RECOGNIZING STANCES IN MANDARIN SOCIAL IDEOLOGICAL DEBATES WITH TEXT AND ACOUSTIC FEATURES

Linchuan Li^{1,2}, *Zhiyong Wu*^{1,2,3}, *Mingxing Xu*^{1,2}, *Helen Meng*^{1,3}, *Lianhong Cai*^{1,2}

1 Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Shenzhen Key Laboratory of Information Science and Technology,
Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
2 Tsinghua National Laboratory for Information Science and Technology (TNList),
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
3 Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China
lilinchuan318@gmail.com, zywu@sz.tsinghua.edu.cn, xumx@tsinghua.edu.cn,
hmmeng@se.cuhk.edu.hk, clh-dcs@tsinghua.edu.cn

ABSTRACT

Recognizing stances is of great importance to understand intention of human beings. While previous related researches mainly focused on text modality, in this paper, we aim to combine textual and acoustic features to automatically recognize stances in social debates. For acoustic features, we find that speaking rate is an indispensable feature to distinguish whether speaker is taking a stance or not. In addition, we also demonstrate that emphasis information is helpful for recognizing stances. For textual modality, we present a novel Support Topic Feature (STF) and use it to recognize which stance the speaker is taking (positive, negative or neutral). Experiments on four debate datasets confirm that the performance of STF is much better than that of n-gram features. When using STF only, F1-measure can be improved by 3% $\sim$ 8% as compared to the baseline. Combining acoustic features with STF leads to even better performance by improving F1-measure with 2% $\sim$ 8% further.

Index Terms— Stance, topic, debate, multimodal, emphasis

1. INTRODUCTION

Over recent years, research on stance recognition has received a growing amount of interests. Stances, or a speaker's subjective attitudes or opinions on the topic of discussion [1] [2], are integrated parts of activities involving collaboration, negotiation, and decision making. The goal of stance recognition in debate is to determine the side (i.e. affirmative, negative or neutral) one participant is taking.

The foundational work has aroused a heated research on the stance-taking recognition field. However, much work has relied on textual materials, such as those presented in [3] and [4]. Predominantly drawing on textual information, a text-based classifying approach concentrates on topic extraction by using Latent Dirichlet Allocation (LDA) [5] method. Some other researches have focused on linguistic statistical features, such as n-grams, linguistic score and recognition confidence measure [3]. Recently, there is much work on recognizing stances in online debate forums [6] [7] [8]. Those approaches are constrained by the textual modality of their data corpus, so they have not taken speech, which carries much emotional information, into account.

There is also a great deal of work investigating issues of subjectivity, sentiment or stance in spontaneous speech, primarily by exploiting existing conversational dyadic or multiparty meeting corpora. Even having speech data, some work like [9] still leveraged word n-gram features, which did not make full use of acoustic information. Regarding this, [10] investigated the combination of several sources of information and proved that a fusion of acoustic and textual features can yield the best performance.

After observing data, we notice that given a specified viewpoint, topics themselves have emotional tendencies towards the viewpoint. For example, in the debate "whether money is the root of evils," a participant on the side "money is the root of evils" may refer to quantities of crimes, including robberies, larcenies, drug crimes, etc. All these crimes are caused by money. Alternatively, the other side will not mention them. They prefer talking about "rape" and "domestic violence" which has nothing to do with money. Basing on the idea, in this paper we present a novel feature named Support Topic Feature (STF) to classify stances in debates. In STF, we introduce support degree of topics to capture the topics' emotional tendencies towards viewpoints. In addition, we discover two interesting phenomena. One is, with only text modality, the ambiguity of text may mislead us in understanding the real intentions of the spoken utterance. For example, the sentence spoken in an ironic may have the totally opposite meaning. The other is speaker usually tends to emphasize the key word when taking a stance. Hence, in addition to textual features, we take acoustic features and emphasis information into account for stance recognition. As for experiments, we construct a dataset of over a thousand labeled utterances extracted from the debate competition videos. The experimental results confirm the accuracy of the proposed STF, achieving up to 8% improvement compared with the challenging baseline of n-gram features. Combining STF with acoustic features and emphasis leads to even better performance by improving F1-measure with 2% $\sim$ 8% further.

2. DEBATE GENRES AND MOTIVATIONS

It is pointed out the debate genre poses significant challenges to stance analysis [11]. Participants on both sides debate issues, express their opinions and argue why their viewpoint is right and why the opposite's is wrong. In addition to expressing positive sentiments about own side, a key strategy is to express negative sentiments about the opposite side. In other words, it is of extraordinary significance to refute the adverse point of view. However many simple stance recognizing approaches just find positive and negative words in a sentence, and aggregate their counts to determine the sentence polarity, without taking the opinion targets into account. It is obvious that these methods will not work well. As is shown in [11], to recognize stances, we need to consider not only which opinions are positive and negative, but also what the opinions are about (their targets).

As we all know that besides directly expressing opinions, we more often refer to related aspects which strongly prove our opinions. In the example described in Section 1, while the affirmative indicates robberies, larcenies and smugglings that are all caused by money, the negative mainly talks about the crimes which have nothing to do with money, e.g. rape. By observing the data, we find out an important rule that **topics carry emotional tendencies towards a specific viewpoint**.

There is an interesting phenomenon that, rather than state their opinions flatly, participants usually employ rhetorical question in debates. The rhetorical question is one kind of rhetorical devices in Mandarin and it can strengthen question intonation. Note the meaning of sentence “Is it right?” in rhetorical question mostly equals to “it is not right” in Mandarin. So ignoring the intonation information may lead to the opposite result.

Complicating the picture further, there are some techniques in debates. People sometimes repeat the utterance their opponent just spoke and find the hidden error in logic to controvert opposite thesis. Besides, they will also make assumptions that the opponent's viewpoint is set up firstly, and then move forward by the logic, finally inferring to a totally impossible situation.

In some cases, participants may acknowledge the opposing side's opinions, which, however, doesn't mean they endorse the rival opinion. We call this situation concession. Uniform treatment of all opinions in an utterance would obviously cause error in such cases.

All the debate genres presented above will bring challenges to recognizing what the real stance speaker is taking. To tackle the problems, we propose STF in Section 4 and further incorporate it with acoustic derived features for stance recognition.

3. DATASETS

3.1. Data collection

The datasets are composed of four debate competitions available online. We investigate existing datasets of debates or multiparty meetings and could hardly find available ones in Mandarin. We download four debate videos of International Varsity Debate. Unlike other stances recognition researches, our debates are triple-sided. Besides the affirmative and negative, we attribute presenter and jury to a neutral side. There are 10 participants in each debate, 4 affirmative, 4 negative, 1 presenter and 1 jury. The resolutions of four debates are “Whether money is the root of evils,” “Human nature is good or evil,” “Whether is starting business more good than harm to college students,” and “Difficult to know and easy to do or difficult to do and easy to know” respectively. The duration of each debate amounts about 52 minutes.

3.2. Segmentation

We first extract audio streams from the video data. Since the alternate statement of each side, we further segment the audio into pieces, each of which is an utterance spoken by only one speaker. Meanwhile we annotate the segmented utterances which stances they belong to. We also remove the noisy segments of applause and laugh. Finally we get 1254 effective utterances in total.

3.3. Speech to text

Acoustic features can help us distinguish subjective utterances from non-subjective ones. However, we need the textual information to further recognize which stance the speaker is taking. With the benefit from automatic speech recognition (ASR), speech to text is easy to do with the IBM Speech-to-Text interface[1]. We also find the scripts of these debates on the debate website[2] so that we can proofread the text corpus to improve the accuracy of ASR. Unlike English word, Chinese words don't have space between each other, so we divide the

[1] http://www.ibm.com/smarterplanet/us/en/ibmwatson/developercloud/speech-to-text.html

[2] http://www.bianlun.net/thread-11887-1-1.html

Table 1. The detailed information of datasets. “Money” is short for the dataset “Whether money is the root of evils,” “Nature” short for “Human nature is good or evil,” “Business” short for “Whether is starting business more good than harm to college students,” and “Knowing” short for “Difficult to know and easy to do or difficult to do and easy to know”. “Aff”, “Neu”, “Neg” represents Affirmative, Neutral and Negative respectively.

Datasets	Money			Nature			Business			Knowing		
Stances	Aff	Neu	Neg	Aff	Neu	Neg	Aff	Neu	Neg	Aff	Neu	Neg
Number of utterances	100	81	110	121	89	111	124	58	126	104	109	121
Total durations(s)	770	543	774	944	678	932	1089	300	1122	966	971	1009
Avg duration per utterance(s)	7.7	6.7	7.0	7.8	7.6	8.4	8.8	5.2	8.9	9.3	8.9	8.3
Total number of words	4017	1943	4263	4988	2500	4335	5434	1288	5896	4225	3666	4894
Avg number of words per utterance	40	24	39	41	28	39	44	22	47	41	34	40

utterance into tokens using Jieba tokenizer[3]. We also obtain the Part-Of-Speech (POS) tag and sentence dependencies by employing the Stanford parser[4]. Table 1 shows the detailed information of our datasets.

3.4. Filter

As the corpus comes from spoken debates, in which ambiguous and fragmentary utterances are inevitable. We set two layers of word filter. Firstly we filter the word by POS tagging. After applying the Stanford parser we find that the notional words are tagged NN or VV. We ignore the structural words such as prepositions and conjunctions. The second filter is relied on TF-IDF. We obtain TF_{ij} by calculating TF-IDF value of $word_i$ in $document_j$. We delete $word_i$ s.t.

$$\prod_{j=1}^{4} TF_{ij} > 0 \tag{1}$$

That means $word_i$ occurring in all 4 documents are not considered.

4. SUPPORT TOPIC FEATURE

4.1. Support topic feature (STF)

Support topic feature (STF) is a sentence-level feature vector. We represent each sentence by STF with the form of $[T_1, T_2, ...T_n]$, where n is dependent on the number of topics extracted from documents. We calculate T_i by

$$T_i = (\sum_{j=1}^{\omega} p_j^d c(i,j)) \cdot S_i, 1 \leq i \leq n \tag{2}$$

where ω is the number of words in the analyzed sentence. $c(i,j)$ denotes the correlation coefficient of $word_j$ to $topic_i$ and is computed by using word2vec[5] to calculate cosine distance of the two words. p_j^d denotes the polarity of the words that describe $word_j$ within a distance of d. We call S_i support degree of $topic_i$, which indicates $topic_i$'s backing intensity to the affirmative stance. The algorithm to evaluate $c(i,j)$, p_j^d and S_i will be introduced below.

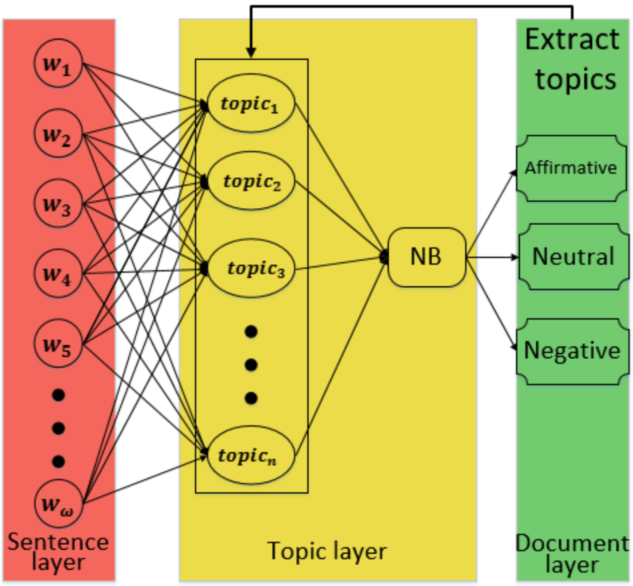

Fig. 1. The hierarchical model to construct support topic features.

4.2. Constructing support topic features

We build a hierarchical model to construct support topic features, as shown in Figure 1. The model can be divided into 3 layers: document layer, topic layer and sentence layer. At document layer, we extract topics of each stance; at topic layer, we compute the topics' support degree S_i to each stance; and at sentence layer we calculate the correlation coefficient of each word in sentence to each topic $c(i,j)$.

Document layer: At document layer, we first classify corpus into 3 sub-datasets according to the stance (affirmative, neutral, negative). Then we employ Latent Dirichlet Allocation (LDA) [5] on affirmative and negative datasets to extract

[3] https://github.com/fxsjy/jieba

[4] http://nlp.stanford.edu/software/index.shtml

[5] https://code.google.com/archive/p/word2vec/

topics respectively. By using Gibbs sampling methods[6] we finally obtain the topics and their occurring probabilities. The occurring probabilities of the topics are regarded as the scores below (i.e. $score(topic_i|+)$ or $score(topic_i|-)$).

Topic layer: At topic layer, we calculate support degree S_i of each topic by

$$S_i = \frac{score(topic_i\ |+) - score\ (topic_i\ |-)}{score(topic_i\ |+) + score\ (topic_i\ |-)} \tag{3}$$

$score(topic_i|+)$ and $score(topic_i|-)$ mean the score (i.e. occurring probability) of $topic_i$ in affirmative and negative datasets respectively. A higher $score(topic_i|+)$ and a lower $score(topic_i|-)$ denote $topic_i$ is a stronger backing to the affirmative stance. If $topic_i$ is not a topic in negative dataset, $score(topic_i|-)$ will be 0 and S_i equals to 1. Similarly, S_i=-1 means $topic_i$ is just a topic in negative corpus. In other words, S_i takes the value in $[-1, 1]$, with $S_i > 0$ meaning $topic_i$ is for affirmative stance and $S_i < 0$ for negative stance.

Algorithm 1 Calculate p_j^d

```
$p_j^d = 1$
if $j \geq d$ then
    $st = j - d$
else
    $st = 0$
end if
if $j + d < len(words[])$ then
    $en = j + d$
else
    $en = len(words[]) - 1$
end if
for $i = st$ to $en$ do
    if $i == j$ then
        continue
    else if $word[i]$ in DENIAL or ASSUMPTION or REPEAT then
        $p_j^d = -p_j^d$
    else if $word[i]$ in CONCESSION then
        $p_j^d = 0$
        break
    else if $word[i]$ in QUESTION then
        $p_j^d = -p_j^d$
        break
    end if
end for
return $p_j^d$
```

Sentence layer: At sentence layer, we calculate the correlation coefficient of each word to each topic $c(i, j)$ and p_j^d. We use word2vec to compute vector representations of words and calculate word cosine distance. As for p_j^d, we firstly establish a trigger expression dictionary based on the debate genres as presented in Section 2. Table 2 lists some common expressions of each debate genre. Then we search the trigger expressions in the vicinity of $word_j$ which is restricted by parameter d. Finally we employ the algorithm to compute its value, as illustrated in Algorithm 1.

Table 2. Trigger expressions of debate genres

Debate genre	trigger expressions
Denial	not(不),deny(否认),cannot(无法), etc
Question	interrogative particle(难道, 吗, 呢, 么, etc)
Assumption	if(如果, 假设, 的话, etc)
Concession	although(即使, 尽管, 就算, 虽然, etc)
Repeat	as you said(如您所说),rival(对方), etc

5. EXPERIMENTS

To explore how acoustic features and textual information influence stance-taking, we conduct three experiments. First, we focus on acoustic features to distinguish between subjective and non-subjective utterances (i.e. whether speaker is taking a stance or not). Second, we employ support topic feature (STF) to recognize stances in text modality. Lastly, we combine acoustic (including emphasis information) and textual features to examine whether the performance of stance classification can be further improved.

5.1. Acoustic experiment

In this experiment, we formulate the work as a dual-side classification task: Side 0 represents the neutral group, including presenter and jury; Side 1 represents the subjective: affirmative and negative. We extract features including F0 (fundamental frequency), Loudness, VoiceProb (voicing probability), ZCR (zero-crossing rate), MFCCs (Mel-Frequency Cepstral Coefficients) with openSmile and Speaking rate by praat[7] from the acoustic speech of datasets. OpenSmile calculates features for each frame. Statistical results of these features are further computed for each utterance, including max, min, range, maxPos (the position of maximum value), minPos, mean, stddev, skewness, and kurtosis.

Experimental results are shown in Table 3. As [12] has suggested, MFCCs are better than any other feature, and combining all features does not improve the performance significantly. A reason may be that MFCCs give a more comprehensive measure of voice source. It should be noted the performance of speaking rate, though only one attribute, is second to MFCCs. It is reasonable that people may speed up while they express their opinions eagerly when taking stance. Note that features like MFCCs and F0 are closely related to speaker, e.g., the F0 of women is usually higher than men,

[6] http://gibbslda.sourceforge.net/

[7] http://www.praat.org

Table 3. F1-Measure for acoustic features in distinguishing subjective and non-subjective utterances with three-fold cross-validation

Features	#Attributes	Side 0	Side 1	Average
Speaking rate	1	0.608	0.840	0.763
F0	9	0.529	0.846	0.741
Loudness	9	0.430	0.853	0.738
VoiceProb	9	0.559	0.834	0.743
ZCR	9	0.183	0.798	0.593
MFCCs	108	0.767	0.889	0.848
Combined	145	0.788	0.888	0.853

Table 4. F1-Measure of different textual features in recognizing stances (affirmative, negative and neutral) with three-fold cross-validation

Datasets / Features	Money	Nature	Business	Knowing
Unigram	0.821	0.798	0.803	0.774
Bigram	0.833	0.714	0.779	0.723
STF	**0.848**	**0.829**	**0.827**	**0.853**

which results in some inferences to distinguish subjective and non-subjective utterances from different speakers.

5.2. Textual experiment

We have verified that by using acoustic features we can distinct the utterances of stances (including affirmative and negative) from neutrals. To recognize stances in depth, we conduct experiments with textual information.

Much previous work has attempted to recognize stances or sentiment polarities with text features. [13] concludes related work and proves that Naive Bayes (NB) outperforms Support Vector Machines (SVMs) for short snippet sentiment tasks. It also points out the usefulness of n-gram models has been underappreciated. In our work, we choose unigram and bigram as baseline features and NB as classifier.

The results are shown in Table 4, which indicate that STF has the highest performance for all datasets. The F1-measure using STF can be improved by 3% $\sim$ 8% as compared to unigram or bigram baseline features. We also find that unigram outperforms bigram on our debate datasets.

Many factors can influence the results, such as the number of topics n and parameter d in p_j^d. We conduct a series of experiments to study how these parameters affect classification accuracy. Further experiments prove that with the increase of topics number, STF performs better and better. In addition, we figure out that $d = 4$ leads to the best performance. Hence, in all experiments, we set $d = 4$.

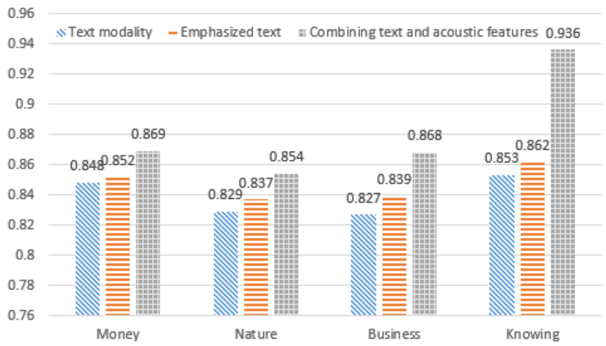

Fig. 2. Comparisons of text modality and combining text and acoustic features.

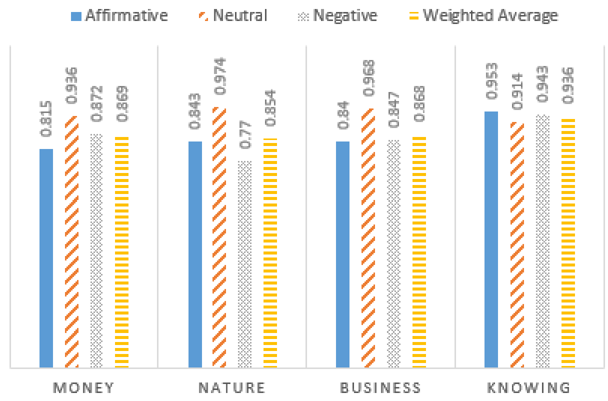

Fig. 3. F1-Measure comparisons of three sides for affirmative, negative and neutral

The fact that the unigram model performs better than bigram model, which is different from the conclusion presented in [13], should be further investigated. The reason may be that we put word, rather than Chinese character, as the unit. The bigram information captured in [13] has already been represented by the word unit.

5.3. Bimodal experiment

We further conduct the bimodal experiment to investigate whether performance can be improved when taking into account both acoustic and textual features. And if it is, we want to figure out whether emphasis information is helpful in recognizing stances.

We conduct experiments in two cases: a) simply concatenating acoustic features to STF and b) extracting STF again using corpus considering the emphasis information of words (i.e. whether the word is acoustically emphasized). In b) whether a word is emphasized is automatically detected from acoustic features [14] and manually checked and corrected by subjective listening. When extracting topics and transforming sentence to support topic feature, we treat the emphasized word as occurring twice.

Figure 2 shows that compared with the best results of text modality only, the emphasis information in case b) actually improve the classification accuracy slightly. It agrees with common sense that speakers always emphasize the keyword to highlight it to others. As for case a), STF combined with acoustic features can further improve the performance by at most 8.3%. This is because acoustic information corrects the error generated by the ambiguity of text. For instance, in the last of each debate competition, jury will comment on the performance of both sides, utterances in which may contain many topic and sentiment words. Judging only by text modality is likely to classify these utterances to the subjective side, while with acoustic features they can be correctly judged to be neutral.

Figure 3 depicts detailed F1-measure comparisons of different sides (affirmative, negative and neutral) on the four debate competition datasets. It should be noted the F1-measures for different sides vary a lot. Except for the "Knowing" dataset, F1-measure of the neutral side is the highest. Such phenomena may be explained by the performance differences introduced by acoustic features on different sides. As shown in Section 5.1, acoustic features demonstrate good performance in distinguishing non-subjective (i.e. neutral) utterances from subjective ones; but can hardly tell affirmative stances from negative ones (that usually can be distinguished from semantic meaning of text modality). As for the "Knowing" dataset, we compare its audio with the others and find, in "Knowing", four participants of affirmative are men while the participants of negative are all women, which leads to the high accuracy of distinguishing affirmative from negative because of the gender information inferred from acoustic features. This interesting phenomenon suggests speaker-related information may probably have been utilized in recognizing stances that should be avoided with better modeling technologies in future work.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we study the problem of recognizing stances and polarities in debates. We investigate the debate genres and discover many deceiving features, including rhetorical question, repeated rival statement, concession, etc. To address the challenges we propose a novel feature named Support Topic Feature (STF) based on the fact that topics carry emotional tendencies in debates. Experiments on four debate competitions validate the effectiveness of STF and combining STF with acoustic features can improve performance further. On the other hand, in English the assertive sentence has different grammatical structure from the rhetorical question. However in Mandarin, they may be literally identical: two utterances may look the same except question mark, while their meanings are opposite. In this situation, determining the intonation (e.g. question) from acoustic features would be very crucial for recognizing stances. In the future, we plan to enhance the method in such aspect.

7. ACKNOWLEDGEMENT

This work is supported by the National Basic Research Program of China (2012CB316401), the National High Technology Research and Development Program of China (2015AA016305), the National Natural Science Foundation of China (NSFC) (61375027, 61433018, 61370023 and 61171116), the joint fund of NSFC-RGC (Research Grant Council of Hong Kong) (61531166002, N_CUHK404/15) and the Major Program for National Social Science Foundation of China (13&ZD189).

References

[1] D. Biber, S. Johansson, G. Leech, S. Conrad, E. Finegan, R. Quirk, *Longman grammar of spoken and written English*, vol. 2, MIT Press, 1999.

[2] P. Haddington, "Stance taking in news interviews," *SKY Journal of Linguistics*, vol. 17, pp. 101-142, 2004.

[3] B. Pang, L. Lee, S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proc. ACL*, 2012.

[4] J. Wiebe, T. Wilson, C. Cardie, "Annotating expressions of opinions and emotions in language," *Language resources and evaluation*, vol. 39, no. 2-3, pp. 165-210, 2005.

[5] D. Blei, A. Ng, M. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp.993-1022, 2003.

[6] D. Sridhar, L. Getoor, M. Walker, "Collective stance classification of posts in online debate forums," in *Proc. ACL*, 2014.

[7] D. Sridhar, J. Foulds, B. Huang, L. Getoor, M. Walker, "Joint models of disagreement and stance in online debate," in *Proc. ACL*, 2015.

[8] K. Hasan, V. Ng, "Why are you taking this stance? identifying and classifying reasons in ideological debates," in *Proc. EMNLP*, 2014.

[9] G. Murray, G. Carenini, "Detecting subjectivity in multiparty speech," in *Proc. INTERSPEECH*, 2009.

[10] S. Raaijmakers, K. Truong, T. Wilson, "Multimodal subjectivity analysis of multiparty conversation," in *Proc. EMNLP*, 2008.

[11] S. Somasundaran, J. Wiebe, "Recognizing stances in online debates," in *Proc. ACL*, 2009.

[12] D. Neiberg, K. Elenius, K. Laskowski, "Emotion recognition in spontaneous speech using gmms," in *Proc. INTERSPEECH*, 2006.

[13] S. Wang, C. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *Proc. ACL*, 2012.

[14] Y. Ning, Z. Wu, X. Luo, H. Meng, J. Jia, L. Cai, "Using tilt for automatic emphasis detection with Bayesian Networks," in *Proc. INTERSPEECH*, 2015.