Simplicity of Positive Reviews and Diversity of Negative Reviews in Hotel Reputation

Sachio Hirokawa
Research Institute for Information Technology
Kyushu University
Fukuoka, JAPAN
hirokawa@cc.kyushu-u.ac.jp

Kiyota Hashimoto
ESSAND
Prince of Songkla University
Phuket, Thailand
kiyota.h@phuket.psu.ac.th

***Abstract*—User's review on products and services is valuable information for both users and providers. The present paper conducted polarity estimation of 73,589 reviews on hotels in Europe. Users rated one to five points for seven aspects (Value, Rooms,Location, Cleanliness, Checkin, Service, Business, Overall). In this paper, we predicted the polarity (positive/negative) of each aspect by using machine learning method SVM (Support Vector Machine) and feature selection, with more than 4 points being positive and less than 3 being negative. As a result, positive reviews with respect to six aspects, other than Business, were able to achieve 74% prediction performance (F-measure) with only 20 feature words. On the other hand, for negative reviews, optimal prediction performance could not be obtained unless almost all words were used, and on average F-measure was only 27%. The results indicate that positive reviews are simple, meanwhile negative reviews are diverse and hard to predict mechanically**

Keywords— Sentiment Analysis; Hotel Review; SVM; Feature Selection

I. Introduction

. It is not easy for users to choose goods and services among abundant candidates. A product or service provider has the similar difficulties in communicating its merits to users. With the development of the Web, users can publish their impressions and opinions as reviews. These reviews are useful information resources for users and providers as well. In particular, the opinions of other users, which can be confirmed in advance, are valuable for products that can not be checked by hand or services that can only be experienced if not visited there. Hotel review is a typical case.

A study of sentinment analysis that mechanically performs a human evaluation on a subject rather than just a search began with [7]. In the early days, there were many studies of polarity, i.e. the quantification of whether a word represents a positive emotion or a negative emotion. Positive/Negative table [6] and polar phrase dictionary [2] have been constructed as the collection of case examples showing emotion progresses

In many reputation information sites, two types of information are provided, including free description and ranked evaluation with respect to several aspects. Therefore, simple evaluation analysis of whether the reputation is positive or negative has become insufficient. It is necessary to analyze which aspect is evaluated positive and what is the reason for that. Even the same word may have different polarity if the aspect is different. For example, [4] formulated an evaluation expression as a pattern that contains adjectives. Also, instead of polarity evaluation for general purpose, it is required to analyze according to purpose of use, for example, reputation information required by small hotels [1]. In other words, applying fixed evaluation criteria as sentiment analysis has become insufficient. Further, imbalance in amounts of positive reputation and that of negative reputation is recognized as an big issue. Therefore, in this paper, we predict the polarity of each reputation by machine learning and evaluate its prediction performance. We also extract feature words characterizing positive and negative reviews.

II. Seven Viewpoints in Hotel Review

[9] provides 246,399 hotel reviews. In this paper, we analyzed 73,589 hotels in Europe in the same dataset. Prices and places are available as basic information of each hotel, but we used only location information, in this paper. Users' reviews are provided for each hotel. In each review, five levels of evaluation are given for the seven aspects in Table I below in addition to the user's comment text. In this paper, we analyzed with more than 4 points being positive and less than 3 points being negative. For Overall, Cleanliness, Service, Rooms, and Value, 50% or more is a good evaluation. No bad evaluation is 10% in any aspect as we can see in Table I and Fig. I.

Table I Positive Reviews & Negative Reviews in Seven Aspects

	count			ratio		
aspect	pos	Neg	neu	pos	neg	neu
Overall	7359	59545	6585	0.10	0.81	0.09
Value	4545	39071	29873	0.06	0.53	0.41
Rooms	5497	47308	20684	0.07	0.64	0.28
Location	1359	31668	40462	0.02	0.43	0.55
Cleanliness	3164	51794	18531	0.04	0.70	0.25
Check in/ front desk	2582	29383	41524	0.04	0.40	0.57
Service	4639	47746	21104	0.06	0.65	0.29
Business	2195	16614	54680	0.03	0.23	0.74

Figure I Size of Positive Reviews & Negative Reviews

III. Prediction Performance by Support Vector Machine

Applying machine learning SVM (Support Vector Machine) and feature selection method [5] to each review, we predicted whether the review is positive or negative in each aspect. In the first step of [5], they adopted a linear kernel of SVM, and generated a linear model that distinguishes between positive examples and negative examples using all words. They called the coefficient of each word in the linear model as the SVM-score of the word. They used the top N positive words and the top N negative words, with respect to their SVM-score, for feature selection to represents a document as a vecotor of words. They changed N to find the optimum N. Note that negative cases are extremely few. So, accuracy is not appropriate as an evaluation index as an indicator of prediction performance. In this paper, F-measure is used as a measure of optimal feature selection.

A. Precition Performance with All Words of Positive Reviews

Table II and Table III show the discrimination performance when all words are used. As for the discrimination performance, the F-measure is 53% on average on positive reviews and 22% on negative reviews. In other words, it means that mechanical identification can not be done with the naive method.

Table II Prediction Performance of Positive Reviews

aspect	prec	recll	F-meas	acc
Overall	0.94	0.87	0.91	0.85
Value	0.77	0.48	0.59	0.65
Rooms	0.80	0.59	0.68	0.64
Location	0.71	0.29	0.41	0.64
Cleanliness	0.83	0.57	0.67	0.61
Check in/front desk	0.71	0.29	0.41	0.67
Service	0.79	0.58	0.67	0.63
Business	0.37	0.57	0.45	0.68
average	0.66	0.47	0.53	0.60

Table III Prediction Peformance of Negative Reviews

aspect	prec	recll	F-meas	acc
Overall	0.35	0.85	0.50	0.83
Value	0.27	0.75	0.40	0.86
Rooms	0.20	0.76	0.32	0.76
Location	0.08	0.41	0.13	0.90
Cleanliness	0.13	0.71	0.22	0.78
Check in/front desk	0.15	0.58	0.24	0.87
Service	0.17	0.75	0.28	0.75
Business	0.10	0.43	0.16	0.86
average	0.15	0.52	0.22	0.66

B. Prediction Performance with Feature Selection

Table IV and Table V show the discrimination performance when SVM and feature selection[Sakai 2010] are applied. Except for Busines, high discrimination performance is obtained with N=10, that is, 10 positive words and 10 negative words or less.

Table IV prediction peformance of positive reviews

aspect	N	prec	recll	F-meas	acc
Overall	500	0.92	0.90	0.91	0.86
Value	9	0.66	0.79	0.72	0.67
Rooms	1	0.64	1.00	0.78	0.64
Location	3	0.56	0.91	0.69	0.65
Cleanliness	9	0.72	0.98	0.83	0.72
Check in/front desk	2	0.52	0.91	0.66	0.63
Service	9	0.67	0.98	0.79	0.67
Business	4000	0.38	0.62	0.47	0.69
average		0.64	0.90	0.74	0.69

Table V prediction performance of negative reviews

	N	prec	recll	F-meas	acc
aspect	8000	0.35	0.85	0.50	0.83
Overall	8000	0.27	0.75	0.40	0.86
Value	5000	0.21	0.77	0.32	0.76
Rooms	1000	0.14	0.55	0.22	0.93
Location	5000	0.13	0.71	0.22	0.78
Cleanliness	5000	0.13	0.71	0.22	0.78
Check in/front desk	3000	0.19	0.64	0.29	0.89
Service	3000	0.19	0.78	0.30	0.76

Business	4000	0.11	0.49	0.18	0.86
average		0.17	0.63	0.27	0.75

It is natural that accuracy is high because there are only a few reviews of negative reviews. Therefore, it is necessary to look at the F-measure for the good or bad of the identification performance.Regarding the positive reviews, except for the aspect of Business, the optimum discrimination performance N is 10 or less and the F-measureis 74% on average. On the other hand, as for the identification of negative reviews, identification performance increases as the number of words used for identification increases, but still the F-measure is only 27% on average.

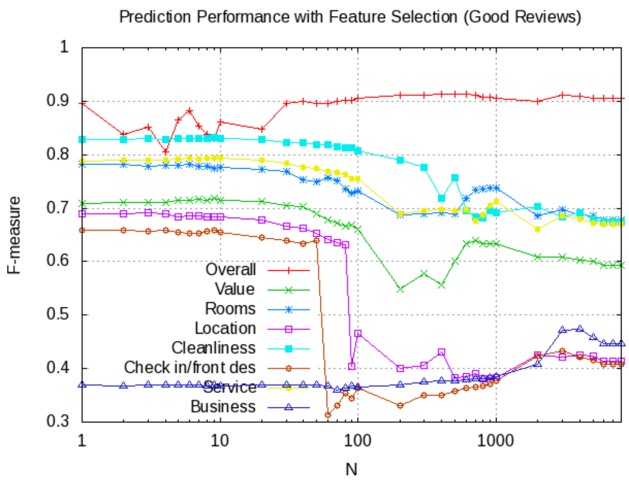

Figure II Prediction Performance with Feature Selection (Positive Reviews)

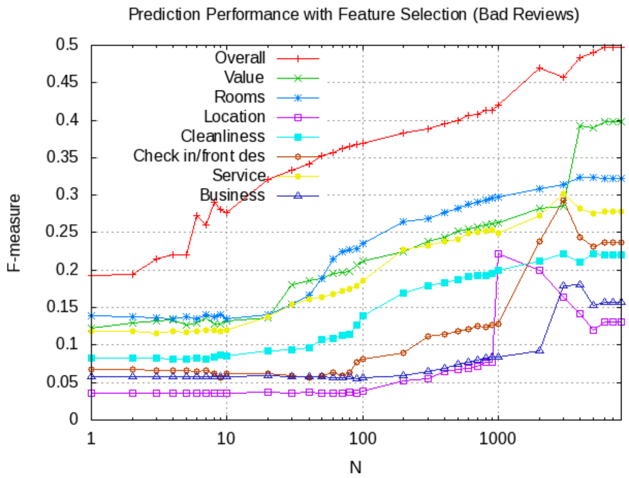

Figure III Prediction Performance with Feature Selection (Negative Reviews)

Fig. II and Fig. III display F-measures when changing the number N of feature selections. Regarding prediction performance of positive reviews (Fig. II), optimal performance ranges from 60% to 90% depending on aspect. However, for all aspects, within the range where the value of N is small, it is almost the same as the optimum discrimination performance. In other words, the positive reviews can be identified with a few feature words. On the other hand, looking at the negative review of Fig. III, the prediction performance increases as the number of feature increases. However, it is less than 50% at maximum.

Positive reviews are as simple as characterized with a few words and high discrimination performance. On the other hand, negative reviews can not be characterized with a few words, and even if many words are used, the discrimination performance is low. In other words, it can be said that a negative review has various factors.

IV. Features of Good Reviews and Bad Reviews in Eight Aspects

A. Positive Words and Negative Words

Table VI and Table VII represent the ten words with respec to the SVM-scores in the linear model obtained when positive reviews and negative review are used as positive cases in training. We can observe that adjectives appear most frequently in either positive reviews or negative rReview. This is what we expected.

Table VI Top 10 Feature Words of Positive Review

Overall	perfect excellent fantastic minor combination hesitate london great gem loved
Value	excelente muy struttura albergo est una soggiorno avons ottima perfecto
Rooms	lcd hoxton x niggles downside visible sandwiches recommander forum disco
Location	alessandra soho magda nadia wifi mate sumner ta fantastico marco
Cleanliness	sandwiches recognised correspondence southern insulated annoyance mugs shortcomings conseiller unknown
Check in/front desk	wifi soho nadia alessandra magda hoxton fantastico mate wireless michelangiolo
Service	lane technology maker carte owed opulent recommander pickpockets pint obliging
Business	hoxton davanzati wifi jays imac wireless fabrizio sumner sharmila wi

Table VII Top 10 Feature Words of Negative Reviews

Overall	dirty unfriendly unhelpful terrible worst shabby rude poor awful dump
Value	worst joke dirty overpriced inch filthy rip woken unfriendly broken

Rooms	dirty tiny worst smallest uncomfortable filthy stained prison mould worn
Location	outskirts estate homeless council dodgy motorway cap newspaper pavement dormitory
Cleanliness	dirty filthy mildew mold hair cleaner mouse smelly dust mould
Check in/front desk	eventually usd unhelpful dismissive unfriendly proof surly rude worst incompetent
Service	unfriendly rude disinterested unhelpful joke unpleasant attitude uninterested nasty indifferent
Business	intermittent pattern signal dirt spiders texture horrendous recognize incorrect abandoned

B. *Feature Words of Positive Reviews with respect to Value*

A lot of words such as Spanish and Italian other than English appear in the characteristic word of the positive reviews. So, we confirmed the number of sentences that contain such word and concrete examples for the words within the top ten (Table VIII). It was confirmed that they are Spanish, Italian, French except "excelent" of 1st place.

Table VIII Positive Words with respect to Value

rank	word	freq	Sample
1	excelent	8833	Great choice! Excelent choice in Paris.
2	mui	270	Muy buen hotel, con excelentes habitaciones.
3	struttura	105	Buona struttura e in posizione.
4	albergo	274	Ottimo albergo qualit prezzo
5	est	187	Excelente estancia en este centrico hotel
6	una	216	Para disfrutar de una buena estancia.
7	soggiorno	64	Un soggiorno indimenticabile
8	avon	45	Nous avons eu une chambre type
9	ottima	259	Ottima la posizione nei pressi
10	perfecto	67	Perfecto para visitar Florencia

After reviewing the location of those reviews, it was found that Barcelona and Florence accounted for 42% (Table IX). It was surprising that words other than English appeared as feature words in positive reviews for the aspect Value, since we made no choice or limiting the language in this paper. On the other hand, in the aspect of Value, all of the feature words of negative reviews are English only. Therefore, it seems that there is an influence of the mother tongue to give a high evaluation to the aspect Value.

Table IX Location with Positive Value

location	Count	ratio
Barcelona_Catalonia	9450	0.2419
Florence_Tuscany	7172	0.1836
Paris_Ile_de_France	6138	0.1571
Amsterdam_Noord_Holland	5950	0.1523
Berlin	3288	0.0842
London_England	2684	0.0687
Madrid	2429	0.0622
Venice_Veneto	1304	0.0334
Frankfurt_Hesse	656	0.0168

V. Conclusion and Further Work

User's review on products and services is valuable information for both users and providers. Users' evaluation, for example, a five-point evaluation, gives a direct interpretation of users' reaction. In addition, detailed analysis can be made possible by segmentation of evaluation aspects. However, correspondence between free descriptive text and 5-step evaluation is not easily understood and remains an important subject of machine learning. Furthermore, even if a high discrimination performance is achieved by a model of machine learning, it is meaningless unless human beings can understand the model.

In this paper, we applied SVM and attribute selection to predict if a review is positive or negative for 73,589 hotel reviews in Europe with respect to seven aspects (Value, Rooms, Location, Cleanliness, Checkin, Service, Business and Overall).

For positive reviews, we achieved a hight discrimination performance 74% (F-measure) only for 20 characteristic words with respec to 6 aspects except for Buesiness. On the other hand, for negative reviews, optimum prediction performance can not be obtained unless almost all words are used, and on the average F-measure is only 27%.

The analysis in the present paper is limited to hotels in Europe. We plan to analyze the review of other areas in the future. In addition, it is necessary to analyze the language limited to English. The discovery of the characteristic difference of positive reviews and negative review should be an interesting start point that needs further investigation. We could think of several reasons to explain the phenomena. For example, the number of netative reviews is very small. Thus the imbalancedness would be one reason for the poor prediction performance. However, it does not expain the diversity. We used SVM. We wonder if other machine learning methods, for example, random forest or Naïve Bayes, conditional random fields, would yield the same result.

References

[1] S. Hirokawa, M. Okada, K. Hashimoto, “Extraction of Hints and Advice from Hotel Reviews for Improving Small Hotel Management”, Proc. CEC(IEEE 14th International Conference on Commerce and Enterprise Computing), pp.166-170, 2012

[2] N. Kaji, M. Kitsuregawa, Building “Lexicon for Sentiment Analysis from Massive HTML Documents”, Proc. EMNLP-CoNLL 2007, pp.1075-1083, 2007

[3] N. Kobayashi, K. Inui, Y. Matsumoto, K. Tateishi, “Collecting evaluative expressions for opinion extraction”, Journal of Natural Language Processing, Vol. 12, No. 3, pp. 203-222, 2005

[4] M. Okada, K. Takeuchi, K. Hashimoto, “The Effectiveness of Japanese Estimation Patterns for Automatic Classification of Costumer Reviews”, Proc. ACIS2014, pp.366-369

[5] T. Sakai, S. Hirokawa, “Feature Words that Classify Problem Sentence in Scientific Article”,Proc. iiWAS2012, pp.360-367, 2012

[6] H. Takamura, T. Inui, M. Okumura, “Extracting Semantic Orientations of Words using Spin Model”, Proc. ACL2005, pp. 13-140, 2005

[7] P. D. Turney, “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews”, Proc. 40th Annual Meeting of the Association for Computational Linguisitics(ACL), pp.417-424, 2002

[8] H. Wang, Y. Lu, C. Zhai, “Latent Aspect Rating Analysis without Aspect Keyword Supervision”, The 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'2011), pp.618-626, 2011

[9] H. Wang, Y. Lu, C. Zhai, “Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach”. The 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'2010), pp.783-792, 2010.