Combining Social Cognitive Theories with Linguistic Features for Multi-genre Sentiment Analysis

Hao Li[1], **Yu Chen**[2], **Heng Ji**[1], **Smaranda Muresan**[3], **Dequan Zheng**[2]
[1]Computer Science Department and Linguistics Department,
Queens College and Graduate Center, City University of New York, U.S.A
[2]School of Computer Science and Technology, Harbin Institute of Technology, China
[3]School of Communication and Information, Rutgers University, U.S.A
{haoli.qc,hengjicuny}@gmail.com
{chenyu,dqzheng}@mtlab.hit.edu.cn
smuresan@rci.rutgers.edu

Abstract

With the rapid development of social media and social networks, spontaneously user generated content like tweets and forum posts have become important materials for tracking people's opinions and sentiments online. In this paper we investigate the limitations of traditional linguistic-based approaches to sentiment analysis when applied to these informal genres. Inspired by various social cognitive theories, we combine local linguistic features and global social evidence in a propagation scheme to improve sentiment analysis results. Without using any additional labeled data, this new approach obtains significant improvement (up to 12% higher accuracy) for various genres in the domain of presidential election.

1 Introduction

Sentiment analysis is an important step for both Natural Language Processing (NLP) tasks such as opinion question answering (Yu and Hatzivassiloglou, 2003) and practical applications such as commercial product reputation mining (Morinaga et al., 2002), movie review mining (Pang et al., 2002) and political election prediction (Tumasjan et al., 2010).

With the prevalence of social media, spontaneously user generated content such as tweets or forum posts have become an invaluable source of people's sentiments and opinions. However, as with other NLP tasks, sentiment analysis on such informal genres presents several challenges: (1) informal text expressions; (2) lexical diversity (e.g., for example, in our training data only 10% of words in the discussion forums and tweets appear more than ten times, while in movie reviews over 20% of words appear more than ten times); (3) unpredictable shift in topics/issues. The prevalence of debate in both forum posts and tweets leads to the use of more complicated discourse structures involving multiple targets and sentiments, as well as the second-person voice. These difficulties are magnified in tweets due to necessarily compressed contexts (tweets are limited to 140 characters).

In this paper, we tackle these challenges from two perspectives. First, we approach the sentiment analysis task by identifying not only a specific "target" (e.g., presidential candidate) but also its associated "issues" (e.g., foreign policy) before detecting sentiment. This approach is similar to the idea of modeling "aspect" in product reviews (Titov and McDonald, 2008; Wang et al., 2011).

Second, a detailed error analysis has shown that currently available sentiment lexicons and various shallow linguistic features are not sufficient to advance simple bag-of-words baseline approaches due to the diverse ways in which sentiment can be expressed as well as the prevalence of debate in social media. Fortunately, documents in informal genres are often embedded in very rich social structures. Therefore, augmenting the context available for a target and an issue based on social structures is likely to provide a much richer context. We propose three hypotheses based on social cognitive theories and incorporate these hypotheses into a new framework of propagating consistent sentiments across documents. Without using any additional labeled data

26th Pacific Asia Conference on Language,Information and Computation pages 127–136

this new approach obtained significant improvement (up to 12% higher accuracy).

2 Related Work

Most sentiment analysis has been applied to movie/product reviews, blogs and tweets. Very little work has been conducted on discussion forums. Hassan et al. (2010) identified the attitudes of participants toward one another in an online discussion forum using a signed network representation of participant interaction. In contrast, we are interested in discovering the opinions of participants toward a public figure in light of their stance on various political issues.

Sentiment Analysis can be categorized into target-independent and target-dependent. The target-independent work mainly focused on exploring various local linguistic features and incorporating them into supervised learning based systems (Pang and Lee, 2004; Zhao et al., 2008; Narayanan et al., 2009) or unsupervised learning based systems (Joshi et al., 2011). Recent target-dependent work has focused on automatically extracting sentiment expressions for a given target (Godbole et al., 2007; Chen et al., 2012), or incorporating target-dependent features into sentiment analysis (Liu et al., 2005; Jiang et al., 2011). In this paper we focus on the task of jointly extracting sentiment, target and issue in order to provide richer and more concrete evidence to describe and predict the attitudes of online users. This bares similarity to the idea of modeling aspect rating in product reviews (Titov and McDonald, 2008; Wang et al., 2011).

When sentiment analysis is applied to social media, feature engineering is a crucial step (Agarwal et al., 2011; Kouloumpis et al., 2011). Most previous work based solely on lexical features suffers from data sparsity. For example, Saif et al. (2012) observed that 90% of words in tweets appear less than ten times. The semantic clustering approach they have proposed (e.g. grouping "Iphone" , "Ipad" and "Itouch" into "Apple Product") can alleviate the bottleneck, but it tends to ignore the fine-grained distinctions among semantic concepts. To address the lexical diversity problem, we take advantage of the information redundancy in rich social network structures. Unlike most previous work which only exploited user-user relations (Speriosui et al., 2011; Conover et al., 2011) or document-document relations (Tan et al., 2011; Jiang et al., 2011), we use user-document relations derived from social cognitive theories to design global features based on the interrelations among the users, targets and issues. Guerra et al. (2011) measured the bias of social media users on a topic, and then transferred such knowledge to improve sentiment classification. In this paper, we mine similar knowledge such as the bias of social media users on target-issue pairs and target-target pairs.

3 Experimental Setup

Our task focuses on classifying user contributed content (e.g., tweets and forum posts) as "Positive" or "Negative", for the domain of political elections. Tweet messages usually contain sentiments related to specific targets (e.g., presidential candidates), while forum posts often contain both specific targets and related issues (e.g., foreign policy) because participants often debate with each other and thus need to provide concrete evidence. Therefore, we define the sentiment analysis task as *target dependent* for tweets and *target-issue dependent* for forum posts. Consequently, we automatically extract targets and issues before conducting sentiment analysis. Table 1 presents some examples labeled as "Positive" or "Negative" for each genre.

3.1 Data

The tweet data set was automatically collected by retrieving positive instances with #Obama2012 or #GOP2012 hashtags[1], and negative instances with #Obamafail or #GOPfail hashtags. Similar to Gonzalez-Ibanez et al (2011), we then filtered all tweets where the hashtags of interest were not located at the very end of the message.

The discussion forum data set was adapted from the "Election & Campaigns" board of a political forum[2], where political candidates, campaigns and elections are actively discussed. We have collected the most recent posts from March 2011 to December 2011. About 97.3% posts contain either positive or

[1] "GOP" refers to the U.S. republican party which includes presidential candidates such as Ron Paul and Mitt Romney

[2] http://www.politicalforum.com/elections-campaigns

Genre	Sentiment	Target	Issue	Example
Review	Positive	N/A	N/A	The film provides some great insight into the neurotic mindset of all comics -- even those who have reached the absolute top of the game.
	Negative	N/A	N/A	Star trek was kind of terrific once, but now it is a copy of a copy of a copy.
Tweet	Positive	Ron Paul	Foreign Policy	Ron Pauls Foreign Policy Puts War Profiteers out of Business http://t.co/VGWfqcbs #ronpaul #tcot #tlot #gop2012 #FITN
	Negative	Mitt Romney	Economics	Mitt Romney said the "economy is getting better" fool!!! \#GOPFAIL
Forum	Positive	Ron Paul	Foreign Policy	I also find it interesting that so many people ridicule Ron Paul's foreign policy yet the people that are directly affected by it, our troops, support Ron Paul more than any other GOP candidate combined and more than Obama.
	Negative	Barack Obama	Economics	Obama screwed up by not fixing the economy first and leaving health care reform for a second term.

Table 1: Sentiment Examples of Different Genres

negative sentiments as opposed to neutral, therefore we only focus on the polarity classification problem.

We also used a more traditional set for sentiment analysis — the movie review polarity data set shared by (Pang et al., 2002) — to highlight the challenges of more informal texts.

Table 2 summarizes the statistics of data sets used for each genre. All experiments in this paper are based on three-fold cross-validation.

Genre	**Positive**	**Negative**
Review	5691	5691
Tweet	2323	2323
Forum	381	381

Table 2: Statistics of Data Sets

4 Linguistic-based Approach

In this section, we present our baseline approach using only linguistic features.

4.1 Pre-processing

We have applied the tool developed by Han and Baldwin (2011) together with the following additional steps to perform normalization for informal documents (tweets and forum posts).

- Replace URLs with "@URL".
- Replace @username with "@USERNAME".
- Replace negation words with "NOT" based on the list derived from the LIWCLexicon (Pennebaker et al., 2001).
- Normalize slang words (e.g. "LOL" to "laugh out loud") (Agarwal et al., 2011).
- Spelling correction using WordNet (Fellbaum, 2005) (e.g. "cooooool" to "cool")

In addition, each document has been tokenized and annotated with Part-of-speech tags (Toutanova et al., 2003).

4.2 Target and Issue Detection

After pre-processing, the first step is to detect documents which include popular targets and issues. A popular target is an entity that users frequently discuss, such as a product (e.g. "*Iphone4*"), a person (e.g. "*Ron Paul*") or an organization (e.g. "*Red Cross*"). A popular issue is a related aspect associated with a target, such as "*display function*" or "*economic issue*".

We have applied a state-of-the-art English entity extraction system (Li et al., 2012; Ji et al., 2005) that includes name tagging and coreference resolution to detect name variants from each document (e.g. "*Ron*", "*Paul*", "*Ron Paul*" and "*RP*" are all name variations for the presidential candidate Ron Paul). In order to detect issues, we mined common keywords from the U.S. presidential election web sites. The two most frequent issues are "*Economic*" which includes 647 key phrases such as "Debt", "Deficit", "Money", "Market", "Tax" and "unemployment", and "*Foreign Policy*" which includes 27 key phrases such as "military", "isolationism", "foreign policy", "Israel", "Iran" and "China". Sentiment analysis is applied on the documents that include at least one target and one issue.

We have evaluated the target and issue detection performance and the accuracy scores obtained 99.0% and 92.0%, respectively.

4.3 Sentiment Detection

We have developed a supervised learning model based on Support Vector Machines to classify sentiment labels for each document (a post, a tweet message or a movie review document), incorporating several features such as N-grams, POS, various lexicons, punctuation, capitalization (see Table 3).

Feature	Description
N-grams	All unique unigrams, bigrams and trigrams
Part-of-Speech	Part-Of-Speech tags generated by Stanford Parser (Toutanova et al., 2003)
Gazetteer	Lexical matching based on (Joshi et al., 2011), SentiWordNet (Baccianella et al., 2010), Subjectivity Lexicon (Wiebe et al., 2004), Inquirer (Stone et al., 1966), Taboada (Taboada and Grieve, 2004), UICLexicon (Hu and Liu, 2004), LIWCLexicon (Pennebaker et al., 2001)
Word Cluster	Use synset information provided by Wordnet to expand the entries of each gazettteer; Lexical matching based on the expanded gazetteers
Punctuation	Whether the document includes any exclamation mark or question mark
Capitalization	Unique words which include all capitalized letters

Table 3: Linguistic Features Used in the Baseline System

The classification results are normalized to probability based confidence values via a sigmoid kernel function (Wu et al., 2004).

4.4 Results and Analysis

Figure 1 presents the performance of the baseline system as we add each feature category. In general, N-gram based features provide a strong baseline, and thus it is difficult for local linguistic features (e.g., POS, gazetteers, punctuation) to make significant improvement. In addition, discussion forums prove to be the most challenging among these three genres. We provide a more detailed analysis for the impact of N-gram features as well as a discussion of the "long-tail" problem prevalent for informal genres.

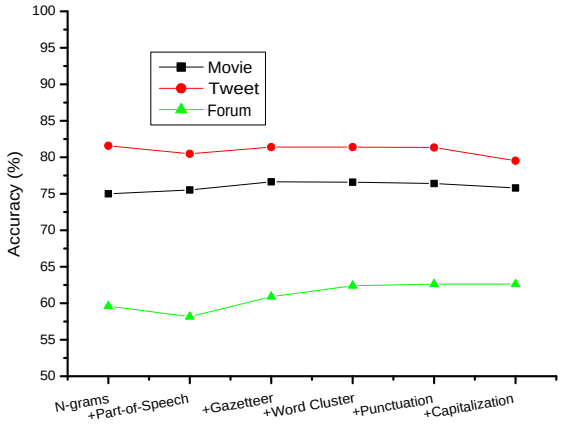

Figure 1: Baseline Performance

N-gram Features. Table 4 investigated various combinations of n-gram (n=1, 2 and 3) features. The unigram features were proven to be dominant for reviews and tweets, which is consistent with the observations by previous work on these two genres (Bermingham and Smeaton, 2010; Pak and Paroubek, 2010). However, bigram and trigram features significantly outperformed unigram features for the forum data, because forum posts tend to be longer and contain more complicated linguistic structures used to formulate arguments.

Features	**Forum**	**Tweet**	**Review**
Unigram	54.3%	81.6%	75.0%
Bigram	58.9%	79.3%	70.6%
Unigram+Bigram	58.2%	83.7%	**75.8%**
Unigram+Trigram	58.3%	**84.0%**	75.6%
Bigram+Trigram	**59.6%**	79.7%	69.7%

Table 4: Impact of N-gram Features on Accuracy

"Long-Tail" Problem. The limited gain (1%-2%) from gazetteer based features is due to long-tailed distribution of lexicon coverage. 53.3% of gazetteer entries do not cover any movie review documents, but about 87% of entries do not cover any forum posts or tweets, which clearly indicates that social media includes more diverse way to express sentiment. Similarly, 16% of entries cover 1 movie review document, but only about 6%-7% of entries cover 1 tweet message or 1 forum post; 6% of entries cover more than 10 movie review documents, but only about 0.8%-0.9% of entries cover more than 10 tweet messages or forum posts. All of the various gazetteers only cover 16.5% of movie documents, 12.4% of tweets and 17.6% of forum posts. The Word Cluster features (see Table 3) can cover more documents and achieved slight improvement (0.83% for forum posts and 0.40% for tweets) but it may require much deeper understanding and global knowledge to generalize to diverse lexical contexts.

5 Combining Linguistic Features with Global Social Evidence

The linguistic-based approach provided discouraging results. Fortunately, sentiment analysis is an inter-disciplinary task in that it attempts to capture people's social behavior. Sentiment differences within a group can result in social mitosis, leading to the emergence of two groups (Wang and Thorngate, 2003). In this section, we explore a different direction by applying social cognitive theories and propose three hypotheses that take user behavior into account in order to improve sentiment analysis.

5.1 Hypotheses based on Social Cognitive Theories

We formulate the following three hypotheses based on social cognitive theories, which we aim to prove for the domain of presidential election:

Hypothesis 1 (One sentiment per Indicative Target-Issue Pair). *The sentiment for a particular target is globally consistent across users because of the target's stance on some particular issue.*

The impression formation theory (Hamilton and Sherman, 1996) postulates a global coherence in perception, namely that users assume consistency in traits and behavior, such that observations about current behavior lead to causal attributions regarding past and future behaviors. Certain target-issue pairs are consistently associated with a particular sentiment across most users. For example, when a user is commenting on the target "*Ron Paul*" about his policy on "*Economy*" issue, the post usually indicates a positive sentiment. In contrast, the sentiments toward "*Barack Obama*"'s policy on "*Foreign Issue*" are usually negative.

Hypothesis 2 (One sentiment per Indicative Target-Target Pair). *The sentiment for a particular target is globally consistent when he or she is compared with another particular target.*

The social categorization process (Mason and Marcae, 2004) states that we mentally categorize people into different groups based on common characteristics. As a result, when commenting on an individual target, a user often compares the target with another target to express implicit sentiments or strengthen the opinions, which brings additional challenges for detecting the boundaries of sentiment words associated with specific targets. For example, the following sentence: "*NONE of the GOP candidates have a significant advantage on national polls against* ***Obama***." includes two different targets "*Obama*" and "*GOP*" and therefore a mixture of positive words (e.g. "significant" and "advantage") and negative words (e.g. "against" and "NONE"). However, some common pairs often retain consistent sentiments. For example, when compared to "*McCain*" or "*Nixon*", the sentiment towards "*Barack Obama*" is usually positive, while compared to "*Washington*", the sentiment is mostly negative.

In order to incorporate the above two hypotheses, we use a simple propagation approach. For each unique target-target pair or unique target-issue pair in the training data, we count the frequency of the sentiment labels in the training data, f_p for positive and f_n for negative. Then we adopt the following confidence metric to measure the degree of sentiment consistency for this pair:

$$c = max(f_p, f_n)/(f_p + f_n) \quad (1)$$

$Confidence$ value ranges from 0.5 to 1 and higher confidence value implies higher probability that the learned indicative pair is correct. If the $confidence$ value is larger than a threshold δ ($\delta = 0.8$ results in the best performance), we consider it as an indicative pair. Then we re-label all of the corresponding test instances which include this indicative pair with its most frequent sentiment.

Hypothesis 3 (One sentiment per User-Target-Issue during a short time). *One user's sentiment toward one target or his/her stance on one issue tends to be consistent during a short period of time.*

The social balance theory (Heider, 1946) aims to analyze the interpersonal network among social agents and see how a social group evolves to a possible balance state. Situngkir and Khanafiah (2004) extended Heider's theory to many agents. Example of possible balance states are given in Figure 2, where *"+"* means positive relations/sentiments among agents, while *"-"* means negative relations/sentiments among agents.

When applying social balance theory to our domain of presidential election, we consider the user as one agent and the two presidential candidates (tar-

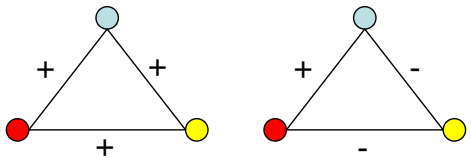

Figure 2: Social Balance Theory: Balanced States among Three People

gets) as the other two agents (see Figure 3). Since the two targets are competing in the election we assume the sentiment between them is negative; therefore, the only balanced state consists of two mutual negative and one mutual positive sentiment. In addition, a user often imposes sentiment upon a target because his or her stance on a particular political issue. The extended theory is presented in Figure 3.

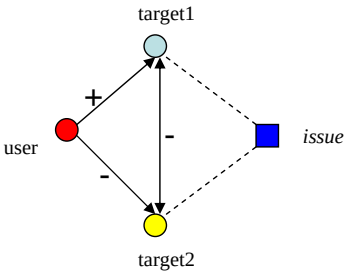

Figure 3: Balanced States for Presidential Election Domain

The Halo Effect or Halo Error theory (Thorndike, 1920) states that there exists a cognitive bias in which once we form a general impression of someone, we tend to assume that additional information will be consistent with that first impression. Abelson (1968) has proposed theories of cognitive consistency, which suggest that people will try to maintain consistency among their beliefs. Based on these social cognitive theories we have formulated Hypothesis 3. This hypothesis is valid for 90% of the training instances. The consistency of a user's sentiment regarding a target's stance on an issue is not a property of a single document, and it depends on the label for each document that mentions the target-issue pair in question. Therefore this property is not appropriately expressed as an SVM feature; instead, we incorporate Hypothesis 3 as follows: we cluster the documents authored by the same user and target (for tweets) or the same user, target, and issue (for forum posts) into one cluster. Then, within each cluster we apply one of three ways of correcting baseline results:

Approach	Accuracy
(1). Baseline	83.97%
(2). (1) + Propagating the Most Confident Sentiment	84.87%
(3). (1) + Majority Voting	84.87%
(4). (1) + Weighted Majority Voting	**85.35%**

Table 5: Impact of Hypothesis 3 on Tweets

- **Most Confident Sentiment Propagation**: within each cluster, propagate the most confident sentiment through all instances.
- **Majority Voting**: within each cluster, re-label all the instances with the sentiment that appears most often.
- **Weighted Majority Voting**: the same as Majority Voting, but use the confidence values from the baseline system for possible sentiment labels during voting.

5.2 Experiment Results

In the following we will present the performance of the enhanced approach on tweets and forum posts.

5.2.1 Impact on Tweets

The contexts of tweets are artificially compressed (each tweet message limited to 140 characters), so each single tweet message rarely includes a target-target pair or a pair target-issue pair. Therefore in this section we focus on evaluating the impact of Hypothesis 3 on tweets. The experimental results of applying Hypothesis 3 are presented in Table 5.

The results demonstrate that each voting method can provide consistent gains, with the majority voting method achieving significant gains at 99% confidence level over the baseline (using Wilcoxon Matched-Pairs Signed-Rank test). For example, the following three tweet messages about the target "*Obama*" were sent by the same user:

1. *#Obama rebuilding America using Chinese workers! http://t.co/Pk4HkvtL*
2. *But we had to rush #Obamacare thru? In the pipeline? Obama has it both ways on a controversial plan http://t.co/rb65LIx3*
3. *Small business owners confirm #Obamacare is a job killer: http://t.co/If7yNqVo*

Approach	Accuracy
Baseline	59.61%
+ Hypothesis 1	62.89%
+ Hypothesis 2	62.64%
+ Hypothesis 3	67.24%
+ Hypothesis 1+2	64.21%
+ Hypothesis 1+2+3	**71.97%**

Table 6: Impact of New Hypotheses on Forum Data

The baseline approach misclassified the first message as "Positive", but correctly classified the other two as "Negative" with high confidence. Therefore the voting approach successfully fixed the sentiment of the first message to "Negative".

5.2.2 Impact on Forum Posts

We conducted a systematic evaluation on the enhanced approach by gradually adding each hypothesis to improve sentiment analysis of the forum posts. As we have shown in Section 4, the baseline results for forum data are worse than for tweets. Applying the majority voting methods based on Hypothesis 3 to forum data would lead to compounding errors. Therefore, we only use the "most confident sentiment propagation" to incorporate Hypothesis 3. Table 6 presents the experimental results and shows that each hypothesis provides significant gain over the baseline. The overall new approach achieves up to 12.3% improvement in accuracy.

For the following post: "*If I threw you in a room with 400 corrupt politicians who each had mandates to expand government spending, I guarantee you that you could shout all you wanted for 20 years about cutting the deficit and they wouldn't hear you. Does that make Paul wrong? Does it make him a failure?*", the baseline system mistakenly labeled the sentiment for the target "*Ron Paul*" as "negative" because of the context words such as "shout", "wouldn't", "wrong" and "failure". However, based on Hypothesis 1, since in most cases the posts including the target "*Ron Paul*" and the issue "*Economics*" indicate a positive sentiment, we can correct the label successfully.

Similarly, Hypothesis 2 can correct instances when local linguistic features are misleading. For example, in the following post: "*Actually I see Newt as being more of an effective leader than Mitt with this speakership role and all, but Mitt has the business realm sealed tightly in his hip pocket, and jobs and economic progress are what we desperately need now.*", simply incorporating the context entity features from the first sub-sentence, this baseline system mistakenly labeled the sentiment on the target "*Mitt Romney*" as "negative". In addition, due to the lack of discourse features, the baseline system failed to recognize the scope of identification (the second sub-sentence). However more than 80% instances in the training data indicate that the sentiment on "*Mitt Romney*" is positive when he is compared to '*Newt*", therefore we can correct the sentiment of this post to "positive".

Hypothesis 3 can effectively exploit information redundancy and propagate the high-confidence results from posts with relatively simpler linguistic structures to those posts with more complicated structures. For example, it is difficult for the baseline system to determine the sentiment on the target "*Mitt Romney*" from the following post: "*Paul is the complete opposite of Romney. Romney has a political history that can be examined..and debated.. Paul has 22 years of voting No..but nothing else. Romney has 30 years of business experience. Paul was a doctor a long time ago.*" But the same user posted other messages that include simpler structures and therefore the baseline system can detect correct "positive" sentiment with high confidence: "*Romney saved failed business and political models. Paul merely participated.*". As a result, the sentiment analysis results of all the posts within the same cluster (posted by the same user, and including the same target and issue) can be corrected.

5.2.3 Parameter Tuning

Figure 4 shows the overall performance of our approaches when the indicative pairs are learned from training data with different thresholds set for confidence estimation given in 1. Figure 4 shows consistent performance improvement as the threshold is larger than 0.5. We also noticed that when the threshold is low (0.5), the overall approach performs a little worse than the baseline due to the propagation of erroneous results with low confidence values.

6 Remaining Challenges

Although the proposed approach based on social cognitive theories has significantly enhanced the

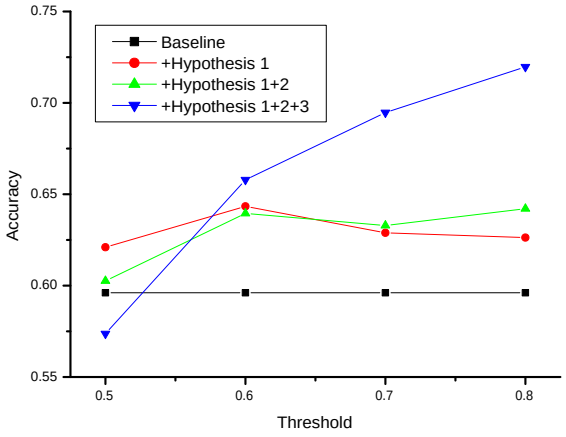

Figure 4: Impact of Parameters

performance of sentiment analysis, some challenges remain. We analyze the major sources of the remaining errors as follows.

Sarcasm Detection. For both tweets and forum posts, some remaining errors require accurate detection of sarcasm (Davidov et al., 2010; Gonzalez-Ibanez et al., 2011). For example, "*LOL..remember Obama chastising business's for going to Vegas. Vegas would have cost a third of what these locations costs. But hey, no big deal...* " contains sarcasm, which leads our system to misclassify this post.

Domain-specific Latent Sentiments. The same word or phrase might indicate completely different sentiments in various domains. For example, "*big*" usually indicates positive sentiment, but it indicates negative sentiment in the following sentence: "*tell me how the big government, big bank backing, war mongering Obama differs from Bush?*". Most of these domain-specific phrases do not exist in the currently available semantic resources and thus a system is required to conduct deep mining of such latent sentiments.

Thread Structure. A typical online forum discussion consists of a root post and the following posts which form a tree structure, or thread. Performing sentiment analysis at post level, without taking into account the thread context might lead to errors. For example, if a post disagree with another post, and the first post expresses "Positive" sentiment, we can infer that the second post should be "Negative". Identifying who replies to whom in a forum might not be straightforward (Wang et al., 2011). In addition, we would need to identify agreement/disagreement relations among posts.

Multiple Sentiments. Due to the prevalence of debate in discussion forums, the users tend to list multiple argument points to support their overall opinions. As a result, a single post often contains a mixture of sentiments. For example, the following post indicates "Positive" sentiment although it includes negative words such as "disagreement": "*....As a huge Ron Paul fan I have my disagreements with him.....but even if you disagree with his foreign policy.....the guy is spot on with everything and anything else.....*". This requires a sentiment analyzer to go beyond lexical level analysis and conduct global logic inferences. This is not a challenge in social media genres that impose stringent length restrictions such as Twitter.

Figure 5 summarizes the distributions of the remaining errors for tweets and forum posts.

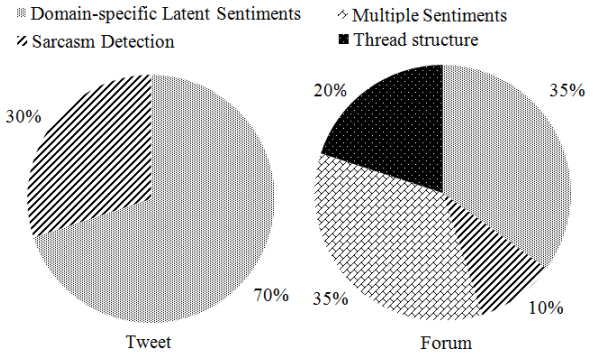

Figure 5: Remaining Challenges

7 Conclusion and Future Work

We have presented a novel approach to social cognitive theories to enhance sentiment analysis for user generated content in social media. We have investigated the limitations of approaches based solely on shallow linguistic features. We have introduced three hypotheses that incorporate global consistency within the rich social structures consisting of users, targets and associated issues, and have shown that using such social evidence improve the results of sentiment analysis on informal genres such as tweets and forum posts.

In the future, we aim to address the remaining challenges discussed in Section 6, especially

to exploit the implicit global contexts by analyzing thread structures and discovering cross-post agreement/disagreement relations.

Acknowledgements

This work was supported by the U.S. ARL grant W911NF-09-2-0053, the U.S. NSF Grants IIS-0953149 and IIS-1144111 and the U.S. DARPA BOLT program. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmen-t purposes notwithstanding any copyright notation here on.

References

Robert Adelson. 1968. *Theories of Cognitive Consistency Theory*. Rand McNally.

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the 49th Annual Meeting of Association for Computational Linguistics Workshop on Languages in Social Media*.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 6th international conference on Language Resources and Evaluation*.

Adam Bermingham and Alan Smeaton. 2010. Classifying sentiment in microblogs: is brevity an advantage is brevity an advantage? In *Proceedings of the 19th ACM Conference on Information and Knowledge Management*.

Lu Chen, Wenbo Wang, Meenakshi Nagarajan, Shaojun Wang, and Amit P. Sheth. 2012. Extracting diverse sentiment expressions with target-dependent polarity from twitter. In *Proceedings of the 6th International Conference on Weblogs and Social Media*.

Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. In *Proceedings of the 5th International Conference on Weblogs and Social Media*. The AAAI Press.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Christiane Fellbaum. 2005. Wordnet and wordnets. *Encyclopedia of Language and Linguistics*.

Namrata Godbole, Manjunath Srinivasaiah, and Steven Skiena. 2007. Large-scale sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media*.

Roberto Gonzalez-Ibanez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: A closer look. In *Proceedings of the 49th Annual Meeting of Association for Computational Linguistics*.

Pedro Henrique Calais Guerra, Adriano Veloso, Wagner Meira Jr., and Virgilio Almeida. 2011. From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

David L. Hamilton and Steven J. Sherman. 1996. Perceiving persons and groups. *Psychological Review*.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of Association for Computational Linguistics*.

Ahmed Hassan, Vahed Qazvinian, and Dragomir R. 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*.

F. Heider. 1946. Attitudes and cognitive organization. *Journal of Psychology*, pages (21):107–112.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Heng Ji, David Westbrook, and Ralph Grishman. 2005. Using semantic relations to refine coreference decisions. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of Association for Computational Linguistics*.

Aditya Joshi, Balamurali A. R., Pushpak Bhattacharyya, and Rajat Kumar Mohanty. 2011. C-feel-it: A sentiment analyzer for micro-blogs. In *Proceedings of the 49th Annual Meeting of Association for Computational Linguistics (Demo)*.

E. Kouloumpis, T. Wilson, and J. Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the 5th International Conference on Weblogs and Social Media*.

Qi Li, Haibo Li, Heng Ji, Wen Wang, Jing Zheng, and Fei Huang. 2012. Joint bilingual name tagging for paral-

lel corpora. In *Proceedings of the 21th ACM Conference on Information and Knowledge Management*.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*.

Malia F. Mason and C. Neil Marcae. 2004. Catagorizing and individuating others: The neural substrates of person perception. *Journal of Cognitive Neuroscience*.

Satoshi Morinaga, Kenji Yamanishi, Kenji Tateishi, and Toshikazu Fukushima. 2002. Mining product reputations on the web. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

R. Narayanan, B. Liu, and A. Choudhary. 2009. Sentiment analysis of conditional sentences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the 6th international conference on Language Resources and Evaluation*.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42th Annual Meeting of Association for Computational Linguistics*.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *CoRR*, volume cs.CL/0205070.

J.W. Pennebaker, M.E. Francis, and R.J. Booth. 2001. Linguistic inquiry and word count: Liwc2001. In *http://www.liwc.net/*.

Hassan Saif, Yulan He, and Harith Alani. 2012. Alleviating data sparsity for twitter sentiment analysis. In *The 2nd Workshop on Making Sense of Microposts*.

Hokky Situngkir and Deni Khanafiah. 2004. Social balance theory: Revisiting heider's balance theory for many agents. *Technical Report*.

Michael Speriosui, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. 2011. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.

Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

Maite Taboada and Jack Grieve. 2004. Analyzing appraisal automatically. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*.

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 2011 Conference on Empirical Methods in Natural Language Processing*.

E. L. Thorndike. 1920. A constant error in psychological ratings. *Journal of Applied Psychology*, pages 4(1):25–29.

Ivan Titov and Ryan T. McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Andranik Tumasjan, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the 4th International Conference on Weblogs and Social Media*.

Zhigang Wang and Warrant Thorngate. 2003. Sentiment and social mitosis: Implications of heider's balance theory. *Journal of Artificial Societies and Social Simulation*.

Hongning Wang, Chi Wang, ChengXiang Zhai, and Jiawei Han. 2011. Learning online discussion structures by conditional random fields. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. In *Computional Linguistics*.

Ting-Fan Wu, Chih-Jen Lin, and Ruby C. Weng. 2004. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*.

J. Zhao, K. Liu, and G Wang. 2008. Adding redundant features for crfs-based sentence sentiment classification. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*.