A Corpus for Research on Deliberation and Debate

Marilyn A. Walker, Pranav Anand, Jean E. Fox Tree, Rob Abbott, Joseph King

University of California Santa Cruz
Computer Science Department, Linguistics Department
maw@soe.ucsc.edu, panand@ucsc.edu, abbott@soe.ucsc.edu

Abstract

Deliberative, argumentative discourse is an important component of opinion formation, belief revision, and knowledge discovery; it is a cornerstone of modern civil society. Argumentation is productively studied in branches ranging from theoretical artificial intelligence to political rhetoric, but empirical analysis has suffered from a lack of freely available, unscripted argumentative dialogs. This paper presents the Internet Argument Corpus (IAC), a set of 390,704 posts in 11,800 discussions extracted from the online debate site 4forums.com. A 2866 thread/130,206 post extract of the corpus has been manually sided for topic of discussion, and subsets of this topic-labeled extract have been annotated for several dialogic and argumentative markers: degrees of agreement with a previous post, cordiality, audience-direction, combativeness, assertiveness, emotionality of argumentation, and sarcasm. As an application of this resource, the paper closes with a discussion of the relationship between discourse marker pragmatics, agreement, emotionality, and sarcasm in the IAC corpus.

Keywords: online, dialogue, debate

1. Introduction

A critical function of public discourse and debate is the expression and formation of opinions related to current social and political issues. Public discourse and debate has always occurred in both informal and formal settings – from conversations in social or workplace settings – to classrooms, courtrooms, or formal organized debates. See Fig. 1. What is different now is that: (1) these debates, whether formal or informal, are typically captured and made part of a permanent record, (2) the context of debate is often a virtual social community (members of a forum, or viewers of a televised debate), rather than a community sharing the same space and time in the real world, and (3) the scale of participation and exposure to other views in the virtual community can be much larger than in the past.

Figure 1: Formal and informal debate over time.

Our primary aim is to use these permanent records of deliberation and debate to deepen our theoretical and practical understanding of deliberation, how people argue, how they decide what they believe on issues of relevance to their lives and their country, how linguistic structures in debate dialogues reflect these processes, and how debate and deliberation affect people's choices and their actions in the public sphere. These conversations range from current political topics such as national health care to religious questions such as the meaning of biblical passages. We are also interested in understanding how differences in affordances of different online forums shape how opinions are expressed. One affordance type that has a major effect are CONTEXTUAL AFFORDANCES, such as the ability to quote another person's post or to break a previous post up into dialogic bites, and then responding to each part of the post. See Figure 2.

Topic	Q-R: Post
Evolution	**Q:** How can you say such things? The Bible says that God CREATED over and OVER and OVER again! And you reject that and say that everything came about by evolution? If you reject the literal account of the Creation in Genesis, you are saying that God is a liar! If you cannot trust God's Word from the first verse, how can you know that the rest of it can be trusted? **R:** It's not a literal account unless you interpret it that way.
Gay marriage	**Q:** Gavin Newsom- I expected more from him when I supported him in the 2003 election. He showed himself as a family-man/Catholic, but he ended up being the exact oppisate, supporting abortion, and giving homosexuals marriage licenses. I love San Francisco, but I hate the people. Sometimes, the people make me want to move to Sacramento or DC to fix things up. **R:** And what is wrong with giving homosexuals the right to settle down with the person they love? What is it to you if a few limp-wrists get married in San Francisco? Homosexuals are people, too, who take out their garbage, pay their taxes, go to work, take care of their dogs, and what they do in their bedroom is none of your business.
Abortion	**Q:** Equality is not defined by you or me. It is defined by the Creator who created men. **R:** Actually I think it is defined by the creator who created all women. But in reality your opinion is gibberish. Equality is, like every other word, defined by the people who use the language. Currently it means "the same". People aren't equal because they are not all the same. Any attempt to argue otherwise is a display of gross stupidity.

Figure 2: Sample Quote/Response Pairs

This paper details the Internet Argument Corpus (IAC), a collection of 390,704 posts in 11,800 discussions extracted from the online debate site 4forums.com. The IAC should

support a much deeper understanding of argumentative dialogue. Below, we describe the IAC in more detail and some of the initial results that can be derived from it. We are making it available at *http://nlds.soe.ucsc.edu/software*.

2. Corpus Description

2.1. Overview

The IAC was scraped from 4forums.com, a website for political debate and discourse. The site is a fairly typical internet forum where people post some discussion topic, other people post responses and a conversation ensues. The entire corpus consists of $390,704$ posts in $11,800$ discussions (aka threads) by $3,317$ authors. A subset of these discussions fall into our list of topics.

The forum is a shallow tree of sub-forums with each discussion posted under a specific topic (e.g. $\gg$ Topics $>$ Economic Debates $>$ Tax Debates). These sub-forums cover a broad range of topics relevant to the US political landscape.

Each discussion has a tree structure enabling people to respond to posts out of order and engage in side conversations. The forum software facilitates this by providing an option to view the discussion in "threaded mode".

Some discussions contain polls and some polls contain a list of users and how they vote. This information could be used for stance classification.

One important feature in this forum (and others like it) is a mechanism for quoting another post. A poster may decide to link to and replicate a previous post in whole or in part. This establishes very precise context - something very useful for forum participants and NLP applications alike. Quotes need not be derived from other posts, the same mechanism may be repurposed for external or even original content. Someone may use "quotes" to quote a news article, cite a study, include an excerpt from Wikipedia, religious texts, or the US constitution, or they may even use it to satirize a previous post or opposing viewpoint. Because the site's layout encourages quoting, 72.3% of all posts contain at least one quote. We will refer to quotes and the immediately following response text up until the next quote or the end of the post as a *quote-response pair* or Q-R pair. A post may have more than one Q-R pair. Although rare, quotes sometimes nest within one another. On the site, quotes sometimes lack information referencing their originating post; because of this and the usefulness in knowing these associations, we go to great lengths to find the original post. Interestingly, this means that our corpus has higher quality information about the reply structure than the original site!

The site has a number of additional affordances which we capture. For discussions we capture the title, a reference URL, breadcrumbs indicating the sub-forum it belongs to, and poll information if present. For posts we capture text, author, timestamp (with minute resolution), reply structure, links, formating (i.e. bold, italic, color, etc.), quotes, and post title if present. We don't capture attachments or tags & discussion ratings which are rarely used.

The forum provides a set of stylized and sometimes animated emoticons which participants may choose to use. These range from the standard smiley face :) to the somewhat less standard clown face. We include these in the text with their own markers.

2.2. Annotations

Scraping a website and organizing it into a database for processing does not require extensive effort. The value of the IAC is in the annotations. We selected a set of contentious issues and hand-labeled discussions for topic from this set (note: more recently acquired discussions may be missing this annotation). See Table 1 above for topics and details. We used Amazon's Mechanical Turk to gather annotations on a large set of Q-R pairs and post triples for various dialogic interactions. The final annotation we present here used Mechanical Turk to label participant stance relative to several of the aforementioned topics.

For our dialogic interaction surveys we used two sets of data. One set involved 10,003 Quote-Response (Q-R) pairs as defined above; the other involved 6,797 chains of three posts defined as the series P1, P2, and P3 such that P3 is a response to P2 which is itself a response to P1 (henceforth we will refer to this task & data as P123). For the P123 data we stripped quotes from each post.

These annotation tasks were motivated by our interest in how specific discourse markers were used in this medium so we biased the selection of post triples and Q-R pairs in favor of certain keywords. For Q-R pairs we selected 5000 examples which started with one of the selected terms, 2003 which had a term starting in the first 10 tokens, and 3000 which did not have any term starting in the first 10 tokens. The 5000 term initial examples were selected according to a distribution defined by hand based on data availability and interest. For the P123 set we required that all posts either start with one of the terms or do not contain any term in the first 10 tokens. Thus resampling is necessary if one desires a natural distribution for some analysis or NLP task. The excluded data is perfectly usable as part of a development set and/or training set. The distribution we used means that it is possible to derive a natural distribution from the Q-R set of around 3000 examples in size, but this is not possible for the P123 set. 1717 of 6797 P123 triples have no terms starting in the first 10 tokens for any of the 3 posts

For both sets we restricted ourselves to the topic list described above. Due to problems scraping, which have since been overcome, we only used posts up to a depth of 5 in the discussion tree, meaning chains of length at most six from root to leaf. Note that this does not preclude posts from very far into a large discussion if they have a short path to the discussion's root post. The text presented to the Turkers was also missing emoticons represented as images. We chose not to enforce that quotes come from a prior post, and thus some quotes from external sources are included.

For the Q-R and P123 annotation tasks we constructed a number of HITs (Human Intelligence Tasks), each consisting of seven pairs of *context* and *response* text. We asked the Turkers (Mechanical Turk Workers) to judge the *response* given the *context* according to several measures

Topic	Discs	Posts	NumA	P/A	A>1P	PL	Samp	Agree	Sarcasm	Emote	Attack	Nasty
Evolution	871	39199	744	53	80%	430	1224	11%	11%	19%	19%	13%
Abortion	564	35721	755	47	73%	338	733	13%	9%	30%	15%	10%
Gun Control	824	27122	514	53	70%	323	589	12%	13%	23%	18%	14%
Gay Marriage	305	15678	435	36	72%	362	256	14%	14%	25%	16%	9%
Existence of God	105	5914	321	18	72%	347	211	12%	14%	35%	22%	16%
Healthcare	81	1810	150	12	67%	293	33	15%	15%	27%	15%	18%
Death Penalty	25	1485	185	8	68%	350	22	5%	0%	5%	0%	0%
Climate Change	40	1470	170	9	58%	375	42	12%	14%	17%	14%	10%
Communism vs Capitalism	38	1154	148	8	60%	333	16	31%	12%	12%	6%	6%
Marijuana Legalization	13	653	140	5	60%	298	32	12%	16%	25%	16%	16%
None	8934	260498	2824	92	60%	330	0					
All	11800	390704	3317	118	63%	341	3158	12%	12%	24%	17%	12%

Table 1: Characteristics of Different Topics. **KEY**: Number of discussions and posts on the topic (**Discs, Posts**). Number of authors (**NumA**). Posts per author (**P/A**). Authors with more than one post (**A > 1P**). Median post Length in Characters (**PL**) after stripping quotes. Percentage of resampled Q-R pairs (**Samp**) that agree (**Agree**), use sarcasm (**Sarcasm**), are emotional (**Emote**), attack the previous poster (**Attack**), and are nasty (**Nasty**). The scalar values are threshholded at ± 1 inclusive, for sarcasm we required that at least half of all annotators marked it sarcastic.

such as Agreement/Disagreement as shown in Figure 2. In any given HIT and for each Turker, we took steps to avoid showing multiple *context-response* pairs derived from the same discussion or posts. The Q-R and P123 data were not mixed. For the Q-R pair data set, the *context* portion consisted of the quote and the *response* portion was the Q-R pair's response text. For the P123 set we presented each of the following three pairs as (*context*, *response*): (P1, P2), (P2, P3), and (P1, P3). Note that P3 is not a response to P1 making the (P1, P3) pair a little different. We used the (P1, P3) pair to gauge how specific language influenced judgments or whether the content of the *context* and *response* mattered. If (P1, P2) is labeled as a disagreement and (P2, P3) is as well, does it follow that (P1, P3) is labeled agreement? In what cases is this relation intransitive and when is it transitive? Did the Turkers notice something fishy and signal their uncertainty by marking "unsure" in addition to their assessment?

Most of our measures were scalar; we chose to do this because previous work suggests that taking the means of scalar annotations can reduce noise in Mechanical Turk annotations (Snow et al., 2008). Turkers were not given additional definitions of the meaning of any measures, e.g. we let Turkers to use their native intuitions about what it means for a post to be sarcastic, since previous work suggests that non-specialists tend to collapse all forms of verbal irony under the term sarcastic (Bryant and Fox Tree, 2002; Gibbs, 2000). The scalar judgments were on an 11 point scale [-5,5] implemented with a slider. The annotators were also able to signal uncertainty with a CAN'T TELL option. Each of the HITs and their associated *context-response* pairs were annotated by 5-7 Turkers.

As indicated by Figure 2, our questions were split into two separate surveys, Survey 1 and Survey 2. Survey 1 used both sets of data (Q-R and P123) while Survey 2 only used the Q-R set. Furthermore, for the questions in Survey 2 we were only interested in instances where the *response* was a disagreement so we first asked the agree/disagree binary question and only showed the remaining questions if the Turker marked it as a disagreement. It should be noted that Survey 2 was conducted before we had results from Survey 1 and thus the results from Survey 1 were unavailable for use as a filter. For the binary agree/disagree question, the question of what to do if the annotator was uncertain was raised on the Mechanical Turk forums; we decided to instruct them to mark it as disagreement so we could have more annotations and better coverage.

For the Q-R and P123 tasks we required that Turkers use a US IP address and have high approval ratings. We paid between \$0.20-\$0.36 USD per HIT. Due to the complexity of our HITs and setup, we hosted the HITs on our own server. We believe that this had a side effect of making it more difficult for bots to automatically submit answers. We have not attempted to remove unreliable annotators. By using the mean rating of all annotators, we get fairly reliable annotations for this task, but filtering out unscrupulous annotators may lead to even higher quality annotations. The raw data is included in the IAC.

In addition to annotating these features of particular posts, we leveraged Mechanical Turk to annotate the stances of 6144 posters within 380 threads across our 10 topics ($\mu_{posters} = 16.2$). There were 1054 unique poster, topic

Type	α	Survey Question
		Survey 1
S	0.62	**Agree/Disagree**: Does the respondent agree or disagree with the prior post?
S	0.32	**Fact/Emotion**: Is the respondent attempting to make a fact based argument or appealing to feelings and emotions?
S	0.42	**Attack/Insult**: Is the respondent being supportive/respectful or are they attacking/insulting in their writing?
B	0.22	**Sarcasm**: Is the respondent using sarcasm?
S	0.46	**Nice/Nasty**: Is the respondent attempting to be nice or is their attitude fairly nasty?
		Survey 2
B		**Agree/Disagree**: Does the respondent agree or disagree with the prior post? (No Unsure option)
S		**Audience**: Is the respondent 's arguments intended more to be interacting directly with the original poster OR with a wider audience?
S		**Undercutting**: Is the argument of the respondent targeted at the entirety of the original poster 's argument OR is the argument of the respondent targeted at a more specific idea within the post?
S		**Negotiate/attack**: Does the respondent seem to have an argument of their own OR is the respondent simply attacking the original poster 's argument?
S		**Question/Assert**: Is the respondent questioning the original poster OR is the respondent asserting their own ideas?

Table 2: Mechanical Turk Annotations (Binary = B and Scalar = S) and level of agreement as Krippendorff's α.

Topic	Average κ
Abortion	0.54
Climate Change	0.50
Communism vs Capitalism	0.22
Death Penalty	0.48
Evolution	0.41
Existence of God	0.42
Gay Marriage	0.57
Gun Control	0.46
Healthcare	0.50
Marijuana Legalization	0.60

Figure 3: Interannotator agreement for poster siding task. Many scores are pulled down by particularly difficult threads.

elements. For each of these threads, annotators were provided with a visualization of the thread's conversational reply tree and were asked to side each poster as either *pro* or *con* the issue in question. Because many posters on forums have either unclear or complicated positions on these issues, we also allowed annotators to select an *other* category. In earlier rounds of annotations, we found that Turkers often over-selected the *other* category, and directions were iteratively altered to advise workers to select *other* only as a last resort, and additionally required that such answers be accompanied by a short free-text justification. These strategies reduced selection of *other* to 11% of annotations, of which approximately 8% were justified as unclear instances (the justifications included in the IAC). Turkers were provided with one hour to complete this task, and averaged 8.4 minutes (or, roughly 2 poster annotations per minute). They were paid $\$0.035$ per poster, or on average $\$0.567$.

Initial experiments revealed that Turkers found this task difficult and highly subjective, and average κ across threads was 0.25 We thus devised a two-level training/filtering scheme. In the first stage, 212 debates with between 7 and 38 posters ($\mu = 17.0$) were hand-selected from our five most prolific topics (gun control, gay marriage, abortion, evolution, existence of God). For each of these posts, we hand-annotated gold standard siding for one *pro* and *con* poster and used this to filter Turkers. Of the 293 Turkers who attempted this stage, 124 (42%) were eliminated. The remaining 169 were used to annotate the 212 threads (with 5 annotators per thread). In the second round of annotation, we selected those Turkers who correctly annotated gold standard data on 4 or more threads from the original 212. These annotators were invited to annotate the remaining 178 threads, which we did not examine or determine gold standard labels for. Of these threads, 97 comprised all of these topic-annotated for the least prolific 5 topics (climate change, communisms vs. capitalism, dealth penalty, healthcare, marijuana legalization) and the remaining 81 were distributed roughly uniformly across the most prolific topics. Figure 3 provides the Cohen's κ values for each topic.

3. Analysis

Previous computational work has drawn from two distinct threads of research. One thread approaches the problem from the perspective of social structures and social network models (Agrawal et al., 2003; Murakami and Raymond, 2010). The other thread draws on computational natural language text processing techniques, e.g. building on research in text classification and topic modeling and applying it to both face-to-face and online debate and deliberation (Lin et al., 2006; Thomas et al., 2006; Somasundaran and Wiebe, 2009; Somasundaran and Wiebe, 2010; Bansal et al., 2008; Greene and Resnik, 2009; Anand et al., 2011; Walker et al., ; Forsyth and Martell, 2007; Awadallah et al., 2010). Our theoretical and empirical orientation draws on both of these lines of previous work, as well as incorporating our previous work on human dialogue in both mediated and non-mediated settings. As described above, we have collected annotations on Mechanical Turk for agreement, stance side, and emotive and affective elements such as sarcasm.

Figure 4 provides examples from the end points and means of the annotations for three of the questions, Respect/Insult, Sarcasm, and Fact/Emotion. Nice/Nasty and Respect/Insult are strongly correlated by worker annotations $r(54003) = 0.84$, $p <$ 2.2e-16 and both weakly correlated with Agree/Disagree ratings ($r(54003) = 0.32$ and $r(54003)=0.36$, respectively; $p <$ 2.2e-16) and Fact/Emotion ratings ($r(54003) = 0.32$ and $r(54003)=0.31$, respectively; $p <$ 2.2e-16), while Agree/Disagree and Fact/Emotion ratings show the smallest correlation, $r(54003)=0.11$, $p <$ 2.2e-16. For the linguistic marker correlations discussed below we averaged scores across annotators, a process which sharpened correlations (e.g., Respect/Insult means correlate with Agree/Disagree means more strongly ($r(5393) = 0.51$) as well as Nice/Nasty means ($r(5393) = 0.91$); Agree/Disagree is far less correlated with Fact/Emotion ($r(5393) = 0.07$). Interannotator agreement was computed using Krippendorff's α (due to the variability in number of annotators that completed each hit), assuming an ordinal scale for all measures except sarcasm; see Figure 2. The low agreement for Sarcasm accords with native intuition – it is the class with the least dependence on lexicalization and the most subject to interspeaker stylistic variation. The relatively low results for Fact/Emotion is perhaps due to the emotional charge many ideological arguments engender; informal examination of posts that showed the most disagreement in this category often showed a cutting comment or a snide remark at the end of a post, which was was ignored by some annotators and evidence for others (one Emotional post in Figure 4 is clearly an insult, but was uniformly labeled as -5 by all annotators).

Discourse Markers. Because both psychological research on discourse processes (Fox Tree and Schrock, 1999; Fox Tree and Schrock, 2002; Groen et al., 2010) and computational work on agreement (Galley et al., 2004) indicates that discourse markers are strongly associated with particular pragmatic functions, we first tested the role of turn-initial markers in predicting upcoming content (Fox Tree and Schrock, 2002; Groen et al., 2010). Based on manual

Class	Very High Degree	Neutral	Very Low Degree
Insult or Attack	Well, you have proven yoruself to be a man with no brain, that is for sure. The definition that was given was the one that scientists use, not the layperson.	The empire you defend is tyrannical. They are responsible for the death of millions.	Very well put.
Sarcasm	My pursuit of happiness is denied by trees existing. Let's burn them down and destroy the environment. It's much better than me being unhappy.	An interesting analysis of that article you keep quoting from the World Net Daily [url]	I would suggest you look at the faero island mouse then. That is a new species, and it is not man doing it, but rather nature itself.
Emotion-based Argument	Really! You can prove that most pro-lifers don't care about women?...it is idiotic thinking like this that makes me respect you less and less.	Fine by me. First, I don't consider having a marriage recognized by government to be a "right". Second, I've said many times I don't think government should be in the marriage business at all.	Sure. Here is an explanation. The 14C Method. That is from the Radiocarbon WEB info site by the Waikato Radiocarbon Dating Lab of the University of Waikato (New Zeland).

Figure 4: Sample Responses for the Insult, Sarcasm, and Fact/Feeling spectrums

inspection of a subset of the IAC, we constructed a list of discourse markers; 17 of these occurred at least 50 times in a quote response (upper bound of 700 samples): *actually*, *and*, *because*, *but*, *I believe*, *I know*, *I see*, *I think*, *just*, *no*, *oh*, *really*, *so*, *well*, *yes*, *you know*, *you mean*.

The top discourse markers highlighting disagreement were *really* (67% read a response beginning with this marker as prefacing a disagreement with a prior post), *no* (66%), *actually* (60%), *but* (58%), *so* (58%), and *you mean* (57%). At this point, the next most disagreeable category was the unmarked category, with about 50% of respondents interpreting an unmarked post as disagreeing. On the other hand, the most agreeable marker was *yes* (73% read a response beginning with this marker as prefacing an agreement) followed by *I know* (64%), *I believe* (62%), *I think* (61%), and *just* (57%). The other markers were close to the unmarked category: *and* (50%), *because* (51%), *oh* (51%), *I see* (52%), *you know* (54%), and *well* (55%).

The overall agreement on sarcasm was low, as in other computational work on recognizing sarcasm (Davidov et al., 2010). At most, only 31% of respondents agreed that the material after a discourse marker was sarcastic, with the most sarcastic markers being *you mean* (31%), *oh* (29%), *really* (24%), *so* (22%), and *I see* (21%). Only 15% of respondents rated the unmarked category as sarcastic (e.g., fewer than 1 out of 6 respondents). The cues *I think* (10%), *I believe* (9%), and *actually* (10%) were the least sarcastic markers.

Taken together, these ratings suggest that the cues *really*, *you mean*, and *so* can be used to indicate both disagreement and sarcasm. However, *but*, *no*, and *actually* can be used for disagreement, but not sarcasm. And *I know* (14% sarcastic, similar to None), *I believe*, and *I think* can be used for non-sarcastic agreement.

From informal analyses, we hypothesized that *really* and *oh* might indicate sarcasm. While we found evidence supporting this for *really*, it was not the case for *oh*. Instead, *oh* was used to indicate emotion; it was the discourse marker with the highest ratings of feeling over fact.

Despite the fact that it would seem that disagreement would be positively correlated with sarcasm, disagreement and sarcasm were not related. There were two tests possible. One tested the percentage of people who identified an item as disagreeing against the percentage of people who identified it as sarcasm, $r(16)$ = -.27, p = .27 (tested on 17 discourse markers plus the None category). The other tested the degree of disagreement (from -5 to +5) against the percentage of people who identified the post as sarcastic, $r(16)$ = -.33, p = .18.

However, we did observe relationships between sarcasm and other variables. Two results support the argument that sarcasm is emotional and personal. The more sarcastic, the nastier (rather than nicer), $r(16)$ = .87, $p < .001$. In addition, the more sarcastic, the more emotional (over factual) respondents were judged to be, $r(16)$ = .62, p = .006, and the more the respondents were judged to be questioning the poster rather than asserting their own ideas, $r(16)$ = .76, $p < .001$. Finally, the more the respondents' argument was judged to be directed at the single poster rather than a broader audience, the more sarcastic, $r(16)$ = .74, $p < .001$. Taken together, these analyses suggest that sarcasm is emotional and personal, but not necessarily a sign of disagreement.

4. Conclusion

This paper presents a new corpus with useful annotations for exploring issues pertaining to online debate, a medium not well represented by existing corpora. We are actively using, expanding, and improving this corpus as we explore research which this dataset enables. The dataset has broad applicability and we look forward to seeing what others do with it.

The Internet Argument Corpus is available at http://nlds.soe.ucsc.edu/software

5. Acknowledgements

This work was funded by the Intelligence Advanced Research Projects Activity (IARPA) through the Army Research Laboratory, and by the Naval Postgraduate School Grant NPS-BAA-03 to UCSC. We'd like to thank Jason Aumiller, Robeson Bowmani, Ricky Grant, Michael Minor, and Constantine Perepelitsa for programming assistance. We'd also like to thank Craig Martell and the anonymous reviewers for their feedback and effort.

6. References

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

Pranav Anand, Marilyn Walker, Rob Abbott, Jean E. Fox Tree, Robeson Bowmani, and Michael Minor. 2011. Cats Rule and Dogs Drool: Classifying Stance in Online Debate. In *Proceedings of the ACL Workshop on Sentiment and Subjectivity*.

R. Awadallah, M. Ramanath, and G. Weikum. 2010. Language-model-based pro/con classification of political text. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 747–748. ACM.

M. Bansal, C. Cardie, and L. Lee. 2008. The power of negative thinking: Exploiting label disagreement in the min-cut classification framework. *Proceedings of COLING: Companion volume: Posters*, pages 13–16.

G.A. Bryant and J.E. Fox Tree. 2002. Recognizing verbal irony in spontaneous speech. *Metaphor and symbol*, 17(2):99–119.

D. Davidov, O. Tsur, and A. Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 107–116. Association for Computational Linguistics.

E.N. Forsyth and C.H. Martell. 2007. Lexical and discourse analysis of online chat dialog. *IEEE Computer Society*.

J.E. Fox Tree and J.C. Schrock. 1999. Discourse Markers in Spontaneous Speech: Oh What a Difference an Oh Makes. *Journal of Memory and Language*, 40(2):280–295.

J.E. Fox Tree and J.C. Schrock. 2002. Basic meanings of you know and I mean. *Journal of Pragmatics*, 34(6):727–747.

M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg. 2004. Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 669–es. Association for Computational Linguistics.

R.W. Gibbs. 2000. Irony in talk among friends. *Metaphor and Symbol*, 15(1):5–27.

S. Greene and P. Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 503–511. Association for Computational Linguistics.

M. Groen, J. Noyes, and F. Verstraten. 2010. The Effect of Substituting Discourse Markers on Their Role in Dialogue. *Discourse Processes: A Multidisciplinary Journal*, 47(5):33.

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

A. Murakami and R. Raymond. 2010. Support or Oppose? Classifying Positions in Online Debates from Reply Activities and Opinion Expressions. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 869–875. Association for Computational Linguistics.

R. Snow, B. O'Connor, D. Jurafsky, and A.Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263. Association for Computational Linguistics.

S. Somasundaran and J. Wiebe. 2009. Recognizing stances in online debates. In *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: Volume 1-Volume 1*, pages 226–234. Association for Computational Linguistics.

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

M. Thomas, B. Pang, and L. Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 327–335. Association for Computational Linguistics.

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