Joint Models of Disagreement and Stance in Online Debate

Dhanya Sridhar,[1] **James Foulds,**[1] **Bert Huang,**[2] **Lise Getoor,**[1] **Marilyn Walker**[1]
[1]Department of Computer Science, University of California Santa Cruz
{dsridhar, jfoulds, getoor, mawalker}@ucsc.edu
[2]Department of Computer Science, Virginia Tech
bhuang@vt.edu

Abstract

Online debate forums present a valuable opportunity for the understanding and modeling of dialogue. To understand these debates, a key challenge is inferring the stances of the participants, all of which are interrelated and dependent. While collectively modeling users' stances has been shown to be effective (Walker et al., 2012c; Hasan and Ng, 2013), there are many modeling decisions whose ramifications are not well understood. To investigate these choices and their effects, we introduce a scalable unified probabilistic modeling framework for stance classification models that 1) are collective, 2) reason about disagreement, and 3) can model stance at either the author level or at the post level. We comprehensively evaluate the possible modeling choices on eight topics across two online debate corpora, finding accuracy improvements of up to 11.5 percentage points over a local classifier. Our results highlight the importance of making the correct modeling choices for online dialogues, and having a unified probabilistic modeling framework that makes this possible.

1 Introduction

Understanding stance and opinion in dialogues can provide critical insight into the theoretical underpinnings of discourse, argumentation, and sentiment. Systems for predicting the stances of individuals can potentially have positive social impact and are of practical interest to non-profits, governmental organizations, and companies. For example, stance predictions may be used to target public awareness and advocacy campaigns, direct political fundraising and get-out-the vote efforts, and improve personalized recommendations.

Dialogue Turns	Stance
User 1: 18. That's the smoking age thats the shooting age. Why do you think they call it ATF?	ANTI
User 2: Shooting age? I know 7 year old shooters. 18 should be the gun purchasing age, but there is really no "shooting" age.	ANTI
User 1: I know. I was just pointing out that the logic used to propose a 21 year "shooting age" was inconsistent.	ANTI
User 2: I see. I dont think its really fair that you can join the army at 18 and use handguns and military weapons, but you cant purchase a handgun until 21.	ANTI

Figure 1: Example of a debate dialogue turn between two users on the *gun control* topic, from 4FORUMS.COM.

Online debate websites are a particularly rich source of argumentative dialogic data (Fig. 1). On these websites, users debate and share their opinions on a variety of social and political issues. Previous work (Somasundaran and Wiebe, 2010; Walker et al., 2012c) has shown that stance classification in online debates is a challenging problem. While collective approaches that jointly predict user stance seem promising (Walker et al., 2012c; Hasan and Ng, 2013), the rich structure of online debate forums necessitates many modeling choices. For example, users publish opinions and reply and respond to each others' posts. In so doing, they may agree or disagree with either all or a portion of another user's post, suggesting that collective classifiers for stance may benefit from text-based disagreement modeling. Furthermore, one can model stance either at the author level—assuming that an author's stance is based on all of their posts on a topic (Burfoot et al., 2011)—or at

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics
and the 7th International Joint Conference on Natural Language Processing, pages 116–125,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

the post level—assuming that an author's stance is post-specific and may vary across posts (Hasan and Ng, 2013). These decisions can drastically change the nature of stance models, so understanding their implications is critical.

In this paper, we develop a flexible modeling framework for stance classification using probabilistic soft logic (PSL) (Bach et al., 2013; Bach et al., 2015), a recently introduced probabilistic modeling framework.[1] PSL is a probabilistic programming system that allows models to be specified using a declarative, rule-like language. The resulting models are a special form of conditional random field, called a hinge-loss Markov random field, which admits highly scalable exact inference (Bach et al., 2013). Modeling stance in large, richly connected online debate forums requires a careful exploration of many modeling choices. This complex domain especially benefits from PSL's flexibility and scalability. PSL makes it easy to develop model variations and extensions, as one can readily incorporate new factors capturing additional intuitions about dependencies in a domain.

We evaluate our models on data from two debate sites, 4FORUMS and CREATEDEBATE (Walker et al., 2012b; Hasan and Ng, 2013), which we describe in detail in Section 2. Our experimental results show that there are important ramifications of several modeling decisions, including whether to use collective or non-collective models, to represent stance at the post level or the author level, and how to model disagreement. We find that with appropriate modeling choices, our approach leads to improvements of up to 11.5 percentage points of accuracy over simple classification approaches.

Our contributions include (1) a flexible, unified framework for modeling online debates, (2) extensive experimental study of many possible models on eight forum datasets, collected across two different debate websites, and (3) general modeling recommendations resulting from our empirical studies.

2 Online Debate Forums

Online debate forums represent richly structured argumentative dialogues. On these forums, users debate with each other in discussion threads on a variety of topics or issues, such as *gun control*, *gay marriage*, and *marijuana legalization*. Each discussion consists of a number of posts, which are short text documents authored by users of the forum. A post is either a reply to a previous post, or it is the start (root) of a thread. As users engage with each other, a thread branches out into a tree of argumentative interactions between the users. Forum users often post numerous times and across multiple discussions and topics, which creates a richly structured interaction graph. Online debates present different challenges than more controlled dialogic settings such as congressional debates. Posts are short and informal, there is limited external information about authors, and debate topics admit many modes of argumentation ranging from serious, to tangential, to sarcastic. The reply graph in online debates also has substantially different semantics to networks in other debate settings, such as the graph of speaker mentions in congressional debates. To illustrate this setting, Fig. 1 shows an example dialogue between two users who are debating their opinions on the topic of gun control.

In the context of online debate forums, *stance classification* (Thomas et al., 2006; Somasundaran and Wiebe, 2009) is the task of assigning stance labels with respect to a discussion topic, either at the level of the user or the level of the post. Stance is typically treated as a binary classification problem, with labels PRO and ANTI. In Fig. 1, both users' stances toward gun control are ANTI.

Previous work on stance in online debates has shown that contextual information given by reply links is important for stance classification (Walker et al., 2012a), and that collective classification often outperforms methods which treat each post independently. Hasan and Ng (2013) use conditional random fields (CRFs) to encourage opposite stances between sequences of posts, and Walker et al. (2012c) use MaxCut over explicitly given rebuttal links between posts to separate them into PRO and ANTI clusters. Sridhar et al. (2014) use hinge-loss Markov random fields (HL-MRFs) to encourage consistency between post level stance labels and observed post-level textual agreements and disagreements.

While the first two approaches leverage rebuttal or reply links, they model reply links as being indicative of opposite stances. However, as shown in Fig. 1, responses—even rebuttals—can occur be-

[1] PSL is an open-source Java toolkit, available here: http://psl.cs.umd.edu.

tween users with the same stance, which suggests the benefit of a more nuanced treatment of reply links. The approach of Sridhar et al. (2014) considers text-based agreement annotations between posts, though it requires that reply links are labeled. Accurate reply polarity labels are likely to be as expensive to obtain as the stance labels that we aim to predict. Noisy or sparse reply labels are cheaper, though likely to reduce performance. In this work, we show how to reason over uncertain reply label predictions to improve stance classification.

Also in the online debate setting, Hasan and Ng (2014) show the benefits of joint modeling to classify post-level stance and the authors' reasons for their stances. In contrast, in this work we focus on the dependencies between stance and polarity of replies.

In the context of opinion subgroup discovery, Abu-Jbara and Radev (2013) demonstrate the effectiveness of clustering users by opinion-target similarity. In contrast, Murakami and Raymond (2010) use simple recurring patterns such as "*that's a good idea*" to categorize reply links as *agree*, *disagree* or *neutral*, prior to using MaxCut for subgroup clustering of comment streams on government websites. This approach improves over a MaxCut approach that casts all reply links as disagreements. Building on this work, Lu et al. (2012) model unsupervised discovery of supporting and opposing groups of users for topics in online military forums. They improve upon a MaxCut baseline by formulating a linear program (LP) to combine multiple textual and reply-link signals, suggesting the benefits of jointly modeling textual and reply-link features.

In a different line of work, while Somasundaran and Wiebe (2010) do not use relational information between users or posts, their approach shows the benefit of modeling opinions and their targets at a fine-grained level using relational sentiment analysis techniques. Similarly, Wang and Cardie (2014) demonstrate the effectiveness of using sentiment analysis to identify disputes on Wikipedia Talk pages. Boltužić and Šnajder (2014) and Ghosh et al. (2014) study various linguistic features to model stance and agreement interactions respectively.

In the congressional debate setting, approaches using CRFs and similar collective techniques such as minimum-cut have also leveraged reply link polarity for improvements in stance classification (Thomas et al., 2006; Bansal et al., 2008; Balahur et al., 2009; Burfoot et al., 2011). However, these methods rely heavily on features specific to the congressional setting in order to predict link polarity, and make little use of textual features. In contrast, Abbott et al. (2011) use a range of linguistic features from the text of posts and their parents to classify agreement or disagreement between posts on the online debate website 4Forums.com, without the goal of classifying stance.

	4Forums	CreateDebate
Users per topic	336	311
Posts per user, per topic	19	4
Words per user, per topic	2511	476
Words per post	134	124
Distinct reply links per user, per topic	6	3
Stance labels given for	Users	Posts
%Post-level reply links have opposite-stance users	71.6	73.9
%Author-level reply links have opposite-stance users	52.0	68.9

Table 1: Structural statistics averages for 4Forums and CreateDebate.

In this work, we study datasets from two online debate websites: 4Forums.com, from the Internet Argument Corpus (Walker et al., 2012b), and CreateDebate.com (Hasan and Ng, 2013). Table 1 shows statistics about these datasets including the average number of users per discussion topic and average number of posts authored. The best stance classification accuracy to date for online debate forums ranges from 70.1% on ConvinceMe.net to 75.4% on CreateDebate.com (Walker et al., 2012c; Hasan and Ng, 2013). The web interface for ConvinceMe.net enforces opposite stances for reply posts, making this dataset inapplicable for text-based disagreement modeling, and so we do not consider it in our experiments. In the more typical online debate forum corpora that we study, the presence of a reply, or even a textual disagreement between posts, does not necessarily indicate opposite stance (e.g. in gun control debates on 4Forums, 23% of disagreements correspond with same stance).

For our unified framework, we specify a hinge-loss Markov random field to reason jointly about stance and reply-link polarity labels. A closely related line of work focuses on improving structured prediction with domain knowledge modeled as constraints in the objective function (Chang et al., 2012; Ganchev et al., 2010; Mann and McCallum, 2010). Though more often used in semi-supervised settings, constraint-based learning can be especially appropriate for supervised learning when commonly used feature functions for linear models do not capture the richness of the data. Our HL-MRF formulation admits highly expressive features while maintaining a convex objective, thereby enjoying both tractability and a fully probabilistic interpretation.

3 Modeling Choices

We face multiple modeling decisions that may impact predictive performance when classifying stance in online debates. A key contribution of this work is the exploration of the ramifications of these choices. We consider the following variations on modeling: collective (**C**) versus local (**L**) classifiers, whether to explicitly model disagreement (**D**), and author-level (**A**) versus post-level (**P**) models.

Collective versus Local. Both collective and non-collective methods for stance prediction require a strong local text classifier. The methods proposed in this paper build upon the state-of-the-art local classification approach of Walker et al. (2012a), which trains a supervised classifier using features including n-grams, lexical category counts, and text lengths. We use logistic regression for the local classifier. These models will be referred to as *local* (**L**). In *collective* (**C**) classification approaches for stance prediction, the stance labels are all predicted jointly, leveraging relationships along the graph of replies. The simplest way to make use of reply links is to encode that the stance of posts (or authors) that reply to each other is likely to be opposite (Walker et al., 2012c; Hasan and Ng, 2013). Collective approaches attempt to find the most likely joint stance labeling that is consistent with both the local classifier's predictions and the alternation of stance along response threads. The alternating stance assumption is not necessarily a hard constraint, and may potentially be overridden by the local predictions. **C** and **L** models can be constructed with **A** or **P**-level granularity as described below, resulting in four modeling combinations.

Modeling Disagreement. As seen in Fig. 1 and Table 1, the assumption that reply links correspond to opposite stance is not always correct. This suggests the potential benefit of more nuanced models of agreement and disagreement. A natural disagreement modeling approach is to predict the polarity of reply links jointly with stance.

There are two variants of reply link polarity to consider. In *textual disagreement*, replying posts are coded as expressing agreement or disagreement with the text of the parent post. This may not correspond to a disagreement in stance *relative to the thread topic*. Some forum interfaces support user self-labeling of post reply links as rebuttals or agreements, thereby explicitly providing textual disagreement labels for posts. Alternatively, in the *stance disagreement* variant, reply links denote either same or opposite *stance* between users (posts). In Fig. 1, User 1 and User 2 disagree in text but have the same stance. For collective modeling of stance and disagreement, it is useful to consider the stance disagreement variant which identifies opposite and same-stance reply links, and jointly encourage stance predictions to be consistent with the disagreement predictions.

As with the local classification of stance, we can construct local classifiers for stance disagreement. In this work, for each reply link instance, we use a copy of the local stance classification features for each author/post at the ends of the reply link. The linguistic features further include discourse markers such as "actually" and "because" from the disagreement classifier of Abbott et al. (2011). Additionally, we use textual disagreement as a feature for stance disagreementwhen available. When reply links are not explicitly labeled as rebuttals or agreements, or only rebuttals are known, we instead predict textual disagreement using the features given above, trained on a separate data set with textual-disagreement labels.

Finally, with a stance disagreement classifier in hand, we can build collective models that predict stance based on predicted stance disagreement polarity. We denote these models as *disagreement* (**D**). When applied at one of **A** or **P**-level modeling, this yields two more possible modeling configurations. These models are certainly more complex than others we consider, but their design is consistent with intuition about the nature of discourse, so the added complexity may yield better accuracy.

All models:		**Collective models only:**		**Disagreement models only:**	
localPro(X1)	→ *pro*(X1)	*disagree*(X1, X2) ∧ *pro*(X1)	→ ¬ *pro*(X2)	*localDisagree*(X1, X2)	→ *disagree*(X1, X2)
¬ *localPro*(X1)	→ ¬ *pro*(X1)	*disagree*(X1, X2) ∧ ¬ *pro*(X1)	→ *pro*(X2)	¬ *localDisagree*(X1, X2)	→ ¬ *disagree*(X1, X2)
		¬ *disagree*(X1, X2) ∧ *pro*(X1)	→ *pro*(X2)	*pro*(X1) ∧ ¬ *pro*(X2)	→ *disagree*(X1, X2)
		¬ *disagree*(X1, X2) ∧ ¬ *pro*(X1)	→ ¬ *pro*(X2)	*pro*(X1) ∧ *pro*(X2)	→ ¬ *disagree*(X1, X2)
		disagree(X1, X2)	= 1	¬ *pro*(X1) ∧ ¬ *pro*(X2)	→ ¬ *disagree*(X1, X2)

Figure 2: PSL rules to define the collective classification models, both for post-level and author-level models. Each X is an author or a post, depending on the level of granularity that the model is applied at. The *disagree*(X_1, X_2) predicates apply to post reply links, and to pairs of authors connected by reply links.

Author-Level versus Post-Level. When modeling debates, stance classifiers can predict either the stance of a debate participant (i.e. an *author* (**A**)) (Burfoot et al., 2011), or the stance expressed by a specific dialogue act (i.e. a *post* (**P**)) (Hasan and Ng, 2013). The choice of prediction target may depend on the downstream goal, such as user modeling or the study of the dialogic expression of disagreement. From a philosophical perspective, authors are individuals who hold opinions, while posts are not. A post is simply a piece of text which may or may not express the opinions of its author.

Nevertheless, given a prediction target, either author or post, it may be beneficial to consider modeling at a different level of granularity. For example, Hasan and Ng (2013) find that post-level prediction accuracy can be improved by "clamping" all posts by a given author to the same stance in order to smooth their labels. Alternatively, author-level predictions may potentially be improved by first treating each post separately, thereby effectively giving a classifier more training examples, i.e. the number of *posts* instead of the number of *authors*. With this procedure, a final author-level prediction can be obtained by averaging the predictions over the posts for the author, trading the noisiness of post-level instances against the smoothing afforded by the final aggregation. When designing a stance classifier, the modeler must decide the level of granularity for the prediction target and find the best model therein.

4 A Collective Classification Framework

To study these choices, we build a flexible stance classification framework that implements the above variations using probabilistic soft logic (PSL) (Bach et al., 2015; Bach et al., 2013), a recently introduced probabilistic programming system. Like other probabilistic modeling frameworks, notably Markov logic (Richardson and Domingos, 2006), PSL uses a logic-like language for defining the potential functions for a conditional random field. However, unlike Markov logic, PSL makes inference tractable, even in the loopy author-level networks and the very large post-level networks of online debates.

PSL's tractability arises from the use of a special class of conditional random field models referred to as hinge-loss MRFs (HL-MRFs), which admit efficient, scalable and exact maximum a posteriori (MAP) inference (Bach et al., 2013). These models are defined over continuous random variables, and MAP inference is a convex optimization problem over these variables. Formally, a hinge-loss MRF defines a probability density function of the form

$$P(\mathbf{Y}|\mathbf{X}) = \frac{1}{\mathcal{Z}} \exp\Big(- \sum_{r=1}^{M} \lambda_r \phi_r(\mathbf{Y}, \mathbf{X})\Big) , \quad (1)$$

where the entries of $\mathbf{Y}$ and $\mathbf{X}$ are in $[0, 1]$, λ is a vector of weight parameters, $\mathcal{Z}$ is a normalization constant, and

$$\phi_r(\mathbf{Y}, \mathbf{X}) = (\max\{l_r(\mathbf{Y}, \mathbf{X}), 0\})^{\rho_r} \qquad (2)$$

is a *hinge-loss potential* specified by a linear function l_r and optional exponent $\rho_r \in \{1, 2\}$. Given a collection of first-order PSL rules, each instantiation of the rules maps to a hinge-loss potential function as in Equation 2, and the potential functions define an HL-MRF model. For example, $a \Rightarrow b \triangleq \max(a - b, 0)$, where a and b are ground variables, and $\max(a - b, 0)$ is a convex relaxation of logical implication, and which can be understood as its *distance to satisfaction*. For a full description of PSL, see (Bach et al., 2015).

The models we introduce are specified by the PSL rules in Fig. 2, with both post-level and author-level models following the same design. We denote the different modeling choices with the letters defined in Section 3. First, local logistic regression classifiers output stance probabilities based on textual features of posts or authors. All of the models begin with these real-valued stance predictions, encoded by the observed predicate *localPro*(X_i). The rules listed for all models encourage the inferred global predictions *pro*(X_i) to match these local predictions.

This defines the *local classification* models **L**, which are HL-MRFs with node potentials and no edge potentials, and which are equivalent to the local classifiers. The collective models extend the **L** models by adding edge potentials which encourage the stance labels to respect disagreement relationships along reply links. Specifically, every reply link between authors (for author-level models) or between posts (for post-level models) x_1 and x_2 is associated with a latent variable *disagree*$(x_1,\ x_2)$. The rules encourage the global stance variables to respect the polarity of the disagreement variables (same stance, or opposite stance) and while also trying to match the stance classifiers. For the models that do not explicitly model disagreement, it is assumed that every reply edge constitutes a disagreement, i.e. *disagree*$(x_1, x_2) = 1$. These models are denoted **C**.

Otherwise, the disagreement variables are encouraged to match binary-valued predictions from the local disagreement classifiers. We binarize the predictions of the disagreement classifiers to encourage propagation. The disagreement variables are modeled jointly with the stance variables, and label information propagates in both directions between stance and disagreement variables. The full joint stance/disagreement collective models are denoted **D**. In the following, the models are denoted by pairs of letters according to their collectivity level and modeling granularity. For example, **AC** denotes collective classification performed at the author level, without joint modeling of disagreement. To train these models and use them for prediction, weight learning and MAP inference are performed using the structured perceptron algorithm and ADMM algorithm of Bach et al. (2013).

5 Experimental Evaluation

The goals of our experiments were to validate the proposed collective modeling framework, and to make substantive conclusions about the merits of the different possible modeling options described in Section 3. To this end, we evaluated the models on eight topics from 4FORUMS.COM (Walker et al., 2012b) and CREATEDEBATE.COM (Hasan and Ng, 2013), for classification tasks at both the author level and the post level. With comparison to Hasan and Ng (2013), our collective models (**C**) are essentially equivalent to their CRF, up to the form of the CRF potential function, which is not explicitly specified in the paper. A further goal of our experiments was to determine whether the modeling options in our more general CRF could improve performance over models with this structure.

On average, each topic-wise data set contains hundreds of authors and thousands of posts. The 4FORUMS data sets are annotated for stance at the author level, while CREATEDEBATE has stance labels at the post level. To perform post-level evaluations on 4FORUMS we apply author labels to the posts of each author, and on CREATEDEBATE we computed author labels by selecting the majority label of their posts. For 4FORUMS, since post-level stance labels correspond directly to author-level stance labels, we use averages of post-level predictions as the local classifier output for authors. Section 2 includes an overview of these debate forum data sets.

In the experiments, classification accuracy was estimated via five repeats of 5-fold cross-validation. In each fold, we ran logistic regression using the scikit-learn software package,[2] using the default settings, except for the L1 regularization trade-off parameter C which was tuned on a within-fold hold-out set consisting of 20% of the discussions within the fold. For the collective models, weight learning was performed on the same in-fold tuning sets. We trained via 700 iterations of structured perceptron, and ran the ADMM MAP inference algorithm to convergence at test time. On average, weight learning and inference took around 1 minute per fold.

The full results for author-level and post-level predictions are given in Table 2 and Table 3, respectively. In the tables, entries in bold identify statistically significant differences from the local classifier baseline under a paired t-test with significance level $\alpha = 0.05$. These results are summarized in Fig. 3, which shows box plots for the six possible models, computed over the final cross-validated accuracy scores of each of the four data

[2] Available at `http://scikit-learn.org/`.

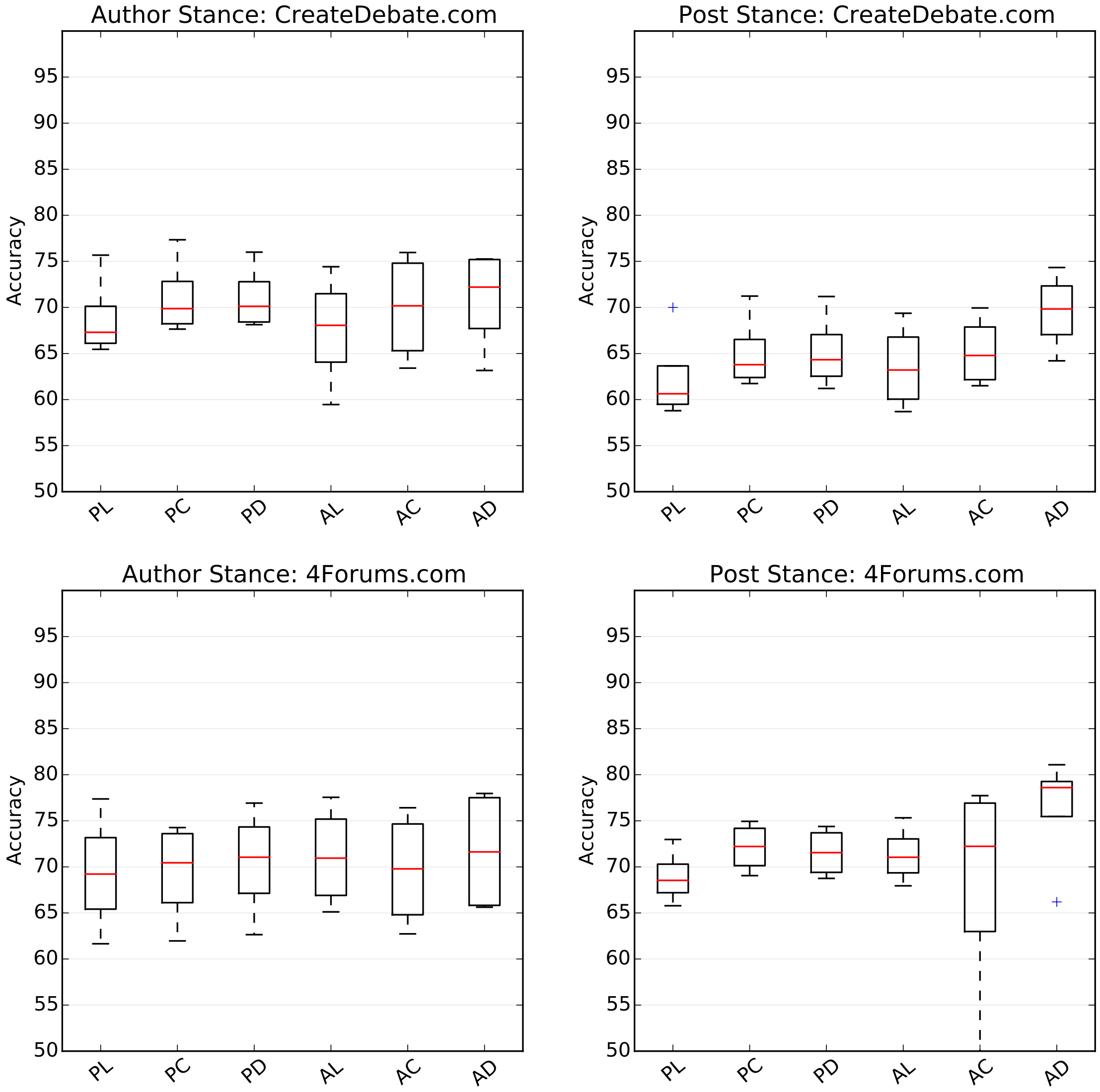

Figure 3: Overall accuracies per model for the author stance prediction task, computed over the final results for each of the four data sets per forum. Note that we expect significant variation in these plots, as the data sets are of varying degrees of difficulty.

sets from each forum. The overall trends can be seen by reading the box plots in each figure from left to right. In general, collective models outperform local models, and modeling disagreement further improves accuracy. Author-level modeling is typically better than post-level, even for the post-level prediction task. The improvements shown by collective models and author-level models are consistent with Hasan and Ng (2013)'s conclusion about the benefits of user-level constraints. This may suggest that posts only provide relatively noisy observations of the underlying author-level stance. Modeling at the author level results in more stable predictions, as noisy posts are pooled together. But here we also show that the full joint disagreement model at the author level, **AD**, performs the best overall, for both prediction tasks and for both forums, gaining up to 11.5 percentage points of post-level accuracy over the local post-level classifier.

A closer analysis reveals some subtleties. When comparing **D** models with **C** models in Fig. 3, disagreement modeling makes a much bigger difference at the author level than at the post level. This is likely impacted by the level of class imbalance for *disagreement* classification in the different levels of modeling. Disagreement, rather than agreement, between authors prompts many responses. Thus, reply links are more likely disagreements when measured at the post level, as seen in Ta-

	4Forums				CreateDebate			
Models	Abortion	Evolution	Gay Marriage	Gun Control	Abortion	Gay Rights	Marijuana	Obama
PL	61.9 ± 4.3	76.6 ± 3.9	72.0 ± 3.6	66.4 ± 4.6	66.4 ± 5.2	70.2 ± 5.0	74.1 ± 6.5	**63.8 ± 8.7**
PC	63.4 ± 5.9	74.6 ± 4.1	73.7 ± 4.3	68.3 ± 5.5	**68.7 ± 5.7**	**72.6 ± 5.6**	75.4 ± 7.4	**66.1 ± 8.5**
PD	63.0 ± 5.4	76.7 ± 4.2	73.7 ± 4.6	67.9 ± 5.0	**69.5 ± 5.7**	**73.2 ± 5.9**	74.7 ± 7.0	**66.1 ± 8.5**
AL	64.9 ± 4.2	77.3 ± 2.9	74.5 ± 2.9	67.1 ± 4.5	65.2 ± 6.5	69.5 ± 4.4	74.0 ± 6.6	59.0 ± 7.5
AC	**66.0 ± 5.0**	74.4 ± 4.2	75.7 ± 5.1	61.5 ± 5.6	65.8 ± 7.0	**73.6 ± 3.5**	73.9 ± 7.6	**62.5 ± 8.3**
AD	**65.8 ± 4.4**	**78.7 ± 3.3**	**77.1 ± 4.4**	67.1 ± 5.4	**67.4 ± 7.5**	**74.0 ± 5.3**	74.8 ± 7.5	**63.0 ± 8.3**

Table 2: Author stance classification accuracy and standard deviation for 4Forums (*left*) and CreateDebate (*right*), estimated via 5 repeats of 5-fold cross-validation. Bolded figures indicate statistically significant ($\alpha = 0.05$) improvement over AL, the baseline model for the author stance classification task.

	4Forums				CreateDebate			
Models	Abortion	Evolution	Gay Marriage	Gun Control	Abortion	Gay Rights	Marijuana	Obama
PL	66.1 ± 2.5	72.4 ± 4.2	69.0 ± 2.7	67.8 ± 3.5	60.2 ± 3.2	62.7 ± 4.4	68.1 ± 6.1	59.4 ± 6.0
PC	**70.5 ± 2.5**	**74.1 ± 3.8**	**73.2 ± 3.1**	**69.1 ± 3.0**	**62.8 ± 3.8**	**66.1 ± 4.9**	68.7 ± 7.9	**61.1 ± 6.6**
PD	**69.7 ± 2.5**	**73.9 ± 4.0**	**72.5 ± 3.0**	**68.8 ± 3.0**	**62.6 ± 4.1**	**66.2 ± 5.4**	69.1 ± 7.4	**61.0 ± 6.6**
AL	**74.7 ± 7.1**	73.0 ± 5.7	70.3 ± 6.0	68.7 ± 5.3	61.6 ± 9.8	63.7 ± 5.3	66.7 ± 6.7	59.7 ± 13.6
AC	**76.8 ± 8.1**	68.3 ± 5.3	72.7 ± 11.1	46.9 ± 8.0	63.4 ± 12.4	**71.2 ± 8.4**	66.9 ± 9.0	63.7 ± 15.6
AD	**77.0 ± 8.9**	**80.3 ± 5.5**	**80.5 ± 8.5**	65.4 ± 8.3	**66.8 ± 12.2**	**72.7 ± 8.9**	69.0 ± 8.3	63.5 ± 16.3

Table 3: Post stance classification accuracy and standard deviations for 4Forums (*left*) and CreateDebate (*right*), estimated via 5 repeats of 5-fold cross-validation. Bolded figures indicate statistically significant ($\alpha = 0.05$) improvement over PL, the baseline model for the post stance classification task.

ble 1. Therefore, enforcing disagreement may be a better assumption at the post level, and the nuanced disagreement model is not necessary in this case. The overall improvements in accuracy from disagreement modeling for post-level models were small.

Text	**Stance**
Post: I agree with everything except the last part. Safe gun storage is very important, and sensible storage requirements have two important factors.	ANTI
Reply: I can agree with this. And in case it seemed otherwise, I know full well how to store guns safely, and why it's necessary. My point was that I don't like the idea of such a law, especially when you consider the problem of enforcement.	ANTI

Figure 4: A post-reply pair by 4Forums.com authors whose gun control stance is correctly predicted by **AD**, but not by **AC**.

On the other hand, the assumption that reply edges constitute disagreement is less accurate when modeling at the author level (see Table 1). In this case, the full joint disagreement model is necessary to obtain good performance. In an extreme example, the two datasets with the lowest disagreement rates at the author level are evolution (44.4%) and gun control (50.7%) from 4Forums. The **AC** classifier performed very poorly for these data sets, dropping to 46.9% accuracy in one instance, as the "opposite stance" assumption did not hold (Tables 2 and 3). The full joint disagreement model **AD** performed much better, in fact achieving an outstanding accuracy rates of 80.3% and 80.5% for posts on evolution and gay marriage respectively. To illustrate the benefits of author-level disagreement modeling, Fig. 4 shows a post for an author whose stance towards gun control is correctly predicted by **AD** but not the **AC** model, along with a subsequent reply. The authors largely agree with each other's views, which the joint disagreement model leverages, while the simpler collective model encourages opposite stance due to the presence of reply links between them.

To summarize our conclusions from these experiments, the results suggest that author-level modeling is the preferred strategy, regardless of the prediction task. In this scenario, it is essential to explicitly model disagreement in the collective classifier. Our top performing **AD** model statistically significantly outperforms the respective prediction task baseline on 6 out of 8 topics for both tasks with p-values less than 0.001. Based on our experimental results, we recommend the full author-disagreement model **AD** as the classifier of choice.

6 Discussion and Future Work

The prediction of user stance in online debate forums is a valuable task, and modeling debate dialogue is complex and requires many decisions such collective or non-collective reasoning, nuanced or naive use of disagreement information, and post versus author-level modeling granularity. We systematically explore each choice, and in doing so build a unified joint framework that incorporates each salient decision. Our method uses a hinge-loss Markov random field to encourage consistency between local classifier predictions for stance and disagreement information. We find that modeling at the author level gives better predictive performance regardless of the granularity of the prediction task, and that nuanced disagreement modeling is of particular importance for author-level collective modeling. The resulting collective classifier gives improved predictive performance over both the simple non-collective and standard collective approaches, with a running time overhead of only a few minutes, thanks to the efficient nature of hinge-loss MRFs.

There are many directions for future work. Our results have found that collective reasoning can also be beneficial at the post level, as previously reported by Hasan and Ng (2013). It is likely that a multi-level model for a combination of post- and author-level collective modeling of both stance and disagreement could bring further improvements in performance. It would also be informative to explore dynamic models which elucidate trends of opinions over time. Another direction is to model influence between users in online debate forums, and to identify the most influential users who are able to convince other users to change their opinions. Finally, we note that stance and disagreement classification are both challenging and important problems, and going forward, there is likely to be much room for improvement in these prediction tasks.

Acknowledgments

This work was supported by NSF grant IIS1218488, and IARPA via DoI/NBC contract number D12PC00337. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

Rob Abbott, Marilyn Walker, Jean E. Fox Tree, Pranav Anand, Robeson Bowmani, and Joseph King. 2011. How can you say such things?!?: Recognizing disagreement in informal political argument. In *ACL Workshop on Language and Social Media*.

Amjad Abu-Jbara and Dragomir R Radev. 2013. Identifying opinion subgroups in Arabic online discussions. In *ACL*.

Stephen H. Bach, Bert Huang, Ben London, and Lise Getoor. 2013. Hinge-loss Markov random fields: Convex inference for structured prediction. In *Uncertainty in Artificial Intelligence (UAI)*.

S. H. Bach, M. Broecheler, B. Huang, and L. Getoor. 2015. Hinge-loss Markov random fields and probabilistic soft logic. arXiv:1505.04406 [cs.LG].

Alexandra Balahur, Zornitsa Kozareva, and Andres Montoyo. 2009. Determining the polarity and source of opinions expressed in political debates. *Computational Linguistics and Intelligent Text Processing*.

Mohit Bansal, Claire Cardie, and Lillian Lee. 2008. The power of negative thinking: Exploiting label disagreement in the min-cut classification framework. *COLING*.

Filip Boltužić and Jan Šnajder. 2014. Back up your stance: recognizing arguments in online discussions. In *ACL Workshop on Argumentation Mining*.

Clinton Burfoot, Steven Bird, and Timothy Baldwin. 2011. Collective classification of congressional floor-debate transcripts. In *ACL*.

Ming-Wei Chang, Lev Ratinov, and Dan Roth. 2012. Structured learning with constrained conditional models. *Machine learning*, 88(3):399–431.

Kuzman Ganchev, Joao Graça, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *Machine Learning*, 11:2001–2049.

Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. 2014. Analyzing argumentative discourse units in online interactions. In *ACL Workshop on Argumentation Mining*.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. *International Joint Conference on Natural Language Processing*.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? Identifying and classifying reasons in ideological debates. In *EMNLP*.

Y. Lu, H. Wang, C. Zhai, and D. Roth. 2012. Unsupervised discovery of opposing opinion networks from forum discussions. In *CIKM*.

Gideon S Mann and Andrew McCallum. 2010. Generalized expectation criteria for semi-supervised learning with weakly labeled data. *Machine Learning*, 11:955–984.

Akiko Murakami and Rudy Raymond. 2010. Support or Oppose? Classifying positions in online debates from reply activities and opinion expressions. In *ACL*.

Matthew Richardson and Pedro Domingos. 2006. Markov logic networks. *Machine learning*, 62(1-2).

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *ACL and AFNLP*.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*.

Dhanya Sridhar, Lise Getoor, and Marilyn Walker. 2014. Collective stance classification of posts in online debate forums. In *ACL Joint Workshop on Social Dynamics and Personal Attributes in Social Media*.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *EMNLP*.

Marilyn Walker, Pranav Anand, Rob Abbott, Jean E. Fox Tree, Craig Martell, and Joseph King. 2012a. That's your evidence?: Classifying stance in online political debate. *Decision Support Sciences*.

Marilyn Walker, Pranav Anand, Robert Abbott, and Jean E. Fox Tree. 2012b. A corpus for research on deliberation and debate. In *LREC*.

Marilyn Walker, Pranav Anand, Robert Abbott, and Richard Grant. 2012c. Stance classification using dialogic properties of persuasion. In *NAACL*.

Lu Wang and Claire Cardie. 2014. A piece of my mind: A sentiment analysis approach for online dispute detection. In *ACL*.