Weakly-Supervised Hierarchical Models for Predicting Persuasion Strategies

Jiaao Chen, Diyi Yang
School of Interactive Computing
Georgia Institute of Technology
{jchen896,dyang888}@gatech.edu

Abstract

Modeling persuasive language has the potential to better facilitate our decision-making processes. Despite its importance, computational modeling of persuasion is still in its infancy, largely due to the lack of benchmark datasets that can provide quantitative labels of persuasive strategies to expedite this line of research. To this end, we introduce a large-scale multi-domain text corpus for modeling persuasive strategies in good-faith text requests. Moreover, we design a hierarchical weakly-supervised latent variable model that can leverage partially labeled data to predict such associated persuasive strategies for each sentence, where the supervision comes from both the overall document-level labels and very limited sentence-level labels. Experimental results showed that our proposed method outperformed existing semi-supervised baselines significantly. We have publicly released out code at https://github.com/GT-SALT/Persuasion_Strategy_WVAE.

Introduction

Persuasive communication has the potential to bring significant positive and pro-social factors to our society (Hovland, Janis, and Kelly 1971). For instance, persuasion could largely help fundraising for charities and philanthropic organizations or convincing substance-abusing family members to seek professional help. Given the nature of persuasion, it is of great importance to study how and why persuasion works in language. Modeling persuasive language is challenging in the field of natural language understanding since it is difficult to quantify the persuasiveness of requests and even harder to generalize persuasive strategies learned from one domain to another. Although researchers from social psychology have offered useful advice on how to understand persuasion, most of them have been conducted from a qualitative perspective (Bartels 2006; Popkin and Popkin 1994). Computational modeling of persuasion is still in its infancy, largely due to the lack of benchmarks that can provide unified, representative corpus to facilitate this line of research, with a few exceptions like (Luu, Tan, and Smith 2019b; Atkinson, Srinivasan, and Tan 2019; Wang et al. 2019).

Most existing datasets concerning persuasive text are either (1) too small (e.g., in the order of hundreds) for current machine learning models (Yang et al. 2019) or (2) not representative for understanding persuasive strategies by only looking at one specific domain (Wang et al. 2019). To make persuasion research and technology maximally useful, both for practical use and scientific study, a generic and representative corpus is a must, which can represent persuasive language in a way that is not exclusively tailored to any one specific dataset or platform. To fill these gaps, building on theoretical work on persuasion and these prior empirical studies, we first introduce **a set of generic persuasive strategies and a multi-domain corpus** to understand different persuasion strategies that people use in their requests for different types of persuasion goals in various domains.

However, constructing a large-scale dataset that contains persuasive strategies labels is often time-consuming and expensive. To mitigate the cost of labeling fine-grained sentence persuasive strategy, we then introduce **a simple but effective weakly-supervised hierarchical latent variable model** that leverages mainly global or document-level labels (e.g., overall persuasiveness of the textual requests) alongside with limited sentence annotations to predict sentence-level persuasion strategies. Our work is inspired by prior work (Oquab et al. 2015) in computer vision that used the global image-level labels to classify local objects. Intuitively, our model is hierarchically semi-supervised, with sentence-level latent variables to reconstruct the input sentence and all latent variables of sentences are aggregated to predict document-level persuasiveness. Specifically, at the sentence-level, we utilize two latent variables representing persuasion strategies and context separately, in order to disentangle information pertaining to label-oriented and content-specific properties to do reconstructions; at the document level, we encode those two latent variables together to predict the overall document labels in the hope that it could supervise the learning of sentence-level persuasive strategies. To sum up, our contributions include:

1. A set of generic persuasive strategies based on theoretical and empirical studies and introducing a relatively large-scale dataset that includes annotations of persuasive strategies for three domains.
2. A hierarchical weakly-supervised latent variable model to predict persuasive strategies with partially labeled data.
3. Extensive experimental results that test the effectiveness

of our models and visualize the importance of our proposed persuasion strategies.

Related Work

There has been much attention paid to computational persuasive language understanding (Guo, Zhang, and Singh 2020; Atkinson, Srinivasan, and Tan 2019; Lukin et al. 2017; Yang and Kraut 2017; Shaikh et al. 2020). For instance, Tan et al. (2016) looked at how the interaction dynamics such as the language interplay between opinion holders and other participants predict the persuasiveness via ChangeMyView subreddit. Althoff, Danescu-Niculescu-Mizil, and Jurafsky (2014) studied donations in Random Acts of Pizza on Reddit, using the social relations between recipient and donor plus linguistic factors like narratives to predict the success of these altruistic requests. Although prior work offered predictive and insightful models, most research determined their persuasion labels or variables without reference to a taxonomy of persuasion techniques. Yang et al. (2019) identified the persuasive strategies employed in each sentence among textual requests from crowdfunding websites in a semi-supervised manner. Wang et al. (2019) looked at utterance in persuasive dialogues and annotated a corpus with different persuasion strategies such as self-modeling, foot-in-the-door, credibility, etc., together with classifiers to predict such strategies at a sentence-level. These work mainly focused on a small subset of persuasion strategies and the identification of such strategies in a specific context. Inspired by those work, we propose a generic and representative set of persuasion strategies to capture various persuasion strategies that people use in their requests.

Recently many semi-supervised learning approaches have been developed for natural language processing, including adversarial training (Miyato, Dai, and Goodfellow 2016), variational auto-encoders (Kingma et al. 2014; Yang et al. 2017; Gururangan et al. 2019), consistency training (Xie et al. 2020; Chen, Wu, and Yang 2020; Chen, Yang, and Yang 2020) and various pre-training techniques (Kiros et al. 2015; Dai and Le 2015). The contextual word representations (Peters et al. 2018; Devlin et al. 2019) have emerged as powerful mechanisms to make use of large scale unlabeled data. Most of these prior works focus on semi-supervised learning, in which the labels are partially available and the supervisions for labeled and unlabeled data are both on the sentence-levels. In contrast, our work is hierarchical weakly supervised and we aim to predict **sentence-levels labels, not document-level persuasiveness**. To our best knowledge, weakly supervised learning has been explored much less in natural language processing except for a few recent work (Lee, Chang, and Toutanova 2019; Min et al. 2019) in question answering. There are a few exceptions: Yang et al. (2019) utilized a small amount of hand-labeled sentences together with a large number of requests automatically labeled at the document level for text classification. Pryzant, Chung, and Jurafsky (2017) proposed an adversarial objective to learn text features highly predictive of advertisement outcomes. Our work has an analog task in computer vision–weakly supervised image segmentation (Papandreou et al. 2015; Pinheiro and Collobert 2015)– which uses image labels or bounding boxes information to predict pixel-level labels. Similar to image segmentation, obtaining global/document/image level labels for persuasive understanding is much cheaper than local/sentence/pixel level labels. Different from multi-task learning where models have full supervisions in each task, our proposed model is fully supervised at the document level while partially supervised at the sentence level.

Persuasion Taxonomy and Corpus

Previous work modeling persuasion in language either focus on a small subset of strategies or look at a specific platform, hard to be adapted to other contexts. To fill this gap, we propose a set of generic persuasive strategies based on widely used persuasion models from social psychology. Specifically, we leverage Petty and Cacioppo's elaboration likelihood model (1986) and Chiaken's social information processing model (Chaiken 1980), which suggest that people process information in two ways: either performing a relatively deep analysis of the quality of an argument or relying on some simple superficial cues to make decisions (Cialdini 2001). Guided by these psychology insights, we examine the aforementioned computational studies on persuasion and argumentation (Wang et al. 2019; Yang et al. 2019; Durmus, Cardie, and Durmus 2018; Vargheese, Collinson, and Masthoff 2020a; Carlile et al. 2018), and further synthesize these theoretical and practical tactics into eight unified categories: *Commitment, Emotion, Politeness, Reciprocity, Scarcity* that allow people to use simple inferential rules to make decisions, and *Credibility, Evidence, Impact* that require people to evaluate the information based on its merits, logic, and importance. As shown in Table 1, our taxonomy **distilled, extended**, and **unified** existing persuasion strategies. Different from prior work that introduced domain-specific persuasion tactics with limited generalizability, our generic taxonomy can be easily plugged into different text domains, making large-scale understanding of persuasion in language across multiple contexts comparable and replicable.

Dataset Collection & Statistics

We collected our data from three different domains related to persuasion. (1) **Kiva**[1] is a peer-to-peer philanthropic lending platform where persuading others to make loans is a key to success (no interest), (2) subreddit "r/Random_Acts_of_Pizza[2]" (**RAOP**) where members write requests to ask for free pizzas (social purpose, no direct money transaction), and (3) subreddit "r/borrow[3]" (**Borrow**) that focuses on writing posts to borrow money from others (with interest). After removing personal and sensitive information, we obtained 40,466 posts from Kiva, 18,026 posts from RAOP, and 49,855 posts from Borrow.

We sampled 5% documents with document length ranging from 1 to 6 from Kiva, 1 to 8 from RAOP and 1 to 7 from Borrow to annotate, as documents with at most 6 sentences account for 89% in Kiva, 86% posts in RAOP has no more

[1] www.kiva.org

[2] www.reddit.com/r/Random_Acts_Of_Pizza

[3] www.reddit.com/r/borrow

Strategy	Definition and Examples	Connection with Prior Work
Commitment	The persuaders indicating their intentions to take acts or justify their earlier decisions to convince others that they have made the correct choice. e.g., *I just lent to Auntie Fine's Donut Shop.* (*Kiva*)	*Commitment* (Yang et al. 2019), *Self-modeling* (Wang et al. 2019), *Commitment* (Vargheese et al. 2020b)
Emotion	Making request full of emotional valence and arousal affect to influence others. e.g., *Guys I'm desperate.* (*Borrow*) *I've been in the lowest depressive state of my life.* (*RAOP*)	*Ethos* (Carlile et al. 2018), *Emotion appeal* (Carlile et al. 2018), *Sentiment* (Durmus et al. 2018), *Emotion words* (Luu et al, 2019a), *Emotion* (Asai et al. 2020)
Politeness	The usage of polite language in requests. e.g., *Your help is deeply appreciated!* (*Borrow*)	*Politeness* (Durmus et al. 2018), *Politeness* (Althoff et al. 2014), *Politeness* (Nashruddin et al. 2020)
Reciprocity	Responding to a positive action with another positive action. People are more likely to help if they have received help themselves. e.g., *I will pay 5% interest no later than May 1, 2016.* (*Borrow*) *I'll pay it forward with my first check.* (*RAOP*)	*Reciprocity* (Althoff et al. 2014), *Reciprocity* (Roethke et al. 2020), *Reciprocity* (Vargheese et al. 2020b)
Scarcity	People emphasizing on the urgency, rare of their needs. e.g., *Need this loan urgently.* (*Borrow*) *I haven't ate a meal in two days.* (*RAOP*) *Loan expiring today and still needs $650.* (*Kiva*)	*Scarcity* (Vargheese et al. 2020b), *Scarcity* (Yang et al. 2019), *Scarcity* (Lawson et al. 2020)
Credibility	The uses of credentials impacts to establish credibility and earn others' trust. e.g., *Can provide any documentation needed.* (*Borrow*) *She has already repaid 2 previous loans.* (*Kiva*)	*Credibility appeal* (Wang et al. 2019), *Social Proof* (Roethke et al. 2020), *Social Proof* (Vargheese et al. 2020a)
Evidence	Providing concrete facts or evidence for the narrative or request. e.g. *My insurance was canceled today.* (*Borrow*) *There is a Pizza Hut and a Dominos near me.* (*RAOP*) *$225 to go and 1 A+ member on the loan.* (*Kiva*)	*Evidentiality* (Althoff et al. 2014), *Evidence* (Carlile et al. 2018), *Evidence* (Stab and Gurevych 2014), *Concreteness* (Yang et al. 2019) *Evidence* (Durmus et al. 2018)
Impact	Emphasizing the importance or impact of the request. e.g., *I will use this loan to pay my rent.* (*Borrow*) *This loan will help him with his business.* (*Kiva*)	*Logos* (Carlile et al. 2018), *Logic appeal* (Wang et al. 2019) *Impact* (Yang et al. 2019)

Table 1: The generic taxonomy of persuasive strategies, their definitions, example sentences, and connections with prior work.

than 8 sentences, and 85% posts in Borrow has at most 7 sentences. We recruited four research assistants to label persuasion strategies for each sentence in sampled documents. Definitions and examples of different persuasion strategies were provided, together with a training session where we asked annotators to annotate a number of example sentences and walked them through any disagreed annotations. To assess the reliability of the annotated labels, we then asked them to annotate the same 100 documents with 400 sentences and computed Cohen's Kappa coefficient to measure inter-rater reliability. We obtained an average score of 0.538 on Kiva, 0.613 on RAOP, and 0.623 on Borrow, which indicates moderate agreement (McHugh 2012). Annotators then annotated the rest 1200 documents by themselves independently.

The dataset statistics are shown in Table 2, and the sentence-level label distribution in each dataset is shown in Figure 1. We merge rare strategies into the Other category. Specifically, we merge Commitment, Scarcity, and Emotion in Borrow, Credibility and Commitment in RAOP, Reciprocity and Emotion in Kiva, as Other. We utilized whether the requester received pizzas or loans from the subreddits as the document-level labels for RAOP and Borrow. 30.1% of people successfully got pizzas on RAOP and 48.5% of people received loans on Borrow. In Kiva, we utilized the number of people who lent loans as the document-level labels. The numbers are further labeled based on buckets: $[1, 2)$, $[2, 3)$, $[3, 4)$, $[4, \infty)$, accounting for 44.1%, 20.3%, 12.4% and 33.2% of all documents.

Method

To alleviate the dependencies on labeled data, we propose a hierarchical weakly-supervised latent variable model to leverage partially labeled data to predict sentence-level persuasive strategies. Specifically, we introduce a sentence-level latent variable model to reconstruct the input sentence and predict the sentence-level persuasion labels spontaneously, supervised by the global or document-level labels (e.g., overall persuasiveness of the documents). The overall architecture of our method is shown in Figure 2.

Weakly Supervised Latent Model

Given a corpus of N documents $\mathbf{D} = \{\mathbf{d}_i\}_{i=1}^N$, where each document $\mathbf{d}$ consists of M sentences $\mathbf{d}_i = \{\mathbf{s}_i^j\}_{j=1}^M$. For each document $\mathbf{d}_i \in \mathbf{D}$, its document level label is denoted as $\mathbf{t}_i$, representing the overall persuasiveness of the documents. We divide the corpus into two parts: $\mathbf{D} = \mathbf{D}_L \cup \mathbf{D}_U$, where $\mathbf{D}_L$ ($\mathbf{D}_U$) denotes the set of documents with (without) *sentence* labels. For each document $\mathbf{d}_i \in \mathbf{D}_L$, the corresponding sentence labels are $\{\mathbf{y}_i^j\}_{j=1}^M$, where $\mathbf{y}_i^j \in \mathbf{C} = \{c_k\}_{k=1}^K$ and represents the persuasive strategy of a given sentence. In many practical scenarios, getting document-level labels $\{\mathbf{t}_i\}$ is much easier and cheaper than the fine-

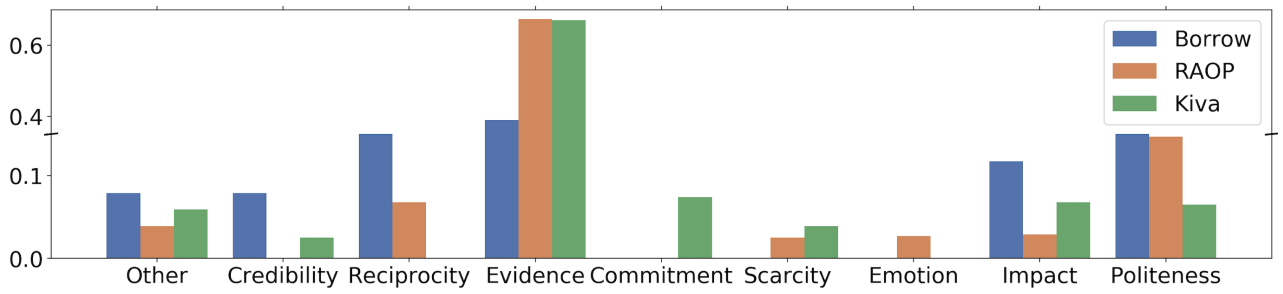

Figure 1: The distribution of each persuasion strategy in three annotated three datasets.

	# Docs	# Sents w/ label	# Sents w/o label	Doc Labels	Sent Labels
Borrow	49,855	5,800	164,293	Success or not	Evidence, Impact, Politeness, Reciprocity, Credibility
RAOP	18,026	3,600	77,517	Success or not	Evidence, Impact, Politeness, Reciprocity, Scarcity, Emotion
Kiva	40,466	6,300	135,330	# People loaned	Evidence, Impact, Politeness, Credibility, Scarcity, Commitment

Table 2: Dataset statistics. For strategies that are rare, we merged them into an *Other* category.

grained sentence labels $\{\mathbf{s}_i^j\}$ since the number of sentences M in a document $\mathbf{d}_i$ can be very large. Similarly, in our setting, the number of documents with fully labeled sentences is very limited, i.e., $|\mathbf{D}_L| \ll |\mathbf{D}|$. To this end, we introduce a novel hierarchical weakly supervised latent variable model that can leverage both the document-level labels and the small amount of sentence-level labels to discover the sentence persuasive strategies. Our model is **weakly supervised** since we will utilize document labels to facilitate the learning of sentence persuasive strategies. The intuition is that global documents labels of persuasiveness carry useful information of local sentence persuasive strategies, thus can provide supervision in an **indirect** manner.

Sentence Level VAE Following prior work on semi-supervised variational autoencoders (VAEs) (Kingma and Welling 2013), for an input sentence $\mathbf{s}$, we assume a graphical model whose latent representation contains a continuous vector $\mathbf{z}$, denoting the content of a sentence, and a discrete persuasive strategy label $\mathbf{y}$:

$$p(\mathbf{s}, \mathbf{z}, \mathbf{y}) = p(\mathbf{s}|\mathbf{z}, \mathbf{y})p(\mathbf{z})p(\mathbf{y}) \tag{1}$$

To learn the semi-supervised VAE, we optimize the variational lower bound as our learning objective. For unlabeled sentence, we maximize the evidence lower bound as:

$$\begin{aligned}\log p(\mathbf{s}) \geq \mathbb{E}_{\mathbf{y}\sim q(\mathbf{y}|\mathbf{s})}[\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{s},\mathbf{y})}[\log p(\mathbf{s}|\mathbf{z}, \mathbf{y})] \\ - \mathrm{KL}[q(\mathbf{z}|\mathbf{s}, \mathbf{y})||p(\mathbf{z})]] - \mathrm{KL}[q(\mathbf{y}|\mathbf{s})||p(\mathbf{y})]\end{aligned} \tag{2}$$

where $p(\mathbf{s}|\mathbf{y}, \mathbf{z})$ is a decoder (generative network) to reconstruct input sentences and $q(\mathbf{y}|\mathbf{s})$ is an encoder (an inference or a predictor network) to predict sentence-level labels.

For labeled sentences, the variational lower bound is:

$$\begin{aligned}\log p(\mathbf{s}, \mathbf{y}) \geq \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{s},\mathbf{y})}[\log p(\mathbf{s}|\mathbf{z}, \mathbf{y})] \\ - \mathrm{KL}[q(\mathbf{z}|\mathbf{s}, \mathbf{y})||p(\mathbf{z})] + \text{constant}\end{aligned} \tag{3}$$

In addition, for sentences with labels, we also update the inference network $q(\mathbf{y}|\mathbf{s})$ via minimizing the cross entropy loss $\mathbb{E}_{(\mathbf{s},\mathbf{y})}[-\log q(\mathbf{y}|\mathbf{s})]$ directly.

Document Level VAE Different from sentence-level VAEs, we model the input document $\mathbf{d}$ with sentences $\{\mathbf{s}^j\}_{j=1}^M = \mathbf{s}^{1:M}$ as a whole and assume that the document-level label $\mathbf{t}$ depends on the sentence-level latent variables. Thus we obtain the document-level VAE model as:

$$p(\mathbf{d}, \mathbf{t}, \mathbf{y}, \mathbf{z}) = p(\mathbf{d}, \mathbf{t}|\mathbf{y}, \mathbf{z}) \prod_{j=1}^{M} p(\mathbf{y}^j) \prod_{j=1}^{M} p(\mathbf{z}^j) \tag{4}$$

where $p(\mathbf{d}, \mathbf{t}|\mathbf{y}^{1:M}, \mathbf{z}^{1:M})$ is the generative model for all sentences in the document $\mathbf{d}$ and the document label $\mathbf{t}$.

For simplicity, we further assume conditional independence between the sentences $\mathbf{s}^{1:M}$ in $\mathbf{d}$ and its label $\mathbf{t}$ given the latent variables: $p(\mathbf{d}, \mathbf{t}|\mathbf{y}^{1:M}, \mathbf{z}^{1:M}) = p(\mathbf{t}|\mathbf{y}^{1:M}, \mathbf{z}^{1:M}) \prod_{j=1}^{M} p(\mathbf{s}^j|\mathbf{y}^j, \mathbf{z}^j)$. Since the possible number of the sentence label combinations is huge, simply computing the marginal probability becomes intractable. Thus we optimize the evidence lower bound. By using mean field approximation (Jain, Koehler, and Mossel 2018), we factorize the posterior distribution as: $q(\mathbf{z}^{1:M}, \mathbf{y}^{1:M}|\mathbf{d}, \mathbf{t}) = \prod_{j=1}^{M} q(\mathbf{z}^j|\mathbf{y}^j, \mathbf{s}^j, \mathbf{t}) \prod_{j=1}^{M} q(\mathbf{y}^j|\mathbf{s}^j, \mathbf{t})$. That is, the posterior distribution of latent variables $\mathbf{y}^j$ and $\mathbf{z}^j$ only depends on the sentence $\mathbf{s}^j$ and the document label $\mathbf{t}$. For documents without sentence labels, the evidence lower bound is:

$$\begin{aligned}\log p(\mathbf{d}, \mathbf{t}) \geq \mathbb{E}_{\mathbf{y}\sim q(\mathbf{y}|\mathbf{s},\mathbf{t})}[\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{s},\mathbf{y},\mathbf{t})}[\log p(\mathbf{t}|\mathbf{y}, \mathbf{z}) \\ + \sum_{i=1}^{N} \log p(\mathbf{s}^j|\mathbf{y}^j, \mathbf{z}^j)] - \sum_{j=1}^{M} \mathrm{KL}[q(\mathbf{z}^j|\mathbf{s}^j, \mathbf{y}^j, \mathbf{t})||p(\mathbf{z}^j)]] \\ - \sum_{j=1}^{M} \mathrm{KL}[q(\mathbf{y}^j|\mathbf{s}^j, \mathbf{t})||p(\mathbf{y}^j)] = U(\mathbf{d}, \mathbf{t})\end{aligned} \tag{5}$$

For document with sentence labels, the variational lower

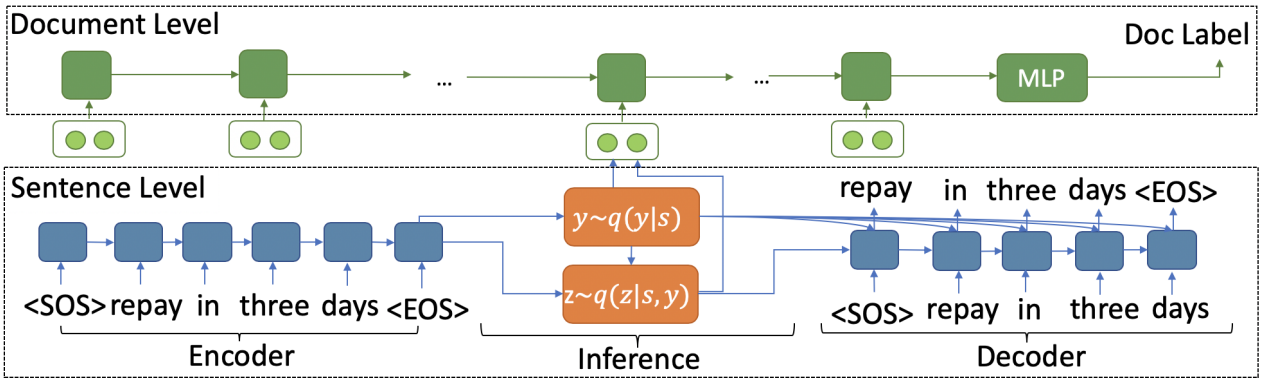

Figure 2: Overall architecture. At sentence-level, the input sentences are first encoded into two latent variables: y representing strategies and z containing context information; the decoder reconstructs the input sentences. At document-level, a predictor network aggregates the latent variables within the input document to predict document-level labels. For labeled documents, labels are directly used for the reconstruction and prediction; for unlabeled ones, latent variables y are used.

bound can be adapted from above as:

$$\begin{aligned}\log p(\mathbf{d},\mathbf{t},\mathbf{y}) &\geq \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{s},\mathbf{y},\mathbf{t})}[\log p(\mathbf{t}|\mathbf{y},\mathbf{z}) \\ &+ \sum_{i=1}^{N}\log p(\mathbf{s}^j|\mathbf{y}^j,\mathbf{z}^j)] - \sum_{j=1}^{M}\mathrm{KL}[q(\mathbf{z}^j|\mathbf{s}^j,\mathbf{y}^j,\mathbf{t})||p(\mathbf{z}^j)] \\ &= L(\mathbf{d},\mathbf{t},\mathbf{y}) + \text{constant}\end{aligned} \tag{6}$$

Combining the loss for document with and without sentence labels, we obtain the overall loss function:

$$\begin{aligned}L = \quad & \mathbb{E}_{\mathbf{d}\in\mathbf{D}_U}U(\mathbf{d},\mathbf{t}) + \mathbb{E}_{\mathbf{d}\in\mathbf{D}_L}L(\mathbf{d},\mathbf{t},\mathbf{y}^{1:M}) \\ & + \alpha\cdot\mathbb{E}_{\mathbf{d}\in\mathbf{D}_L}\prod_{j=1}^{M}\log q(\mathbf{y}^j|\mathbf{s}^j,\mathbf{t})\end{aligned} \tag{7}$$

Here, $\mathbb{E}_{\mathbf{d}\in\mathbf{D}_L}\prod_{j=1}^{M}\log q(\mathbf{y}^j|\mathbf{s}^j,\mathbf{t})$ represents the discriminative loss for sentences with labels and α controls the trade-off between generative loss and discriminative loss [4].

Compared to sentence-level VAE (S-VAE) that only learns sentence representation via a generative network $p(\mathbf{s}|\mathbf{y},\mathbf{z})$, document-level VAE utilizes the contextual relations between sentences by aggregating multiple sentences in a document and further predicting document-level labels via a predictor network $p(\mathbf{t}|\mathbf{y}^{1:M},\mathbf{z}^{1:M})$. Document-level weakly supervised VAE (WS-VAE) incorporates both direct sentence-level supervision and indirect document-level supervision to better make use of unlabeled sentences, thus can further help the persuasion strategies classification. Note that our hierarchical weakly-supervised latent variable model presents a generic framework to utilize dependencies between sentence-level and document-level labels, and can be easily adapted to other NLP tasks where document-level supervision is rich and sentence-level supervision is scarce.

Training Details

In practice, we parameterize the inference network $q(\mathbf{y}|\mathbf{s},\mathbf{t})$ and $q(\mathbf{z}|\mathbf{y},\mathbf{s},\mathbf{t})$ using a LSTM or a BERT which encodes the sentences (and document label) to get the posterior distribution. We used another LSTM as the decoder to model the the generative network $p(\mathbf{s}|\mathbf{z},\mathbf{y})$. At the document level, each sentence's content vector and strategy vector is fed as input to a LSTM to model the predictor network $p(\mathbf{t}|\mathbf{z}^{1:M},\mathbf{y}^{1:M})$.

Reparametrization: It is challenging to back-propagate through random variables as it involves non-differentiable sampling procedures. For latent variable $\mathbf{z}$, we utilized the reparametrization technique proposed by Kingma and Welling (2013) to re-parametrize the Gaussian random variable $\mathbf{z}$ as $\mu + \sigma\epsilon$, where $\epsilon \sim N(0,I)$, μ and σ are deterministic and differentiable. For discrete latent variable $\mathbf{y}$, we adopted Gumbel softmax (Jang, Gu, and Poole 2017) to approximate it continuously:

$$y_k = \frac{\exp\left(\left(\log\left(\pi_k\right)+g_k\right)/\tau\right)}{\sum_{k=1}^{K}\exp\left(\left(\log\left(\pi_k\right)+g_k\right)/\tau\right)}$$

where $\pi_{1:K}$ are the probabilities of a categorical distribution, g_k follows Gumbel$(0,1)$ and τ is the temperature. The approximation is accurate when $\tau \to 0$ and smooth when $\tau > 0$. We gradually decrease τ in the training process.

Prior Estimation: Classical variational models usually assume simple priors such as uniform distributions. We performed a Gaussian kernel density estimation over training data to estimate the prior for $\mathbf{y}$, and assumed the latent variable z follows a standard Gaussian distribution.

Dataset	Train	Dev	Test
Borrow	900	400	400
RAOP	300	200	300
Kiva	1000	400	400

Table 3: Split statistics about train, dev, and test set.

Experiment and Result

Experiment Setup: We randomly sampled from the labeled documents to form the maximum labeled train set, the development, and test set to train and evaluate models, and we

[4] The influence of α is discussed in Section 4 in Appendix.

Dataset	Model	Sentence-level Persuasion Strategy Prediction F1 Score				Doc-Level Accuracy
		20	50	100	Max	
Kiva	LSTM	26.1 ± 0.8	37.6 ± 1.0	43.3 ± 1.0	54.6 ± 2.0	-
	SH-Net	29.1 ± 0.4	38.8 ± 0.9	43.4 ± 0.9	54.8 ± 0.9	34.8 ± 1.0
	BERT	28.6 ± 4.0	38.5 ± 0.7	44.6 ± 3.0	57.0 ± 1.0	-
	S-VAE	30.9 ± 1.0	40.3 ± 0.7	43.6 ± 0.9	55.7 ± 1.0	-
	WS-VAE	31.5 ± 0.8	40.9 ± 1.0	44.0 ± 1.0	55.4 ± 0.8	35.5 ± 1.0
	WS-VAE-BERT	$\mathbf{34.2 \pm 0.2}$	$\mathbf{43.0 \pm 0.9}$	$\mathbf{45.2 \pm 0.9}$	$\mathbf{59.1 \pm 0.9}$	$\mathbf{36.7 \pm 2.0}$
RAOP	LSTM	28.5 ± 1.0	37.7 ± 1.0	42.5 ± 1.0	47.8 ± 0.9	-
	SH-Net	30.0 ± 1.0	39.1 ± 1.0	42.8 ± 1.0	48.1 ± 1.0	66.6 ± 1.0
	BERT	30.6 ± 2.0	39.5 ± 2.0	43.4 ± 2.0	54.0 ± 1.0	-
	S-VAE	31.7 ± 0.7	40.1 ± 1.0	43.2 ± 1.0	48.8 ± 2.0	-
	WS-VAE	32.1 ± 0.9	39.9 ± 0.9	43.8 ± 0.9	49.1 ± 2.0	65.3 ± 1.0
	WS-VAE-BERT	$\mathbf{41.0 \pm 0.8}$	$\mathbf{45.6 \pm 2.0}$	$\mathbf{51.2 \pm 0.8}$	$\mathbf{58.3 \pm 2.0}$	$\mathbf{67.8 \pm 1.0}$
Borrow	LSTM	53.4 ± 0.9	62.6 ± 0.9	68.1 ± 0.8	74.4 ± 2.0	-
	SH-Net	53.7 ± 1.0	63.2 ± 1.0	68.0 ± 0.7	74.5 ± 1.0	56.5 ± 2.0
	BERT	56.7 ± 1.0	64.1 ± 3.0	68.5 ± 1.0	74.6 ± 0.4	-
	S-VAE	59.2 ± 0.7	65.3 ± 0.4	68.8 ± 0.6	74.6 ± 0.5	-
	WS-VAE	59.5 ± 1.0	66.0 ± 0.7	68.9 ± 1.0	74.7 ± 0.3	56.5 ± 0.9
	WS-VAE-BERT	$\mathbf{62.6 \pm 2.0}$	$\mathbf{68.5 \pm 1.0}$	$\mathbf{70.4 \pm 1.0}$	$\mathbf{75.9 \pm 0.7}$	$\mathbf{57.5 \pm 0.8}$

Table 4: Sentence-level persuasion strategy prediction performance (Macro F1 Score) and document-level prediction performance (Accuracy). Models are trained with documents amount of 20 (81 sentences in Kiva, 99 sentences in RAOP and 59 sentences in Borrow), 50 (200 sentences in Kiva, 236 sentences in RAOP and 168 sentences in Borrow), 100 (355 sentences in Kiva, 480 sentences in RAOP and 356 sentences in Borrow), and all the training set (3512 sentences in Kiva, 1382 sentences in RAOP and 3136 sentences in Borrow). The results are averaged after 5 different runs, with the 95% confidence interval.

utilized all the unlabeled documents as training data as well. The data splits are shown in Table 3. We utilized NLTK (Bird, Klein, and Loper 2009) to split the documents into sentences and tokenize each sentence with BERT-base uncased tokenizer (Devlin et al. 2019). We added a special CLS token at the beginning of each sentence and a special SEP token at the end of each sentence. We used BERT (Devlin et al. 2019) as the discriminative network, LSTM as the generative network and predictor network. The inference network is a 2-layer MLP. We trained our model via AdamW (Loshchilov and Hutter 2017) and tuned hyper-parameters on the development set.

Baselines and Model Settings[5]

We compared our model on strategy classification for each sentence with several baselines: (1) **LSTM** (Hochreiter and Schmidhuber 1997): LSTM is utilized as the encoder for sentences. We use the last layer's hidden states as the representations of sentences to classify the persuasion strategies. Only labeled sentences are used here. (2) **SH-Net** (Yang et al. 2019): SH-Net utilized a hierarchical LSTM to classify strategies with the supervision from both sentence-level and document-level labels, thus both labeled documents and unlabeled documents being used. We followed their implementation and modified the document-level inputs as concatenations of latent variables y and z. (3) **BERT** (Devlin et al. 2019): We used the pre-trained BERT-base uncased model and fine-tuned it for the persuasion strategy classification. BERT only utilized labeled sentences. (4) **S-VAE**: Sentence-level VAE applied variational autoencoders in classifications by reconstructing the input sentences while learning to classify them. Both labeled and unlabeled sentences are used.

WS-VAE denotes our proposed weakly supervised latent variable model that made use of sentence-level labels and document-level labels at the same time, as well as reconstructing input documents. We further showed that our proposed WS-VAE model is orthogonal to pre-trained models like BERT as well by utilizing pre-trained BERT as the discriminative network to encode the input sentences and then using 2-layer LSTM as the generative network and predictor network, denoted as **WS-VAE-BERT**, a special case (based on pre-trained transformer models) of WS-VAE.

Results

Varying the Number of Labeled Documents We tested the models with varying amount of labeled documents from 20 to the maximum number of labeled training documents, and summarized the results in Table 4. The simple LSTM classifier showed the worst performance over three datasets, especially when limited labeled documents were given. After simply adding document-level supervision as well as unlabeled documents, SH-Net got better Macro F1 scores as well as lower variance, showing the impact of document-level supervision on sentence-level learning. BERT fine-tuned on persuasion strategy classification tasks showed better performance than LSTM and SH-Net with limited labeled data in most cases.

By leveraging the reconstruction of each input sentence using corresponding persuasion strategies and context latent variables, S-VAE showed a significant performance boost

[5]Parameters details are stated in Section 5 in the Appendix.

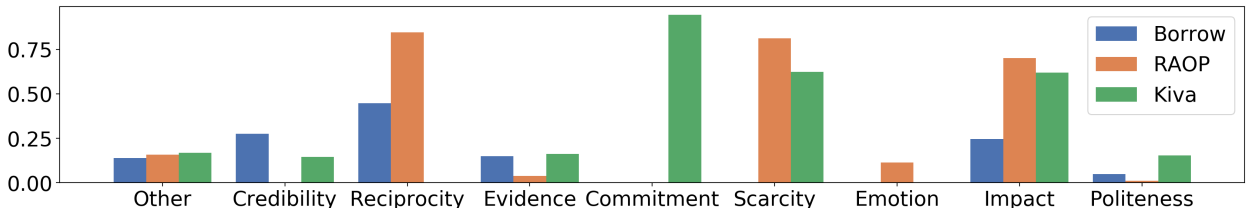

Figure 3: Average attention weight learned in the predictor network for different strategies in three datasets.

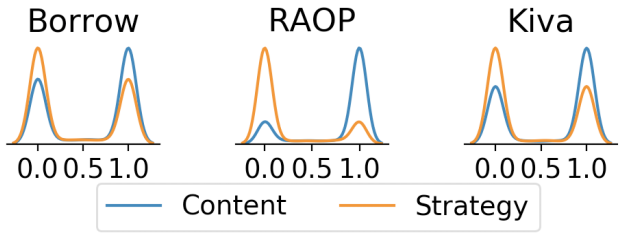

Figure 4: Attention weight for content vectors and strategy vectors when predicting document-level labels in the predictor network.

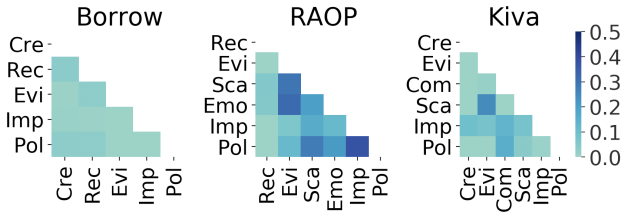

Figure 5: Cosine similarities between different persuasive strategies (**Cre**dibility, **Rec**iprocity, **Evi**dence, **Com**mitment, **Sca**rcity, **Emo**tion, **Imp**act and **Pol**iteness).

comparing to only utilizing indirect supervision from the document-level labels. This indicated that by incorporating the extra supervision directly from the input sentence itself, we can gain more help than hierarchical supervision from document levels. By utilizing the hierarchical latent variable model, which not only utilized the sentence reconstruction but also document-level predictions to assist the sentence-level classifications, WS-VAE outperformed S-VAE. When combining with the state-of-the-art pre-trained models like BERT, our WS-VAE-BERT achieved the best performance over three datasets. This suggests that such improvement does not only come from large pre-trained models, but also the incorporation of our hierarchical latent variable model.

Note that we also showed the document-level prediction accuracy for models that used all the labeled documents. Even though the document-level predictions were not our goals, we observed a consistent trend that higher document-level performance correlated with the higher sentence-level accuracy, suggesting that the global document-level supervision helped the sentence-level predictions.

Importance of Strategies vs Content To better understand how these persuasive strategies and the text content jointly affect the success of text requests, we added an attention layer over content latent variable z and strategy latent variable y in the predictor network to visualize the importance of persuasive strategies and text content in the WS-VAE-BERT, as shown in Figure 4. In all three domains, we found that content vectors tend to have larger weights than strategy vectors. This suggests that when people are writing requests to convince others to take action, content is relatively the more important component than persuasion strategies. However, leveraging proper persuasive strategies can further boost the likelihood of their requests being fulfilled.

Attention Weight We further calculated the average attention weights learned in the predictor network (attended over strategy latent variable y and content latent variable z to predict the document-level labels) for different strategies in three datasets which is shown in Figure 3. We observed that *Reciprocity*, *Commitment*, *Scarcity* and *Impact* seemed to play more important roles, while *Credibility*, *Evidence*, *Emotion* and *Politeness* had lower average attention weights, which indicated that simple superficial strategies might be more influential to overall persuasiveness in online forums than strategies that required deeper analysis.

Relation between Persuasive Strategies To explore possible relations among different persuasive strategies, we utilized the embeddings for each persuasive strategy from the predictor network and visualized their pairwise similarities in Figure 5. All the similarities scores were below 0.5, showing those strategies in our taxonomy are generally orthogonal to each other and capture different aspects of persuasive language. However, some strategies tend to demonstrate relatively higher relations; for example, *Scarcity* highly correlates with *Evidence* on RAOP and Kiva, indicating that people may often use them together in their requests.

Conclusion and Future Work

This work introduced a set of generic persuasive strategies based on theories on persuasion, together with a large-scale multi-domain text corpus annotated with their associated persuasion strategies. To further utilize both labeled and unlabeled data in real-world scenarios, we designed a hierarchical weakly-supervised latent variable model to utilize document-level persuasiveness supervision to guide the learning of specific sentence-level persuasive strategies. Experimental results showed that our proposed method outperformed existing semi-supervised baselines significantly on three datasets. Note that, we made an assumption that

the document-level persuasiveness label only depended on the sentence-level information. However there are other factors closely related to the overall persuasiveness such as requesters/lenders' backgrounds or their prior interactions (Valeiras-Jurado 2020; Longpre, Durmus, and Cardie 2019). Future work can investigate how these audience factors further affect the predictions of both sentence- and document-level labels. As an initial effort, our latent variable methods disentangle persuasion strategies and the content, and highlight the relations between persuasion strategies and the overall persuasiveness, which can be further leveraged by real-world applications to make textual requests more effective via different choices of persuasion strategies.

Acknowledgment

We would like to thank Jintong Jiang, Leyuan Pan, Yuwei Wu, Zichao Yang, the anonymous reviewers, and the members of Georgia Tech SALT group for their feedback. We acknowledge the support of NVIDIA Corporation with the donation of GPU used for this research. DY is supported in part by a grant from Google.

References

Althoff, T.; Danescu-Niculescu-Mizil, C.; and Jurafsky, D. 2014. How to Ask for a Favor: A Case Study on the Success of Altruistic Requests. In *Proceedings of ICWSM*.

Asai, S.; Yoshino, K.; Shinagawa, S.; Sakti, S.; and Nakamura, S. 2020. Emotional Speech Corpus for Persuasive Dialogue System. In *Proceedings of The 12th Language Resources and Evaluation Conference*, 491–497. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.62.

Atkinson, D.; Srinivasan, K. B.; and Tan, C. 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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2904–2914.

Bartels, L. M. 2006. Priming and persuasion in presidential campaigns. *Capturing campaign effects* 78–112.

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition. ISBN 0596516495, 9780596516499.

Carlile, W.; Gurrapadi, N.; Ke, Z.; and Ng, V. 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 (Volume 1: Long Papers)*, 621–631. Melbourne, Australia: Association for Computational Linguistics. doi:10.18653/v1/P18-1058. URL https://www.aclweb.org/anthology/P18-1058.

Chaiken, S. 1980. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of personality and social psychology* 39(5): 752.

Chen, J.; Wu, Y.; and Yang, D. 2020. Semi-supervised Models via Data Augmentation for Classifying Interactive Affective Responses. In *AffCon@AAAI*.

Chen, J.; Yang, Z.; and Yang, D. 2020. MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2147–2157. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.194. URL https://www.aclweb.org/anthology/2020.acl-main.194.

Cialdini, R. 2001. 6 principles of persuasion. *Arizona State University, eBrand Media Publication* .

Dai, A. M.; and Le, Q. V. 2015. Semi-supervised sequence learning. In *Advances in neural information processing systems*, 3079–3087.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Durmus, E.; Cardie, C.; and Durmus, E. 2018. Exploring the Role of Prior Beliefs for Argument Persuasion. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1035–1045. New Orleans, Louisiana: Association for Computational Linguistics. doi:10.18653/v1/N18-1094. URL https://www.aclweb.org/anthology/N18-1094.

Guo, Z.; Zhang, Z.; and Singh, M. 2020. In Opinion Holders' Shoes: Modeling Cumulative Influence for View Change in Online Argumentation. In *Proceedings of The Web Conference 2020*, 2388–2399.

Gururangan, S.; Dang, T.; Card, D.; and Smith, N. A. 2019. Variational Pretraining for Semi-supervised Text Classification. *arXiv preprint arXiv:1906.02242* .

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Comput.* 9(8): 1735–1780. ISSN 0899-7667. doi:10.1162/neco.1997.9.8.1735. URL http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Hovland, C. I.; Janis, I. L.; and Kelly, H. 1971. Communication and persuasion. *Attitude change* 66–80.

Jain, V.; Koehler, F.; and Mossel, E. 2018. The Mean-Field Approximation: Information Inequalities, Algorithms, and Complexity. *CoRR* abs/1802.06126. URL http://arxiv.org/abs/1802.06126.

Jang, E.; Gu, S.; and Poole, B. 2017. Categorical Reparameterization with Gumbel-Softmax. URL https://arxiv.org/abs/1611.01144.

Kingma, D. P.; Mohamed, S.; Rezende, D. J.; and Welling, M. 2014. Semi-supervised learning with deep generative

models. In *Advances in neural information processing systems*, 3581–3589.

Kingma, D. P.; and Welling, M. 2013. Auto-Encoding Variational Bayes. URL http://arxiv.org/abs/1312.6114. Cite arxiv:1312.6114.

Kiros, R.; Zhu, Y.; Salakhutdinov, R. R.; Zemel, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, 3294–3302.

Lawson, P.; Pearson, C. J.; Crowson, A.; and Mayhorn, C. B. 2020. Email phishing and signal detection: How persuasion principles and personality influence response patterns and accuracy. *Applied Ergonomics* 86: 103084.

Lee, K.; Chang, M.-W.; and Toutanova, K. 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering. *arXiv preprint arXiv:1906.00300* .

Longpre, L.; Durmus, E.; and Cardie, C. 2019. Persuasion of the Undecided: Language vs. the Listener. In *Proceedings of the 6th Workshop on Argument Mining*, 167–176.

Loshchilov, I.; and Hutter, F. 2017. Fixing Weight Decay Regularization in Adam. *CoRR* abs/1711.05101. URL http://arxiv.org/abs/1711.05101.

Lukin, S.; Anand, P.; Walker, M.; and Whittaker, S. 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: Volume 1, Long Papers*, 742–753.

Luu, K.; Tan, C.; and Smith, N. 2019a. Measuring Online Debaters' Persuasive Skill from Text over Time. *Transactions of the Association for Computational Linguistics* 7(0): 537–550. ISSN 2307-387X. URL https://transacl.org/index.php/tacl/article/view/1639.

Luu, K.; Tan, C.; and Smith, N. A. 2019b. Measuring Online Debaters' Persuasive Skill from Text over Time. *Transactions of the Association for Computational Linguistics* 7: 537–550.

McHugh, M. 2012. Interrater reliability: The kappa statistic. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB* 22: 276–82. doi:10.11613/BM.2012.031.

Min, S.; Chen, D.; Hajishirzi, H.; and Zettlemoyer, L. 2019. A discrete hard em approach for weakly supervised question answering. *arXiv preprint arXiv:1909.04849* .

Miyato, T.; Dai, A. M.; and Goodfellow, I. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725* .

Nashruddin, N.; Alam, F. A.; and Harun, A. 2020. Moral Values Found in Linguistic Politeness Patterns of Bugis Society. *Edumaspul: Jurnal Pendidikan* 4(1): 132–141.

Oquab, M.; Bottou, L.; Laptev, I.; and Sivic, J. 2015. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 685–694.

Papandreou, G.; Chen, L.-C.; Murphy, K. P.; and Yuille, A. L. 2015. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, 1742–1750.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* .

Petty, R. E.; and Cacioppo, J. T. 1986. The elaboration likelihood model of persuasion. In *Communication and persuasion*, 1–24. Springer.

Pinheiro, P. O.; and Collobert, R. 2015. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1713–1721.

Popkin, S. L.; and Popkin, S. L. 1994. *The reasoning voter: Communication and persuasion in presidential campaigns*. University of Chicago Press.

Pryzant, R.; Chung, Y.; and Jurafsky, D. 2017. Predicting Sales from the Language of Product Descriptions. In *eCOM@ SIGIR*.

Roethke, K.; Klumpe, J.; Adam, M.; and Benlian, A. 2020. Social influence tactics in e-commerce onboarding: The role of social proof and reciprocity in affecting user registrations. *Decision Support Systems* 131: 113268. ISSN 0167-9236. doi:https://doi.org/10.1016/j.dss.2020.113268. URL http://www.sciencedirect.com/science/article/pii/S0167923620300233.

Shaikh, O.; Chen, J.; Saad-Falcon, J.; Chau, D. H.; and Yang, D. 2020. Examining the Ordering of Rhetorical Strategies in Persuasive Requests. *Findings of EMNLP* .

Stab, C.; and Gurevych, I. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 46–56.

Tan, C.; Niculae, V.; Danescu-Niculescu-Mizil, C.; and Lee, L. 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 '16, 613–624. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-4143-1.

Valeiras-Jurado, J. 2020. Genre-specific persuasion in oral presentations: Adaptation to the audience through multimodal persuasive strategies. *International Journal of Applied Linguistics* 30(2): 293–312.

Vargheese, J. P.; Collinson, M.; and Masthoff, J. 2020a. Exploring susceptibility measures to persuasion. In *International Conference on Persuasive Technology*, 16–29. Springer.

Vargheese, J. P.; Collinson, M.; and Masthoff, J. 2020b. Exploring Susceptibility Measures to Persuasion. In Gram-Hansen, S. B.; Jonasen, T. S.; and Midden, C., eds., *Persuasive Technology. Designing for Future Change*, 16–29.

Cham: Springer International Publishing. ISBN 978-3-030-45712-9.

Wang, X.; Shi, W.; Kim, R.; Oh, Y.; Yang, S.; Zhang, J.; and Yu, Z. 2019. Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good. *arXiv preprint arXiv:1906.06725* .

Xie, Q.; Dai, Z.; Hovy, E.; Luong, M.-T.; and Le, Q. V. 2020. Unsupervised Data Augmentation for Consistency Training. URL https://openreview.net/forum?id=ByeL1R4FvS.

Yang, D.; Chen, J.; Yang, Z.; Jurafsky, D.; and Hovy, E. 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: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3620–3630.

Yang, D.; and Kraut, R. E. 2017. Persuading teammates to give: Systematic versus heuristic cues for soliciting loans. *Proceedings of the ACM on Human-Computer Interaction* 1: 114.

Yang, Z.; Hu, Z.; Salakhutdinov, R.; and Berg-Kirkpatrick, T. 2017. Improved variational autoencoders for text modeling using dilated convolutions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3881–3890. JMLR. org.

Appendix

Dataset & Annotation Details

In different contexts, people tend to write documents with different numbers of sentences, which might be associated with different sets of persuasion strategies.

The mean and std for number of sentences per document are 4.68 and 4.63 in Borrow, 5.10 and 4.40 in RAOP, and 3.83 and 4.12 in Kiva.

We recruited two graduate and two undergraduate students to label the persuasion strategies for each sentence in given documents which were randomly sampled from the whole corpus. Definitions and examples of different persuasion strategies were provided to the annotators. We also conducted a training session where we asked annotators to annotate 50 example sentences and walked through them any disagreements or confusions they had. Annotators then annotated 1200 documents by themselves independently.

To assess the reliability of the annotated labels, the same set of documents which contained 100 documents with 400 sentences was given to annotators to label and we computed the Cohen's Kappa coefficient. We obtained an average score of 0.538 on Kiva, 0.613 on RAOP and 0.623 on Borrow, which indicated moderate agreement and reasonable annotation quality (McHugh 2012).

WS-VAE

Sentence level VAE Based on prior work on semi-supervised VAEs (Kingma and Welling 2013), for an input sentence $\mathbf{s}$, we assume a graphical model whose latent representation contains a continuous vector $\mathbf{z}$, denoting the content of a sentence, and a discrete persuasive strategy label $\mathbf{y}$:

$$p(\mathbf{s}, \mathbf{z}, \mathbf{y}) = p(\mathbf{s}|\mathbf{z}, \mathbf{y})p(\mathbf{z})p(\mathbf{y}).$$

To learn the semi-supervised VAE, we optimize the variational lower bound as our learning objective. For unlabeled sentence, we maximize:

$$\begin{aligned}\log p(\mathbf{s}) &= \log \mathbb{E}_{\mathbf{y}\sim p(\mathbf{y})}\mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}[p(\mathbf{s}|\mathbf{z}, \mathbf{y})] \\ &\geq \mathbb{E}_{\mathbf{y}\sim q(\mathbf{y}|\mathbf{s})}[\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{s},\mathbf{y})}[\log p(\mathbf{s}|\mathbf{z}, \mathbf{y})] \\ &\quad - \mathrm{KL}[q(\mathbf{z}|\mathbf{s}, \mathbf{y})||p(\mathbf{z})]] \\ &\quad - \mathrm{KL}[q(\mathbf{y}|\mathbf{s})||p(\mathbf{y})],\end{aligned}$$

where $p(\mathbf{s}|\mathbf{y}, \mathbf{z})$ is a decoder (generative network) to reconstruct input sentences and $q(\mathbf{y}|\mathbf{s})$ is an encoder (an inference or a predictor network) to predict sentence-level labels. For labeled sentences, the variational lower bound becomes:

$$\begin{aligned}\log p(\mathbf{s}, \mathbf{y}) &= \log \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}[p(\mathbf{s}|\mathbf{z}, \mathbf{y})p(\mathbf{y})] \\ &\geq \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{s},\mathbf{y})}[\log p(\mathbf{s}|\mathbf{z}, \mathbf{y})] \\ &\quad - \mathrm{KL}[q(\mathbf{z}|\mathbf{s}, \mathbf{y})||p(\mathbf{z})] + \text{constant}\end{aligned}$$

In addition, for sentences with labels, we also update the inference network $q(\mathbf{y}|\mathbf{s})$ via minimizing the cross entropy loss $\mathbb{E}_{(\mathbf{s},\mathbf{y})}[-\log q(\mathbf{y}|\mathbf{s})]$ directly.

Document level VAE Different from sentence-level VAEs, we model the input document $\mathbf{d}$ with sentences $\{\mathbf{s}^j\}_{j=1}^M = \mathbf{s}^{1:M}$ as a whole and assume that the document-level label $\mathbf{t}$ depends on the sentence-level latent variables. Thus we obtain the document-level VAE model as:

$$\begin{aligned}&p(\mathbf{d}, \mathbf{t}, \mathbf{y}^{1:M}, \mathbf{z}^{1:M}) = \\ &\quad p(\mathbf{d}, \mathbf{t}|\mathbf{y}^{1:M}, \mathbf{z}^{1:M}) \prod_{j=1}^{M} p(\mathbf{y}^j) \prod_{j=1}^{M} p(\mathbf{z}^j),\end{aligned}$$

where $p(\mathbf{d}, \mathbf{t}|\mathbf{y}^{1:M}, \mathbf{z}^{1:M})$ is the generative model for all sentences in the document $\mathbf{d}$ and the document label $\mathbf{t}$. For simplicity, we further assume conditional independence between the sentences $\mathbf{s}^{1:M}$ in $\mathbf{d}$ and its label $\mathbf{t}$ given the latent variables:

$$\begin{aligned}&p(\mathbf{d}, \mathbf{t}|\mathbf{y}^{1:M}, \mathbf{z}^{1:M}) = \\ &\quad p(\mathbf{t}|\mathbf{y}^{1:M}, \mathbf{z}^{1:M}) \prod_{j=1}^{M} p(\mathbf{s}^j|\mathbf{y}^j, \mathbf{z}^j).\end{aligned}$$

Since the possible number of the sentence label combinations is huge, simply computing the marginal probability becomes intractable. Thus we optimize the evidence lower bound. By using mean field approximation (Jain, Koehler, and Mossel 2018), we factorize the posterior distribution as:

$$\begin{aligned}&q(\mathbf{z}^{1:M}, \mathbf{y}^{1:M}|\mathbf{d}, \mathbf{t}) \\ &\quad = q(\mathbf{z}^{1:M}|\mathbf{y}^{1:M}, \mathbf{s}^{1:M}, \mathbf{t})q(\mathbf{y}^{1:M}|\mathbf{s}^{1:M}, \mathbf{t}) \\ &\quad = \prod_{j=1}^{M} q(\mathbf{z}^j|\mathbf{y}^j, \mathbf{s}^j, \mathbf{t}) \prod_{j=1}^{M} q(\mathbf{y}^j|\mathbf{s}^j, \mathbf{t}),\end{aligned}$$

That is, the posterior distribution of latent variables $\mathbf{y}^j$ and $\mathbf{z}^j$ only depends on the sentence $\mathbf{s}^j$ and the document label $\mathbf{t}$. For documents without sentence labels, the variational lower bound $U(\mathbf{d}, \mathbf{t})$ is:

$$\begin{aligned}
\log p(\mathbf{d}, \mathbf{t}) &= \log \mathbb{E}_{\mathbf{y}\sim p(\mathbf{y})}\mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}[p(\mathbf{t}|\mathbf{z}^{1:M}, \mathbf{y}^{1:M}) \\
&\qquad \prod_{j=1}^{M} p(\mathbf{s}^j|\mathbf{z}^j, \mathbf{y}^j) \prod_{j=1}^{M} p(\mathbf{y}^j) \prod_{j=1}^{M} p(\mathbf{z}^j)] \\
&\geq \mathbb{E}_{\mathbf{y}^{1:M}\sim q(\mathbf{y}^{1:M}|\mathbf{s}^{1:M},\mathbf{t})}[\mathbb{E}_{\mathbf{z}^{1:M}\sim q(\mathbf{z}^{1:M}|\mathbf{s}^{1:M},\mathbf{y}^{1:M},\mathbf{t})} \\
&\quad [\log p(\mathbf{t}|\mathbf{y}^{1:M}, \mathbf{z}^{1:M}) + \sum_{i=1}^{N} \log p(\mathbf{s}^j|\mathbf{y}^j, \mathbf{z}^j)] \\
&\qquad - \sum_{j=1}^{M} \mathrm{KL}[q(\mathbf{z}^j|\mathbf{s}^j, \mathbf{y}^j, \mathbf{t})||p(\mathbf{z}^j)]] \\
&\qquad - \sum_{j=1}^{M} \mathrm{KL}[q(\mathbf{y}^j|\mathbf{s}^j, \mathbf{t})||p(\mathbf{y}^j)] \\
&\qquad = U(\mathbf{d}, \mathbf{t})
\end{aligned}$$

For document with sentence labels, the variational lower bound can be adapted from above as:

$$\begin{aligned}
&\log p(\mathbf{d}, \mathbf{t}, \mathbf{y}^{1:M}) \\
&= \log \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}[p(\mathbf{t}|\mathbf{z}^{1:M}, \mathbf{y}^{1:M}) \\
&\qquad \prod_{j=1}^{M} p(\mathbf{s}^j|\mathbf{z}^j, \mathbf{y}^j) \prod_{j=1}^{M} p(\mathbf{y}^j) \prod_{j=1}^{M} p(\mathbf{z}^j)] \\
&\geq \mathbb{E}_{\mathbf{z}^{1:M}\sim q(\mathbf{z}^{1:M}|\mathbf{s}^{1:M},\mathbf{y}^{1:M},\mathbf{t})} \\
&\quad [\log p(\mathbf{t}|\mathbf{y}^{1:M}, \mathbf{z}^{1:M}) + \sum_{i=1}^{N} \log p(\mathbf{s}^j|\mathbf{y}^j, \mathbf{z}^j)] \\
&\quad - \sum_{j=1}^{M} \mathrm{KL}[q(\mathbf{z}^j|\mathbf{s}^j, \mathbf{y}^j, \mathbf{t})||p(\mathbf{z}^j)] + \text{constant} \\
&\quad = L(\mathbf{d}, \mathbf{t}, \mathbf{y}^{1:M}) + \text{constant}
\end{aligned}$$

Combining the loss for document with and without sentence labels, we obtain the overall loss function:

$$\begin{aligned}
L = \quad & \mathbb{E}_{\mathbf{d}\in\mathbf{D}_U} U(\mathbf{d}, \mathbf{t}) + \mathbb{E}_{\mathbf{d}\in\mathbf{D}_L} L(\mathbf{d}, \mathbf{t}, \mathbf{y}^{1:M}) \\
& + \alpha \cdot \mathbb{E}_{\mathbf{d}\in\mathbf{D}_L} \prod_{j=1}^{M} \log q(\mathbf{y}^j|\mathbf{s}^j, \mathbf{t})
\end{aligned}$$

Here, $\mathbb{E}_{\mathbf{d}\in\mathbf{D}_L} \prod_{j=1}^{M} \log q(\mathbf{y}^j|\mathbf{s}^j, \mathbf{t})$ represents the discriminative loss for sentences with persuasive strategy labels and α controls the trade-off between generative loss and discriminative loss.

Threshold on KL Divergence

Yang et al. (2017) found that VAEs might easily get stuck in two local optimums: the KL term on $\mathbf{y}$ is very large and all samples collapse to one class or the KL term on $\mathbf{y}$ is very small and $q(\mathbf{y}|\mathbf{s})$ is close to the prior distribution. Thus we minimize the KL term only when it is larger than a threshold w:

$$\mathrm{KL}_{\mathbf{y}} = \mathbf{max}(w, \mathrm{KL}[q(\mathbf{y}|\mathbf{s})||p(\mathbf{y})])$$

Influence of the Trade-off Weight α

The overall loss function of our proposed weakly-supervised hierarchical latent variable model is:

$$\begin{aligned}
L = \quad & \mathbb{E}_{\mathbf{d}\in\mathbf{D}_U} U(\mathbf{d}, \mathbf{t}) + \mathbb{E}_{\mathbf{d}\in\mathbf{D}_L} L(\mathbf{d}, \mathbf{t}, \mathbf{y}^{1:M}) \\
& + \alpha \cdot \mathbb{E}_{\mathbf{d}\in\mathbf{D}_L} \prod_{j=1}^{M} \log q(\mathbf{y}^j|\mathbf{s}^j, \mathbf{t})
\end{aligned}$$

Here, the α is a parameter that controls the balance of reconstruction loss and supervised sentence classification loss. When α is small, the sentence level classifications are not well learned. When α is large, the model tends to only learn the sentence level classification tasks and ignore the reconstructions and document level predictions. In experiments, we set α to 5 through a grid search from the set $\{1, 5, 10, 20\}$.

Model Implementation Details

S-VAE

For **S-VAE** - the sentence-level latent variable model, which applies variational autoencoderes in sentence-level classifications by reconstructing the input sentences while learning to classify them, which encourages the model to assign input sentences to a label y such that the reconstruction loss is low. S-VAE is a special case (only performing operations at sentence levels) of our proposed WS-VAE. The weight for the reconstruction term is 1, the weight for the classification term is 5 and the weight for KL divergence terms are annealing from a small value to 1 through the training process. The learning rate is 0.001.

WS-VAE

For **WS-VAE** - our proposed weakly supervised latent variable model, takes advantage of sentence-level labels and document-level labels at the same time, as well as reconstructing input documents. The weight for the reconstruction term is 1, the weight for the classification term is 5, the weight for KL divergence terms are annealing from a small value to 1 through the training process, and the weight for predictor term is 0.5. The threshold for KL regularization on $q(y|s)$ is 1.2. The learning rate is 0.001.

WS-VAE-BERT

For **WS-VAE-BERT** - a special case (based on pre-trained transformer models) of WS-VAE, combines ES-VAE with recent pre-trained BERT. The weight for the reconstruction term is 1, the weight for the classification term is 5, the weight for KL divergence terms are annealing from a small value to 1 through the training process, and the weight for predictor term is 0.1. The threshold for KL regularization on $q(y|s)$ is 1.2. The learning rate is 0.00001.

Datasets	Threshold on y	Macro F1
Kiva	0	0.228
	1.2	0.315
	2.0	0.305
RAOP	0	0.274
	1.2	0.321
	2.0	0.316
Borrow	0	0.485
	1.2	0.595
	2.0	0.542

Table 5: Macro F1 Score with different threshold on y in KL regularization term for SH-VAE. Models are trained on three datasets with 20 labeled documents (81 sentences in kiva, 99 sentences in RAOP and 59 sentences in Borrow).

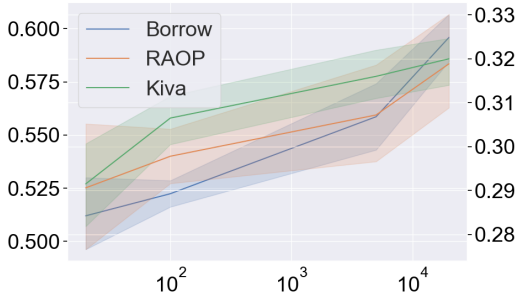

Figure 6: Macro F1 scores with 20 documents with sentence labels and different numbers of documents without sentence labels for WS-VAE. Results on Borrow follow the left y-axis, while RAOP and Kiva follow the right y-axis.

Impact of Variational Regularization

To show the importance of variational regularization on the latent variable y (the threshold on KL divergence w) mentioned in Section , we performed ablation study for the KL term for y. We tested WS-VAE with different values of threshold on three datasets using 20 labeled documents and the results were shown in Table 5. When the threshold is small like 0, which meant we added large regularization on y, the performance is bad because the $q(y|s)$ was so close to estimated prior distributions and barely learned from objective functions. When the threshold was large like 2, which meant there did not exist any regularization on y, we got lower F1 scores as well. When there is a appropriate threshold such as 1.2 to offer regularization, WS-VAE could achieve the best performance.

Varying the Number of Unlabeled Documents

We visualized WS-VAE's performances on three datasets when varying the amount of unlabeled data in Figure 6: macro F1 scores increased with more unlabeled data, demonstrating the effectiveness of the introduction of unlabeled sentences, and our hierarchical weakly-supervised model.