Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions

Chenhao Tan Vlad Niculae Cristian Danescu-Niculescu-Mizil Lillian Lee
Cornell University
{chenhao|vlad|cristian|llee}@cs.cornell.edu

ABSTRACT

Changing someone's opinion is arguably one of the most important challenges of social interaction. The underlying process proves difficult to study: it is hard to know how someone's opinions are formed and whether and how someone's views shift. Fortunately, ChangeMyView, an active community on Reddit, provides a platform where users present their own opinions and reasoning, invite others to contest them, and acknowledge when the ensuing discussions change their original views. In this work, we study these interactions to understand the mechanisms behind persuasion.

We find that persuasive arguments are characterized by interesting patterns of interaction dynamics, such as participant entry-order and degree of back-and-forth exchange. Furthermore, by comparing similar counterarguments to the same opinion, we show that language factors play an essential role. In particular, the interplay between the language of the opinion holder and that of the counterargument provides highly predictive cues of persuasiveness. Finally, since even in this favorable setting people may not be persuaded, we investigate the problem of determining whether someone's opinion is susceptible to being changed at all. For this more difficult task, we show that stylistic choices in how the opinion is expressed carry predictive power.

1. INTRODUCTION

Changing a person's opinion is a common goal in many settings, ranging from political or marketing campaigns to friendly or professional conversations. The importance of this topic has long been acknowledged, leading to a tremendous amount of research effort [9, 15, 17, 42, 46, 47]. Thanks to the increasing number of social interactions online, *interpersonal persuasion* has become observable at a massive scale [19]. This allows the study of interactive persuasion *in practice, without elicitation*, thus bypassing some limitations of laboratory experiments and leading to new research questions regarding dynamics in real discussions. At the same time, the lack of the degree of experimental control offered by lab trials raises new methodological challenges that we address in this work.

It is well-recognized that multiple factors are at play in persuasion. Beyond (*i*) the characteristics of the arguments themselves, such as intensity, valence and framing [1, 2, 4, 6, 23], and (*ii*) social aspects, such as social proof and authority [7, 10, 33], there is also (*iii*) the relationship between the opinion holder and her belief, such as her certainty in it and its importance to her [44, 45, 54, 59]. Thus, an ideal setting for the study of persuasion would allow access to the reasoning behind people's views in addition to the full interactions. Furthermore, the outcome of persuasion efforts (e.g., which efforts succeed) should be easy to extract.[1]

One forum satisfying these desiderata is the active Reddit subcommunity `/r/ChangeMyView` (henceforth CMV).[2] In contrast to general platforms such as Twitter and Facebook, CMV requires posters to state the reasoning behind their beliefs and to reward successful arguments with explicit confirmation. Moreover, discussion quality is monitored by moderators, and posters commit to an openness to changing their minds. The resulting conversations are of reasonably high quality, as demonstrated by Figure 1, showing the top portion of a discussion tree (an original post and all the replies to it) about legalizing the "tontine".[3] In the figure, Reply B.1 branches off to an extended back-and-forth between the blue original poster (OP) and the orange user; as it turns out, neither ends up yielding, although both remain polite. Reply A.1, on the other hand, is successful, as the OP acknowledges at A.2. The example suggests that content and phrasing play an important role (A.1 does well on both counts), but also that interaction factors may also correlate with persuasion success. Examples include time of entry relative to others and amount of engagement: the discussion at B.1 started earlier than that at A.1 and went on for longer.

Outline and highlight reel. This work provides three different perspectives on the mechanics of persuasion. First, we explore how interaction dynamics are associated with a successful change of someone's opinion (§3). We find (example above to the contrary) that a challenger that enters the fray before another tends to have a higher likelihood of changing the OP's opinion; this holds even for first-time CMV challengers, and so is not a trivial consequence of more experienced disputants contriving to strike first. Although engaging the OP in some back-and-forth is correlated with higher chances of success, we do not see much OP conversion in extended conversations. As for opinion conversion rates, we find that the more participants there are in the effort to persuade the OP, the larger the likelihood of the OP changing her view; but, interestingly, the relationship is sublinear.

[1] One might think that the outcome is trivially "no one ever changes their mind", since people can be amazingly resistant to evidence contravening their beliefs [8, 32, 36]. But take heart, change does occur, as we shall show.

[2] `https://reddit.com/r/changemyview`

[3] It is not necessary for the reader to be familiar with tontines, but a brief summary is: a pool of money is maintained where the annual payouts are divided evenly among all participants still living.

WWW 2016, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4143-1/16/04.
http://dx.doi.org/10.1145/2872427.2883081.

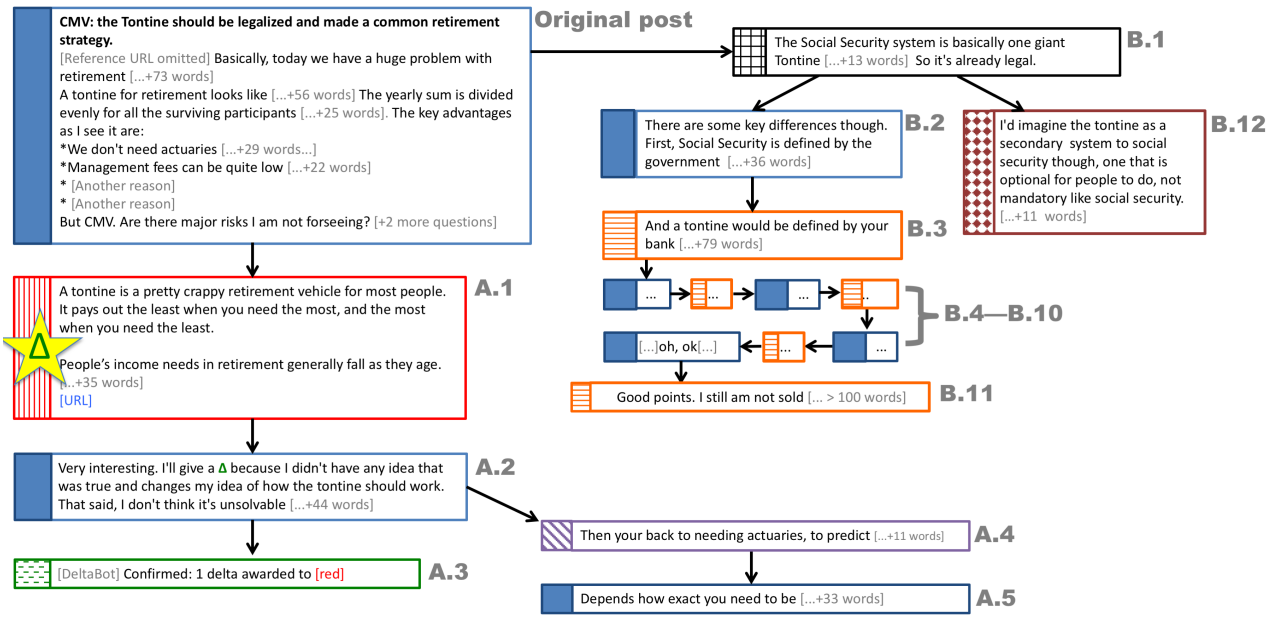


Figure 1: A fragment of a "typical" `/r/ChangeMyView` *discussion tree*—typical in the sense that the full discussion tree has an average number of replies (54), although we abbreviate or omit many of them for compactness and readability. Colors indicate distinct users. Of the 17 replies shown (in our terminology, every node except the original post is a reply), the OP explicitly acknowledged only one as having changed their view: the starred reply A.1. The explicit signal is the "Δ" character in reply A.2. (The full discussion tree is available at `https://www.reddit.com/r/changemyview/comments/3mzc6u/cmv_the_tontine_should_be_legalized_and_made_a/`.)

Besides interaction dynamics, language is a powerful tool that is in the full control of the challengers. In §4 we explore this perspective by tackling the task of predicting which of two *similar* counterarguments will succeed in changing the same view. By comparing similar arguments we focus on the role of stylistic choices in the presentation of an argument (identifying reasoning strategies is a separate problem we do not address). We experiment with style features based solely on the counterargument, as well as with features reflecting the interplay between the counterargument and the way in which the view is expressed. Style features and interplay features both prove useful and outperform a strong baseline that uses bag-of-words. In particular, interplay features alone have strong predictive power, achieving an improvement of almost 5% in accuracy over the baseline method (65.1% vs 59.6%) in a *completely fresh* heldout dataset. Our results also show that it is useful to include links as evidence—an interesting contrast to studies of the *backfire effect*: "When your deepest convictions are challenged by contradictory evidence, your beliefs get stronger" [8, 32, 36]. However, it hurts to be too intense in the counterargument. The feature with the most predictive power of successful persuasion is the dissimilarity with the original post in word usage, while existing theories mostly study matching in terms of attitude functions or subject self-discrepancy [43, 56].

In the majority of cases, however, opinions are not changed, even though it takes courage and self-motivation for the original poster to post on CMV and invite other people to change her opinion. Can we tell whether the OP is unlikely to be persuaded from the way she presents her reasoning? In §5, we turn to this challenging task. In our pilot study, humans found this task quite difficult in a paired setting and performed no better than random guessing. While we can outperform the random baseline in a realistic imbalanced setting, the AUC score is only 0.54. Our feature analysis is consistent with existing theories on self-affirmation [11, 12] and shows that malleable beliefs are expressed using more self-confidence and more organization, in a less intense way.

While we believe that the observations we make are useful for understanding persuasion, we do not claim that any of them have causal explanations.

In §6, we discuss other observations that may open up future directions, including attempts to capture higher-level linguistic properties (e.g., semantics and argument structure); §7 summarizes additional related work and §8 concludes.

2. DATASET

We draw our data from the `/r/ChangeMyView` subreddit (CMV), which has over 211,000 subscribers to date. It is self-described[4] as "dedicated to the civil discourse [sic] of opinions". CMV is well-suited to our purposes because of its setup and mechanics, the high quality of its arguments, and the size and activity of its user base. We elaborate below.

The mechanics of the site are as follows. Users that "accept that they may be wrong or want help changing their view" submit *original posts*, and readers are invited to argue for the other side. The original posters (*OPs*) explicitly recognize arguments that succeed in changing their view by replying with the *delta* (Δ) character (an example is node A.2 in Figure 1) and including "an explanation as to why and how" their view changed. A Reddit bot called the DeltaBot confirms deltas (an example is A.3 in Figure 1) and maintains a leaderboard of per-user Δ counts.[5] The experimental

[4]Quotations here are from the CMV wiki.

[5]Although non-OPs can also issue deltas, in this work, we only count deltas given by a user in their OP role. A consequence is that we only consider discussion trees where the OP's Reddit account

(OP) Title: I believe that you should be allowed to drive at whatever speed you wish as long as you aren't driving recklessly or under extenuating circumstances CMV.

I think that if you feel comfortable driving 80 mph or 40 mph you should be allowed to do so, as long as you aren't in a school or work zone, etc. because there are a lot more risks in those areas. I think when you're comfortable driving you will be a better driver, and if you aren't worrying about the speed limit or cops you are going to be more comfortable. However, I think that you should only be allowed to drive at whatever speed you wish as long as you aren't driving recklessly. If you're weaving in and out of traffic at 90, you probably shouldn't be allowed to go 90, but if you just stay in the fast lane and pass the occasional person I don't think there is a problem. CMV.

(C1) Some issues with this:

1. Who's to say what is reckless driving? Where do you draw the line? Speed is the standard that ensures we know what is considered to be reckless. The idea of driving any speed you want creates a totally subjective law.
2. How do you judge whether to pass other drives and such? There are a lot of spatial awareness issues with the roads being so unpredictable.
3. How do you expect insurance and courts to work out who's at fault for an accident?

A: "Yeah this guy was going 100 mph!"

B: "But I wasn't driving recklessly - you were!"

It's simply not realistic and creates some serious legal issues.

(C2) They're many issues I have with this idea but I'll start with the most pressing one. Think of the amount of drivers you pass by every day. Imagine all of them going at whatever speed they choose. How would this work? You cannot have a driver going 35 and a driver who wants to go 65 in the same lane.

Now lets take this onto the highway and you can see how horrific this could get quickly. They're too many drivers out on the road for everyone to choose there own speed.

Speed limits protect us all because it gives us a reasonable expectation in whatever area we're driving in. Have you ever been on the highway being a driver going 40mph? If you're doing the speed limit (65) you catch up to them so fast you barely have time to react before an accident occurs. You aren't expecting this low speed when everyone is going at similar speeds to yours.

Drivers need to know the speed expectations so they can drive and react accordingly. If everyone goes at whatever speed they want it will only cause many many accidents.

Figure 2: An *original post* and a pair of *root replies* **C1** and **C2** contesting it, where **C1** and **C2** have relatively high vocabulary overlap with each other, but only one changed the OP's opinion. (§4 reveals which one.)

advantages of this setup include:
(1) Multiple users make different attempts at changing the same person's mind on the same issue based on the same rationale, thus controlling for a number of variables but providing variation along other important aspects. Figure 2, for example, presents in full two counter-arguments, C1 and C2. They both respond to the same claims, but differ in style, structure, tone, and other respects.
(2) The deltas serve as explicit persuasion labels that are (2a) provided by the actual participants and (2b) at the fine-grained level of individual arguments, as opposed to mere indications that the OP's view was changed.
(3) The OP has, in principle, expressed an openness to other points of view, so that we might hope to extract a sufficient number of view-changing examples.
These advantages are not jointly exhibited by other debate sites, such as CreateDebate.com, ForandAgainst.com, or Debate.org.

The high quality of argumentation makes CMV a model site for seeing whether opinion shifts can at least occur under favorable conditions. Moderators enforce CMV rules, making sure that OPs explain why they hold their beliefs and do so at reasonable length (500 characters or more), and that OPs engage in conversation with challengers in a timely fashion. Other rules apply to those who contest the original post. There are rules intended to prevent "low effort" posts, such as "Posts that are only a single link with no substantial argumentation", but "Length/conciseness isn't the determining [criterion]. Adequate on-topic information is."[6] Figure 2 shows an example where indeed, the OP described their point in reasonable detail, and the responders raised sensible objections.

Table 1: Dataset statistics. The disjoint training and test date ranges are 2013/01/01–2015/05/07 and 2015/05/08–2015/09/01.

	# discussion trees	# nodes	# OPs	# uniq. participants
Training	18,363	1,114,533	12,351	69,965
Heldout	2,263	145,733	1,823	16,923

The high amount of activity on CMV means that we can extract a large amount of data. We process all discussion trees created at any time from January 2013, when the subreddit was created, to August 2015, saving roughly the final 4 months (May–August 2015) for held-out evaluation. Some size statistics are given in Table 1. Monthly trends are depicted in Figure 3:[7] after the initial startup, activity levels stabilize to a healthy, stable growth in average number of replies and challengers, as, gratifyingly, do OP conversion rates, computed as the fraction of discussion trees wherein the OP awarded a Δ (Figure 3d). For posts where the OP gave at least one delta, the OP gave 1.5 deltas on average. This dataset is available at `https://chenhaot.com/pages/changemyview.html`.

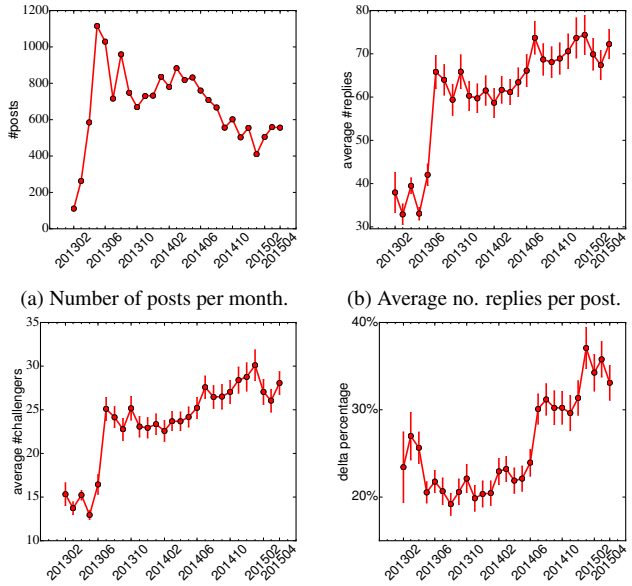
3. INTERACTION DYNAMICS

Changing someone's opinion is a complex process, often involving repeated interactions between the participants. In this section we investigate the relation between the underlying dynamics and the chances of "success", where "success" can be seen from the perspective of the challenger (did she succeed in changing the OP's opinion?), as well as from that of the set of challengers (did anyone change the OP's view?).

had not been deleted—i.e., the original post is not attributed to the ambiguous name "[deleted]"— at the time of crawl.

[6]It is worth noting that, as in many online communities, not all these rules were in place at the site's creation. It is a separate and interesting research question to understand what effects these rules have and why they were put in place. The currently enforced set of rules is available at `https://www.reddit.com/r/changemyview/wiki/rules`.

[7]We omit the first month as the DeltaBot may not have been set up.



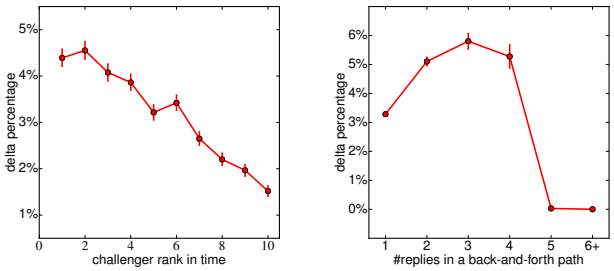
(a) Number of posts per month.

(b) Average no. replies per post.

(c) Average no. challengers per post.

(d) Average delta percentage.

Figure 3: Monthly activity over all full months represented in the training set. The *delta percentage* is the fraction of discussion trees in which the OP awarded a delta.



(a) Delta ratio vs. entry order.

(b) Delta ratio vs. degree of back-and-forth exchanges.

Figure 4: Figure 4a shows the ratio of a person eventually winning a delta in a post with at least 10 challengers depending on the order of her/his entry. *Early entry is more likely to win a delta.* Figure 4b presents the probability of winning a delta given the number of comments by a challenger in a back-and-forth path with OP. With 6 or more replies in a back-and-forth path, *no* challengers managed to win a delta among our 129 data points (with 5 replies, the success ratio is 1 out of 3K). In both figures, error bars represent standard errors (sometimes 0).

In order to discuss the relation between interaction dynamics and success, we now introduce corresponding terminology using the example illustrated in Figure 1:

- An original statement of views (*original post*) together with all the replies form a *discussion tree*.
- A direct reply to an original post is called a *root reply* (A.1 and B.1 in Figure 1). The author of a root reply is a *root challenger*.
- A *subtree* includes a root reply and all its children (B.1–B.12 form one of the two subtrees in Figure 1).
- A *path* constitutes all nodes from root reply to a leaf node. Figure 1 contains four paths: P_1: A.1, P_2: A.1, A.2, A.4, A.5, P_3: B.1–B.11 and P_4: B.1, B.12. Note that when a Δ is awarded, the DeltaBot automatic reply (A.3) and the OP's post that triggers it (A.2) are not considered part of the path.

In order to focus on discussions with non-trivial activity, in this section we only consider discussion trees with at least 10 replies from challengers and at least one reply from the OP.

3.1 Challenger's success

A challenger is successful if she manages to change the view of the OP and receive a Δ. We now examine how the interaction patterns in a discussion tree relate to a challenger's success.

Entry time. How does the time when a challenger enters a discussion relate to her chances of success? A late entry might give the challenger time to read attempts by other challengers and better formulate their arguments, while an early entry might give her the first-mover advantage.[8] Even for original posts that eventually attract attempts by 10 unique challengers, the first two challengers are 3 times more likely to succeed as the 10th (Figure 4a).

One potential explanation for this finding is that dedicated expert users are more likely to be more active on the site and thus see posts first. To account for this, we redo the analysis only for users that are participating for the first time on CMV. We observe that even after controlling for user experience, an earlier entry time is still more favorable. We omit the figure for space reasons.

Back-and-forth. After entering a discussion, the challenger can either spend more effort and engage with the OP in a back-and-forth type of interaction or call it quits. Figure 4b shows the relation between the likelihood of receiving a Δ and degree of back-and-forth, defined as the number of replies the root challenger made in a path involving only her and the OP.[9] We observe a non-monotonic relation between back-and-forth engagement and likelihood of success: perhaps while some engagement signals the interest of the OP, too much engagement can indicate futile insistence; in fact, after 5 rounds of back-and-forth the challenger has virtually no chance of receiving a Δ.

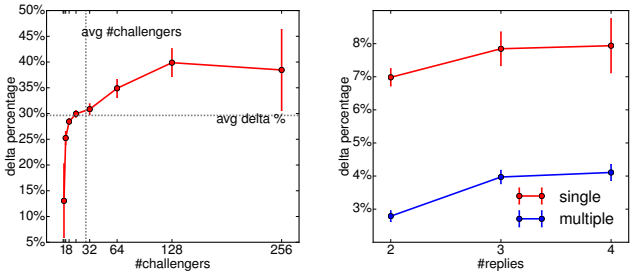
3.2 OP's conversion

From the perspective of an original post, conversion can happen when any of the challengers participating in the discussion succeeds in changing the OP's view. We now turn to exploring how an OP's conversion relates to the volume and type of activity her original post attracts.

Number of participants. It is reasonable to expect that an OP's conversion is tied to the number of challengers [7, 10]. For instance, the OP might be persuaded by observing the sheer number of people arguing against her original opinion. Moreover, a large number of challengers will translate into a more diverse set of arguments, and thus higher likelihood that the OP will encounter the ones that best fit her situation. Indeed, Figure 5a shows that the likelihood of conversion does increase with the number of unique

[8]Note that although reply display order is affected by upvotes, entry time is an important factor when the OP follows the post closely.

[9]If a subtree won a Δ, we only consider the winning path; otherwise, other conversations would be mistakenly labeled unsuccessful. For instance, the path A.1, A.2, A.4, A.5 in Figure 1 is not considered.



(a) Delta percentage vs. number of unique challengers.

(b) Single-challenger subtree vs. multiple-challenger subtree controlled by the number of replies.

Figure 5: Probability that a submitted view will be changed, given (a) the total number of unique challengers binned using $\log_2$, and (b) the number of replies in a subtree.

challengers. Notably, we observe a saturation in how much value each new challenger adds beyond a certain point.

Sheer number of challengers or diversity of counterarguments? To distinguish between the two possible explanations proposed in the previous paragraph, we control for the diversity of counterarguments by focusing only on subtrees, in which challengers generally focus on the same argument. To make a fair comparison, we further control the number of total replies in the subtree. In light of Figure 4b, we only consider subtrees with between 2 and 4 replies. Figure 5b shows that single-challenger subtrees consistently outperform multiple-challenger subtrees in terms of conversion rate. This observation suggests that the sheer number of challengers is not necessarily associated with higher chances of conversion. The fact that multiple-challenger subtrees are less effective might suggest that when talking about the same counterargument, challengers might not be adding value to it, or they might even disagree (e.g., B.12 vs. B.2 in Figure 1); alternatively, root replies that attract multiple challengers might be less effective to begin with.

4. LANGUAGE INDICATORS OF PERSUASIVE ARGUMENTS

The interaction dynamics studied in the previous section are to a large extent outside the challenger's influence. The language used in arguing, however, is under one's complete control; linguistic correlates of successful persuasion can therefore prove of practical value to aspiring persuaders. In order to understand what factors of language are effective, we set up paired prediction tasks to explore the effectiveness of textual discussion features, in the context of CMV.

4.1 Problem setup

In order to study an individual's success in persuasion, we consider the collection of arguments from the same person in the same line of argument. We focus on arguments from root challengers since the root reply is what initiates a line of argument and determines whether the OP will choose to engage. We define all replies in a path by the root challenger as a *rooted path-unit*, e.g., reply A.1 and B.1 in Figure 1.

As shown in §3, situations where there is more than one reply in a rooted path-unit correspond to a higher chance that the OP will be persuaded. So, while the challenger's opening argument should be important, statements made later in the rooted path-unit could be more important. To distinguish these two cases, we consider two related prediction tasks: *root reply*, which only uses the challenger's opening argument in a rooted path-unit, and *full path*, which considers the text in all replies within a rooted path-unit.

In response to the same original post, there are many possible ways to change someone's view. We aim to find linguistic factors that can help one formulate her/his argument, rather than to analyze reasoning strategies.[10] Hence, for each rooted path-unit that wins a Δ, we find the rooted path-unit in the same discussion tree that did not win a Δ but was the most "similar" in topic. We measure similarity between rooted path-units based on Jaccard similarity in the root replies after removing stopwords (as defined by Mallet's dictionary [31]).

$$\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|},$$

where A, B are the sets of words in the first reply of each of the two rooted path-units. This leads to a balanced binary prediction task: which of the two lexically similar rooted path-units is the successful one? With this setup, we attempt to roughly de-emphasize *what* is being said, in favor of *how* it is expressed.

We further avoid trivial cases, such as replies that are not arguments but clarifying questions, by removing cases where the root reply has fewer than 50 words. In order to make sure that there are enough counterarguments that the OP saw, motivated by the results in §3.2, we also require that there are at least 10 challengers in the discussion tree and at least 3 unsuccessful rooted path-units before the last reply that the OP made in the discussion tree.

In an ideal world, we would control for both length [13] and topic [25, 52], but we don't have the luxury of having enough data to do so. In our pilot experiments, annotators find that Jaccard-controlled pairs are easier to compare than length-matched pairs, as the lexical control is likely to produce arguments that make similar claims. Since length can be predictive (for instance, **C2** won a Δ in Figure 2), this raises the concern of false positive findings. Hence we develop a post-mortem "dissection" task (labelled *root truncated*) in which we only consider the root reply and truncate the longer one within a pair so that both root replies have the same number of words. This forcefully removes all length effects.

Disclaimer: Features that lose predictive power in the *root truncated* setting (or "reverse direction"[11]) are not necessarily false positives (or non-significant), as truncation can remove significant fractions of the text and lead to different distributions in the resultant dataset. Our point, though, is: if features retain predictive power *even in* the root truncated settings, they must be indicative beyond length.

We extract pairs from the training and heldout periods respectively as training data (3,456 pairs) and heldout testing data (807 pairs). Given that our focus is on language, we only use text-based features in this section.[12] In preprocessing, we remove explicit edits that users made after posting or commenting, and convert quotations and URLs into special tokens.

4.2 Features

In order to capture characteristics of successful arguments, we explore two classes of textual features: (§4.2.1) features that describe the interplay between a particular challenger's replies and the original post, and (§4.2.2) features that are solely based on his/her

[10] That is an intriguing problem for future work that requires a knowledge base and sophisticated semantic understanding of language.

[11] E.g., more of feature f is significantly better for *root reply*, but less f is significantly better in *root truncated*.

[12] An entry order baseline only achieves 54.3% training accuracy.

Table 2: Significance tests on interplay features. Features are sorted by average p-value in the two tasks. In all feature testing tables, the number of arrows indicates the level of p-value, while the direction shows the relative relationship between positive instances and negative instances, ↑↑↑↑: $p < 0.0001$, ↑↑↑: $p < 0.001$, ↑↑: $p < 0.01$, ↑: $p < 0.05$. T in the *root reply* column indicates that the feature is also significant in the *root truncated* condition, while T^R means that it is significant in *root truncated* but the direction is reversed.

Feature name	*root reply*	*full path*
reply frac. in all	↓↓↓↓ (T)	↓↓↓↓
reply frac. in content	↓↓↓↓ (T)	↓↓↓↓
OP frac. in stopwords	↑↑↑↑ (T^R)	↑↑↑↑
#common in stopwords	↑↑↑↑ (T^R)	↑↑↑↑
reply frac. in stopwords	↓↓↓↓	↓↓↓↓
OP frac. in all	↑↑↑↑ (T^R)	↑↑↑↑
#common in all	↑↑↑↑ (T^R)	↑↑↑↑
Jaccard in content	↓↓↓↓ (T)	↓↓↓↓
Jaccard in stopwords	↑↑↑↑ (T^R)	↑↑↑↑
#common in content	↑↑↑↑ (T^R)	↑↑↑↑
OP frac. in content	↑ (T^R)	↑↑↑↑
Jaccard in all	↓ (T)	

replies. We present those features that are statistically significant in the training data under the paired t-test with Bonferroni correction for multiple comparisons.

4.2.1 Interplay with the original post: Table 2

The context established by the OP's statement of her view can provide valuable information in judging the relative quality of a challenger's arguments. We capture the interplay between arguments and original posts through similarity metrics based on word overlap.[13] We consider four variants based on the number of unique words in common between the argument (A) and the original post (O):

- number of common words: $|A \cap O|$,
- reply fraction: $\frac{|A \cap O|}{|A|}$,
- OP fraction: $\frac{|A \cap O|}{|O|}$,
- Jaccard: $\frac{|A \cap O|}{|A \cup O|}$.

While stopwords may be related to how challengers coordinate their style with the OP [14, 35], content words can be a good signal of new information or new perspectives. Thus, inspired by previous results distinguishing these vocabulary types in studying the effect of phrasing [52], for each of the four variants above we try three different word sets: stopwords, content words and all words.

The features based on interplay are all significant to a certain degree. Similar patterns occur in *root reply* and *full path*: in number of common words and OP fraction, persuasive arguments have larger values because they tend to be longer, as will be shown in §4.2.2; in reply fraction and Jaccard, which are normalized by reply length, persuasive arguments are more dissimilar from the original post in content words but more similar in stopwords. Keeping in mind that the pairs we compare are chosen to be similar to each other, our analysis indicates that, under this constraint, persuasive arguments use a more different wording from the original post in content, while at the same time matching them more on stopwords.

[13] We also tried *tf-idf*, topical, and word embedding–based similarity in cross validation on training data. We defer discussion of potentially useful features to §6.

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. 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. We remind the reader of the *root truncated* disclaimer from §4.

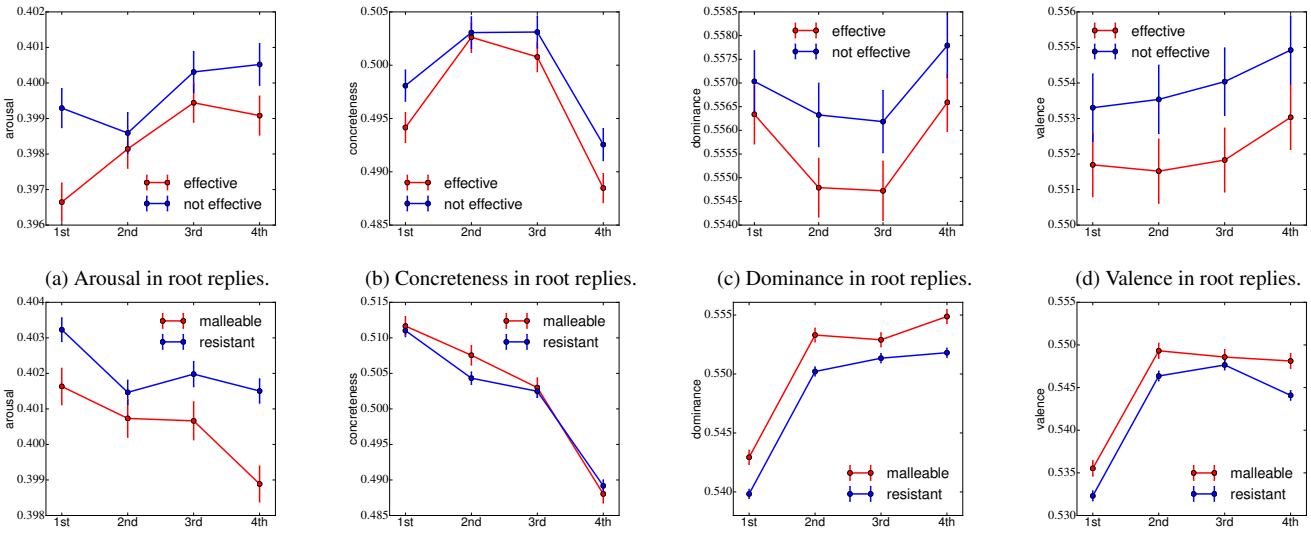
Feature name	*root reply*	*full path*
#words	↑↑↑↑	↑↑↑↑
Word category–based features		
#definite articles	↑↑↑↑	↑↑↑↑
#indefinite articles	↑↑↑↑	↑↑↑↑
#positive words	↑↑↑↑ (T^R)	↑↑↑↑
#2nd person pronoun	↑↑↑↑	↑↑↑↑
#links	↑↑↑↑ (T)	↑↑↑↑
#negative words	↑↑↑↑	↑↑↑↑
#hedges	↑↑↑↑	↑↑↑↑
#1st person pronouns	↑↑↑↑	↑↑↑↑
#1st person plural pronoun	↑↑↑↑	↑↑↑↑
#`.com` links	↑↑↑↑ (T)	↑↑↑↑
frac. links	↑↑↑↑ (T)	↑↑↑↑
frac. `.com` links	↑↑↑↑ (T)	↑↑↑↑
#examples	↑	↑↑↑↑
frac. definite articles	↑ (T)	↑↑
#question marks	↑ —	↑↑↑↑
#PDF links	↑	↑↑↑
#`.edu` links		↑
frac. positive words	↓	
frac. question marks	—	↓
#quotations		↑↑↑↑
Word score–based features		
arousal	↓ (T)	↓↓↓
valence	↓	
Entire argument features		
word entropy	↑↑↑↑	↑↑↑↑
#sentences	↑↑↑↑ —	↑↑↑↑
type-token ratio	↓↓↓↓ (T^R)	↓↓↓↓
#paragraphs	↑↑↑↑ —	↑↑↑↑
Flesch-Kincaid grade levels	—	↓↓↓
Markdown formatting		
#italics	↑↑↑↑ —	↑↑↑↑
bullet list	↑↑↑↑ —	↑↑↑↑
#bolds	↑↑ —	↑↑↑↑
numbered words	↑	↑↑↑↑
frac. italics	↑ —	↑

If we instead use truncation to (artificially) control for reply length, persuasive arguments present lower similarity in all metrics, suggesting that effects might differ over local parts of the texts. However, it is consistent that successful arguments are less similar to the original post in content words.

4.2.2 Argument-only features: Table 3

We now describe cues that can be extracted solely from the replies. These features attempt to capture linguistic style and its connections to persuasion success.

Number of words. A straightforward but powerful feature is the number of words. In both *root reply* and *full path*, a larger number of words is strongly correlated with success. This is not surprising: longer replies can be more explicit [37, 38] and convey more information. But naïvely making a communication longer does not automatically make it more convincing (indeed, sometimes, more succinct phrasing carries more punch); our more advanced features attempt to capture the subtler aspects of length.



(a) Arousal in root replies. (b) Concreteness in root replies. (c) Dominance in root replies. (d) Valence in root replies.

(e) Arousal in original posts. (f) Concreteness in original posts. (g) Dominance in original posts. (h) Valence in original posts.

Figure 6: Style features in different quarters. The first row shows how arousal, concreteness, dominance and valence change in different quarters of the root reply, while the second row shows the same features in the original posts. The descending concreteness trend suggests that opinions tend to be expressed in a particular-to-general way; replies notably differ by having both the opening and the closing be abstract, with a concrete middle. These differences are indicative of the functions that the two forms of utterances serve: a CMV rule is that original posts should not be "like a persuasive essay". Error bars represent standard errors.

Word category–based features. As suggested by existing psychology theories and our intuitions, the frequency of certain types of words may be associated with persuasion success. We consider a wide range of categories (see §9 for details), where for each, we measure the raw number of word occurrences and the length-normalized version.

Word score–based features. Beyond word categories, we employ four scalar word-level attributes [5, 57]:

- Arousal captures the intensity of an emotion, and ranges from "calm" words (*librarian, dull*) to words that excite, like *terrorism* and *erection*.
- Concreteness reflects the degree to which a word denotes something perceptible, as opposed to abstract words which can denote ideas and concepts, e.g., *hamburger* vs. *justice*.
- Dominance measures the degree of control expressed by a word. Low-dominance words can suggest vulnerability and weakness (*dementia, earthquake*) while high-dominance words evoke power and success (*completion, smile*).
- Valence is a measure of how pleasant the word's denotation is. Low-valence words include *leukemia* and *murder*, while *sunshine* and *lovable* are high-valence.

We scale the four measures above to lie in $[0, 1]$.[14] We extend these measures to texts by averaging over the ratings of all content words. Table 3 shows that it is consistently good to use calmer language. Aligned with our findings in terms of sentiment words (§9), persuasive arguments are slightly less happy. However, no significant differences were found for concreteness and dominance.

[14]While the resources cover most common words, out-of-vocabulary misses can occur often in user-generated content. We found that all four values can be extrapolated with high accuracy to out-of-vocabulary words by regressing on dependency-based word embeddings [29] (median absolute error of about 0.1). Generalizing lexical attributes using word embeddings was previously used for applications such as figurative language detection [55].

Characteristics of the entire argument. We measure the number of paragraphs and the number of sentences: persuasive arguments have significantly more of both. To capture the lexical diversity in an argument, we consider the *type-token ratio* and *word entropy*. Persuasive arguments are more diverse in *root reply* and *full path*, but the *type-token ratio* is surprisingly higher in *root truncated*: because of correlations with length and argument structure, lexical diversity is hard to interpret for texts of different lengths. Finally, we compute Flesch-Kincaid grade level [26] to represent readability. Although there is no significant difference in *root reply*, persuasive arguments are more complex in *full path*.

Formatting. Last but not least, discussions on the Internet employ certain writing conventions enabled by the user interface. Since Reddit comments use Markdown[15] for formatting, we can recover the usage of bold, italic, bullet lists, numbered lists and links formatting.[16] While these features are not applicable in face-to-face arguments, more and more communication takes place online, making them highly relevant. Using absolute number, most of them are significant except numbered lists. When it comes to normalized counts, though, only italicizing exhibits significance.

4.2.3 *They hold no quarter, they ask no quarter*

Understanding how a line of argument might evolve is another interesting research problem. We investigate by quartering each argument and measuring certain feature values in each quarter, allowing for finer-grained insight into argument structure.

Word score–based features in quarters. (Figure 6) With the exception of arousal, effective arguments and ineffective arguments present similar patterns: the middle is more concrete and less dominant than the beginning and end, while valence rises slightly over the course of an argument. We also see interesting differences in psycholinguistic patterns between original posts and replies. (We

[15] `https://daringfireball.net/projects/markdown/`

[16]We also consider numbered words (*first*, *second*, *third*, etc.) as the textual version of numbered lists.

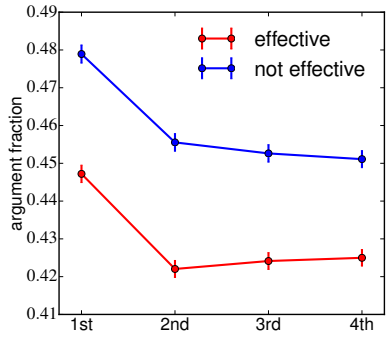


Figure 7: Similarity between each quarter of an argument and the entire original post.

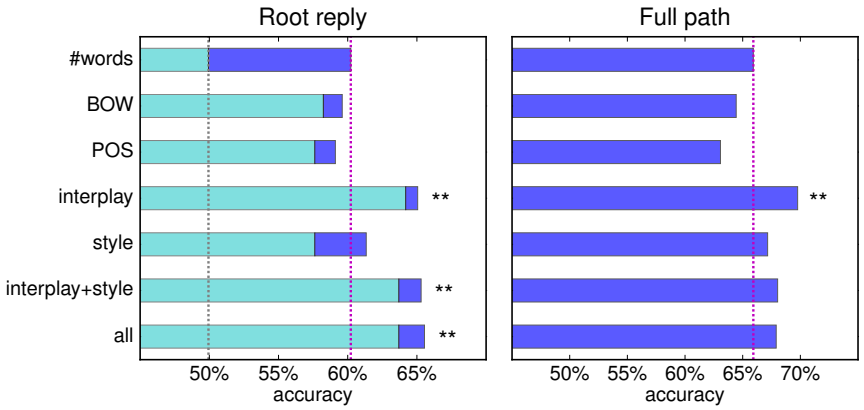


Figure 8: **Prediction results.** The cyan fraction in the left figure shows the performance in *root truncated*, and the purple bar shows the performance in *root reply*. The magenta line shows the performance of *#words* in *root reply*, while the gray line shows the performance of *#words* in *root truncated*, which is the same as random guessing. The figure on the right gives the performance in *full path* (the magenta line gives the performance of *#words*). The number of stars indicate the significance level compared to the *#words* baseline according to McNemar's test. (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.)

defer detailed discussion to §5.) In terms of arousal, however, successful arguments begin by using calmer words.

Interplay with the original post. (Figure 7) To capture partial overlap and possible divergence from the OP's view, we divide both the original post and the rooted path-unit into quarters, and measure similarity metrics between all subdivisions (including the full unit).[17] Since the reply fraction in content words is the most significant interplay feature, in Figure 7 we only show the fraction of common content words in different quarters of replies vs. the original post. Both effective and ineffective arguments start off more similar with the original post; effective arguments remain less similar overall.

4.3 Prediction results

We train logistic regression models with ℓ_1 regularization on the training set and choose parameters using five cross-validation folds, ensuring that all pairs of arguments that share the same OP are in the same fold.[18] All features are standardized to unit variance, and missing values are imputed using the training fold sample mean. We evaluate using pairwise accuracy in the *heldout dataset*, where we restricted ourselves to *a single experimental run* (after holding our collective breath) to further reduce the risk of overfitting. The results are, in fact, in line with what we describe in the training-data analysis here.

Feature sets. As shown in §4.2, the number of words is very predictive, providing a strong baseline to compare against. Bag-of-words features (*BOW*) usually provide a strong benchmark for text classification tasks. We restrict the size of the vocabulary by removing rare words that occurred no more than 5 times in training and ℓ_2-normalize term frequency vectors. Since part-of-speech tags may also capture properties of the argument, we also use normalized term frequency vectors by treating part-of-speech tags as words (*POS*). Features in §4.2.1 are referred to as *interplay*; features in §4.2.2 constitute the feature set *style*. Finally, we use a combination of style and interplay, as well as a combination that includes all the above features (*all*). Note that style and interplay are dense and very low-dimensional compared to *BOW*.

Interplay with the OP plays an essential role. (Figure 8) *#words* is indeed a very strong baseline that achieves an accuracy of 60% in *root reply* and 66% in *full path*. As a sanity check, in *root truncated*, it indeed gets only 50%. In comparison, *BOW* achieves similar performance as *#words*, while *POS* gives even worse performance. However, interplay features lead to a 5% absolute improvement over the *#words* baseline in *root reply* and *full path*, and a 14% absolute improvement in *root truncated*. In fact, the performance of interplay is already close to using the combination of interplay and style and using all features. In *root truncated*, although the performance of style features drops significantly, interplay achieves very similar performance as in *root reply*, demonstrating the robustness of the interplay features.

5. "RESISTANCE" TO PERSUASION

Although it is a good-faith step for a person to post on CMV, some beliefs in the dataset are still "resistant" to changes, possibly depending on how strongly the OP holds them and how the OP acquired and maintained them [45, 54, 59]. Since CMV members must state their opinion and reasons for it in their own words, we can investigate differences between how resistant and malleable views are expressed. In this section, we seek linguistic and style patterns characterizing original posts in order to better understand the mechanisms behind attitude resistance and expression, and to give potential challengers a sense of which views may be resistant before they engage.

However, recognizing the "malleable" cases is not an easy task: in a pilot study, human annotators perform at chance level (50% on a paired task to distinguish which of two original posts is malleable). In light of our observation that persuasion is unsuccessful in 70% of the cases from §3, we set up an imbalanced prediction task. We focus on cases where at least 10 challengers attempt counterarguments, and where the OP replied at least once,[19] alleviating

[17] In prediction, we also take the maximum and minimum of these quarter-wise measures as an order-independent way to summarize fragment similarity.

[18] We also tried ℓ_2 regularization, random forests and gradient boosting classifiers and found no improvement beyond the cross-validation standard error.

[19] Although in preprocessing we replaced all explicit edits, we also remove all posts containing the word *"changed"*, to avoid including post-hoc signals of view change.

Table 4: Opinion malleability task: statistically significant features after Bonferroni correction.

Feature name	More malleable?
#1st person pronouns	↑↑↑↑
frac. 1st person pronoun	↑↑↑↑
dominance	↑↑↑↑
frac. 1st person plural pronoun	↓↓↓
#paragraphs	↑↑
#1st person plural pronoun	↓↓
#bolds	↑
arousal	↓
valence	↑
bullet list	↑

the concern that an opinion appears resistant simply because there was little effort towards changing it. This brings us 10,743 original posts in the training data and 1,529 original posts in the heldout data. We then analyze systematic expression patterns that characterize malleable beliefs and that signal open-mindedness.

5.1 Stylistic features for open-mindedness

We employ the same set of features from §4.2.2 to capture the characteristics of original posts. Among them, only a handful are significantly predictive of malleability, as shown in Table 4.

Personal pronouns and self-affirmation. First person pronouns are strong indicators of malleability, but first person plural pronouns correlate with resistance. In psychology, self-affirmation has been found to indicate open-mindedness and make beliefs more likely to yield [11, 12]. Our result aligns with these findings: 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. Note that it was also found in other work that openness is negatively correlated with first person singular pronouns [39].

Formatting. The use of more paragraphs, bold formatting, and bulleted lists are all higher when a malleable view is expressed. Taking more time and presenting the reasons behind an opinion in a more elaborated form can indicate more engagement.

Word score–based features. Dominance is the most predictive of malleability: the average amount of control expressed through the words used is higher when describing a malleable view than a resistant one. The same holds for happiness (captured by valence). In terms of arousal, malleable opinions are expressed significantly more serenely, ending on a particularly calm note in the final quarter, while stubborn opinions are expressed with relatively more excitement.

5.2 Prediction performance

We use weighted logistic regression and choose the amount and type of regularization (ℓ_1 or ℓ_2) by grid search over 5 cross-validation folds. Since this is an imbalanced task, we evaluate the prediction results using the area under the ROC curve (AUC) score. As in §4, we use the number of words as our baseline. In addition to the above features that characterizes language style (*style*), we use bag-of-words (*BOW*), part-of-speech tags (*POS*) and a full feature set (*all*). The holdout performance is shown in Figure 9.

The classifiers trained on bag of words features significantly outperforms the *#words* baseline. Among words with largest coefficients, resistant views tend to be expressed using more decisive words such as *anyone*, *certain*, *ever*, *nothing*, and *wrong*, while *help* and *please* are malleable words. The *POS* classifier significantly outperforms random guessing, but not the baseline. Nevertheless, it yields an interesting insight: comparative adjectives and adverbs are signs of malleability, while superlative adjectives suggest stubbornness. The full feature set (*all*) also significantly outperform the *#words* baseline. The overall low scores suggest that this is indeed a challenging task for both humans and machines.

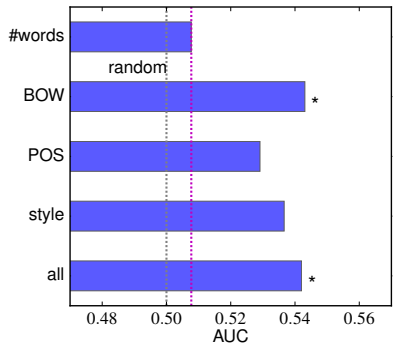


Figure 9: **Opinion malleability prediction performance**: AUC on the heldout dataset. The purple line shows the performance of *#words*, while the gray line gives the performance of random guessing. The *BOW* and *all* feature sets perform significantly better than the *#words* baseline, according to one-sided paired permutation tests. *BOW*, *POS*, *style* and *all* outperform random guessing using bootstrapped tests. (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.)

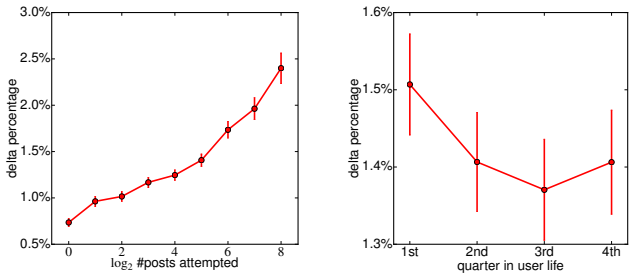
6. FURTHER DISCUSSION

Here we discuss other observations that may open up avenues for further investigation of the complex process of persuasion.

Experience level. Beyond the interactions within a discussion tree, CMV is a community where users can accumulate experience and potentially improve their persuasion ability. Figure 10a shows that a member's success rate goes up with the number of attempts made. This observation can be explained by at least two reasons: the success rate of frequent challengers improves over time, and/or frequent challengers are better at persuasion from the beginning. To disentangle these two possible reasons, for challengers who attempted to change at least 16 views, we split all the attempts into 4 equal chunks sorted by time. Figure 10b presents how the success rate changes over a challenger's life, suggesting that the success rate of frequent challengers does not increase.[20] It is worth noting that this lack of apparent improvement might be explained by a gradual development of a "taste" for original posts that are harder to address [30]. Such community dynamics point to interesting research questions for future work.

Attempts to capture high-level linguistic properties. We experimented with a broader set of features in cross validation, which we still deem interesting but did not discuss in depth for space reasons. One important class are attempts to capture the semantics of original statements and arguments. We experimented with using topic models [3] to find topics that are the most malleable (*topic: food, eat, eating, thing, meat* and *topic: read, book, lot, books, women*), and the most resistant (*topic: government, state, world, country, countries* and *topic: sex, women, fat, person, weight*). However, topic model based features do not seem to bring predictive power to either of the tasks. For predicting persuasive arguments, we attempted to capture interplay with word embeddings for text similar-

[20]In terms of the correlation between previous success (lifetime deltas) and success rate, the result is similar: beyond 4–5 deltas there is no noticeable increase.



(a) Delta ratio vs. $\log_2$ #posts. (b) No growth over life.

Figure 10: Effect of experience.

ity using both the centroid distance and the word mover's distance [27]. Both distances proved predictive by themselves, but were not able to improve over the features presented in the paper in cross validation. More generally, better semantic models applicable to online discussions could open up deeper investigations into effective persuasion strategies.

Sequential argument structure. Another promising direction is to examine the structure of arguments via the sequence of discourse connectors. For instance, we can recover interesting structures such as "$first_0$–but_1–$because_2$" and "now_1–$then_2$–$instead_3$", where the subscripts indicate which quarter the discourse connector occurred in. These features did not perform well in our tasks due to low recall, or lack of argumentative structure in the data, but they deserve further exploration.

7. ADDITIONAL RELATED WORK

A few lines of research in natural language processing are related to our work. Argumentation mining focuses on fine-grained analysis of arguments and on discovering the relationships, such as support and premise, between different arguments [34]. Studies have also worked on understanding persuasive essays [18, 41, 51] and opinion analysis in terms of agreement and ideology [53, 49, 22, 50, 48]. Another innovative way of using Internet data to study mass persuasion is through AdWords [20].

8. CONCLUSION

In this work, in order to understand the mechanisms behind persuasion, we use a unique dataset from `/r/ChangeMyView`. In addition to examining interaction dynamics, we develop a framework for analyzing persuasive arguments and malleable opinions. We find that not only are interaction patterns connected to the success of persuasion, but language is also found to distinguish persuasive arguments. Dissimilarity with the wording in which the opinion is expressed turns out to be the most predictive signal among all features. Although members of CMV are open-minded and willing to change, we are still able to identify opinions that are resistant and to characterize them using linguistic patterns.

There are many possible extensions to our approach for representing arguments. In particular, it would be interesting to model the framing of different arguments and examine the interplay between framing of the original post and the replies. For instance, is benefit-cost analysis the only way to convince a utilitarian?

Furthermore, although this novel dataset opens up potential opportunities for future work, other environments, where people are not as open-minded, can exhibit different kinds of persuasive interactions; it remains an interesting problem how our findings generalize to different contexts. It is also important to understand the effects of attitude change on actual behavior [44].

Finally, beyond mechanisms behind persuasion, it is a vital research problem to understand how community norms encourage such a well-behaved platform so that useful rules, moderation practices, or even automated tools can be deployed in future community building.

9. APPENDIX

In this section we explain the features based on word categories.

- (In)definite articles (inspired by [13]). These are highly correlated with length, so they are both highly significant in terms of absolute numbers. However, in terms of word ratios, definite articles (e.g., "the" instead of "a") are preferred, which suggests that specificity is important in persuasive arguments.
- Positive and negative words. We use the positive and negative lexicons from LIWC [40]. In absolute numbers, successful arguments are more sentiment-laden in both *root reply* and *full path*. When truncating, as well as when taking the frequency ratio, persuasive opening arguments use *fewer* positive words, suggesting more complex patterns of positive emotion in longer arguments [23, 58].
- Arguer-relevant personal pronouns. We consider 1st person pronouns (*me*) 2nd person pronouns (*you*) and 1st person plural pronouns (*us*). In both *root reply* and *full path*, persuasive arguments use a significantly larger absolute number of personal pronouns.
- Links. Citing external evidence online is often accomplished using hyperlinks. Persuasive arguments use consistently more links, both in absolute and in per-word count. We make special categories for interesting classes of links: those to `.com` and `.edu` domains, and those to PDF documents. Maybe due to high recall, `.com` links seem to be most powerful. Features based on links also tend to be significant even in the *root truncated* condition.
- Hedging. Hedges indicate uncertainty; an example is "It could be the case". Their presence might signal a weaker argument [16], but alternately, they may make an argument easier to accept by softening its tone [28]. We curate a set of hedging cues based on [21, 24]. Hedging is more common in persuasive arguments under *root reply* and *full path*.
- Examples. We consider occurrences of "for example", "for instance", and "e.g.". The absolute number of such example markers is significantly higher in persuasive arguments.
- Question marks. Questions can be used for clarification or rhetorical purposes. In terms of absolute number, there are more in *root reply* and *full path*. But when it comes to ratio, if anything, it seems better to avoid using question marks.
- Quotations. One common practice in argumentation is to quote the other party's words. However, this does not seem to be a useful strategy for the root reply.

Acknowledgments. We thank X. Chen, J. Hessel, S. Hu, S. Kundu, N. Li, M. Raghu, N. Rojas, A. Schofield, T. Shi, and J. Zhang for participating in our pilot annotation experiments. We thank Y. Artzi, J. Carpenter, P. Grabowicz, J. Hessel, Y. Hua, J. Park, M. Rooth, A. Schofield, A. Singh, and the anonymous reviewers for helpful comments. This work was supported in part by NSF grant IIS-0910664, a Google Research Grant, a Google Faculty Research Award and a Facebook Fellowship.

References

[1] T. Althoff, C. Danescu-Niculescu-Mizil, and D. Jurafsky. How to ask for a favor: A case study on the success of altruistic requests. In *Proceedings of ICWSM*, 2014.

[2] M. Bailey, D. J. Hopkins, and T. Rogers. Unresponsive and unpersuaded: The unintended consequences of voter persuasion efforts. In *Meeting of the Society for Political Methodology at the University of Virginia*, 2014.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[4] C. J. Bryan, G. M. Walton, T. Rogers, and C. S. Dweck. Motivating voter turnout by invoking the self. *Proceedings of the National Academy of Sciences*, 108(31):12653–12656, 2011.

[5] M. Brysbaert, A. B. Warriner, and V. Kuperman. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911, 2014.

[6] M. Burgoon, S. B. Jones, and D. Stewart. Toward a message-centered theory of persuasion: Three empirical investigations of language intensity. *Human Communication Research*, 1 (3):240–256, 1975.

[7] S. Chaiken. The heuristic model of persuasion. In *Social influence: The Ontario Symposium*, 1987.

[8] M. J. Chambliss and R. Garner. Do adults change their minds after reading persuasive text? *Written Communication*, 13(3): 291–313, 1996.

[9] R. B. Cialdini. *Influence: The Psychology of Persuasion*. HarperCollins, 1993.

[10] R. B. Cialdini, W. Wosinska, D. W. Barrett, J. Butner, and M. Gornik-Durose. Compliance with a request in two cultures: The differential influence of social proof and commitment/consistency on collectivists and individualists. *Personality and Social Psychology Bulletin*, 25(10):1242–1253, 1999.

[11] G. L. Cohen, J. Aronson, and C. M. Steele. When beliefs yield to evidence: Reducing biased evaluation by affirming the self. *Personality and Social Psychology Bulletin*, 26(9): 1151–1164, 2000.

[12] J. Correll, S. J. Spencer, and M. P. Zanna. An affirmed self and an open mind: Self-affirmation and sensitivity to argument strength. *Journal of Experimental Social Psychology*, 40(3): 350–356, 2004.

[13] C. Danescu-Niculescu-Mizil, J. Cheng, J. Kleinberg, and L. Lee. You had me at hello: How phrasing affects memorability. In *Proceedings of ACL*, 2012.

[14] C. Danescu-Niculescu-Mizil, L. Lee, B. Pang, and J. Kleinberg. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of WWW*, 2012.

[15] J. P. Dillard and L. Shen. *The Persuasion Handbook: Developments in Theory and Practice*. SAGE Publications, 2014.

[16] A. M. Durik, M. A. Britt, R. Reynolds, and J. Storey. The effects of hedges in persuasive arguments: A nuanced analysis of language. *Journal of Language and Social Psychology*, 2008.

[17] A. H. Eagly and S. Chaiken. *The Psychology of Attitudes*. Harcourt Brace Jovanovich College Publishers, 1993.

[18] N. Farra, S. Somasundaran, and J. Burstein. Scoring persuasive essays using opinions and their targets. In *Proceedings of the NAACL Workshop on Innovative Use of NLP for Building Educational Applications*, 2015.

[19] B. J. Fogg. Mass interpersonal persuasion: An early view of a new phenomenon. In *Persuasive Technology*, 2008.

[20] M. Guerini, C. Strapparava, and O. Stock. Evaluation metrics for persuasive NLP with Google AdWords. In *Proceedings of LREC*, 2010.

[21] D. A. Hanauer, Y. Liu, Q. Mei, F. J. Manion, U. J. Balis, and K. Zheng. Hedging their mets: The use of uncertainty terms in clinical documents and its potential implications when sharing the documents with patients. In *Proceedings of the AMIA Annual Symposium*, 2012.

[22] K. S. Hasan and V. Ng. Why are you taking this stance? Identifying and classifying reasons in ideological debates. In *Proceedings of EMNLP*, 2014.

[23] C. R. Hullett. The impact of mood on persuasion: A meta-analysis. *Communication Research*, 32(4):423–442, 2005.

[24] K. Hyland. *Hedging in Scientific Research Articles*. John Benjamins Publishing, 1998.

[25] A. Jaech, V. Zayats, H. Fang, M. Ostendorf, and H. Hajishirzi. Talking to the crowd: What do people react to in online discussions? In *Proceedings of EMNLP*, pages 2026–2031, 2015.

[26] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom. Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for Navy enlisted personnel. *Chief of Naval Technical Training, Naval Air Station, Research Branch Report 8-75*, 1975.

[27] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger. From word embeddings to document distances. In *Proceedings of ICML*, 2015.

[28] G. Lakoff. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic*, 2, 1975.

[29] O. Levy and Y. Goldberg. Dependency-based word embeddings. In *Proceedings of ACL*, 2014.

[30] J. J. McAuley and J. Leskovec. From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. In *Proceedings of WWW*, 2013.

[31] A. K. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[32] D. McRaney. The backfire effect. `http://youarenotsosmart.com/2011/06/10/the-backfire-effect/`, 2011. Accessed: 2015-10-15.

[33] T. Mitra and E. Gilbert. The Language that Gets People to Give: Phrases that Predict Success on Kickstarter. In *Proceedings of CSCW*, 2014.

[34] R. Mochales and M.-F. Moens. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22, 2011.

[35] K. G. Niederhoffer and J. W. Pennebaker. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360, 2002.

[36] B. Nyhan and J. Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2): 303–330, 2010.

[37] D. J. O'Keefe. Standpoint explicitness and persuasive effect: A meta-analytic review of the effects of varying conclusion articulation in persuasive messages. *Argumentation and Advocacy*, 34(1):1–12, 1997.

[38] D. J. O'Keefe. Justification explicitness and persuasive effect: A meta-analytic review of the effects of varying support articulation in persuasive messages. *Argumentation and Advocacy*, 35(2):61–75, 1998.

[39] J. W. Pennebaker and L. A. King. Linguistic styles: language use as an individual difference. *J Pers Soc Psychol*, 77(6): 1296–1312, 1999.

[40] J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic inquiry and word count: LIWC 2007. Technical report, 2007.

[41] I. Persing and V. Ng. Modeling argument strength in student essays. In *Proceedings of ACL*, 2015.

[42] R. E. Petty and J. T. Cacioppo. *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*. Springer Science & Business Media, 2012.

[43] R. E. Petty and D. T. Wegener. Matching versus mismatching attitude functions: Implications for scrutiny of persuasive messages. *Personality and Social Psychology Bulletin*, 24(3): 227–240, 1998.

[44] R. E. Petty, D. T. Wegener, and L. R. Fabrigar. Attitudes and attitude change. *Annual Review of Psychology*, 48(1):609–647, 1997.

[45] E. M. Pomerantz, S. Chaiken, and R. S. Tordesillas. Attitude strength and resistance processes. *Journal of Personality and Social Psychology*, 69(3):408, 1995.

[46] S. L. Popkin. *The Reasoning Voter: Communication and Persuasion in Presidential Campaigns*. University of Chicago Press, Chicago, 1994.

[47] K. K. Reardon. *Persuasion in Practice*. Sage, 1991.

[48] S. Rosenthal and K. McKeown. I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *Proceedings of SIGdial*, 2015.

[49] S. Somasundaran and J. Wiebe. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 2010.

[50] D. Sridhar, J. Foulds, B. Huang, L. Getoor, and M. Walker. Joint models of disagreement and stance in online debate. In *Proceedings of ACL*, 2015.

[51] C. Stab and I. Gurevych. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of EMNLP*, 2014.

[52] C. Tan, L. Lee, and B. Pang. The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. In *Proceedings of ACL*, 2014.

[53] M. Thomas, B. Pang, and L. Lee. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of EMNLP*, 2006.

[54] Z. L. Tormala and R. E. Petty. What doesn't kill me makes me stronger: The effects of resisting persuasion on attitude certainty. *Journal of Personality and Social Psychology*, 83 (6):1298, 2002.

[55] Y. Tsvetkov, L. Boytsov, A. Gershman, E. Nyberg, and C. Dyer. Metaphor detection with cross-lingual model transfer. In *Proceedings of ACL*, 2014.

[56] O. Tykocinski, E. T. Higgins, and S. Chaiken. Message framing, self-discrepancies, and yielding to persuasive messages: The motivational significance of psychological situations. *Personality and Social Psychology Bulletin*, 20(1):107–115, 1994.

[57] A. B. Warriner, V. Kuperman, and M. Brysbaert. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207, 2013.

[58] D. T. Wegener and R. E. Petty. Effects of mood on persuasion processes: Enhancing, reducing, and biasing scrutiny of attitude-relevant information. In L. L. Martin and A. Tesser, editors, *Striving and Feeling: Interactions Among Goals, Affect, and Self-regulation*. Psychology Press, 1996.

[59] J. R. Zuwerink and P. G. Devine. Attitude importance and resistance to persuasion: It's not just the thought that counts. *Journal of Personality and Social Psychology*, 70(5):931, 1996.