Offer a Different Perspective: Modeling the Belief Alignment of Arguments in Multi-party Debates

Suzanna Sia* Johns Hopkins University ssia1@jhu.edu

Kokil Jaidka* National University of Singapore jaidka@nus.edu.sg

Hansin Ahuja IIT Ropar hansinahuja@gmail.com

Niyati Chhaya Adobe Research India nchhaya@adobe.com

Kevin Duh Johns Hopkins University kevinduh@cs.jhu.edu

Abstract

In contexts where debate and deliberation are the norm, the 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 Change My View and Intelligence Squared debates datasets. We adopt a hierarchical generative Variational Autoencoder as our model and impose structural constraints that reflect competing hypotheses about the nature of argumentation. Our findings suggest that in most settings, predictive models that anticipate winning arguments to be further from the initial argument of the opinion holder are more likely to succeed.

1 Introduction

On social media, individuals are often exposed to information that conflicts with their beliefs, which may result in them experiencing cognitive dissonance (Festinger, 1962; Festinger et al., 2017). In the context of multi-party online debates, the Commenter (C) tries to change the Opinion Holder's (O) point of view. We may observe a confirmation bias, when exposed to information that conflicts with their beliefs, people may seek out and favor supporting arguments that are closest to their own beliefs (Taber and Lodge, 2006; Festinger et al., 2017), leading to polarized online spaces (Bail et al., 2018). However, there is also evidence to support that novel information could work in favor of changing the O's view. Based on evidence from three different online experiments, Guess and Coppock (2020) reported that "when people are exposed to (new) information, they update their views in the expected or 'correct' direction, on average."

Which paradigm better describes the norms of online and offline debates? Work on modeling persuasion in online forums has focused on identifying debate winners (Potash and Rumshisky, 2017; Zhang et al., 2016; Wang et al., 2017; Prabhakaran et al., 2013) and winning negotiation games (Keizer et al., 2017). Most of this work focuses on engineering or learning features to predict argument success. Previous work has suggested that to be persuasive, the commenter (C) should exert influence (Hidey and McKeown, 2018), and perform effective social interaction (Wei et al., 2016; Jo et al., 2018) and language interplay (Tan et al., 2016). Many of these studies suggest that interplay with the original opinion holder (O) would support a competing argument. However, it is not clear whether in this interaction, C should try to reflect O's views and build arguments from there, or try to directly bring in 'new' counter-arguments. In this study, we explore whether implementing the interplay of O and C via modeling constraints in Variational Autoencoders can identify winning counter-arguments. Our contributions are:

- We demonstrate how neural network architectures can be applied to test competing social science hypotheses related to how opinion holders would react to new information.
- We introduce *distance-based structural modeling constraints* into the feature learners for Variational Auto Encoding architectures, which operationalize our hypotheses.
- We show that predictive models that expect the winning counter-arguments to be further away from the opinion holder's initial argument were more likely to have better performance.

Our study addresses this puzzle through a computational linguistic analysis of Reddit Change My View discussions. We also test the generalizability

*Suzanna Sia & Kokil Jaidka co-lead this work.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11939 - 11950
December 7-11, 2022

of our framework to a second dataset (IQ2 Debates).

2 Related Work

The state of the art in the space of argument modeling typically focuses on data representation. For instance, Padó et al. (2019) propose the construction of "discourse networks" using news coverage of political debates to analyze how discursive elements influence policy making. Sawhney et al. (2020) utilize semantic language representations and dynamics between debate transcripts, topics and interlocutors. Our purpose is to reformulate the problem in terms of the adversarial nature of arguments posed during the debate.

On the other hand, computational methods to study the idea of confirmation bias and the proclivity to change one's opinion have been explored with a focus on author and text attributes (Workman, 2018; Thornhill et al., 2019; Mensah et al., 2019). Extensive study has also been done with the primary dataset that we use for experimentation – the Change My View (CMV) dataset (Tan et al., 2016). Öcal et al. (2021) investigate people's reasoning behaviour in online forums and Dayter and Messerli (2021) discuss the degree of formality in persuasive speech. These approaches mainly focus on characterizing the linguistic properties of convincing arguments (Dayter and Messerli, 2021) or the characteristics of an influential or susceptible opinion holder (Tan et al., 2016). Instead, our study suggests that the latent spatial projection representing winning arguments are best interpreted in relation to the their distance from the original opinion holder's comment, as well as from other losing or irrelevant arguments.

Our work focuses on the inter-relation of the arguments to understand persuasion in debates. Work that follows similar intuition as ours has proposed an attention mechanism that identifies the more malleable sections of the opinion holder's speech and interaction encodings which establish relationships between the opinion holder and commenter's speech (Jo et al., 2018). Instead, our hierarchical Variational Autoencoder (VAE) model characterizes the commenter's relative alignment (or misalignment) with the opinion holder.

Chen and Yang (2021) propose a weakly-supervised hierarchical latent variable model that utilises broader persuasiveness of textual requests along with limited sentence annotations to predict persuasion strategies on a sentence level. Their work discretizes these persuasion strategies along specific axes (commitment, reciprocity, etc.). Our work, on the other hand, focuses on the persuader and their alignment with that of the person being persuaded.

To summarize, although there are many different ways in which prior work has conceptualized persuasive arguments in online contexts, the contribution of this work is in proposing a new way to model conversation stance in adversarial paradigms, by representing any argument relative to each other. We benchmark our approach against transformers (Devlin et al., 2019; Conneau et al., 2017) and models inter-speech dynamics (Jo et al., 2018) reported in previous work.

3 Problem Formulation

Our problem formulation relies on the notion of beliefs, recently operationalized as the latent space projections of a semantic representation into lower dimensions for the purpose of representing and conveying certain ideas and concepts (Vu et al., 2022). Therefore, in our primary task, we pose the following assumptions on the data generation process, i.e., how CMV debates occur: (a) all arguments in a "thread of conversation" are co-dependent for context, and constitute a "belief"; (b) irrelevant beliefs differ in topic but not intent, so constitute arguments on other topics; (c) the winning argument or belief is independent of others and is decided by the original poster $\mathtt{O}$, and (d) all threads are independent of each other. We discuss these assumptions again and their possible relaxation in our generalizability test.

Let the latent states of Opinion Holders ($\mathtt{O}$) and Commenters ($\mathtt{C}$) be depicted by the latent spaces occupied by the text of their arguments. Their beliefs are thus the hidden vectors in a hierarchical Variational Autoencoder (VAE). That is, we denote the Opinion Holder's and Commenter's text instances $X^{\mathtt{O}}$ and $X^{\mathtt{C}}$, and the beliefs modelled with hidden vectors (Figure 1) as $Z^{\mathtt{O}}$ and $Z^{\mathtt{C}}$.

The goal is to predict whether the $\mathtt{O}$ has been persuaded by $\mathtt{C}$. In the "Change My View" (CMV) subreddit, we indicate successful counter-arguments with a Δ and non-successful counter-arguments with $\emptyset$. We abuse notation for convenience but without loss of generality, and refer to both the opinion holder, as well as the *opinion holder's opinions* as $\mathtt{O}$. The notation's position as a superscript indicates

Hypothesis	Model assumption	Explanation
Main hypothesis and counter-hypothesis: Winning arguments Δ are close to or far from original post $\circ$ than the losing arguments $\emptyset$		
h_1	$d(Z, Z^{\emptyset}) \geq d(Z, Z^{\Delta})$	**Confirmation bias**: Δ are closer to $\circ$ than $\emptyset$
h_2	$d(Z, Z^{\Delta}) \geq d(Z, Z^{\emptyset})$	**Alternate hypothesis**: Δ are further from $\circ$ than $\emptyset$
Additional scoping constraints: Irrelevant arguments *Irr* are further from $\circ$ than Δ and $\emptyset$ AND adhere to h_1 OR h_2		
h_3	$d(Z, Z^{irr}) \geq [d(Z, Z^{\Delta}), d(Z, Z^{\emptyset})]$	Δ and $\emptyset$ are far from *Irr*
$h_4 = h_1 + h_3$	$d(Z, Z^{irr}) \geq d(Z, Z^{\emptyset}) \geq d(Z, Z^{\Delta})$	Δ are closer to $\circ$ than $\emptyset$ AND Δ are far from *Irr*
$h_5 = h_2 + h_3$	$d(Z, Z^{irr}) \geq d(Z, Z^{\Delta}) \geq d(Z, Z^{\emptyset})$	Δ are further from $\circ$ than $\emptyset$ AND both are closer to $\circ$ than *Irr*

Table 1: Various hypotheses tested through imposing constraints on the model. d is a distance metric; which we adopt $L2$ loss in our experiments.

whether we are referring to this as a text instance (X) or with hidden vectors (Z).

We frame competing hypotheses to test whether Opinion Holders penalize or reward Commenters who conflict with their worldview. Table 1 reports the hypotheses to be tested:

- h_1 will test whether $\circ$ manifests a **confirmation bias**: successful arguments Δ are "closer" to the original opinion than the losing counter-argument $\emptyset$.
- h_2 will test whether $\circ$ is persuaded through **dissonance**: successful arguments Δ are further away from the original opinion than the losing counter-argument $\emptyset$.
- $h_{3,4,5}$ will add scoping constraints: successful arguments Δ are closer (h_4) OR further (h_5) to the original opinion than the losing counter-argument $\emptyset$ and the irrelevant argument *Irr* is always further away from the original opinion than Δ and $\emptyset$.

Although irrelevant counter-arguments are never directly fed into the Δ predictor, the presence of irrelevant counter-arguments serves as an additional distance constraint for the relevant counter-arguments so that they are not arbitrarily far away from the original post.

4 Modeling Approach

We adopt a hierarchical generative model to model constraints on the latent spaces depicting the beliefs of $\circ$ and $\subset$, denoted Z°, and $Z^{\subset}$ respectively. The hierarchical generative framework (Kingma et al., 2014; Serban et al., 2017) was applied to argument modeling by Chen and Yang (2021), but we extended the work by Yang et al. (2019) to consider multi-party interactions. This is because, within each main thread, there can be multiple commenters $\subset$ trying to obtain a Δ from the $\circ$.

In subsection 4.1, we describe the implementation of hypothesis-specific modeling constraints, which is the main contribution of this paper. In subsection 4.2, we explain the general form of posterior inference over the latent states Z given the observed content X, $p(Z|X)$, which is common to all hypotheses.

4.1 Modeling Hypotheses with Constraints

Argumentation hypotheses (h_1 to h_5) are modeled using constraints designed to test the relationships between the original post (Z°), and Z^{Δ}, $Z^{\emptyset}$, where Δ are arguments which have successfully changed $\circ$'s view, and $\emptyset$ are arguments which are unsuccessful. For example, in Table 1, h_1 tests if the distance between the latent spaces of Z° and Z^{Δ} is greater than Z° and $Z^{\emptyset}$. This is operationalized as $\mathcal{L}^{\text{dist}}_{h_1}$, using an $L2$ loss. α_b is a hyperparameter representing the margin of loss:

$$\mathcal{L}_{h_1} = || Z^{\circ} - Z^{\emptyset} ||_2^2 - || Z^{\circ} - Z^{\Delta} ||_2^2 \quad (1)$$

$$\mathcal{L}^{\text{dist}}_{h_1} = \max(\mathcal{L}_{h_1} + \alpha_{\text{b}}, 0) \quad (2)$$

The formulation of the loss function above is similar to a triplet loss (Hoffer and Ailon, 2015). However, unlike the triplet loss, we do not know in advance whether winning arguments are 'closer' or further away from the original post. By implementing different hypotheses on the relationships between $\circ$, Δ, $\emptyset$ counter-arguments (Table 1), we demonstrate a way to operationalise and thus test for these relationships. First, if the pairwise Euclidean distances satisfy the modeling assumptions,

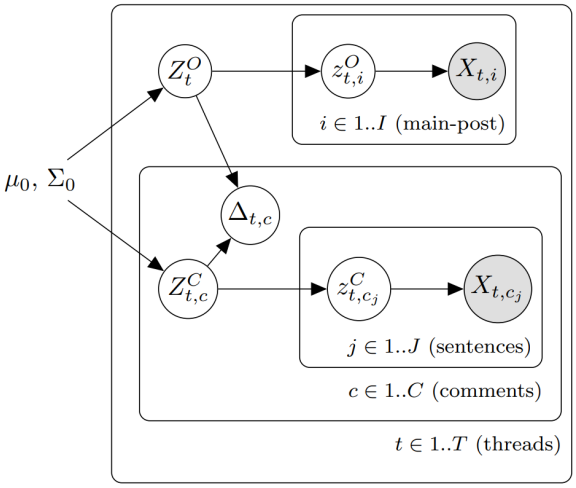

Figure 1: Probabilistic Graphical Model showing the hierarchical organization of arguments and counter-arguments inside the Reddit ChangeMyView forum. Multiple Counter-Argumenters (C) within a thread try to obtain a Δ from the Opinion Holder (O), signalling a change in view. Conditioned on the post level latent states, Z^{C} and Z^{O}, the model is trained to infer whether a Δ has been awarded to the Commenter.

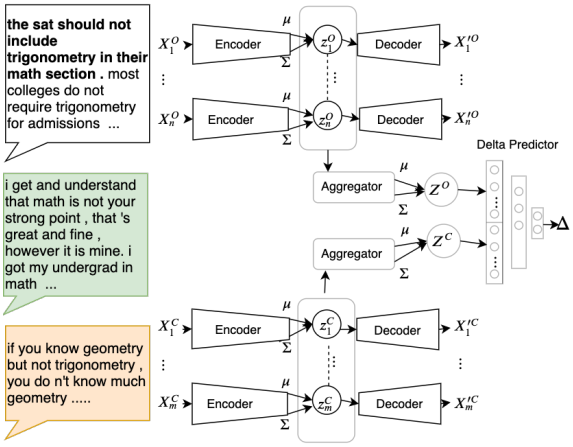

Figure 2: Inference network for approximating the posterior distribution over latent states for the Graphical Model (Figure 1). The top text is an example of the original argument x^{O}, and below that are examples of counter-arguments x^{C}. The middle text is an example of a winning (Δ) counter-argument and the bottom is an unsuccessful ($\emptyset$) counter-argument.

then it validates the modeling implementation. Second, if a particular model constraint results in a better performance for the downstream Δ prediction task, then it offers support for that argumentation hypothesis. Applying distance constraints to the latent states enable us to investigate the following research question: **Are winning arguments closer to or farther away from O's original post?**

4.2 Posterior Inference for $p(Z|X)$

Consider Figure 1. The latent states Z^{O} and Z^{C} that we apply distance constraints to in subsection 4.1 are responsible for generating the observed content, X^{O} and X^{C} respectively. $X^{\text{O}} = [x_1^{\text{O}}, \cdots, x_n^{\text{O}}]$ denotes O's post with n sentences, and $X^{\text{C}} = [x_1^{\text{C}}, \cdots, x_m^{\text{C}}]$ denotes C's post with m sentences. Given the observed sentences X^{O} and X^{C}, the first step is to find $p(Z|X)$: the posterior over the latent states. A 'good' representation of Z

- can generate the observed sequences X^{O} and X^{C}. Each observed sentence is generated conditioned on the latent state. E.g, x_1^{O} is generated conditioned on z_1^{O}.
- is useful for the subsequent binary classification task of predicting whether a counter-argument is successful with respect to the Original opinion $f(Z^{\text{O}}, Z^{\text{C}}) \rightarrow \{\Delta, \emptyset\}$.

A hierarchical model is a natural choice to aggregate each latent state from the observed sentences belonging to a single thread. i.e, thread level latent state representations are aggregated from sentence level, $g([z_1^{\text{O}}, \cdots, z_n^{\text{O}}]) \rightarrow Z^{\text{O}}$ and $g([z_1^{\text{C}}, z_2^{\text{C}}, \cdots, z_m^{\text{C}}]) \rightarrow Z^{\text{C}}$, where g is a recurrent model which takes variable number of sentences, such as an RNN-LSTM.[1]

4.3 Inference Network

Variational Autoencoders (VAE; Kingma and Welling (2014)) are a neural model and variational inference method that links autoencoders with mean field variational Bayes through non-linear approximations. The VAE outputs an approximation $q_\phi(z|x)$ by training the variational parameters ϕ mapping from x to a mean μ and (diagonal of) covariance σ, which are in turn parameters to a multivariate Gaussian. $(\mu, \sigma) = \text{NeuralNet}_\phi(x)$. $q(z|x) = \mathcal{N}(z; \mu, diag(\sigma))$.

As seen in Figure 2, the high-level overview of our approach is to apply VAEs to each sentence x and obtain a distribution over sentence-level latent states by approximating the posterior $p(z|x)$. We then sample from $p(z|x)$, a latent state z for each sentence, and aggregated this to a thread level Z using bidirectional RNN-LSTM to handle a varying number of sentences.[2] A Δ predictor for successful

[1] We aggregate latent state representations at the sentence level because it is not feasible to reconstruct long paragraphs. We adopt RNN-LSTM instead of the more modern Transformers, because argument datasets are not typically available on a large scale.

[2] We also experimented with other forms of aggregation such as CNNs, simple feedforward attention and Pairwise attention between C and O sentence level latent states, but find that these do not outperform the RNN.

(Δ) and unsuccessful $(\emptyset)$ arguments is then trained by concatenating Z^{O} and Z^{C} as input to a two-layer feedforward network (Figure 2).

VAE (Encoder-Decoder): We have experimented with various models for the Encoder (q_ϕ) and have reported results with a 2-layer RNN-LSTM (Hochreiter and Schmidhuber, 1997) for the Decoder (p_θ). We adopted the standard reconstruction error and the KL-Divergence loss from Kingma and Welling (2014) which includes all counter-arguments and original opinion as:

$$\mathcal{L}^{\text{VAE}} = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i^{\text{VO}} + \sum_{c=1}^{|C|}\frac{1}{m_c}\sum_{i=1}^{m_c}\mathcal{L}_i^{V_c} \quad (3)$$

$$\mathcal{L}_i^{V_O} = D(q_\phi(z|x_i) \,||\, p(z)) - \mathbb{E}_{q_\phi}[\log p_\theta(x_i|z_i)] \quad (4)$$

where n is the number of sentences in the main opinion, and m_c is the number of sentences in counter-argument c. $\mathcal{L}_i^{V_c}$ follows the same form as $\mathcal{L}_i^{V_O}$.

Δ Predictor: We train a classification head which takes in Z^{O} and Z^{C} as concatenated input $f(Z^{\text{O}}, Z^{\text{C}}) \rightarrow \{\Delta, \emptyset\}$. Since the encoder for mapping X to Z should be expressive, we only apply a simple 2-layer feedforward network as the classification head, and minimize the Binary Cross entropy Loss, referring to this as $\mathcal{L}^{\Delta}(Z^{\text{O}}, Z^{\text{C}})$.

As the data is skewed towards $\emptyset$ (Table 2) for Reddit, we adopt the margin ranking loss similar to Jo et al. (2018), $\mathcal{L}^{\text{Margin}}(Z^{\Delta}, Z^{\emptyset})$ to drive the predicted probability of Δ counter-arguments to be greater than $\emptyset$, where α_m is the margin threshold.

$$\mathcal{L}^{\text{Margin}}(Z^{\Delta}, Z^{\emptyset}) = \max(P(\Delta|Z^{\emptyset}, Z) - P(\Delta|Z^{\Delta}, Z) + \alpha_m, 0) \quad (5)$$

Note that for datasets which are balanced (such as the debate dataset), we do not apply a margin ranking loss on the final prediction.

4.4 Overall objective function

Combining the above loss functions, the overall objective function is

$$\mathcal{L}^{\text{thread}} = \mathcal{L}^{\text{VAE}} + \frac{1}{N_c}\sum_{j=1}^{N_c}(\mathcal{L}_j^{\Delta} + \mathcal{L}_{h,j}^{\text{dist}}) + \mathbb{E}_{c_1,c_2}[\mathcal{L}^{\text{Margin}}(Z^{c_1}, Z^{c_2})] \quad (6)$$

where we take one SGD step for each thread with mini-batch size equal to the number of thread counter-arguments, N_c. In this equation, we involve the hypothesized distances between the latent states from Equation (2) mentioned as $\mathcal{L}_{h,j}^{\text{dist}}$, the standard KL Divergence and reconstruction error $\mathcal{L}^{\text{VAE}}$ from Equation (3), the distance for the delta $\mathcal{L}_j^{\Delta}$ from Equation (4) and the margin loss $\mathcal{L}^{\text{Margin}}(Z^{c_1}, Z^{c_2})$ from Equation (5). The loss terms are rescaled to be in the same range, typical of multi-task learning.

5 Experiments

5.1 Datasets

Change My View (CMV) dataset We have used the Change My View (CMV) dataset (Tan et al., 2016) which is a subreddit[3] where Opinion Holders (O) post their views on various issues and challenge the community to try and change their view. O signals when their view has been changed by awarding a Δ to counter-arguments. The original dataset composes 18,363 discussions from January 1, 2013 - May 7, 2015, for training data and 2263 discussions from May 8-September 1, 2015, for test data. First, we did not consider threads where a Δ has not been awarded, as our modeling constraints and margin loss formulation involves contrasting successful and unsuccessful arguments *within* a thread. Second, we removed any text in the counter-arguments which was a quote from the original posts. Third, direct replies from the same commenter to any subthread following the original post are considered valid counterarguments. Fourth, we truncated each sentence to 100 tokens and removed sentences with less than five words to reduce length effects.

	Train	Val	Test(ID)	Test(CD)
Successful (Δ)	1705	202	485	1026
Unsuccessful $(\emptyset)$	38519	4599	6502	16965

Table 2: Statistics of CMV dataset for successful and unsuccessful comments that can change the Opinion Holder's view. In-domain (ID) and cross-domain (CD) splits obtained from Jo et al. (2018).

Intelligence Squared (IQ2) Debates dataset The Intelligence Squared (IQ2) Corpus[4] (Zhang et al., 2016; Wang et al., 2017) contains transcripts of debates annotated with the speaker, audience

[3] https://www.reddit.com/r/changemyview

[4] https://convokit.cornell.edu/documentation/iq2.html

pre- and post-vote, and final outcome, where an aggregate opinion change was calculated based on the number of votes before and after the debate. We adopted Δ for arguments by the winning team and $\emptyset$ for the arguments by the losing team. Wins and losses were calculated based on the change in the pre-and post-votes for a stance (of at least 0.25 standard deviations), yielding 45 debates with 225 argument pairs.

Irrelevant Arguments For the CMV dataset, irrelevant comments are randomly sampled from an unrelated discussion thread. For the IQ2 dataset, we relax the criterion for irrelevant arguments to allow them to be semantically uninformative, so they constitute a randomly sampled comment by the moderator within the same debate.

5.2 Model Settings

We finally adopted a two hidden layer RNN-LSTM with 128 latent dimensions, and 256 hidden dimensions. We applied 0.4 word dropout for the decoder and a standard variational prior of a Gaussian with mean 0 and diagonal covariance, $\mathcal{N}(0, I)$, similar to Bowman et al. (2016) which is the standard normal prior for the most general case. To avoid KL collapse, we applied cyclic annealing of the KL loss (Fu et al., 2019).

Following Jo et al. (2018), we used 40000 vocabulary size, and set ranking margin α_{m} to 0.5. The contrastive margin, α_{b} was set to 0.01. This is the hyperparameter used in the loss functions for our "hypothesis testing", and we selected the best α_{b} from 1e-4 to 1e-1. We used the Adam Optimizer (Kingma and Ba, 2015) with 1e-3 as the initial learning rate, weight decay 1e-4, and enabled re-training of the GloVe embeddings (Pennington et al., 2014). Training stopped after 10 epochs if the validation AUC of the last five epochs fell continuously.[5]

5.3 Evaluation Setup

Benchmarking experiments. First, we benchmarked the performance of our model architecture against baselines in the form of pre-trained and fine-tuned BERT transformers, as well as the state of the art by Jo et al. (2018) which also models Opinion Holder (O) and Commenter (C) relations. We test three versions of obtaining sentence embeddings for the hierarchical VAE model. We experiment with a BERT Sentence Encoder (Devlin et al., 2019) referred to as VAE-BERT, Infersent Encoder (Conneau et al., 2017) referred to as VAE-Infersent, and the two layer RNN (VAE-RNN).

Hypothesis testing and cross-domain validity. Next, we tested our hypotheses regarding O and C by evaluating the efficacy of the VAE architecture under different modeling constraints. Following the evaluation setup and the labels available from Jo et al. (2018), we separated the test split into subsets with 'in-domain' (ID) and 'cross-domain' (CD) data. Based on similarity scores, the in-domain subset comprised the test set with similar topics as the training set. The cross-domain subset comprised 13 dissimilar topics in the test set.

	CMV	
Model	**In-Domain**	**Cross-Domain**
BERT (Devlin et al., 2019)	69.2	68.3
AIM (Jo et al., 2018)	**70.5**	67.5
VAE-BERT	70.1	**69.7**
VAE-Infersent	68.6	67.9
VAE-RNN (This study)	70.3	68.6

Table 3: Benchmarking experiments: Predictive performance as Area Under the Curve (AUC). Our model (Hierarchical VAE with no modeling constraints) to predict the winning argument compares favorably to modern neural baselines even before we add modeling constraints.

Model validity. It is essential to evaluate whether our VAEs are indeed able to apply the modeling constraints and learn the distances between the arguments. Therefore, we tested the distance measurements generated from each model against its corresponding hypotheses. First, for each thread, we extracted the hidden states generated by the Aggregator of our model. We then verified whether the pairwise Euclidean distances between the original posts and the counter-arguments adhered to the modeling constraints that had been imposed (Table 1).

Generalizability. A generalizability test allows us to relax our strong primary assumptions about the nature of irrelevance (assumption b in Section 3) and the determiner of winning arguments (assumption c), and evaluate whether our model is still appropriate. In the IQ2 Debates dataset, within each debate, there are multiple speakers on each team. First, we consider O to be the debate team

[5]Code will be made available at https://github.com/suzyahyah/modeling_belief_alignment_arguments

	Modeling constraints in the VAE model	CMV - In-Domain	CMV - Cross-Domain	IQ2 Debates
h_0	No distance assumptions	70.3	68.6	53.0
h_1	Δ are close to $\circ$	**70.6**$\uparrow$	68.8$\uparrow$	45.2$^*\Downarrow$
h_2	Δ are far from $\circ$	69.9 $\downarrow$	68.6$\updownarrow$	**60.0**$^*\Uparrow$
	Additional scoping constraints: Irrelevant arguments *Irr* are further from $\circ$ than Δ and $\emptyset$ AND adhere to h1 OR h2			
h_3	Δ and $\emptyset$ are far from *Irr*	69.2$\downarrow$	68.3$\downarrow$	53.8$\uparrow$
$h_4 = h_1 + h_3$	Δ are closer to $\circ$ than $\emptyset$ AND Δ are far from *Irr*	68.3$\downarrow$	68.4$\downarrow$	58.4$^*\Uparrow$
$h_5 = h_2 + h_3$	Δ are further from $\circ$ than $\emptyset$ AND Δ are far from *Irr*	69.7$\downarrow$	**69.7**$^*\Uparrow$	55.5$^*\Uparrow$

Table 4: Area Under the Curve (AUC) of models applying different hypotheses for predicting the winning arguments on the test set. Bold text signifies the best-performing model for the setting. * and $\Uparrow$ signifies $p < 0.05$ in t-tests against h_0.

arguing on the side of the debate topic. Second, we made the simplifying assumption that all speakers on the winning team have provided the winning argument. Third, we only considered arguments from the speakers if they were more than three sentences long and truncated each paragraph to the first 100 tokens. Fourth, because of the issue of data sparsity, we generated five random train, test, and validation splits using 70%, 20%, 10% of the data and report results aggregated across the splits. Finally, we subset our data to debates with pre-vote $\rightarrow$ post-vote changes.

6 Results

Benchmarking against the state of the art. The benchmarking results in Table 3 offer a sanity check on the modeling approach with h_0 (no constraints) against previous work on CMV. The VAE-RNN model performs at par 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 the VAE with pre-trained BERT Sentence Encoder (Devlin et al., 2019). We opt to use the VAE-RNN model in subsequent experiments as we are able to train this from scratch to remove any confounds of the pretrained model when testing the different hypothesis (h_1 to h_5).

Hypothesis testing and cross-domain validity. In Table 4, we have reported results for testing hypotheses h_1 to h_5. Asterisks indicate that the results were better than the model with no modeling assumptions ($p < 0.05$), based on repeated runs and a pairwise t-test of the model AUCs.

Modeling constraints improve predictive performance across all three contexts, by as little as 0.3% AUC in the in-domain online setting, and as much as 1.1% in the cross-domain setting. In in-domain validation, models that constrain latent space representations of winning arguments to be close to those of the initial argument than losing arguments outperform all others (AUC = 70.6 vs 70.3 in the no-constraints setup). On the other hand, in the cross-domain setting, modeling latent space representations of Δ arguments as further from those of the original opinion, and incorporating additional hypotheses about latent space representations of irrelevant comments as being further from those of the real arguments offers a significant predictive advantage (AUC = 69.7 vs 68.6, $p < 0.05$).

Generalizability. When we replicate our analysis on the IQ2 dataset, the h_2 model showed a statistically significant improvement of 7% over h_0 (AUC = 60.0 vs 53.0, $p < 0.05$). Once again, we observe that arguments with novel information were more likely to win. Modeling the irrelevant counter-arguments (h_3 and h_5) did not provide a substantial advantage (AUC = 53.8).

Model validity. We calculated the pairwise Euclidean distances between the hidden states corresponding to the original post on the one hand, and the Δ, $\emptyset$ and *Irr* on the other. Figure 4 reports the answer when we input the distances back into the modeling assumptions for each VAE variant. This figure gives us confidence that any performance difference we are observing in Table 4 is indeed because of the modeling constraints we specified, as in all cases except h_1, the assumptions are held up by the actual data.

7 Analysis

7.1 Visualizing the latent spaces

In Figure 3, we visually demonstrate how modeling constraints facilitate the cleavage of the irrelevant counter-arguments from losing and winning

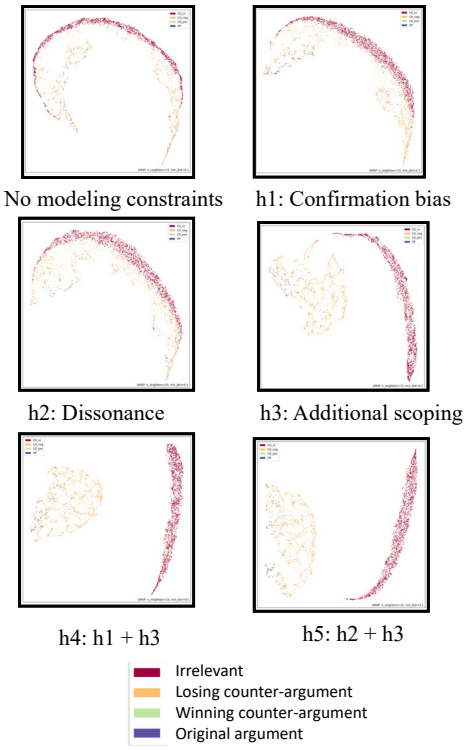

Figure 3: Visualization of latent space using UMAP

counter-arguments.

We applied Uniform Manifold Approximation and Projection (UMAP, McInnes et al. (2018)) [6] to represent the validation dataset's global structure in terms of a 2-dimensional projection of the clusters of original posts and each group of counter-arguments. UMAP is often preferred over alternatives such as t-Distributed Stochastic Neighbor Embedding (t-SNE) for visualizing hidden states (Van der Maaten and Hinton, 2008). Finally, we obtained the visualizations reported in Figure 3. Proceeding from left to right, distance between the latent spaces of the irrelevant counter-arguments (red dots) and those of the losing and winning counter-arguments (the orange and green dots) appears to increase, with the cleavage appearing to be larger in h_5 as compared to h_3 and h_4.

In the second image for h_1 as compared to the third image for h_2, the green dots corresponding to the winning counter-arguments are closer to the blue dots corresponding to the ○. Therefore, the visualization coheres with our expectations based on the modeling constraints.

[6] https://umap-learn.readthedocs.io/en/latest/

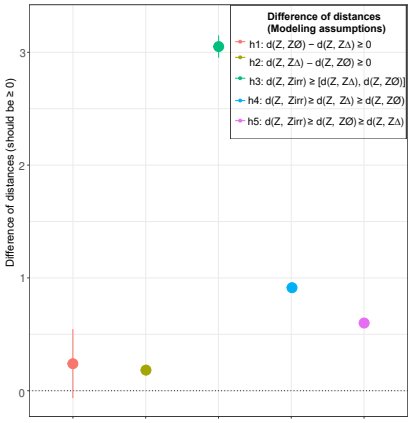

Figure 4: Validating the modeling assumptions with corresponding pairwise distance data

7.2 Error Analysis

A few illustrative examples of the errors encountered under various specifications in the Cross-Domain dataset are reported in Table 5, together with the pairwise cosine similarity between the latent spaces representing the arguments and the original post. First and foremost, the green checkmarks suggest the benefits of using a distance-based approach to model the relationship between ○ and ⊂ arguments as they indicate where modeling constraints based on the hypothesis h_5 successfully modeled the Δ as further away from the ○ than the ∅. In the case of the first two rows, doing so resolved some persisting false negatives (from other models) possibly because they invoked **new ideas** (*"gender-neutral pronoun,"*) in the Δ.

Second, the next couple of rows suggest that the model has difficulty when the Δ is too far away from O, or the distance metric isn't useful at distinguishing Δ and ∅. For instance, the winning counter-argument in the third row recontextualized women's rights *("female genital mutilation")*. In this cases, our model incorrectly detected that the counter-argument was digressing from the topic. In contrast, in the last row, both arguments appeared to paraphrase ○. Our posthoc speculation is of a sweet spot between how novel and how relevant a counter-argument must be to win a Δ.

7.3 Discussion

Are the findings driven by syntax or semantics? Semantic similarity between participants in a conversation is termed "linguistic accommodation." It is a well-known persuasion technique documented in other studies (Tan et al., 2016; Niculae et al., 2015); therefore, it formed the basis of our mod-

CMV Cross-Domain dataset - the false negatives *resolved* by h_5 that persist in models from h_1 to h_4		
Topic & Original Post (○)	**Losing counter-argument (∅)**	**Winning counter-argument (△)**
He, she, they. that's all the pronouns you are getting. My proposition is that pronouns past the three most common ones are not necessary and are actively harmful. (...) 1. Pronouns are always dealing with a spectrum (...) 2. (...) we would very likely end up with a sheer infinite amount of pronouns. (...)	**(Cosine similarity = 0.93)** New words are added to languages all the time to describe new concepts and new ideas. Hundreds of years ago, nobody knew what an automobile was (...) Why can't we add new words to our language (...)?	**(Cosine similarity = 0.92)**✔A new single, **gender-neutral pronoun** could be introduced (...) "they" already basically fills this role (...) even if correct it is potentially confusing because that word can also be used as a plural pronoun. (...) language isn't really a **top-down process** like that, but much more **organic**. (...)
Prosecuting elderly germans who allegedly aided the holocaust is counter-productive I believe that the German laws allowing old people to be prosecuted for crimes committed during the Holocaust provide few benefits and may cause harmful effects instead. (...)	**(Cosine similarity = 0.94)** But that statue of limitations was created by society (...) blaming the entire holocaust on them mean the judge was not impartial, or the lawyer was incompetent (...)	**(Cosine similarity = 0.93)**✔I understand what you're trying to say (...) But those that are on trial are not innocent, they did play a role in the Holocaust. They were the **accountants, the guards, the drivers**, they did make it possible. (...)
CMV Cross-Domain dataset - the false negatives *not resolved* by h_5 that also persist in earlier models		
CMV: The Kurds are the good guys, and western nations should be engaging in a hands-down alliance with them in the middle east.(...) the current Middle East conflicts goes only to reinforce my ever-growing belief that the Kurds are the only group involved with any moral high ground. (...)	**(Cosine similarity = 0.92)** First of all, there are a number of other actors which does subscribe to woman's rights. (...) Kurdish middlemen (...) are quite willing to work with ISIS when there's material gains involved. (...)	**(Cosine similarity = 0.80)**✔"The Kurds"are an ethnic group, with a similar mix of people as any other (...) For example, you cite women's rights, when Kurdish areas have some of the highest rates of **Female Genital Mutilation** of anywhere in the world. (...)
I like anecdotal evidence. Whenever I want to know more about an idea, product, etc. I look to someone who has that idea or uses that product to learn more about it (...) you would likely not get cancer from cigarettes either because of your genetic background?	**(Cosine similarity = 0.93)** It seems that your view is basically anecdotal evidence is nice sometimes. I can't think how this could ever be disproven. Could you tell us what would change your view?	**(Cosine similarity = 0.93)**✗In that specific example, no, you wouldn't want to count on not getting cancer. Lung cancer only kills 1/7 of **people who smoke** two packs a day or more, so if there were 15 people in your family (...)

Table 5: Error diagnostics for a random selection of false negatives from the cross-domain data.

eling constraints. Our findings only suggest that winning arguments are further away from original posts in *either or both* syntax and semantics. Based on our empirical findings and previous work, we can attribute at least some contribution of the finding to the difference in semantics. Figure 3 suggests that our hypotheses bear out in terms of the cosine distances between the latent spaces corresponding to the different arguments and the original posts. Most importantly, Table 5 gives our inference face-validity, as we see that the arguments with very different phrasing (i.e., semantically distant) are also more distant from the original post when we rely on the "belief"-based calculus. Prior research has discussed how novel information is always more popular and shareworthy (Vosoughi et al., 2018), which might also explain why can often cause more people to change their minds. However, when we change the underlying assumptions and operationalizations of Δ and $\emptyset$, we see an effect in the relative performance of different modeling constraints. For instance, in the case of the IQ2 dataset, the winning arguments was chosen based on the audience vote.

8 Conclusion

A hierarchical VAE model was applied to model belief alignment in multi-party interactive arguments. The findings suggest that in debating contexts, participants may be open to changing their views when they encounter novel information. An evident limitation is that our findings are correlational and may only apply when participants are open to changing their point of view. Yet, they offer the potential to model dialogue in new contexts invoking persuasion, influence, and cognitive dissonance, such as exposure to misinformation.

9 Limitations

The limitation of this paper is the weakness of association between actual human belief states and what is being modeled by a VAE. In addition there could be an arbitrary many number of other modeling choices instead of what we have used in this paper. As such many of the findings here should be taken as suggestions or indications, rather than non-negotiable scientific claims. Also, our findings are correlational and may only apply in a setting (ChangeMyView Subreddit, or in debates) when participants are open to changing their point of view and less applicable to when participants are less open to view change.

10 Acknowledgements

We thank David Mueller, Alexandra Delucia, Lynnette Ng, and the anonymous reviewers for comments and helpful suggestions.

References

Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*.

Jiaao Chen and Diyi Yang. 2021. Weakly-supervised hierarchical models for predicting persuasive strategies in good-faith textual requests. *arXiv preprint arXiv:2101.06351*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Daria Dayter and Thomas C Messerli. 2021. Persuasive language and features of formality on the r/changemyview subreddit. *Internet Pragmatics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*.

Leon Festinger. 1962. *A Theory of Cognitive Dissonance*, volume 2. Stanford University Press.

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

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. 2019. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*.

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

Christopher Thomas Hidey and Kathleen McKeown. 2018. Persuasive influence detection: The role of argument sequencing. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Elad Hoffer and Nir Ailon. 2015. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer.

Yohan Jo, Shivani Poddar, Byungsoo Jeon, Qinlan Shen, Carolyn Rosé, and Graham Neubig. 2018. Attentive interaction model: Modeling changes in view in argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 103–116, New Orleans, Louisiana. Association for Computational Linguistics.

Simon Keizer, Markus Guhe, Heriberto Cuayáhuitl, Ioannis Efstathiou, Klaus-Peter Engelbrecht, Mihai Dobre, Alex Lascarides, and Oliver Lemon. 2017. Evaluating persuasion strategies and deep reinforcement learning methods for negotiation dialogue agents. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Humphrey Mensah, Lu Xiao, and Sucheta Soundarajan. 2019. Characterizing susceptible users on reddit's changemyview. In *Proceedings of the 10th International Conference on Social Media and Society*, pages 102–107.

Vlad Niculae, Srijan Kumar, Jordan Boyd-Graber, and Cristian Danescu-Niculescu-Mizil. 2015. Linguistic harbingers of betrayal: A case study on an online strategy game. *arXiv preprint arXiv:1506.04744*.

Ayşe Öcal, Lu Xiao, and Jaihyun Park. 2021. Reasoning in social media: insights from reddit "change my view" submissions. *Online Information Review*.

Sebastian Padó, André Blessing, Nico Blokker, Erenay Dayanik, Sebastian Haunss, and Jonas Kuhn. 2019. Who sides with whom? towards computational construction of discourse networks for political debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2841–2847.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Peter Potash and Anna Rumshisky. 2017. Towards debate automation: a recurrent model for predicting debate winners. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Vinodkumar Prabhakaran, Ajita John, and Dorée D Seligmann. 2013. Who had the upper hand? ranking participants of interactions based on their relative power. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 365–373.

Ramit Sawhney, Arnav Wadhwa, Shivam Agarwal, and Rajiv Shah. 2020. Gpols: A contextual graph-based language model for analyzing parliamentary debates and political cohesion. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4847–4859.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Charles S. Taber and Milton Lodge. 2006. Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, 50(3):755–769.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624. International World Wide Web Conferences Steering Committee.

Calum Thornhill, Quentin Meeus, Jeroen Peperkamp, and Bettina Berendt. 2019. A digital nudge to counter confirmation bias. *Frontiers in big data*, 2:11.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Huy Vu, Salvatore Giorgi, Jeremy D. W. Clifton, Niranjan Balasubramanian, and H. Andrew Schwartz. 2022. Characterizing social spambots by their human traits. In *International AAAI Conference on Web and Social Media: ICWSM 2022*, Online. Association for the Advancement of Artificial Intelligence.

Lu Wang, Nick Beauchamp, Sarah Shugars, and Kechen Qin. 2017. Winning on the merits: The joint effects of content and style on debate outcomes. *Transactions of the Association for Computational Linguistics*, 5:219–232.

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Michael Workman. 2018. An empirical study of social media exchanges about a controversial topic: Confirmation bias and participant characteristics. *The Journal of Social Media in Society*, 7(1):381–400.

Diyi Yang, Jiaao Chen, Zichao Yang, Dan Jurafsky, and Eduard Hovy. 2019. Let's make your request more persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in Oxford-style debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*.

A In-Domain and Cross-Domain evaluation setup

Previous work (Jo et al., 2018) has created an evaluation setup that involved separating the test set into in-domain and cross-domain observations. The authors first obtained topics by running Latent Dirichlet Allocation on the entire data with 100 topics, taking each post/counter-argument as a document. The training set is seven topics that have the highest $\Delta : \emptyset$ ratios. The similarity of topics was computed by applying a cosine similarity metric to compare the topic distributions between training and test set data.

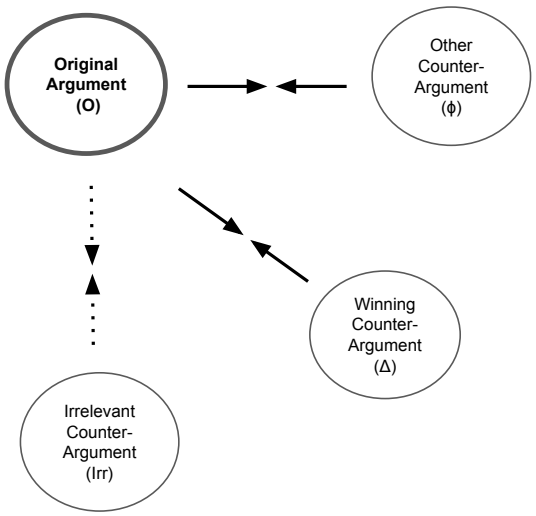

Figure 5: Symbolic representation of the original post (O) and counter-arguments in the ChangeMyView Reddit.

B Symbolic representation of the modeling constraints

A symbolic representation of the original post (O) and counter-arguments in the ChangeMyView Reddit is reported in Figure 5.