Decision Support Systems 53 (2012) 719–729

Contents lists available at SciVerse ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

That is your evidence?: Classifying stance in online political debate

Marilyn A. Walker [a,*], Pranav Anand [a], Rob Abbott [a], Jean E. Fox Tree [a], Craig Martell [b], Joseph King [a]

[a] *Natural Language and Dialogue Systems Lab, University of California Santa Cruz, USA*
[b] *Natural Language Processing Lab, Naval Postgraduate School, USA*

ARTICLE INFO

Available online 22 May 2012

Keywords:
Online debate
Argumentation strategies
Dialogue
Stance
Automatic classification

ABSTRACT

A growing body of work has highlighted the challenges of identifying the stance that a speaker holds towards a particular topic, a task that involves identifying a holistic subjective disposition. We examine stance classification on a corpus of 4731 posts from the debate website ConvinceMe.net, for 14 topics ranging from the playful to the ideological. We show that ideological debates feature a greater share of rebuttal posts, and that rebuttal posts are significantly harder to classify for stance, for both humans and trained classifiers. We also demonstrate that the number of subjective expressions varies across debates, a fact correlated with the performance of systems sensitive to sentiment-bearing terms. We present results for classifying stance on a per topic basis that range from 60% to 75%, as compared to unigram baselines that vary between 47% and 66%. Our results suggest that features and methods that take into account the dialogic context of such posts improve accuracy.

1. Introduction

Recent work has highlighted the challenges of identifying the stance that a speaker holds towards a particular political, social or technical topic [5,7,8,14,20,21,31,32,34]. Stance is defined as an overall position held by a person towards an object, idea or position [31]. Stance is similar to point of view or perspective, and has been treated as identifying the "side" that a speaker is on, e.g. for or against capital punishment, as illustrated in Fig. 1. Its classification involves identifying a holistic subjective disposition, beyond the word or sentence.

This paper utilizes 104 two-sided debates from Convinceme.net for 14 different debate topics. On Convinceme, a person starts a debate by posting a topic or a question and providing sides such as *for* vs. *against*. Debate participants can then post arguments for one side or the other, essentially self-labeling their post for stance. These debates may be heated and emotional, discussing weighty issues such as euthanasia and capital punishment, as in Fig. 1, but they also appear to be a form of entertainment via playful debate. Popular topics on Convinceme.net over the past 4 years include discussions of the merits of Cats vs. Dogs, or Pirates vs. Ninjas (almost 1000 posts) (see Fig. 2). The full corpus consists of 2902 debates and 36,307 posts by 3637 authors. As indicated above, this work focuses on a subset of these.

Our long term goal is to understand the discourse and dialogic structure of such conversations. This could be useful for: (1) creating automatic summaries of each position on an issue [16,30]; (2) gaining a deeper understanding of what makes an argument persuasive [18,23]; and (3) identifying the linguistic reflexes of perlocutionary acts such as persuasion and disagreement [14,22,32,35,36]. While it seems unlikely that summaries of playful topics would be useful, we believe it is very useful to compare and contrast the dialogic structure of the idealogical topics with that of the playful or technical topical debates. Table 1 provides an overview of our corpus.

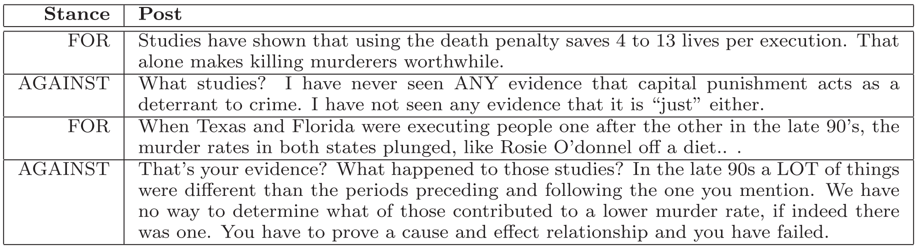
Convinceme provides three possible sources of dialogic structure: (1) the side that a post is placed on indicates the poster's stance with respect to the original debate title and its framing initial posts, and thus can be considered as a response to the title and framing posts; (2) rebuttal links between posts which are explicitly indicated by the poster using the affordances of the site; and (3) the temporal context of the debate, i.e. the state of the debate at a particular point in time, which a debate participant orients to in framing their post. Convinceme provides no way to explicitly indicate agreement with a prior speaker, beyond placing a post on the same side; this does not imply any specify reply-to structure, as rebuttal links do.

Convinceme's support for rebutting a previous post allows the speaker to explicitly mark some debate posts as fundamentally dialogic, while other posts make less use of the immediate context and thus have fewer dialogic properties [6,9,11]. Compare the dialogic aspects of the death penalty debate in Fig. 1 to that of the same topic without rebuttal links in Fig. 3. As shown in the rebuttals column of Table 1, the percentage of rebuttals by topic varies from 34% to 80%. Ideological topics (below the line) have a much higher percentage of rebuttals. We show below that the performance of automatic stance classifiers is better for discussions containing many rebuttal links when the dialogue context is included in the feature set provided to the classifier.

Section 2 first describes related work. Section 3 discusses our corpus in more detail. Given the dialogic nature of our data, as indicated by the high percentage of rebuttals in the ideological debates, we first aim to determine how difficult it is for humans to side an individual post

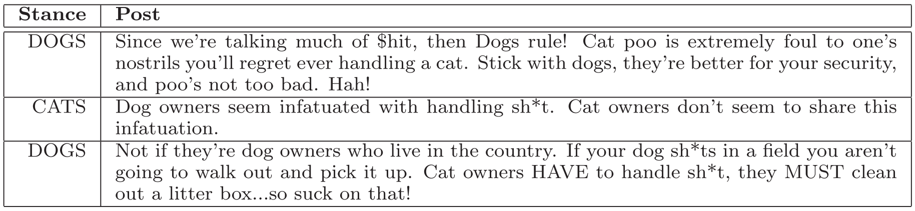
[*] Corresponding author.
E-mail addresses: maw@soe.ucsc.edu (M.A. Walker), panand@soe.ucsc.edu (P. Anand), abbott@soe.ucsc.edu (R. Abbott).

0167-9236/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2012.05.032



Stance	Post
FOR	Studies have shown that using the death penalty saves 4 to 13 lives per execution. That alone makes killing murderers worthwhile.
AGAINST	What studies? I have never seen ANY evidence that capital punishment acts as a deterrant to crime. I have not seen any evidence that it is "just" either.
FOR	When Texas and Florida were executing people one after the other in the late 90's, the murder rates in both states plunged, like Rosie O'donnel off a diet.. .
AGAINST	That's your evidence? What happened to those studies? In the late 90s a LOT of things were different than the periods preceding and following the one you mention. We have no way to determine what of those contributed to a lower murder rate, if indeed there was one. You have to prove a cause and effect relationship and you have failed.

Fig. 1. Dialogic death penalty discussion with posts explicitly linked via rebuttal links. The discussion topic was "Death Penalty," and the argument was framed as yes we should keep it vs. no we should not.



Stance	Post
DOGS	Since we're talking much of $hit, then Dogs rule! Cat poo is extremely foul to one's nostrils you'll regret ever handling a cat. Stick with dogs, they're better for your security, and poo's not too bad. Hah!
CATS	Dog owners seem infatuated with handling sh*t. Cat owners don't seem to share this infatuation.
DOGS	Not if they're dog owners who live in the country. If your dog sh*ts in a field you aren't going to walk out and pick it up. Cat owners HAVE to handle sh*t, they MUST clean out a litter box...so suck on that!

Fig. 2. Cats vs. Dogs discussions with posts linked by rebuttal links.

from a debate without context. Section 3 presents the results of a human debate-side classification task conducted on Mechanical Turk. Section 4 describes experiments for automatically determining stance, and presents our results. Our overall results show that using sentiment, subjectivity, dependency and dialogic features, we can achieve debate-side classification accuracies, on a per topic basis, that range from 60% to 75%, as compared to unigram no-context baselines that vary between 47% and 66%. We show that even a naive representation of context uniformly improves results across all topics. We also conduct an experiment to classify rebuttals, as a type of disagreement discourse relation, and show that we can identify rebuttals with 63% accuracy.

2. Related work

There are several threads of related work that focuses on classifying a speaker's "side" or "stance" toward a debate topic in either formal or informal debate settings, such as congressional floor debates or in conversations from online forums and debate websites [3,34,38].

The research most strongly related to our own is that of Somasundaran and Wiebe [31,32], who also report results for automatically determining the stance of a debate participant in online forums. The websites that their corpus was collected from apparently did not support dialogic threading, so that there are no explicitly linked rebuttals in their corpus. They present different results for stance classification for ideological vs. non-ideological topics, and utilize a number of different approaches, including an unsupervised method that finds relevant terms from the web, and an inductive logic programming approach that builds on the assumption that speakers are self-consistent with respect to their stance on a particular topic and its attributes. They also show that discourse relations such as concessions and the identification of argumentation triggers improves performance over sentiment features alone. Their best performance for siding ideological debates is approximately 64% accuracy over all topics, for a collection of 2nd Amendment, Abortion, Evolution, and Gay Rights debate posts [32]. Their best performance is 70% for the 2nd amendment topic. Their work, along with others, indicates that for such tasks it is difficult to beat a unigram baseline [26].

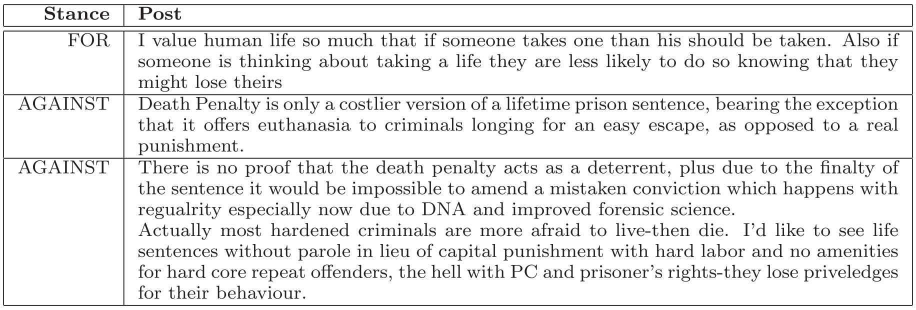
The other significant body of work that we build on classifies the speaker's side in a corpus of congressional floor debates, using the speaker's final vote on the bill as a labeling for side [4,5,34,39]. This work infers agreement between speakers based on cases where one speaker mentions another by name, and a simple algorithm for determining the polarity of the sentence in which the mention occurs. This work shows that even with the resulting sparsely connected agreement structure, the MinCut algorithm can improve over stance classification based on textual information alone.

Other work has utilized the reply structure of online forums, either with or without textual features of particular posts [2,21,24,25]. The threading structure of these debates does not distinguish between agreement and disagreement responses, so Agrawal et al. [2] assume that adjacent posts always disagree, based on the results of Mishne and Glance [24] who showed that most replies to blog posts are disagreements. Murakami and Raymond [25] show that simple rules for identifying disagreement, defined on the textual content of the post, can improve over Agarwal's results. Malouf and Mullen [21] also show that a combination of textual and response structure features provides the best performance.

Other related work analyzes forum quote/response structures [1,37]. Quote/response pairs have a similar discourse structure to the rebuttal post pairs in Convinceme, but are often shorter and more targeted; this may mean that they are easier to classify because the linguistic reflexes of stance are expressed very locally. Wang and Rose [37] use unlabelled data, and do not attempt to distinguish between the agreement and disagreement discourse relations across quote/response pairs. Rather they show that they can use a variant of LSA to identify a parent post, given a response post, with approximately 70% accuracy.

Table 1
Threading characteristics of different topics. Topics below the line are considered "ideological." Key: number of posts on the topic (posts), percent of posts linked by rebuttal links (rebuttals), posts per author (P/A). Authors with more than one post (A > 1P). Average post length in characters (length).

Topic	Discussions	Posts	Rebuttals	P/A	A > 1p	Length
Cats vs. dogs	3	162	40%	1.68	26%	242
Firefox vs. IE	2	233	40%	1.28	16%	167
Mac vs. PC	7	126	47%	1.85	24%	347
Superman/Batman	4	146	34%	1.41	21%	302
2nd Amendment	6	134	59%	2.09	45%	385
Abortion	10	607	70%	2.82	43%	339
Climate change	6	207	69%	2.97	40%	353
Communism vs. capitalism	6	207	70%	3.03	47%	348
Death penalty	12	331	62%	2.44	45%	389
Evolution	16	818	76%	3.91	55%	430
Exist God	16	852	77%	4.24	52%	336
Gay marriage	6	560	65%	2.12	29%	401
Healthcare	5	112	80%	3.24	56%	280
Marijuana legalization	5	236	52%	1.55	26%	423



Stance	Post
FOR	I value human life so much that if someone takes one than his should be taken. Also if someone is thinking about taking a life they are less likely to do so knowing that they might lose theirs
AGAINST	Death Penalty is only a costlier version of a lifetime prison sentence, bearing the exception that it offers euthanasia to criminals longing for an easy escape, as opposed to a real punishment.
AGAINST	There is no proof that the death penalty acts as a deterrent, plus due to the finalty of the sentence it would be impossible to amend a mistaken conviction which happens with regualrity especially now due to DNA and improved forensic science. Actually most hardened criminals are more afraid to live-then die. I'd like to see life sentences without parole in lieu of capital punishment with hard labor and no amenities for hard core repeat offenders, the hell with PC and prisoner's rights-they lose priveledges for their behaviour.

Fig. 3. Monologic Death Penalty discussion. The posts have no explicit link structure. The discussion topic was "Death Penalty," and the argument was framed as yes we should keep it vs. no we should not.

Abbott et al. [1] examine agreement and disagreement relations across quote/response pairs in online forum discussions for a range of ideological and nonideological topics as we do here. Their corpus has been hand-labeled for agreement using Mechanical Turk. They achieve a best accuracy of 68% for classifying whether a post is an agreement or a disagreement with the prior post. Their results also indicate that contextual features improve performance for identifying the agreement relation between quotes and responses.

3. Corpus description and analysis

Our corpus consists of two-sided debates from Convinceme.net for 14 topics that range from playful debates such as Cats vs. Dogs to more heated political topics such as the Death Penalty. Table 1 provides an overview of our corpus; the topics above the line are either technical or playful, while the topics below the line are ideological. As discussed above, Convinceme provides three possible sources of dialogic structure, side, rebuttal links and temporal context. However some of the temporal context is lost when creating a corpus from the Convinceme site: the timestamps for posts are only available by day; thus rather than a total order from time of post, it is only possible to calculate a partial order on posts by day, plus order within day only via rebuttals. In addition, as mentioned above, there are no agreement links. In total the corpus consists of 2902 two-sided debates (36,307 posts), totaling 3,080,874 words; the topic labeled debates which we use in our experiments contain 575,818 words.

Each of our fourteen topics consists of more than one debate. Interestingly, the user who initiates a debate frames the way the two sides are expressed, by specifying the debate title and two originating framing posts for each side. Table 2 shows a sample of debate titles that were mapping to the Death Penalty topic, and Table 3 shows sample debate titles for Evolution. In both tables, the Pro Framing Post and the Con Framing Post columns provide the text that the debate initiator used to frame each side of the current debate discussion. Because debates can frame an issue differently, each debate was mapped by hand to its topic, and the two sides were mapped by hand to the Pro and Con sides, as in Ref. [32]. The N column gives the number of posts in each debate, which might vary quite a lot. Thus, for example, the 331 posts for the Death Penalty topic in Table 1 (9th row), originally consisted of posts in 12 different debate discussions, some of which are shown in Table 2.

This mapping to topic obviously increases the number of posts on each topic, thus increasing the chance of learning a good model for stance classification. On the other hand, the topic mapping means that different discussions on the same topic orient to a somewhat different context because they orient to the debate title and originating posts. This might mean that they focus on slightly different aspects of an issue, or make reference to different originating posts, and the arguments that they express, as exemplified by the different originating posts shown in Tables 2 and 3.

Topics vary a great deal in terms of their dialogic structure and linguistic expression. In Tables 1 and 4, the columns providing counts for different variables were selected to illustrate ways in which topics differ in the form and style of the argument and in its subjective content. As mentioned above, one important variable is the percentage of the topic posts that are linked into a rebuttal dialogic structure (column Rebuttals in Table 1).

Ideological topics display more author investment; people feel more strongly about these issues. This is shown by the fact that there are more rebuttals per topic in the topics below the line in Table 4. All of the ideological topics have more than 50% rebuttals. It follows that these topics have a much higher degree of context-dependence in each post, since posts respond directly to the parent post. Rebuttals exhibit more markers of dialogic interaction: greater pronominalization, especially *you* (Rebuttals $\bar{x}=9.6$ and Non-Rebuttals $\bar{x}=8.5$, $t(27)=24.94$, $p<.001$), as well as propositional anaphora such as *that* and *it*, ellipsis, and dialogic cue words, such as *well* and *so*. Examples of arguments illustrating some of these differences by topic (correlated with the percentage of rebuttals) as shown in Table 4 can be observed by comparing dialogic (rebuttal links) and monologic (no rebuttal links) posts. For example, compare Figs. 1 and 3 for the Death Penalty topic and Figs. 2 and 4 for the Cats vs. Dogs topic.

Another indication of author investment is the percentage of authors with more than one post (A>1P) and the number of posts per author (P/A) in Table 1. The A>1P percentage ranges from 16% for Firefox vs. IE to 56% for Healthcare; it is significantly greater for ideological topics ($t(12)=4.27$, $p=.001$). The posts per author (P/A) variable is also greater for idealogical topics; it ranges from 1.28 for

Table 2
Example of mapping multiple discussions, each initiated with distinct debate titles and framing posts as above, to a single debate topic: death penalty.

Debate title	N	Pro framing post	Con framing post
Should the U.S. continue death penalty executions?	39	Kill them all!	Let them rot in prison!
Should child molesters face the death penalty	28	Yes, fry the bastards	No, just imprison them.
Death Penalty; justice?	22	They should be put to death. An eye for an eye.	They should have life in prison. Two wrongs don't make a right.
Is the death penalty morally correct as it is supposed to be used in the United States?	12	Yes, most of the times it is.	No, it is never morally correct.
Whether to abolish death penalty	31	Execution	Abolish death penalty = Life without parole

Table 3
Example of mapping multiple discussions, each initiated with distinct debate titles and framing posts as above, to a single debate topic: evolution.

Debate title	*N*	Pro framing post	Con framing post
Evolution exists. Get used to it.	4	There is no longer an excuse besides ignorance or overwhelming blind faith to deny evolution in modern society.	Evolution surely exists but the evolution of mankind is still a question.
Evolution	25	Evolution exists	Evolution exists through intelligent design.
Evolution vs. Creation	4	That evolution adequately explains the forms of life we see on Earth.	That God did create the earth and is a more logical explanation of the earth.
Evolution vs. Young Earth Creationism	8	Evolution and science have dismantled young earth creationism	The theory of evolution (not science; science can be useful to discover God's creation) has replaced creation taught in schools but has not dismantled any bit of the concept of creation in terms of evidence or pragmatic regularity.
Evolution vs. Creation (Intelligent Design)	22	Pro-evolution	Evolution is not a fact. That's ******* retarded to assume that everything came from nothing. Order doesn't happen by chance. That's a fact. Check w/ the 2nd law of physics: all things head towards entropy/chaos. It requires an order agent (aka "God")

Firefox vs. IE to 4.24 for discussions about the Existence of God. It appears that for some topics, speakers simply state their position, and feel no need to support their position further, or to rebut others' positions, while for many of the idealogical topics, the speakers engage in an ongoing dialogic argument.

On the other hand, median post length (Length) might be predicted to show author investment, but Length is not correlated with the percentage of authors with more than one post ($r(12)=.36$, $p=.20$). Thus, those who contribute more posts are not necessarily contributing longer (or shorter) posts. However median post length is strongly negatively correlated with words of positive emotion ($r(12)=-.81$, $p<.001$), the longer the post, the fewer positive emotion words. Also contrary to intuition, there were more swear words in the nonideological topics (nonideological $\bar{x} = .19$, ideological $\bar{x} = .06$, $p=.002$). Positive emotion and swear words were also highly correlated ($r(12)=.85$, $p<.001$).

Table 4
Characteristics of different topics. Topics below the line are considered "ideological." Shown are the normalized LIWC variable *z*-scores for each topic. *Z*-scores are significant when more than 1.94 standard deviations away from the mean (two-tailed). Key: Pro = percent of the words as pronominals, WPS = words per sentence, 6LTR = percent of words that are longer than 6 letters, PosE positive emotion words, NegE negative emotion words.

Topic	Pro	WPS	6LTR	PosE	NegE
Cats vs. dogs	3.30	−1.95	−2.43	1.70	0.30
Firefox vs. IE	−0.11	−0.84	0.53	1.23	−0.81
Mac vs. PC	0.52	0.28	−0.85	−0.11	−1.05
Superman/Batman	−0.57	−1.78	−0.43	1.21	0.99
2nd Amendment	−1.38	1.74	0.58	−1.04	0.38
Abortion	0.63	−0.27	−0.41	−0.95	0.68
Climate change	−0.74	1.23	0.57	−1.25	−0.63
Communism vs. capitalism	−0.76	−0.15	1.09	0.39	−0.55
Death penalty	−0.15	−0.40	0.49	−1.13	2.90
Evolution	−0.80	−1.03	1.34	−0.57	−0.94
Exist God	0.43	−0.10	0.34	−0.24	−0.32
Gay marriage	−0.13	0.86	0.85	−0.42	−0.01
Healthcare	0.28	1.54	0.99	0.14	−0.42
Marijuana legalization	0.14	0.37	0.53	−0.86	0.50

Other factors we examined were words per sentence (WPS), the length of words used (6LTR) which typically indicates scientific or low frequency words, the use of pronominal forms (Pro), and the use of positive and negative emotion words (PosE, NegE) [27]. For example, the significant *z*-score values in Table 4 indicate that discussions about Cats vs. Dogs consist of short simple words (*z*-score of LTR is −2.45, more than two standard deviations below the mean for all topics) in short sentences (*z*-score of WPS is −1.95, almost two standard deviations below the mean for all topics), with relatively high usage of positive emotion words (*z*-score of PosE is 1.70), and pronouns (*z*-score of 3.30, more than three standard deviations above the mean), whereas 2nd amendment debates use relatively longer sentences (*z*-score of WPS is 1.74), and death penalty debates (unsurprisingly) use a lot of negative emotion words (*z*-score of NegE is 2.90).

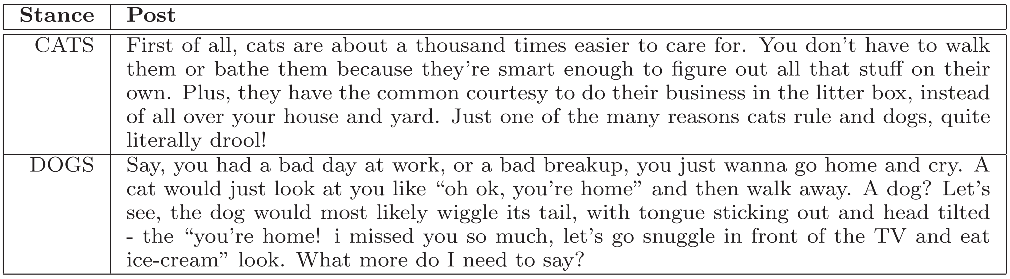
Grouping topics revealed that non-ideological topics had shorter words (6LTR, $t(12)=3.16$, $p=.008$) and more words expressing positive emotion (PosE, $t(12)=4.51$, $p=.001$), but a similar number of negative emotion words (NegE), words per sentence (WPS), and pronominals (PRO). While we hypothesized that different sides of the same debate might use different types of language, e.g. one side might emphasize scientific evidence, while the other side might make an emotional argument, we found no evidence for this. An analysis of debate by side showed that each side of the debates had similar numbers of pronominals, words per sentence, six letter words, positive emotion words, and negative emotion words. Fig. 5 shows the interaction between positive and negative emotion words and whether or not the topics were ideological (main effect of ideology: $F(1, 12)=4.92$, $p=.047$, main effect of type of emotion: $F(1, 12)=10.85$, $p=.006$, interaction: $F(1, 12)=6.89$, $p=.022$). Perhaps surprisingly for a debate website, across both types of topics, there were more positive emotion words than negative emotion words.

3.1. Human topline

To our knowledge, none of the previous work on debate side classification has attempted to establish a human topline for classifying debate posts by side. When examining results from prior work, we were surprised that the best accuracies were around 70%, so we decided to see how well humans would perform at the stance classification task when given the post to be classified, and the post and sides that originated the debate, as illustrated in Tables 2 and 3, but no other context. We believed that this task was the best approximation of what automatic stance classifiers were being asked to do in previous work [31,32,34,39], and thus this task (without context) would provide the best estimate of a human topline.

We set up a Mechanical Turk task by randomly selected a subset of our data excluding the first post on each side of a debate (because we use it to set context) and debates with fewer than 6 posts on either side. We selected equal numbers of posts for each topic for each side, and created 132 tasks (Mechanical Turk HITs). Each HIT consisted of choosing the correct side for 10 posts divided evenly, and selected randomly without replacement, from two debates. For each debate we presented a title, side labels, and the initial post on each side. For each post we presented the first 155 characters with a See More button which expanded the post to its full length. Fig. 6 shows a sample HIT for one discussion for the Death Penalty topic. The top of the page shows the originator's Debate Title. Underneath the debate title on each side is a color-coded layout of the originator's framing of the Pro vs. Con sides of the debate. The sides shown below with each post are color-coded to match the framing posts.

Each HIT was judged by 9 annotators with each annotator restricted to at most 30 HITS (300 judgments). Since many topics were specific to issues with U.S. politics, and we wanted annotators with a good grasp of English, we required Turkers to have a U.S. IP address.



Stance	Post
CATS	First of all, cats are about a thousand times easier to care for. You don't have to walk them or bathe them because they're smart enough to figure out all that stuff on their own. Plus, they have the common courtesy to do their business in the litter box, instead of all over your house and yard. Just one of the many reasons cats rule and dogs, quite literally drool!
DOGS	Say, you had a bad day at work, or a bad breakup, you just wanna go home and cry. A cat would just look at you like "oh ok, you're home" and then walk away. A dog? Let's see, the dog would most likely wiggle its tail, with tongue sticking out and head tilted - the "you're home! i missed you so much, let's go snuggle in front of the TV and eat ice-cream" look. What more do I need to say?

Fig. 4. Monologic Cats vs. Dogs Posts, i.e. there are no explicit rebuttal links between these posts.

Fig. 7 plots the number of annotators over all topics who selected the "true siding" as the side that the post was on. We defined "true siding" for this purpose as the side that the original poster placed their post. Fig. 7 illustrates that humans often placed the post on the wrong side. The majority of posters agreed with the true siding 78.26% of the time. The Fleiss' kappa statistic was .2656.

Importantly and interestingly, annotator accuracy varied across topics in line with rebuttal percentage. Annotators correctly labeled 94 of 100 posts for Cats vs. Dogs but only managed 66 of 100 for the Climate Change topic. This suggests that posts may be difficult to side without context, which is what one might expect given their dialogic nature. Rebuttals were clearly harder to side: annotators correctly sided non-rebuttals 87% of the time, but only managed 73% accuracy for rebuttals. Since all of the less serious topics consisted of ≤50% rebuttals while all of the more serious ideological debates had >50% rebuttals, 76% of ideological posts were sided correctly, while 85% of non-ideological posts were correctly sided (see Table 5).

Looking at the data by hand revealed that when nearly all annotators agreed with each other but disagreed with the self-labeled side, the user posted on the wrong side (either due to user error, or because the user was rebutting an argument the parent post raised, not the actual conclusion).

The difficult-to-classify posts (where only 4–6 annotators were correct) were more complex. Fig. 8 provides some examples of posts that were hard for humans to side. The first is anti-death penalty, but the speaker's stance is due to concerns about the legal system, not the actual punishment. The first sentence, which often can be counted on to provide a summary of a speaker's position, actually states that the speaker *has no problem with killing someone who is a mass murderer*. The last sentence, which does summarize the speaker's position, i.e. that *mistrust of government prevents me from supporting the death penalty*, appears to contradict the first sentence. Only a close reading of this carefully qualified argument supports correct inference of the speaker's position. Some classifiers would also have difficulty siding a post that explicitly appears to support both sides. Of 9 annotators, 8 marked this post incorrectly.

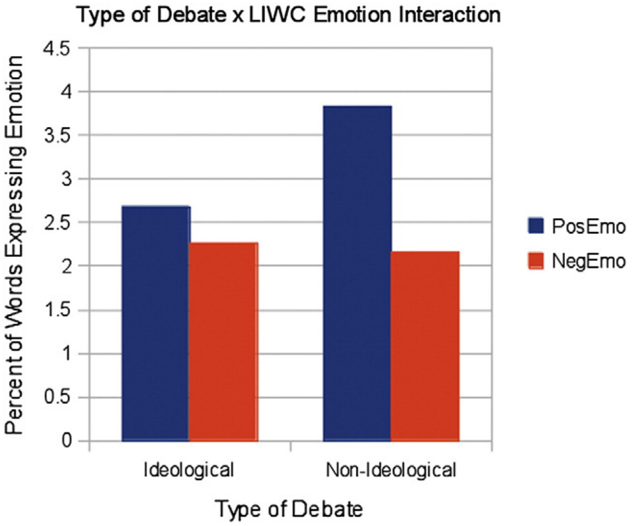


Fig. 5. Use of emotional language vs. debate topic category.

The second post in Fig. 8 from the Superman/Batman topic misspells *kriptonite* which in any case requires domain specific knowledge. The author only reluctantly concedes that Batman has the advantage indicating they are not completely pro-Batman. Automated systems and humans should understand the misspelled term, translate "just cause" to "just because," "batmans" to "Batman's" and, most importantly, recognize that although "superman would own batmans ass" implies Superman > Batman, it is conditioned on there being no kryptonite. Of 9 annotators, 7 marked this post incorrectly.

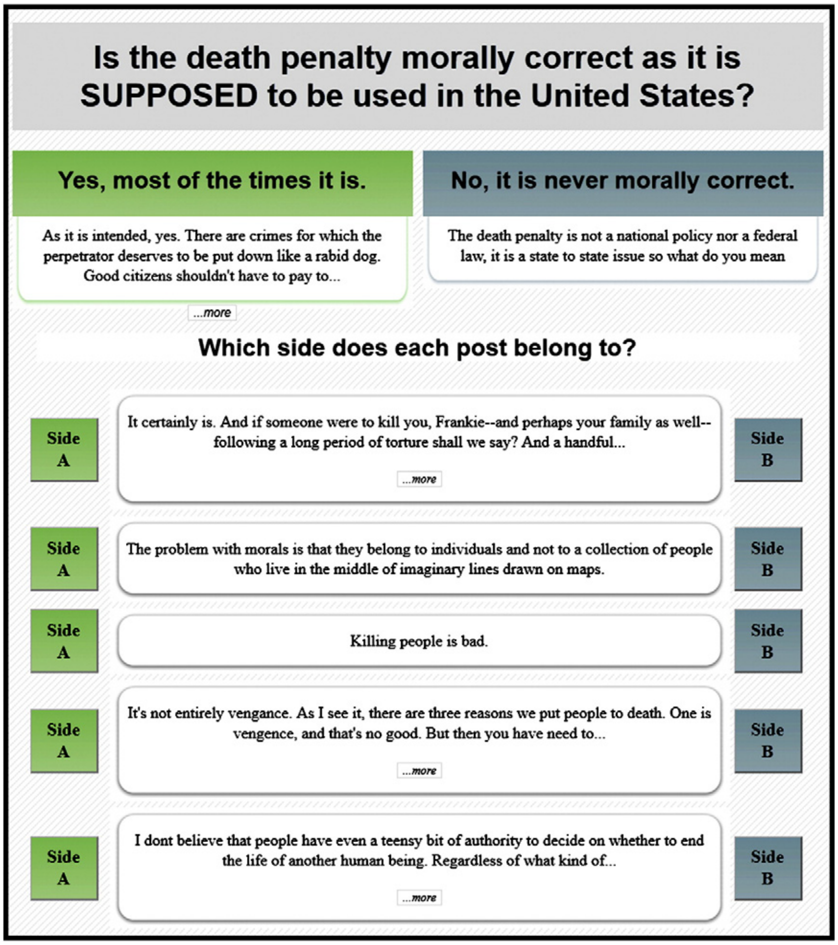
The third post from the Mac vs. PC debate contains a number of concessions at the beginning of the post, where speakers often state their main claim. Moreover, these concessions are not explicitly marked by any concessive cue words [28,29,33]. The fragment *the biggest problem I have with Apple* is the first cue to the reader that these statements are indeed concessive. This post, interestingly, also contains an analogy of Lexus to Mac and PC to Honda, which might be difficult for a machine to understand. Finally the real statement of position *if there was something I could not accomplish on Windows that I could do on an Apple, then that would be a compelling argument* requires an inference that none of the preceding arguments for Apple are compelling. Of 9 annotators, 7 marked this post incorrectly.

Of the remainder of the cases that we classified as "hard to side" on the basis of lack of agreement between annotators, our analysis suggests that 28% of the time the annotators were simply wrong, perhaps only skimming a post when the stance indicator was buried deep inside it. Our decision to show only the first 155 characters of each post by default (with a Show More button) undoubtedly contributed to this error (see Fig. 6). An additional 39% was short comments or ad hominem responses that showed disagreement, but no indication of side, and 17% were ambiguous out of context. A remaining 10% were meta-debate comments, either about whether there were only two sides, or whether the argument was meaningful.

4. Experimental setup and results

We also conducted two types of experiments to classify posts. In one type, we attempted to classify posts based on features drawn from the post and (optionally) its parent alone. For this class of experiments, we trained Naive Bayes, JRIP, and SVM learners. Our Naive Bayes and JRIP experiments were conducted with the Weka toolkit, while our SVM experiments used LibLinear and discovered best parameter settings using grid-search.[1] All results are from 10 fold cross-validation on a balanced test set. In the hand examination of Mechanical Turk annotators' siding performance, 101 posts were

[1] An earlier, exploratory investigation on three of the feature sets below found that a linear kernel outperformed a radial basis kernel. We elected to continue with a linear kernel throughout.



Is the death penalty morally correct as it is SUPPOSED to be used in the United States?

Yes, most of the times it is.

As it is intended, yes. There are crimes for which the perpetrator deserves to be put down like a rabid dog. Good citizens shouldn't have to pay to...

...more

No, it is never morally correct.

The death penalty is not a national policy nor a federal law, it is a state to state issue so what do you mean

Which side does each post belong to?

Side A | It certainly is. And if someone were to kill you, Frankie--and perhaps your family as well--following a long period of torture shall we say? And a handful... ...more | Side B

Side A | The problem with morals is that they belong to individuals and not to a collection of people who live in the middle of imaginary lines drawn on maps. | Side B

Side A | Killing people is bad. | Side B

Side A | It's not entirely vengance. As I see it, there are three reasons we put people to death. One is vengence, and that's no good. But then you have need to... ...more | Side B

Side A | I dont believe that people have even a teensy bit of authority to decide on whether to end the life of another human being. Regardless of what kind of... ...more | Side B

Fig. 6. An example of a Mechanical Turk HIT for the Death Penalty topic.

determined to have incorrect self-labeling for side. We eliminated these posts and their descendants from the experiments detailed below. This resulted in a data set of 3546 posts.

In a second, subsequent series of experiments, inspired by Thomas et al. [34], we investigated the extent to which classification could be improved by classifying all posts by a speaker simultaneously, and to what extent applying the MinCut algorithm exactly as described in Ref. [5] would improve our results. For MinCut, we used the python-graph library, and both Naive Bayes and SVM to provide values for the *Ind* function.

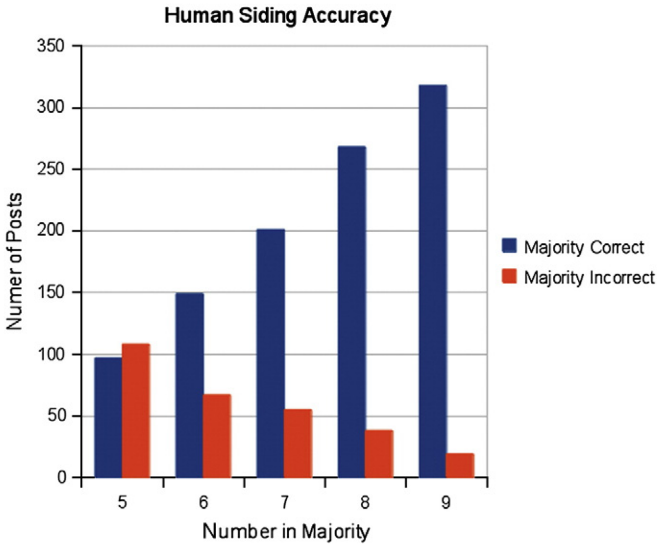


Fig. 7. Accuracies of Human Mechanical Turk judges at selecting the true siding of a post without context.

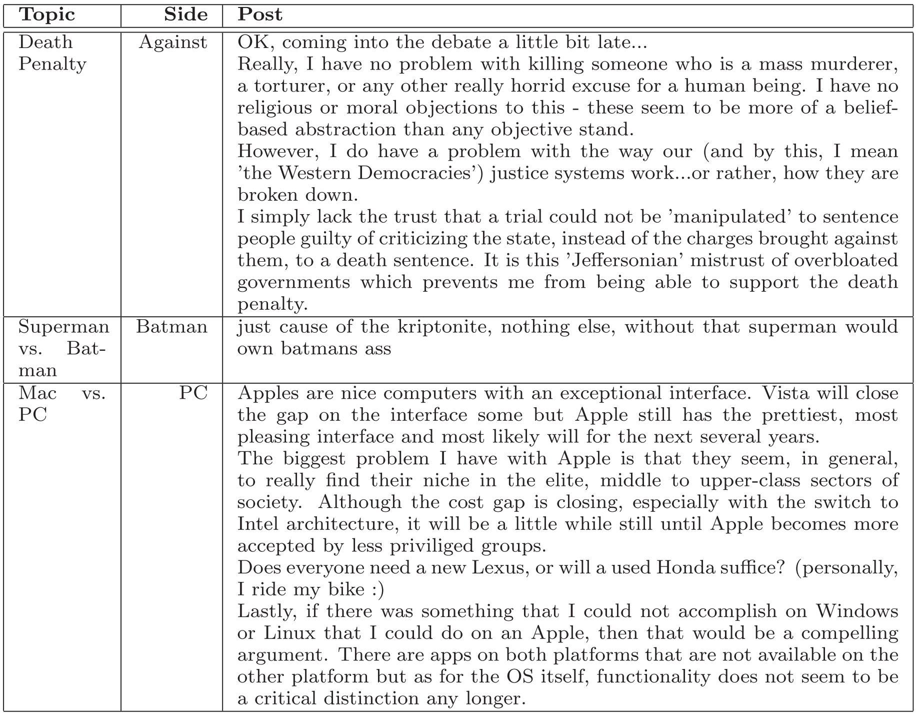
However, below we only report our Naive Bayes results for with and without context for different feature sets. This is because none of the other algorithms improved accuracy. We found this very surprising; clearly more research is needed. First, we were expecting SVM to give us a non-trivial boost in results, especially since we conducted a grid search for the best value of the cost parameter. Although there were isolated cases where SVM beat Naive Bayes, they were few and not systematic.

Second, we tried the simple thing of concatenating all the posts by the same speaker and conducting stance classification on the speaker-based documents. This improved results for Thomas et al. [34], almost as much as using MinCut. Sadly, it did not improve our results.

Third, we expected MinCut to greatly improve classification performance, based on the results reported by Thomas et al. [34] and Bansal et al. [5]. However, at present our results are much worse when we apply MinCut. We hypothesize that a possible explanation for the fact that Bansal et al.'s [5] algorithms (Set-To, Inc-By with MinCut) do not

Table 5
Human agreement on rebuttal classification.

Class	Correct	Total	Accuracy
Rebuttal	606	827	0.73
Non-rebuttal	427	493	0.87



Topic	Side	Post
Death Penalty	Against	OK, coming into the debate a little bit late... Really, I have no problem with killing someone who is a mass murderer, a torturer, or any other really horrid excuse for a human being. I have no religious or moral objections to this - these seem to be more of a belief-based abstraction than any objective stand. However, I do have a problem with the way our (and by this, I mean 'the Western Democracies') justice systems work...or rather, how they are broken down. I simply lack the trust that a trial could not be 'manipulated' to sentence people guilty of criticizing the state, instead of the charges brought against them, to a death sentence. It is this 'Jeffersonian' mistrust of overbloated governments which prevents me from being able to support the death penalty.
Superman vs. Batman	Batman	just cause of the kriptonite, nothing else, without that superman would own batmans ass
Mac vs. PC	PC	Apples are nice computers with an exceptional interface. Vista will close the gap on the interface some but Apple still has the prettiest, most pleasing interface and most likely will for the next several years. The biggest problem I have with Apple is that they seem, in general, to really find their niche in the elite, middle to upper-class sectors of society. Although the cost gap is closing, especially with the switch to Intel architecture, it will be a little while still until Apple becomes more accepted by less priviliged groups. Does everyone need a new Lexus, or will a used Honda suffice? (personally, I ride my bike :) Lastly, if there was something that I could not accomplish on Windows or Linux that I could do on an Apple, then that would be a compelling argument. There are apps on both platforms that are not available on the other platform but as for the OS itself, functionality does not seem to be a critical distinction any longer.

Fig. 8. Examples of posts which proved difficult for Mechanical Turk annotators.

work for us, is that our graphs are much more highly connected than those resulting from analysis of the Congressional Floor Debate corpus, which uses a weak method to infer agreement and disagreement links, based on mentions of another speaker by name, and a calculation of the average polarity around that speaker's name. Our debates produce a much more highly connected graph. See the sample graph for one of the discussions in our corpus in Fig. 9. In addition, we made several assumptions to augment these with agreement links, such as: (1) a speaker always agrees with him or herself; and (2) if two speakers (i, j) both disagree with speaker k, then i and j agree with one another. Our conclusion is that a highly connected graph such as ours may require a very different method, edge pruning, more complex edge weighting, or a different way of tuning MinCut parameters than that reported in previous work.

4.1. Features

Table 6 provides a summary of the features we extract for each post. We describe and motivate these feature sets below.

4.1.1. Post info

This set of features includes basic count features from a post: the number of characters, number of words, and number of sentences in the post, as well as the average words per sentence (WPS) and average word length.

4.1.2. Unigrams, bigrams

Previous work suggests that the unigram baseline can be difficult to beat for certain types of debates [32]. Thus we derived both unigrams and bigrams as features. These were extracted as frequency counts within the post (i.e., normalized by the total number of unigrams or bigrams in the post.)

4.1.3. Cue words

We represent each post's initial unigram, bigram and trigram sequences to capture the usage of cue words to mark responses of particular type, such as *oh really*, *so*, and *well*; these features were based on both previous work and our examination of the corpus [12,13,15].

4.1.4. Repeated punctuation

Our informal analyses suggested that repeated sequential use of particular types of punctuation such as !! and ?? did not mean the same thing as simple counts or frequencies of punctuation across a whole post. Thus we developed distinct features for a subset of these repetitions. These were normalized by the number of unigrams in the post.

4.1.5. LIWC

We also derived features using the Linguistics Inquiry Word Count tool [27]. LIWC provides meta-level conceptual categories for words to use in word counts. Some LIWC features that we expect to be important are words per sentence (WPS), pronominal forms (Pro), and positive and negative emotion words (PosE) and (NegE) (see Table 4). These were normalized by frequency.

4.1.6. Syntactic dependency

Previous research in this area suggests the utility of dependency structure to determine the target of an opinion word [17,31,32]. The dependency parse for a given sentence is a set of triples, composed of a grammatical relation and the pair of words for which the grammatical relation holds (rel$_i$, w_j, w_k), where rel$_i$ is the dependency relation among words w_j and w_k. The word w_j is the head of the dependency relation. We use the Stanford parser to parse the utterances in the posts and extract dependency features [10,19].

4.1.7. POS generalized dependency

To create generalized dependencies, we "back off" the head word in each of the above features to its part-of-speech tag [17]. Joshi and Rosé's results suggested that this approach would work better than either fully lexicalized or fully generalized dependency features. We call these POS generalized dependencies in the results below.

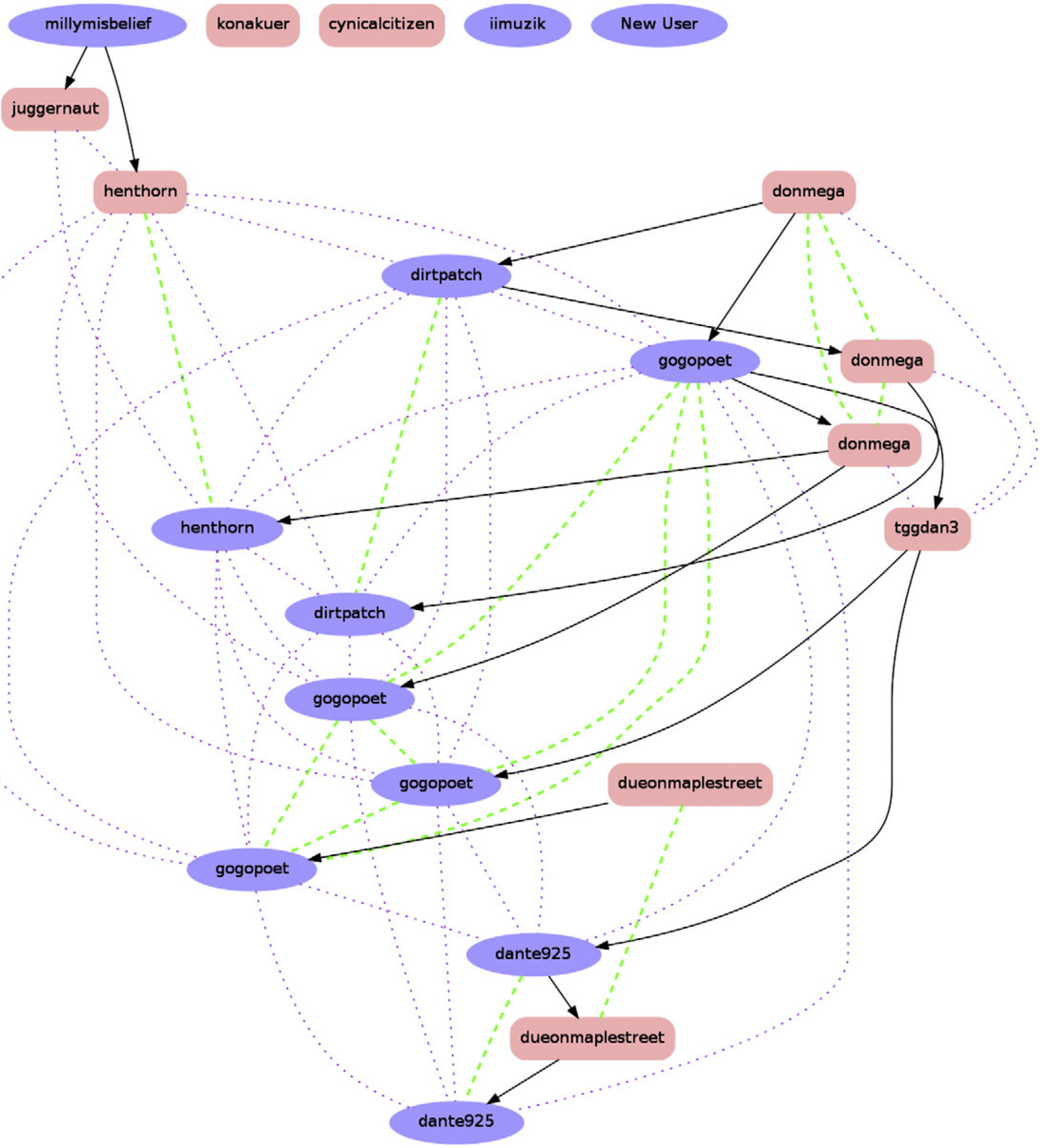


Fig. 9. Post graph showing rebuttal (disagreement) links in black solid line, same-author (agreement) links in green large dashed line, and inferred agreement links in purple dotted line. These links are all used by the *Assoc* function for MinCut. Each post also has links not shown, the *Ind* links used by the MinCut algorithm that bias a post towards being classified on either the Pro or the Con side of the debate. Pro posts are shown in blue ellipses and Con posts are shown in red rectangles.

Table 6
Feature sets, descriptions, and examples.

Set	Description/examples
Post info	Number of characters, number of words, number of sentences, WPS, average word length
Unigrams	Word frequencies
Bigrams	Word pair frequencies
Cue words	Initial unigram, bigram, and trigram
Repeated punctuation	Collapsed into one of the following: ??, !!, ?!
LIWC	LIWC measures and frequencies
Dependencies	Dependencies derived from the Stanford Parser.
POS generalized dependencies	Dependency features generalized with respect to POS of the head word.
Opinion generalized dependencies	Dependency features generalized using polarity from MPQA. Each word is generalized independently.
LIWC dependencies	Generalized dependencies with each of the lexical items replaced with its LIWC set yielding potentially many permutations for a single dependency.
Context features	Corresponding features from the parent post.

4.1.8. Opinion generalized dependencies

Somasundaran and Wiebe [31] introduced features that identify the target of opinion words. Inspired by this approach, we used their MPQA dictionary of opinion words to select the subset of dependency features in which those opinion words appear. For these features we replace the opinion words with their positive or negative polarity equivalents [20]. Similar generalized features have also been used by Pitler et al. [28]. Note that one dependency could spawn more than one generalized dependency if both words map to opinion polarities.

4.1.9. LIWC generalized dependencies

In addition to backing off to part of speech and sentiment class, we hypothesized that backing off to word topics could potentially capture generalizations about particular types of predications that the dependencies alone may not. For example, in the topic *The existence of god*, posters may convey the same stance via a variety of assertions that have no clear lexically-encoded sentiment term: *God exists*, *God is alive*, *God is real*. We attempted to capture this generalization by creating a type of generalized dependency triples where words in the triples were replaced by their LIWC classes. Note that one dependency could spawn many generalized dependencies if either or both words map to several LIWC categories.

4.1.10. Context features

Given the difficulty annotators had in reliably siding rebuttals as well as their prevalence in the corpus, we hypothesize that features representing the parent post could be helpful for classification. Here, we use a naive representation of context, where for all the feature types in Table 6, we construct both parent features and post features. For top-level parentless posts, the parent features were null.

After extracting all features, for each topic we eliminated features that were not present in more than 2% of posts (bounded by a minimum of 2 posts). Previous work suggests that eliminating infrequent features will improve accuracy [28].

4.2. Results

The primary aim of our experiments was to determine the potential contribution to debate side classification performance of contextual dialogue features, such as linguistic reflexes indicating a poster's orientation to a previous post or information from a parent post. In order to examine the utility of different feature sets, we fit a generalized linear model to the data shown in Tables 7 and 8, using the SPSS GLM package for univariate ANOVA, with topic, feature set and context as independent variables and accuracy as the dependent variable.

The optimal model showed a main effect for context ($p<8.06e-05$). When examining results by topic for a main effect we used Abortion as a reference topic. We found significant improvements over the Abortion topic for the 2nd Amendment ($p<.0182$) and Cats vs. Dogs ($p<.008$) topics.

Surprisingly, there are no main effects for other types of features. In general, our results indicate that if the data are aggregated over all topics, that indeed it is very difficult to beat the unigram baseline. In fact, dependency features do significantly worse than unigrams for a subset of topics: dep for Firefox vs. IE (due to the low 43.75% no context accuracy) and GdepO for Superman vs. Batman (38.71% and 49.19% for no context and context, respectively); dep, GdepL, and GdepP for Climate Change, and all four dependency features for 2nd Amendment. It is also interesting to note that in general the unigram accuracies without context (54.09% in Table 7) are significantly below what Somasundaran and Wiebe achieve (who report overall unigram accuracy of 62.5%). This suggests a difference between the debate posts in their corpus and the Convinceme data we used which may be related to the proportion of rebuttals.

Table 7
Accuracies achieved for no context classifiers using different feature sets and 10-fold cross validation as compared to the human topline from MTurk. Best accuracies are shown in bold for each topic in each row. Key: human topline results (Turk), unigram features (Uni), linguistics inquiry word count features (LIWC), dependency features (dep), generalized dependency features containing POS tags (GdepP), MPQA terms (GdepO), LIWC classes (GdepL) and all features combined (All). Naive Bayes was used, with features appearing less than 2% of the time pruned. The weighted average reflecting the differential number of posts per topic is in the last row.

	Turk	Uni	LIWC	dep	GdepP	GdepO	GdepL	All
Cats vs. dogs	94	66.15	56.92	63.08	**67.69**	61.54	70.00	65.38
Firefox vs. IE	74	52.50	**63.75**	43.75	52.50	**63.75**	55.00	56.25
Mac vs. PC	76	47.50	45.83	54.17	**60.00**	57.50	**60.00**	47.50
Superman vs. Batman	89	56.45	42.74	53.23	55.65	38.71	52.42	**58.06**
2nd Amendment	69	60.26	53.85	56.41	47.44	55.13	53.85	**65.38**
Abortion	75	51.95	52.14	**59.14**	55.64	50.19	55.06	53.11
Climate change	66	**58.33**	56.77	42.19	50.00	50.52	49.48	**58.33**
Comm. vs. capitalism	68	48.73	53.16	46.20	**56.33**	51.27	50.00	50.00
Death penalty	79	49.64	**54.32**	49.28	52.16	47.84	52.16	49.64
Evolution	72	54.11	48.36	**57.73**	**57.73**	56.58	55.76	55.43
Existence of God	73	52.14	51.42	54.13	**55.56**	53.42	53.28	54.13
Gay marriage	88	61.39	56.67	61.11	59.72	**62.22**	61.39	60.83
Healthcare	86	46.81	48.94	47.87	**58.51**	54.26	55.32	45.74
MJ legalization	81	53.70	**58.33**	45.37	51.85	49.07	54.63	55.56
Weighted avg.	77	54.09	52.30	54.68	**56.18**	53.87	55.43	55.12

Table 8
Accuracies achieved for context classifiers using different feature sets and 10-fold cross validation as compared to the human topline from MTurk. Best accuracies are shown in bold for each topic in each row. Key: human topline results (Turk), unigram features (Uni), linguistics inquiry word count features (LIWC), dependency features (dep), generalized dependency features containing POS tags (GdepP), MPQA terms (GdepO), LIWC classes (GdepL) and all features combined (All). Naive Bayes was used, with features appearing less than 2% of the time pruned. The weighted average reflecting the differential number of posts per topic is in the last row.

	Turk	Uni	LIWC	dep	GdepP	GdepO	GdepL	All
Cats vs. dogs	94	69.23	65.38	**75.38**	70.77	69.23	72.31	70.00
Firefox vs. IE	74	57.50	**66.25**	60.00	57.50	60.00	57.50	58.75
Mac vs. PC	76	55.00	55.83	52.50	55.83	55.83	**60.00**	56.67
Superman vs. Batman	89	60.48	59.68	**61.29**	58.87	49.19	55.65	**61.29**
2nd Amendment	69	**69.23**	58.97	52.56	60.26	56.41	55.13	**69.23**
Abortion	75	58.17	57.59	**65.95**	65.18	60.12	64.20	60.12
Climate change	66	65.10	59.90	53.65	54.69	51.56	56.77	**67.19**
Comm. vs. capitalism	68	61.39	**63.29**	59.49	67.09	56.33	51.27	61.39
Death penalty	79	58.63	**61.15**	**61.15**	**61.15**	**61.15**	55.40	57.55
Evolution	72	59.87	54.77	58.06	**64.14**	58.72	61.68	61.18
Existence of God	73	58.40	54.27	**59.54**	57.69	54.70	55.13	59.26
Gay marriage	88	65.28	56.39	66.11	63.06	64.72	65.56	**65.83**
Healthcare	86	63.83	61.70	62.77	60.64	**69.15**	65.96	64.89
MJ legalization	81	60.19	55.56	64.81	**66.67**	62.96	62.96	55.56
Weighted avg.	77	60.57	57.43	61.25	**61.70**	58.81	60.14	61.39

The main effect for context is perhaps unsurprising when comparing the results in Table 7 (no context) to those in Table 8 (context). Out of the reported accuracies for the 108 feature-topic pairs in each table, only 5 feature-topic pairs show a decrease from no context to context.

Our analysis shows that the biggest effect of context is on the rebuttals—across all feature sets, 81.8% of the posts that the context features shift to correct classification are rebuttals.

The overall lack of impact for either the POS generalized dependency features (GDepP) or the Opinion generalized dependency features (GDepO) is surprising given that they improve accuracy for other similar tasks [17,32]. While our method of extracting the GDepP features is identical to Joshi and Penstein-Rosé [17], our method for extracting GDepO is an approximation of the method of Somasundaran and Wiebe [32], that does not rely on using a development set to learn patterns indicating the topics of arguing.

We hypothesized that rebuttals would be harder to side than non-rebuttals. Table 9 shows classification accuracy averaged across topic and feature set for both classifiers sensitive to context and those that were not. Interestingly, the accuracy for non-rebuttals was not appreciably greater than rebuttals for classifiers without context (55.06% vs. 54.88%). As context features assisted rebuttals overwhelmingly, classifiers sensitive to context do much better on rebuttals than non-rebuttals (63.51% vs. 55.82%). Table 10 shows the accuracy for rebuttals conditioned on the type and accuracy of the parent post. Two important points emerge. First, accuracy is lower when the parent is a rebuttal than when it is a non-rebuttal, indicating that posts further down a dialogic chain are harder to side. Second, somewhat surprisingly, posts whose parents the system incorrectly classifies

Table 9
The average accuracy of non-rebuttals and rebuttals for classifiers provided with context features as opposed to those that were not.

Post type	Average context accuracy	Average without context accuracy
Non-rebuttal	55.82	55.06
Rebuttal	63.51	54.88

Table 10
The average accuracy on a rebuttal given its parent type (rebuttal vs. non-rebuttal) and accuracy on the parent (incorrect vs. correct).

Parent type	Parent correctness	Average context accuracy	Average without context accuracy
Non-rebuttal	Incorrect	74.9	64.85
Non-rebuttal	Correct	68.84	50.04
Rebuttal	Incorrect	66.62	59.72
Rebuttal	Correct	52.3	47.05

are easier to side than those it correctly classifies; this effect is persistent across classifiers with and without context, suggesting that there is something potentially easier about such rebuttals. A manual inspection of 10% of the rebuttals did not produce any obvious patterns, and we leave further investigation of this issue to future work.

4.2.1. Rebuttal classification results

We then conducted an experiment to see if we could automatically identify rebuttals. This was partly to examine the features that distinguish rebuttals from other posts, and partly to see whether the linguistic reflexes of rebuttals are similar to those for disagreement as reported in Ref. [1]. The rule-based JRip classifier on a 10-fold cross-validation achieved 63% accuracy. Fig. 10 illustrates a sample model learned for distinguishing rebuttals from non-rebuttals across all topics for the full data set. The figure shows that, although we used the full complement of lexical and syntactic features detailed in Section 1, the learned rules were almost entirely based on LIWC and unigram lexical features, such as 2nd person pronouns (7/8 rules), quotation marks (4/8 rules), question marks (3/8), and negation (4/8), all of which correlated with rebuttals. Other features that are used at several places in the tree are LIWC social processes, LIWC references to people, and LIWC inclusive and exclusive. One tree node reflects the particular concern with bodily functions that characterizes the Cats vs. Dogs debate as illustrated in Fig. 2.

5. Discussion

This paper reports on a number of experiments on stance classification in online debates. We first carry out an experiment to establish how difficult stance classification is for human annotators, showing that on average humans only achieve 77% accuracy. These are the first results that we are aware of that establish a human topline for debate side classification.

Our results for automatic stance classification are mixed. Our overall accuracies are respectable when compared with previous work [5,31,32,34]. We show that using sentiment, subjectivity, dependency and dialogic features, we can achieve debate-side classification accuracies, on a per topic basis, that range from 60% to 75%, as compared to unigram baselines that vary between 55% and 69%. We show that even a naive representation of context uniformly improves results across all topics. However, when we combine our data across all topics, we are not able to show that LIWC, and various kinds of dependency features that have been useful in prior work can beat the unigram baseline.

We believe that there are critical features that could capture better what is going on in these debates that we have not yet implemented. In descriptive error analysis, we examined 386 incorrectly classified posts from the abortion, death penalty, evolution, existence of god, and gay marriage topics. Almost one-fifth of the incorrectly sided posts required richer context to side: 6% of posts were determined to be statements of pure disagreement, and an additional 12% of posts were found to be contentful posts that were nonetheless hard to side without additional context. An additional 12% of posts contained quotations, which indicate a textual source of contextual information. Currently, our feature sets do not exploit this context, and, in fact, assume that information in quotes reflects the poster's own stance. A final 33% were qualitatively labeled as “hard,” due to factors similar to those discussed above; correctly siding such posts will require more robust inference.

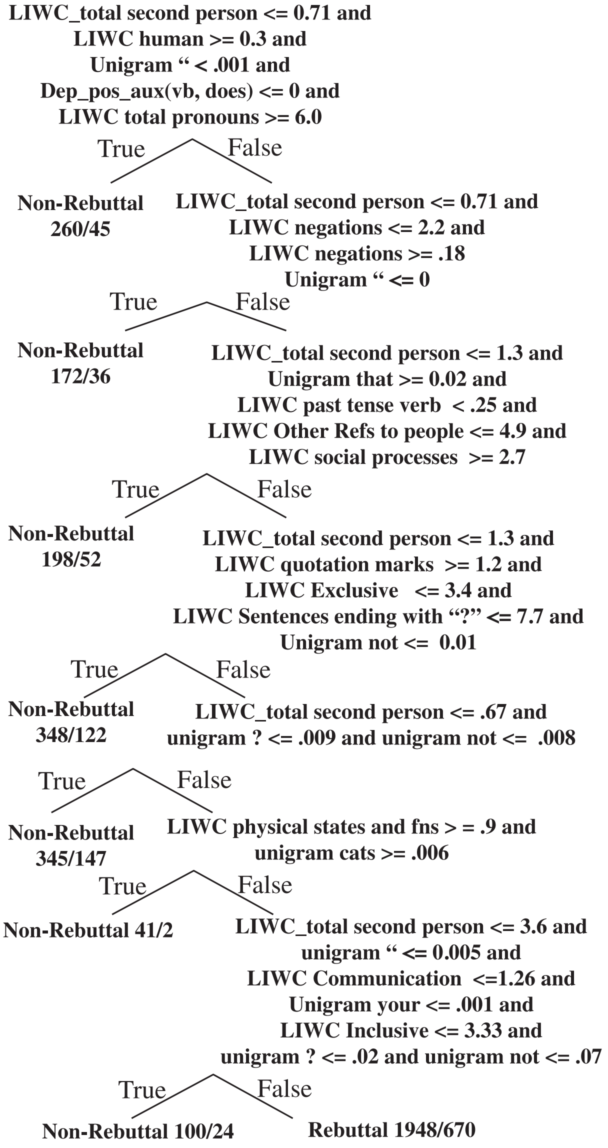

LIWC_total second person <= 0.71 and
LIWC human >= 0.3 and
Unigram “ < .001 and
Dep_pos_aux(vb, does) <= 0 and
LIWC total pronouns >= 6.0
True
False
Non-Rebuttal 260/45
LIWC_total second person <= 0.71 and
LIWC negations <= 2.2 and
LIWC negations >= .18
Unigram “ <= 0
True
False
Non-Rebuttal 172/36
LIWC_total second person <= 1.3 and
Unigram that >= 0.02 and
LIWC past tense verb < .25 and
LIWC Other Refs to people <= 4.9 and
LIWC social processes >= 2.7
True
False
Non-Rebuttal 198/52
LIWC_total second person <= 1.3 and
LIWC quotation marks >= 1.2 and
LIWC Exclusive <= 3.4 and
LIWC Sentences ending with “?” <= 7.7 and
Unigram not <= 0.01
True
False
Non-Rebuttal 348/122
LIWC_total second person <= .67 and
unigram ? <= .009 and unigram not <= .008
True
False
Non-Rebuttal 345/147
LIWC physical states and fns > = .9 and
unigram cats >= .006
True
False
Non-Rebuttal 41/2
LIWC_total second person <= 3.6 and
unigram “ <= 0.005 and
LIWC Communication <=1.26 and
Unigram your <= .001 and
LIWC Inclusive <= 3.33 and
unigram ? <= .02 and unigram not <= .07
True
False
Non-Rebuttal 100/24
Rebuttal 1948/670

Fig. 10. Model for distinguishing rebuttals vs. nonrebuttals across all topics.

In future work, we hope to improve our results with more intelligent features for representing context, discourse and rhetorical structure, and dialogic structure. We also plan to make our corpus available to other researchers in the hopes that it will stimulate further work analyzing the dialogic structure of such debates.

Acknowledgments

This work was funded by Naval Postgraduate School Grant NPS-BAA-03 to UCSC, and by the Intelligence Advanced Research Projects Activity (IARPA) through the Army Research Laboratory. We would like to thank Jason Aumiller, Robeson Bowmani, Michael Minor and Constantine Perepelitsa for programming assistance. We would also like to thank the anonymous reviewers for useful feedback, and Alexandra Balahur for organizing the WASSA workshop where this work was first presented.

References

[1] R. Abbott, M. Walker, P. Anand, J. Tree, R. Bowmani, J. King, How can you say such things?!?: recognizing disagreement in informal political argument, ACL HLT 2011, 2011, p. 2.

[2] R. Agrawal, S. Rajagopalan, R. Srikant, Y. Xu, Mining newsgroups using networks arising from social behavior, Proceedings of the 12th international conference on World Wide Web, ACM, 2003, pp. 529–535.

[3] X. Bai, Predicting consumer sentiments from online text, Decision Support Systems 50 (4) (2011) 732–742.

[4] A. Balahur, Z. Kozareva, A. Montoyo, Determining the polarity and source of opinions expressed in political debates, Computational Linguistics and Intelligent Text Processing, 2009, pp. 468–480.

[5] M. Bansal, C. Cardie, L. Lee, The power of negative thinking: exploiting label disagreement in the min-cut classification framework, Proceedings of COLING: Companion Volume: Posters, 2008, pp. 13–16.

[6] D. Biber, Variation Across Speech and Writing, Cambridge Univ Pr, 1991.

[7] S. Blair-Goldensohn, K. McKeown, O. Rambow, Building and refining rhetorical-semantic relation models, Proceedings of NAACL HLT, 2007, pp. 428–435.

[8] Y. Choi, C. Cardie, Learning with compositional semantics as structural inference for subsentential sentiment analysis, Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2008, pp. 793–801.

[9] D. Crystal, Language and the Internet, Cambridge Univ Pr, 2001.

[10] M. De Marneffe, B. MacCartney, C. Manning, Generating typed dependency parses from phrase structure parses, LREC 2006, Citeseer, 2006.

[11] J. Fox Tree, Discourse markers across speakers and settings, Language and Linguistics Compass 4 (2010) 269–281.

[12] J. Fox Tree, J. Schrock, Discourse markers in spontaneous speech: oh what a difference an oh makes, Journal of Memory and Language 40 (1999) 280–295.

[13] J. Fox Tree, J. Schrock, Basic meanings of you know and I mean, Journal of Pragmatics 34 (2002) 727–747.

[14] S. Greene, P. Resnik, More than words: syntactic packaging and implicit sentiment, Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2009, pp. 503–511.

[15] M. Groen, J. Noyes, F. Verstraten, The effect of substituting discourse markers on their role in dialogue, Discourse Processes: A Multidisciplinary Journal 47 (2010) 33.

[16] K. Jones, Automatic summarizing: factors and directions, Advances in Automatic Text Summarization, 1999, pp. 1–12.

[17] M. Joshi, C. Penstein-Rosé, Generalizing dependency features for opinion mining, Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, Association for Computational Linguistics, 2009, pp. 313–316.

[18] C. Kiss, M. Bichler, Identification of influencers—measuring influence in customer networks, Decision Support Systems 46 (2008) 233–253.

[19] D. Klein, C. Manning, Accurate unlexicalized parsing, Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, vol. 1, Association for Computational Linguistics, 2003, pp. 423–430.

[20] W. Lin, T. Wilson, J. Wiebe, A. Hauptmann, Which side are you on?: identifying perspectives at the document and sentence levels, Proceedings of the Tenth Conference on Computational Natural Language Learning, Association for Computational Linguistics, 2006, pp. 109–116.

[21] R. Malouf, T. Mullen, Taking sides: user classification for informal online political discourse, Internet Research 18 (2008) 177–190.

[22] D. Marcu, Perlocutions: the Achilles' heel of speech act theory, Journal of pragmatics 32 (2000) 1719–1741.

[23] G. Marwell, D. Schmitt, Dimensions of compliance-gaining behavior: an empirical analysis, Sociometry (1967) 350–364.

[24] G. Mishne, N. Glance, Leave a reply: an analysis of weblog comments, Third Annual Workshop on the Weblogging Ecosystem, Citeseer, 2006.

[25] A. Murakami, R. Raymond, Support or oppose?: classifying positions in online debates from reply activities and opinion expressions, Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, 2010, pp. 869–875.

[26] B. Pang, L. Lee, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval 2 (2008) 1–135.

[27] J. Pennebaker, M. Francis, R. Booth, Linguistic Inquiry and Word Count: Liwc2001, 2001.

[28] E. Pitler, A. Louis, A. Nenkova, Automatic sense prediction for implicit discourse relations in text, Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, vol. 2, Association for Computational Linguistics, 2009, pp. 683–691.

[29] R. Prasad, A. Joshi, B. Webber, Realization of discourse relations by other means: alternative lexicalizations, Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, 2010, pp. 1023–1031.

[30] S. Purpura, C. Cardie, J. Simons, Active learning for e-rulemaking: public comment categorization, Proceedings of the 2008 International Conference on Digital Government Research, Digital Government Society of North America, 2008, pp. 234–243.

[31] S. Somasundaran, J. Wiebe, Recognizing stances in online debates, Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, vol. 1, Association for Computational Linguistics, 2009, pp. 226–234.

[32] S. Somasundaran, J. Wiebe, Recognizing stances in ideological on-line debates, Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, Association for Computational Linguistics, 2010, pp. 116–124.

[33] S. Somasundaran, J. Wiebe, J. Ruppenhofer, Discourse level opinion interpretation, Proceedings of the 22nd International Conference on Computational Linguistics, vol. 1, Association for Computational Linguistics, 2008, pp. 801–808.

[34] M. Thomas, B. Pang, L. Lee, Get out the vote: determining support or opposition from congressional floor-debate transcripts, Proceedings of the 2006 conference on empirical methods in natural language processing, Association for Computational Linguistics, 2006, pp. 327–335.

[35] M. Walker, Rejection by implicature, Proceedings of the 20th Meeting of the Berkeley Linguistics Society, Citeseer, 1994.

[36] M. Walker, Inferring acceptance and rejection in dialog by default rules of inference, Language and Speech 39 (1996) 265.

[37] Y. Wang, C. Rosé, Making conversational structure explicit: identification of initiation-response pairs within online discussions, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2010, pp. 673–676.

[38] J. Yearwood, A. Stranieri, The generic/actual argument model of practical reasoning, Decision Support Systems 41 (2006) 358–379.

[39] A. Yessenalina, Y. Yue, C. Cardie, Multi-level structured models for document-level sentiment classification, Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2010, pp. 1046–1056.

Dr. Marilyn Walker is Professor of Computer Science at University of California Santa Cruz, founder and director of the Natural Language and Dialogue Systems Lab. Dr. Walker received her B.A. degree in Computer and Information Science from the University of California Santa Cruz and her M.S. and Ph.D. degrees in Computer Science from Stanford University and the University of Pennsylvania. Dr. Walker has worked on many aspects of dialogue interaction, both in algorithms for dialogue management and language generation for dialogue systems, as well as computational analysis of human-human dialogue. She conducted the first experiments using reinforcement learning to adapt a dialogue system to human users. Her work on personalization of dialogue systems involves algorithms for user tailoring, individual adaptation of linguistic style using boosting, and generation using theories of politeness and personality. She has authored over a hundred technical papers and is holder of more than 10 patents. She has edited a book on Centering Theory and special issues of journals on empirical methods in discourse and dialogue and on spoken language generation for dialogue systems. Her H-index, a measure of research excellence, is 42.

Pranav Anand is an Assistant Professor in the Linguistics Department at University of California Santa Cruz. He received a B.A. in Mathematics from Harvard College and a Ph.D. in Linguistics from the Massachusetts Institute of Technology, concentrating in formal semantics. His work concentrates on the effects of discourse and grammatical context on interpretation.

Rob Abbott is a graduate student in Computer Science at the University of California, Santa Cruz where he studies natural language processing under the supervision of Dr. Marilyn Walker. He received a BS in Mathematics and Computer Science from California State University, Chico.

Jean E. Fox Tree is a Professor of Psychology at the University of California Santa Cruz. She received her AB from Harvard University in linguistics, her MSc from the University of Edinburgh in cognitive science, and her PhD from Stanford University in psychology. She studies the comprehension and production of spontaneous speech.

Craig Martell is an Associate Professor of Computer Science at the Naval Postgraduate School. He received a Ph.D. in Computer and Information Sciences from the University of Pennsylvania. He also holds an MA in Philosophy from the New School for Social Research, an MA in Political Science from Pennsylvania State University, and a BA in Politics from Catholic University of America. His research interests include Machine Learning, Natural Language Processing, and Gesture Analysis.

Joseph King received a B.A. in Linguistics from UCSC in 2011. He has done research on the discourse structure of multi-party chat, the recognition of persuasion in blogs and forums, the online processing of English reflexives, and the syntactic and semantic representations of tag questions.