Detecting Subgroups in Online Discussions by Modeling Positive and Negative Relations among Participants

Ahmed Hassan
Microsoft Research
Redmond, WA
hassanam@microsoft.com

Amjad Abu-Jbara
University of Michigan
Ann Arbor, MI
amjbara@umich.edu

Dragomir Radev
University of Michigan
Ann Arbor, MI
radev@umich.edu

Abstract

A mixture of positive (friendly) and negative (antagonistic) relations exist among users in most social media applications. However, many such applications do not allow users to explicitly express the polarity of their interactions. As a result most research has either ignored negative links or was limited to the few domains where such relations are explicitly expressed (e.g. Epinions trust/distrust). We study text exchanged between users in online communities. We find that the polarity of the links between users can be predicted with high accuracy given the text they exchange. This allows us to build a signed network representation of discussions; where every edge has a sign: positive to denote a friendly relation, or negative to denote an antagonistic relation. We also connect our analysis to social psychology theories of balance. We show that the automatically predicted networks are consistent with those theories. Inspired by that, we present a technique for identifying subgroups in discussions by partitioning singed networks representing them.

1 Introduction

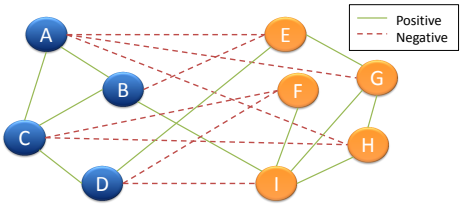
Most online communities involve a mixture of positive and negative relations between users. Positive relations may indicate friendship, agreement, or approval. Negative relations usually indicate antagonism, opposition, or disagreement.

Most of the research on relations in social media applications has almost exclusively focused on positive links between individuals (e.g. friends, fans, followers, etc.). We think that one of the main reasons, of why the interplay of positive and negative links did not receive enough attention, is the lack of a notion for explicitly expressing negative interactions. Recently, this problem has received increasing attention. However, all studies have been limited to a handful of datasets from applications that allow users to explicitly label relations as either positive or negative (e.g. trust/distrust on Epinion (Leskovec et al., 2010b) and friends/foes on Slashdot (Kunegis et al., 2009)).

Predicting positive/negative relations between discussants is related to another well studied problem, namely debate stance recognition. The objective of this problem is to identify which participants are supporting and which are opposing the topic being discussed. This line of work does not pay enough attention to the relations between participants, rather it focuses on participant's stance toward the topic. It also assumes that every participant either supports or opposes the topic being discussed. This is a simplistic view that ignore the nature of complex topics that has many aspects involved which may result in more than two subgroups with different opinions.

In this work, we apply Natural Language Processing techniques to text correspondences exchanged between individuals to identify the underlying signed social structure in online communities. We present a method for identifying user attitude and for automatically constructing a signed social network representation of discussions. We apply the proposed methods to a large set of discussion posts. We evaluate the performance using a manually labeled dataset. We also conduct a large scale evaluation by showing that predicted links are consistent with the principals of social psychology theories, namely the Structural Balance Theory (Heider, 1946). The balance theory has been shown to hold both theoretically (Heider, 1946) and empirically (Leskovec et al., 2010c) for a variety of social community settings. Finally, we present a method for identifying subgroups in online discussions by identifying groups with high density of intra-group positive relations and high density of inter-group negative relations. This method is capable of identifying subgroups even if the community splits into more than two subgroups which is more general than stance recognition which assumes that only two groups exist.

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 59–70, Jeju Island, Korea, 12–14 July 2012.



Source	Target	Sign	Evidence from Text
A	E	-	I have to disagree with what you are saying.
G	A	-	You are missing the entire point, he is putting lives at risk.
D	I	-	and you manufacture lies for what reason?
E	G	+	you have explained your position very well.
C	H	+	I am neutral on this, but I agree with your assessment!

Figure 1: An example showing a signed social network along with evidence from text that justifies edge signs.

The input to our algorithm is a set of text correspondences exchanged between users (e.g. posts or comments). The output is a *signed network* where edges signify the existence of an interaction between two users. The resulting network has polarity associated with every edge. Edge polarity is a means for indicating positive or negative affinity between two individuals.

Figure 1 shows a signed network representation for a subset of posts from a long discussion thread. The thread discussed the November 2010 Wikileaks cable release. We notice that participants split into two groups, one supporting and one opposing the leak. We also notice that most negative edges are between groups, and most positive edges are within groups. It is worth mentioning that networks generated from larger datasets (i.e. with thousands of posts) have much more noise compared to this example.

The rest of the paper is structured as follows. In section 2, we review some of the related prior work on mining sentiment from text, mining online discussions, extracting social networks from text, and analyzing signed social networks. We define our problem and explain our approach in Section 3. Section 4 describes our dataset. Results and discussion are presented in Section 5. We present a method for identifying subgroups in online discussions in Section 3.3. We conclude in Section 6.

2 Related Work

In this section, we survey several lines of research that are related to our work.

2.1 Mining Sentiment from Text

Our general goal of mining attitude from one individual toward another makes our work related to a huge body of work on sentiment analysis. One such line of research is the well-studied problem of identifying the polarity of individual words (Hatzivassiloglou and McKeown, 1997; Turney and Littman, 2003; Kim and Hovy, 2004; Takamura et al., 2005). Subjectivity analysis is yet another research line that is closely related to our general goal of mining attitude. The objective of subjectivity analysis is to identify text that presents opinion as opposed to objective text that presents factual information (Wiebe, 2000; Hatzivassiloglou and Wiebe, 2000; Banea et al., 2008; Riloff and Wiebe, 2003). Our work is different from subjectivity analysis because we are not only interested in discriminating between opinions and facts. Rather, we are interested in identifying the polarity of interactions between individuals. Our method is not restricted to phrases or words, rather it generalizes this to identifying the polarity of an interaction between two individuals based on several posts they exchange.

2.2 Stance Classification

Perhaps the closest work to this paper is the work on stance classification. We notice that most of these methods focus on the polarity of the written text assuming that anyone using positive text belongs to one group and anyone using negative text belongs to another. This works well for single-aspect topics or entities like the ones used in (Tan et al., 2011) (e.g. Obama, Sara Palin, Lakers, etc.). In this simple notion of topics, it is safe to assume that text polarity is a good enough discriminator. This unfortunately is not the case in online discussions about complex topics having many aspects (e.g. abortion, health care, etc.). In such complex topics, people use positive and negative text targeting different aspects of the topic, for example in the health care bill topic, discussants expressed their opinion regarding many aspects including: the enlarged coverage, the insurance premiums, Obama, socialism, etc. This shows that simply looking at text polarity is not enough to identify groups.

Tan et al. (2011) studied how twitter following relations can be used to improve stance classification. Their main hypothesis is that connected users are more likely to hold similar opinions. This may be correct for the twitter following relations, but it is not necessarily correct for open discussions where

no such relations exist. The only criterion that can be used to connect discussants is how often they reply to each other's posts. We will show later that while many people reply to people with similar opinions, many others reply to people with different opinions as well.

Thomas et al. (2006) address the same problem of determining support and opposition as applied to congressional floor-debates. They assess the agreement/disagreement between different speakers by training a text classifier and applying it to a window surrounding the names of other speakers. They construct their training data by assuming that if two speaker have the same vote, then every reference connecting them is an agreement and vice versa. We believe this will result in a very noisy training/testing set and hence we decided to recruit human annotators to create a training set. We found out that many instances with references to other discussants were labeled as neither agreement nor disagreement regardless of whether the discussants have similar or opposing positions. We will use this system as a baseline and will show that the existence of positive/negative words close to a person name does not necessarily show agreement or disagreement with that person.

Hassan et al. (2010) use a language model based approach for identifying agreement and disagreement sentences in discussions. This work is limited to sentences. It does not consider the overall relation between participants. It also does not consider subgroup detection. We will use this method as a baseline for one of our components and will show that the proposed method outperforms it.

Murakami and Raymond (2010) present another method for stance recognition. They use a small number of hand crafted rules to identify agreement and disagreement interactions. Hand crafted rules usually result in systems with very low recall causing them to miss many agreement/disagreement instances (they report 0.26 recall at the 0.56 precision level). We present a machine learning system to solve this problem and achieve much better performance. Park et al. (2011) propose a method for finding news articles with different views on contentious issues. Mohit et al. (2008) present a set of heuristics for including disagreement information in a minimum cut stance classification framework. Galley et al. (2004) show the value of using durational and structural features for identifying agreement and disagreement in spoken conversational speech. They use features like duration of spurts, speech rate, speaker overlap, etc. which are not applicable to written language.

Our approach is different from agreement/disagreement identification because we not only study sentiment at the local sentiment level but also at the global level that takes into consideration many posts exchanged between participants to build a signed network representation of the discussion. Research on debate stance recognition attempts to perform classification under the "supporting vs. opposing" paradigm. However such simple view might not always be accurate for discussions on more complex topics with many aspects. After building the signed network representation of discussions, we present a method that can detect how the large group could split into many subgroups (not necessarily two) with coherent opinions.

2.3 Extracting Social Networks from Text

Little work has been done on the front of extracting social relations between individuals from text. Elson et al. (2010) present a method for extracting social networks from nineteenth-century British novels and serials. They link two characters based on whether they are in conversation or not. McCallum et al. (2007) explored the use of structured data such as email headers for social network construction. Gruzd and Hyrthonthwaite (2008) explored the use of post text in discussions to study interaction patterns in e-learning communities. Extracting social power relations from natural language (i.e. who influences whom) has been studied in (Bramsen et al., 2011; Danescu-Niculescu-Mizil et al., 2011).

Our work is related to this line of research because we employ natural language processing techniques to reveal embedded social structures. Despite similarities, our work is uniquely characterized by the fact that we extract signed social networks with both positive and negative links from text.

2.4 Signed Social Networks

Most of the work on social networks analysis has only focused on positive interactions. A few recent papers have taken the signs of edges into account.

Brzozowski et al. (2008) study the positive and negative relationships between users of Essembly. Essembly is an ideological social network that distinguishes between ideological allies and nemeses. Kunegis et al. (2009) analyze user relationships in

the Slashdot technology news site. Slashdot allows users of the website to tag other users as friends or foes, providing positive and negative endorsements. Leskovec et al. (2010b) study signed social networks generated from Slashdot, Epinions, and Wikipedia. They also connect their analysis to theories of signed networks from social psychology. A similar study used the same datasets for predicting positive and negative links given their context (Leskovec et al., 2010a).

All this work has been limited to analyzing a handful of datasets for which an explicit notion of both positive and negative relations exists. Our work goes beyond this limitation by leveraging the power of natural language processing to automate the discovery of signed social networks using the text embedded in the network.

The research presented in this paper extends this previous work in a number of ways: (i) we present a method based on linguistic analysis that finds instances of showing positive or negative attitude between participants (ii) we propose a technique for representing discussions as signed networks where a sign is associated with every edge to denote whether the relation is friendly or antagonistic (iii) we evaluate the proposed methods using human annotated data and also conduct a large scale evaluation based on social psychology theories; (iv) finally we present a method for identifying subgroups that globally splits the community involved in the discussion by utilizing the dynamics of the local interactions between participants.

3 Approach

3.1 Identifying Attitude from Text

To build a signed network representation of discussants, we start by trying to identify sentences that show positive or negative attitude from the writer to the addressee. The first step toward identifying attitude is to identify words with positive/negative semantic orientation. The semantic orientation or polarity of a word indicates the direction the word deviates from the norm (Lehrer, 1974). We use OpinionFinder (Wilson et al., 2005a) to identify words with positive or negative semantic orientation. The polarity of a word is also affected by the context where the word appears. For example, a positive word that appears in a negated context should have a negative polarity. Other polarized words sometimes appear as neutral words in some contexts. To identify contextual polarity of words, a large set of features is used including words, sentences, structure, and other features similar to the method described in (Wilson et al., 2005b).

Our overall objective is to find the direct attitude between participants. Hence after identifying the semantic orientation of individual words, we move on to predicting which polarized expressions target the addressee and which do not.

Text polarity alone cannot be used to identify attitude between participants. Sentences that show an attitude are different from subjective sentences. Subjective sentences are sentences used to express opinions, evaluations, and speculations (Riloff and Wiebe, 2003). While every sentence that shows an attitude is a subjective sentence, not every subjective sentence shows an attitude toward the recipient.

In this method, we address the problem of identifying sentences with attitude as a relation detection problem in a supervised learning setting. We study sentences that has mentions to the addressee and polarized expressions (negative/positive words or phrases). Mentions could either be names of other participants or second person pronouns (you, your, yours) used in text posted as a reply to another participant. Reply structure (i.e. who replies to whom) is readily available in many discussion forums. In cases where reply structure is not available, we can use a method like the one in (Lin et al., 2009) to recover it.

We predict whether the mention is related to the polarized expression or not. We regard the mention and the polarized expression as two entities and try to learn a classifier that predicts whether the two entities are related or not.

The text connecting the two entities offers a very condensed representation of the information needed to assess whether they are related or not. For example the two sentences *"you are completely unqualified"* and *"you know what, he is unqualified ..."* show two different ways the words *"you"*, and *"unqualified"* could appear in a sentence. In the first case the polarized word "unqualified" refers to the word "you". In the second case, the two words are not related. The information in the shortest path between two entities in a dependency tree can be used to assert whether a relationship exists between them (Bunescu and Mooney, 2005).

The sequence of words connecting the two entities is a very good predictor of whether they are related or not. However, these paths are completely

lexicalized and consequently their performance will be limited by data sparseness. To alleviate this problem, we use higher levels of generalization to represent the path connecting the two tokens. These representations are the part-of-speech tags, and the shortest path in a dependency graph connecting the two tokens. We represent every sentence with several representations at different levels of generalization. For example, the sentence "*your ideas are very inspiring*" will be represented using lexical, polarity, part-of-speech, and dependency information as follows:

LEX: "*YOUR ideas are very POS*"
POS: "*YOUR NNS VBP RB JJ_POS*"
DEP: "*YOUR poss nsubj POS*"

The set of features we use are the set of unigrams, and bigrams representing the words, part-of-speech tags, and dependency relations connecting the two entities. For example the following features will be set for the previous example:

YOUR_ideas, YOUR_NNS, YOUR_poss, poss_nsubj, ..., etc.

We use Support Vector Machines (SVM) as a learning system because it is good with handling high dimensional feature spaces.

3.2 Extracting the Signed Network

In this subsection, we describe the procedure we used to build the signed network given the component we described in the previous subsection. This procedure consists of two main steps. The first is building the network without signs, and the second is assigning signs to different edges.

To build the network, we parse our data to identify different threads, posts and senders. Every sender is represented with a node in the network. An edge connects two nodes if there exists an interaction between the corresponding participants. We add a directed edge $A \rightarrow B$, if A replies to B's posts at least n times in m different threads. We set m, and n to 2 in all of our experiments. The interaction information (i.e. who replies to whom) can be extracted directly from the thread structure. Alternatively, as mentioned earlier, we can use a method similar to the one presented in (Lin et al., 2009) to recover the reply structure if it is not readily available.

Once we build the network, we move to the more challenging task in which we associate a sign with every edge. We have shown in the previous section how sentences with positive and negative attitude can be extracted from text. Unfortunately the sign of an interaction cannot be trivially inferred from the polarity of sentences. For example, a single negative sentence written by A and directed to B does not mean that the interaction between A and B is negative. One way to solve this problem would be to compare the number of negative sentences to positive sentences in all posts between A and B and classify the interaction according to the plurality value. We will show later, in our experiments section, that such a simplistic method does not perform well in predicting the sign of an interaction.

Participant Features
Number of posts per month for A (B)
Percentage of positive posts per month for A (B)
Percentage of negative posts per month for A (B)
gender
Interaction Features
Percentage/number of positive (negative) sentences per post
Percentage/number of positive (negative) posts per thread
Discussion Domain (e.g. politics, science, etc.)

Table 1: Features used by the Interaction Sign Classifier.

As a result, we decided to pose the problem as a classical supervised learning problem. We came up with a set of features that we think are good predictors of the interaction sign, and we trained a classifier using those features on a labeled dataset. Our features include numbers and percentages of positive/negative sentences per post, posts per thread, and so on. A sentence is labeled as positive/negative if a relation has been detected in this sentence between a mention referring to the addressee and a positive/negative expression. A post is considered positive/negative based on the majority of relations detected in it. We use two sets of features. The first set is related to A only or B only. The second set is related to the interactions between A and B. The features are summarized in Table 1.

3.3 Sub-Group Detection

In any discussion, different subgroups may emerge. Members of every subgroup usually have a common focus (positive or negative) toward the topic being discussed. Each member of a group is more likely to show positive attitude to members of the same group, and negative attitude to members of opposing groups. The signed network representation could prove to be very useful for identifying those subgroups. To detect subgroups in a discussion thread,

we would like to partition the corresponding signed network such that positive intra-group links and negative inter-group links are dense.

This problem is related to the constrained clustering (Wagstaff et al., 2001) and the correlation clustering problem (Bansal et al., 2004). In constrained clustering, a pairwise similarity metric (which is not available in our domain), and a set of must-link/cannot-link constraints are used with a standard data clustering algorithm. Correlation clustering operates in a scenario where given a signed graph $G = (V, E)$ where the edge label indicates whether two nodes are similar (+) or different (-), the task is to cluster the vertices so that similar objects are grouped together. Bansal et. al (2004) proved NP-hardness and gave constant-factor approximation algorithms for the special case in which the graph is complete (full information) and every edge has weight +1 or -1 which is not the case in our network. Alternatively, we can use a greedy optimization algorithm to find partitions. A criterion function for a local optimization partitioning procedure is constructed such that positive links are dense within groups and negative links are dense between groups.

For any potential partition C, we seek to optimize the following function: $P(C) = \alpha \sum_n + (1-\alpha) \sum_p$ where $\sum_n$ is the number of negative links between nodes in the same subgroup, $\sum_p$ is the number of positive links between nodes in different subgroups, and α is a trade factor that represents the importance of the two terms. We set α to 0.5 in all our experiments.

Clusters are selected such that: $C^* = \arg\min P(C)$. A greedy optimization framework is used to minimize $P(C)$. Initially, nodes are randomly partitioned into t different clusters and the criterion function P is evaluated for that cluster. Every cluster has a set of neighbors in the cluster space. A neighbor cluster is obtained by moving one node from one cluster to another, or by exchanging two nodes in two different clusters. Neighbor partitions are evaluated, and if one with a lower value for the criterion function is found, it is set as the current partition. This greedy procedure is repeated with random restarts until a minimal solution is found. To determine the number of subgroups t, we select t that minimizes the optimization function $P(C)$. In all experiments we used an upper limit of $t = 5$. This technique was able to identify the correct number of subgroups in 77% of the times. In the rest of the cases, the number was different from the correct number by at most 1 except for a single case where it was 2.

4 Data

4.1 Signed Network Extraction

Our data consists of a large amount of discussion threads collected from online discussion forums. We collected around $41,000$ topics (threads) and 1.2M posts from the period between the end of 2008 and the end of 2010. All threads were in English and had 5 posts or more. They covered 11 different domains including: politics, religion, science, etc. The average number of participants per domain is 1320 and per topic is 52. The data was tokenized, sentence-split, and part-of-speech tagged with the OpenNLP toolkit. It was parsed with the Stanford parser (Klein and Manning, 2003).

We randomly selected around 5300 posts (1000 interactions), and asked human annotators to label them. Our annotators were instructed to read all the posts exchanged between two participants and decide whether the interaction between them is positive or negative. We used Amazon Mechanical Turk for annotations. Following previous work (Callison-Burch, 2009; Akkaya et al., 2010), we took several precautions to maintain data integrity. We restricted annotators to those based in the US to maintain an acceptable level of English fluency. We also restricted annotators to those who have more than 95% approval rate for all previous work. Moreover, we asked three different annotators to label every interaction. The label was computed by taking the majority vote among the three annotators. We refer to this data as the *Interactions Dataset*.

We ran a different annotation task where we selected sentences including mentions referring to discussants (names or pronouns) and polarized expressions. Annotators were asked to select sentences where the polarized attribute is referring to the mention and hence show a positive or negative attitude toward other discussion participants. This resulted in a set of 5000 manually annotated sentences. We refer to this data as the *Sentences Dataset*.

We asked three different annotators to label every instance. The kappa measure between the three groups of annotations was 0.62 for the *Interactions Dataset* and 0.64 for the *Sentences Dataset*. To better assess the quality of the annotations, we asked a trained annotator to label 10% of the data. We measured the agreement between the expert annotator

	Class	Pos.	Neg.	Weigh. Avg.
Logistic Reg.	Precision	0.848	0.724	0.809
	Recall	0.884	0.657	0.812
	F-Measure	0.866	0.689	0.81
	Accuracy	-	-	**0.812**
SVM	Precision	0.906	0.71	0.844
	Recall	0.847	0.809	0.835
	F-Measure	0.875	0.756	0.838
	Accuracy	-	-	**0.835**

Table 2: Interaction sign classifier performance.

Classifier	Random	Thresh-Num	Thresh-Perc.	SVM
Accuracy	65%	69%	71%	83.5%

Table 3: A comparison of different sign interaction classifiers.

and the majority label from Mechanical Turk. The kappa measure was 0.69 for the *Interactions Dataset* and 0.67 for the *Sentences Dataset*.

4.2 Sub-group Detection

We used a dataset of more than 42 topics and approximately 9000 posts collected from two political forums (Createdebate[1] and Politicalforum[2]). The forum administrators ran a poll asking participants to select their stance from a set of possible answers and hence the dataset was self-labeled with respect to groups. We also used a set of discussions from the Wikipedia discussion section. When a topic on Wikipedia is disputed, the editors of that topic start a discussion about it. We collected 117 Wikipedia discussion threads. The threads contain a total of 1,867 posts. The discussions were annotated by an expert annotator (a professor in sociolinguistics, not an author of the paper) who was instructed to read each of the Wikipedia discussion threads in its entirety and determine whether the discussants split into subgroups, in which case he was asked to identify the subgroup membership for each discussant. In total, we had 159 topics with an average of approximately 500 posts, 60 participants and 2.7 subgroups per topic. Examples of the topics include: Arizona immigration law, airport security, oil spill, evolution, Ireland partitions, abortion and many others.

5 Results and Discussion

We performed experiments on the data described in the previous section. We trained and tested the sentence with the attitude detection classifiers described in Section 3.1 using the *Sentences Dataset*. We also trained and tested the interaction sign classifier described in Section 3.2 using the *Interactions Dataset*. We built one signed social network for every domain (e.g. politics, economics, etc.). We decided to build a network for every domain as opposed to one single network because the relation between any two individuals may vary across domains (e.g. politics vs. science). In the rest of this section, we will describe the experiments we did to assess the performance of the sentences with attitude detection and interaction sign prediction steps.

In addition to classical evaluation, we evaluate our results using the structural balance theory which has been shown to hold both theoretically (Heider, 1946) and empirically (Leskovec et al., 2010c). We validate our results by showing that the *automatically* extracted networks mostly agree with the theory. We evaluated the approach using the structural balance theory because it presents a global (pertaining to relations between multiple edges) and large-scale (used millions of posts and thousands of users) evaluation of the results as opposed to traditional evaluation which is local in nature (only considers one edge at a time) and smaller in scale (used thousands of posts).

5.1 Identifying Sentences with Attitude

We compare the proposed methods to two baselines. The first baseline is based on the work of (Thomas et al., 2006). We used the speaker agreement component presented in (Thomas et al., 2006) as a baseline. The speaker agreement component is one step in their approach. In this component, they used an SVM classifier trained using a window of text surrounding references to other speakers to predict agreement/disagreement between speakers.

We build an SVM text classifier trained on the sentence at which the mention referring to the other participant occurred. We refer to this baseline as the *T*ext Classification approach. The second baselines adopts the language model approach presented in (Hassan et al., 2010). Two language models are trained using a stream of words, part-of-speech tags, and dependency relations, one for sentences that show an attitude and one for sentences that do not. New sentences are classified based on generation likelihoods. We refer to this baseline as the *L*anguage Models approach.

We tested this component using the *Sentences Dataset* described in Section 4. We compared the performance of the proposed method and the two

[1] www.createdebate.com

[2] www.politicalforum.com

	Extracted Networks				Random Networks			
Domain	$(+++)$	$(++-)$	$(+--)$	$(---)$	$(+++)$	$(++-)$	$(+--)$	$(---)$
abortion	51.67	26.31	18.92	0.48	35.39	43.92	18.16	2.52
current-events	67.36	22.26	8.76	0.23	54.08	36.90	8.39	0.64
off-topic-chat	65.28	23.54	9.45	0.25	58.07	34.59	6.88	0.46
economics	72.68	18.30	7.77	0.00	66.50	29.09	4.22	0.20
political opinions	60.60	24.24	12.81	0.43	45.97	40.79	12.06	1.19
environment	47.46	32.54	17.26	0.30	37.38	43.61	16.89	2.12
latest world news	58.29	22.41	16.33	0.62	42.26	42.20	13.98	1.56
religion	47.17	25.89	22.56	1.42	39.68	42.94	15.51	1.87
science-technology	57.53	26.03	14.33	0.00	50.14	38.93	10.05	0.87
terrorism	64.96	23.36	9.46	0.73	41.54	42.42	14.36	1.68

Table 4: Percentage of different types of triangles in the extracted networks vs. the random networks.

Method	Accuracy	Precision	Recall	F1
Text Classification	60.4	61.1	60.2	60.6
Language Models	80.3	81.0	79.4	80.2
Relation Extraction	82.3	82.3	82.3	82.3

Table 5: Comparison of attitude identification methods.

baselines. Table 5 compares the precision, recall, F1, and accuracy for the three methods. The text classification based approach does much worse than others. The reasons is that it ignores the structure and uses much less information (part-of-speech tags and dependency trees are not used) compared to the other methods. Additionally, the short length of the sentences compared to what is typical in text classification may have had a bad effect on the performance. Both other models try to learn the characteristics of the path connecting the mention and the polarized expression. We notice that optimizing the weights for unigram and bigrams features using SVM results in a better performance compared to language models because it does not have the constraints imposed by the former model on the learned weights.

We evaluated the importance of the feature types (i.e. dependency vs. pos tags vs words) by measuring the chi-squared statistic for every feature with respect to the class. Dependency features were most helpful, but other types of features helped improve the performance as well.

5.2 Interaction Sign Classifier

We used the relation detection classifier described in Section 3.1 to find sentences with positive and negative attitude. The output of this classifier was used to compute the features described in Section 3.2, which were used to train a classifier that predicts the sign of an interaction between any two individuals.

We used both Support Vector Machines (SVM) and logistic regression to train the sign interaction classifier. We report several performance metrics for them in Table 2. We notice that the SVM classifier performs better with an accuracy of 83.5% and an F-measure of 81%. All results were computed using 10 fold cross validation on the labeled data. To better assess the performance of the proposed classifier, we compare it to a baseline that labels the relation as negative if the percentage of negative sentences exceeds a particular threshold, otherwise it is labeled as positive. The thresholds were empirically estimated using a separate development set. The accuracy of this baseline is only 71%.

To better assess the performance of the proposed classifier, we compare it to three baselines. The first is a random baseline that predicts an interaction as positive with probability p that equals the proportion of positive instances to all instances in the training set. The second classifier (Thresh-Num) labels the edge as negative if the number of negative instances exceeds a threshold T_n. The third classifier (Thresh-Perc) labels the edge as negative if the percentage of negative instances to all instances exceeds a threshold T_p. The cutoff thresholds were estimated using a separate development set.

The 3 baselines were tested using the entire labeled dataset. The SVM classifier was tested using 10 fold cross validation. The accuracy of the random classifier, the two based on a cut off number and percentage , and the SVM classifier are shown in Table 3. We notice that the random classifier performs worst, and the classifier based on percentage cutoff outperforms the one based on number cutoff. The SVM classifier significantly outperforms all other classifiers. We tried to train a classifier using both the number and percentage of negative and positive posts. The improvement over using the baseline using the percentage of negative posts was not statistically significant.

We evaluated the importance of the features listed

in Table 1 by measuring the chi-squared statistic for every feature with respect to the class. We found out that the features describing the interaction between the two participants are more informative than the ones describing individuals characteristics. The later features are still helpful though and they improve the performance by a statistically significant amount. We also noticed that all features based on percentages are more informative than those based on counts. The most informative features are: percentage of negative posts per tread, percentage of negative sentences per post, percentage of positive posts per thread, number of negative posts, and discussion domain.

5.3 Structural Balance Theory

The structural balance theory is a psychological theory that tries to explain the dynamics of signed social interactions. It has been shown to hold both theoretically (Heider, 1946) and empirically (Leskovec et al., 2010c). In this section, we study the agreement between the theory and our *automatically* extracted networks. The theory has its origins in the work of Heider (1946). It was then formalized in a graph theoretic form by (Cartwright and Harary, 1956). The theory is based on the principles that "the friend of my friend is my friend", "the enemy of my friend is my enemy", "the friend of my enemy is my enemy", and variations on these. The structural balance theory states that triangles that have an odd number of positive signs (+ + + and + - -) are balanced, while triangles that have an even number of positive signs (- - - and + + -) are not.

In this section, we compare the predictions of edge signs made by our system to the structural balance theory by counting the frequencies of different types of triangles in the predicted network. Table 4 shows the frequency of every type of triangle for 10 different domains. To better understand these numbers, we compare them to the frequencies of triangles in a set of random networks. We shuffle the signs for all edges on every network keeping the fractions of positive and negative edges constant. We repeat shuffling for 1000 times and report the average.

We find that the all-positive triangle $(+ + +)$ is overrepresented in the generated network compared to chance across all domains. We also see that the triangle with two positive edges $(+ + -)$, and the all-negative triangle $(- - -)$ are underrepresented compared to chance across all domains. The triangle with a single positive edge is slightly overrepresented in most but not all of the topics compared to chance. This shows that the predicted networks mostly agree with the structural balance theory. The slightly non standard behavior of the triangle with one positive edge could be explained in light of the weak balance theory. In this theory, Davis (1967) states that this triangle, which corresponds to the "enemy of enemy is my friend" proposition, holds only if the network can be partitioned into exactly two subsets, but not when there are more than two. In general, the percentage of balanced triangles in the predicted networks is higher than in the shuffled networks, and hence the balanced triangles are significantly overrepresented compared to chance showing that our *automatically* constructed network is similar to explicit signed networks in that they both mostly agree with the balance theory.

5.4 Sub-Group Detection

We compare the performance of the sub-group detection method to three baselines. The first baseline uses graph clustering (GC) to partition a network based on the frequency of interaction between participants. We build a graph where each node represents a participant. Edges link participants if they exchange posts, and edge weights are based on the number of posts exchanged. The second baseline (TC) is based on the premise that participants with similar text are more likely to belong to the same subgroup. We measure text similarity by computing the cosine similarity between the tf-idf representations of the text in a high dimensional vector space. We tried two methods for partitioning those graphs: spectral partitioning (Luxburg, 2007) and a hierarchical agglomeration algorithm which works by greedily optimizing the modularity for graphs (Clauset et al., 2004). The third baseline is based on stance classification approaches (e.g. (Tan et al., 2011)). In this baseline we put all the participants who use more positive text in one subgroup and the participants who use more negative text in another subgroup. Text polarity is identified using the method described in Section 3.1.

Table 6 shows the average purity (*Purity*), entropy (*Entropy*), Normalizes Mutual Information (*NMI*), and Rand Index (*RandIndex*) values of the method based on signed networks and the baselines using different partitioning algorithms. The differences in the results shown in the table are statistically significant at the 0.05 level (as indicated by a 2-tailed

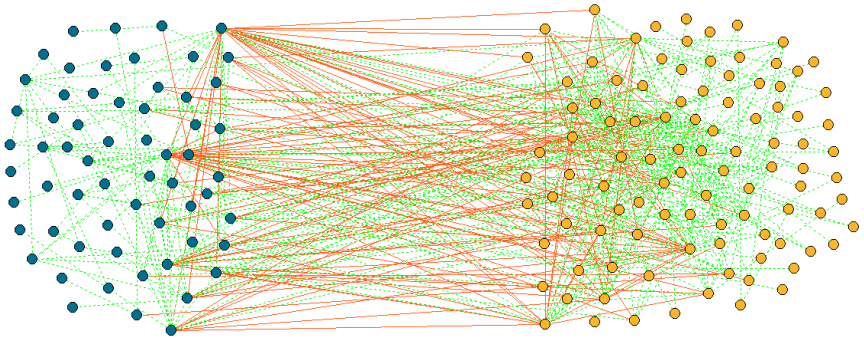


Figure 2: A signed network representing participants in a discussion about the "Health Care Reform Bill". Blue (dark) nodes represent participants with the bill, Yellow (light) nodes represent participants against the bill, red (solid) edges represent negative attitude, while green (dashed) edges represent positive attitude.

	Createdebate				Politicalforum				Wikipedia			
Method	Purity	Entropy	NMI	RandIndex	Purity	Entropy	NMI	RandIndex	Purity	Entropy	NMI	RandIndex
GC - Spectral	0.50	0.85	0.28	0.40	0.50	0.88	0.27	0.39	0.49	0.89	0.33	0.35
GC - Hierarchical	0.48	0.86	0.30	0.41	0.47	0.89	0.31	0.40	0.49	0.87	0.38	0.39
TC - Spectral	0.50	0.85	0.31	0.43	0.48	0.90	0.30	0.45	0.51	0.87	0.40	0.46
TC - Hierarchical	0.49	0.90	0.35	0.46	0.48	0.91	0.33	0.49	0.53	0.80	0.40	0.49
Text Polarity	0.55	0.80	0.38	0.49	0.54	0.91	0.31	0.38	0.34	0.95	0.30	0.40
Signed Networks	**0.64**	**0.74**	**0.46**	**0.59**	**0.58**	**0.80**	**0.43**	**0.55**	**0.65**	**0.54**	**0.51**	**0.60**

Table 6: Comparison of the sub-group detection method to baseline systems

paired t-test).

We notice that partitioning the signed network that was automatically extracted from text results in significantly better partitions on the three datasets as indicated by the higher Purity, NMI, and RandIndex and the lower Entropy values it achieves. We believe that the first two baselines performed poorly because the interaction frequency and the text similarity are not key factors in identifying subgroup structures. Many people would respond to people they disagree with more, while others would mainly respond to people they agree with most of the time. Also, people in opposing subgroups tend to use very similar text when discussing the same topic and hence text clustering does not work as well. The baseline that classifies the stance of discussants based on the polarity of their text performed bad too because it overlooks the fact that most of the discussed topics in our datasets have multiple aspects and a discussant may use both positive and negative text targeting different aspects of the topic. An example of a signed network and the corresponding subgtoups as extracted from real data is showm in Figure 2.

6 Conclusions

In this paper, we have shown that natural language processing techniques can be reliably used to extract signed social networks from text correspondences. We believe that this work brings us closer to understanding the relation between language use and social interactions and opens the door to further research efforts that go beyond standard social network analysis by studying the interplay of positive and negative connections. We rigorously evaluated the proposed methods on labeled data and connected our analysis to social psychology theories to show that our predictions mostly agree with them. Finally, we presented potential applications that benefit from the automatically extracted signed network.

Acknowledgments

This research was funded in part by the Office of the Director of National Intelligence, Intelligence Advanced Research Projects Activity. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI or the U.S. Government

References

Cem Akkaya, Alexander Conrad, Janyce Wiebe, and Rada Mihalcea. 2010. Amazon mechanical turk for subjectivity word sense disambiguation. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 195–203.

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2008. A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In *LREC'08*.

Nikhil Bansal, Avrim Blum, and Shuchi Chawla. 2004. Correlation Clustering. *Machine Learning*, 56(1):89–113.

Mohit Bansal, Claire Cardie, and Lillian Lee. 2008. The power of negative thinking: Exploiting label disagreement in the min-cut classification framework. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*.

Philip Bramsen, Martha Escobar-Molano, Ami Patel, and Rafael Alonso. 2011. Extracting social power relationships from natural language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pages 773–782.

Michael J. Brzozowski, Tad Hogg, and Gabor Szabo. 2008. Friends and foes: ideological social networking. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 817–820, New York, NY, USA.

Razvan C. Bunescu and Raymond J. Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 724–731, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chris Callison-Burch. 2009. Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 286–295.

Dorwin Cartwright and Frank Harary. 1956. Structure balance: A generalization of heiders theory. *Psych. Rev.*, 63.

Aaron Clauset, Mark E. J. Newman, and Cristopher Moore. 2004. Finding community structure in very large networks. *Phys. Rev. E*, 70:066111.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon M. Kleinberg. 2011. Echoes of power: Language effects and power differences in social interaction. *CoRR*.

J. A. Davis. 1967. Clustering and structural balance in graphs. *Human Relations*, 20:181–187.

David Elson, Nicholas Dames, and Kathleen McKeown. 2010. Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147, Uppsala, Sweden, July.

Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth 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*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Anatoliy Gruzd and Caroline Haythornthwaite. 2008. Automated discovery and analysis of social networks from threaded discussions. In *Proceedings of the International Network of Social Network Analysis (INSNA)*, St. Pete Beach, Florida.

Ahmed Hassan, Vahed Qazvinian, and Dragomir Radev. 2010. What's with the attitude?: identifying sentences with attitude in online discussions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1245–1255.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *EACL'97*, pages 174–181.

Vasileios Hatzivassiloglou and Janyce Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *COLING*, pages 299–305.

Fritz Heider. 1946. Attitudes and cognitive organization. *Journal of Psychology*, 21:107–112.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *COLING*, pages 1367–1373.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *ACL'03*, pages 423–430.

Jérôme Kunegis, Andreas Lommatzsch, and Christian Bauckhage. 2009. The slashdot zoo: mining a social network with negative edges. In *Proceedings of the 18th international conference on World wide web*, pages 741–750, New York, NY, USA.

Adrienne Lehrer. 1974. Semantic fields and lezical structure. North Holland, Amsterdam and New York.

Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010a. Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World wide web*, pages 641–650, New York, NY, USA.

Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010b. Signed networks in social media. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 1361–1370, New York, NY, USA.

Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010c. Signed networks in social media. In *CHI 2010*, pages 1361–1370, New York, NY, USA. ACM.

Chen Lin, Jiang-Ming Yang, Rui Cai, Xin-Jing Wang, and Wei Wang. 2009. Simultaneously modeling semantics and structure of threaded discussions: a sparse coding approach and its applications. In *SIGIR '09*, pages 131–138.

Ulrike Luxburg. 2007. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, December.

Andrew McCallum, Xuerui Wang, and Andrés Corrada-Emmanuel. 2007. Topic and role discovery in social networks with experiments on enron and academic email. *J. Artif. Int. Res.*, 30:249–272, October.

Akiko Murakami and Rudy 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.

Souneil Park, KyungSoon Lee, and Junehwa Song. 2011. Contrasting opposing views of news articles on contentious issues. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pages 340–349.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *EMNLP'03*, pages 105–112.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2005. Extracting semantic orientations of words using spin model. In *ACL'05*, pages 133–140.

Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. 2011. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 1397–1405.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *In Proceedings of EMNLP*, pages 327–335.

Peter Turney and Michael Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21:315–346.

Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schrödl. 2001. Constrained k-means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 577–584.

Janyce Wiebe. 2000. Learning subjective adjectives from corpora. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 735–740.

Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005a. Opinionfinder: a system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, HLT-Demo '05, pages 34–35, Stroudsburg, PA, USA. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005b. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT/EMNLP'05*, Vancouver, Canada.

Bo Yang, William Cheung, and Jiming Liu. 2007. Community mining from signed social networks. *IEEE Trans. on Knowl. and Data Eng.*, 19(10):1333–1348.