Computational Linguistics and Chinese Language Processing
Vol. 19, No. 2, June 2014, pp. 17-32

Modeling the Helpful Opinion Mining of Online Consumer Reviews as a Classification Problem

Yi-Ching Zeng[*], Tsun Ku[+], Shih-Hung Wu[*],

Liang-Pu Chen[#], and Gwo-Dong Chen[+]

Abstract

The paper addresses an opinion mining problem: how to find the helpful reviews from online consumer reviews via the quality of the content. Since there are too many reviews, efficiently identifying the helpful ones earlier can benefit both consumers and companies. Consumers can read only the helpful opinions from helpful reviews before they purchase a product, while companies can acquire the true reasons a product is liked or hated. A system is built to assess the difficulty of the problem. The experimental results show that helpful reviews can be distinguished from unhelpful ones with high precision.

Keywords: Helpful Opinion Mining, Online Consumer Review, Online Customer Reivew, Text Quality.

1. Introduction

Online consumer (or customer) review is a very important information source for many potential consumers to decide whether to buy a product or not. Li *et al.* (2011) shows that, compared to an expert product review, "the consumer product review in the online shopping environment will be perceived by consumers to be more credible." This fact makes opinion mining of consumer reviews more interesting since it shows that opinions from other consumers are more appreciated than those from experts. Nevertheless, some reviews are not

[*] Department of Computer Science and Information Engineering, Chaoyang University of Technology, Taichung, Taiwan, R.O.C
E-mail: st9506522@gmail.com; shwu@cyut.edu.tw
The author for correspondence is Shih-Hung Wu.
[+] Department of Computer Science and Information Engineering, National Central University, Taiwan
E-mail: cujing@gmail.com; chen@csie.ncu.edu.tw
[#] Institute for Information Industry, Taiwan
E-mail: eit@iii.org.tw

very helpful, as we can see from the voting results on each consumer review from readers on Amazon.com.

This paper will address an opinion mining problem: how to find the helpful reviews from online consumers' reviews before mining the information from them. This task can benefit both consumers and companies. Consumers can read the opinions from useful reviews before they purchase a product, while companies can acquire the true reasons a product is liked or hated. Both save time from reading meaningless opinions that do not show good reasons. Figure 1 shows a clip image of an Amazon.com customer review. Each review has been labeled with stars by the author and people who found the review helpful and has been labeled with the number of total votes. A three-class classification problem is defined to model this application. A system is designed to find the helpful positive reviews for finding good reasons to buy a product; to find the helpful negative reviews for finding reasons not to buy a product; and to filter out the unhelpful reviews, no matter whether they are positive or negative.

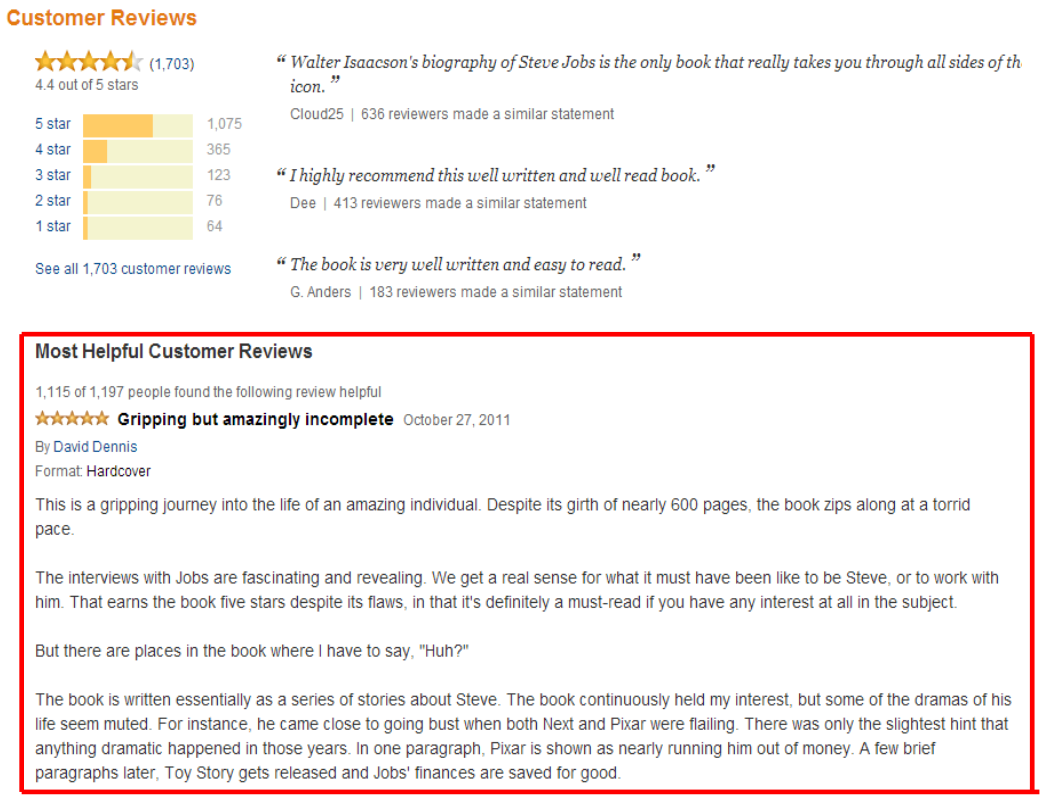


Figure 1. A clip image of an Amazon.com customer review.

The paper is organized as follows. Section 2 describes the related works. Section 3 describes the features that can be used to classify the reviews as helpful or unhelpful. Section

4 describes the data collection of this study. Section 5 reports and discusses the experiment. The final section gives conclusions and future work.

2. Related Works

Early works on opinion mining focused on the polarity of opinion, positive or negative; this kind of opinion mining is called sentiment analysis. Another type of opinion mining focused on finding the detailed information of a product from reviews; this approach is a kind of information extraction (Hu & Liu, 2004). Recent research has focused on assessing the review quality before mining the opinion. Kim *et al*. (2006) explored the use of some semantic features for review helpfulness ranking. They found that some important features of a review, including length, unigrams, and stars, might provide the basis for assessing the helpfulness of reviews. Siersdorfer *et al*. (2010) presented a system that could automatically structure and filter comments for YouTube videos by analyzing dependencies between comments, views, comment ratings, and topic categories. Their method used the SentiWordNet thesaurus, a lexical WordNet-based resource containing sentiment annotations. Moghaddam *et al*. (2011) proposed the Matrix Factorization Model and Tensor Factorization Model to predict of the quality of online reviews, and they evaluated the models on a real-life database from Epinions.com. Lu (2010) exploited contextual information about authors' identities and social networks to improve review quality prediction. Lu's method provided a generic framework to incorporate social context information by adding regularization constraints to the text-based predictor. Xiong and Litman (2011) investigated the utility of incorporating specialized features tailored to peer-review helpfulness. They found that structural features, review unigrams, and meta-data combination were useful in modeling the helpfulness of both peer reviews and product reviews.

3. Classification Features

3.1 Observation

Observation is necessary to find features for the helpful/unhelpful classification. Connors *et al*. (2011) gave a list of common ideas related to helpfulness and unhelpfulness, shown in Table 1, which was collected from 40 students, with each student reading 20 online reviews about a single product and giving comments on the reviews. The study provided 15 reasons people think a consumer review is helpful and 10 reasons why it is unhelpful. These ideas can be considered as features for a classifier. Nevertheless, some of them are difficult to implement and require clear definition. For example, mining comparative sentences from text requires considerable knowledge of the language. (Jindal & Liu, 2006).

Table 1. The 15 reasons that people think a customer review helpful and the 10 reasons they think it to be unhelpful (Connors et al., 2011).

Helpfulness	**Times Mentioned**
Pros and Cons	36
Product Usage Information	30
Detail	24
Good Writing Style	13
Background Knowledge of Product	12
Personal Information about Reviewer	12
Comparisons	10
Layman's Terms	9
Conciseness	8
Lengthy	7
Use of Ratings	7
Authenticity	5
Honesty	5
Miscellaneous	4
Unbiased	4
Accuracy	3
Relevancy	3
Thoroughness	3
Unhelpfulness	**Times Mentioned**
Overly Emotional/Biased	24
Lack of Information	17
Irrelevant Comments	9
Not Enough Detail	6
Poor Writing Style	6
Using Technical Language	6
Low Credibility	5
Problems with Quantitative Rating	5
Too Much Detail	5

3.2 Features

Table 2 lists the features that we implement in this study. Compared with the features used in Kim *et al.* (2006), we add more features, based on the observation of Connors *et al.* (2011), especially the degree of detail. The first three features are common n-grams used between a review and the corresponding product description. We believe that they are effective since a good review should contain more relevant information and use exact terminology. The fourth feature is the length of the review. A very short review cannot give much information, and a long review might give more useful information. The fifth feature is whether or not the review makes a comparison among things. A good review should compare similar products. Our program detects whether the string "compare to/with" or the pattern "ADJ+er than" exists in the review or not, with the help of a list of comparative adjectives. The sixth feature is the degree of detail, which is a combination of length and n-gram. The degree of detail has not been defined well in previous works. Our definition is only a tentative one. We define the degree of detail of a review as:

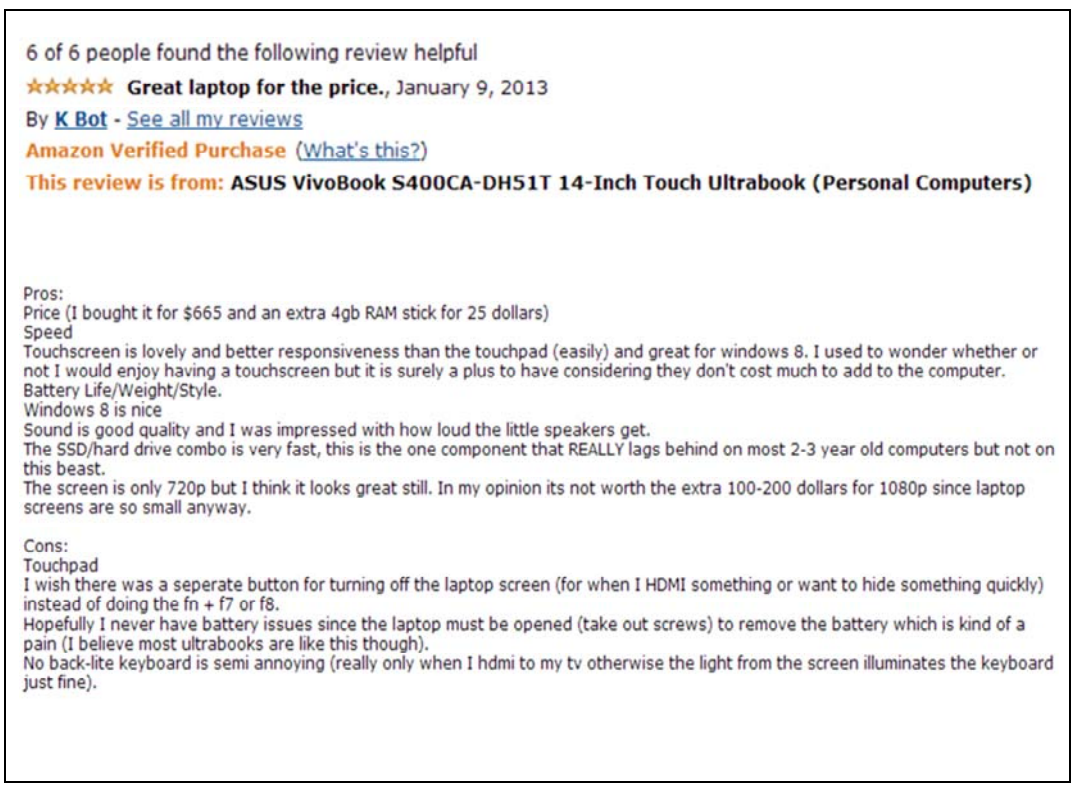
$$log_{10}(\text{Unigram+Bigram+Trigram+Length}) \quad (1)$$

where unigram, bigram, and trigram are the common n-grams between a review and the corresponding product description. Length is the length of the review. The seventh feature is the number of stars given by the review author. The eighth feature is whether the review contains "Pros" and "Cons" or not. Our system detects whether the string "Pros" and "Cons" exist in the review or not.

Table 2. Eight Features used in our system.

Feature	**Description**
Unigram (Product Description)	The number of unigrams used between the review and the corresponding product description
Bigram (Product Description)	The number of bigrams used between the review and the corresponding product description
Trigram (Product Description)	The number of trigrams used between the review and the corresponding product description
Length	The length of a review
Comparisons	The review uses the string "compare to" or "ADJ + er than"
Degree of detail	Defined by formula (1)
Use of Ratings	The "Star" ratings of the review
Pros and Cons	The review contains exact the strings "Pros" and "Cons"

We use an example to show the eight feature values. Consider the review in Figure 2, where the "pros_cons" value is 1, since we can see the author explicitly lists the pros and cons. The "Detail" value is 1.17760, as defined in Formula (1). The "Length" value is 568, which is the number of words in the review. The "Compare" value is 4, because the author really makes a comparison of this product with other products. The "Star" value is 5, since the author gave five stars to the product. The "Unigram" value is 15. The "Bigram" value is 0, since we found no common bigrams between the review and the corresponding product description (not shown here). Hence, the "Trigram" value is also 0.


6 of 6 people found the following review helpful

★★★★★ **Great laptop for the price.**, January 9, 2013

By **K Bot** - See all my reviews

Amazon Verified Purchase (What's this?)

This review is from: ASUS VivoBook S400CA-DH51T 14-Inch Touch Ultrabook (Personal Computers)

Pros:
Price (I bought it for $665 and an extra 4gb RAM stick for 25 dollars)
Speed
Touchscreen is lovely and better responsiveness than the touchpad (easily) and great for windows 8. I used to wonder whether or not I would enjoy having a touchscreen but it is surely a plus to have considering they don't cost much to add to the computer.
Battery Life/Weight/Style.
Windows 8 is nice
Sound is good quality and I was impressed with how loud the little speakers get.
The SSD/hard drive combo is very fast, this is the one component that REALLY lags behind on most 2-3 year old computers but not on this beast.
The screen is only 720p but I think it looks great still. In my opinion its not worth the extra 100-200 dollars for 1080p since laptop screens are so small anyway.

Cons:
Touchpad
I wish there was a seperate button for turning off the laptop screen (for when I HDMI something or want to hide something quickly) instead of doing the fn + f7 or f8.
Hopefully I never have battery issues since the laptop must be opened (take out screws) to remove the battery which is kind of a pain (I believe most ultrabooks are like this though).
No back-lite keyboard is semi annoying (really only when I hdmi to my tv otherwise the light from the screen illuminates the keyboard just fine).

Figure 2. Example of review

4. Data Collection

In order to test the idea, we collected online customer reviews manually from Amazon.com in March and April 2013. The reviews were from eight different product domains: Book, Digital Camera, Computer, Food & Drink, Movie, Shoes, Toys, and Cell phone. Without any special selection criterion in each domain, we collected the first available 1000+ reviews with an equal number of reviews of one to five stars. The average length was 80.63 words. The summary of our data collection is listed in Table 3.

Table 3. The summary of our data collection of 8 classifications and 8,690 reviews.

Product	Reviews	Total Reviews Words	Average Length	s.d.
Book	1,065	93,497	87.79	1.8
Digital Camera	1,028	93,404	90.85	2.7
Computer	1,067	83,708	78.45	2.1
Foods & Drink	1,025	71,027	69.29	1.7
Movies	1,097	94,037	88.13	2.5
Shoes	1,000	75,237	75.23	1.6
Toys	1,100	85,196	77.45	1.7
Cell Phone	1,308	101,957	77.88	2.0
Total / Average	8,690	884,964	80.63	2.02

The helpfulness score is given by the readers. As shown in Figure 1, the reviewer labeled the number of stars and other users voted the review as helpful or unhelpful. We take the confidence in being helpful as an index to sort the reviews. Figure 3 shows the distribution of polarity (from 1 to 5 stars) and the helpful/unhelpful confidence, where the y-axis is the confidence score. Note that the confidence score in previous works has been defined as:

$$\text{Confidence}=100\% \times \left(\frac{\#\text{of Think helpful vote}}{\#\text{of Total vote}} \right) \tag{2}$$

Nevertheless, since there are some high confidence reviews with very little support, the reviews might not be very helpful. We discount the confidence of them by redefining the confidence score as the log-support confidence (LSC):

$$\text{LSC}=log_{10}\left[\begin{array}{l} \#\text{of Think Help ful vote} * \\ (\#\text{of Think Help ful vote}/\ \#\text{ of Total vote}) \end{array} \right] \tag{3}$$

Figure 3 shows the data distribution. The positive reviews (with 4 or 5 stars) get higher helpfulness confidence in most product categories. This fact shows that readers think other consumers are credible. The confidence of helpfulness is lower for the negative reviews. The average LSC confidence scores for each product category are listed in Table 4.

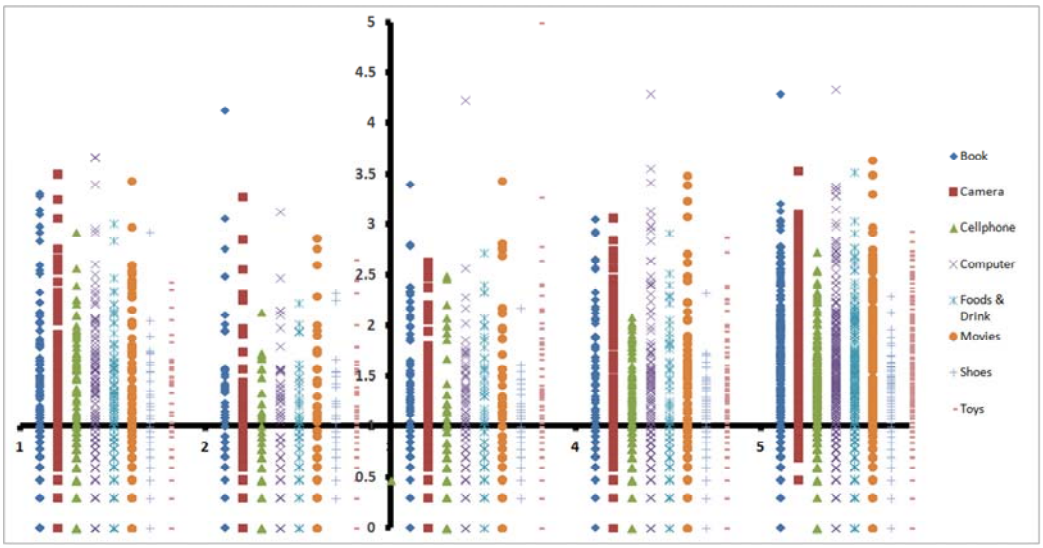


Figure 3. Stars vs. helpfulness distribution of our data collection. The x-axis is the number of stars of customer reviews; the y-axis is the confidence score LSC.

Table 4. The average LSC Confidence scores of the eight product categories.

Product	**Average LSC Confidence score**
Book	1.134
Digital Camera	1.373
Computer	1.140
Foods & Drink	0.932
Movies	1.116
Shoes	0.808
Toys	0.807
Cell Phone	1.005
Total average	1.039

4.1 The Three-class Classification Problem

Instead of finding the correlation between the ranking of helpfulness and the prediction, we define the problem as a three-class classification problem. The three classes are: the helpful

positive reviews, for finding good reasons to buy a product; the helpful negative reviews, for finding reasons not to buy a product; and the unhelpful reviews.

Since there is no distinct boundary between the helpful and the unhelpful and since one purpose of the system is to filter out the most unhelpful reviews, the sizes of the three classes can be adjusted by setting different thresholds. A higher threshold filters out more data. We can control the filtering level by setting different thresholds.

In our experiments, Class 1 includes positive reviews with 4 or 5 stars and the helpfulness confidence higher than the threshold. Class 2 includes negative reviews with 1 to 3 stars and the helpfulness confidence higher than the threshold. Class 3 is the remaining reviews, which are regarded as unhelpful, where the helpfulness confidence is lower than the threshold.

5. Experiments

The goal of the experiment is to test the filter accuracy of the three-class classification problem with different thresholds. We use the libSVM[1] toolkit to build the classifier, based on the features described in Section 2.2.

5.1 Experimental Design

We divide the data into a training set and test set, consisting of 7,690 reviews and 1,000 reviews, respectively. The class distribution of the test data are balanced to one third for each class. The different thresholds tested in our experiment are 1.039, 1.5, and 2.0. The first threshold is the average confidence score in Table 5, which filters out 56.1% of the reviews as unhelpful; the second threshold 1.5, filtering out 79.6%; and the third threshold 2.0, filtering out 91.0%. The numbers of useful (both positive and negative) reviews of each product domain to the three thresholds are listed in Tables 5, 7, and 9. The sizes of classes corresponding to the three thresholds are shown in Tables 6, 8, and 10.

Table 5. Number of reviews over the threshold "1.039"

Product	Reviews
Book	522
Digital Camera	698
Computer	532
Foods & Drink	404
Movies	521

[1] http://www.csie.ntu.edu.tw/~cjlin/lib

Shoes	246
Toys	318
Cell Phone	571
Total Reviews	3,812

Table 6. The size of the three classes with the threshold "1.039"

Classes	Reviews	%
Class 1 : Useful Positive	2,712	31.2%
Class 2 : Useful Negative	1,100	12.7%
Class 3 : Not Useful	4,878	56.1%
Total Reviews	8,690	

Table 7. Number of reviews over the threshold "1.5"

Product	Reviews
Book	270
Digital Camera	354
Computer	254
Foods & Drink	189
Movies	341
Shoes	49
Toys	174
Cell Phone	139
Total Reviews	1,770

Table 8. The size of the three classes with the threshold "1.5"

Classes	Reviews	%
Class 1 : Useful Positive	1,265	14.5%
Class 2 : Useful Negative	505	5.8%

Class 3 : Not Useful	6,920	79.6%
Total Reviews	8,690	

Table 9. Number of reviews over the threshold "2.0"

Product	**Reviews**
Book	129
Digital Camera	202
Computer	104
Foods & Drink	72
Movies	160
Shoes	9
Toys	73
Cell Phone	32
Total Reviews	781

Table 10. The size of the three classes with the threshold "2.0"

Classes	**Reviews**	**%**
Class 1 : Useful Positive	604	6.9%
Class 2 : Useful Negative	177	2.0%
Class 3 : Not Useful	7,910	91.0%
Total Reviews	8,690	

We conducted two experiments. The first one was a 10-fold validation on the training set, and the second one was a test on a separated test set.

5.2 Experimental Results

The average accuracy of the 10-fold cross-validation result of each configuration is shown in Table 11. The 7,690 training data were separated into ten folds, and the system used 90% of the data as the training set and the other 10% as the test set. A SVM classifier was trained in each fold and repeated 10 times. The result shows that, with a higher threshold, 1.5 or 2.0, the accuracy of our system is about 72%.

Table 11. The average accuracy result of each data set in the ten-fold cross-validation

Data set	Average Accuracy
LSC threshold 1.039	60.83%
LSC threshold 1.5	72.72%
LSC threshold 2.0	**72.82%**

In the second experiment, we used the 7,690 reviews as a training set and tested the classification on the 1,000 test set, where the number of tests of each class was balanced to 1/3. Note that the actual class of the test was fixed during the test, which corresponds to a threshold 1.039. The classifier was trained with three different class distributions. The confusion matrix of our system is shown in Tables 12 to 14. The precision and the recall of each class are also shown.

Table 12. The confusion matrix (LSC threshold is over 1.039)

Predicted	**Actual**			**Total**	**Precision**
	Class 1	Class 2	Class 3		
Class 1	172	75	46	293	59%
Class 2	80	196	24	300	65%
Class 3	81	62	264	407	65%
Total	333	333	334	1,000	
Recall	52%	59%	79%		

Table 13. The confusion matrix (LSC threshold is over 1.5)

Predicted	**Actual**			**Total**	**Precision**
	Class 1	Class 2	Class 3		
Class 1	213	47	28	288	74%
Class 2	42	257	14	313	82%
Class 3	78	29	292	399	73%
Total	333	333	334	1,000	
Recall	64%	77%	87%		

Table 14. The confusion matrix (LSC threshold is over 2.0)

Predicted	**Actual**			**Total**	**Precision**
	Class 1	Class 2	Class 3		
Class 1	203	45	27	275	74%
Class 2	46	263	10	319	82%
Class 3	84	25	297	406	73%
Total	333	333	334	1,000	
Recall	61%	79%	89%		

5.3 Feature Analysis Result

To compare which features are more important in the classifier, we conducted a series experiments with one less feature each time. The results are shown in Table 15. We can find that the "detail" feature is the most important. Second, third, and fourth are length, star, and unigram. Since detail is a hybrid feature, this result suggests that a hybrid feature works better than the combination of individual ones.

Table 15. Accuracy with all-minus-one features

Features	**Accuracy**
All-(Detail)	38.569%
All-(Compare)	52.152%
All-(Pros_cons)	49.727%
All-(Length)	39.594%
All-(Star)	39.342%
All-(Unigram)	42.493%
All-(Bigram)	55.339%
All-(Trigram)	49.469%

5.4 Discussion on the Experimental Result

Table 11 shows that the average accuracy numbers of the three data sets are 60.83%, 72.72%, and 72.82%. We find that setting the threshold to 1.5 is expected to prune 79.6% of data; our system can get 72.72% accuracy on the helpful/unhelpful classification. This is a considerable reduction of human labor to find better mining candidates.

From the confusion matrix in Table 13, we find that choosing the threshold as 1.5 enables our system to classify the three classes with precision 74%, 82%, and 73%; while the

system recall for the three classes are 64%, 77%, and 87%. We also can find a similar result in Table 14, where the threshold is 2.0. The precision is almost the same, and the recall is slightly different.

From Table 15, we can find that the “detail” feature is the most important. Without it, the accuracy drops from 60.83% to 38.57%. Nevertheless, each feature helps the performance, so no one feature can be omitted. This result also suggests that more features might be necessary to attain higher performance.

6. Conclusion and Future Works

The paper reports how a system can find helpful online reviews, and the system is tested on a three-class classification problem. The threshold of helpful/unhelpful reviews can be decided according to the amount of data that the users want to prune. The overall accuracy of the three-class problem is about 73%. Helpful negative reviews can be found with 82% precision and 77% recall. Helpful positive reviews can be found with 74% precision and 64% recall. Unhelpful reviews can be filtered out automatically from the consumer reviews with a high recall rate of about 87% with 73% precision. Considering the original data distribution (only 20% as useful), the system performance is quite high.

Currently, our system is based on features observed by humans in previous works, and we only implement some of them. In the future, we will try to implement more features and attempt to extract features from the training corpus automatically.

Acknowledgements

This study was conducted under the "Online and Offline Integrated Smart Commerce Platform (1/4)" of the Institute for Information Industry, which is subsidized by the Ministry of Economic Affairs of the Republic of China. This study was partially supported by Research Grant NSC 102-2221-E-324 -034 from the Ministry of Science and Technology.

Reference

Connors, L., Mudambi, S. M., & Schuff, D. (2011). Is it the Review or the Reviewer? A Multi-Method Approach to Determine the Antecedents of Online Review Helpfulness. In *Proceedings of the 2011 Hawaii International Conference on Systems Sciences (HICSS)*, January.

Hu, M., & Liu, B. (2004). Mining opinion features in customer reviews. In *Proceedings of the 19th national conference on Artifical intelligence (AAAI'04)*, Anthony G. Cohn (Ed.). AAAI Press 755-760.

Jindal, N., & Liu, B. (2006). Mining comparative sentences and relations. In *proceedings of the 21st national conference on Artificial intelligence - Volume 2* (AAAI'06), Anthony Cohn (Ed.), Vol. 2. AAAI Press 1331-1336.

Kim, S.-M., Pantel, P., Chklovski, T., & Pennacchiotti, M**. (**2006). Automatically Assessing Review Helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 423-430.

Li, M., Huang, L., Tan, C., & Wei, K. (2011) Assessing The Helpfulness Of Online Product Review: A Progressive Experimental Approach. In *Proceedings of PACIS*.

Lu, Y., Tsaparas, P., Ntoulas, A., & Polanyi, L. **(**2010). Exploiting Social Context for Review Quality Prediction. In *Proceedings of the 19th international conference on World wide web*, 691-700.

Moghaddam, S., Jamali, M., & Ester, M. (2010). Review Recommendation: Personalized Prediction of the Quality of Online Reviews. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2249-2252.

Mudambi, S. M., & Schuff, D. (2010). What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com. *MIS Quarterly*, *34*(1), 185-200.

Siersdorfer, S., Chelaru, S., & San Pedro, J. (2010). How useful are your comments?: analyzing and predicting youtube comments and comment ratings. In *Proceedings of the 19th international conference on World wide web*, 891-900.

Xiong, W., & Litman, D. (2011). Automatically Predicting Peer-Review Helpfulness. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 502-507.