Do Users Rate or Review? Boost Phrase-level Sentiment Labeling with Review-level Sentiment Classification[*]

Yongfeng Zhang, Haochen Zhang, Min Zhang, Yiqun Liu, Shaoping Ma
State Key Laboratory of Intelligent Technology and Systems
Department of Computer Science & Technology, Tsinghua University, Beijing, 100084, China
{zhangyf07,rukyzhc}@gmail.com, {z-m,yiqunliu,msp}@tsinghua.edu.cn

ABSTRACT

Current approaches for contextual sentiment lexicon construction in phrase-level sentiment analysis assume that the numerical star rating of a review represents the overall sentiment orientation of the review text. Although widely adopted, we find through user rating analysis that this is not necessarily true. In this paper, we attempt to bridge the gap between phrase-level and review/document-level sentiment analysis by leveraging the results given by review-level sentiment classification to boost phrase-level sentiment polarity labeling in contextual sentiment lexicon construction tasks, using a novel constrained convex optimization framework. Experimental results on both English and Chinese reviews show that our framework improves the precision of sentiment polarity labeling by up to 5.6%, which is a significant improvement from current approaches.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval - *Classification*

Keywords

Sentiment Analysis; Sentiment Classification; Sentiment Lexicon Construction; Optimization

1. INTRODUCTION

The construction of a sentiment lexicon is of key importance in phrase-level sentiment analysis [7] and many other tasks such as recommender systems [10], where each entry in the lexicon is a Feature-Opinion (F-O) word pair together with the corresponding Sentiment polarity (S), represented by (F,O,S) [5]. For example, the entries (*service, excellent, positive*) and (*phone quality, perfect, positive*) could be extracted from the textual review of Figure 1.

However, current phrase-level sentiment lexicon construction approaches may only give sentiment polarity labeling (assigning the S for an F-O pair) precisions of around 70% $\sim$ 80% [4]. We find through large-scale user behavior analysis that one of the basic assumptions in current approaches, i.e., the overall numerical rating of a review represents the overall sentiment of the review text, is not necessarily true.

To avoid the biased assumption, we propose to boost the performance of phrase-level sentiment polarity labeling in a reverse way, which is to use unsupervised review-level sentiment classification results instead of the numerical ratings as a heuristic for phrase-level polarity labeling. State-of-the-art review-level sentiment classification techniques, even the unsupervised approaches, can give pretty good precisions of above 90% [9, 6], which could be reliable to boost the performance of phrase-level sentiment polarity labeling.

In general the framework is two-stage. In the first stage, the overall sentiment orientations of the product reviews are labeled using a review-level sentiment classifier. In the second stage, we extract feature-opinion pairs from the corpus [5, 8], then use the overall sentiment orientations of the reviews as constraints to learn the sentiment polarities of these pairs automatically, using a novel optimization framework.

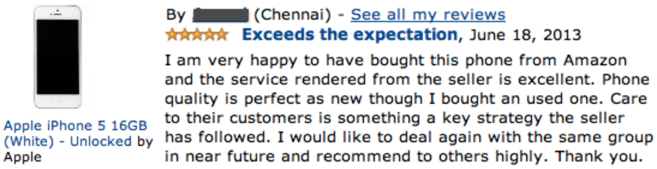
Experimental results on both English and Chinese review datasets show that our framework improves the precision of phrase-level sentiment polarity labeling significantly, which means that the original assumption might be infeasible, and that it might be promising to leverage sentence- or review-level sentiment analysis techniques to boost the performance of phrase-level sentiment analysis tasks.

2. THE FRAMEWORK

The first stage of the framework determines the overall sentiment of each piece of review by conducting review-level sentiment classification, and the second stage leverages the results for sentiment lexicon construction. We use $\mathbf{x} = [x_1, x_2]^T (x_i \geq 0)$ to represent a sentiment vector, where x_1 and x_2 are the *positive* and *negative* degrees, respectively, and use $\mathbf{X} = [\mathbf{x}_1 \mathbf{x}_2 \cdots \mathbf{x}_m]^T$ as the sentiment matrix for a set of m reviews or feature-opinion pairs.

2.1 Review-Level Sentiment Classification

Two possible sentiment vector candidates are used in this stage. If a review is classified as *positive* by a sentiment classification algorithm, then its sentient vector is assigned



Apple iPhone 5 16GB (White) - Unlocked by Apple

By [redacted] (Chennai) - See all my reviews
★★★★★ **Exceeds the expectation**, June 18, 2013
I am very happy to have bought this phone from Amazon and the service rendered from the seller is excellent. Phone quality is perfect as new though I bought an used one. Care to their customers is something a key strategy the seller has followed. I would like to deal again with the same group in near future and recommend to others highly. Thank you.

Figure 1: A sample user review from Amazon.com

[*] This work was supported by Natural Science Foundation (60903107, 61073071) and National High Technology Research and Development (863) Program (2011AA01A205).

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.
Copyright 2014 ACM 978-1-4503-2257-7/14/07 ...$15.00.
http://dx.doi.org/10.1145/2600428.2609501.

as $\mathbf{x} = [1, 0]^T$, otherwise, the corresponding sentiment vector is $\mathbf{x} = [0, 1]^T$. Based on the classification results, a sentiment matrix $\tilde{\mathbf{X}} = [\mathbf{x}_1 \mathbf{x}_2 \cdots \mathbf{x}_m]^T$ is constructed, which will be used as a constraint in the next stage.

We use the sentence orientation prediction approach in [1] for English reviews, and the automatic seed word selection scheme in [9] for Chinese reviews. Both of them are state-of-the-art approaches on the corresponding language.

2.2 Sentiment Lexicon Construction

We consider four kinds of constraints to learn the sentiment lexicon $\mathbf{X}$: 1) Review-level sentiment orientation, 2) General sentiment lexicon, 3) Linguistic heuristics, and 4) Sentential sentiment consistency.

1) Review-level Sentiment Orientation captures the overall sentiment of a review given by the review-level sentiment classification algorithm in the previous stage. We construct a matrix $\mathbf{A}$ to indicate the frequency of each F-O pair in each review: $\mathbf{A}_{ij} = I^{neg}_{ij} \cdot \frac{\text{Freq}(i,j)}{\sum_k \text{Freq}(i,k)}$, where $\text{Freq}(i, j)$ is the frequency of F-O pair j in review i. The matrix I^{neg} is an indication matrix that allows us to take the "negation rules" into consideration. $I^{neg}_{ij} = -1$ if the F-O pair j is modified by a negation word, e.g. "no", "not", "hardly", etc. Otherwise, $I^{neg}_{ij} = 1$.

The sentiments of all the F-O pairs are aggregated to approximate the review-level sentiment polarity, which gives the following objective function: $\mathcal{R}_1 = \|\mathbf{AX} - \tilde{\mathbf{X}}\|^2_F$.

2) General Sentiment Lexicon captures the sentiment of some context-irrelevant opinion words, like *excellent, good* and *bad*. We construct the general sentiment lexicon $\mathbf{X}_0$ by labeling the polarities of the F-O pairs in $\mathbf{X}$ according to the public sentiment corpora MPQA[1] on English, and HowNet[2] on Chinese. An F-O pair is labeled as $[1, 0]^T$ or $[0, 1]^T$ if the opinion word is included in the positive or negative word set, correspondingly. Otherwise, we use $[0, 0]^T$.

We expect the sentiment polarities of the context-irrelevant words in $\mathbf{X}$ to be close to those in the general sentiment lexicon $\mathbf{X}_0$, which corresponds to the objective function $\mathcal{R}_2 = \|\mathbf{G}(\mathbf{X} - \mathbf{X}_0)\|^2_F$, where $\mathbf{G}$ is a diagonal matrix indicating which F-O pairs in $\mathbf{X}$ are "fixed" by the general sentiment lexicon $\mathbf{X}_0$. Namely, $\mathbf{G}_{ii} = 1$ if the i-th F-O pair has a fixed sentiment, and $\mathbf{G}_{ii} = 0$ otherwise.

3) Linguistic Heuristic captures the linguistic "and" and "but" relationship. It is intuitional that those F-O pairs frequently concatenated with "and" might have similar sentiments, while those frequently connected by "but" tend to have opposite sentiments. To formalize the intuition, we define two $n \times n$ matrices $\mathbf{W}^a$ and $\mathbf{W}^b$ for the "and" and "but" linguistic heuristics, respectively. We set $\mathbf{W}^a_{ij} = \mathbf{W}^a_{ji} = 1$ or $\mathbf{W}^b_{ij} = \mathbf{W}^b_{ji} = 1$ if pair i and$j$ are concatenated by "and" or "but" for a minimal number of times, correspondingly, otherwise, we set $\mathbf{W}^a_{ij} = \mathbf{W}^a_{ji} = 0$. The objective function regarding both "and" and "but" linguistic heuristic is:

$$\begin{aligned}\mathcal{R}_3 &= \text{tr}(\mathbf{X}^T\mathbf{D}^a\mathbf{X} - \mathbf{X}^T\mathbf{W}^a\mathbf{X}) + \text{tr}(\mathbf{X}^T\mathbf{D}^b\mathbf{X} - \mathbf{X}^T\mathbf{W}^b\mathbf{XE}) \\ &= \text{tr}(\mathbf{X}^T\mathbf{DX} - \mathbf{X}^T\mathbf{W}^a\mathbf{X} - \mathbf{X}^T\mathbf{W}^b\mathbf{XE})\end{aligned}$$

where $\text{tr}(\cdot)$ is the trace of a matrix, $\mathbf{D}^a, \mathbf{D}^b \in \mathbb{R}^{n\times n}$ are diagonal matrices where $\mathbf{D}^a_{ii} = \sum_{j=1}^n \mathbf{W}^a_{ij}$, and $\mathbf{D}^b_{ii} = \sum_{j=1}^n \mathbf{W}^b_{ij}$. $\mathbf{E} = \left[\begin{smallmatrix}0 & 1\\ 1 & 0\end{smallmatrix}\right]$ is an anti-diagonal matrix that serves as a column permutation function to reverse the columns of $\mathbf{X}$, and $\mathbf{D} = \mathbf{D}^a + \mathbf{D}^b$. The underlying intuition is that the sentiment vectors of two pairs should be similar if they are frequently linked by "and" and opposite if by "but", or a penalty would be introduced to the loss function.

4) Sentential Sentiment Consistency captures the sentiment consistency in sentences [3], i.e., similar opinion orientations are usually expressed in consecutive sentences.

To formalize the heuristic, a sentential similarity matrix $\mathbf{W}^s \in \mathbb{R}^{n\times n}$ is introduced, which leverages the sentential distance between F-O pairs in corpus to estimate their sentential similarities. For example, consider two pairs i and j, if they co-occur in the same piece of review in the corpus, then we calculate their sentential similarity in this review, and the final similarity between i and j is the average of all their intra-review similarities. More formally, suppose pair i and pair j co-occur in the same review for N_{ij} times, and the k-th co-occurrence happens in review t_{i_k}, then $\mathbf{W}^s_{ij}$ and $\mathbf{W}^s_{ji}$ are defined as:

$$\mathbf{W}^s_{ij} = \mathbf{W}^s_{ji} = \begin{cases} 0, \ \mathit{if} \ N_{ij} = 0 \text{ or } \mathbf{W}^a_{ij} \neq 0 \text{ or } \mathbf{W}^b_{ij} \neq 0 \\ \frac{1}{N_{ij}} \sum\limits_{k=1}^{N_{ij}} \left(1 - \frac{\mathit{dist}(i,j)}{\mathit{length}(r_{i_k})}\right), \ \mathit{else} \end{cases}$$

where the length of a review $\mathit{length}(r_{i_k})$ is the number of words (punctuations excluded) in the review, and the distance $\mathit{dist}(i, j)$ of pair i and j in the review is the number of words between the two feature words of the pair. The corresponding objective function is $\mathcal{R}_4 = \text{tr}(\mathbf{X}^T\mathbf{D}^s\mathbf{X} - \mathbf{X}^T\mathbf{W}^s\mathbf{X})$, where $\mathbf{D}^s$ is also a diagonal matrix, and $\mathbf{D}^s_{ii} = \sum_{j=1}^n \mathbf{W}^s_{ij}$.

2.3 The Unified Model for Polarity Labeling

With the above constraints from different information sources and aspects, we adopt the following objective function to learn the contextual sentiment lexicon $\mathbf{X}$:

$$\begin{aligned}\min_{\mathbf{X}\geq 0} \mathcal{R} = \ &\lambda_1\|\mathbf{AX} - \tilde{\mathbf{X}}\|^2_F + \lambda_2\|\mathbf{G}(\mathbf{X} - \mathbf{X}_0)\|^2_F \\ &+ \lambda_3 \, \text{tr}(\mathbf{X}^T\mathbf{DX} - \mathbf{X}^T\mathbf{W}^a\mathbf{X} - \mathbf{X}^T\mathbf{W}^b\mathbf{XE}) \\ &+ \lambda_4 \, \text{tr}(\mathbf{X}^T\mathbf{D}^s\mathbf{X} - \mathbf{X}^T\mathbf{W}^s\mathbf{X})\end{aligned} \tag{1}$$

where $\lambda_1, \lambda_2, \lambda_3$ and λ_4 are positive weighing parameters that control the contributions of each information source in the learning process. An important property of the objective function (1) is its convexity, which makes it possible to search for the global optimal solution $\mathbf{X}^*$. We give the updating rule for learning $\mathbf{X}^*$ directly here, as shown in (2). The proof of the updating rule as well as its convergence is similar to the KKT condition approach in [2].

$$\mathbf{X}_{ij} \leftarrow \mathbf{X}_{ij}\sqrt{\frac{[\lambda_1\mathbf{A}^T\tilde{\mathbf{X}} + \lambda_2\mathbf{GX}_0 + \lambda_3\mathbf{W}^a\mathbf{X} + \lambda_3\mathbf{W}^b\mathbf{XE} + \lambda_4\mathbf{W}^s\mathbf{X}]_{ij}}{[\lambda_1\mathbf{A}^T\mathbf{AX} + \lambda_2\mathbf{GX} + \lambda_3\mathbf{DX} + \lambda_4\mathbf{D}^s\mathbf{X}]_{ij}}} \tag{2}$$

In this work, we choose the function $s(\mathbf{x}_i) = x_{i1} - x_{i2}$ to calculate the final sentiment polarity. Pair i is labeled as *positive* if $s(\mathbf{x}_i) \geq 0$, and *negative* if $s(\mathbf{x}_i) < 0$.

3. EXPERIMENTS

We use the MP3 player reviews crawled from Amazon for the experiment on English, which is publicly available[3]. For the Chinese language, we use the restaurant reviews crawled from Dianping[4], which is a famous restaurant rating website

[1] http://mpqa.cs.pitt.edu/corpora/

[2] http://www.keenage.com/

[3] http://sifaka.cs.uiuc.edu/~wang296/Data/

[4] http://www.dianping.com/

in China. Each of the reviews of the two datasets consists of a piece of review text and an overall numerical rating raging from 1 to 5 stars. Some statistical information about these two datasets is shown in Table 1.

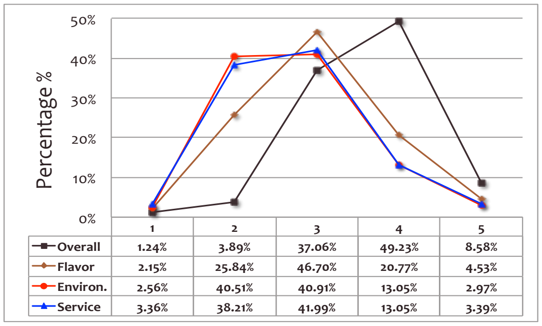
Table 1: Statistics of the two datasets

	#Users	#Items	#Reviews
MP3 Player	26,113	796	55,740
Restaurant	11,857	89,462	510,551

An important property of our restaurant review dataset is that, each review is accompanied with three sub-aspect ratings except for the overall rating. They are users' ratings made on the *flavour*, *environment* and *service* of restaurants, respectively, which makes it possible for us to conduct much detailed user rating analysis on this dataset. The range of the sub-aspect ratings are also from 1 to 5.

3.1 User Rating Analysis

The ratings on three sub-aspects allow us to investigate a user's "true" feelings on more specific aspects of a restaurant beyond the overall rating. For the overall rating and each sub-aspect rating, we calculate the percentage that each of the 5 star ratings takes in the total number of ratings, shown in Figure 2. The x-axis represents 1 star through 5 stars, and the y-axis is the percentage of each star rating.



	1	2	3	4	5
Overall	1.24%	3.89%	37.06%	49.23%	8.58%
Flavor	2.15%	25.84%	46.70%	20.77%	4.53%
Environ.	2.56%	40.51%	40.91%	13.05%	2.97%
Service	3.36%	38.21%	41.99%	13.05%	3.39%

Figure 2: Percentage of each star of overall rating, flavour, environment and service.

We see that user ratings tend to center around 4 stars on overall rating, while they tend to center around 2∼3 stars on the sub-aspect ratings. This implies that the overall rating might not serve as a real reflection of the users' feelings, and users tend to "tell the truth" in much detailed sub-aspects. In order to examine the statistical significance, we calculate the average rating μ and coefficient of variation $c_v = \sigma/\mu$ for the overall rating and the three sub-aspect ratings, where σ is the standard deviation. Table 2 shows the results. We see that users tend to give higher scores on overall rating, and the scores on overall rating are more concentrated.

More intuitionally, we conduct per user analysis. For each user and each kind of rating (overall, flavour, environment and service), we calculate the percentage of 4 or 5 stars that the user made. Then we sort these percentages of the users in descending order, which is shown in Figure 3.

It is clear that user rating behaviours on overall and sub-aspect ratings are different. More than a half of the users made 50% or more 4+ ratings in terms of overall rating, while less than 5% users did so on sub-aspect ratings.

Table 2: Average ratings and coefficient of variation

	Overall	Flavour	Environment	Service
μ	3.6432	3.1547	2.8934	2.8510
c_v	0.1977	0.2522	0.2697	0.2816

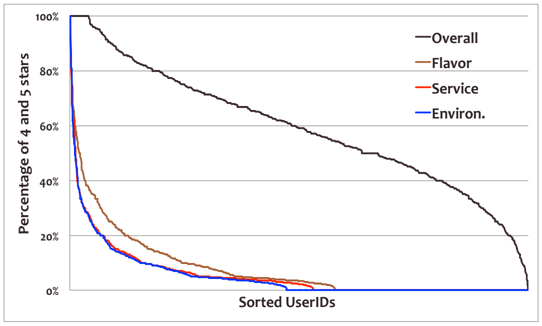


Figure 3: Percentage of ≥ 4 stars made by each user on each kind of rating, sorted in descending order of percentages.

This analysis partly shows that it might not be appropriate to use overall ratings as groundtruth to label the sentiment orientations of review texts, as users tend to act differently when making overall ratings and expressing their true feelings on detailed product features/aspects.

3.2 Phrase-Level Polarity Labeling

We choose the frequently used measures *precision, recall* and *F-measure* to evaluate the performance of polarity labeling, and experiment with the following methods:

- **General**: Predict by querying the polarity of the opinion word in general sentiment opinion word sets. Also, we use MPQA for English and HowNet for Chinese.
- **Optimize**: The optimization approach in [5], which reduces the problem of polarity labeling to the problem of constrained linear programming.
- **Overall**: Use our framework except that the review-level sentiment orientation is determined using the corresponding overall rating.
- **Subaspect**: Use our framework except that sentiment orientations of reviews are determined by averaging the corresponding sub-aspect ratings.
- **Boost**: Use our complete framework, where unsupervised sentiment classification is conducted on reviews to boost phrase-level sentiment polarity labeling.

We use $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 1$ in this experiment, and the results on the two datasets are shown in Table 3. We did not perform the "Subaspect" method on mp3 player reviews as the sub-aspect ratings are absent.

We see that labeling the polarities by querying the general opinion word sets gives the best precision on both of the two datasets. However, the recall of this method is rather low. This implies that there are many "contextual dependent" opinion words which are absent from these word sets.

The "Optimize" method and our "Overall" method are similar in that both of them leverage overall numerical ratings as the groundtruth of review-level sentiment orientations. Though the Optimize method achieves slightly better recall, their overall performance are comparable. Further more, by taking advantage of the sub-aspect ratings in the "Subaspect" method, both precision and recall are improved from "Optimize" and "Overall" methods, which implies that the detailed sub-aspect ratings could be more reliable.

Finally, our "Boost" method achieves the best performance in terms of recall and F-measure, on both of the two datasets.

Table 3: Performance of polarity labeling

	Precision	Recall	F-measure
MP3 Player Data			
General	**0.9238**	0.4201	0.5776
Optimize	0.8269	0.7626	0.7934
Overall	0.8288	0.7525	0.7888
Boost	0.8504	**0.7683**	**0.8073**
Restaurant Review			
General	**0.9017**	0.3571	0.5115
Optimize	0.8405	0.7760	0.8069
Overall	0.8473	0.7468	0.7938
Subaspect	0.8675	0.7561	0.8079
Boost	0.8879	**0.7818**	**0.8315**

Besides, it also achieves the best precision without regard to the "General" method. This further verifies the effect of leveraging review-level sentiment classification in boosting the process of phrase-level polarity labeling.

3.3 Parameter Analysis

In this subsection, we attempt to study the effect of different constraints in our framework by analyzing the four main parameters $\lambda_1 \sim \lambda_4$ in objective function (1).

We first conduct "Knock Out One Term" experiment on these parameters, to see whether all these constraints contribute to the performance of phrase-level polarity labeling. We set one of the four parameters to 0 at a time, and evaluate the F-measure. The results are shown in Table 4.

The experimental result shows that knocking out any of the four parameters decreases the performance of polarity labeling. Besides, removing the constraint on review-level sentiment orientation (λ_1) or general sentiment lexicon (λ_2) decreases the performance to a great extent, which implies that these two information sources are of great importance in constructing the sentiment lexicon.

We further investigate the effect of different constraints by fixing three parameters to 1 and weighing the remaining parameter. The results on restaurant are shown in Figure 4, and the observations on mp3 player dataset are similar.

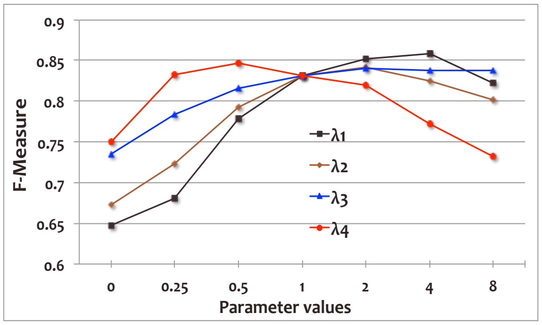


Figure 4: Tune one of the four parameters.

The experimental result shows that giving more weights to the constraints of review-level sentiment orientation and general sentiment lexicon could further improve the performance, which means that these two information sources might be more reliable. However, weighting the constraint on sentential sentiment consistency too much would decrease the performance, this implies that noise could be introduced by this heuristic and it is not as reliable as the linguistic heuristic of "and" and "but".

Table 4: F-measure by knocking out one constraint

	λ_1	λ_2	λ_3	λ_4	MP3 Player	Restaurant
Default	1	1	1	1	0.8073	0.8315
Knock	0	1	1	1	0.6783	0.6476
out	1	0	1	1	0.6332	0.6728
one	1	1	0	1	0.7461	0.7352
term	1	1	1	0	0.7756	0.7504

We tuned the parameters carefully to get the optimal performance. Finally, the optimal result on mp3 player dataset was achieved when using the parameters (4, 2, 1, 0.25), with an F-measure of 0.8237, and on restaurant review dataset (3, 2, 2, 0.5) is used, which gives the F-measure of 0.8584.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we investigated the inconsistency between overall numerical ratings and the sentiment orientations of textual user reviews in real-world datasets, which is an unvalidated assumption but frequently used in previous work. We propose to leverage review-level sentiment classification techniques to boost the performance of phase-level sentiment polarity labeling. Besides, we formalize the phrase-level sentiment polarity labeling problem in a simple convex optimization framework, and designed iterative updating algorithms for model learning. Experimental results on both English and Chinese datasets show that our framework helps to improve the performance in contextual sentiment lexicon construction tasks.

This work is a first step towards bridging the gap between phrase-level and sentence/review-level sentiment analysis. Except for the four kinds of heuristics investigated in this paper, the framework can also integrate various other information sources. Besides, review-level analysis could also be promising to help extract feature or opinion words in phrase-level analysis, except for the polarity labeling task in this work. Additional insights about the bidirectional relationship of phrase- and review-level analysis may also yield more effective heuristics and algorithms for both tasks.

5. REFERENCES

[1] M. Hu and B. Liu. Mining and Summarizing Customer Reviews. *KDD*, pages 168–177, 2004.

[2] X. Hu, J. Tang, H. Gao, and H. Liu. Unsupervised Sentiment Analysis with Emotional Signals. *WWW*, 2013.

[3] H. Kanayama and T. Nasukawa. Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis. *EMNLP*, pages 355–363, 2006.

[4] B. Liu and L. Zhang. A Survey of Opinion Mining and Sentiment Analysis. *Jour. Mining Text Data*, 2012.

[5] Y. Lu, M. Castellanos, U. Dayal, and C. Zhai. Automatic Construction of a Context-Aware Sentiment Lexicon: An Optimization Approach. *WWW*, 2011.

[6] L. Qiu, W. Zhang, C. Hu, and K. Zhao. Selc: A self supervised model for sentiment classification. *CIKM*, 2009.

[7] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguastics*, 37(2), 2011.

[8] Y. Tan, Y. Zhang, M. Zhang, Y. Liu, and S. Ma. A Unified Framework for Emotional Elements Extraction based on Finite State Matching Machine. *NLPCC*, 400:60–71, 2013.

[9] T. Zagibalov and J. Carroll. Automatic Seed Word Selection for Unsupervised Sentiment Classification of Chinese Text. *Coling*, pages 1073–1080, 2008.

[10] Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, and S. Ma. Explicit Factor Models for Explainable Recommendation based on Phrase-level Sentiment Analysis. *SIGIR*, 2014.