

What Changed Your Mind: The Roles of Dynamic Topics and Discourse in Argumentation Process

Jichuan Zeng
The Chinese University of Hong Kong
Hong Kong, China
jczeng@cse.cuhk.edu.hk

Jing Li
The Hong Kong Polytechnic
University, Hong Kong, China
jing-amelia.li@polyu.edu.hk

Yulan He
University of Warwick
United Kingdom
yulan.he@warwick.ac.uk

Cuiyun Gao*
Harbin Institute of Technology
(Shenzhen), China
gcyydxf@gmail.com

Michael R. Lyu
The Chinese University of Hong Kong
Hong Kong, China
lyu@cse.cuhk.edu.hk

Irwin King
The Chinese University of Hong Kong
Hong Kong, China
king@cse.cuhk.edu.hk

ABSTRACT

In our world with full of uncertainty, debates and argumentation contribute to the progress of science and society. Despite of the increasing attention to characterize human arguments, most progress made so far focus on the debate outcome, largely ignoring the dynamic patterns in argumentation processes. This paper presents a study that automatically analyzes the key factors in argument persuasiveness, beyond simply predicting who will persuade whom. Specifically, we propose a novel neural model that is able to dynamically track the changes of latent *topics and discourse* in argumentative conversations, allowing the investigation of their roles in influencing the outcomes of persuasion. Extensive experiments have been conducted on argumentative conversations on both social media and supreme court. The results show that our model outperforms state-of-the-art models in identifying persuasive arguments via explicitly exploring dynamic factors of topic and discourse. We further analyze the effects of topics and discourse on persuasiveness, and find that they are both useful – topics provide concrete evidence while superior discourse styles may bias participants, especially in social media arguments. In addition, we draw some findings from our empirical results, which will help people better engage in future persuasive conversations.

CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**; • **Human-centered computing** → *Social media*.

KEYWORDS

social media, argumentation mining, topic modeling, discourse modeling, dynamic data processing

ACM Reference Format:

Jichuan Zeng, Jing Li, Yulan He, Cuiyun Gao, Michael R. Lyu, and Irwin King. 2020. What Changed Your Mind: The Roles of Dynamic Topics and

* Cuiyun Gao is the corresponding author.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW '20, April 20–24, 2020, Taipei, Taiwan

© 2020 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.

ACM ISBN 978-1-4503-7023-3/20/04.

<https://doi.org/10.1145/3366423.3380223>

Discourse in Argumentation Process. In *Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan*. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3366423.3380223>

1 INTRODUCTION

“The aim of argument, or of discussion, should not be victory, but progress.” – Joseph Joubert

Argumentation process is a turn-taking dialogue mostly held to increase the acceptability of a controversial standpoint. In the process, a series of connected propositions (henceforth **arguments**) are put forward intending to justify or refute a standpoint before a rational judge [40]. It plays an essential role in making decisions, constructing knowledge, and bringing truths and better ideas to life [19]. Consequently, the understanding of argumentation processes will help individuals and human society better engage with conflicting stances and open up their minds to pros and cons [24]. It collides different ideas to form thoughts and knowledge, contributing to advance science and society forward [45]. However, making sense of argumentative conversations is a daunting task for human readers, mostly due to the varied viewpoints and evidence continuously put forward and the complicated interaction structure therein; not to mention huge volume of argumentation data appearing on online platforms every day.

We hence study how to automatically understand argumentation processes, predicting who will persuade whom and figuring out why it happens. To date, much progress made in persuasiveness prediction has focused on individual arguments, the wordings therein [13, 44], and how they locally connect with other arguments [15, 17]. On the contrary, we examine the context and the dynamic progress of argumentative conversations, which is beyond the studies of argument-level persuasiveness. Some research work analyze argument interactions [14, 43] to predict who will win the debate. Most of them focus on the outcome of argumentation instead of diving deep into the argumentation process [19, 37]. The latter, however, is arguably the essence of argumentation, revealing how participants collaborate to reshape and refine ideas.

In light of these missing points, we track the argumentation process and explicitly explore the dynamic patterns of what a discussion is centered around (henceforth **topics**) and how the participants voice their opinion in arguments (henceforth **discourse**), as well as how they affect the persuasion results. To illustrate the interplay of topics and discourse in argument persuasiveness, Figure 1

OP: Translation software already exists and is pretty good.
 ... Most people putting in years of effort to learn a foreign language skill they might only use a couple times in their whole life, and will likely forget.
A₁ [Evidence]: ... There is research that indicates “that *those who spoke two or more languages had significantly better cognitive abilities compared to what would have been expected from their baseline test.*” [\(url\)](#). ... Another study found that “*the language-learning participants ended up with increased density in their grey matter and that their white matter tissue had been strengthened.*” [\(url\)](#)
A₂ [Metaphor]: The common comparison is made to learning music, as /u/awesomeosprey has pointed out. I did some research into the matter. It seems that *learning a musical instrument does have long-lasting benefits ((url)) that relate to “higher-order aspects of cognition.”*
 ...
A₄ [Reference] ... But a quick search and I have other sources: [\(digit\)](#) [\(url\)](#), [\(digit\)](#) [\(url\)](#), [\(digit\)](#) [\(url\)](#). The most interesting study is this one ((url)), but I can’t find a complete version of it, sorry. /n/nNote: Study [\(digit\)](#) has an exceptionally small sample size. It’s still interesting reading.

Figure 1: A ChangeMyView conversation snippet of challengers’ arguments against *OP* raised by the opinion holder concerning “learning a second language isn’t worth it anymore for most people”. The red and italic words indicate the key points resulting in the challengers’ victory. The words in [] are our interpretations of the arguments’ discourse styles.

shows a Reddit conversation snippet from ChangeMyView subreddit.¹ On ChangeMyView, an opinion holder first raises a viewpoint (henceforth *OP* short for original post), followed by challengers’ arguments attempting to change the opinion holder’s mind. This example dialogue is formed with challengers’ arguments against “learning a second language isn’t worth it for most people anymore”, which was the opinion holder’s point of view.

It is seen that the challengers successfully persuaded the opinion holder to change their view in the aforementioned example. The probable reasons are two fold. First, there are strong evidences (reflected by topic words) put forward, such as the research findings on cognitive abilities. Second, they deploy skillful debating styles (captured by discourse words), such as the metaphors with learning music (in *A₂*) and the reference to external information (in *A₄*).

Motivated with these observations, we propose a novel neural framework that explicitly models how the change of discussion topic and discourse styles affect persuasion effectiveness. Our model first explores latent topics and discourse in arguments with word clusters. Furthermore, it tracks topic change and discourse flow in the argumentation process and automatically interprets the key factors indicating the success or failure of the persuasion. Coupling the advantages of neural topic models [28, 49, 51] and dynamic memory networks [23, 46, 54], we are able to explore dynamic topic and discourse representations indicative of persuasiveness in an end-to-end manner with the persuasion outcome prediction. To the best of our knowledge, we are the first to *explicitly model topics*

and discourse in argumentation processes, and investigate how their dynamic patterns contribute to the argument persuasiveness.

We carry out extensive experiments on argumentative conversations gathered from both social media and U.S. supreme court. The results show that our model can significantly outperform state-of-the-art methods on both datasets, which shows its effectiveness in identifying persuasive arguments. For example, we achieve 70.2% accuracy when predicting winners in supreme court debates, compared with 63.1% obtained by logistic regression without explicitly exploiting dynamic topics and discourse features in argumentation processes. Based on the produced topics and discourse, we further analyze how they affect persuasiveness. It is indicated that topics (such as evidence and viewpoints) statistically contribute more on persuasion success while skillful discourse style may sometimes lead to victory. In addition, we summarize the key findings from our empirical results, which will help individuals better engage in future persuasions.

To sum up, our contributions are three fold:

- We are the first to study the argumentation process via dynamic analysis of latent topics and discourse, which reveals the key factors in argument persuasiveness.
- We propose a novel neural model to predict argumentation outcome via tracking dynamic topic and discourse patterns in the dialogue process.
- We provide an extensive empirical study on two real-world datasets that demonstrates the effectiveness of our model and sheds light on a better understanding and development of persuasive augmentations.

2 RELATED WORK

2.1 Argument Persuasiveness

As a fast growing sub-field of computational argumentation mining [35, 41], previous work in this area mostly work on the identification of convincing arguments [13, 44] and viewpoints [14, 19] from varying argumentation genres, such as social media discussions [37], political debates [4], and student essays [6]. In this line, many existing studies focus on crafting hand-made features [37, 44], such as wordings and topic strengths [43, 53], echoed words [2], semantic and syntactic rules [15, 30], participants’ personality [42], argument interactions and structure [29], and so forth. These methods, however, require labor-intensive feature engineering process, and hence have limited generalization abilities to new domains.

Recently, built upon the success of neural models in natural language processing (NLP), neural argumentation mining methods have been proposed to enable end-to-end learning of automatic features and argument persuasiveness. For example, Potash et al. [31] tailor a pointer network architecture to learn argument representations. Lin et al. [26] focus on incorporating external lexicons into an attentive neural network for argumentative component identification. These studies, however, ignore the dynamic nature of argumentation process, where the persuasion features may change in a heated back-and-forth debate. Some other methods consider the modeling of the argument interactions in persuasiveness prediction. Ji et al. [17] explore the argument-level interactions between ChangeMyView original post (*OP*) and its following comments with a co-attention network. Jo et al. [19] investigate the interplay

¹<https://www.reddit.com/r/changemyview/>

between *OP* and its challenger’s argument, explicitly identifying the amenable parts of *OP* that is likely to be affected with good arguments. Compared with these work focusing on interaction between *OP* and comments, we dynamically track the entire argumentation flow and capture how topics and discourse therein change and affect persuasion outcomes. Hidey and McKeown [14] employ sequence modeling to learn implicit persuasiveness signals from chronologically ordered arguments. Different from them, we explicitly capture the dynamic topics and discourse behaviors as discussion process is moved forward, where their roles in shaping the persuasive arguments can be examined.

2.2 Conversation Process Understanding

Our work is also closely related with conversation process understanding. In this line, previous studies have shown the benefits of discovering the latent discourse structure. It shapes how utterances interact with each other and form the discussion flow with the use of dialogue acts (e.g., making a statement, asking a question, and giving an example). Most of them extend Hidden Markov Model (HMM) to produce distributional clusters of words to reflect latent discourse [8, 33]. In discourse learning, features are exploited via modeling of conversation tree structure [25], relative position of sentences [20], topic content [32, 52], and so forth.

In addition, the recent progress in recurrent variational neural networks (non-linear HMM counterpart) enables to capture latent discourse structure in dialogues. For example, latent variable RNN (LVRNN) and variational RNN (VRNN) have been adopted to model the latent conversation states in each turn [18, 34]. Zeng et al. [49] jointly explore the topic content and discourse behavior to better understand conversations by using the word clusters to represent topics and discourse in microblog conversations. However, none of them captures how topics and discourse change in a conversation process and how these dynamic patterns affect argumentation persuasiveness, which is the gap our work fills in.

3 STUDY DESIGN

In this section, we first introduce how we formulate our problem, followed by a detailed discussion on the experimental datasets.

3.1 Problem Formulation

In this paper, we define **argumentation process** C as a dynamic conversation process held by participants. It is formulated as a sequence of turns, denoted as $C = \{x_t\}_{t=1}^T$, where a turn x_t refers to an argument and T the number of turns in the process. As discussed above, our work studies argument persuasiveness in the context of its discussion process, which however relies on subjective judgement. After all, human performance on “yes-or-no” persuasiveness judgement is still close to random guess [37]. In our study, we view argument persuasiveness from a perspective of comparison (instead of answering “yes or no”), and formulate its prediction as a pairwise ranking problem under a debate \mathcal{D} . Concretely, we construct the pairwise comparison settings to take a pair of argumentation process $\langle C_i, C_j \rangle$ as input, where $C_i, C_j \in \mathcal{D}$; Scores y_i and y_j are assigned to measure their persuasiveness respectively. Here $y_i > y_j$ means that C_i has a better chance to win the debate compared with C_j , while $y_i < y_j$ otherwise. The goal of our paper is to predict

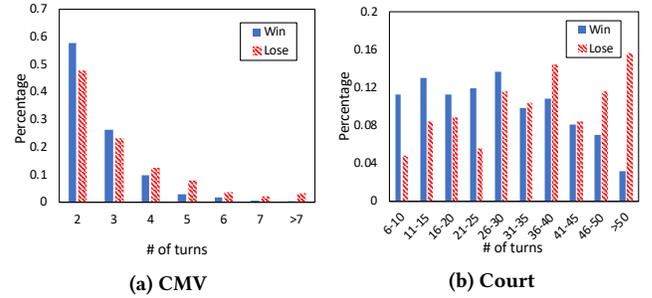


Figure 2: Distributions over the winning and losing augmentation processes concerning their number (range) of turns. (a) is for ChangeMyView (CMV) dataset and (b) supreme court (Court) dataset.

Table 1: Statistics of the ChangeMyView (CMV) and the supreme court (Court) datasets. Here a moot refers to an original post in CMV and a case in Court.

Datasets	# of moots	# of convs	# of turns	avg. words per turn	vocab	# of pairs
CMV	2,396	30,341	109,644	96.2	13,541	12,879
Court	204	655	17,599	46.1	6,260	3,656

which argumentation process from the input pair is relatively more persuasive and analyze the key factors therein to reveal insights for argumentation study.

Our problem setting can fit diverse scenarios to learn what a good persuasion should be. For example, it works for the classic Oxford-style debate involving two sides, where one argues “for” a statement and the other “against”. The arguments from both sides can be defined as $C^f = \{x_t^f\}$ (“for” side) and $C^a = \{x_t^a\}$ (“against” side), which corresponds to our input pair in problem setting.

3.2 Data Description

We conduct our study in two scenarios — social media arguments, which tend to use colloquial and informal languages, and supreme court debates,² exhibiting a more formal language style. The social media arguments are gathered from the ChangeMyView subreddit, where challengers engage in the discussion with attempts to change the opinion holder’s view (pointed out in the original post *OP*) [37]. As a multi-party conversation, a debate there is in tree structure formed with in-reply-to relations (a post can have multiple replies), and a path therein is defined as an argumentation process. We aim to predict which path has a better chance to be awarded a Δ by the opinion holder to indicate successful persuasion. For the supreme court debates, we aim to predict whether the petitioner or respondent will win the case, given their corresponding conversational exchanges with the justices.

The ChangMyView social media dataset (henceforth *CMV*) is built with a corpus released by Tan et al. [37] with argumentative conversations held from Jan 2013 to May 2015. As stated above, each discussion in *CMV* can be organized in a tree structure with in-reply-to relations (henceforth a debate tree), with its root representing the *OP* (the opinion holder’s viewpoint). To construct our input data, following Tan et al. [37], we first filter out the trivial

²https://www.supremecourt.gov/oral_arguments/oral_arguments.aspx

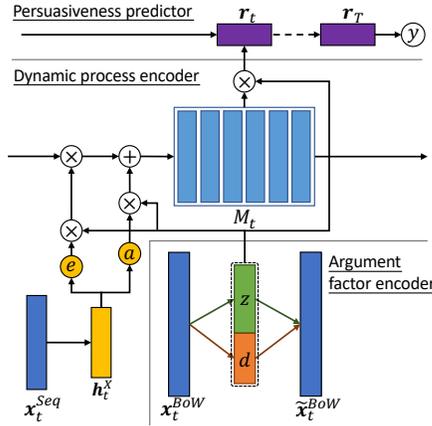


Figure 3: The architecture of our dynamic topic-discourse memory networks (DTDMN) for persuasiveness prediction.

cases by removing the discussions with less than 10 challengers, or those do not contain a Δ . Then, we flatten the debate tree into conversation paths and remove replies with 50 words or less. Also removed are conversation paths involving less than two turns³. Next, all challengers’ replies remained in a conversation path is considered as the turns in argumentation process. For each debate tree, we form a positive candidate set with all the argumentation processes (paths) leading to a Δ , and include those without a Δ into the negative candidate set. To formulate our pairwise inputs, we perform the Cartesian product⁴ on the positive and negative candidate sets, which returns all the possible combinations of successful/unsuccessful argumentation process pairs in the debate.

For the supreme court debate dataset (abbreviated as **Court**), it is gathered by Danescu-Niculescu-Mizil et al. [9] from the U.S. supreme court dialogues⁵. In this corpus, the petitioner and respondent make conversational exchanges to justices to defend for themselves in turn. Here the petitioner’s utterances are taken to form its augmentation process, and so does the respondent’s. For each case, we build the positive candidate set with argumentation processes from the winning side, and negative from its opponent. The pairwise inputs are formed following the similar procedure used for CMV dataset.

In addition, we employ two strategies to further improve the quality of our data. First, to ensure the argumentation processes in an input pair concern relevant topics, their Jaccard similarity are measured over bag-of-words form. After that, following the practice in Tan et al. [37], we remove pairs with < 0.5 Jaccard similarity where the conversation pairs may not be on the same page. Second, as pointed out in previous studies [37], the number of argument turns can largely affect the debate outcome. Here we show the distribution of turn number over winning and losing argumentation processes in Figure 2 and observe that the winning ones tend to be shorter (with smaller turn number). It might result in trivial features of turn number to be learned for persuasiveness prediction. To mitigate the effects of turn number and better study

³In our paper, unless otherwise specified, a **conversation** is used as the short form of a conversation path.

⁴https://en.wikipedia.org/wiki/Cartesian_product

⁵http://www.supremecourt.gov/oral_arguments/

the roles of topics and discourse, we make sure that the pairwise processes fed to the model are equally long (have the same number of argument turns). To this end, we remove pairs with shorter negative process and for the rest, we truncate the longer parts of negative processes. The statistics of our two datasets are shown in Table 1. As can be seen, there are more conversations in CMV than Court. However, the Court debates involve more turns (26.9 vs. 3.6 turns on average per conversation). It might be because court debates are more serious and usually result in a back-and-forth fashion while social media discussions are mostly casual and may end soon.

It is worth noting that we do not feed the words from either opinion holders or justices to avoid the possible bias incurred in persuasiveness prediction. In doing so, we can focus on linguistic features in participants’ arguments that lead to good persuasion. Further, it enables our setting to be easily adapted to scenarios without the third-party engagement (e.g., opinion holders and justices). In addition, for CMV dataset, we consider the engagements of all challengers regardless of their Δ records, which is different from the setting in Tan et al. [37], which only examine the Δ winners. It is because everyone’s efforts may contribute to the final success (or failure) of an argumentation process. Therefore, all the challengers’ argument are taken into account in our persuasiveness analysis. For the same reason, we have more training data instances than those in Tan et al. [37] (12, 879 vs. 4, 263).

4 DTDMN: DYNAMIC TOPIC-DISOURSE MEMORY NETWORKS FOR ARGUMENT PERSUASIVENESS

This section presents our model that predicts persuasiveness, and dynamically discovers the key topic and discourse factors therein to explain the reasons behind. Our model, named as dynamic topic-discourse memory networks (DTDMN), consists of three modules — one to learn latent topic and discourse factors from each argument (henceforth **argument factor encoder**), one to explore the change of topic and discourse factors in argumentation flows (henceforth **dynamic process encoder**), and the last one to identify the more persuasive conversation from the input pair (henceforth **persuasiveness predictor**). The model architecture is shown in Figure 3 with an overview presented in Section 4.1. Then in Section 4.2, 4.3, and 4.4, we describe our three modules in turn, followed by our learning objective discussed in Section 4.5.

4.1 Model Overview

As described in Section 3, our model takes pairwise conversations as input. In training, we feed $\langle C^+; C^- \rangle$ into our model, where C^+ is a positive instance referring to a persuasive conversation. Likewise, C^- , the negative instance, denotes a failed persuasion. During the testing, given two conversations, our model will recognize the one which is more persuasive. Each conversation C is formed with a sequence of argumentative turns (henceforth **arguments**): $C = \langle x_1, \dots, x_T \rangle$, where T denotes the number of arguments in C .

For the t -th argument x_t , we capture argument-level representations, $z_t \in \mathbb{R}^K$ for topic factor and $d_t \in \mathbb{R}^D$ for discourse factor, from the input of bag-of-words vector $x_t^{BoW} \in \mathbb{R}^V$, where K is the number of topics, D discourse, and V the vocabulary size. Then, z_t

and \mathbf{d}_t are fed into the dynamic memory, together with the word index sequence $\mathbf{x}_t^{Seq} \in \mathbb{R}^L$, to update the memory state, where L is the sequence length. The output of the dynamic memory networks is used to predict the persuasiveness score y for each conversation, where higher scores indicate better persuasiveness. Our training target is to have $y^+ > y^-$ for C^+ and C^- .

4.2 Argument Factor Encoder

This section presents how we capture topic and discourse factors at the argument level. The subscript t is omitted for simplicity. As mentioned in Section 4.1, we employ latent variables \mathbf{z} for argument topic factor representation, and \mathbf{d} for discourse. The modeling process is inspired by Zeng et al. [50] and based on variational auto-encoder (VAE) [22] to reconstruct a given argument in the BoW form, \mathbf{x}^{BoW} , conditioned on \mathbf{z} and \mathbf{d} . Here \mathbf{z} is the topic mixture and \mathbf{d} is a one-hot vector denoting the discourse style.⁶ Specifically, the generation process for each word $w_n \in \mathbf{x}^{BoW}$ is defined as:

$$\begin{aligned} \epsilon &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2), \quad \mathbf{z} = \text{softmax}(f_z(\epsilon)), \quad \mathbf{d} \sim \text{Multi}(\boldsymbol{\pi}), \\ \beta_n &= \text{softmax}(f_{\phi^T}(\mathbf{z}) + f_{\phi^D}(\mathbf{d})), \quad w_n \sim \text{Multi}(\beta_n), \end{aligned} \quad (1)$$

where $f_*(\cdot)$ is a neural perceptron that linearly transforms inputs. For both latent topic and discourse factors, we employ word distributions to represent them. Here we consider the weight matrix of $f_{\phi^T}(\cdot)$ (after the softmax normalization) as topic-word distributions, ϕ^T . Likewise, $f_{\phi^D}(\cdot)$'s weight matrix is used to compute the discourse-word distributions, ϕ^D .

For the other parameters $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, and $\boldsymbol{\pi}$, they can be learned from the input \mathbf{x}^{BoW} following the formula below:

$$\begin{aligned} \boldsymbol{\mu} &= f_{\mu}(\tanh(f_e(\mathbf{x}^{BoW}))), \quad \log \boldsymbol{\sigma} = f_{\sigma}(\tanh(f_e(\mathbf{x}^{BoW}))), \\ \boldsymbol{\pi} &= \text{softmax}(f_{\pi}(\mathbf{x}^{BoW})). \end{aligned} \quad (2)$$

4.3 Dynamic Process Encoder

Based on the topic and discourse factors learned at the argument level, here we discuss how to capture their dynamic patterns in the persuasion process. Our dynamic process encoder is inspired by dynamic memory network (DMN) [23, 46, 54] and topic memory mechanism [51], where we capture the indicative dynamic topic and discourse factors to interpret why a conversation can result in successful persuasion.

To be more specific, memory weight $\mathbf{w}_t \in \mathbb{R}^{(K+D)}$ is defined as the concatenation of latent aspects \mathbf{z}_t and \mathbf{d}_t :

$$\mathbf{w}_t = [\mathbf{z}_t; \mathbf{d}_t], \quad (3)$$

where $[\cdot; \cdot]$ represents the concatenation. Once we have the memory weight, DTDMN will retrieve and update the memory according to the memory weight and input argument.

We employ a bidirectional attentive GRU [3, 47] to encode the word index sequence vector input \mathbf{x}_t^{Seq} into hidden states $\mathbf{h}_t^X \in \mathbb{R}^H$:

$$\begin{aligned} \overrightarrow{\mathbf{h}}_{t,j}^X &= \overrightarrow{\text{GRU}}(x_{t,j}; \overrightarrow{\mathbf{h}}_{t,j-1}^X), \quad \overleftarrow{\mathbf{h}}_{t,j}^X = \overleftarrow{\text{GRU}}(x_{t,j}; \overleftarrow{\mathbf{h}}_{t,j+1}^X), \\ \mathbf{h}_t^X &= \text{attn}(\{\overrightarrow{\mathbf{h}}_{t,j}^X; \overleftarrow{\mathbf{h}}_{t,j}^X\}_{j=1}^L), \end{aligned} \quad (4)$$

where $j \in [1, L]$, $x_{t,j}$ is the j -th token in \mathbf{x}_t^{Seq} . $\text{attn}(\cdot)$ is the attention operator [3, 27] to aggregate the representations of tokens to form a vector representation for \mathbf{x}_t^{Seq} .

Similar to Zhang et al. [54], we employ a forget gate to erase the retrieved memory. The erase vector is denoted as $\mathbf{e}_t \in \mathbb{R}^E$, where E is the dimension of memory embeddings. Afterwards, an augment gate is used to strengthen the retrieved memory. The augment vector is denoted as $\mathbf{a}_t \in \mathbb{R}^E$. The overall update formulae for episodic memory are:

$$\begin{aligned} \mathbf{M}_{t,i} &= \mathbf{M}_{t-1,i}[\mathbf{1} - w_{t,i}\mathbf{e}_t] + w_{t,i}\mathbf{a}_t, \\ \mathbf{e}_t &= \text{sigmoid}(\mathbf{W}^{(e)}\mathbf{h}_t^X + \mathbf{b}^{(e)}), \quad \mathbf{a}_t = \text{tanh}(\mathbf{W}^{(a)}\mathbf{h}_t^X + \mathbf{b}^{(a)}), \end{aligned} \quad (5)$$

where $\mathbf{M}_{t,i} \in \mathbb{R}^E$ is the i -th row of the memory matrix \mathbf{M}_t , $\mathbf{1}$ is a row-vector of all 1s. $\mathbf{W}^{(e)}, \mathbf{W}^{(a)} \in \mathbb{R}^{E \times H}$ and $\mathbf{b}^{(e)}, \mathbf{b}^{(a)} \in \mathbb{R}^E$ are the weight matrices and bias vectors for computing \mathbf{e}_t and \mathbf{a}_t , respectively. The read content $\mathbf{r}_t \in \mathbb{R}^E$ of the episodic memory \mathbf{M}_t is the weighted sum of the memory matrix:

$$\mathbf{r}_t = \sum_{i=1}^{K+D} w_{t,i}\mathbf{M}_{t,i}. \quad (6)$$

4.4 Persuasiveness Predictor

For each conversation, DTDMN dynamically summarizes the read contents of the previous arguments in a conversation $\{\mathbf{r}_t\}_{t=1}^{T'}$ via an attentive GRU at the argument level:

$$\begin{aligned} \mathbf{h}_t^R &= \text{GRU}(\mathbf{r}_t; \mathbf{h}_{t-1}^R), \\ \mathbf{h}^R &= \text{attn}(\{\mathbf{h}_t^R\}_{t=1}^{T'}). \end{aligned} \quad (7)$$

Then we map \mathbf{h}^R to a score value y :

$$y = \mathbf{W}^{(r)}\mathbf{h}^R + b^{(r)}, \quad (8)$$

where $\mathbf{W}^{(r)} \in \mathbb{R}^{1 \times E}$ and $b^{(r)} \in \mathbb{R}^1$ are weight and bias for computing y .

4.5 Learning Objective

Argument Factor Learning. To model topic and discourse factors, in learning, we maximize the variational lower bound \mathcal{L}_z for \mathbf{z} and \mathcal{L}_d for \mathbf{d} . The corresponding functions are defined as:

$$\begin{aligned} \mathcal{L}_z &= \mathbb{E}_{q(\mathbf{z} | \mathbf{x})}[\log p(\mathbf{x} | \mathbf{z})] - D_{KL}(q(\mathbf{z} | \mathbf{x}) || p(\mathbf{z})), \\ \mathcal{L}_d &= \mathbb{E}_{q(\mathbf{d} | \mathbf{x})}[\log p(\mathbf{x} | \mathbf{d})] - D_{KL}(q(\mathbf{d} | \mathbf{x}) || p(\mathbf{d})), \end{aligned} \quad (9)$$

where $p(\mathbf{z})$ is the standard normal prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and $p(\mathbf{d})$ the uniform distribution $\text{Unif}(0, 1)$. $q(\mathbf{z} | \mathbf{x})$ and $q(\mathbf{d} | \mathbf{x})$ are posterior probabilities to approximate how \mathbf{z} and \mathbf{d} are generated from the arguments. $p(\mathbf{x} | \mathbf{z})$ and $p(\mathbf{x} | \mathbf{d})$ represent the corpus likelihoods conditioned on these topic and discourse factors.

The overall argument factor learning objective is to maximize:

$$\mathcal{L}_{\text{Factor}} = \mathcal{L}_z + \mathcal{L}_d + \mathcal{L}_x - \lambda \mathcal{L}_{MI}, \quad (10)$$

where \mathcal{L}_x is for reconstructing the argument \mathbf{x} from \mathbf{z} and \mathbf{d} , \mathcal{L}_{MI} is the mutual information (MI) penalty (for separating topic and discourse words). The hyperparameter λ is the trade-off parameter for balancing between the \mathcal{L}_{MI} and the other learning objectives. We leave out the details and refer the readers to Zeng et al. [50].

⁶We follow the setting of Zeng et al. [50], and apply Gumbel-Softmax relaxation for \mathbf{d} .

Persuasiveness Prediction Learning. In our setting, we aim to identify which conversation is more persuasive given an input of two conversations. Therefore, our goal is to have C^+ scored higher than C^- . We apply the pairwise cross-entropy loss to maximize the margin of y^+ and y^- for C^+ and C^- , which equals to minimize:

$$\mathcal{L}_{Pred} = \log(1 + \exp(y^- - y^+)). \quad (11)$$

Overall learning Objective. The three components of our model can be jointly optimized by minimizing the objective function:

$$\mathcal{L} = \mathcal{L}_{Pred} - \sum_t (\mathcal{L}_{Factor}^t), \quad (12)$$

where \mathcal{L}_{Factor}^t is for argument turn level.

5 EXPERIMENTAL SETUP

Data Preprocessing. We randomly split the dataset with 80% for training and 20% for test. Then, 20% of the training data is randomly selected for validation. For preprocessing, we take the the following steps. First, non-English terms were filtered out. Then, quotations, digits, and links were replaced with generic tags ‘<quote>’, ‘<digit>’, and ‘<url>’, respectively. Next, we employed the natural language toolkit (NLTK) for tokenization⁷. After that, all letters were converted to lowercase. Finally, words occurred less than 10 times were filtered out from the data.

Parameter Setting. We use Gated Recurrent Unit (GRU) as the RNN cell. The hidden size of GRU is set to 512 with the word dropout rate of 0.2. The dimensions of word embeddings and memory embeddings are both set to 200. $\lambda = 0.01$ following the setting of Zeng et al. [50] for balancing the MI loss. For all the other hyperparameters, we tune them on the validation set by grid search. Optimization is performed using Adam [21]. In the learning process, we alternatively update the parameters of the argument factor encoder and the rest of our model. We run our model for 80 epochs with early-stop strategy applied [7].

Comparison Baselines. Tan et al. [37] uses logistic regression with bag-of-words features in the pairwise persuasiveness prediction tasks, achieving good performance when compared with most of the handcrafted features. Here we implement logistic regression with TfIdf-weighted n -grams features (LR-TFIDF). Similar to [37], we adopt ℓ_1 regularization on the training stage to avoid overfitting. Joint topic-discourse model (JTDM) [50] extracts topics and discourse features in an unsupervised way and can be used to place our argument factor encoder. We use the mean of each argument’s topic-discourse mixture as the feature of an input conversation without considering the dynamics. Hierarchical attention recursive neural network (HATT-RNN) [48] uses bi-directional GRU as sequence encoder, including two levels of attention mechanisms (i.e., word level and argument level) while constructing the representation of a conversation. Dynamic memory network (DMN) [23] is a neural sequence model that can encode the contextual history into the episodic memory component. Dynamic key-value memory network (DKVMN) [54] improves upon DMN using one static matrix

Table 2: Pairwise classification results on persuasiveness prediction. Best results in bold. Paired t-test is conducted between our full model and baselines/ablations (: $p < 0.01$, *: $p < 0.05$).**

Models	CMV		Court	
	Acc.	F1	Acc.	F1
Baselines				
LR-TFIDF	0.571**	0.557**	0.631**	0.608**
JTDM	0.615**	0.586**	0.642**	0.625**
HATT-RNN	0.632**	0.641**	0.571**	0.538**
DMN	0.688**	0.673**	0.637**	0.602**
DKVMN	0.696**	0.698**	0.648**	0.629**
DTDMN				
W/O TOPIC	0.713**	0.710**	0.653**	0.655**
W/O DISCOURSE	0.749*	0.745*	0.671**	0.682**
W/O MEMORY	0.707**	0.696**	0.616**	0.589**
FULL MODEL	0.751	0.748	0.702	0.694

as key to compute the memory reading weights and one dynamic matrix as value for updating the memory states.

6 EXPERIMENTAL RESULTS

This section presents the how models perform on persuasiveness prediction. We reports the main comparison results on persuasiveness prediction in Section 6.1, followed by topic and discourse interpretations in Section 6.2. Afterwards, we analyze the major parameters and errors in Section 6.3 and Section 6.4 respectively.

6.1 Persuasiveness Prediction Comparison

We follow Tan et al. [37] to conduct pairwise classification. For the CMV dataset, we predict which conversation can win Δ , and for the Court dataset, which side will win the case. In Table 2, we report the pairwise accuracy and F1 scores. For our models, we also display ablation results without considering topic, discourse, and memory structure, respectively. It is observed that:

- **Topic and discourse factors are useful.** By exploiting pre-learned latent topic and discourse factors, JTDM outperforms LR-TFIDF baseline on both datasets. It even performs better than HATT-RNN on Court debates. This observation implies that topic and discourse factors can be indicative of persuasiveness arguments.
- **Neural models generally outperform the non-neural baselines.** This indicates that neural models are able to learn deep persuasiveness features. We also find that the improvement upon non-neural models is less significant on the Court compared to CMV. This may be partly attributed to the sparse training instances in the Court dataset as shown in Table 1, which may result in overfitting. Nevertheless, our models can well alleviate such sparsity and achieve significantly better performance on both datasets.
- **Process modeling is important to predict argument persuasiveness.** We observe that LR-TFIDF and JTDM, with only word features encoded, perform worse compared to other methods that explore dynamic patterns in argumentation process. This shows that persuasion outcomes are also dependent on a dynamic process beyond word features.
- **Dynamic memory mechanism is effective.** Our FULL MODEL obtains better results than its W/O MEMORY variant. Also, DMN

⁷<https://www.nltk.org/>

Table 3: Top 10 representative words of example topics learned from CMV and Court. We manually label the topical content according to their associated words.

Topic	Words
<i>Academic (CMV)</i>	phd sociology genetics predetermined biology quantum field classical influences refers
<i>Foreign culture (CMV)</i>	japanese europeans european spanish french africans german indian native heritage
<i>Ecocrisis (CMV)</i>	chernobyl fukushima warming nirvana tolerance hydroelectric dangers grunge swastika warnings
<i>Criminal histroy (Court)</i>	juvenile evasion adult youth records olds history sims thorough court
<i>Commerce (Court)</i>	income billion lawful cents marketplace revenue supply descriptive rate transactions
<i>Political party (Court)</i>	voting voters attract republican republicans challenger democrats poaching vote candidate

and DKVMN outperform other baselines without dynamic memory mechanism. The above observations indicate that dynamic memory mechanism is effective for the argumentation process.

• **Both dynamic topic and discourse factors contribute to argument persuasiveness.** It is observed that our FULL MODEL achieves better results than the w/o TOPIC and w/o DISCOURSE ablation, which considers only dynamic discourse or topic factors. Though the slightly better performance of w/o DISCOURSE than w/o TOPIC shows that topic factors might contribute more to persuasiveness, coupling the topics and discourse exhibiting the best performance.

6.2 Interpretation on Topics and Discourse

To analyze the latent topics and discourse produced our model, we carry out a qualitative analysis to investigate their interpretability. Here we select the top 10 words from the distributions of some example topics and discourse factors and list them in Table 3 and Table 4, respectively. It can be observed that there are some meaningful word clusters reflecting varying debate topics and discourse skills on the two datasets. Interestingly, we observe that latent discourse from CMV and Court, though learned separately, exhibit some overlaps in their corresponding top 10 words; particularly for “pronoun”, which are used to refer to participants (e.g., “we”) or someone/something else (e.g., “he” and “they”) in the discourse. We also note that the discourse style of “statistic” is represented by very different words. The reason is that Court debates are likely to involve lawsuit-related statistical evidence, hence exhibiting the prominence of words like “records” and “proximate”.

We further explore why coherent topics and discourse styles can be learned with the example conversation in Figure 1. In Figure 4, we visualize the topics and discourse assignments where we highlight the topic words (with $p(w|z) > p(w|d)$) in red, the rest in blue to indicate discourse style. The shade indicates the confidence level of such word assignment. We can see that our model can identify the topic words, e.g., “*language beyond*”, “*found*”, and “*learning participants ended*”, and also discourse words, e.g., “<digit>” and “[”.

Table 4: Top 10 representative words of example discourse styles learned from CMV and Court. The discourse styles of the word clusters are manually assigned according to their associated words.

Discourse Role	CMV	Court
<i>Question</i>	what <quote> ? why were how would has could our	can ? you if further him are we go take
<i>Pronoun</i>	my he him his we one am on 've myself	his he was no after did they not this became
<i>Conjecture</i>	as there per more from less often). than many	no there i further don either your any many be
<i>Quotation</i>	[' <url> ^[[][/]][[^ (< > : < :] (.] ” i don if you find would .)
<i>Statistic</i>	more <digit> was than less from could been ' had	resulting <digit> regs records sims definitional proximate thorough counts instruction

It is seen that topics and discourse words can be well distinguished, which allows us to discover meaningful latent factors and analyze reasons behind persuasive arguments.

6.3 Parameter Analysis

Here we study how the two important hyper-parameters in our model, the number of topics (K) and discourse (D) affect our model performance. In Figure 5, we show the persuasiveness prediction accuracy given varying K in (a) and varying D in (b).

As can be seen, for both topic and discourse, the curves corresponding model performance are not monotonic. In particular, better accuracies are achieved given relatively larger topic numbers for CMV with the best result observed at $K = 50$. While for Court, the optimum topic number is $K = 20$. This may be due to the relatively more centralized topics in Court debates, whereas wider range of topics discussed in social media, CMV. For discourse, we observed a similar trend in both CMV and Court datasets. The best score is achieved when $D = 10$ for CMV and $D = 8$ for Court dataset. This implies that the discourse styles used in both CMV and Court are somewhat limited.

6.4 Error Analysis

In this section, we look into the errors produced by our model in predicting the argumentation persuasiveness, where three types of major errors are observed.

Error Type I: Wrong separation of the topic words and discourse words. The errors occur in distinguishing topic and discourse words may result in erroneous persuasiveness prediction results. For example, as shown in Figure 5, the word “*cognition*” should be considered as a topic word yet erroneously inferred to reflect discourse. Because of the cascading failure, the model output might be affected.

Error Type II: Preconception held by opinion holders. Sometimes opinion holders hold preconception towards the debating subject and their views are difficult to be changed by others. As shown in Figure 6, the opinion holder raised an issue related to the “*abortion ban act*”. Although the challengers provide arguments

there several benefits come from learning second (or third or more) language beyond ability converse new language specifically language learning has been shown measurable cognitive benefits ; there research indicates [* those who spoke two or more languages had significantly better cognitive abilities compared what would been expected from their baseline test *](<url>). effects present those who learn new language young or old * another study found [* language - learning participants ended up with increased density their grey matter their white matter tissue had been *](<url>) can also read [study](<url>) more information

<quote> common comparison made learning music as / u / has pointed out did some research into matter seems learning musical instrument does [long - lasting benefits](<url>) relate * higher - order aspects cognition * cannot find study compares benefits language learning vs musical training back this up my intuition benefits from language learning would greater (more widely available) since humans wired language music on other hand skill everyone manages pick up no matter how hard they practice so degree difficulty would make language - learning better prospect those looking increase cognitive benefits from learning new skills please note my point about language learning focused on cognitive benefits necessarily increased capacity communicate with others

<quote> * * possible quantify mental benefits read study linked learn about benefits language learning music training benefits also documented couldn't find head - - head comparison two my * feeling * language learning would easier see benefits from <quote> easy say when talking young people benefits increased grey matter density later onset dementia which isn big concern people 're talking they don value something could significantly improve their cognition makes them ignorant enlightened edit [spelling

apologies do currently access full versions articles linked above quick search other sources [digit >](<url>) , [digit >](<url>) & ; [digit >](<url>) . most interesting study [this one](<url>) , can find complete version sorry note [study # < digit >] has an exceptionally small sample size still interesting reading

Figure 4: Visualization of the topic-discourse assignment of CMV conversation in Figure 1. The annotated blue words are prone to be discourse words, and the red are topic words. The shade is the word-level confidence of current assignment.

with concrete evidences against the *OP*, they fail to obtain a Δ due to the opinion holder's preconception. Such cases are prominent on social media, posing a challenge to understand opinion holders' prior beliefs for a better prediction of argumentation outcome.

Error Type III: Lack of knowledge for judging the sufficiency of the evidence. In the court scenario, successful debates depend on how the lawyers make use of their persuasive skills to present their evidence or interpret their opponents' evidence. The judgement of the evidence sufficiency is beyond what the current model can capture. The logic and sufficiency of evidence could not be easily determined without external knowledge, e.g., law terminology and clauses. In future work, we will strengthen the reasoning process of the model by incorporating external knowledge sources.

7 DISCUSSION

In Section 6, we have shown the superior performance of our proposed model to identify persuasiveness arguments. Here, we discuss how the latent topics and discourse signal argumentation outcome. From our results, we further draw three suggestions, which might help individuals better engage in argumentative dialogues.

7.1 The Roles of Topics and Discourse in Argumentation Process

In Section 6.1, topics have shown slightly stronger effects on successful persuasion than discourse. Here we further analyze their roles in affecting persuasion outcome. Similar trends are observed on both datasets and we only discuss the results on CMV dataset due to the space limitation.

To investigate topic effects, we follow Wang et al. [43] to identify *strong* argument topics when the topic likelihood is larger than a pre-defined threshold (set to 0.2 here).⁸ Then in Figure 7(a), we show how the number of strong argument topics distribute over winning arguments compared with the losing ones. For discourse, we similarly show the discourse factor distributions on winning and losing arguments in Figure 7(b). Here we display the discourse

⁸To set the threshold, we first sample 100 arguments and manually align them to the strongly related latent topic. Then we set the threshold resulting in the smallest errors compared with human annotations.

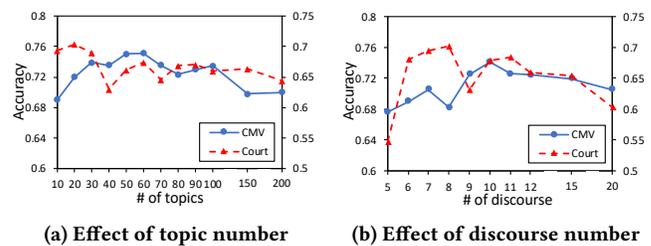


Figure 5: The impact of topic number (a) and discourse number (b) on our model for persuasiveness prediction. For both (a) and (b), the blue and solid line shows the results on CMV with left vertical axis, and the red and dashed line Court with the right vertical axis.

Title: Abortion is not a woman's rights issue ...
OP: I have never met a person who said they were against abortion because they did not think woman should have autonomy over their own body. ...
Winning Side: ... When one right is a person's right to be alive, I think the woman's right is unimportant. ...
Losing Side: ... Abortion is fundamentally a "prioritizing" of rights. The rights of the fetus vs. the rights of the woman. ...
 <quote> No. We can't know what the world would be like if men could get pregnant. ...

Figure 6: An example of inaccurate prediction caused by the preconception of the opinion holder. The losing side put forward stronger argument and far more evidences compared to the winning side (15 vs. 5 sentences). We only show partial arguments here due to the space constraint.

factors with our interpretations on the discourse styles according to their associated word distributions. In the following we discuss the findings from topics and discourse in turn.

Topic Roles. As can be seen in Figure 7(a), the winning side tends to put forward fewer topics in the argumentative process. This indicates that strong and focused argument points are more closely related to successful arguments than diverse topics, because arguing with too many things might overwhelm the opinion holder,

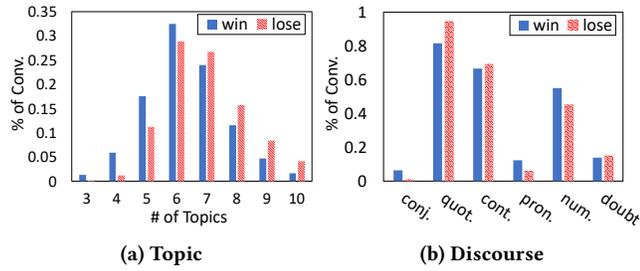


Figure 7: Distributions of winning and losing persuasion over the number of *strong* argument topics involved in (a) and varying discourse factors in (b). For (b), we display the discourse factors with our interpretation (“conj.”-conjunction, “quot.”-quotation, “cont.”-contrast, “pron.”-personal pronoun, and “num.”-statistic). Two-sided Mann-Whitney rank test shows that the two distributions are significantly different for both sides ($p < 0.01$).

which may lead to the persuasion failure. Similar trend can also be observed in Court dataset.

Discourse Roles. From Figure 7(b), we can see discourse styles vary in their effects over the persuasiveness results. Specifically, personal pronoun and statistic are more likely to appear in the winning side than the losing side. Their positive effects have also been previously reported [37, 43]. Moreover, we find that conjunction, though not widely used, is obviously more endorsed by winning sides. The benefit of conjunction may be related with the better logic it renders. For the losing side, they are more in favor of the quotation discourse, which is used in CMV to quote and attack others’ weak points. People may dislike such criticism, which renders the negative impact on persuasiveness.

7.2 Discourse Effects over Turn Number

To provide more insights, we further study the change of discourse effects over argumentation process with varying conversation length (the number of turns). Here we focus on discourse effects instead of topics because discourse styles are commonly used in diverse arguments and exhibit shared patterns on the two dataset, while topics vary in different scenarios. Specifically, we investigate the effects of the example discourse (listed in Table 4) over argumentation processes with varying turn number. The persuasiveness scores (computed with Eq. 8) are employed to measure the discourse effects and the results on the two datasets are displayed in Figure 8. Here comes our observations.

First, we find that in general all the discourse styles exhibit a decreasing trend in terms of persuasiveness scores with more argumentative turns coming in, although they appear to be more important in the initial few turns. Second, same discourse style may demonstrate varying persuasiveness impacts in different debate scenarios. For example, the pronouns usually express more personal emotion and tend to arouse empathy from others. Such discourse shows a positive effect on debates in the social media scenario, as shown in Figure 8 (a), but its effectiveness in the Court scenario is less apparent. A similar observation can be found for the quotations. Finally, we also observe that on the Court dataset, various discourse styles exhibit very similar effects on the persuasion, as shown in

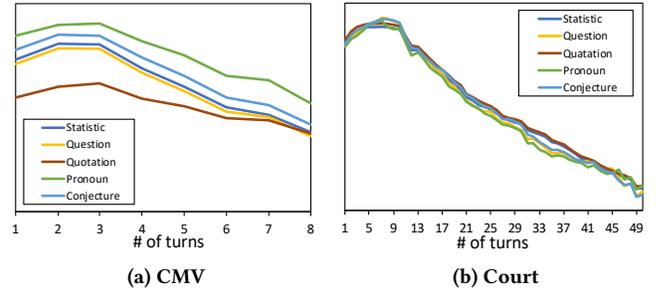


Figure 8: Effects of the example discourse styles (in Table 4) on the ultimate persuasiveness as the argumentation process continues. The horizontal axis indicates the number of argumentative turns, and the vertical axis the dynamic persuasiveness score (given by Eq. 8).

Figure 8 (b), while in CMV, the effects of various discourse styles can be easily distinguished.

7.3 Case Study

Our DTDMN is designed for capturing the topic shifting and discourse flow in an argumentation conversation, which allows us to interpret argument persuasiveness from the perspectives of topic and discourse dynamics. Here we take the CMV discussion in Figure 1 as an example to look into its persuasion process. Recall that the challengers put forward viewpoints centered around “*the advantage to learn a second language*”, and they successfully change the opinion holder’s mind with good arguments delivered. In Figure 9(a), we visualize the dynamic memory weights w^t (see Eq. 3) for each turn. It is observed that our model highlights the ‘*cognition*’ topic factor, which suggests the cognitive research evidence (e.g., learning a musical instrument) might help challengers win. For discourse, the model highlights latent factors represented by words like ‘*url*’, ‘*digits*’, and ‘*more*’. This suggests that effective discourse styles, such as quotation of links (‘*url*’) and statistic (‘*digits*’), may also play an important role in persuasiveness.

To further study how each topic and discourse alone contributes to this example’s persuasion, we disable the effects from other topics and discourse via masking w^t , and map the prediction score y in Eq. 8 to $[0, 1]$ range. We visualize the prediction scores in Figure 9(b) to depict the effect of persuasiveness from each topic and discourse. We observe that the “*cognition*” topic is still highlighted for all turns. It implies our model still recognizes this topic to be important, without taking the discourse effects into account. For discourse, we notice that the *quotation* and *statistic* skills are considered useful for the first few turns, whose impacts however later change to be negative. It might be because people tend to be tired of excessive URL links and statistics without providing more insightful opinions.

7.4 Suggestions on Argumentation

From the results, we draw some general suggestions on argumentation, which might help participants behave better in debates.

Topics are more important than discourse styles. In an argumentative conversation, opponents attempt to establish the validity of two positions by convincing each other and trying to win points

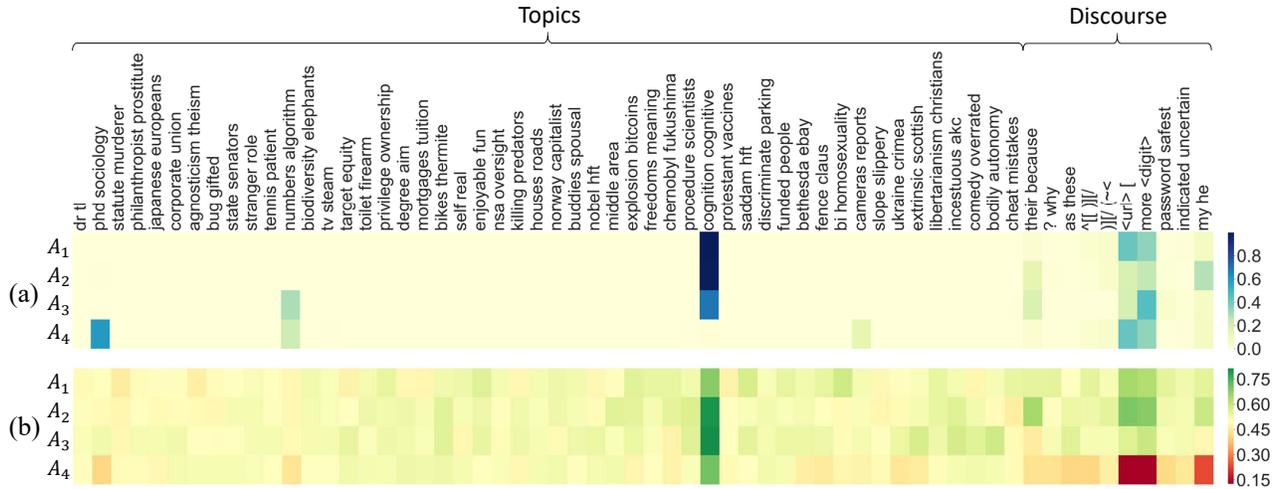


Figure 9: The heatmap visualizing the dynamic memory weights and persuasiveness on topic and discourse factors for the conversation in Figure 1. The vertical axis shows the turn id (from A_1 to A_4), and horizontal axis shows the latent topics and discourse displayed with their top 2 words. (a) Dynamic memory weights w^t that indicate topics shift and discourse flow. (b) Persuasiveness effect from each topic or discourse. For (a) and (b), darker colors indicate higher impacts. For (b), green indicate positive impacts while red negative.

in the debate [36]. Our study shows that topics contribute slightly more on persuasiveness than discourse. It happens especially in later stage of the argumentation process, which is suggested by the decreasing effects of discourse over turn number (see Figure 8). This is consistent with the discovery in Van Dijk et al. [39], which points out that style and rhetoric are not the dominant factors to determine debate outcome.

Strong and focused argument points are better than diverse topics. Strong arguments that are well-supported with evidence and/or reasoning, generally deliver more persuasive messages to audience [5]. Our study reveals that successful argumentation usually conveys fewer and focused topics. Diverse topics could only distract audience and expose more vulnerable points to the opponent.

Well organize the points and address them in a modest and concrete way. Argument discourse represents the cultural and situational realities of human reasoning, and is more sensitive to audience in conversational debates [10]. Amossy [1] also claims that argumentativity constitutes an inherent feature of discourse. This advice works particularly well in social media arguments, where the amateur debaters from general public are likely to be affected by opponents’ discourse skills. As a result, we see that personal pronoun (modest), statistic (concrete), and conjunction (well-organized) discourse are more likely to appear in winning side.

7.5 Limitations of Our Study

In this paper, we use the CMV dataset following the previous augmentation mining setting [14, 17, 26, 37]. In them, only the CMV dataset is used in evaluation. To better evaluate the generalization performance, we also include the Supreme Court dataset, which exhibits different data statistics from CMV (e.g., fewer argumentation processes and more turns involved in a process). However, it might not guarantee the generalization capability of our proposed model on other argumentation genres. Also, our findings are drawn from

the experimental results of the CMV and the Court datasets. These findings are consistent with prior studies in social linguistics, and provide some additional details. To further evaluate if our model and empirical results are applicable in other scenarios, more experimental study is required on a diverse range of debate data to better understand human arguments.

In addition, we mainly consider the topic and discourse factors in the modeling of the argumentation process. There are other factors that may relate to the persuasiveness of an argumentative conversation, such as age [12], culture [38], gender [16] of participants. For example, earlier research [11] on argumentation suggests that adults use advanced discourse strategies more consistently, frequently, and flexibly than adolescents do. Due to the unavailability of such metadata in our datasets, we could not easily incorporate these factors. Future research can consider building debate datasets with side information such as demographics data included.

8 CONCLUSION AND FUTURE WORK

In this paper, we propose to dynamically track both topics and discourse factors in conversational argumentation for persuasiveness prediction. The proposed neural model not only identifies persuasive arguments more accurately, but also provides insights into the usefulness of topics and discourse for a successful persuasion. The findings concluded in this paper can facilitate the argument persuasiveness analysis.

ACKNOWLEDGEMENTS

The work described in this paper is supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (No. CUHK 14210717 of the General Research Fund and No. CUHK 2410021 of the Research Impact Fund R5034-18). Jing Li is supported by the Hong Kong Polytechnic University Internal Fund (1-BE2W). YH is partly funded by the EPSRC grant EP/T017112/1.

REFERENCES

- [1] Ruth Amossy. 2009. Argumentation in discourse: A socio-discursive approach to arguments. *Informal Logic* 29, 3 (2009), 252–267.
- [2] David Atkinson, Kumar Bhargav Srinivasan, and Chenhao Tan. 2019. What Gets Echoed? Understanding the ‘Pointers’ in Explanations of Persuasive Arguments. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, EMNLP 2019, Hong Kong, China, November 3-7, 2019*.
- [3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- [4] Amparo Elizabeth Cano Basave and Yulan He. 2016. A Study of the Impact of Persuasive Argumentation in Political Debates. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*. 1405–1413.
- [5] William L Benoit. 1987. Argumentation and credibility appeals in persuasion. *Southern Journal of Communication* 52, 2 (1987), 181–197.
- [6] Winston Carlile, Nishant Gurrupadi, Zixuan Ke, and Vincent Ng. 2018. Give Me More Feedback: Annotating Argument Persuasiveness and Related Attributes in Student Essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*. 621–631.
- [7] Rich Caruana, Steve Lawrence, and C. Lee Giles. 2000. Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping. In *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems, NIPS 2000 (Denver, CO, USA)*. 402–408.
- [8] Ananlada Chotimongkol. 2008. *Learning the structure of task-oriented conversations from the corpus of in-domain dialogs*. Ph.D. Dissertation. Carnegie Mellon University, Language Technologies Institute, School of Computer Science.
- [9] Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon M. Kleinberg. 2012. Echoes of power: language effects and power differences in social interaction. In *Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012*. 699–708.
- [10] Donald G Ellis. 2015. Argument Discourse. *The International Encyclopedia of Language and Social Interaction* (2015), 1–6.
- [11] Mark Felton and Deanna Kuhn. 2001. The development of argumentative discourse skill. *Discourse processes* 32, 2-3 (2001), 135–153.
- [12] Mark K Felton. 2004. The development of discourse strategies in adolescent argumentation. *Cognitive Development* 19, 1 (2004), 35–52.
- [13] Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.
- [14] Christopher Hidey and Kathleen R. McKeown. 2018. Persuasive Influence Detection: The Role of Argument Sequencing. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. 5173–5180.
- [15] Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the Semantic Types of Claims and Premises in an Online Persuasive Forum. In *Proceedings of the 4th Workshop on Argument Mining, ArgMining@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*. 11–21.
- [16] Allan Jeong and Gayle V. Davidson-Shivers. 2006. The Effects of Gender Interaction Patterns on Student Participation in Computer-Supported Collaborative Argumentation. *Educational Technology Research and Development* 54, 6 (01 Dec 2006), 543–568.
- [17] Lu Ji, Zhongyu Wei, Xiangkun Hu, Yang Liu, Qi Zhang, and Xuanjing Huang. 2018. Incorporating Argument-Level Interactions for Persuasion Comments Evaluation using Co-attention Model. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*. 3703–3714.
- [18] Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A Latent Variable Recurrent Neural Network for Discourse Relation Language Models. *CoRR abs/1603.01913* (2016).
- [19] Yohan Jo, Shivani Poddar, Byungsoo Jeon, Qinlan Shen, Carolyn Penstein 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, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*. 103–116.
- [20] Shafiq R. Joty, Giuseppe Carenini, and Chin-Yew Lin. 2011. Unsupervised Modeling of Dialog Acts in Asynchronous Conversations. In *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*. 1807–1813.
- [21] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR abs/1412.6980* (2014).
- [22] Diederik P. Kingma and Max Welling. 2013. Auto-Encoding Variational Bayes. *CoRR abs/1312.6114* (2013). arXiv:1312.6114
- [23] Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Roman Paulus, and Richard Socher. 2016. Ask Me Anything: Dynamic Memory Networks for Natural Language Processing. In *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*. 1378–1387.
- [24] Mark R Leary, Kate J Diebels, Erin K Davison, Katrina P Jongman-Sereno, Jennifer C Isherwood, Kaitlin T Raimi, Samantha A Deffler, and Rick H Hoyle. 2017. Cognitive and interpersonal features of intellectual humility. *Personality and Social Psychology Bulletin* 43, 6 (2017), 793–813.
- [25] Jing Li, Yan Song, Zhongyu Wei, and Kam-Fai Wong. 2018. A Joint Model of Conversational Discourse and Latent Topics on Microblogs. *Computational Linguistics* 44, 4 (2018).
- [26] Jian-Fu Lin, Kuo Yu Huang, Hen-Hsen Huang, and Hsin-Hsi Chen. 2019. Lexicon Guided Attentive Neural Network Model for Argument Mining. In *Proceedings of the 6th Workshop on Argument Mining, ArgMining@ACL 2019, Florence, Italy, August 1, 2019*. 67–73.
- [27] Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*. 1412–1421.
- [28] Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering Discrete Latent Topics with Neural Variational Inference. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*. 2410–2419.
- [29] Vlad Niculae, Joonsuk Park, and Claire Cardie. 2017. Argument Mining with Structured SVMs and RNNs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*. 985–995.
- [30] Isaac Persing and Vincent Ng. 2017. Why Can’t You Convince Me? Modeling Weaknesses in Unpersuasive Arguments. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*. 4082–4088.
- [31] Peter Potash, Alexey Romanov, and Anna Rumshisky. 2016. Here’s My Point: Argumentation Mining with Pointer Networks. *CoRR abs/1612.08994* (2016).
- [32] Kechen Qin, Lu Wang, and Joseph Kim. 2017. Joint Modeling of Content and Discourse Relations in Dialogues. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*. 974–984.
- [33] Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised Modeling of Twitter Conversations. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA*. 172–180.
- [34] Weiyang Shi, Tiancheng Zhao, and Zhou Yu. 2019. Unsupervised Dialog Structure Learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. 1797–1807.
- [35] Christian Stab and Iryna Gurevych. 2014. Identifying Argumentative Discourse Structures in Persuasive Essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. 46–56.
- [36] Nancy L Stein and Elizabeth R Albro. 2001. The origins and nature of arguments: Studies in conflict understanding, emotion, and negotiation. *Discourse processes* 32, 2-3 (2001), 113–133.
- [37] 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, WWW 2016, Montreal, Canada, April 11 - 15, 2016*. 613–624.
- [38] Sonja Tirkkonen-Condit. 1996. Explicitness vs. implicitness of argumentation: An intercultural comparison.
- [39] Teun Adrianus Van Dijk, Walter Kintsch, and Teun Adrianus Van Dijk. 1983. Strategies of discourse comprehension. (1983).
- [40] Frans H Van Eemeren, Rob Grootendorst, Ralph H Johnson, Christian Plantin, and Charles A Willard. 2013. *Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments*. Routledge.
- [41] Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017. Argumentation Quality Assessment: Theory vs. Practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*. 250–255.
- [42] Marilyn A. Walker, Pranav Anand, Stephanie M. Lukin, and Steve Whittaker. 2017. Argument Strength is in the Eye of the Beholder: Audience Effects in Persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EAACL 2017, Valencia, Spain, April 3-7,*

- 2017, *Volume 1: Long Papers*. 742–753.
- [43] Lu Wang, Nick Beauchamp, Sarah Shugars, and Kechen Qin. 2017. Winning on the Merits: The Joint Effects of Content and Style on Debate Outcomes. *TACL 5 (2017)*, 219–232.
- [44] 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, ACL 2016, August 7–12, 2016, Berlin, Germany, Volume 2: Short Papers*.
- [45] Charles Arthur Willard. 1996. *Liberalism and the problem of knowledge: A new rhetoric for modern democracy*. University of Chicago Press.
- [46] Caiming Xiong, Stephen Merity, and Richard Socher. 2016. Dynamic Memory Networks for Visual and Textual Question Answering. In *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19–24, 2016*. 2397–2406.
- [47] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6–11 July 2015*. 2048–2057.
- [48] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. Hierarchical Attention Networks for Document Classification. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12–17, 2016*. 1480–1489.
- [49] Jichuan Zeng, Jing Li, Yulan He, Cuiyun Gao, Michael R. Lyu, and Irwin King. 2019. What You Say and How You Say it: Joint Modeling of Topics and Discourse in Microblog Conversations. *CoRR abs/1903.07319 (2019)*.
- [50] Jichuan Zeng, Jing Li, Yulan He, Cuiyun Gao, Michael R. Lyu, and Irwin King. 2019. What You Say and How You Say it: Joint Modeling of Topics and Discourse in Microblog Conversations. *Arxiv Preprint abs/1903.07319 (2019)*.
- [51] Jichuan Zeng, Jing Li, Yan Song, Cuiyun Gao, Michael R. Lyu, and Irwin King. 2018. Topic Memory Networks for Short Text Classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*. 3120–3131.
- [52] Ke Zhai and Jason D. Williams. 2014. Discovering Latent Structure in Task-Oriented Dialogues. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22–27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*. 36–46.
- [53] Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational Flow in Oxford-style Debates. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12–17, 2016*. 136–141.
- [54] Jiani Zhang, Xingjian Shi, Irwin King, and Dit-Yan Yeung. 2017. Dynamic Key-Value Memory Networks for Knowledge Tracing. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3–7, 2017*. 765–774.