Persuasive Argument Prediction: Toy Analysis

Rui Wang

September 25, 2023

1 Introduction and Dataset

This project was a toy endeavor into understanding social interaction dynamics in persuasive arguments and understanding malleability and resistance in opinion expression through language indicators and writing mannerisms, heavily inspired by this project. The dataset used in this analysis is available here. This project analyzes a collection of posts and comments on the popular Reddit community `r/ChangeMyView`, which has 3.4 million members at present, from 2013 to 2015. A sufficient knowledge of the mechanics behind `r/ChangeMyView` is assumed. The analysis is comprised of 3 different tasks and objectives: Interaction Dynamics, Language Indicators, and Openness to Persuasion. For each task in the analysis, a random sample of 500 entries from the dataset of that respective task was chosen, with a proportionate amount chosen for any holdout and testing.

2 Interaction Dynamics Between OP and Challenger

The first analysis task is regarding the social dynamics behind the interactions between the original poster (OP henceforth) and their challengers in terms of the success of the challenger in changing the OP's view.

From our random sample of 500 posts and their comments, we follow the original paper's refinement and only select those that have more than 10 comments as well as at least one reply from the OP in the comments.

2.1 Number of Comments

From a surface-level view, it can be sensible to intuit that a successful challenger tends to provide their opinion more and writes more comments. As a preliminary understanding, we first look to understand how the number of comments a challenger makes impacts their chances of success, measured by delta percentage. We see from Figure 1 that writing more comments generally leads to more success, but after a certain point it becomes less effective. However, this analysis is crude and may be conflated with other comment-related factors, such as the effect of interacting with the OP vs. other fellow challengers, or even the order at which a challenger enters the discussion, as earlier challengers will have the opportunity to write more comments. Thus, we continue to a more fine-tuned analysis on factors of challenger success in the subsequent entries for a better understanding.

2.2 Delta Chances by Entry

Mirroring the analysis done in the original paper, we endeavor to understand how the chances of OP conversion are impacted by the order at which a challenger enters the discussion. We see in Figure 2 that the chances of OP conversion are the highest for the first couple challengers to the discussion, mirroring the conclusions in the original paper.

This can be explained through a combination of earlier challengers having the ability to be the first to craft novel and meaningful rebuttals to the OP's claims as well as generally being the first replies the OP will read, as `r/ChangeMyView` mandates that the OP be quick to respond to any replies within 3 hours.

However, the levelling out for the later entries rather than a continued decrease can be accounted for with a similar idea, in that the later challengers can elaborate or develop on the earlier challengers' arguments,

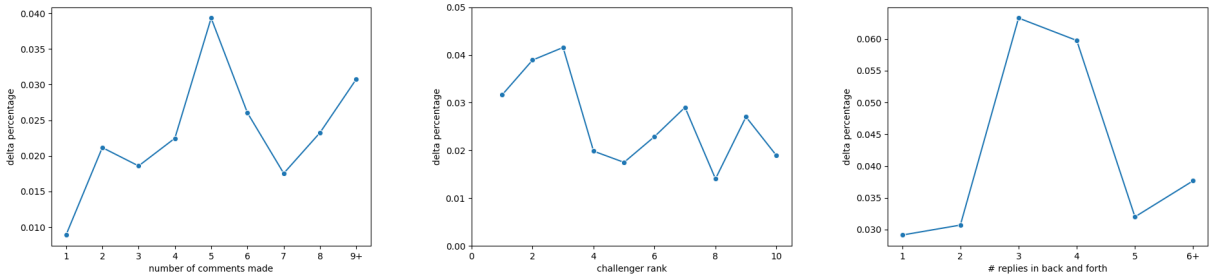


Fig. 1 (left) is a graph of delta percentage for challengers based on the total number of comments they write. Fig. 2 (middle) is a graph of delta percentage for challengers based on order of their entry to the discussion. Fig. 3 (right) is graph of the delta percentage for challengers based on the length of conversation with the OP.

leading to a more cohesive argument. Additionally, since `r/ChangeMyView` allows for an OP to award multiple Deltas per post, the contribution of a later challenger developing off of an earlier challenger's argument can plausibly represent the push over the edge for an OP, leading to the OP awarding Deltas to both the earlier challenger for their initial argument, and the later challenger.

2.3 Level of Back-and-forth

Continuing the analyses as with the original paper, we endeavor to understand how the degree of back-and-forth conversations between an OP and a challenger impacts the success of the challenger. Figure 3 details the length of a conversation between an OP and a challenger and the delta ratio for the challenger. We see that excessive communication between an OP and a challenger may signify more resistance by the OP and confidence in their perspective and are thus not productive, while too brief of an interaction is also not indicative of success, which is consistent with the conclusions presented in the original paper.

3 Language Characteristics of Successful Arguments

Continuing with the analyses conducted in the original paper, we endeavor to understand some key characteristics of the format, language, and style used by challengers in their arguments and their effect on their success.

This analysis is comprised of three sections, with the first being a study into how the challenger stylizes their argument to be similar or dissimilar to the OP's writing, which we name as interplay, following the original paper's terminology. Secondly, we analyze several text-based and language features of solely the argument, which we name as argument/style. Lastly, we aim to understand the evolution of an argument's features, studied quarter-by-quarter.

3.1 Interplay Between OP and Challenger

We capture several features that describe the interplay and similarity between the OP (O) and challenger's argument (A). Similar to the features presented in the original paper, the base feature metrics are:

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

Additionally, analysis is performed on both the root reply (just the original reply), as well as the full path (the original reply plus all subsequent child replies) made by a challenger. We then compute each of these bases on the versions of the argument and original post:

- without removing stopwords
- in just the stopwords
- after removing stopwords (we henceforth call this the content words)

Disclaimer: One key difference between this toy endeavor and the original paper is the exclusion of an analysis of the above features in the *root-truncated* case. The simple truncation of a root reply to account for the impact of length was weighed against the potential loss of critical aspects and information of the argument, and ultimately was deemed a useful but still too brute method - especially considering the aspect of the evolution of an argument as studied in the quarter-by-quarter analysis; as well as the limited sample size of which the effects of such truncation may lead to wildly different distributions and results compared to applying this method on a larger dataset.

Similar to the original post, we then conduct paired t-tests, with Bonferroni correction, on these features to determine significance. We see from Table 1 that for the features not normalized by length, such as the number of common words, persuasive arguments tend to have more overlap, which is sensible due to persuasive arguments being longer. However, in the features that are normalized for by length, we see that persuasive arguments tend to be more similar to the OP's argument in their usage for stopwords, while more dissimilar in its content words and in general. This is consistent with the conclusions of the original paper, and hints that successful arguments are generally dissimilar to the OP's in terms of idea and word usage, except in the presentation of the stopwords which may provide the OP a sense of lexical comfort.

Feature Name	root reply	full path
reply frac. in stopwords	↑↑↑↑	↑↑↑↑
OP frac. in all	↓↓↓↓	↓↓↓↓
# common in stopwords	↑↑↑↑	↑↑↑↑
OP frac. in content	↓↓↓↓	↓↓↓↓
reply frac. in all	↑↑↑↑	↑↑↑↑
Jaccard in stopwords	↑↑↑↑	↑↑↑↑
# common in all	↑↑↑↑	↑↑↑↑
OP frac. in stopwords	↓↓	↓↓↓↓
reply frac. in content		↑↑↑↑
# common in content		↑↑↑↑

Table 1: Significant features for interplay between an OP and a challenger, through a paired t-test with Bonferroni correction. Features area ranked on their average score across the root reply and full path cases. Borrowing the notation of the original paper, the direction of the arrow represents the relationship between the successful (positive) and unsuccessful (negative) instances. The number of arrows represents the level of significance: ↑↑↑↑: $p < 0.0001$, ↑↑↑: $p < 0.001$, ↑↑: $p < 0.01$, ↑: $p < 0.05$.

3.2 Argument Features

Now we examine the features of the argument specifically, rather than examining its interaction/similarity to the OP. These can give insight into specific linguistic cues that are aligned or indicative of successful persuasion as well as characteristics of persuasive arguments.

Following the process in the original paper, in our preprocessing we replace all quotes and links present in the argument with special tokens.

We break up the features into different categories. The first category is a simple measure of the word count of the argument. At first glance, it seems intuitive that persuasive arguments are generally longer,

and we will endeavor to investigate this. To continue with the original paper's breakdown, we investigate word-category related features, word-score related features, and finally entire-text related features.

Disclaimer: The original paper has a robust selection of word-category related and entire-text based features, and in this toy study we will only consider a handful of them.

Word-category based features: We measure the number of occurrences of certain word types in the argument. These include: indefinite and definite articles, first-person pronouns, singular and plural, second-person pronouns, links, quotes, and questions.

Word-score based features: We adopt the same metrics used in the original paper. Please refer to the original paper for the sources of these measures. We measure the following word-score features:

- Valence, which measures how pleasant the word's denotation is (high: joyous, rainbow, low: pandemonium, murder)
- Arousal, which measures the intensity of a word (high: ecstasy, low: mundane)
- Dominance, which measures the extent of control within a word between vulnerability/weakness to power/strength (high: strength, low: crippled)
- Concreteness, which measures the perceptibility/tangibility of a word (high: bus, low: virtue)

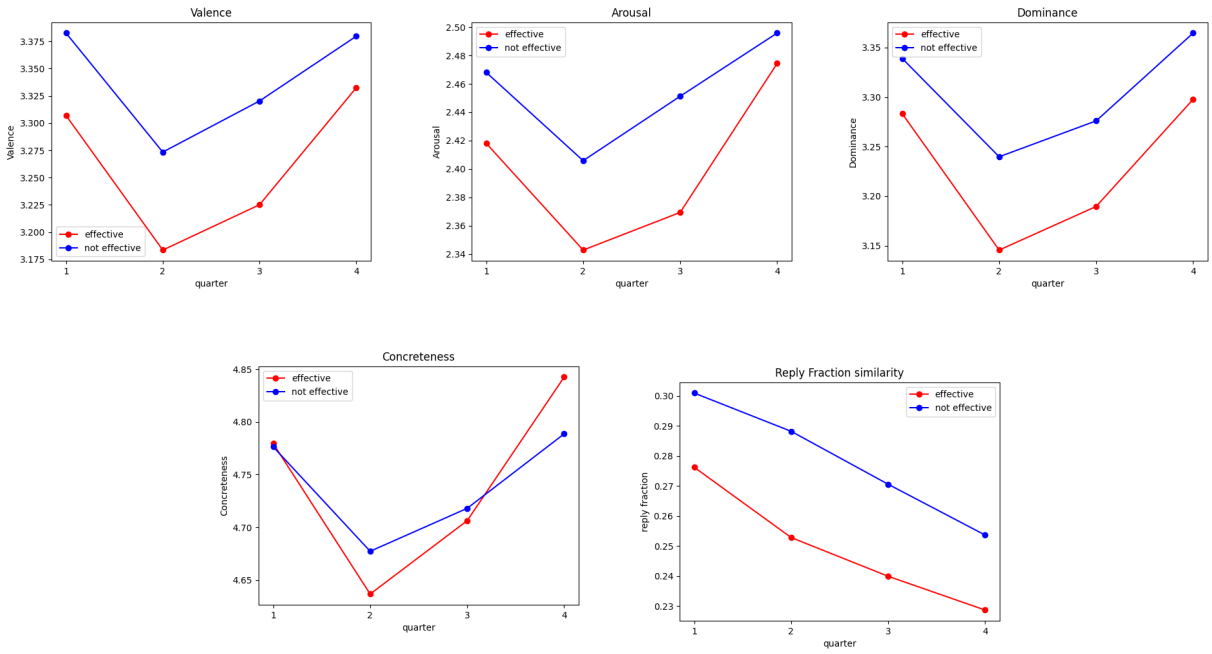
Entire-text based features: We then measure higher-level features that measure certain aspects of the entire argument. These include: number of sentences, number of paragraphs, and the Flesch-Kincaid readability score of the argument.

Table 2 lists out the significant features for each of these categories of features that passed a paired t-test with Bonferroni correction. We can see from Table 2 that most of the word-category based features are

Feature Name	root reply	full path
# words	↑↑↑↑	↑↑↑↑
Word-category features		
# definite articles	↑↑↑↑	↑↑↑↑
# indefinite articles	↑↑↑↑	↑↑↑↑
# 1st person plural pronouns	↑↑↑↑	↑↑↑↑
# of links	↑↑↑	↑↑↑↑
# 1st person pronouns	↑↑↑	↑↑↑↑
# 2nd person pronouns	↑	↑↑↑↑
# of quotes		↑↑
Word-score features		
Valence	↓↓	↓↓
Dominance	↓↓	↓↓
Arousal		↓
Entire-text features		
# of sentences	↑↑↑↑	↑↑↑↑
Flesch-Kincaid Readability	↑↑↑↑	↑↑↑↑
# of paragraphs	↑↑↑	↑↑↑↑

Table 2: Significant features for arguments, through a paired t-test with Bonferroni correction. Features area ranked on their average score across the root reply and full path cases. ↑↑↑↑: $p < 0.0001$, ↑↑↑: $p < 0.001$, ↑↑: $p < 0.01$, ↑: $p < 0.05$.

significant, which are most indicative and identifiable to an OP reading an argument. Additionally, these, alongside the number of words, are the most significant features, which is sensible due to their more noticeable presence within the argument. Additionally we see that persuasive arguments tend have lower values in the word-score metrics, which implies that less emotional and lower sentiment words are more often found within persuasive arguments.



Graphs that measure the word-score features, and the reply fraction at each quarter of the root reply.

3.3 Quarter-by-Quarter Analysis

Next, we consider the evolution of an argument as it starts and finishes. We quarter each reply and measure feature scores in each quarter, both argument-focused and interplay-focused. We see that effective arguments tend to have lower word-score features at all points of the argument, yet following a similar structure of dipping and then resurfacing to that level. Except in Concreteness, where we can see that effective arguments dip lower than not effective arguments before bookmarking and rebounding to a higher level. The consistent shape of each metric's graph can be explained with the bulk of the argument's opinions and reasonings being in the middle portion, where these word scores would tend to be lower. The rebound of Concreteness at the end can be explained by effective arguments bookmarking and ending on a tangible or actionable conclusion that helps sway an OP.

Next, we evaluate a singificant interplay feature for effective and not effective arguments at each quarter. Following the original paper, we choose to evaluate the reply fraction similarity between the root reply and the original post, which exhibits the same shape but corroborates our finding that effective persuasive arguments tend to be less similar than the OP's post.

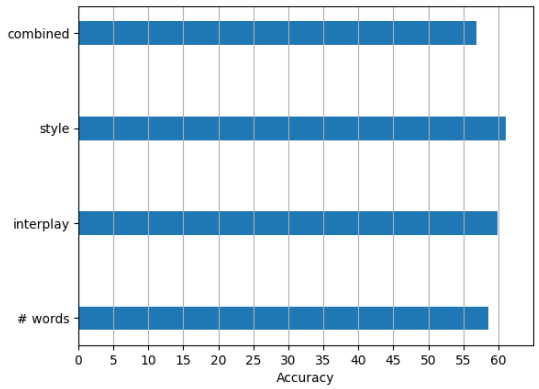
3.4 Prediction Results

Following the original paper, we train logistic regression predictive models with l_1 regularization with 5 cross-validation folds, ensuring that positive and negative argument entries with the same OP are in the same fold, and standardize each feature to unit variance.

We also separate our features into distinct feature sets. We first use only the word length as a baseline. The other feature sets are the significant interplay features as presented in Table 1, and the significant argument features as presented in Table 2, which we denote as style features. Then, our last feature set is the combination of all of these features. Please refer to Figure 5 for our prediction results.

We see that just like in the original paper, *# words* is a strong baseline for predicting, resulting in around 59% accuracy in the root reply case, and around 62% accuracy in the full path case.

From the results as listed on Figure 5, we see that interplay and style are both important and informative. However, in our sample case, style seems to have a more impactful effect on prediction than interplay, which is in contrast to the original paper which saw more improvements with interplay. Additionally, we see that



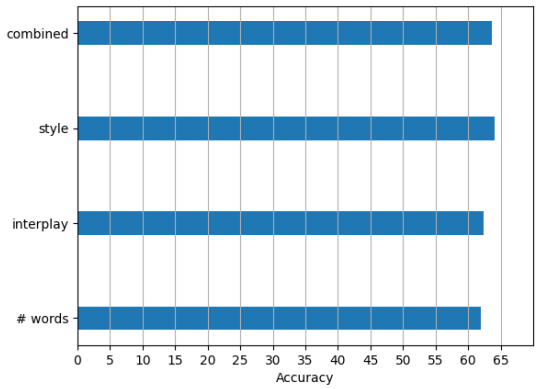


Figure 5a (left): Graph of accuracy results for the root reply case.
Figure 5b (right): Graph of accuracy results for the full path case.

the predictions on the combined features perform slightly worse than just style or just interplay, especially so in the root reply case.

4 Openness and Resistance to Persuasion

Our next analysis task is regarding the malleability and resistance of OP's changing their opinion. Following the paper, we use our argument features as outlined in the original paper to evaluate feature scores for our random sample of 500 OP posts. However, when conducting our two-sample t-tests between the malleable and resistant OPs, we did not find any of the features to be significant after Bonferroni correction. However, we will list some of the top features ordered by their p-value as well as their general direction.

# of links	$\downarrow$
Flesch-Kincaid readability	$\uparrow$
# of sentences	$\uparrow$
Dominance	$\uparrow$
# of first person pronouns	$\uparrow$
# of first person plural pronouns	$\downarrow$

These features, while not significant in the sample case perhaps due to the limited sample size, have directions consistent with the findings of the original paper. For example, the number of links in an OP's post may signify the conviction and evidence supporting the OP's perspective, thus perhaps making them less open to having their mind changed due to the presence of evidence.

4.1 GPT-3.5

Because of our lack of any significant features in the sample for the malleability and resistance of the OP, we utilize GPT-3.5 API to make predictions about whether or not a user will change their mind given their opinion.

Due to cost, resource, and time reasons, the initial objective of fine-tuning either GPT-3.5 or another pre-trained LLM such as BeRT or LLaMa was not feasible.

We ask GPT-3.5 to generate a prediction about whether or not an OP will change their mind by feeding in the prompt:

> Please predict if the user who wrote the following opinion had their opinion changed.
>
> **USER POST HERE**
>
> Prediction:

with the system message that it is trying to predict whether or not a user will change their mind about a topic. We remove any links present in the user post as part of preprocessing before feeding it in to the prompt and remove the mandated `r/ChangeMyView` footer. We have GPT-3.5 predict on both our training sample of 500 entries, as well as the holdout sample of 71 entries.

4.2 GPT-3.5 results

GPT-3.5 achieved an accuracy of 60% on the training sample, but only an accuracy of roughly 46% on the holdout sample. When asked to explain its predictions, GPT-3.5 offered some simple reasons for its prediction, of which we will list sample predictions that it got wrong and right.

Wrong prediction (result is true):

> Based on the user's opinion, it is unlikely that their opinion will change. They express a clear understanding of different perceptions of equality and acknowledge the evolution of feminism beyond just wanting equality. They also mention feeling off-put by certain aspects of feminism and the lack of misogyny-free communities for discussing men's issues. These factors suggest that the user has a well-formed opinion and is unlikely to change it.

Correct prediction (result is true):

> Based on the user's initial opinion, it is possible that their opinion could be changed. However, it would require a persuasive argument to convince them to be more accepting of flamboyant gay people.

Wrong prediction (result is false):

> Based on the user's statement, it is possible that their opinion could be changed. They express openness to hearing other points of view and acknowledge their own ignorance on racial issues. Therefore, there is a chance that they may be receptive to new information and perspectives that could potentially change their opinion.

Correct prediction (result is false):

> Based on the strong language and negative tone towards Reddit's views on women, sexuality, and race, it is unlikely that the user's opinion will be changed. They seem to have a firm belief that Reddit is being hypocritical in their treatment of Kim Kardashian and other women.

There are also several examples where GPT-3.5 was unable to predict whether or not the user had their opinion changed and supplied an answer that described why it found it difficult to predict. For example:

No prediction (result is true):

> Based on the given opinion, it is difficult to determine whether the user's opinion will change. However, it is possible that the user may have their opinion changed if they are presented with compelling arguments or evidence that supports the benefits or safety of lane splitting for motorcycles.

No prediction (result is false):

> Based on the information provided, it is difficult to determine if the user's opinion has changed. The user presents a detailed analysis of President Bush's decisions and policies, highlighting both positive and negative aspects. However, without additional information about the user's initial opinion and any potential shifts in their perspective, it is not possible to make an accurate prediction.

We see from GPT-3.5's analyses that it has a basic knowledge of what constitutes an attitude that presents itself as malleable vs resistant to persuastion, as it identifies characteristics such as tone, as well as evidence-based supporting within the post. However, it is unable to understand deeper factors or any specific language-based characteristics.

5 Further Work and Ideas

This topic of understanding persuasive arguments is evidently a rich and robust question that still requires understanding and analysis. Due to time and lack of resources, the original objective of finetuning GPT-3.5 or another pre-trained LLM was not feasible to be completed, but given such resources and time, is a direction that could yield fruitful results. We see from GPT-3.5's explanations that there is potential for it to understand deeper connections behind persuasion and malleability, but without fine-tuning for this specific task, will remain at this level of understanding.

Another interesting area of exploration is to examine whether or not these qualities discussed in this analysis, such as persuasive arguments, resistance to persuasion, etc. are different across different topics and conversations. For example, is the way an OP is persuaded on a political issue different than on some more mundane topic? How is this reflected in their approaches? Of course, this question is wildly broad and not currently answer-able without controlling for the people who post about their opinions, but is still nonetheless an interesting question.

Additionally, the sample and dataset for this project was derived from `r/ChangeMyView` from 2013-2015. An interesting new idea would be to endeavor to understand characteristics and qualities such as persuasive arguments, as well as malleability/resistance to persuasion across many different time periods, especially considering the seeming increase in extreme political opinions, as well as social media usage and activity since then. `r/ChangeMyView` remains today an active community with several posts uploaded daily, and the changes in discussion-worthy topics, writing mannerisms, interactions/conversations are all factors that have possibly changed significantly since 2013 to 2015. The aim of understanding how these qualities that compromise persuasive arguments or resistance/openness to persuasion across different time periods can provide a better examination of how language and social dynamics have changed, as well as their potential reasons and causes.