∆Time: What Can We Learn From an Online Community That Has Sustained Good-faith Discussions for over 7 years?

Mir Masood Ali, Joel Miller, Chris Kanich
University of Illinois at Chicago

Abstract

Since its creation in 2013, Reddit's ChangeMyView (CMV) subreddit has hosted a discussion platform where participants who are willing to consider opposing arguments post their views online. Remarkably, CMV has maintained its reputation for hosting calm, thoughtful debate despite seeing a tremendous amount of growth in the past seven years. Discussions from the first two years of CMV were studied by (Tan et al. 2016). They generated features that could serve as potential indicators of successful persuasion, and identified factors that underlie good-faith debate.

In this paper we replicate and expand upon their analysis of the CMV subreddit. We analyze over 7 years of discussions, exploring interaction dynamics, the impact of stylistic choices, language and topic indicators, and content moderation.

We observed that results of significant features from the earlier study largely held up over time. We reasoned about the changes in results against the composition of the community, its moderation, and the topics that drove discussions.

Our results provide insights into the working of persuasive arguments and the subreddit's ability to maintain good-faith discourse. We provide recommendations of factors to consider when developing communities for discussion online.

Introduction

For all its pervasiveness and popularity, Internet-mediated conversation as a widespread method of communication is relatively young. Over the last few years, researchers have been attempting to understand factors that make and break such discourse. The team of (Tan et al. 2016) analyzed two years (2013–2015) of posts on a structured online debate community looking for indicators of persuasive arguments. They evaluated the strength of various conversational features and features of the overall discussion as indicators of a change of opinion, and were able to determine what types of language leads to productive discussions within this community - everything from the length of the argument to the use of italicization formatting.

While Tan et al.'s conclusions are compelling, and two years is a substantial amount of time to analyze, large communities on the Internet are highly dynamic. The population, topics, and access modalities for such a community will almost certainly shift over time, and even a change of seven years' time on the Internet can likely lead to a qualitative change in the community itself. Due to the central importance of civil, productive discussion to maintaining a healthy body politic, replicating this work to validate its scientific value and extending this analysis up to the present day is a worthwhile task. Furthermore, such a replication can help us investigate whether a community can maintain a civil and effective space over extended periods of time; indeed, the long-term health of such conversation media is almost certainly tied to the long-term health of the human communities that are increasingly using online conversation platforms to shape their public discourse.

Tan et al.'s analysis focuses on the ChangeMyView (CMV) subreddit: for the last seven years, it has allowed people with controversial opinions to share their view and perspective with other active members, providing a platform for numerous people to challenge these opinions. What would otherwise appear to be a recipe for disaster has, surprisingly, emerged as a platform for balanced, nuanced discussion that has stood the test of time.

While the CMV community's success shows that civil and seemingly productive discourse is possible, the sustainability of such efforts is unclear. The ever-increasing volume of conversations online, along with the dynamic nature of the platform and the ever-evolving set of topics under discussion make it far from a foregone conclusion that such effective conversation will continue in perpetuity.

Accordingly, we seek to answer the following research questions in this work:

- **RQ1:** Has the content/style of interactions in the community remained consistent across time, providing indicators for communities to encourage in good-faith discussions?
  1. Does an interplay between opinions and arguments provide insight into the homogeneity of participants of discussions in the community?
  2. Which linguistic/stylistic features, adopted by the participants of the community, provide insight into its ability consistently deliver changes in opinion?
  3. Does the content of a post provide definitive indicators that an opinion is malleable?

- **RQ2:** Can external higher-level trends in the community provide insights into the its sustainability?
  1. Does effective moderation of content within the community, which has oft been attributed to its success, explain the ability of CMV to sustain civilized debate?
  2. How does a topic's popularity line up against the malleability of opinions held within the same topic? Can activity and moderation actions indicate the community's willingness to engage in difficult deliberation?

We explore the above questions not only to find indicators that explain successful discussions in CMV, but to also provide a set of recommendations that communities can encourage and adopt to replicate effective participation in debates. (See Discussion on Page 8.)

In the process of answering these questions, we also successfully replicate the results of Tan et al.'s 2016 study of conversational dynamics in CMV between 2013 and 2015, and extend this analysis to the years 2013-2020.

Background and Related Work

Social networks are dynamic platforms that host all manner of online communications. The US Agency for International Development has identified that these platforms are increasingly becoming the *de facto* medium for communication among citizens, between politicians and their constituents, and between businesses and customers (Yesayan 2014).

While most social networks may have been created with the intention of facilitating productive and friendly discussion, they are infamous for being cesspools for harassment, trolling and bullying (Justin 2018), and occasionally large swaths of fake users (Twitter Public Policy 2018) or even entire social networks (The British Broadcasting Corporation 2014) have been created in covert attempts to inflame or influence a population.

Websites like Twitter, Wikipedia, and StackExchange have been widely studied in attempts to detect the signs and causes of conversations gone awry. A related vein of research has sought to understand the factors that influence people to decisively change long-held opinions, sometimes with the help of machine learning models (Habernal and Gurevych 2016; Zhang, Culbertson, and Paritosh 2017; Dutta, Chakraborty, and Das 2019; Luu, Tan, and Smith 2019; Dutta, Das, and Chakraborty 2020; Gleize et al. 2019; Shi et al. 2020).

Alongside conventional social media platforms, websites like Create Debate[1] and iDebate[2] were specifically built with the intention of hosting debates. Unfortunately, these websites are not necessarily paragons of effective discourse (Sridhar et al. 2015; Abbott et al. 2011; Wachsmuth, Syed, and Stein 2018). They have been further studied in (Shepherd et al. 2015), (Zhang et al. 2018), and (Sridhar et al. 2015).

The CMV community on Reddit has also been examined by researchers (Tan et al. 2016; Xiao and Khazaei 2019; Srinivasan et al. 2019) and covered in the media (Heffernan 2018) as a platform known for harboring good-faith discussions.

The CMV subreddit "is built around the idea that in order to resolve our differences, we must first understand them." Discussions in the subreddit proceed as follows: first, a member of the community posts a potentially controversial view and their reasoning for holding it (in the context of this discussion, this community member is referred to as the original poster, or OP). Other users then engage in back-and-forth discussion with the OP and other members of the community. If a user manages to change the OP's view, they reward them by replying to the view-changing comment with the Δ character.

Here's a brief primer on terminologies that recur in this paper:

- An initial, original statement of views or beliefs posted in the community is an *Original Post*.
- Users of CMV that reply to the original post are referred to as *challengers*. The content of their reply is usually referred to as an argument or a comment.
- An original post along with all the arguments under it are referred to as a *discussion tree*.
- A direct reply to the original post is called a *root reply*.
- A thread of replies between the OP and a challenger from a root reply to the end of a conversation is called a *full path*.

Several factors enhance the quality of discussions on CMV. Firstly, the community abides by 10 rules that influence the nature of discourse on the platform. For example, one rule states that any user that challenges the OP's opinion is required to do so with a comment of at least 500 characters, and other rules mandate that discussions be in good faith. Active moderators regularly delete posts and comments that they feel violate these rules. Users are provided with reasoning for the deletions and an opportunity to appeal them, but repeat offenders are ultimately banned after a maximum of three violations. On the other hand, the number of Δs awarded to each user is tracked on a leaderboard, which positively incentivizes users to engage with each other in accordance to the rules of the community. Lastly and perhaps most importantly, users are strongly encouraged to only post views if they are truly open to hearing other's opinions and having their minds changed.

All of these elements make CMV stand out as an online space that harbors particularly high-quality discussion. Accordingly, the record of CMV posts and comments can serve as a viable dataset to explore persuasion strategies.

Describing the Data

The dataset used in this paper is extracted from the ChangeMyView community. It's a popular reddit community: with 1,159,276 community members (or "subscribers"), it ranks as the 347^{th} most popular subreddit on the platform.[3]

[1] Create Debate: https://www.createdebate.com/

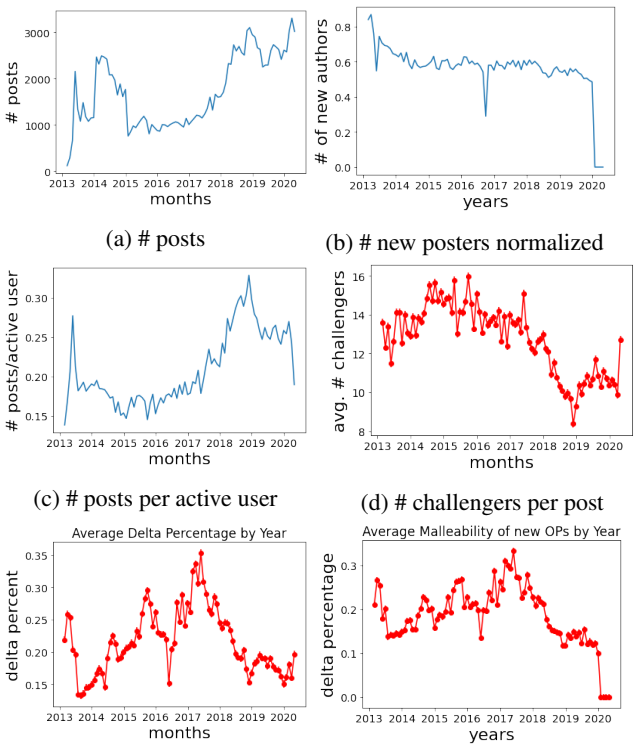
[2] iDebate: https://idebate.org/

[3] Subreddit Stats: https://subredditstats.com/r/changemyview

The dataset includes posts and comments made in the community between its inception, 01 January 2013, to 10 May, 2020. We cover 7 years and 4 months of data, while the community has remained active now for close to 8 years.

Number of discussion trees :	137,573
Number of nodes in the tree :	4,702,009
Number of OPs :	78,923

Table 1: Overall Dataset Statistics. The dataset contains posts in CMV between 01 Jan. 2013 and 10 May 2020.



(a) # posts (b) # new posters normalized

(c) # posts per active user (d) # challengers per post

(e) avg. Δs over all posters (f) avg. Δs for new posters

Figure 1: Descriptive Statistics. These graphs provide a baseline understanding of activity on CMV over time. Future Sections refer to these statistics to draw inferences from observations.

Figure 1 provides an overview of the dataset that the paper uses for its analysis in the sections that follow. The raw number of posts made in the community, shown in Figure 1a, roughly remained consistent through 2017, with an average of 1576 posts every month. However, the community saw an increase in activity starting in 2018 which has taken the average number of posts per month up to 2506 (for the 2018-2020 period).

But this increase in the number of posts is not necessarily due to new subscribers alone. The number of posts made by first-time posters each month, as shown in Figure 1b, has been consistent, averaging at around 56% of the total posts made each month. Therefore, it would appear that both first-time users and repeat users are posting in greater quantities over time.

We consider an active user to be any participant in the community (for a specific month) that made at least one post or comment in that month. Figure 1c shows the number of posts made each month as a ratio of the number of active users on the community. CMV's Subreddit Stats page shows that, since 2018, the rate at which the community has been recruiting new subscribers has increased by around 300K every year. This increase is reflected in Figures 1a and 1c.

We observed that each post in our dataset received 35 replies on average. We additionally observed a dip in the number of replies following 2018, bringing the average down to 28.9%, significantly lower than the 38.7% average of posts prior to 2018.

Figure 1e shows the probability that a post in our dataset receives a Δ. We observed that an average of 21.5% of posts made between 2013 and 2020 received a valid Δ from the OP. Although Figure 1f shows that posts made by first-time posters are roughly just as malleable, the average Δ percent is at a slightly lower, at a value of 18.67%.

Interaction Dynamics

In this section we analyze the effects that dynamics, community, and time have on the possibility of receiving a Δ. We only considered discussion trees that have received replies from at least 10 unique challengers and in which the OP has replied at least once.

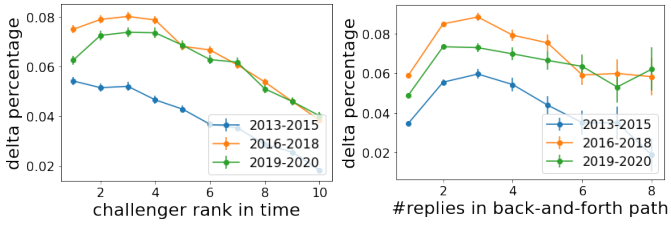
Challenger's Success

In this subsection, we explore the relationship between how challengers interact with the OP and their chance of success.

Entry Order. Does the order in which a challenger enters the discussion affect their chance at winning a Δ? (Tan et al. 2016) found that an earlier entry time is favorable, even when controlled for user experience. They stated that the first two challengers are 3 times more likely to succeed as the 10^{th}. This result still holds true: figure 2a shows that the second, third, and fourth challenger to the OP enjoy the highest probability of receiving a Δ, and have around twice as much a chance of receiving a Δ as the the tenth challenger to enter the discussion, regardless of the timeframe in which we examine the data.

Degree of Back-and-forth. Once a challenger manages to enter into a discussion with the OP, they can then engage with them in a back-and-forth interaction, expressing their points against those of the OP's. Figure 2b shows the relation between the probability of the challenger receiving a Δ given the number of such back-and-forth interactions they had with the OP.

We observe a similar pattern across year bins. The challenger has the best of chance of success at changing the OP's view in 2 or 3 back-and-forth interactions, after which the probability starts to drop. (Tan et al. 2016) had suggested that such a trend might be due to discussions becoming repetitive after a certain point, with both sides re-iterating the same arguments. We offer an alternate and compatible



(a) Δ ratio vs. entry order (b) Δ ratio vs. degree of back-and-forth exchanges

Figure 2: Figure 2a shows the probability of a commenter winning a Δ depending on their order of entry. We only considered posts that had >10 unique challengers. Figure 2b shows the probability of a commenter winning a Δ depending on the number of back-and-forth exchanges they have with the OP.

explanation: people often start debates by painting their arguments in broad strokes, and then get into discussing the details. If the big picture of an argument doesn't convince someone, it is less likely that examining the finer points of it will.

OP's Conversion

The malleability of the OP's opinion may not solely have to do with the challenger, as much as the dynamics of the discsussion as a whole. This section explores such factors and the impact they may have on the chances of a change in opinion.

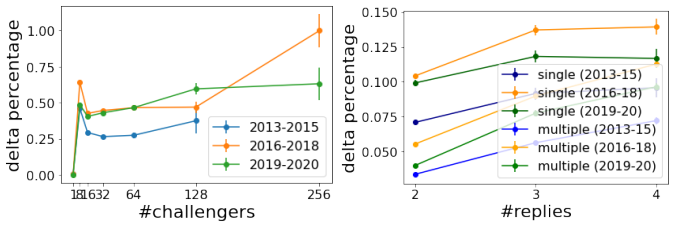
Number of Challengers. The sheer number of users attempting to change an OP's opinion might influence their malleability. Figure 3a shows a higher Δ percentage with an increase in the number of challengers. However, such an increase is only observable up to a certain point, after which the Δ percentage plateaus.

While (Tan et al. 2016) observed and reported such saturation, they witnessed a peak in Δ% much later than we did. Our results show a peak at around 8 challengers, following which we noticed a slight dip before our results displayed a plateau. We second Tan et al.'s reasoning that the value that each new challenger brings diminishes beyond a point.

Single challenger vs. Multiple Challenger Subtrees. The sheer number of challengers may not be the only factor that influences a change in view. An OP may be confronted with multiple points of view. To observe such a factor, we focused on subtrees alone, where arguments are similar and about the same topic.

We observe the relation between the Δ percentage and the number of back-and-forth replies in single challenger and multiple challenger subtrees. The number of replies are controlled to between 2 and 4 replies.

In all three year bins, the single challenger subtrees have a higher chance of winning a Δ as compared to multiple challenger subtrees. While observing similar results, (Tan et al. 2016) reasoned that when making the same argument, multiple challengers might not be adding value to the argument and may even be disagreeing with each other, thereby reducing the chance of the OP changing their mind.



(a) Δ ratio vs. # of unique challengers (b) Δ ratio in single-challenger subtrees and multiple challenger subtrees

Figure 3: Figure 3a shows the probability of a change of view of the OP in comparison to the total number of unique users that participate in the discussion under their post. Figure 3b shows the probability of an OP changing their view in a discussion involving a single challenger as opposed to one involving multiple challengers. The graph is controlled by limiting the number of replies.

Language Indicators

This section explores the relationship between the content of the arguments made against the OP and the OP's likelihood of changing their view. We consider stylistic features, language, and conversational interplay with the OP as potential factors that might influence a change in opinion. We track the significance of these features over time use those trends to draw conclusions about the community's sustainability.

Problem Setup

Our problem setup closely follows that of (Tan et al. 2016): we start with the posts in which the OP awarded a Δ to a challenger, and add each succesful argument to our dataset. For each successful argument, we also add the closest unsuccessful argument to our dataset, where closeness is measured in Jaccardian distance.

Like Tan, we also analyzed the same factors on a dataset where all winning argument/losing argument pairs are truncated so that their lengths are equal. In Table 2 and Table 3, these the truncated arguments are referred to as *root truncated* comments.

Interplay between OP and Arguments (Table 2)

The way the OP voices their opinion can influence how challengers respond and interact in the discussion tree. Approaching the OP using similar vocabulary and quoting parts of the original post may help make an argument more convincing.

The similarity metrics we explore in this subsection analyze the overlap between the content of the argument made by challengers and the OP themselves. Letting A denote the set of words in an argument and O denote the set of words posted by the OP, the metrics we consider are given below.

- number of common words: $\mid A \cap O \mid$

Feature Name	root reply			full path		
	2013-2015	2016-2018	2019-2020	2013-2015	2016-2018	2019-2020
all						
#common in all	↑↑↑↑(TR)	↑↑↑↑	↑↑↑↑(T)	↑↑↑↑	↑↑↑↑	↑↑↑↑
OP frac. in all	↑↑↑↑(TR)	↑↑↑↑	↑↑↑↑(T)	↑↑↑↑	↑↑↑↑	↑↑↑↑
Reply frac. in all	↓↓↓↓	↓↓↓↓(TR)	↓↓↓↓(TR)	↓↓↓↓	↓↓↓↓	↓↓↓↓
Jaccard in all	↑↑↑↑(TR)	↑↑↑↑	↑↑↑↑(T)	↑↑↑↑	↑↑↑↑	↑↑↑↑
content words						
#common in content	↑↑↑↑(TR)	↑↑↑↑	↑↑↑↑(T)	↑↑↑↑	↑↑↑↑	↑↑↑↑
OP frac. in content	↑↑↑↑(TR)	↑↑↑↑	↑↑↑↑(T)	↑↑↑↑	↑↑↑↑	↑↑↑↑
Reply frac. in content	↓↓↓↓	↓↓↓↓(TR)	↓↓↓↓(TR)	↓↓↓↓	↓↓↓↓	↓↓↓↓
Jaccard in content	↑↑↑↑(TR)	↑↑↑↑	↑↑↑↑(T)	↑↑↑↑	↑↑↑↑	↑↑↑↑
stopwords						
#common in stopwords	↑↑↑↑(TR)	↑↑↑↑	↑↑↑↑(T)	↑↑↑↑	↑↑↑↑	↑↑↑↑
OP frac. in stopwords	↑↑↑↑(TR)	↑↑↑↑	↑↑↑↑(T)	↑↑↑↑	↑↑↑↑	↑↑↑↑
Reply frac. in stopwords	↓↓↓↓	↓↓↓↓(TR)	↓↓↓↓(TR)	↓↓↓↓	↓↓↓↓	↓↓↓↓
Jaccard in stopwords	↑↑↑↑	↑↑↑↑	↑↑↑↑(T)	↑↑↑↑	↑↑↑↑	↑↑↑↑

Table 2: Significance tests on interplay features. Features are sorted by average p-value in the two tasks. **Legend:** (1) Direction indicates winning (↑) or losing (↓) argument; (2) ↑↑↑↑: $p < 0.0001$, ↑↑↑: $p < 0.001$, ↑↑: $p < 0.01$, ↑: $p < 0.05$; (3) (T) in the *root reply* column indicates that the feature is significant in the *root truncated* condition. (TR) indicates significance in the in the reverse direction;

- reply fraction: $\frac{|A \cap O|}{|A|}$
- OP fraction: $\frac{|A \cap O|}{|O|}$
- Jaccard: $\frac{|A \cap O|}{|A \cup O|}$

We analyze the similarities of the overall arguments via these measures, and like Tan we also consider similarities only between sets of stopwords (which provide an insight into stylistic similarities) and content words (which provide a clearer picture of new information).

Table 2 shows the results of significance tests on the factors evaluated between the original post and the arguments across three year bins (2013-2015, 2016-2018, and 2019-2020).

Our significance tests for the 2013-2015 year bin, which corresponds with the (Tan et al. 2016) dataset, produced similar results for the # of common words, OP fraction, and Reply fraction across all three word categories. We, however, found that successful arguments had a higher Jaccard distance from the OP, contrary to their initial observations. We reason that this might be the result of differences in word tokenization.

The implications of various trends in these features across time are discussed at the end of this section, in the subsection titled *Language Indicators and Community Longevity*.

Argument-only Features (Table 3)

In table 3, we explore the stylistic and linguistic factors of the argument alone, without considering its interplay with the OP. The factors we evaluate follow (Tan et al. 2016), and we refer the reader there for a complete explanation of the factors.

Our results for the 2013-2015 year bin are similar to those of (Tan et al. 2016), except for in the case of word-category-based features of root-truncated arguments. Again, this might be a result of different tokenization techniques, since arguments are tokenized before they are truncated. The implications of trends in the significance of these features over time for community sustainability are discussed at the end of this section.

OP-only Features (Table 4)

Lastly, it stands to reason that a change in opinion is not purely dependent on the persuasiveness of the challenger. Despite the understanding that every OP *should* only post an opinion if they are open to seeing opposing arguments, their opinion may not be malleable, and undoubtedly some OPs may be less willing to change their minds than others.

Accordingly, we evaluated the same features as we did in the previous section, but for the content of posts (instead of arguments), in order to see what factors significantly correlated with an OPs handing out a Δ. Below is a summary of our findings:

- We observed significance results in malleable OPs in the same patterns that we observed significances in persuasive arguments. These results, however, only hold up in the 2013-2015 and 2016-2018 year bins. However, only four of the factors that we evaluated showed significance across all three bins: indefinite articles, 2^{nd} person pronouns, hedge words, and arousal.
- Posts made in the 2019-20 bin showed significance in only 4 of all the categories evaluated. This period correlates with a slight decrease in the number of new OPs

Feature Name	root reply			full path		
	2013-15	2016-18	2019-20	2013-15	2016-18	2019-20
#words	↑↑↑↑ —	↑↑↑↑ —	↑↑↑↑ —	↑↑↑↑	↑↑↑↑	↑↑↑↑
word category-based features						
definite articles; indefinite articles	↑↑↑↑(T)	↑↑↑↑(T)	↑↑↑↑(T)	↑↑↑↑	↑↑↑↑	↑↑↑↑
positive words; negative words; 1st pers (sing. & plural) pro.	↑↑↑↑	↑↑↑↑	↑↑↑↑	↑↑↑↑	↑↑↑↑	↑↑↑↑
2nd person pronouns; #links	↑↑↑↑(TR)	↑↑↑↑(TR)	(TR)	↑↑↑↑	↑↑↑↑	
# .com links	↑↑↑↑(TR)	(TR)	↓↓↓↓(T)	↑↑↑↑	↑	↓↓
# .edu links	↑↑↑↑	↑↑↑↑	↑↑↑↑	↑↑↑↑	↑↑↑↑	↑↑↑↑
# .pdf links	↑↑↑↑(T)	↑↑↑↑(T)	↑↑↑↑	↑↑↑↑	↑↑↑↑	↑↑↑↑
# hedges	↑↑↑↑(TR)	↑↑↑↑(TR)	↑↑↑↑	↑↑↑↑	↑↑↑↑	↑↑↑↑
# examples	↑↑↑↑	↑↑↑↑	↑↑↑↑(T)	↑↑↑↑	↑↑↑↑	↑↑↑↑
# questions marks; # quotations	↑↑↑↑ —	↑↑↑↑ —	↑↑↑↑ —	↑↑↑↑	↑↑↑↑	↑↑↑↑
Entire argument features						
word entropy	↑↑↑↑	↑↑↑↑(T)	↑↑↑↑(T)	↑↑↑↑	↑↑↑↑	↑↑↑↑
type-token ratio	↓↓↓↓	↓↓↓↓(T)	↓↓↓↓(T)	↓↓↓↓	↓↓↓↓	↓↓↓↓
# sentences; # paragraphs; Flesch-Kincaid Grade Level	↑↑↑↑ —	↑↑↑↑ —	↑↑↑↑ —	↑↑↑↑	↑↑↑↑	↑↑↑↑
markdown formatting						
# italics; bullet lists; numbered lists	↑↑↑↑ —	↑↑↑↑ —	↑↑↑↑ —	↑↑↑↑	↑↑↑↑	↑↑↑↑
# bolds	↑↑↑↑ —	↑↑ —	—	↑↑↑↑	↑↑↑↑	
Word Score-based Features						
concreteness			↓↓↓↓(T)			↓↓↓↓
arousal		↑↑↑↑(TR)	↑↑↑↑	↑	↑↑↑↑	↑↑↑↑
valence			(TR)			↓
dominance		↓↓↓↓(T)	↓↓↓↓(T)		↓↓↓↓	↓↓↓↓

Table 3: Argument-only features that pass a Bonferroni-corrected significance test. Features are sorted within each group by average p-value over the two tasks. **Legend:** (1) Indicators are the same as in Table 2; (2) Due to our simple truncation based on words, some features, such as those based on complete sentences, cannot be extracted in root truncated; these are indicated by a dash (—).

Feature Name	Signficance by Year Bins		
	2013-2015	2016-2018	2019-2020
indefinite articles	↑↑	↑↑↑↑	↑↑
1st person pronouns	↑↑↑↑	↑↑↑↑	↑↑↑↑
1st person plural	↑↑↑↑	↑↑↑↑	
# hedges	↑↑↑↑	↑↑↑↑	↑↑↑↑
# paragraphs	↑↑↑↑	↑↑↑↑	
# bolds	↑↑↑↑	↑↑↑↑	
arousal	↑↑↑↑	↑↑↑↑	↓↓
valence	↑↑↑↑	↑↑↑↑	
dominance	↑↑↑↑	↑↑↑↑	

Table 4: Features of the OP's argument analyzed for significance in malleable and non-malleable posts. The table only includes features that showed satistical significance across $\geq$2 year bins. **Legend:** (1) Malleable (↑) or Non-Mallable (↓) Original Post. (2) Significance and direction indicators are the same as Table 2.

(Figure 1b), a significant decrease in the average replies each post received, and a significant decrease in average malleability. Since most features analyzed were in favour of malleable OPs, we reason that a fall in significance is a result of a change in community composition and interaction patterns.

- Malleable OPs show a level of uncertainty in their stance, using significantly more hedge words. They also refer to themselves as being the 'holder' of the opinion, indicated via significantly more occurrences of 1st person pronouns.

Language Indicators and Community Longevity

The relationship between language indicators and malleability over time can provide several clues to the community's longevity. In particular, since users in r/CMV are encouraged to change others' views (via leaderboards and indicators of the number of Δ's users have been awarded), language indicators correlating positively with malleability could be seen as features that are valued in the community. Therefore, trends in the significance of features over time can provide clues to the types of arguments valued by the community over time, which can in turn help us understand why r/CMV has remained popular for so many years. Our analyses are given below:

- *Dominant arguments have gotten less popular over time.* Table 3 shows that the dominance of arguments was not correlated with malleability at all in the early years of r/CMV, but was strongly negatively correlated with malleability from 2016 onward. Since many dominant words may indicate a less cautious and more forceful argument,

we can interpret this trend as reflective of a community that chose to reject heavy-handed and domineering arguments in favor of more reasoned approaches. The declining significance of bold words in persuasive arguments also mirrors this trend. Similarly, the continued positive correlation between hedge words (which reflect caution) and malleability is further evidence that r/CMV does not give much creedence to overly forceful arguments. This attitude towards discourse has undoubtedly helped create a community where conversations are less likely to derail, making it more likely that users will return to the platform.

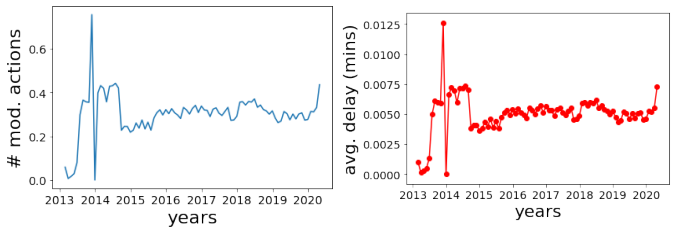
- *Good sources are valued.* Links to `.edu` domains and PDF files are highly correlated with malleability across all year bins, suggesting that the community appreciates links to reputable sources of information. Further supporting this hypothesis is the fact that `.com` have, over time, become negatively correlated with malleability. In light of the continued value of `.edu` and `.pdf` links, it seems reasonable to assume that a decline in the effectiveness of `.com` links can potentially be attributed to changing perceptions of the trustworthiness of the average Internet link: the years since 2015 have seen a rise in questionable news sources and other forms of online disinformation (Bradshaw and Howard 2017; Bessi and Ferrara 2016). Therefore, the trends in these link types seem to clearly reflect that the community has a clear and consistent definition of *good* sources. The community expectation for references, especially to documents that are likely to be primary sources, is in line with the generally accepted standards for well sourced research. A community with an affinity for good sources as an indicator for stronger arguments could be one of many reasons that users keep coming back.
- *A diversity of ideas.* Table 2 shows that persuasive arguments are more similar to the original post across all similarity metrics, except for *reply fraction*. We therefore infer that persuasive arguments repeat and reuse content from the original post significantly more than unsuccessful arguments. However, the results concerning *reply fraction* indicate that non-persuasive arguments include little more than those the OP has already made.

  While reusing using the original post to guide their argument, persuasive arguments add significantly more new content to drive home their point. This suggests that discussion communities might do well to aid their participants in forming arguments by providing options to quote/reuse content from the OP, while also encouraging the addition of uncovered perspectives.

Community Moderation

Active moderation of content on ChangeMyView has been credited for the maintenance of civil discourse within the subreddit (Heffernan 2018). The community uses content removal as their primary moderation strategy for rule violations.

Content in the ChangeMyView community is governed by 10 rules, all aimed at ensuring that users engage in civil discourse and preventing conversations from turning abusive. After 3 strikes (content removals), participants are removed from the community.



(a) # of moderation actions per post over time

(b) Average time taken by moderators to remove rule-violating content

Figure 4: Moderation stats. Figure 4a gives an indicator that moderation actions remain consistent with activity in the community. In Figure 4b, we observe that the average time for a moderation action is unaffected by an increase in the activity in the community post 2018.

Each time a moderator finds content to be in violation of the community, they take down the comment, provide an explanation (including the rule that was violated), and provide a link for the action to be contested.

Figure 4 shows statistics reflecting the ways moderation has been carried out on CMV over time. We observe that on average, around 487 moderation actions are taken every month, and each month around 30% of posts see a removal of at least one comment (or of the post itself). The community used a total of 48 unique moderators over the last 7 years, of which 3 were bots.

Content on CMV has always been monitored by multiple moderators at any given time, and we observed that the composition of the group of moderators has changed over time. We also observed that the overall load on a single moderator did not increase in the years following 2018 when the community witnessed a surge in subscribers. They did, however, manage to maintain their average response delay by balancing the load across moderators. Only a few select moderators took on an (outlier) active role, performing between 250 and 300 content removal actions each month in the last 3 years.

While these observations provide an insight into the activity and composition of moderators, it does not provide give us an understanding of the context in which these actions were implemented – a factor that could indicate the effectiveness of their actions. We therefore clubbed discussions by high-level topics in the next section, and included context for moderation actions below.

Topics and Malleability

After a pre-processing step in which we removed all moderator posts, deleted posts, and posts with empty bodies or titles, we used MALLET[4] (a topic modelling tool which implements LDA (Blei, Ng, and Jordan 2003)) to classify the entire corpus of original posts and titles into 14 topics (the

[4]MALLET: http://mallet.cs.umass.edu/

number of topics was decided via manual review of models parameterized with between 10 and 20 topics). Two of the authors determined topic names independently, and found that our names matched.

For each topic, we measure:

1. *Popularity*: what fraction of posts belong to this topic?
2. *Malleability*: (as it is used in the rest of the paper) within this topic, what fraction of posts receiving at least one comment feature a Δ?
3. *Activity*: within this topic, how many comments does a post receive on average (including replies by the OP)?
4. *Moderator Actions*: within this topic, what fraction of comments are removed by moderators?

In order to be confident that we were only calculating these measures for posts which were firmly within any given topic, we first discarded posts which did not receive a confidence score of at least 50% for any topic.

Figure 5 shows malleability for all topics across all seven years of the dataset. We observe that conversations about food have the highest malleability (42.6%), 11.5% higher than the topic with the next highest malleability (Education, malleability = 38.2%), and 17% higher than the dataset-wide average of 36.4% (among all posts with at least one comment that passed our pre-processing phase). These observations can likely be attributed to the fact that people may feel less strongly about food than they do about other issues (one post achieving a 98% confidence score for the *Food* topic had the title *"CMV: Tortillas should be square"*). We also observed that people are less willing to change their opinions on *International Relations*

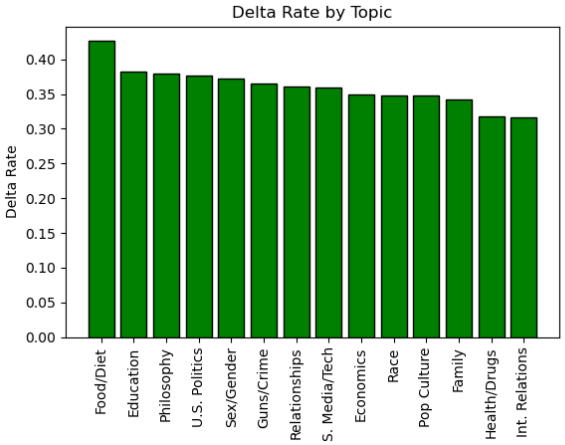


Figure 5: Malleability across topics. Posts are considered members of a topic if MALLET gives a the post a confidence score of at least 50% in that topic.

Figure 7 shows the popularity, malleability, activity, and moderation density for topics across time. Since the subreddit-wide malleability, moderation, and activity levels have changed over time (Figure 1), we normalize for these trends by plotting the ratio between a topic's malleability/activity/moderation activity and the subreddit-wide average for each year.

Of particular interest are the changes in time for the *U.S. Politics* topic, visible in Figure 6. In the data we observe,

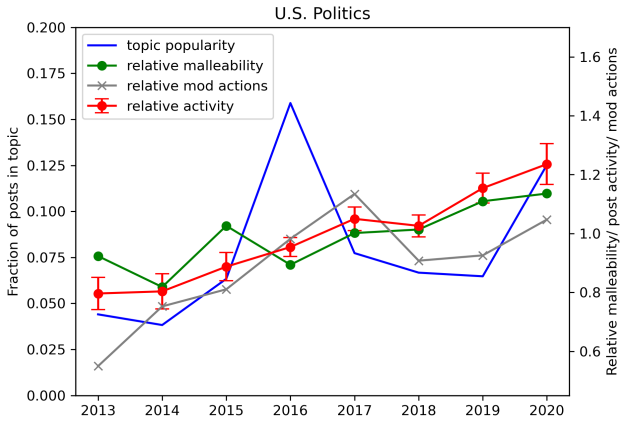


Figure 6: Popularity, relative malleabilitiy, and relative activity of U.S. Politics posts over time.

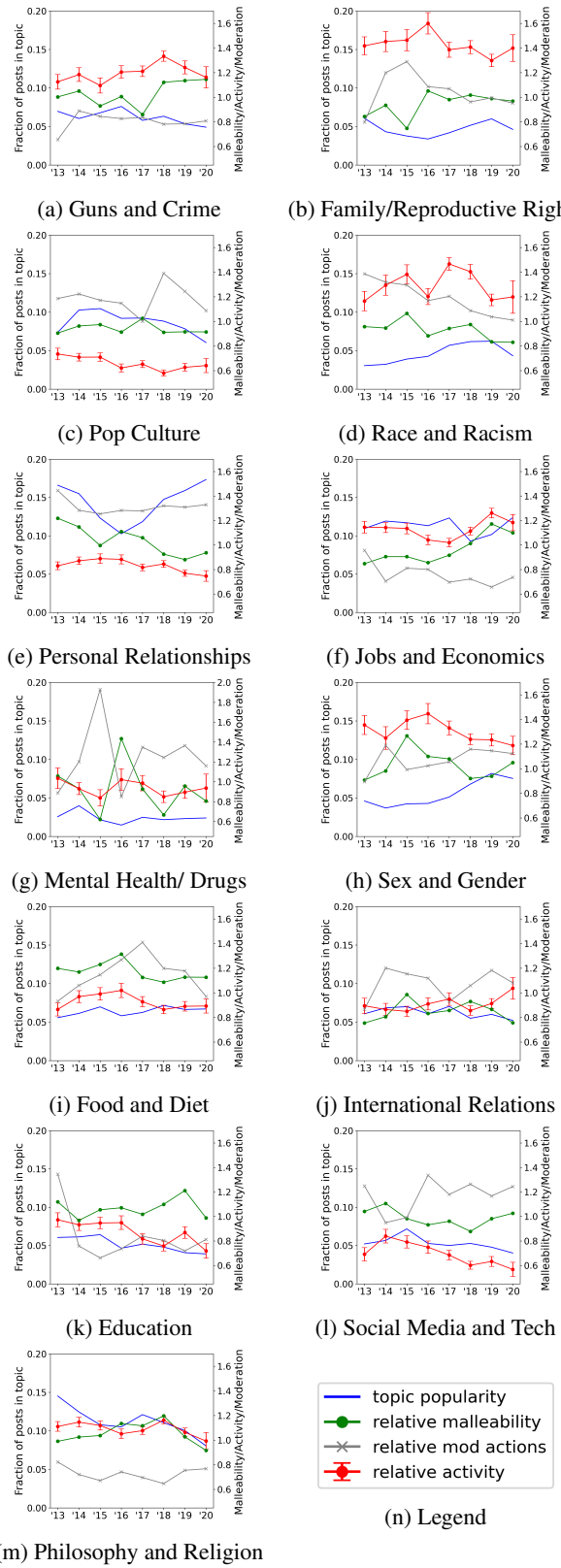
posts about *U.S. Politics* account for a dramatically greater share of CMV posts during election years. Despite a dip in malleability around the time of the contentious 2016 election, the malleability trends upwards over time, and since 2018 posts about politics have enjoyed a higher malleability than the subreddit-wide average. Also steadily trending upward are the average number of comments and moderation actions per comment (again, expressed as ratios to the subreddit-wide average). These suggest a trend of increased engagement with and malleability around political discussions, and the coupling of high numbers of comment removals and high malleability suggests that the "touchiness" of a subject need not reflect how willing one is to actually change their minds about it – within CMV, at least. Discussions about *Guns and Crime*, another subject that is traditionally thought of as "touchy", also enjoy an upward trend in malleability. On the other hand, conversations about *Race and Racism* have seen a drop in malleability since 2018.

Discussion

The paper performs an analyses of activity within the ChangeMyView community over five years. Our efforts towards understanding the factors that contribute to the communities sustainability were bound by the research questions we set out to answer.

Content-based Features

Interplay between OP and Arguments. We observed that the interplay between the Original Posts and Arguments under these posts have remained consistent over the last 7 years. In any argument that an OP follows closely, early challengers have a greater advantage, and challengers who



(a) Guns and Crime (b) Family/Reproductive Rights

(c) Pop Culture (d) Race and Racism

(e) Personal Relationships (f) Jobs and Economics

(g) Mental Health/ Drugs (h) Sex and Gender

(i) Food and Diet (j) International Relations

(k) Education (l) Social Media and Tech

(m) Philosophy and Religion (n) Legend

Figure 7: Popularity, malleability, and activity over time for all topics. Note that figure 7g uses a different scale for its right vertical axis. All other figures share the same scale for their right vertical axes.

engage the OP in 2 or 3 back-and-forth interactions have a higher chance of changing a view. Increasing the activity is helpful only upto a certain point (8 challengers), and individual challengers interacting back-and-forth with the OP is more useful than multiple challengers dividing their attention.

Active communities where participants engage with posts with little delay, and where they are willing to engage in a back-and-forth interaction directly with the OP (and not multiple other challengers/the community at large) can observe healthier success rates in opinion changes. Communities that encourage individual conversations once a challenge has been made might be more successful than those that encourage numerous challengers to jump into an ongoing interaction.

Argument-only Language Indicators. The indicators of language and style of a winning challenger's argument has largely remained consistent over the years. Challengers making longer arguments, reusing words from the OP, displaying uncertainty and malleability through hedge words and limited dominance in their arguments, are favored to successfully change an opinion.

While CMV encourages participants to post reasoning for their arguments and does not limit their length (unlike Twitter), it does not nudge participants to limit asserting dominance, quote or reuse words from the OP, or post more `.edu` and `.pdf` links in their arguments. The challengers who do abide by these practices are therefore mixed in with numerous unsuccessful challengers who display different linguistic styles, thereby reducing the possibility of a change in opinion.

OP-only Language Indicators. For discussion communities to observe a healthy success rate, they need participants with malleable opinions. Our analysis of CMV revealed 1^{st} person pronouns and the use of hedge words (displaying uncertainty), to be the only definitive indicators of opinion malleability. CMV encourages posters to use these indicators indirectly by requiring that OPs provide reasoning for their stated belief.

Community-wide Features

Moderation Actions. Our analysis of moderation on r/CMV indicated that the community had an active, but changing group of moderators. The activity was distributed among multiple people, and their reply delay indicates that the moderators were prompt and consistent. An importance placed towards enforcing rules, providing reasoning for an action, and a link to contest the action, are factors that helps r/CMV ensure that their participants engage in civil discourse. Communities that intend to infer from r/CMV can benefit from enforcing similarly active moderation.

The effectiveness of content removal as a moderation strategy is debatable. (Srinivasan et al. 2019) found compliance to rules as the only verifiable result of such actions. They, however, did not find a causal relationship between content removal and increase in participation or decrease in toxicity of the affected individual.

High-level Topics. High-level topics do not provide an inference for malleability by themselves. Discussion com-

munities attempting to replicate the success of r/CMV do not need to shy away from difficult topics – the lack of correlation between malleability and topic activity means that one does not need to restrict discussions to avoid sensitive topics in an effort to maintain an active user base. In fact, we conjecture that r/CMV has maintained its popularity at least in part because of its ability to host discussions on low-malleability topics.

Limitations

We acknowledge that numerous external factors influence the probability that an OP will award a Δ, and our analysis cannot capture these factors. Even though the CMV community strongly encourages OPs to only post if they are willing to have their view changed, these guidelines cannot mandate a change of view. OP's regularly engage with their challengers but the rules in place cannot govern an extent of commitment in this regard.

We believe that our work is generalizable to all types of well-moderated online discussions, and hope that these results can guide the creation of deliberation platforms in general. Positive results about the malleability and engagement of users in political discussions on CMV suggest that the creation of better deliberation platforms could increase cooperation and reduce animosity in conversations about controversial subjects.

References

Abbott, R.; Walker, M.; Anand, P.; Fox Tree, J. E.; Bowmani, R.; and King, J. 2011. How Can You Say Such Things?!? Recognizing Disagreement in Informal Political Argument. In *Proceedings of the Workshop on Languages in Social Media*.

Bessi, A.; and Ferrara, E. 2016. Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday* 21(11).

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3(Jan).

Bradshaw, S.; and Howard, P. 2017. Troops, trolls and troublemakers: A global inventory of organized social media manipulation. In *CompProp, OII, Working Paper*.

Dutta, S.; Chakraborty, T.; and Das, D. 2019. *How Did the Discussion Go: Discourse Act Classification in Social Media Conversations*, 137–160. Cham: Springer International Publishing. ISBN 978-3-030-01872-6. doi:10.1007/978-3-030-01872-6_6.

Dutta, S.; Das, D.; and Chakraborty, T. 2020. Changing views: Persuasion modeling and argument extraction from online discussions. *Information Processing & Management* 57(2).

Gleize, M.; Shnarch, E.; Choshen, L.; Dankin, L.; Moshkowich, G.; Aharonov, R.; and Slonim, N. 2019. Are You Convinced? Choosing the More Convincing Evidence with a Siamese Network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Habernal, I.; and Gurevych, I. 2016. What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in Web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Heffernan, V. 2018. Our Best Hope for Civil Discourse on the Internet Is on ... Reddit. https://www.wired.com/story/free-speech-issue-reddit-change-my-view/.

Justin, W. 2018. Day of the trope: White nationalist memes thrive on Reddit's r/The_Donald. Southern Poverty Law Center.

Luu, K.; Tan, C.; and Smith, N. A. 2019. Measuring Online Debaters' Persuasive Skill from Text over Time. *Transactions of the Association for Computational Linguistics* 7.

Shepherd, T.; Harvey, A.; Jordan, T.; Srauy, S.; and Miltner, K. 2015. Histories of Hating. *Social Media + Society* 1(2).

Shi, W.; Wang, X.; Oh, Y. J.; Zhang, J.; Sahay, S.; and Yu, Z. 2020. Effects of Persuasive Dialogues: Testing Bot Identities and Inquiry Strategies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.

Sridhar, D.; Foulds, J.; Huang, B.; Getoor, L.; and Walker, M. 2015. Joint Models of Disagreement and Stance in Online Debate. In *Proceedings of ACL*.

Srinivasan, K. B.; Danescu-Niculescu-Mizil, C.; Lee, L.; and Tan, C. 2019. Content Removal as a Moderation Strategy: Compliance and Other Outcomes in the ChangeMyView Community 3(CSCW).

Tan, C.; Niculae, V.; Danescu-Niculescu-Mizil, C.; and Lee, L. 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-Faith Online Discussions. In *Proceedings of the 25th International Conference on World Wide Web*.

The British Broadcasting Corporation. 2014. US created 'Cuban Twitter' to stir unrest.

Twitter Public Policy. 2018. Update on Twitter's review of the 2016 US election.

Wachsmuth, H.; Syed, S.; and Stein, B. 2018. Retrieval of the Best Counterargument without Prior Topic Knowledge. In *Proceedings of the ACL*.

Xiao, L.; and Khazaei, T. 2019. Changing Others' Beliefs Online: Online Comments' Persuasiveness. In *Proceedings of the 10th International Conference on Social Media and Society*.

Yesayan, T. 2014. Social networking: A guide to strengthening civil society through social media. https://www.usaid.gov/sites/default/files/documents/1866/SMGuide4CSO.pdf.

Zhang, A. X.; Culbertson, B.; and Paritosh, P. 2017. Characterizing Online Discussion Using Coarse Discourse Sequences. In *Proceedings of the 11th International AAAI Conference on Weblogs and Social Media*, ICWSM '17.

Zhang, J.; Chang, J.; Danescu-Niculescu-Mizil, C.; Dixon, L.; Hua, Y.; Taraborelli, D.; and Thain, N. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of the 56th ACL*.