submission history to account for affirmation sources. For instance, comments that the author makes in their own submissions can have their scores and first person pronoun useage extracted to provide a more complete view of external and self-affirmation in the context of each individual submission. The subreddit gini index used here, which only utilized submissions, could be pooled and/or compared with the *comment* subreddit gini index. In general, including comments is a promising way to extend the analysis of a user's past Reddit participation, as users typically have more comments than submissions.

One aspect of that past Reddit participation that could prove important, is a CMV participant's "reputation" within the subreddit. The number of deltas the user is awarded over time are displayed next to that user's name when they submit a post, signalling the user's experience and success in changing views. CMV has a "deltaboard" to track its userbases history of delta awards, so it is possible to determine the reputation the user at certain points in time. With this information, the effect of success in changing other's views on changing one's one views could be evaluated.

Finally, although the data's size and features could be heavily extended, the presented analysis of the submissions and authors in r/ChangeMyView explores a unique dataset that contains, explicitly coded, the success or failure of persuasion, as well as a nearly exhaustive history of author participation in Reddit. Using this data, I provide evidence that a user's collective past instances of self-affirmation (as gathered from their submission content) correlates with more malleables opinions on CMV, and evidence that other aspects of a Redditor's submission history, like their expressed sentiment, removed posts, and diversity of subreddit

participation, can also be leveraged for predictive power. Although a larger set of data ought to be tested to verify this fact, I also outperform previous model performance in predicting opinion malleability. Further study into CMV looks promising, and hopefully more light can be shed on persuasion, opinion malleability, and self-affirmation in an online debate space.

References

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.

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

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.