# Modeling Constraints Can Identify Winning Arguments in Multi-Party Interactions (Student Abstract)

**Suzanna Sia,** * [1] **Kokil Jaidka,** [2] **Niyati Chayya**[3] and **Kevin Duh** [1]

[1]Johns Hopkins University
[2] National University of Singapore
[3] Adobe Research

## Abstract

In contexts where debate and deliberation is the norm, participants are regularly presented with new information that conflicts with their original beliefs. When required to update their beliefs (belief alignment), they may choose arguments that align with their worldview (confirmation bias). We test this and competing hypotheses in a constraint-based modeling approach to predict the winning arguments in multi-party interactions in the Reddit ChangeMyView dataset. We impose structural constraints that reflect competing hypotheses on a hierarchical generative Variational Auto-encoder. Our findings suggest that when arguments are further from the initial belief state of the target, they are more likely to succeed.

## Introduction

Individuals are often exposed to information that conflicts with their beliefs, which may result in them experiencing cognitive dissonance (Festinger, Riecken, and Schachter 2017). In some cases, the dissonance works in the favor of the Commenter (C) providing new information which can succeed in changing the view of the Opinion Holder (O). Based on evidence from three different online experiments "when people are exposed to information, they update their views in the expected or 'correct' direction, on average" (Guess and Coppock 2020).

On the other hand, individuals may choose belief confirmation. The exposure to conflicting information may cause them to seek out and favor supporting arguments while rejecting contrary information (Festinger, Riecken, and Schachter 2017), leading to heightened opinion and affective polarization (Bail et al. 2018). Which paradigm better describes the norms of online and offline debates?

This work aims to ground the computational linguistic analysis of Reddit discussions using modeling constraints based on the cognitive dissonance theory. Would opinion holders be persuaded by arguments that present new and conflicting information, or by those that build on their existing beliefs? Our preliminary experiments address these questions. We make the following contributions:

- We introduce *distance-based structural modeling constraints* to test hypotheses within the confirmation bias paradigm of how individuals react to new information.
- We find that in an online forum, winning arguments are farther away from the user's initial belief, indicating that people are open to change when the argument presents new information.

## Problem Formulation

We denote the Opinion Holder's and Commenter's (text) arguments as $X^{\mathsf{O}}$ and $X^{\mathsf{C}}$, and the latent beliefs modelled with hidden vectors as $Z^{\mathsf{O}}$ and $Z^{\mathsf{C}}$. The goal is to predict whether the Opinion Holder O has been persuaded by the Commenter C. In the "Change My View" (CMV) subreddit, we indicate successful comments with a $\Delta$ and non-successful comments with $\emptyset$. Similarly, we adopt $\Delta$ for the winning team and $\emptyset$ for the losing team in debates. We model the sentences, labels and latent belief states of the participants jointly under a hierarchical generative framework.

### Modeling Approach

A hierarchical generative model is applied to model constraints on the latent belief states of O as $Z^{\mathsf{O}}$, and C as $Z^{\mathsf{C}}$. They generate the observed content, $X^{\mathsf{O}}$ and $X^{\mathsf{C}}$ respectively. $X^{\mathsf{O}} = [x_1^{\mathsf{O}}, \cdots, x_n^{\mathsf{O}}]$ denotes O's post with $n$ sentences. Constraints on the latent belief states enable us to investigate the following research question: **Are winning arguments closer to or farther away from O's original belief?** The 'hierarchical' formulation comes from aggregating each belief state from the observed sentences that belong to a single thread. Within each main thread, there can be multiple C trying to obtain a $\Delta$ from the O. Given the observed sentences, the first step is to find $p(Z|X)$: the posterior over the latent belief states. Subsequently, we have learned a model $f(Z^{\mathsf{O}}, Z^{\mathsf{C}}) \rightarrow \{\Delta, \emptyset\}$.

Argumentation hypotheses ($h_1$ to $h_5$) are modeled using constraints. Each constraint tests the relationships between the 'anchor' $Z^{\mathsf{O}}$, and $Z^{\Delta}$, $Z^{\emptyset}$. For example, in Table 1, $h_1$ tests if the distance between $Z^{\mathsf{O}}$ and $Z^{\Delta}$ is greater than $Z^{\mathsf{O}}$

*Corresponding author ssia1@jhu.edu, Malone Hall, Suite 340 3400 North Charles Street Baltimore, MD 21218

| | |
|---|---|
| $h_1$ | **Alternate hypothesis**; Successful arguments are "far" from the original opinion ○. |
| $h_2$ | **Confirmation bias**; Successful arguments are "close" to the original opinion ○. |
| $h_{3,4,5}$ | Successful arguments are not irrelevant ($h_3$), AND far ($h_4$) OR close ($h_5$) to ○. |

Table 1: Competing hypotheses

and $Z^{\emptyset}$. This is operationalised as $\mathcal{L}_{h_1}^{\text{bel}}$, where $\alpha_b$ represents the margin of loss: $\mathcal{L}_{h_1} = || Z^{\bigcirc} - Z^{\emptyset} ||_2^2 - || Z^{\bigcirc} - Z^{\Delta} ||_2^2$ , $\mathcal{L}_{h_1}^{\text{bel}} = \max(\mathcal{L}_{h_1} + \alpha_{\text{b}}, 0)$.

If a particular model constraint/assumption results in a better performance for the downstream $\Delta$ prediction task, then it offers support for that argumentation hypothesis.

### Argumentation Hypotheses

There is a lack of consensus on whether ○ would adjust their beliefs when presented with new information. A (dis)confirmation bias would imply that ○ favor arguments which somewhat align with their own beliefs (Bail et al. 2018). ○ might subconsciously penalize ⊂ who provide new information, and so more likely to conflict with their worldview. We devise the following modeling constraints (operationalised in Table 1). to test competing hypotheses. Note that only one of $h_1$ or $h_2$ may be true at a time.

## Experiments

**CMV Dataset:** We have used the CMV dataset processed by Jo et al. (2018) to 'in-domain' (ID) and 'cross-domain' (CD) topics with respect to their training split. [1] We truncated each sentence to 100 tokens and removed sentences with less than five words to reduce length effects.

**Model Settings:** We adopted a 2 hidden layer RNN-LSTM with 128 latent dimensions, and 256 hidden dimensions. We applied 0.4 word dropout for the decoder, and cyclic annealing of the KL loss against a standard variational prior of normal distribution with mean 0 and Identity Covariance $\mathcal{N}(0, I)$. We used 40000 vocabulary size, and set ranking margin $\alpha_{\text{m}}$ to 0.5. The contrastive margin, $\alpha_{\text{b}}$ was set to 0.01.[2] We used the Adam Optimizer with 0.001 as the initial learning rate, weight decay 0.0001, and enabled re-training of the GloVe embeddings. Training stopped after 10 epochs if the validation AUC of the last 5 epochs fell continuously.

### Results and Discussion

As a sanity check, we first benchmarked our modeling approach with $h_0$ (no constraints) against previous work on CMV and obtained comparable results with the state of the art in the In-Domain (AUC = 70.3 vs 70.5) for Attention-Interactive model (Jo et al. 2018), and Cross-Domain settings (AUC = 68.6 vs 69.7) for Pre-trained BERT Sentence Encoder (Devlin et al. 2019).

| | $h_0$ | $h1$ | $h2$ | $h3$ | $h4$ | $h5$ |
|---|---|---|---|---|---|---|
| ID | 70.3 | 69.9 | 70.6 | 69.2 | 69.7 | 68.3 |
| CD | 68.6 | 68.6 | 68.8 | 68.3 | **69.7**$^*$ | 68.4 |

Table 2: Comparison between models applying different hypotheses ($h_0 - h_5$ from Table 1) for predicting the winning arguments. $*p < 0.05$ for t-test against null hypothesis $h_0$.

In Table 2, we have reported results for testing hypotheses $h_1$ to $h_5$.

We observe no significant differences for ID and significant difference for $h_4$(AUC = 69.7) compared to $h_0$. This suggests that firstly, it is more appropriate to model $\Delta$ arguments as dissimilar from the original opinion ($h_1$). Additionally, constraining both winning and non-winning arguments to be closer to the original opinion than irrelevant comments improved representation and predictive performance.

## Conclusion

A hierarchical generative model was applied to model multiparty interactions in arguments. We are replicating our experiments on a face-to-face setting. Early findings suggest that in platforms intended for debate and deliberation, arguments grounded in new information are more persuasive.

Our framework offers promising directions for building an interdisciplinary understanding of argumentation and persuasion. Future research could examine how this approach would generalize to other paradigms invoking cognitive dissonance, such as exposure to misinformation.

## References


Bail, C. A.; Argyle, L. P.; Brown, T. W.; Bumpus, J. P.; Chen, H.; Hunzaker, M. F.; Lee, J.; Mann, M.; Merhout, F.; and Volfovsky, A. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* 115(37):9216–9221.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*.

Festinger, L.; Riecken, H.; and Schachter, S. 2017. *When prophecy fails: A social and psychological study of a modern group that predicted the destruction of the world*. Lulu Press, Inc.

Guess, A., and Coppock, A. 2020. Does counter-attitudinal information cause backlash? results from three large survey experiments. *British Journal of Political Science* 50(4):1497–1515.

Jo, Y.; Poddar, S.; Jeon, B.; Shen, Q.; Rosé, C.; and Neubig, G. 2018. Attentive interaction model: Modeling changes in view in argumentation. In *Proceedings of the ACL*, 103–116. Association for Computational Linguistics.


[1] ID refers to training and testing within the same high-level topics such as Sports, while CD refers to training and testing with different topics. Please refer to Jo et al. (2018) for details.

[2] This is the hyperparameter used in the loss functions for our "hypothesis testing", and we selected the best $\alpha_{\text{b}}$ from 0.001, 0.005, 0.01, 0.05 and 0.1.