Cats Rule and Dogs Drool!: Classifying Stance in Online Debate

Pranav Anand, Marilyn Walker, Rob Abbott, Jean E. Fox Tree,
Robeson Bowmani, and Michael Minor
University of California Santa Cruz

Abstract

A growing body of work has highlighted the challenges of identifying the stance a speaker holds towards a particular topic, a task that involves identifying a holistic subjective disposition. We examine stance classification on a corpus of 4873 posts across 14 topics on ConvinceMe.net, 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 identifing rebuttals with 63% accuracy, and for identifying stance on a per topic basis that range from 54% to 69%, as compared to unigram baselines that vary between 49% and 60%. Our results suggest that methods that take into account the dialogic context of such posts might be fruitful.

1 Introduction

Recent work has highlighted the challenges of identifying the STANCE that a speaker holds towards a particular political, social or technical topic. Classifying stance involves identifying a holistic subjective disposition, beyond the word or sentence (Lin et al., 2006; Malouf and Mullen, 2008; Greene and Resnik, 2009; Somasundaran and Wiebe, 2009; Somasundaran and Wiebe, 2010). Our work is inspired by the large variety of such conversations now freely available online, and our observation that the contextual affordances of different debate and discussion websites vary a great deal. One important contextual variable, discussed at length below, is the percentage of posts that are **rebuttals** to previous posts, which varies in our data from 34% to 80%. The ability to explicitly rebut a previous post gives these debates both monologic and dialogic properties (Biber, 1991; Crystal, 2001; Fox Tree, 2010); Compare Figure 1 to Figure 2. We believe that discussions containing many rebuttal links require a different type of analysis than other types of debates or discussions.

Dialogic Capital Punishment
Studies have shown that using the death penalty saves 4 to 13 lives per execution. That alone makes killing murderers worthwhile.
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.
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.. .
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.

Figure 1: Capital Punishment discussions with posts linked via rebuttal links.

This paper utilizes 1113 two-sided debates (4873 posts) from Convinceme.net for 14 different debate topics. See Table 1. 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-labelling their post for stance. These debates may be heated and emotional, discussing weighty issues such as euthanasia and capital punishment, such as the example in Figure 1. But they also appear to be a form of entertainment via playful

Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, ACL-HLT 2011, pages 1–9,
24 June, 2011, Portland, Oregon, USA ©2011 Association for Computational Linguistics

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 Figure 3.

Monologic Capital Punishment
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
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.
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.

Figure 2: Posts on the topic Capital punishment without 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.

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 (Sparck-Jones, 1999); (2) gaining a deeper understanding of what makes an argument persuasive (Marwell and Schmitt, 1967); and (3) identifying the linguistic reflexes of perlocutionary acts such as persuasion and disagreement (Walker, 1996; Greene and Resnik, 2009; Somasundaran and Wiebe, 2010; Marcu, 2000). As a first step, in this paper we aim to automatically identify rebuttals, and identify the speaker's stance towards a particular topic.

Dialogic Cats vs. 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!
Dog owners seem infatuated with handling sh*t. Cat owners don't seem to share this infatuation.
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!

Figure 3: Cats vs. Dogs discussions with posts linked by rebuttal links.

The most similar work to our own is that of Somasundaran & Wiebe (2009, 2010) who also focus on automatically determining the stance of a debate participant with respect to a particular issue. Their data does not provide explicit indicators of dialogue structure such as are provided by the rebuttal links in Convinceme. Thus, this work treats each post as a monologic text to be classified in terms of stance, for a particular topic. They show that discourse relations such as concessions and the identification of argumentation triggers improves performance over sentiment features alone (Somasundaran and Wiebe, 2009; Somasundaran and Wiebe, 2010). This work, along with others, indicates that for such tasks it is difficult to beat a unigram baseline (Pang and Lee, 2008).

Other similar related work analyzes Usenet forum quote/response structures (Wang and Rosé, 2010). We believe quote/response pairs have a similar discourse structure to the rebuttal post pairs in Convinceme, but perhaps with the linguistic reflexes of stance expressed even more locally. However agreement vs. disagreement is not labelled across quote/response pairs and Wang & Rose (2010) do not attempt to distinguish these different discourse relations. Rather they show that they can use a variant of LSA to identify a parent post, given a response post, with approximately 70% accuracy. A recent paper by (Abbott et al., 2011) examines agreement and disagreement in quote/response pairs in idealogical and nonidealogical online forum discussions, and shows that you can distinguish the agreement relation with 68% accuracy. Their results indicate that contextual features do improve performance for identifying the agreement relation between quotes and responses.

Other work has utilized the social network structure of online forums, either with or without textual features of particular posts (Malouf and Mullen, 2008; Mishne and Glance, 2006; Murakami and Raymond, 2010; Agrawal et al., 2003). However this work does not examine the way that the dialogic structure varies by topic, as we do, and the threading structure of their debates does not distinguish between agreement and disagreement responses. (Mishne and Glance, 2006) show that most replies to blog posts are disagreements, while Agarwal's work assumed that adjacent posts always disagree, and did not use any of the information in the text. Murakami & Raymond (2010) show that simple rules for identifying disagreement, defined on the textual content of the post, can improve over Agarwal's results and (Malouf and Mullen, 2008) show that a combination of textual and social network features provides the best performance. We leave the incorporation of social network information for stance classification to future work.

Section 3 discusses our corpus in more detail, and presents the results of a human debate-side classification task conducted on Mechanical Turk. Section 3 describes two different machine learning experiments: one for identifying rebuttals and the other for automatically determining stance. Section 4 presents our results. We show that we can identify rebuttals with 63% accuracy, and that using sentiment, subjectivity and dialogic features, we can achieve debate-side classification accuracies, on a per topic basis, that range from 54% to 69%, as compared to unigram baselines that vary between 49% and 60%.

2 Corpus Description and Analysis

Table 1 provides an overview of our corpus. Our corpus consists of 1113 two-sided debates (4873 posts) from Convinceme.net for 12 topics ranging from playful debates such as Cats vs. Dogs to more heated political topics such as Capital Punishment. In Table 1, the topics above the line are either technical or playful, while the topics below the line are ideological. In total the corpus consists of 2,722,340 words; the topic labeled debates which we use in our experiments contain 507,827 words.

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 topic, and thus can be considered as a response to that post; (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.

Topics vary a great deal in terms of their dialogic structure and linguistic expression. In Table 1, the columns providing counts for different variables are selected to illustrate ways in which topics differ in the form and style of the argument and in its subjective content. One important variable is the percentage of the topic posts that are linked into a rebuttal dialogic structure (**Rebuttals**). Some of these differences can be observed by comparing the dialogic and monologic posts for the Capital Punishment topic in Figures 1 and 2 to those for the Cats vs. Dogs topic in Figures 3 and 4. 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 and more posts per author (**P/A**) in the topics below the line in Table 1. 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* as well as propositional anaphora such as *that* and *it*), ellipsis, and dialogic cue words; Figure 5 shows the difference in counts of 'you' between rebuttals and non-rebuttals (Rebuttals $\bar{x} = 9.6$ and Non-Rebuttals $\bar{x} = 8.5$, $t(27) = 24.94$, $p < .001$). Another indication of author investment is the percentage of authors with more than one post (**A > 1P**). Post Length (**PL**), on the other hand, is not significantly correlated with degree of investment in the topic.

Monologic Cats vs. Dogs
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!
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?

Figure 4: Posts on the topic Cats vs. Dogs without explicit rebuttal links.

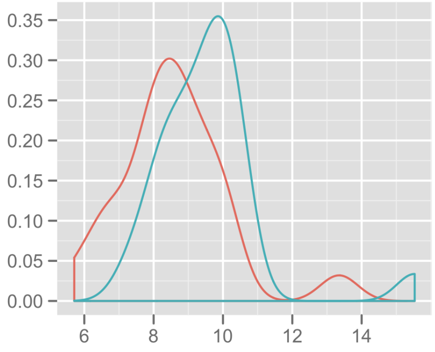

Figure 5: Kernel density estimates for 'you' counts across rebuttals (green) and non-rebuttals (red).

Other factors we examined were words per sen-

	Post and Threading Variables					Normalized LIWC Variables				
Topic	**Posts**	**Rebuttals**	**P/A**	**A $>$ 1p**	**PL**	**Pro**	**WPS**	**6LTR**	**PosE**	**NegE**
Cats v. Dogs	148	40%	1.68	26%	242	3.30	-1.95	-2. 43	1.70	.30
Firefox vs. IE	218	40%	1.28	16%	167	-0.11	-0.84	0.53	1.23	-0.81
Mac vs. PC	126	47%	1.85	24%	347	0.52	0.28	-0.85	-0.11	-1.05
Superman/Batman	140	34%	1.41	21%	302	-0.57	-1.78	-0.43	1.21	.99
2nd Amendment	134	59%	2.09	45%	385	-1.38	1.74	0.58	-1.04	0.38
Abortion	594	70%	2.82	43%	339	0.63	-0.27	-0.41	-0.95	0.68
Climate Change	202	69%	2.97	40%	353	-0.74	1.23	0.57	-1.25	-0.63
Communism vs. Capitalism	212	70%	3.03	47%	348	-0.76	-0.15	1.09	0.39	-0.55
Death Penalty	324	62%	2.44	45%	389	-0.15	-0.40	0.49	-1.13	2.90
Evolution	798	76%	3.91	55%	430	-0.80	-1.03	1.34	-0.57	-0.94
Exist God	844	77%	4.24	52%	336	0.43	-0.10	0.34	-0.24	-0.32
Gay Marriage	505	65%	2.12	29%	401	-0.13	.86	.85	-0.42	-0.01
Healthcare	110	80%	3.24	56%	280	0.28	1.54	.99	0.14	-0.42
Marijuana Legalization	214	52%	1.55	26%	423	0.14	0.37	0.53	-0.86	0.50

Table 1: Characteristics of Different Topics. Topics below the line are considered "ideological". Normalized LIWC variable z-scores are significant when more than 1.94 standard deviations away from the mean (two-tailed).
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**). Post Length in Characters (**PL**). **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.

tence (**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**) (Pennebaker et al., 2001). For example, Table 1 shows that discussions about Cats vs. Dogs consist of short simple words in short sentences with relatively high usage of positive emotion words and pronouns, whereas 2nd amendment debates use relatively longer sentences, and death penalty debates (unsurprisingly) use a lot of negative emotion words.

Human Topline. The best performance for siding ideological debates in previous work is approximately 64% accuracy over all topics, for a collection of 2nd Amendment, Abortion, Evolution, and Gay Rights debate posts (Somasundaran and Wiebe, 2010). Their best performance is 70% for the 2nd amendment topic. The website that these posts were collected from apparently did not support dialogic threading, and thus there are no explicitly linked rebuttals in this data set. 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 from a debate *without context*. To our knowledge, none of the previous work on debate side classification has attempted to establish 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 and debates with fewer than 6 posts on either side. Each of our 12 topics consists of more than one debate: each debate was mapped by hand to the topic and topic-siding (as in (Somasundaran and Wiebe, 2010)). 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. Each HIT was judged by 9 annotators using Mechanical Turk with each annotator restricted to at most 30 HITS (300 judgments). Since many topics were US specific and we wanted annotators with a good grasp of English, we required Turkers to have a US IP address.

Figure 6 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. Figure 6 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 0.2656.

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 Cli-

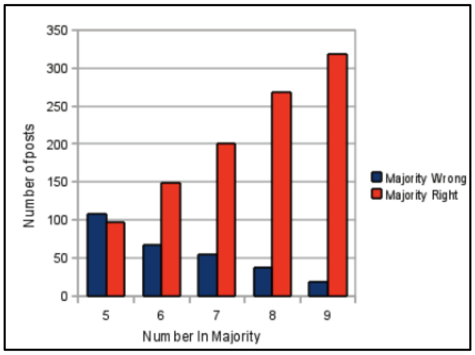

Figure 6: Accuracies of Human Mechanical Turk judges at selecting the True Siding of a post without context.

mate 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 2.

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

Table 2: Human Agreement on Rebuttal Classification

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. Our analysis suggests that in 28% of these cases, 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) may have contributed to this error. An additional 39% were 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. Given the differences in siding difficulty depending on rebuttal status, in Section 4 we present results for both rebuttal and stance classification.

3 Features and Learning Methods

Our experiments were conducted with the Weka toolkit. All results are from 10 fold cross-validation on a balanced test set. In the hand examination of annotators siding performance, 101 posts were determined to have incorrect self-labeling for side. We eliminated these posts and their descendants from the experiments detailed below. This resulted in a dataset of 4772 posts. We used two classifiers with different properties: NaiveBayes and JRip. JRip is a rule based classifier which produces a compact model suitable for human consumption and quick application. Table 3 provides a summary of the features we extract for each post. We describe and motivate these feature sets below.

Set	Description/Examples
Post Info	IsRebuttal, Poster
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.
Generalized Dependencies	Dependency features generalized with respect to POS of the head word and opinion polarity of both words.
Opinion Dependencies	Subset of Generalized Dependencies with opinion words from MPQA.
Context Features	Matching Features used for the post from the parent post.

Table 3: Feature Sets, Descriptions, and Examples

Counts, Unigrams, Bigrams. Previous work suggests that the unigram baseline can be difficult to beat for certain types of debates (Somasundaran and Wiebe, 2010). Thus we derived both unigrams and bigrams as features. We also include basic counts such as post length.

Cue Words. We represent each posts **initial** unigram, bigram and trigram sequences to capture the useage 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 (Fox Tree and Schrock, 1999; Fox Tree and Schrock, 2002; Groen et al., 2010).

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.

LIWC. We also derived features using the Linguistics Inquiry Word Count tool (LIWC-2001) (Pennebaker et al., 2001). 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 1.

Syntactic Dependency. Previous research in this area suggests the utility of dependency structure to determine the TARGET of an opinion word (Joshi and Penstein-Rosé, 2009; Somasundaran and Wiebe, 2009; Somasundaran and Wiebe, 2010). 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 (De Marneffe et al., 2006; Klein and Manning, 2003).

Generalized Dependency. To create generalized dependencies, we "back off" the head word in each of the above features to its part-of-speech tag (Joshi and Penstein-Rosé, 2009). Joshi & Rose'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.

Opinion Dependencies. Somasundaran & Wiebe (2009) introduced features that identify the TARGET of opinion words. Inspired by this approach, we used the MPQA dictionary of opinion words to select the subset of dependency and generalized dependency features in which those opinion words appear. For these features we replace the opinion words with their positive or negative polarity equivalents (Lin et al., 2006).

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 3, we construct both **parent** features and **post** features. For top-level parentless posts, the **parent** features were null.

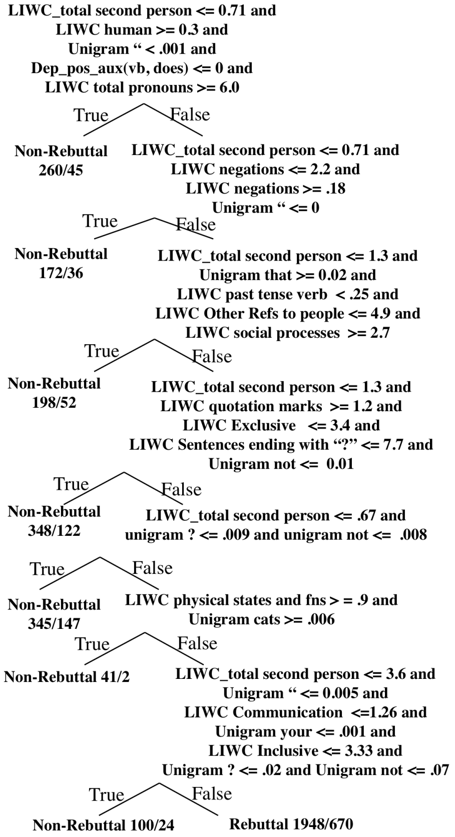
LIWC_total second person <= 0.71 and
LIWC human >= 0.3 and
Unigram " < .001 and
Dep_pos_aux(vb, does) <= 0 and
LIWC total pronouns >= 6.0
True
False
Non-Rebuttal 260/45
LIWC_total second person <= 0.71 and
LIWC negations <= 2.2 and
LIWC negations >= .18
Unigram " <= 0
True
False
Non-Rebuttal 172/36
LIWC_total 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

Figure 7: Model for distinguishing rebuttals vs. nonrebuttals across all topics.

4 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. Because we believed that identification of whether a post is a rebuttal or not might be helpful in the long term for debate-side classification, we also establish a baseline for rebuttal classification.

4.1 Rebuttal Classification Results

The differences in human performance for siding depended on rebuttal status. Our experiments on rebuttal classification using the rule-based JRip classifer on a 10-fold cross-validation of our dataset produced 63% accuracy. Figure 7 illustrates a sample model learned for distinguishing rebuttals from non-rebuttals across all topics. The Figure shows that, although we used the full complement of lexical and syntactic features detailed above, 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 Figure 3.

4.2 Automatic Debate-Side Classification Results

We first compared accuracies using Naive Bayes to JRip for all topics for all feature sets. A paired t-test showed that Naive Bayes over all topics and feature sets was consistently better than JRip ($p < .0001$). Thus the rest of our analysis and the results in Table 4 focus on the Naive Bayes results.

Table 4 presents results for automatic debate side classification using different feature sets and the Naive Bayes learner which performs best over all topics. In addition to classifying using only post-internal features, we ran a parallel set of experiments adding contextual features representing the parent post, as described in Section 3. The results in Table 4 are divided under the headers *Without Context* and *With Context* depending on whether features from the parent post were used if it existed (e.g. in the case of rebuttals).

We conducted paired t-tests over all topics simultaneously to examine the utility of different feature sets. We compared unigrams to LIWC, opinion generalized dependencies, POS generalized dependencies, and all features. We also compared experiments using context features to experiments using no contextual 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. Across all topics there are generally no significant differences between experiments conducted with unigrams and other features. The mean accuracies across all topics for unigrams vs. LIWC features was 54.35% for unigrams vs. 52.83% for LIWC. The mean accuracies for unigram vs POS generalized dependencies was 54.35% vs. 52.64%, and for unigrams vs. all features was Unigram 54.35% vs 54.62%. The opinion generalized dependencies features actually performed significantly worse than unigrams with an accuracy of 49% vs. 54.35% ($p < .0001$).

It is interesting to note that in general the unigram accuracies are significantly below what Somasundaran and Wiebe achieve (who report overall unigram 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.

The overal lack of impact for either the POS generalized dependency features (**GDepP**) or the Opinion generalized dependency features (**GDep0**) is surprising given that they improve accuracy for other similar tasks (Joshi and Penstein-Rosé, 2009; Somasundaran and Wiebe, 2010). While our method of extracting the **GDepP** features is identical to (Joshi and Penstein-Rosé, 2009), our method for extracting **GDepO** is an approximation of the method of (Somasundaran and Wiebe, 2010), that does not rely on selecting particular patterns indicating the topics of arguing by using a development set.

The LIWC feature set, which is based on a lexical hierarchy that includes social features, negative and positive emotion, and psychological processes, is the only feature set that appears to have the potential to systematically show improvement over a good range of topics. We believe that further analysis is needed; we do not want to handpick topics for which particular feature sets perform well.

Our results also showed that context did not seem to help uniformly over all topics. The mean performance over all topics for contextual features using the combination of all features and the Naive Bayes learner was 53.0% for context and 54.62% for no context ($p = .15\%$, not significant). Interesting, the use of contextual features provided surprisingly greater performance for particular topics. For example for 2nd Amendment, unigrams with context yield a performance of 69.23% as opposed to the best performing without context features using LIWC of 64.10%. The best performance of (Somasundaran and Wiebe, 2010) is also 70% for the 2nd amendment topic. For the Healthcare topic, LIWC with context features corresponds to an accuracy of 60.64% as opposed to GDepP without context performance of 54.26%. For Communism vs. Capitism, LIWC with context features gives an accuracy of 56.55% as opposed to accuracies actually

	Without Context						With Context				
	Turk	**Uni**	**LIWC**	**GdepO**	**GdepP**	**All**	**Uni**	**LIWC**	**GdepO**	**GdepP**	**All**
Cats v. Dogs	94	59.23	55.38	56.15	61.54	**62.31**	50.77	56.15	55.38	60.77	50.00
Firefox vs. IE	74	51.25	**53.75**	43.75	48.75	50.00	51.25	**53.75**	52.50	52.50	51.25
Mac vs. PC	76	53.33	**56.67**	55.00	50.83	**56.67**	53.33	55.83	**56.67**	49.17	54.17
Superman Batman	89	54.84	45.97	42.74	45.97	54.03	50.00	**57.26**	43.55	50.81	53.23
2nd Amendment	69	56.41	64.10	51.28	58.97	57.69	**69.23**	61.54	44.87	52.56	67.95
Abortion	75	50.97	51.56	50.58	52.14	51.17	51.36	**53.70**	51.75	**53.70**	50.78
Climate Change	66	53.65	**58.33**	38.02	46.35	50.52	48.96	56.25	38.02	38.54	48.96
Comm vs. Capitalism	68	48.81	47.02	46.43	47.02	48.81	45.83	**56.55**	47.02	51.19	48.81
Death Penalty	79	51.80	53.96	46.76	49.28	52.52	51.80	56.12	56.12	**57.55**	53.24
Evolution	72	**57.24**	48.36	54.93	56.41	**57.24**	54.11	46.22	50.82	52.14	52.96
Existence of God	73	52.71	51.14	49.72	52.42	51.99	52.28	52.28	50.14	**53.42**	51.42
Gay Marriage	88	**60.28**	56.11	56.11	58.61	59.44	56.94	52.22	54.44	53.61	54.72
Healthcare	86	52.13	51.06	51.06	54.26	52.13	45.74	**60.64**	59.57	57.45	53.19
MJ Legalization	81	57.55	46.23	43.40	53.77	**59.43**	52.83	46.23	49.06	49.06	50.94

Table 4: Accuracies achieved 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**). Generalized dependency features containing MPQA terms (**GdepO**) & POS tags (**GdepP**). NaiveBayes was used, no attribute selection was applied.

below the majority class baseline for all of the features without context.

Should we conclude anything from the fact that 6 of the topics are idealogical, out of the 7 topics where contextual features provide the best performance? We believe that the significantly greater percentage of rebuttals for these topics should give a greater weight to contextual features, so it would be useful to examine stance classification performance on the subset of the posts that are rebuttals. We believe that context is important; our conclusion is that our current contextual features are naive – they are not capturing the relationship between a post and a parent post. Sequential models or at least better contextual features are needed.

The fact that we should be able to do much better is indicated clearly by the human topline, shown in the column labelled **Turk** in Table 4. Even without context, and with the difficulties siding rebuttals, the human annotators achieve accuracies ranging from 66% to 94%.

5 Discussion

This paper examines two problems in online-debates: rebuttal classification and debate-side or stance classification. Our results show that we can identify rebuttals with 63% accuracy, and that using lexical and contextual features such as those from LIWC, we can achieve debate-side classification accuracies on a per topic basis that range from 54% to 69%, as compared to a unigram baselines that vary between 49% and 60%. These are the first results that we are aware of that establish a human topline for debate side classification. These are also the first results that we know of for identifying rebuttals in such debates.

Our results for stance classification are mixed. While we show that for many topics we can beat a unigram baseline given more intelligent features, we do not beat the unigram baseline when we combine our data across all topics. In addition, we are not able to show across all topics that our contextual features make a difference, though clearly use of context should make a difference in understanding these debates, and for particular topics, classification results using context are far better than the best feature set without any contextual features. In future work, we hope to develop more intelligent features for representing context and improve on these results. 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 Grant NPS-BAA-03 to UCSC and Intelligence Advanced Research Projects Activity (IARPA) through the Army Research Laboratory to UCSC by subcontract from the University of Maryland. We'd like to thank Craig Martell and Joseph King for helpful discussions over the course of this project, and the anonymous reviewers for useful feedback. We would also like to thank Jason Aumiller for his contributions to the database.

References

Rob Abbott, Marilyn Walker, Jean E. Fox Tree, Pranav Anand, Robeson Bowmani, and Joseph King. 2011. How can you say such things?!?: Recognizing Disagreement in Informal Political Argument. In *Proceedings of the ACL Workshop on Language and Social Media*.

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.

D. Biber. 1991. *Variation across speech and writing*. Cambridge Univ Pr.

David Crystal. 2001. *Language and the Internet*. Cambridge University Press.

M.C. De Marneffe, B. MacCartney, and C.D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454. Citeseer.

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.

J. E. Fox Tree. 2010. Discourse markers across speakers and settings. *Language and Linguistics Compass*, 3(1):113.

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.

M. Joshi and C. Penstein-Rosé. 2009. Generalizing dependency features for opinion mining. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 313–316. Association for Computational Linguistics.

D. Klein and C.D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

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.

R. Malouf and T. Mullen. 2008. Taking sides: User classification for informal online political discourse. *Internet Research*, 18(2):177–190.

Daniel Marcu. 2000. Perlocutions: The Achilles' heel of Speech Act Theory. *Journal of Pragmatics*, 32(12):1719–1741.

G. Marwell and D. Schmitt. 1967. Dimensions of compliance-gaining behavior: An empirical analysis. *sociomety*, 30:350–364.

G. Mishne and N. Glance. 2006. Leave a reply: An analysis of weblog comments. In *Third annual workshop on the Weblogging ecosystem*. Citeseer.

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.

B. Pang and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

J. W. Pennebaker, L. E. Francis, and R. J. Booth, 2001. *LIWC: Linguistic Inquiry and Word Count*.

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.

Karen Sparck-Jones. 1999. Automatic summarizing; factors and directions. In Inderjeet Mani and Mark Maybury, editors, *Advances in Automatic Text Summarization*. MIT Press.

Marilyn A. Walker. 1996. Inferring acceptance and rejection in dialogue by default rules of inference. *Language and Speech*, 39-2:265–304.

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