Capturing the stars: predicting ratings for service and product reviews

Narendra Gupta, Giuseppe Di Fabbrizio and **Patrick Haffner**
AT&T Labs - Research, Inc.
Florham Park, NJ 07932 - USA
{ngupta,pino,haffner}@research.att.com

Abstract

Bloggers, professional reviewers, and consumers continuously create opinion–rich web reviews about products and services, with the result that textual reviews are now abundant on the web and often convey a useful overall rating (number of stars). However, an overall rating cannot express the multiple or conflicting opinions that might be contained in the text, or explicitly rate the different aspects of the evaluated entity. This work addresses the task of automatically predicting ratings, for given aspects of a textual review, by assigning a numerical score to each evaluated aspect in the reviews. We handle this task as both a regression and a classification modeling problem and explore several combinations of syntactic and semantic features. Our results suggest that classification techniques perform better than ranking modeling when handling evaluative text.

1 Introduction

An abundance of service and products reviews are today available on the Web. Bloggers, professional reviewers, and consumers continuously contribute to this rich content both by providing text reviews and often by assigning useful overall ratings (number of stars) to their overall experience. However, the overall rating that usually accompanies online reviews cannot express the multiple or conflicting opinions that might be contained in the text, or explicitly rate the different aspects of the evaluated entity. For example, a restaurant might receive an overall great evaluation, while the service might be rated below average due to slow and discourteous wait staff. Pinpointing opinions in documents, and the entities being referenced, would provide a finer–grained sentiment analysis and a solid foundation to automatically summarize evaluative text, but such a task becomes even more challenging when applied to a generic domain and with unsupervised methods. Some significant contributions by Hu and Liu (2004), Popescu and Etzioni (2005), and Carenini et al. (2006) illustrate different techniques to find and measure opinion orientation in text documents. Other work in sentiment analysis (often referred as *opinion mining*) has explored several facets of the problem, ranging from predicting binary ratings (e.g., *thumbs up/down*) (Turney, 2002; Pang et al., 2002; Dave et al., 2003; Yu and Hatzivassiloglou, 2003; Pang and Lee, 2004; Yi and Niblack, 2005; Carenini et al., 2006), to more detailed opinion analysis methods predicting multi–scale ratings (e.g., *number of stars*) (Pang and Lee, 2005; Snyder and Barzilay, 2007; Shimada and Endo, 2008; Okanohara and Tsujii, 2005).

This paper focuses on multi–scale multi–aspect rating prediction for textual reviews. As mentioned before, textual reviews are abundant, but when trying to make a buy decision on a specific product or service, getting sufficient and reliable information can be a daunting and time consuming task. On one hand, a single overall rating does not provide enough information and could be unreliable, if not supported over a large number of independent reviews/ratings. From another standpoint, reading through a large number of textual reviews in order to infer the aspect ratings could be quite time con-

Proceedings of the NAACL HLT 2010 Workshop on Semantic Search, pages 36–43,
Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

suming, and, at the same time, the outcome of the evaluation could be biased by the reader's interpretation. In this work, instead of a single overall rating, we propose to provide ratings for multiple aspects of the product/service. For example, in the case of restaurant reviews, we consider ratings for five aspects: *food*, *atmosphere*, *value*, *service* and *overall experience*. In Lu et al. (2009) such aspect ratings are called *rated aspect summaries*, in Shimada and Endo (2008) they have been referred to as *seeing stars* and in Snyder and Barzilay (2007) they are referred to as *multi–aspect ranking*. We use supervised learning methods to train predictive models and use a specific decoding method to optimize the aspect rating assignment to a review.

In the rest of this paper, we overview the previous work in this research area in Section 2. We describe the corpus used in the experiments in Section 3. In Section 4 we present various learning algorithms we experimented with. Section 5 explains our experimental setup, while in Section 6 we provide analysis of our experimental results. Section 7 presents details of modeling and exploiting interdependence among aspect ratings to boost the predictive performance. Finally, we describe the future work in Section 8 and report the concluding remarks in Section 9.

2 Related work

Previous work in sentiment analysis (Turney, 2002; Pang et al., 2002; Dave et al., 2003; Yu and Hatzivassiloglou, 2003; Pang and Lee, 2004; Yi and Niblack, 2005; Carenini et al., 2006) used different information extraction and supervised classification methods to detect document opinion polarity (positive vs. negative).

By conducting a limited experiment with two subjects, Pang and Lee (2005) demonstrated that humans can discern more grades of positive or negative judgments by accurately detecting small differences in rating scores by just looking at review text. In a five–star schema, for instance, the subjects were able to perfectly distinguish rating differences of three *notches* or 1.5 stars and correctly perceive differences of one star with an average of 83% accuracy. This insight confirms that a five–star scale improves the evaluative information and is perceived with the right discriminative strength by the users.

Pang and Lee applied supervised and semi–supervised classification techniques, in addition to *linear, ϵ-insensitive* SVM regression methods, to predict the overall ratings of movie reviews in three and four–class star rating schemes. In the books review domain, Okanohara and Tsujii (2005) show a similar approach with comparable results. Both these contributions consider only overall ratings, which could be sufficient to describe sentiment for movie and book reviews. Two recent endeavors, Snyder and Barzilay (2007) for the restaurants domain, and Shimada and Endo (2008) for video games reviews, exploit multi–aspect, multiple rating modeling. Snyder and Barzilay (2007) assume inter–dependencies among the aspect ratings and capture the relationship between the ratings via the *agreement relation*. The agreement relation describes the likelihood that the user will express the same rating for all the rated aspects. Interestingly, Snyder and Barzilay (2007) show that modeling aspect rating dependencies helps to reduce the rank loss by keeping in consideration the contributions of the opinion strength of the single aspects referred to in the review. They incorporated information about the aspect rating dependencies in a regression model and minimized the loss (overall *grief*) during decoding. Shimada and Endo (2008) exploits a more traditional supervised machine learning approach where features such as word unigrams and frequency counts are used to train classification and regression models. As detailed in Section 4, our approach is similar to (Snyder and Barzilay, 2007) in terms of review domain and algorithms, but we improve on their performances by optimizing classification predictions.

3 Reviews corpus

Labeled data containing textual reviews and aspect ratings are rarely available. For this work, reviews were mined from the `we8there.com` websites around the end of 2008. `we8there.com` is one of the few websites, where, besides textual reviews, numerical ratings for different aspects of restaurants are also provided. Aspects used for rating on this site are: *food*, *service*, *atmosphere*, *value* and *overall experience*. Ratings are given on a scale from 1 to 5; for example, reviewers posting opinions were asked to rank their overall experience by the following prompt: *"On a scale of 1 (poor) to 5 (excellent), please rate your dining experience"*, and then enter a textual description by the prompt: *"Please describe your experience (30 words minimum)"*. At the time of mining, this site had reviews of about 3,800 restaurants with an average of two reviews per restaurant containing around eight sentences per review. A more detailed description is reported in Table 1. Table 2 shows review ratings distribution over the aspects. Rating distributions are evidently skewed toward high ratings with 70% or more reviews appraised as *excellent* (rank 5) or *above average* (rank 4).

Restaurants	3,866
Reviewers	4,660
Reviews	6,823
Average reviews per restaurant	1.76
Number of sentences	58,031
Average sentences per review	8.51

Table 1: Restaurant review corpus

Rating	1	2	3	4	5
Atmosphere	6.96	7.81	14.36	23.70	47.18
Food	8.24	6.72	9.86	18.53	56.65
Value	9.37	7.57	13.61	23.27	46.18
Service	11.83	6.12	11.91	22.00	48.14
Overall	10.48	8.19	10.17	20.47	50.69

Table 2: Restaurant review ratings distribution per aspect

4 Learning algorithms

In this section we review machine learning approaches that can predict ordinal ratings from textual data. The goal is *ordinal regression*, which differs from traditional numeric regression because the targets belong to a discrete space, but also differs from classification as one wants to minimize the rank loss rather than the classification error. The rank loss is the average difference between actual and predicted ratings and is defined as

$$\textit{RankLoss} = \frac{1}{N}\sum_{i}^{N}(|r_{a_i} - r_{p_i}|)$$

where r_{a_i} and r_{p_i} are actual and predicted ratings respectively for the instance i, and N is the number of considered reviews. There are several possible approaches to such a regression problem.

1. The most obvious approach is *numeric regression*. It is implemented with a neural network trained using the back–propagation algorithm.

2. Ordinal regression can also be implemented with *multiple thresholds* ($r - 1$ thresholds are used to split r ranks). This is implemented with a Perceptron based ranking model called *PRank* (Crammer and Singer, 2001).

3. Since rating aspects with values 1, 2, 3, 4 and 5 is an ordinal regression problem it can also be interpreted as a *classification* problem, with one class per possible rank. In this interpretation, ordering information is not directly used to help classification. Our implementation uses binary one-vs-all Maximum Entropy (MaxEnt) classifiers. We will see that this very simple approach can be extended to handle aspect interdependency, as presented in section 7.

In order to provide us with a broad range of rating prediction strategies, we experimented with a numerical regression technique viz. neural network, an ordinal regression technique viz. PRank algorithm, and a classification technique viz. MaxEnt classifiers. Their implementations are straightforward and the run–time highly efficient. After selecting a strategy from the previous list, one could consider more advanced algorithms described in Section 8.

5 Experimental setup

To predict aspect ratings of restaurants from their textual reviews we used the reviews mined from the `we8there.com` website to train different regression and classification models as outlined in Section 4. In each of our experiments, we randomly partitioned the data into 90% for training and 10% for testing. This ensures that the distributions in training and test data are identical. All the results quoted in this paper are averages of 10–fold cross–validation over 6,823 review examples. We conducted repeatedly the same experiment on 10 different training/test partitions and computed the average rank loss over all the test partitions.

Figure 1 illustrates the training process where each aspect is described by a separate predictive model.

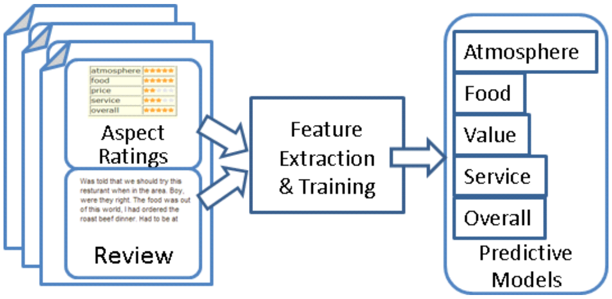

atmosphere
food
price
service
overall
Aspect Ratings
Was told that we should try this resturant when in the area. Boy, were they right. The food was out of this world, I had ordered the roast beef dinner. Had to be at
Review
Feature Extraction & Training
Atmosphere
Food
Value
Service
Overall
Predictive Models

Figure 1: Predictive model training

We introduce the following notation that will be helpful in further discussion. There are m aspects. For our data m is 5. Each aspect can have an integer rating from 1 to k. Once again, for our data k is 5. Each review text document t can have ratings $\mathbf{r}$, which is a vector of m integers ranging 1 to k (bold faced letters indicate vectors). Using the training data $(t_1, \mathbf{r}^1)..(t_i, \mathbf{r}^i)..(t_n, \mathbf{r}^n)$ we train m *rating predictors* $R_j(t_i)$, one for each aspect j. Given text t_i predictor R_j outputs the most likely rating l for the aspect j. In these experiments, we treated aspect rating predictors as independent of each other. For each rated aspect, predictor models were trained independently and were used independently to predict ratings for each aspect.

5.1 Feature Selection

We experimented with different combinations of features, including word unigrams, bigrams, *word chunks*, and *parts–of–speech* (POS) chunks. The assumption is that bag–of–unigrams capture the basic word statistic and that bigrams take into account some limited word context. POS chunks and word chunks discriminate the use of words in the context (e.g., a simple form word sense disambiguation) and, at the same time, aggregate co–occurring words (e.g., collocations), such as *sautéed_onions*, *buffalo_burger*, etc.

Most of the web–based reviews do not usually provide fine–grained aspect ratings of products or services, however, they often give an overall rating evaluation. We therefore also experimented with the overall rating as an input feature to predict the more specific aspect ratings. Results of our experiments are shown in Table 3.

Aspects	Uni-gram	Bi-gram	Word Chunks	Word Chunks POS Chunks	Uni gram Overall Rating
Atmosphere	0.740	0.763	0.789	0.783	0.527
Food	0.567	0.571	0.596	0.588	0.311
Value	0.703	0.725	0.751	0.743	0.406
Service	0.627	0.640	0.651	0.653	0.377
overall	0.548	0.559	0.577	0.583	
Average	0.637	0.652	0.673	0.670	0.405

Table 3: Average ranking losses using MaxEnt classifier with different feature sets

```
Review sentences
<s>Poor service made the lunch unpleasant.</s>
<s>The staff was unapologetic about their mistakes they
just didn't seem to care.</s>
<s>For example the buffalo burger I ordered with sauteed
onions and fries initially was served without either.</s>
<s> The waitress said she'd bring out the onions but had
I waited for them before eating the burger the meat would
have been cold.</s>
<s>Other examples of the poor service were that the
waitress forgot to bring out my soup when she brought out
my friend's salad and we had to repeatedly ask to get our
water glasses refilled.</s>
<s> When asked how our meal was I did politely mention my
dissatisfaction with the service but the staff person's
response was silence not even a simple I m sorry.</s>
<s>I won't return.  </s>
Word Chunks
poor_service made lunch unpleasant
staff unapologetic mistakes n't care
example buffalo_burger ordered sauteed_onions fries served
waitress said bring onions waited eating burger meat cold
other_examples poor_service waitress_forgot bring
soup brought friend salad repeatedly ask_to_get water
glasses_refilled
asked meal politely_mention dissatisfaction service
staff_person response silence not simple sorry
n't return
Parts-of-speech Chunks
NNP_NN VBD NN JJ
NN JJ NNS RB VB
NN NN_NN VBD NN_NNS NNS VBN
NN VBD VB NNS VBD VBG NN NN JJ
JJ_NNS JJ_NN NN_NN VB NN VBD NN NN RB VB_TO_VB NN VBZ_VBN
VBD NN RB_VB NN NN NN_NN NN NN RB JJ JJ
RB VB
```

Table 4: Example of reviews and extracted word chunks

Unigram and bigram features refer to unigram words and bigram words occurring more than 3 times in the training corpus. Word chunks are obtained by only processing Noun (NP), Verb (VP) and Adjective (ADJP) phrases in the review text. We removed modals and auxiliary verbs form VPs, pronouns from NPs and we broke the chunks containing conjunctions. Table 4 shows an example of extracted word and parts–of–speech chunks from review text. As can be seen, word chunks largely keep the information bearing chunks phrases and remove the rest. Parts–of–speech chunks are simply parts–of–speech of word chunks.

In spite of richness of word and parts-of-speech, chunks models using word unigrams perform the best. We can attribute this to the data sparseness, never–the–less, this results is in line with the findings in Pang et al. (2002). Last column of Table 3 clearly shows that use of overall rating as input feature significantly improves the performance. Clearly this validates the intuition that aspect ratings are highly co–related with overall ratings.

For the remaining experiments, we used only the unigram words as features of the review text. Since overall ratings given by reviewers may contain their biases and since they may not always be available, we did not use them as input features. Our hope is that even though we train the predictors using reviewers provided aspect ratings, learned models will be able to predict aspect ratings that depend only on the review text and not on reviewer's biases.

5.2 Results

Table 5 shows the results of our evaluation. Each row in this table reports average rank loss of four different models for each aspect. The baseline rank loss is computed by setting the predicted rank for all test examples to 5, as it is the most frequently occurring rank in the training data (see also Table 2). As shown in Table 5, the average baseline rank loss is greater than one. The third column shows the results from the neural network–based numeric regression. The fourth column corresponds to the Perceptron–based PRank algorithm. The MaxEnt classification results appear in the last column. For these results, we also detail the standard deviation over the 10 cross–validation trials.

Aspects	Base-line	Back-Prop.	Percep-tron	MaxEnt
Atmosphere	1.036	0.772	0.930	0.740 ± 0.022
Food	0.912	0.618	0.739	0.567± 0.033
Value	1.114	0.740	0.867	0.703± 0.028
Service	1.116	0.708	0.851	0.627± 0.033
Overall	1.077	0.602	0.756	0.548± 0.026
Average	1.053	0.694	0.833	0.637± 0.020

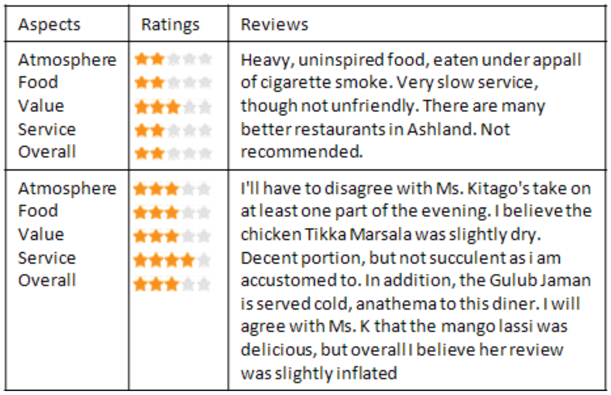
Table 5: Average ranking losses using different predictive models

6 Analysis

As can be seen in table Table 5, *Atmosphere* and *Value* are the worst performers. This is caused by the missing textual support for these aspects in the training data. Using manual examination of small number of examples, we found that only 62% of user given ratings have supporting text for ratings of these aspects in the reviews.

For example, in Figure 2 the first review clearly expresses opinions about food, service and atmosphere (*under appall of cigarette smoke*), but there is no evidence about *value* which is ranked three, two notches above the other aspects. Similarly, the second review is all about food without any reference to *service* rated two notches above the other aspects, or *atmosphere* or *value*.

Because of this reason, we do not expect any predictive model to do much better than 62% accuracy. Manual examination of a small number of examples also showed that 55% of ratings predicted by MaxEnt models are supported by the review text. This is 89% of 62% (a rough upper bound) and can be considered satisfactory given small data set and differences among reviewers rating preference. One way to boost the predictive performance would be to first determine if there is a textual support for an aspect rating, and use only the supported aspect ratings for training and evaluation of the models. This however, will require labeled data that we tried to avoid in this work.



Aspects	Ratings	Reviews
Atmosphere Food Value Service Overall	★★☆☆☆ ★★☆☆☆ ★★★☆☆ ★★☆☆☆ ★★☆☆☆	Heavy, uninspired food, eaten under appall of cigarette smoke. Very slow service, though not unfriendly. There are many better restaurants in Ashland. Not recommended.
Atmosphere Food Value Service Overall	★★★☆☆ ★★★☆☆ ★★★☆☆ ★★★★☆ ★★★☆☆	I'll have to disagree with Ms. Kitago's take on at least one part of the evening. I believe the chicken Tikka Marsala was slightly dry. Decent portion, but not succulent as i am accustomed to. In addition, the Gulub Jaman is served cold, anathema to this diner. I will agree with Ms. K that the mango lassi was delicious, but overall I believe her review was slightly inflated

Figure 2: Example of ratings with partial support in the text review

To our surprise, MaxEnt classification, although it minimizes a classification error, performs best even when evaluated using rank loss. As can be noticed, the performance difference over the second best approach (back–propagation) usually exceeds the standard deviation.

MaxEnt results are also comparable to those presented in Snyder and Barzilay (2007) using the *Good Grief* algorithm. Snyder and Barzilay (2007) also used data from the `we8there.com` website. While we are using the same data source, note the following differences: (i) Snyder and Barzilay (2007) used only 4,488 reviews as opposed to the 6,823 reviews used in our work; (ii) our results are averaged over a 10 fold cross validation. As shown with the baseline results reported in Table 6, the impact on performance that can be attributed to these differences is small. The most significant number, which should minimize the impact of data discrepancy, is the improvement over baseline (labeled as *"gain over baseline"* in Table 6). In that respect, our MaxEnt classification–based approach outperforms *Good Grief* for every aspect. Note also that, while we trained 5 independent predictors (one for each aspect) using only word unigrams as features, the *Good Grief* algorithm additionally modeled the agreements among aspect ratings and used the presence/absence of opposing polarity words in reviews as additional features.

	Our results			Snyder and Barzilay (2007)		
Aspects	Baseline	Max Ent.	Gain over Baseline	Baseline	Good Grief	Gain over Baseline
Atmosphere	1.039	0.740	0.299	1.044	0.774	0.270
Food	0.912	0.567	0.344	0.848	0.534	0.314
Value	1.114	0.703	0.411	1.030	0.644	0.386
Service	1.116	0.627	0.489	1.056	0.622	0.434
Overall	1.077	0.548	0.529	1.028	0.632	0.396

Table 6: Comparison of rank loss obtained from MaxEnt classification and those reported in Snyder and Barzilay (2007)

7 Modeling interdependence among aspect ratings

Inspired by these observations, we also trained MaxEnt classifiers to predict pair–wise absolute differences in aspect ratings. Since the difference in ratings of any two aspects can only be 0,1,2,3 or 4, there are 5 classes to predict. For each test example, MaxEnt classifiers output the posterior probability to observe a class given an input example. In our approach, we use these probabilities to compute the best joint assignment of ratings to all aspects. More specifically, in our modified algorithm we use 2 types of classifiers.

- **Rating predictors** - Given the text t_i, our classifiers $R_j(t_i)$ output vectors $\mathbf{p}^i$ consisting of probabilities p^i_l for text t_i having a rating l for the aspect j.

- **Difference predictors** - These correspond to classifiers $D_{j,k}(t_i)$ which output vectors $\mathbf{p}^{i_{j,k}}$. Elements of these vectors are the probabilities that the difference between ratings of aspects j and k is 0,1,2,3 and 4, respectively. While j ranges from 1 to m, k ranges from 1 to $j-1$. Thus, we trained a total of $m(m-1)/2 = 10$ difference predictors.

To predict aspect ratings for a given review text t_i we use both rating predictors and difference predictors and generate output probabilities. We then select the most likely values of $\mathbf{r}^i$ for text t_i that satisfies the probabilistic constraints generated by the predictors. More specifically:

$$\mathbf{r}_i = \underset{\mathbf{r}\in\mathfrak{R}}{argmax} \sum_{j=1}^{m} log(p^i_{r_j}) + \sum_{j=1}^{m}\sum_{k=1}^{j} log(p^{i_{j,k}}_{|r_j-r_k|})$$

$\mathfrak{R}$ is the set of all possible ratings assignments to all aspects. In our case it contains 5^5 (3,125) tuples. tuples in our case. Like Snyder and Barzilay (2007), we also experimented with additional features indicating presence of positive and negative polarity words in the review text. Besides unigrams in the review text, we also used 3 features: the counts of positive and negative polarity words and their differences. Polarity labels are obtained from a dictionary of about 700 words. This dictionary was created by first collecting words used as adjectives in a corpus of un–related review text. We then retained only those words in the dictionary that, in a context free manner generally conveyed positive or negative evaluation of any object, event or situation. Some examples of negative words are *awful, bad, boring, crude, disappointing, horrible, worst, worthless, yucky* and some examples of positive words are *amazing, beautiful, delightful, good, impeccable, lovable, marvelous, pleasant, recommendable, sophisticated, superb, wonderful, wow*. Table 7 first shows gains obtained from using difference predictors, and then gains from using polarity word features in addition to these difference predictors.

Aspects	MaxEnt	+ Difference predictor	+ Polarity features
Atmosphere	0.740	0.718	0.707
Food	0.567	0.552	0.547
Value	0.703	0.695	0.685
Service	0.627	0.627	0.617
Overall	0.548	0.547	0.528
Average	0.637	0.628	0.617

Table 7: Improved rank loss obtained by using difference predictors and polarity word features

8 Future Work

We have presented 3 algorithms chosen for their simplicity of implementation and run time efficiency. The results suggest that our classification–based approach performs better than numeric or ordinal regression approaches. Our next step is to verify these results with the more advanced algorithms outlined below.

1. For many numeric regression problems, (boosted) classification trees have shown good performance.

2. Several multi–threshold implementations of Support Vector Ordinal Regression are compared in Chu and Keerthi (2005). While they are more principled than the Perceptron–based PRank, their implementation is significantly more complex. A simpler approach that performs regression using a single classifier extracts extended examples from the original examples (Li and Lin, 2007).

3. Among classification–based approaches, nested binary classifiers have been proposed (Frank and Hall, 2001) to take into account the ordering information, but the prediction procedure based on classifier score difference is ad–hoc.

9 Conclusions

Textual reviews for different products and services are abundant. Still, when trying to make a buy decision, getting sufficient and reliable information can be a daunting task. In this work, instead of a single overall rating we focus on providing ratings for multiple aspects of the product/service. Since most textual reviews are rarely accompanied by multiple aspect ratings, such ratings must be deduced from predictive models. Several authors in the past have studied this problem using both classification and regression models. In this work we show that even though the aspect rating problem seems like a regression problem, maximum entropy classification models perform the best. Results also show a strong inter–dependence in the way users rate different aspects.

Acknowledgments

We thank Remi Zajac and his team for their support.

References

Carenini, Giuseppe, Raymond T. Ng, and Adam Pauls. 2006. Interactive multimedia summaries of evaluative text. In *Proceedings of Intelligent User Interfaces (IUI)*. ACM Press, pages 124–131.

Chu, Wei and S. Sathiya Keerthi. 2005. New approaches to support vector ordinal regression. In *Proceedings of the 22nd International Conference on Machine Learning*. Bonn, Germany, pages 145–152.

Crammer, Koby and Yoram Singer. 2001. Pranking with ranking. In *Advances in Neural Information Processing Systems 14*. MIT Press, pages 641–647.

Dave, Kushal, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *WWW '03: Proceedings of the 12th International Conference on World Wide Web*. ACM, New York, NY, USA, pages 519–528.

Frank, Eibe and Mark Hall. 2001. A simple approach to ordinal classification. In *Proceedings* *of the Twelfth European Conference on Machine Learning*. Springer-Verlag, Berlin, pages 145–156.

Hu, Minqing and Bing Liu. 2004. Mining and summarizing customer reviews. In *KDD '04: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, pages 168–177.

Li, Ling and Hsuan-Tien Lin. 2007. Ordinal regression by extended binary classification. In B. Schölkopf, J. C. Platt, and T. Hofmann, editors, *Advances in Neural Information Processing Systems 19*. MIT Press, pages 865–872.

Lu, Yue, ChengXiang Zhai, and Neel Sundaresan. 2009. Rated aspect summarization of short comments. In *WWW '09: Proceedings of the 18th International Conference on World Wide Web*. ACM, New York, NY, USA, pages 131–140.

Okanohara, Daisuke and Jun-ichi Tsujii. 2005. Assigning polarity scores to reviews using machine learning techniques. In Robert Dale, Kam-Fai Wong, Jian Su, and Oi Yee Kwong, editors, *IJCNLP*. Springer, volume 3651 of *Lecture Notes in Computer Science*, pages 314–325.

Pang, Bo and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the Association for Computational Linguistics (ACL)*. pages 271–278.

Pang, Bo and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the Association for Computational Linguistics (ACL)*. pages 115–124.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pages 79–86.

Popescu, Ana-Maria and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*.

Shimada, Kazutaka and Tsutomu Endo. 2008. Seeing several stars: A rating inference task for a document containing several evaluation criteria. In *Advances in Knowledge Discovery and Data Mining, 12th Pacific-Asia Conference, PAKDD 2008*. Springer, Osaka, Japan, volume 5012 of *Lecture Notes in Computer Science*, pages 1006–1014.

Snyder, Benjamin and Regina Barzilay. 2007. Multiple aspect ranking using the Good Grief algorithm. In *Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL)*. pages 300–307.

Turney, Peter. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the Association for Computational Linguistics (ACL)*. pages 417–424.

Yi, Jeonghee and Wayne Niblack. 2005. Sentiment mining in WebFountain. In *Proceedings of the International Conference on Data Engineering (ICDE)*.

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