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].

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

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.