Aspect-Based Rating Prediction on Reviews Using Sentiment Strength Analysis

Conference Paper · June 2017

DOI: 10.1007/978-3-319-60045-1_45

Yinglin Wang
Shanghai University of Finance and Economics

Ming Wang
Shanghai University of Finance and Economics

Aspect-Based Rating Prediction on Reviews Using Sentiment Strength Analysis

Yinglin Wang[1(✉)], Yi Huang[2], and Ming Wang[1]

[1] Department of Computer Science and Technology, Shanghai University of Finance and Economics, Shanghai, China
wang.yinglin@shufe.edu.cn

[2] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
hy890916@gmail.com

Abstract. This paper aims at demonstrating sentiment strength analysis in aspect-based opinion mining. Previous works normally focused on reviewers' sentiment orientation and ignored sentiment strength that users expressed in the reviews. In order to offset this disadvantage, two methods for sentiment strength evaluation were proposed. Experiments on a huge hotel review dataset show how sentiment strength analysis can improve the performance of aspect rating prediction.

Keywords: Opinion mining · Sentiment strength analysis · Aspect-based rating

1 Introduction

With the rapid extension and advancement of internet, increasing number of people are attracted to be involved in e-commerce for its facility and convenience. And most e-commerce websites allow internet users to share their viewpoints and opinions about the products on sale with other users. The reviews, feedbacks or ratings about products play a significant role in the customer choice. Besides, sellers can improve their products or services according to feedbacks from customers.

However, reviews or feedbacks on the web are commonly long and redundant, and users usually care about some certain aspects/features of the products only. It would be tedious and fruitless to scan all of these reviews. As focusing on just the overall ratings will not be sufficient for a user to make decisions, the research of mining different aspects of the reviews is in great demand.

Aspect-based opinion mining aims to extract major aspects of a product and predict the rating of each aspect from the product reviews [1]. Hu and Liu proposed a feature-based opinion mining framework for product reviews [2]. Their research focused on the features of the product, but the features extracted from reviews are complicated and trivial. Aspects are attributes or components of products, and some similar features will be clustered into an aspect (e.g., "breakfast", "snack" in "food" aspect).

In early period, most works about aspect-based opinion mining focus on the feature extraction and aspect identification. The works on aspect-based opinion mining are

S. Benferhat et al. (Eds.): IEA/AIE 2017, Part II, LNAI 10351, pp. 439–447, 2017.
DOI: 10.1007/978-3-319-60045-1_45

feature-based approaches [1–5]. These approaches are mostly based on some rules or constraints to find high-frequency noun phrases and identify product aspects. Apparently, this type of approach would result in a loss of features expressed in low frequency. Moreover, it requires much manual tuning to filter many non-aspect features.

In order to reduce manual cost and improve the adaptability to cross-domain dataset, some works apply automatic learning method to estimating the model parameters from dataset. And most of the current models are based on Latent Dirichlet Allocation(LDA) [6]. LDA methods use the bag-of-words representation of documents and focus on the co-occurrences at the document level. However, some topics generated from the LDA methods are not valid. Additionally, when given some ratable aspects, the topics and the aspects cannot match properly.

Fine-grained aspect extracting system has recently attracted increasing attention. Titov proposed a multi-grain topic models based on LDA [3]. They solved the task in a two-step process. At first they applied LDA method on document level to infer the overall topics which were referred to as global topics. Then they proposed a sliding windows method on sentence level to infer local topics. The local topics are close to the aspects. Wang and Lu proposed a latent variable model to extract sentiments related to each aspect [7]. In this model, they assumed that reviewers first decided which aspects they commented on; and then for each aspect they chose word to express their opinion. The overall rating depends on a weighted sum of all ratings. Their model could estimate the weight and the sentiment strength for each aspect in an unsupervised way.

The main goal of sentiment analysis is to predict the sentiment orientation (*i.e.*, positive or negative) in sentences and documents. To determine which words or phrases are positive or negative, many works are based on sentiment lexicons or manual resources. Turney proposed an unsupervised learning technique based on the mutual information between phrase and the words “excellent” and “poor” [8], whose detail will be mentioned later. To assign the sentiment orientation on document level, many supervised learning techniques were used in early studies. Pang and Lee compared three machine learning methods: Naive Bayes, Maximum Entropy classification and SVM and proved that SVM outperformed the other two learning approaches [9]. Sentence level sentiment analysis is much sophisticated and linguistic knowledge are usually required to get accurate results.

Much work has been completed to build a fine-grained probabilistic model to address the problem of aspect-based opinion mining [1, 3, 4]. In contrast, few researches focus on the sentiment strength that users expressed in the reviews. In most previous work, the sentiment analysis of an aspect in one review is often considered as a binary classification problem: positive or negative, ignoring the sentiment strength. In fact, the word “*wonderful*” apparently expresses more positive sentiment than “*good*”.

Based on sentiment analysis, sentiment strength analysis further explores the emotional intensity of reviewers toward the entity they rate on, which is likely be contributive to the aspect-based opinion mining.

Rather than simply predict whether a review is positive or negative, Pang and Lee [10] proposed an algorithm based on a metric labeling formulation to predict the strength of ratings on a scale of 1 to 5. Experiments show that this algorithm improves the prediction over both multi-class and regression of SVM. And SVM regression performs slightly better than multi-class SVM because SVM regression uses ordering of classes in an implicit way.

Wilson and Wiebe accomplished some work on classifying the intensity of opinions and the subjectivity of deeply nested clause [11]. They defined three levels(low, medium and high) for the subjective sentences or clauses, which is similar to sentiment strength analysis. They employed a wide range of features including syntactic features, and achieved more satisfactory performance.

In this paper, we bridge the gap between aspect-based opinion mining and sentiment strength analysis. Comparing with the binary classification method, we propose two methods to measure the sentiment strength. The experiment demonstrate how sentiment strength analysis can improve the prediction of aspect rating. The rest of the paper is structured as follows. Section 1 describes our aspect identification and clustering method. Section 2 presents the two sentiment strength analysis methods. Section 3 shows the experiment results and according evaluation. Finally, Sect. 4 concludes the paper and provides a summary of our work and a discussion of future work.

2 Feature Identification and Aspect Extraction

Among large amount of features we can extract from the reviews, we need to classify these features into ratable aspects. As outlined in introduction, topic models, including LDA model, are commonly employed to cluster the features.

2.1 Latent Dirichlet Allocation

Latent Dirichlet allocation is a generative probabilistic model for collections of discrete data such as text corpora. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words [6].

LDA assumes the following generative process for each review in the corpus:

(1) Sample ~ *Dir*()
(2) For each of the N words w_n:
    (a) Choose a topic Z_n ~ *Multinomial*
    (b) Choose a word w_n from, $p(w_n \mid z_n)$, a multinomial probability conditioned on the topic z_n.

We apply Latent Dirichlet allocation approach on a hotel review dataset that will be used in the following experiment as well. As for deciding topic number in LDA model training stage, repeated experiments have shown that relatively higher performance can be obtained by 10 topics. Before utilizing LDA model, we removed the opinion words in the corpus to avoid unnecessary workload. An opinion word is an adjective that conveys reviewers' emotion toward an object. For instance, in the sentence "This phone has an amazing and big screen", the "screen" is the opinion target and the "amazing", "big" are opinion words for this particular review.

After feature extraction with certain methods, LDA is employed to conduct features clustering and the results are demonstrated in Table 1.

Table 1. Top 5 groups of words from LDA topics for hotel reviews

I	II	III	IV	V
Service Room Desk Size Ice	Hotel Room Booked Made Reservation	Internet Access Rooms Wedding Wireless	Noise Room Night Floor Street	Room Bed Manager Desk Staff
Breakfast Coffee Room Day Fruit	Parking Hotel Car Room Park	Beach Pool Ocean View Resort	Airport Shuttle Hotel Minutes Bus	Market Downtown Place Location Walk

The result of LDA model reveals its major drawbacks although there are some reasonable topic clusters generated. Specifically, many clusters are not understandable since the model simply formulates various clusters but not provide labels. For example, it is tricky to determine the topic of a cluster that contains *noise*, *room*, *night*, *floor* and *street*. In addition, the model automatically generates clusters without any human choice of topics, which leads to low flexibility.

2.2 Semi-automatic Aspect Extraction

A semi-automatic classification approach can effectively conquer the main shortcoming of the above full automatic method. Concretely, we adopt a bootstrapping approach similar to the aspect segmentation algorithm in Ref. [7].

The bootstrapping approach is carried out with the following steps: first of all, we manually select some keywords that are specific enough to describe the aspect as seed words (*i.e.*, *room, bed* for the *room* aspect). Then given the seed words for each aspect, we split the reviews into sentences and assign each sentence to the aspect that most terms correspond to. After that, we calculate the dependencies between aspects and words by Chi-square statistic with which we could rank the words under each aspect. We join the top words into their corresponding aspect keyword list. The above-mentioned steps will be repeated until the keyword list remains unchanged or the iteration number exceeds the limits.

We apply the bootstrapping approach to the same hotel reviews dataset. The seed words and top word list for each aspect is presented in Table 2. After filtering out stop words, the outcome seems rather reasonable, significantly outperforming the LDA approach. Additionally, given a feature word, we can easily obtain which aspect it belongs to.

Table 2. Top words from aspect segmentation

Aspect	Seed Words	Top Words
Value	Value, price, quality, worth	Cheap, cost, expectation, money, …
Room	Room, suite, view, bed	Sleep, bathroom, bed, size, …
Location	Location, traffic, minute	Market, train, site, facility, …
Cleanliness	Clean, dirty, smell, tidy	Smoke, valet, linen, maintain, …
Service	Service, food, breakfast	Restaurant, food, cafe, drink, …

3 Sentiment Strength Analysis

After feature clustering and aspect identification for each cluster, we need to distinguish opinion words so as to further analyze the precise opinion of reviewers. It is worth mentioning that attention should be paid to negation forms in a review sentence since they can reverse the final analysis result. Based on the above work, we can discuss the process of sentiment strength analysis.

There are various methods for identifying opinion words. Compared with the word distance based or part-of-speech pattern based method, the method we adopt turns out to perform relatively better by considering syntactic patterns. We first apply Stanford Parser to each sentence and then we can extract different dependencies through grammatical relations in a sentence. One type of dependency relation patterns, including adjectival modifiers (AMOD) and nominal subjects (NSUBJ), is effective to detect opinion words in adjective form. Another type containing conjunction (CONJ) prepositional modifier (PREP) is also contributive for opinion words identification when conjunctions and prepositions appear in the review sentences. In terms of negation dependency (NEG) in the sentence, we normally add a "negation" tag to the according opinion words in order to attain the necessary information expressed by reviewers.

With the opinion words identification accomplished, we can explore the sentiment strength of reviews' viewpoint by making use of two methods based on pointwise mutual information and fuzzy set. Comparison between the two methods' performance is made after the experiments.

3.1 PMI-IR Algorithm

The pointwise mutual information (PMI) between two words, *word1(w1)* and *word2(w2)*, is defined as follows:

$$PMI(w1, w2) = \log_2 \left[\frac{p(w1 \ \& \ w2)}{p(w1)p(w2)} \right] \tag{1}$$

In the equation, *p(w1 & w2)* is the probability that *word1* and *word2* co-occur. If *word1* is independent to *word2*, their mutual information value is zero. The ratio between *p(w1 & w2)* and *p(w1)p(w2)* is a measure of the degree of dependence between these two words.

In order to acquire the sentiment strength value of a phrase, the semantic orientation (SO) has defined as a computation form as follows [8]:

$$SO(phrase) \quad PMI(phrase, \text{“}excellent\text{”}) \quad PMI(phrase, \text{“}poor\text{”}) \tag{2}$$

PMI-IR estimates PMI by issuing queries to a search engine(hence the IR in PMI-IR) and returning the number of hits(matching documents). Let hits(*query*) be the number of hits returned, given the *query*, the following estimate of SO can be derived from Eqs. (1) and (2).

$$SO(\text{phrase}) = \log_2 \frac{hits(phraseNEAR\text{“}excellent\text{”})hits(\text{“}poor\text{”})}{hits(phraseNEAR\text{“}poor\text{”})hits(\text{“}excellent\text{”})} \tag{3}$$

Based on the features and their according opinion words, we can employ PMI-IR to generate the sentiment strength. The sentiment strength of some words show in Fig. 1.

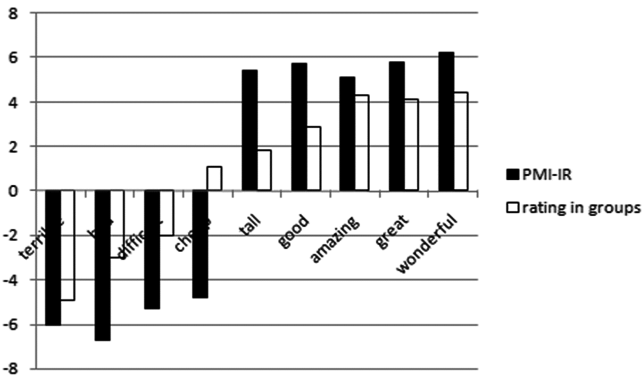

Fig. 1. Sentiment strength score calculated in PMI-IR and Rating in Groups

3.2 Rating in Groups

Another approach of computing sentiment strength of an opinion word is related to the rating of a particular review which it appears in. If an opinion word always appears in reviews with a certain rating, we could intuitively infer its sentiment strength. However, plenty of opinion words appear in reviews with varied ratings. Under this circumstance, their sentiment strength value is closely relevant to the ratio in reviews with different ratings. Based on this assumption, we propose a method to calculate opinion words' sentiment strength.

The initial step is to divide the review dataset into five groups by its overall rating, namely G1, G2, G3, G4, G5, representing a rating of 1 to 5 respectively. We extract the opinion words which are tagged as adjectives (ADJ). Then the frequency distribution among the five groups of each word is calculated. The relative frequency distribution of some frequent opinion words is presented in Table 3. As an example, most '*good*' are located in G4 and G5, which indicates its positive sentiment. By making use of the frequency distribution acquired above, we can utilize the fuzzy sets model to predict sentiment strength of an opinion word and compute the weighted sum of the group rating. Figure 1 shows the sentiment strength of some words calculated by this algorithm.

Table 3. Frequency distribution of some opinion words

Frequency distribution	G1	G2	G3	G4	G5
Good	3.9%	14.4%	12.6%	40.4%	28.8%
Quiet	5.2%	1.7%	0.0%	50.0%	43.1%
Great	2.2%	5.3%	13.0%	38.9%	40.5%
Terrible	25.0%	33.3%	8.3%	16.7%	16.7%

The major issue for now is how to assemble the sentiment strength of all the opinion words belonging to one aspect. The most straightforward approach is adopted, where we simply compute the average sentiment strength of all the opinion words describing the same aspect. Consequently, given a review and an overall rating, we can obtain all the aspect ratings by employing the several methods mentioned above.

4 Experiment Results

4.1 Dataset and Pre-processing

The hotel reviews of TripAdvisor are selected as our dataset. In addition to the free-text reviews and an overall rating, reviewers can also optionally rate predefined 8 aspects of the hotel in each review: service, value, sleep quality, location, cleanliness, room, spa, and breakfast. The ratings can be varied from 5 levels of 1 to 5 stars, which can serve as ground-truth for our aspect rating prediction.

We crawled 198,982 hotel reviews as raw dataset, with some being lack of data integrity. Specifically, among the 8 optional ratable aspects, most people would rate value, room, location, cleanliness and service. Hence, there is a need for data filtering. After that, we attain 111,443 reviews which contain all these 5 aspect ratings. In addition, some other data pre-processings including removing punctuations, stop words and stemming are conducted before we finally carry out the experiments.

4.2 Quantitative Experiments

In previous studies about aspect-based opinion mining, there has been little attention on sentiment strength conveyed by reviewers. The reason is that focusing merely on sentiment orientation can guarantee a relatively high correctness. At the same time, it becomes complicated when evaluating sentiment strength analysis. In order to explore how both sentiment orientation and sentiment strength affect performance, we design the following experiment about aspect rating prediction.

First of all, we set up two baseline methods for comparison. One, namely BASELINE, treats the overall rating given as the rating of every aspect. The other, namely Sentiment Orientation, combine the overall rating and sentiment orientation for each aspect to train a supervised model. As for improvement, we similarly take the overall rating and sentiment strength into consideration to train two models, where sentiment strength is measured by PMI-IR and Rating in groups respectively.

During the experiment, support vector regression (SVR) is chosen as learning method and implemented with LIBSVM toolkit [12]. In order to ensure the comparability of the four different methods, we adopt mean square error (MSE) for quantitative evaluation, which is illustrated in Table 4.

Table 4. Results of experiments

Experiment result	MSE
BASELINE	0.74507
Sentiment orientation	0.62771
PMI-IR	0.59514
Rating in groups	0.55206

4.3 Result Analysis

As shown in the table, transparent improvement triggered by sentiment analysis can be discovered. Concretely, the latter baseline method which takes into account sentiment orientation outperforms the former one. Further, the latter two experiments has proved that sentiment strength is more contributive to aspect rating prediction compared to sentiment orientation.

Besides, in the comparison between PMI-IR and Rating in groups, prediction of the latter is comparatively accurate, about which we can find a hint from Fig. 1. The sentiment strength measured by PMI-IR tends to polarize. In contrast, the strength measured by Rating in Groups is rather refined. However, it should be admitted that the dataset has affected the result to some extent.

5 Conclusions

In this paper, we explore the contribution of sentiment strength analysis to aspect rating prediction with two proposed methods. Experiments reveal that sentiment strength plays an important role in enhancing the performance of aspect rating prediction, despite these two methods provide slightly varied outcomes.

There is space for further improvement. It is probably influential to fully take advantage of those long and complicated sentences with pronoun in it. Besides, some semantic parser methods might also be valuable for this work.

Acknowledgements. We would like to thank the National Natural Science Foundation of China (Grant No. 61375053) for part of the financial support of this paper.

References

1. Moghaddam, S., Ester, M.: On the design of LDA models for aspect-based opinion mining. In: 21st ACM International Conference on Information and Knowledge Management, Maui, Hawaii, USA, pp. 803–812. ACM Press, New York (2012)
2. Hu, M.Q., Liu, B.: Mining and summarizing customer reviews. In: 10th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Seattle, WA, USA, pp. 168–177. ACM Press, New York (2004)
3. Titov, I., McDonald, R.: Modeling online reviews with multi-grain topic models. In: 17th International Conference on World Wide Web, Beijing, China, pp. 111–120. ACM Press, New York (2008)
4. Moghaddam, S., Ester, M.: The FLDA model for aspect-based opinion mining: addressing the cold start problem. In: 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil, pp. 909–918, ACM Press, New York (2013)
5. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: 14th International Conference on Machine Learning, Nashville, TN, USA, pp. 412–420, Morgan Kaufmann Publishers Inc., San Francisco (1997)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
7. Wang, H., Lu, Y., Zhai, C.: Latent aspect rating analysis on review text data: a rating regression approach. In: 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, USA, pp. 783–792, ACM Press, New York (2010)
8. Turney, P.D.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: 40th annual meeting on association for computational linguistics, Philadelphia, PA, USA, pp. 417–424, Association for Computational Linguistics, Stroudsburg (2002)
9. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: ACL-02 Conference on Empirical Methods in Natural Language Processing, EMNLP 2002, Philadelphia, PA, USA, pp. 79–86. Association for Computational Linguistics, Stroudsburg (2002)
10. Pang, B., Lee, L.: Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: 43rd Annual Meeting on Association for Computational Linguistics, Michigan, USA, pp. 115–124. Association for Computational Linguistics, Stroudsburg (2005)
11. Wilson, T., Wiebe, J., Hwa, R.: Recognizing strong and weak opinion clauses. Comput. Intell. **22**(2), 73–99 (2006)
12. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) **2**(3), 389–396 (2011)