The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Persuasive Influence Detection: The Role of Argument Sequencing

Christopher Hidey
Department of Computer Science
Columbia University
New York, NY 10027
chidey@cs.columbia.edu

Kathleen McKeown
Department of Computer Science
Columbia University
New York, NY 10027
kathy@cs.columbia.edu

Abstract

Automatic detection of persuasion in online discussion is key to understanding how social media is used. Predicting persuasiveness is difficult, however, due to the need to model world knowledge, dialogue, and sequential reasoning. We focus on modeling the sequence of arguments in social media posts using neural models with embeddings for words, discourse relations, and semantic frames. We demonstrate significant improvement over prior work in detecting successful arguments. We also present an error analysis assessing novice human performance at predicting persuasiveness.

1 Introduction

Politicians and voters today are increasingly turning to social media to attract others to their cause. Identifying when a post will be influential would be helpful in understanding the appeal of political candidates and the reaction to current events and issues. A writer who is successful in changing the opinions of readers demonstrates *influence* over others and thus detecting persuasive posts that successfully change opinions is part of the overall solution to influence detection (Tan et al. 2016; Jaech et al. 2015).

Predicting persuasion is a difficult task as it requires modeling world knowledge, social interaction, and reasoning. Understanding the sequence of arguments used in online posts is crucial to understanding when a reader's mind has been changed. Empirically, there is evidence to suggest that people change their minds, and we provide evidence that this change is not just caused by new words and concepts but by the way these concepts are presented.

We conduct experiments on "Change My View", a specific "sub-reddit" of the Reddit social media platform, building on previous work using similar data (Tan et al. 2016; Wei, Liu, and Li 2016). "Change My View" (CMV) is a discussion forum where users post their opinions on a topic and their reasons for their beliefs. Other users respond by posting arguments attempting to change the view of the initiator of the discussion. If the views of the original posters are successfully changed, they will indicate this by posting a response with a "delta" character, providing naturally labeled data.

Consider the example in Table 1. In this discussion, the Reddit user "A" states her belief that borders between nations are just a social construct. The user "B" responds with her own argument that even though borders are not a natural occurrence, it is human nature to require this kind of organization. The original poster "A" then responds with a delta and acknowledges that she doesn't have a legitimate counter-argument. The overall structure of the argument is clear: the user begins by introducing evidence, making a concession as a matter of politeness, and finally concluding with a summarization and rhetorical questions.

In this paper, we show that the ordering of arguments is crucial to persuasion. We present a neural model of persuasive influence, modeling words, Penn Discourse Tree Bank (PDTB) relations, and FrameNet semantic frames. The main contributions of our work are: 1) statistically significant improvements over previous work on predicting persuasion by using features representing argument sequences and 2) experiments showing that we outperform novice humans on the same data, illustrating the difficulty of this task.

In the following sections we first discuss prior research in argumentation, persuasion, and influence in Section 2. We then present the Change My View data set and how we preprocess the data for different experiments (Section 3). Section 4 describes the experimental methodology, focusing on how we model the posts using a neural network. Finally we present the results of our experiments in Section 5 and provide an error analysis with respect to human judgments on the same task in Section 6. Code and data for our experiments is available to the research community.[1]

2 Related Work

Some sociolinguistic theorists suggest that persuasive and argumentative discourses are distinct but not disjoint (Nettel and Roque 2012). They assert that argumentation gives reasons in order to provide knowledge about a subject. Persuasion, however, attempts to convince, which may include other rhetorical devices such as emotionally moving the audience. Persuasive argumentation then has the joint goal of providing knowledge and convincing. In the Change My View dataset, all examples are persuasive argumentation, as their stated goal is to change the views of the original posters

[1] https://github.com/chridey/cmv

User	*Post*
Title	CMV: my view is that nations are just lines on a map and not real or useful
A	Nations are just lines on a map and don't exist in reality, here's my reasoning: 1) No one can decide where a nation begins or ends. Everyone's conception of "the South" when talking about America for example, will include different states and regions than the next person. In Europe, Turks claim that Cyprus is part of their nation, while Greece claims that island. Both claim Constantinople. Similarly, ...
B	https://en.wikipedia.org/wiki/Social_fact There is a word for what you are describing. While I'd concede your point is potentially valid, using your line of thinking makes living as a human being really difficult ... social facts make living in a human society possible in the first place. While they might be technically no true/real in a certain sense of the word, they provide structure in an otherwise structureless world. What's better? Have some orientationen, even though it's technically wrong. Or live without any kind of point of orientation, in a structureless world?
A	I'm going to give you a delta because you totally nailed it with the definition and your third paragraph raises points I can't answer: Δ

Table 1: Truncated Discussion Thread in Change My View

and one of the requirements for submissions is that the original posters state the reasoning for their views.

On the computational side, researchers studied the effect of influencers in social media discussions (Rosenthal and Mckeown 2017), where an influencer is a user who posts frequently, attempts to persuade, and is agreed with by others. Another type of influence studied is social power (Prabhakaran and Rambow 2013), which also distinguished dynamics like seniority or popularity. Recent work involved ranking of arguments from social media, attempting to objectively evaluate the quality of an argument posted in online forums (Habernal and Gurevych 2016b; 2016a).

Other work focused specifically on the Reddit social media site. Researchers modeled the rank of comments according to their "karma" score (a Reddit-specific method of rating posts) using linguistic and graph-based features (Jaech et al. 2015). Researchers have also analyzed specific subreddits, smaller communities within the larger Reddit population, such as "Change My View." Some research has focused on ranking comments (Wei, Liu, and Li 2016) while other research has involved predicting whether a post is persuasive (Tan et al. 2016) or identifying persuasive components of argumentation (Hidey et al. 2017).

In other work, researchers examined the linguistic properties of effective formal debates, using features from style and latent content (Wang et al. 2017), a recurrent neural network (Potash and Rumshisky 2017), or semantic frames (Cano-Basave and He 2016). From the perspective of symbolic logic and game theory, Rahwan and Larson developed a mechanism for argumentation (2008). Later work built on this approach to account for agents hiding or lying about their arguments (Rahwan, Larson, and Tohmé 2009). Other researchers modeled uncertainty in strategic argumentation quantitatively (Rienstra, Thimm, and Oren 2013).

In this work, we demonstrate models for predicting persuasion in social media. In other work, data used for predicting convincingness of arguments (Habernal and Gurevych 2016b), consisted of short domain-specific texts of only a few sentences where sequence information is not necessary. In contrast, CMV has longer posts and is open domain; arguments can be about any topic. Furthermore, Change My View involves personalized persuasion, as opposed to requiring an objective standard of convincingness. Compared to previous work on CMV (Wei, Liu, and Li 2016; Tan et al. 2016), we leverage the sequential nature of argumentation.

3 Data

We use a dataset derived from the Change My View subreddit, a persuasive corpus where a user indicates if their view has been changed. As the data is self-labeled by posters, no human annotators are required. In previous work, Tan et al. (2016) collected threads (full discussion trees) submitted between 2013/01/01 and 2015/09/01, and segmented this data into submissions before and after 2015/05/08. This process resulted in 18,363 and 2,263 discussion trees, respectively, for train and test.

We consider three tasks. The first is **influence** prediction where, given a post and response, we attempt to predict whether the user changed their view. For this task, we extract posts and automatically identify positive/negative examples as paths in a discussion tree terminating with/without a delta, respectively. We extract only one path per response to the original poster by following the left-most path in a depth-first search and allowing a single unique response per path. Each datapoint is then an original post and attempted persuasive response, where responses are one or more sequential posts from the same commenter. For training, we require every original post in the data to have at least 1 positive and 1 negative response. The resulting training set has 19516 examples (14849 negative and 4667 positive). The test set contains 2465 examples (1836 negative and 629 positive).

The second and third tasks are the same as previous work (Tan et al. 2016). For the **pairwise** task, we predict which of two responses to the same original poster changed their view, where the two responses are controlled for topic by Jaccard similarity. The third task is **malleability** prediction, where the goal is to predict persuasion given only the original post and no responses.

Tan et al. (2016) distinguished two cases of the path-based prediction: predicting a delta from only the initial post in the response (termed the *root reply*) and including all posts in the response (termed the *full path*). For our experiments, at minimum the root reply and/or original post are available.

4 Methods

We model the posts using a hierarchical deep learning approach. Given a sentence representation $\mathbf{r_s}$ (see Section 4.1), where s is the index of the sentence, we model the document as a Long Short-Term Memory Network (LSTM) (Hochreiter and Schmidhuber 1997) with an attention mechanism over the sentences. We first apply a transformation to $\mathbf{r}_s$ to obtain a hidden state $\mathbf{h}_s$ (see Section 4.2). Next, we compute a document representation with attention over each hidden state, similar to previous work (Yang et al. 2016):

$$\mathbf{h} = \sum_{s \in [1,S]} \alpha_s \mathbf{h}_s \tag{1}$$

where S is the number of sentences in the document and attention is calculated by applying an MLP to the hidden state, $\mathbf{u}_s = \tanh(W_s \mathbf{h}_s + \mathbf{b}_s)$, before calculating the probability distribution over sentences (using $\mathbf{q}$ as a learned parameter vector):

$$\alpha_s = \frac{\exp(\mathbf{u}_s^T \mathbf{q})}{\sum_{i \in [1,S]} \exp(\mathbf{u}_i^T \mathbf{q})} \tag{2}$$

Finally, the document representation $\mathbf{h}$ is then passed through a multi-layer perceptron (MLP) to make a binary prediction of influence, which can be combined with features derived from the document (see Section 5):

$$y = \sigma\left(MLP(\mathbf{h}) + \beta^T \phi\right) \tag{3}$$

4.1 Sentence Representation

We create a sentence representation $\mathbf{r}_s$ by combining features from words, semantic frames, and discourse relations. We first represent each sentence by a weighted average of its word embeddings. Given a sentence at index s with T words and word embedding $\mathbf{x}_{s,t}^{word}$ for $t \in [1, T]$, the vector for s is:

$$\mathbf{v}_s^{word} = \sum_{t \in [1,T]} \alpha_{s,t}^{word} \mathbf{x}_{s,t}^{word} \tag{4}$$

Similarly, we add embeddings for semantic frames. The FrameNet (Ruppenhofer et al. 2006) model of frame semantics provides a method for describing events and relations. It also provides a way to model social interactions that are not captured by discourse structure or explicitly expressed in words such as agreement and disagreement. For example, the verb "agree" may take the "Compatibility" frame, which is shared with similar verbs. We use a FrameNet parser (Das et al. 2010) to predict the labels for lexical units and represent frames as the weighted average of the labels:

$$\mathbf{v}_s^{frame} = \sum_{l \in [1,L]} \alpha_{s,l}^{frame} \mathbf{x}_{s,l}^{frame} \tag{5}$$

where $\mathbf{x}_{s,l}^{frame}$ is the embedding for the l^{th} frame and L is the total number of frames.

Each attention weight α_{sj}^k is calculated for each $\mathbf{x}_{sj}^k$ for $k \in \{word, frame\}$ and $J \in \{T, L\}$, respectively, where T is the number of words and L is the number of frames:

$$\alpha_{sj}^k = \frac{\exp({\mathbf{u}_{s,j}^k}^T \mathbf{q}_k)}{\sum_{a \in [1,J]} \exp({\mathbf{u}_{s,a}^k}^T \mathbf{q}_k)} \tag{6}$$

and $\mathbf{u}_{s,j}^k = \tanh(W_k \mathbf{x}_{s,j}^k + \mathbf{b}_k)$ and $\mathbf{q}_k$ is a parameter vector.

Finally, we augment the sentence representation by incorporating embeddings for PDTB discourse structure. Previous work (Tan et al. 2016) used patterns of connectives such as "but-however-because" as features, but noted that these models suffered from low recall. Thus, modeling implicit discourse relations should improve coverage as implicit discourse is not explicitly captured by the remainder of the model. We use the end-to-end model of Biran and McKeown (2015) to tag PDTB relations rather than alternatives such as RST so that we can incorporate shallow structure into our LSTM. We represent the second-level discourse classes (e.g. Contingency/Causal and Comparison/Concession) for each inter-sentence relation as an embedding for sentence s as $\mathbf{v}_s^{inter}$, indicating the relationship between s and $s-1$.

The final sentence representation is then determined by concatenating each component of the sentence :

$$\mathbf{v}_s = \left[\mathbf{v}_s^{word}; \mathbf{v}_s^{frame}; \mathbf{v}_s^{inter}\right] \tag{7}$$

Given $\mathbf{v}_s$, we could use this representation of each sentence for the input at each timestep of an LSTM, or model feature interaction by applying an MLP to $\mathbf{v}_s$. Instead, we follow previous work in hierarchical language modeling (Kim et al. 2016) and allow the model to decide whether to *carry* features directly to the next layer, in order to allow for interaction between the word, discourse, and frame semantic features derived during this step. We thus obtain our sentence representation by feeding $\mathbf{v}_s$ into a highway network (Srivastava, Greff, and Schmidhuber 2015):

$$\mathbf{r}_s = \mathbf{t}_s \odot \mathbf{z}_s + (1 - \mathbf{t}_s) \odot \mathbf{v}_s \tag{8}$$

where $\mathbf{z}_s = g(W_h \mathbf{v}_s + \mathbf{b}_h)$, a hidden representation of the original vector with a non-linearity g, and $\mathbf{t}_s = \sigma(W_t \mathbf{v}_s + \mathbf{b}_t)$, a prediction of whether to use the original features. The highway network is a mixture of the hidden representation of the vector given by the MLP and the original vector, where the model learns the weight vector $\mathbf{t}_s$. Thus, because of the learned weight $\mathbf{t}_s$, the model decides how to interpolate between the hidden representation and the original vector.

4.2 Dynamic Memory Network

One variant of our model is to use a bi-directional LSTM for $\mathbf{h}_s$ in Equation 1 over the sentences from the reply only. However, this would only allow the attention mechanism to consider the response, rather than the context of the original post. We thus include information about the original post using a dynamic memory network, which has been effective in modeling context (Xiong, Merity, and Socher 2016; Wang and Zhang 2017), to iteratively find abstract representations using information from both the original post and the response. Let $\mathbf{h}_s^r$ be the LSTM state at sentence s in the response and $\mathbf{h}_s^{op}$ the LSTM state at sentence s in the original post. We then create a representation $\mathbf{h}^{op}$ for the entire original post by using the attention mechanism in Equation 1 where $\mathbf{h}_s = \mathbf{h}_s^{op}$. This is concatenated with $\mathbf{h}_s^r$ and the memory representation $\mathbf{v}^t$ to create the input representation: $\mathbf{h}_s^t = [\mathbf{h}_s^r; \mathbf{h}^{op}; \mathbf{v}^t]$. By allowing the attention mechanism to consider the context and the entire response, the model is

able to more accurately predict which sentences are important. This results in a modified version of equation 1, where $\mathbf{h}^t = \sum_{s\in[1,S]} \alpha_s \mathbf{h}_s^t$. After each iteration, the memory $\mathbf{v}^t$ is set to $\mathbf{h}^{t-1}$. The initial memory $\mathbf{v}^0$ is initially set to the average of the hidden states: $\sum_{s\in[1,S]} \mathbf{h}_s^r/S$. We could use $\mathbf{h}^0$ as the final document representation $\mathbf{h}$ in Equation 3, but in practice multiple iterations have been more effective (Xiong, Merity, and Socher 2016), which we validate empirically.

4.3 Hyperparameters and Optimization

We use binary cross-entropy as the loss function and stochastic gradient descent with a mini-batch size of 100 and Nesterov momentum with a coefficient of 0.9. Word embeddings are initialized with pre-trained 300-dimensional GloVe vectors. Out-of-vocabulary words are randomly initialized and optimized during training. We stop training after 30 epochs and perform early stopping on a validation set. The document weights β in Equation 3 were pre-trained using a logistic regression classifier.

We experimented with different settings for various hyper-parameters. For the recurrent and hidden dimensions, we tested values of 50, 100, 200, and 300. For dropout (Srivastava et al. 2014) and word dropout (Iyyer et al. 2015), we used values of 0.25, 0.5, and 0.75 and determined whether to use 1 or 2 hidden layers. We use ReLU as the non-linearity in Equations 3 and 8. We evaluated the number of iterations for the memory networks and found that performance increases up to 3 iterations and begins decreasing after 3. We limit the maximum length of each sentence to 32 words and the maximum length of a post to 40 sentences. Words occurring fewer than 5 times in the training set (including the original post, title, and response) were removed.

5 Results

The results of our experiments on the held-out test set are shown in Tables 2, 3, and 4 for each of the influence, pairwise, and malleability prediction tasks, respectively. For the pairwise and influence prediction subtasks, we report results for both the root reply and full path options and we compare models using sentences from just the response (R) and the response plus the original post (R+OP).

We provide baseline models, from previous work, trained using logistic regression on features from just the response (*bag-of-words*) and from the response plus the original post (*interplay*). In the work of Tan et al. (2016), their best-performing features were derived from the *interplay* (IP) between the original post and the response. They derived 12 features from 4 similarity scores (common words, similar fraction in reply, similar fraction in OP, and Jaccard score) and 3 subsets (all words, stop words, and content words). The interplay provides a strong baseline because we might expect there to be significant overlap between the posts if users are imitating the writing style of the original poster in order to be more persuasive. In addition, we provide a *bag-of-words* (BoW) baseline. We remove words occurring less than 5 times and L2-normalize term frequency vectors.

We present results using only words (word-LSTM) and words, frames, and discourse relations (all-LSTM). For the pairwise and influence tasks, these models consider the response only. For the malleability task, these models consider the original post. When the response and original post are both provided, we use the memory network described in 4.2 (all-LSTM+memory). We also provide results for baseline features combined with our model (all-LSTM+memory+IP), with the features as ϕ in Equation 3.

5.1 Discussion

The LSTM models significantly outperform all baselines, especially when combined with the interplay features. In the **influence** prediction task, the best model using only the response (all-LSTM) outperforms the BoW baseline in both the root reply and full path cases ($p < 0.001$ by a randomized permutation test). Given the response and the original post, the best model (all-LSTM+memory+IP) outperforms the IP baseline in both cases ($p < 0.001$). The difference between the best model and the baseline is also larger in the full path case when compared to the root reply case. This is not surprising, as many responses in our dataset contain only a single sentence, often a clarifying question, so the model is unable to benefit from sequential information when only the root reply is included. We also observe that modeling the context of the original post helps in both scenarios, but the context is more important in the root reply case, obtaining around a 4 point increase from all-LSTM to all-LSTM+memory compared to 2-3 points in the full path case. As the model has limited content to work with in the root reply case it is most likely taking advantage of features in the original post. Additionally, it is surprising that interplay is such a strong baseline, especially in the root reply case. Our all-LSTM+memory model does not outperform the interplay features alone but provides a complementary approach to the interplay features.

For the **pairwise** prediction task, we obtain better performance on accuracy ($p < 0.001$ by McNemar's test, comparing all-LSTM to BoW in both the root reply and full path cases, and $p < 0.01$ comparing all-LSTM+memory+IP to the IP baseline). By controlling for topic in the pairwise dataset, individual words have less influence. Even though the model contains shallow structural features, word embeddings are a central part of the model, so the fact that the model performs well on pairwise prediction even with controlling for topic similarity suggests that the ordering of the document is key. Furthermore, we do not see significant improvement by including context in the pairwise task, which may indicate that the model is learning a bias for features of the original post rather than interacting with the response.

Finally, we would expect BoW to do well on **malleability**, as Tan et al. (2016) showed that common words associated with openness or stubbornness were strong features. However, we see significant gains from sequential models ($p < 0.05$ by a randomized permutation test for all-LSTM), suggesting the ordering of arguments provides some indicator of how and whether they can be convinced.

5.2 Ablation

We present additional results in Table 5 on the full path task for influence, with certain model components from the all-

		Root Reply			*Full Path*		
	Model	***Acc.***	***AUC***	***True F-score***	***Acc.***	***AUC***	***True F-score***
R	BoW	60.4	68.9	47.1	61.9	72.8	50.3
	word-LSTM	71.2	70.5	48.7	72.9	75.1	52.7
	all-LSTM	**72.5**	**70.8**	**48.9**	**75.1**	**75.5**	**53.0**
R+OP	IP	70.5	74.8	52.1	72.7	76.7	54.6
	all-LSTM+memory	75.0	74.9	53.1	74.3	77.3	55.4
	all-LSTM+memory+IP	**77.2**	**79.5**	**58.0**	**81.0**	**82.1**	**60.7**

Table 2: Results of Influence Prediction Task

	Model	***Root Reply***	***Full Path***
R	BoW	59.6	62.3
	word-LSTM	67.0	70.8
	all-LSTM	**67.5**	**71.5**
R+ OP	IP	65.2	69.2
	all-LSTM+memory	67.7	71.6
	all-LSTM+memory+IP	**69.0**	**71.9**

Table 3: Accuracy for Pairwise Prediction Task

Model	***Acc.***	***AUC***	***True F-score***
BoW	51.6	53.3	48.1
word-LSTM	57.7	55.5	**56.5**
all-LSTM	**58.4**	**57.2**	53.2

Table 4: Results of Malleability Prediction Task

Model	***Accuracy***	***AUC***	***True F-score***
all-LSTM	**75.1**	**75.5**	**53.0**
no highway	70.1	74.9	52.6
no lstm	68.8	73.2	50.3
no attention	66.6	74.5	51.3
discourse only	54.6	63.6	43.5
frames only	43.3	66.4	44.2

Table 5: Component Ablation

Model	***Pairwise***	***Influence***
Annotators	54.84	57.14
all-LSTM+IP	71.99	63.00

Table 6: Human Performance

LSTM model ablated to assess their contribution to modeling the sequence of reasoning. We remove the highway network component of the model, indicated in the table as *no highway*, and instead directly use the concatenated embeddings $\mathbf{v}_s$ as the input to the bi-directional LSTM. We also remove the bi-directional LSTM from the model, indicated in the table as *no lstm*, and instead take a weighted average of all the embeddings $\mathbf{v}_s$. Finally, we remove the attention mechanism over the LSTM states (*no attention*) and instead average the LSTM states over each timestep. We also present the impact of discourse and frame embeddings when included in the model without the other embeddings.

As demonstrated in Table 5, the sequential nature of the LSTM contributes to the overall performance of the model. Compared to the full model, the model without an LSTM (which considers the ordering of the content provided) does 2-3 points worse in AUC and F-score, showing that modeling the sequence of arguments helps in predicting persuasion ($p < 0.01$ by a randomized permutation test). We also obtain improvement by including the highway network and the attention mechanism ($p < 0.05$). Removing the highway or the attention component costs the model 0.5 to 1 point of performance. Without the highway layer, the neural network can only consider the sentence features individually and not the interaction between components. Without the attention layer, the model is unable to determine which parts of the sequence are most important to weight in the final prediction. Finally, the frame and discourse embeddings perform poorly on their own, but contribute to the overall model.

6 Analysis

6.1 Human Performance

We also conduct an evaluation of human judgments to compare performance. We set up an experiment on Crowdflower where we ask annotators to view discussions from Change My View. For each discussion thread, we display the original post and title, then display one positive argument and one negative argument in a random order. For each argument, we display all posts from the author of the root response so that the annotators have access to the same data as the model. This is equivalent to the "full path" task in our experiments.

First, for each argument, we ask the annotator whether they believe the original poster would find the argument convincing. Then we ask annotators to rank the arguments, to compare to the pairwise accuracy task. We instruct the annotators to read the original post and both arguments before answering any questions. For each of the three questions, for quality control we require each annotator to provide a justification of their decision of at least 20 words. Justifications that did not meet this requirement or were clearly spam had their judgments removed from the dataset. As an additional quality control, we require annotators to spend at least 300 seconds on each discussion. Annotators are required to give three judgments per thread and we annotate a total of 200 discussion threads. Results are presented in Table 6, showing the majority vote of the annotators along with our model performance on the same subset of data.

It is not surprising that human annotators struggle with both the pairwise prediction task and the influence prediction task. If humans were better at predicting when a post

		Human		*Model*	
Category	***%***	***P***	***I***	***P***	***I***
Government	29	76.3	55.1	64.4	58.5
Sociology	23	71.7	53.3	80.4	68.5
Morality	11	72.7	63.6	77.3	68.2
Economics	9	50.0	50.0	72.2	58.3
Politics	8	62.5	56.3	68.8	62.5
Science	6	66.6	66.6	66.6	62.5
Culture	5.5	54.5	45.5	54.5	63.6

Table 7: Error Analysis on Categorized Data (P: Pairwise I: Influence %: Percentage of Data in Category)

would be persuasive, we would likely see more persuasion in our dataset. Our models significantly outperform human annotators on both tasks. One key distinction is that the annotators received no training in what makes a successful argument, whereas our models are trained on thousands of documents. An expert in persuasive writing may perform very well at this task so we can only claim that our model is better than novice annotators.

6.2 Error Analysis

We categorize examples into several categories to see how our models and the human annotators fare. Then we report performance on each category. We divide all posts in the human-annotated subset into seven broad categories: government (what laws should be implemented), sociology (behavior of groups or discussion of social issues such as feminism), morality (judgments of right and wrong), economics (personal or group decisions to maximize utility), politics (what political parties and candidates should do), science (questions with objective, measurable answers such as whether vaccines are effective), and culture (books, music, games, etc.). Each post is categorized by the first author and any post not clearly belonging to a category is discarded.

In an example of the politics category, an original poster writes: *There is no practical reason for any individual to vote in national elections. By "practical reason," I mean a reason that motivates you to vote by ascribing a cause-effect ... This is a classic example of a collective action problem.*. In a winning argument, a user writes: *Just because it's incredibly unlikely that your vote will make a difference doesn't mean it's never going to happen. ... Depending on a person's valuation of costs and potential benefits, this could very well be enough*. In contrast, another user writes an unconvincing argument: *The same ballot for Presidential and Congressional elections will also have a number of other state and local positions and issues ... Then you are putting in a very low amount of effort for a very low amount of impact.* On this example, the human annotators correctly predict the positive response but not the negative one whereas our model correctly predicts both.

The overall results for accuracy are reported in Table 7. Overall our models perform best on topics in sociology and morality and have issues with discussions in government and economics. We observe that in CMV the former tend to be more emotional (for example, in response to the original poster writing *Weinberg was wrong when he said that "for good people to do evil things, that takes religion"* another user writes *I think that someone isn't a good person if they have an ideology I disagree with*) while the latter tend to be more empirical (for the topic *Countries should have a "no confidence" vote in elections if they want to increase turnout, while achieving a better understanding of the public's perception of the political climate*, another poster responds with facts: *The US state of Nevada has had a choice called "none of these candidates" since 1975*). As the empirical arguments often require world knowledge we would expect our models to struggle in this area. Conversely, our models may pick up on sequential arguments alternating between emotion and logic in other categories. For example, *I think that someone isn't a good person if they have an ideology I disagree with* is followed by *I think nationalists are bad, fascists are bad and so on*. The model correctly identifies the post with these arguments as not receiving a delta, which may be due to the sequence of simplistic, emotional language used. Finally, compared to human performance, our models are worse or at the same level in government and science, suggesting that world knowledge may again be the distinguishing factor.

6.3 Model Evaluation

One advantage of this model is that we can easily see which words, frames, or discourse relations are prominent features according to the attention-based weighting. For the influence task, highly-weighted words include terms such as *objectified*, *stereotyped*, *thesaurus*, and *linguist* which may just indicate that people have strong opinions on these topics. Highly-weighted frames, however, include *research* and *medical_professionals*, which may indicate users providing evidence, or *confronting_problem* and *suasion* (attempts to persuade) which may indicate social interaction. In the malleability case, highly-weighted words include *greetings* and *brigading* (a Reddit term for a group of users coordinating to downvote certain posts), which indicate social aspects of persuasion. Other highly-weighted words include terms such as *protectionism* and *anarcho* (a word in the context of anarcho-capitalism), which is unsurprising as politics is a controversial topic. Highly-weighted frames include social cues such as *contrition* or *hostile_encounter*, which may indicate susceptibility or resistance to persuasion, respectively.

We also conduct a qualitative analysis to evaluate the impact of the *sentence*-level attention weights. We present results showing human judgments of the most important sentences in the response and we compare the results of this annotation task to the attention weights output by the model, as in (Ghosh, Richard Fabbri, and Muresan 2017). We designed an Amazon Mechanical Turk (AMT) task to conduct our experiments. We provide the annotator with an original post and the sentences in the reply. As with the experiment in Section 6.1, the annotators have access to the same data as the model. The annotator is asked to indicate the "most important" sentences in the response. They are then required to select at least one sentence but may select the entire response. We use the same subset of test data as our experiment in section 6.1 and limit the length of the original post

Positive	*Attn*	*Label*
Are you arguing that collage is affordable, or more affordable than people imply?	0.28	0.2
Because while I would agree that there is likely some exaggeration, for many people it is completely unaffordable.	0.29	0.4
Not everyone gets the best case scenario, and if you make less than $30000 a year, then paying minimum of a third of a years pay on education is not feasible.	0.23	0.6
And I don't know how considering alternatives to collage is an argument for the affordability of collages; yes collage is cheep if I do not go to it and take an apprenticeship instead, but I don't know what it would have to do with this discussion.	0.2	0.4
Negative	*Attn*	*Label*
My family made "too much" for FAFSA aide but too little to afford me much assistance with college prices.	0.19	0.6
I went to a school where I was given a full academic scholarship, which included room and board.	0.18	0.4
In order to afford additional fees / books / transportation I still had to take out a Stafford loan every year.	0.16	0.4
On top of that, the government decided that the room and board part of my scholarship qualified as "income", and I then owed the IRS money come tax return time for each of my four years.	0.15	0.4
I'll still be paying off these loans for a few years.	0.16	0
My point: Even with the "best case scenario" of a full scholarship, college still poses a significant financial burden.	0.16	0.4

Table 8: Attention Weights and Human Annotations from a Positive and Negative Example

and reply to be between 3 and 10 sentences to simplify the task for the annotators. This results in 36 positive and 44 negative examples. Each HIT contains one task and 5 annotators were required for each task. Only Master-level annotators were selected.

We first compare the sentence-level weights of the all-LSTM+memory model to the annotators' selections. We find that 32% of the time the highest-weighted sentence from the model is the sentence where the most annotators agree that the sentence is important. We also find that 35% of the time, the highest-weighted sentence from the model is the second-most important sentence from the annotators. Of the remaining 33%, the model selects the first sentence 60% of the time, indicating a bias towards the beginning of the text. Overall, a baseline method of always weighting the first sentence the highest would achieve 20% accuracy compared to the annotators. In this subset of data, the average length of the positive posts is 6.25 sentences and the average length of the negative posts is 6.27 sentences. Even though the posts are the same average length, we find that for positive responses, the Turkers selected 19% of all sentences whereas for negative responses, they selected 16%, indicating that positive responses contain more important content.

We also provide an example of attention weights along with the predictions made by annotators in Table 8. The original post is omitted due to space constraints. The title is "College is not unaffordable in the US." The full text of a response that received a delta and one that did not are both provided, segmented into sentences. The "Labels" column indicates the percentage of annotators that voted for that sentence and the "Attn" column indicates the probability assigned to the sentence by the model. The Attn column will thus sum to 1 but the Labels column will not, so we compare the relative ranking of each sentence. The top-ranked sentence by the annotators is highlighted in bold. In both cases this sentence could act as a summary for the entire argument. However, the attention weights in this example do not reflect this ranking. The overall prediction for both responses was incorrect and a correct prediction may only be possible with world knowledge (about the value of money).

7 Conclusion

We have presented evidence that the ordering of a document is crucial to influential writing. We provided a neural model using words, frames, and discourse relations that effectively predicts persuasiveness in several tasks and significantly improves upon prior work. We have demonstrated that this is a difficult task for humans but we have surpassed non-expert performance.

In future work, we hope to continue work in influence and persuasiveness. This dataset has the advantage of being labeled, but work in *unsupervised* persuasiveness prediction, given only text responses indicating persuasiveness, is one possible direction. Other avenues of research include modeling the interaction between the original post and the responses. Interplay is a simple but effective representation of interaction but modeling threads as dialogues or multi-party discourse rather than monologues may yield further improvements. Finally, the users in Change My View are required to provide an explanation for the reason their view changed and we can analyze these reasons and attempt to predict *why* someone changed their view.

8 Acknowledgments

This work was supported by the DARPA-DEFT program. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. We thank the annotators for their work and the anonymous reviewers for their feedback.

References

Biran, O., and McKeown, K. 2015. Pdtb discourse parsing as a tagging task: The two taggers approach. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 96–104. Prague, Czech Republic: Association for Computational Linguistics.

Cano-Basave, A. E., and He, Y. 2016. A study of the impact of persuasive argumentation in political debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1405–1413. San Diego, California: Association for Computational Linguistics.

Das, D.; Schneider, N.; Chen, D.; and Smith, N. A. 2010. Probabilistic frame-semantic parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 948–956. Los Angeles, California: Association for Computational Linguistics.

Ghosh, D.; Richard Fabbri, A.; and Muresan, S. 2017. The role of conversation context for sarcasm detection in online interactions. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 186–196. Saarbrcken, Germany: Association for Computational Linguistics.

Habernal, I., and Gurevych, I. 2016a. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1214–1223. Austin, Texas: Association for Computational Linguistics.

Habernal, I., and Gurevych, I. 2016b. 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 (Volume 1: Long Papers)*, 1589–1599. Berlin, Germany: Association for Computational Linguistics.

Hidey, C.; Musi, E.; Hwang, A.; Muresan, S.; and McKeown, K. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, 11–21. Copenhagen, Denmark: Association for Computational Linguistics.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural Comput.* 9(8):1735–1780.

Iyyer, M.; Manjunatha, V.; Boyd-Graber, J.; and Daumé III, H. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1681–1691. Beijing, China: Association for Computational Linguistics.

Jaech, A.; Zayats, V.; Fang, H.; Ostendorf, M.; and Hajishirzi, H. 2015. Talking to the crowd: What do people react to in online discussions? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2026–2031. Lisbon, Portugal: Association for Computational Linguistics.

Kim, Y.; Jernite, Y.; Sontag, D.; and Rush, A. M. 2016. Character-aware neural language models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, 2741–2749.

Nettel, A. L., and Roque, G. 2012. Persuasive argumentation versus manipulation. *Argumentation* 26(1):55–69.

Potash, P., and Rumshisky, A. 2017. Towards debate automation: a recurrent model for predicting debate winners. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2455–2465. Copenhagen, Denmark: Association for Computational Linguistics.

Prabhakaran, V., and Rambow, O. 2013. Written dialog and social power: Manifestations of different types of power in dialog behavior. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 216–224. Nagoya, Japan: Asian Federation of Natural Language Processing.

Rahwan, I., and Larson, K. 2008. Mechanism design for abstract argumentation. In *7th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008), Estoril, Portugal, May 12-16, 2008, Volume 2*, 1031–1038.

Rahwan, I.; Larson, K.; and Tohmé, F. A. 2009. A characterisation of strategy-proofness for grounded argumentation semantics. In *IJCAI 2009, Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, California, USA, July 11-17, 2009*, 251–256.

Rienstra, T.; Thimm, M.; and Oren, N. 2013. Opponent models with uncertainty for strategic argumentation. In *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*, 332–338.

Rosenthal, S., and Mckeown, K. 2017. Detecting influencers in multiple online genres. *ACM Trans. Internet Technol.* 17(2):12:1–12:22.

Ruppenhofer, J.; Ellsworth, M.; Petruck, M. R.; Johnson, C. R.; and Scheffczyk, J. 2006. *FrameNet II: Extended Theory and Practice*. Berkeley, California: International Computer Science Institute. Distributed with the FrameNet data.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15(1):1929–1958.

Srivastava, R. K.; Greff, K.; and Schmidhuber, J. 2015. Highway networks. *CoRR* abs/1505.00387.

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 2016, Montreal, Canada, April 11 - 15, 2016*, 613–624.

Wang, Z., and Zhang, Y. 2017. Opinion recommendation using a neural model. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1627–1638. Association for Computational Linguistics.

Wang, L.; Beauchamp, N.; Shugars, S.; and Qin, K. 2017. Winning on the merits: The joint effects of content and style on debate outcomes. *Transactions of the Association for Computational Linguistics* 5:219–232.

Wei, Z.; Liu, Y.; and Li, Y. 2016. Is this post persuasive? ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 195–200. Berlin, Germany: Association for Computational Linguistics.

Xiong, C.; Merity, S.; and Socher, R. 2016. Dynamic memory networks for visual and textual question answering. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, 2397–2406. JMLR.org.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A. J.; and Hovy, E. H. 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.