Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2014 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

2014

THE IDENTIFICATION OF NOTEWORTHY HOTEL REVIEWS FOR HOTEL MANAGEMENT

San-Yih Hwang
National Sun Yat-sen University, syhwang@mis.nsysu.edu.tw

Chia-Yu Lai
National Sun Yat-Sen University, chiayu06@gmail.com

Jia-Jhe Jiang
National Sun Yat-Sen University, js1233211234567@gmail.com

Shanlin Chang
National Sun Yat-Sen University, d004020002@student.nsysu.edu.tw

THE IDENTIFICATION OF NOTEWORTHY HOTEL REVIEWS FOR HOTEL MANAGEMENT

San-Yih Hwang, Department of Information Management, National Sun Yat-sen University, Kaohsiung, Taiwan, syhwang@mis.nsysu.edu.tw

Chia-Yu Lai, Department of Information Management, National Sun Yat-sen University, Kaohsiung, Taiwan, chiayu06@ gmail.com

Jia-Jhe Jiang, Department of Information Management, National Sun Yat-sen University, Kaohsiung, Taiwan , js1233211234567@gmail.com

Shanlin Chang, Department of Information Management, National Sun Yat-sen University, Kaohsiung, Taiwan, d004020002@student.nsysu.edu.tw

Abstract

The rapid emergence of user-generated content (UGC) inspires knowledge sharing among Internet users. A good example is the well-known travel site TripAdvisor.com, which enables users to share their experiences and express their opinions on attractions, accommodations, restaurants, etc. The UGC about travel provide precious information to the users as well as staff in travel industry. In particular, how to identify reviews that are noteworthy for hotel management is critical to the success of hotels in the competitive travel industry. We have employed two hotel managers to conduct an examination on Taiwan's hotel reviews in Tripadvisor.com and found that noteworthy reviews can be characterized by their content features, sentiments, and review qualities. Through the experiments using tripadvisor.com data, we find that all three types of features are important in identifying noteworthy hotel reviews. Specifically, content features are shown to have the most impact, followed by sentiments and review qualities. With respect to the various methods for representing content features, LDA method achieves comparable performance to TF-IDF method with higher recall and much fewer features.

Keywords: Review recommendation, Text mining, Topic model, Sentiment analysis.

1 INTRODUCTION

With the rapid expansion and proliferation of web 2.0 technologies, the advent of user-generated content (UGC) has marked a shift for individuals to create their own content and share their knowledge online. The new style of content sharing enables Internet users to become self-publishing consumers and share their information with others (Raman 2009, Sigala 2008). Online customer reviews are regarded as electronic word of mouth (eWOM) and have been found to have a significant impact on product sales and consumer purchase decision (Duan et al. 2008) . Furthermore, several studies have shown that online users' reviews are particularly important for experience goods, as their qualities are usually hard to determine before consumption (Klein 1998, Zhu & Zhang 2006). Most services and products offered by hotels and travel industry are experience goods. In this context, previous studies had shown the empirical evidence that online reviews play a key role in hotel selection and trip planning (Sparks et al. 2013, Ye et al. 2009).

Online customer reviews on a hotel, especially those with critical comments, may significantly impact its reputation, and subsequently sales. From a business perspective, obtaining a better understanding of how online reviews influence consumer purchasing decision is critical (Ye et al. 2009). Within the context of the hotel industry, many hotel staff now take an active role online by posting their responses to hotel reviews. However, according to a recent survey (Sparks et al. 2013), only 7% of hotels are replying to reviews even though 71% of customers think that a management response has significant influence. Moreover, negative reviews may easily damage the image of the hotels to people, and hotel staffs need to respond to these reviews so as to offset the negative emotions of these tourists in the hope to restorethe hotel's reputation. Therefore, how to quickly identify reviews that may potentially influence the hotel's business performance is important to hotel staff.

However, customer reviews have influence not just on the hotel sales. Customer reviews, if properly utilized, may prove beneficial for the operations of a business in an ever-changing competitive environment. In this paper, we intend to propose a method to automatically identify customer reviews that are noteworthy to hotel managers. To do so, we conducted a preliminary study that involves an interview with several hotel managers, and they reported that noteworthy reviews are always subject to subsequent actions. For example, reviews with negative comments need to be addressed because their spread may harm the reputations of the hotel. However, not every negative review deserves special attention. Negative reviews with reasonable writing are particularly harmful. Some positive reviews may also be noteworthy, especially when they pinpoint some unexpected services that surprise yet please the customers. Such reviews may be used to inspire the employees and motivate them to propose innovative services. Finally, some reviews may provide useful suggestions to the hotel, which may help shape future business strategies of the hotel.

Our study aims to identify the features that are relevant to the noteworthiness of hotel reviews with respect to the hotel management. There have been quite some works proposed for recommending hotels or hotel reviews in the literature (Ghose et al. 2012, Levi et al. 2012, O'Mahony & Smyth 2009). However, their targets of recommendation are mostly customers, not hotel staff. A hotel review that is helpful to prospective customers may or may not concern the hotel staff. For example, a hotel review that describes how much a customer enjoyed or hated the beach that is close to the hotel could be useful for someone who seeks to relax in a hotel near beach but may not be so much concern for the hotel staff because hotel location cannot be easily changed. Ghose and Ipeirotis (2011) explore the factors that affect the sales of various types of products using data from Amazon.com and conclude that features in subjectivity, readability, and informativeness have different degree of impact on different types of products. However, product sales is not the only concern for the hotel staff, and how to identify reviews that contribute or hinder product sales is not explored in their work.

The preliminary study leads us to conjecture that three types of aspects, namely content, sentiment and review quality may affect the noteworthiness of reviews from the perspective of hotel staff. This paper proposes several methods to represent the three types of aspects and identify those that can accurately identify noteworthy hotel reviews for hotel staff. Through the experiments using tripadvisor.com data, we find that all three types of features are important in identifying noteworthy hotel reviews. Specifically, content features are shown to have the most impact, followed by sentiments and review qualities. With respect to the various methods for representing content features, LDA method achieves comparable performance to TF-IDF method with higher recall and much fewer features.

The remainder of this paper is organized as follows. In the next section, we review techniques about content feature identification, sentiment analysis, review quality determination, and hotel and review recommendation. In the third section, we describe our approach to identifying noteworthy hotel reviews. In the fourth section, we examine the empirical data to evaluate our approach. Finally, we conclude with the result of our work and give directions for future research.

2 RELATED WORK

In this section, we firstly review the techniques for the identification of content features from a massive set of documents. Then we present the methods about how to determine the sentiment of a review and the quality of documents. Finally, we describe relevant works in recommending hotels and hotel reviews.

2.1 Content Feature Identification

TF-IDF, term frequency–inverse document frequency, is a typical approach to representing textual features of documents (Chowdhury 2010). The idea is that if a word or a phrase appears in a document with high frequency (term frequency) yet rarely appears in other documents (inverse document frequency), this word or phrase is a good indicator for identifying this document. Let N be the total number of documents and $n_{i,j}$ be the number of times a keyword k_i appears in a document d_j. $TF_{i,j}$ measures the term frequency of keyword k_i in document d_j, as shown below.

$$TF_{i,j} = \frac{n_{ij}}{\sum_{k_l \text{ is a keyword in } d_j} n_{lj}}$$

If a keyword appears in many documents, its importance will decline. $IDF_i = \log\frac{N}{N_i}$ is used to measure the inverse document frequency, where N_i is the number of documents in which keyword k_i appears. The TF-IDF weight for keyword k_i in document d_j, $w_{i,j}$ is then defined as:

$$w_{i,j} = TF_{i,j} \times IDF_i$$

In addition to its basic form shown above, there are several variations about TF-IDF (Chowdhury 2010). As can be imagined, a large number of TF-IDF features (usually thousands) will be needed for representing a massive set of documents. Another approach for concisely representing documents is to use a small number of latent content features, or called topics. Several topic models have recently been proposed to identify a small number of topics inherent in a set of documents. Each document can be subsequently represented as a topic vector. Latent Dirichlet Allocation (LDA) is the most commonly used approach for deriving the topic model (Blei et al. 2003). LDA techniques, such as Gibbs sampling

(Griffiths 2002, Griffiths & Steyvers 2004), take as input a collection of documents, each represented as a bag of words, and produces two kinds of probability distributions: topic probability distributions, one for each document, and word probability distributions, one for each topic. Two parameters, namely α and β, can be set to adjust the concentration parameters of the Dirichlet prior distributions for topic probabilities and word probabilities respectively.

While TF-IDF and LDA model are both powerful techniques in representing documents by their linguistic forms, semantic information is not considered and problems such as synonyms and homonyms may arise. WordNet is a large lexical database of English that describes the mappings between words and senses (i.e., meanings) and the relationships among senses (e.g., is-a relationships, similar senses, etc). Since a word with multiple meanings can be confusing, a method is needed to identify the particular sense for each word appeared in a sentence. This problem is called the word-sense disambiguation (WSD), and there have been many methods proposed for WSD problem (Navigli 2009).

2.2 Sentiment Analysis

Sentiment analysis, also called polarity recognition or opinion mining, aims to determine the sentiment of a document, a sentence, or an entity, being positive, negative, or neutral. According to Pang and Lee (2008), there are generally two approaches for detecting sentiments: supervised approach and unsupervised approach. The supervised approach for determining the sentiment of a review starts by representing a review as a feature vector, e.g., TF-IDF, and builds a classifier using a training data set. There have been many methods that are devoted to the identification of text features relevant to sentiment (Pang & Lee 2008). Several methods have been proposed in the literature for determining the polarity of a review using unsupervised approach (Dave et al. 2003, Taboada et al. 2011, Turney 2002). These methods generally prepare a domain-specific sentiment lexicon and identify a number of linguistic constructs commonly used to express sentiments on certain aspects of products. The sentiment of a sentence is determined by looking at its linguistic constructs and the appeared sentiment word(s). Polarity of a review is an aggregation of the sentiments of its constituent sentences. In Carrillo-de-Albornoz et al. (2010), the authors describe a hybrid approach, in which sentences are first converted into senses based on WordNet using Lesk algorithm (Lesk 1986). By referring to WordNet Affect, which includes senses pertaining to 16 emotions: joy, love, liking, calmness, positive-expectation, hope, fear, sadness, dislike, shame, compassion, despair, anxiety, surprise, ambiguous-agitation and ambiguous-expectation, the emotions pertaining in each review is represented as a 16-tuple, where each element represents the weight of the corresponding emotion (Baccianella et al. 2010). Movie reviews in the corpus are tagged positive or negative polarity, and some machine learning algorithm can be used to train a classifier. The classifier can then be used to classify the polarity (positive or negative) for each incoming review based on its 16 emotional attribute values.

2.3 Quality of Product Review

In Liu et al. (2007), the quality of product reviews is determined by several features, namely sentence level informativeness, word level informativeness, and product feature level informativeness. Sentence level informativeness refers to the number of sentences, the average length of the sentences and the number of the sentence with desired product feature. Word level informativeness refers to the number of words, the number of product names, and the number of brand names. Note that the number of product names and brand names are needed for product reviews but may not be applicable for hotel reviews. In addition, the reputation of the reviewer who wrote a given review is considered as a good indicator about the review quality (Ghose & Ipeirotis 2011, Ghose et al. 2012). Example features for reputation of a reviewer include the number and the average helpfulness score of the reviews s/he has written. Some studies in the past verified quality of reviews by classifying review features based on the readability of the

text, the reputation of the reviewer, the star rating of the review, and various content features based on the review terms (Liu et al. 2007, O'Mahony & Smyth 2009).

2.4 Hotel Recommendation

In commercial applications, hotel recommendations are typically based on hotel ratings given by users. In Adomavicius and Kwon (2007), a regression model is adopted to aggregate ratings on various aspects into a single rating, where aspect ratings for unseen items are predicted using collaborative filtering. Similar approach is adopted by Fuchs and Zanker (2012) for recommending hotels using TripAdvisor data, and they further explore the impact of regression models on different customer segments and exploit penalty-reward-performance model. Jannach et al.(2012) extends the model proposed in Adomavicius and Kwon (2007) by incorporating item-based collaborative filtering, more regression models, and aspects selection. This line of research does not make use of the textual data of hotel reviews.

Ghose et al.(2012) define a measure called *utility gain* for the economic impact of a hotel by considering consumer heterogeneity, hotel characteristics, as well as UGC pertaining to the hotels. It shows that UGC variables, such as text features, subjectivity, and readability, significantly affect the model's predictive power for utility gain. Levi et al.(2012) propose a context-based method for personalized recommendation of hotels based on hotels' reviews and the reviewers' contextual information. Three types of context are identified, namely travel intent, nationality, and preference. Then for each context group, nouns that frequently appeared in the relevant reviews are collected and form a lexicon. A user who seeks hotel recommendation is asked to provide her intent, nationality, and intent, and lexicons of the corresponding context groups are regarded as the traits of the user. Hotels whose reviews contain high positive sentiment on these traits will be recommended. They use data collected from Tripadvisor.com and Venere.com and conduct user study, and the results show 20% higher satisfaction rate than the ratings-based recommendation.

2.5 Hotel Review Recommendation

As shown in previous study, reviews of a hotel play an important role in deciding whether or not to recommend this hotel. However, some reviews are deemed better than the other and more helpful when it comes to decision making. In the past few years, we have seen quite a few works that intend to predict the helpfulness of an incoming review. Product review recommendation was first proposed in the work of (O'Mahony & Smyth 2009, O'Mahony & Smyth 2010). They adopt a supervised learning approach by considering four types of features, namely reputation, content, social, and sentiment. However, for each type of features, relatively simple methods are used for defining sub-features. For example, they use user-supplied rating for determining sentiment, and only linguistic sub-features, such as number of terms and the ratio of upper and lower characters, are used to content feature. Experiments using hotel reviews from TripAdvisor shows reasonable recommendation result. This work is further enhanced by incorporating feature selection and exercising various classification schemes (O'Mahony & Smyth 2010). In Dong et al. (2013), a supervised learning method for identifying helpful reviews is proposed by taking basic features such as age, rating, readability, as well as product features and sentiment features. It is shown based on reviews for various product categories from Amazon.com that both the product features and sentiments expressed in the review are important factors for identifying helpful reviews. In Ghose and Ipeirotis (2011), the authors propose a regression model to predict the helpfulness of a given review by considering readability, subjectivity, and reviewer's reputations. Experiments using Amazon.com's data show that the impact of the three types of features on predicting the usefulness of a review or a product's sale vary across different product types.

Considering the fact that the helpfulness of a review may differ across users, Musat et al. (2013) propose to derive the interest topic profile of a user based on the reviews she wrote. The interest topic profile is subsequently used to filter out less relevant reviews and generate personalized ratings for hotels. Moghaddam et al. (2011) propose a matrix factorization approach for personalized recommendation of product reviews based on review rating data. They have shown from experiments using Epinion.com data that their proposed methods perform better than other non-personalized, textual/social features-based methods. However, all the above mentioned works that intend to recommend either hotels or hotel reviews target at consumers, rather than hotel staff as focused in our work.

3 THE APPROACH

To recommend noteworthy reviews, we extract content features from reviews, measure the sentiment pertaining to each sentence and consider review quality. We then go on to establish a classification model for recommendation, where each review is characterized by three descriptive vectors and a class labeled as "noteworthy" or "no noteworthy". These tasks are processed in sequence and are described in the following subsections

3.1 Content Features

In this work, we consider three methods for representing content features of a review. The baseline method is TF-IDF, denoted T0, which represents each review as a vector of words. We choose approximately 4000 words with the highest TF-IDF values as the content features as it yielded the best performance in our preliminary experiments. The second method is a topic model-based method, denoted T1. It first regards each review as a bag of words by excluding stop words and then applies LDA to generate a topic model. The set of topics are then treated as the content features, and each review is represented as a vector of topics. The third method, denoted T2, uses semantic-based LDA, which utilizes semantic information in the text for determining topics. Specifically, we use Stanford Parser (Klein & Manning 2003) to process each sentence of a review and identify the part-of-speech (POS) of each word. We then extract all the nouns, verbs, adjectives, and adverbs and exclude stop words. The senses of the extracted words in WordNet are determined using the graph based WSD, UKB (Agirre & Soroa 2009). Note that these senses can be further extended by including their hypernyms, hyponyms, or similar senses (line 9 in Figure 1). Finally, each review is represented as a bag of senses, and Latent Dirichlet Allocation (LDA) is used for inferring the topic model. The topic model construction algorithm is shown in Figure 1:

```
Algorithms ：Semantic Topic Model Construction
Input : A set of reviews S
Output: A Topic Inference Mode I
1.    S=Ø
2.    For each review r do
3.       C(r) = Ø
4.       For each sentence s in r do
5.            Use Natural Language Parser(NLP) to identify all nouns, verbs, adjectives, and adverbs in s that are not stopwords.
6.            Use some Word Sense Disambiguation(WSD) technique to find all senses of these terms.
7.            Add theses sense to C(r)
8.            S=S ∪ {C(r)}
9.            Extend S
10.  Apply LDA on S to find the topic inference model I.
```

Figure1. Algorithm for semantics-based topic model construction

For a newly arrived review, we can apply the constructed topic model to infer its topic vector, and the algorithm is shown in Figure 2.

```
Algorithm: Topic Vector Extraction
Input: A review r and a topic inference model I
Output: A topic vector
1.  C(r) = Ø
2.  For each sentence s in r do
3.      Use Natural Language Parser(NLP) to identify all nouns, verbs, adjectives, and adverbs  in s that are not stopwords
4.      Use Word Sense Disambiguation(WSD) techniques to  find all senses of these terms
5.      Add theses senses to C(r)
6.      Extend(C(r))
7.  Apply I to C(r) and return r's topic vector
```

Figure2. Algorithm for semantics-based topic vector determination

3.2 Sentiment Features

Before determining the sentiment of a given review, we need to first identify the sentiments for its sentences. Due to the lack of sentiment training data set, we adopt the unsupervised approach for determining review sentiment. The sentiment lexicon we use in this work is SentiWordNet 3.0 (Baccianella et al. 2010), an extension of WordNet by incorporating emotion values to senses. In SentiWordNet, there are three types of emotions, namely positivity, objectivity, and negativity, and each sense has an emotional value for each type of emotion in the range of [0, 1]. Sentiment terms accompanied by negation words have to be carefully addressed because their sentiments may become opposite (e.g., "not bad" has an opposite sentiment to "bad"). We use Stanford Parser to find the phrase structure tree for each sentence to delimiting the scope of negation, if any (Carrillo-de-Albornoz et al. 2012). We interchange positivity with negativity for each sense in the scope of negation, and their values are multiplied by 0.9 by following the work proposed in (Carrillo-de-Albornoz et al. 2010). For example, "not good" is usually considered less negative than "bad". Finally, we sum the emotional values of every sense in the sentences of a review to determine the overall sentiment score. Note that in our work, the sentiment score of a review is a pair of emotional values for positivity and negativity. The entire algorithm for determining sentiment of a review is shown in Figure 3.

```
Algorithm:Sentiment Score Detection
Input: a review r
Output: sentiment score (positivity, negativity)
1.  P(r) = 0 , N(r) = 0
2.  For each sentence s in r do
3.      Use Natural Language Parser(NLP) to identify all nouns, verbs, adjectives and adverbs in s
4.      Use word sense Disambiguation(WSD) technique to find the set C f all senses of these terms
5.      Find the positive sentiment score P(c) and negative sentiment score N(c) for each c∈C using SentiWordNet3.0 database
6.      Use Stanford Parser to find phrase structure tree in s
7.      For each negation cue in s do
8.          Identify the scope p of negation according to phrase structure tree
9.          For each c∈C in p do
10.             tmp = P(c)
11.             P(c) = 0.9 * N(c)
12.             N(c) = tmp * 0.9
13.     P(s) = ∑c∈C P(c)
14.     N(s) = ∑c∈C N(c)
15.     P(r) = P(r) + P(s); N(r) = N(r) + N(s)
16. Return (P(r)/(P(r) + N(r)), N(r)/(P(r) + N(r)))
```

Figure 3. Algorithm for sentiment score detection

3.3 Quality Features

In this section, we describe our features about the quality of the reviews. The quality features in previous studies can be divided into three categories: sentence level, word level, and user reputation (Cheung et al. 2009, Liu et al. 2007, O'Mahony & Smyth 2009). In this work, we carefully consider each category and identify features that are suitable in our context (i.e., hotel reviews). The following lists the quality features used in this work:

- NumSent: the number of the sentences
- LenSent: the average length of sentences
- NumEmoSent: the number of sentences with non-zero sentiment scores
- NumWord: the total number of words
- NumReview: the number of authoring reviews that is the number of reviews authored by the reviewer.
- MeanHelpReview: mean review helpfulness, which is the mean review helpfulness over all reviews authored by the reviewer.
- STDHelpReview: review helpfulness deviation, which is the standard deviation of review helpfulness over all reviews authored by the reviewer.

3.4 Classification model Construction

As a result, each review can be represented by a vector of words/topics, a sentiment score, and a vector of review qualities. Each review in the training data set is labeled as noteworthy or not-noteworthy. Figure 4 shows the structure of the training data. A classification method can be adopted for training a binary classifier using the training data. Various classification methods will be executed and compared.

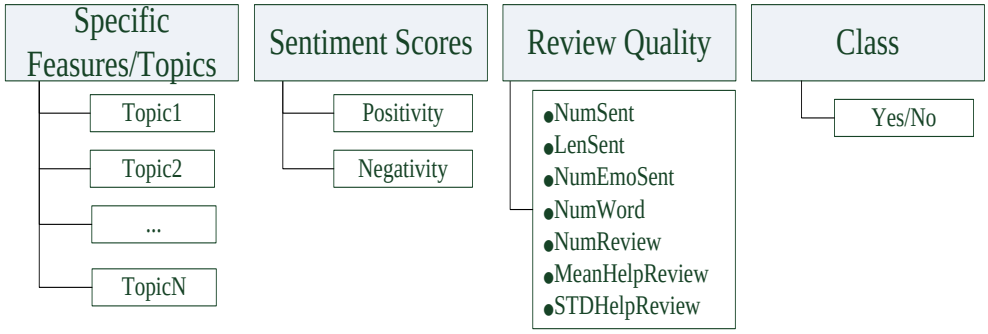


Figure 4. Representations of reviews

4 EMPIRICAL EVALUATION

In this section, we will first describe the data used in our experiments. We then explain the experimental design and the performance metrics. Finally we discuss the experimental results.

4.1 Data collection

We collected data from Tripadvisor.com (http://www.tripadvisor.com/), which is a travel site that provides objective and impartial evaluation of hotels, restaurant recommendations, B&B Reviews, membership information, and travel guides. TripAdvisor, established in 2000, is a pioneer that hosts UGC in tourism. In our work, we focus on hotels and B&Bs reviews. As TripAdvisor is an internationally renowned travel site, many foreign tourists use it to express their opionions and gain feedback. Thus, this study focuses on

foreign tourists, and English reviews for the top ten hotels, up to May, 2013, in each of the following cities in Taiwan: Taipei, Kaohsiung, Taichung, New Taipei City, Hualien, Nantou, and Ilan, are collected.. As a result, we collected 3124 hotel reviews that comprise 28,088 sentences. These 3124 reviews were contributed by 2623 authors. These authors also wrote totally 54,746 posts. Each review contains the hotel name, review author, review title, overall rating, review date, review URL, accommodation types (e.g., travel accommodations with family or business accommodation), value rating, location rating, sleep rating, comfort rating, cleaning rating, service rating, review respondents, replying date and replying content, in addition to review content.

4.2 Training data set

To construct a review noteworthiness prediction model, we need a training data set. To prepare the training data set, we first retrieved some 500 reviews that are diversified in their emotional polarities. Specifically, we chose 179 reviews for each of the following classes: highest positivity, lowest positivity, highest negativity, and lowest negativity, as determined by our sentiment detection method described in Section 3.2. After excluding duplicate reviews, we finally obtained 501 hotel reviews. We gave these reviews to two experts, who are senior managers of renowned hotels in Taiwan. The two experts have been in travel industry for more than six years. Their manual classification is to determine the reviews to be noteworthy or not. Finally, we chose 386 reviews which have the same label classified by the two experts.

4.3 Parameter Settings

In LDA, documents are generated by first picking a Dirichlett probability distribution Dir(α) for generating topics, and then, for each topic, a Dirichlett probability distribution Dir(β) is chosen for generating words. Here α (β) is a hyperparameter specifying the skewness on the topic (word) distribution (Griffiths & Steyvers 2004, Hofmann 2001). Smaller α (β) indicates a bias towards sparsity and results in picking topic (word) distributions favouring just a few topics (words) per document (topic). Thus, based on previous research (Griffiths & Steyvers 2004, Steyvers & Griffiths 2007) , we set the parameters of LDA as follows: α= 1 and β=0.1. α= 1 yields a uniform distribution over a small number of topics. For β=0.1, it is intended that each topic is associated with only a relatively small number of terms out of 4000 terms (Blei et al. 2010). In addition, we compute the perplexity at different number of topics (Blei et al. 2003), and it was found that with topic number being 25, we are able to achieve the lowest (and best) perplexity.

We exercised several classification techniques, and SVM exhibited the best results. In the following, we present our performance results running using SVM. With respect to the parameter settings of SVM, we keep the default value and use Platt's Sequential Minimal Optimization (SMO) algorithm for training a support vector classifier. We perform 10-fold cross-validation and use average precision and recall on noteworthy reviews as the performance measures. F-measure also serves as a combinational measure. Precision, recall and F-measure are defined as follows, where TP, FP, and FN denote true positive, false positive and false negative respectively:

$$\text{Precision} = \frac{TP}{TP+FP}, \text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F-measure}=2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$$

4.4 Preliminary Experiment

Figure 5 shows the F-measure for T1+S+Q under different number of topics, where T1, S and Q denote the incorporations of topic features, sentiment features and quality features respectively. The result is consistent with our previous experiment using perplexity as the measure in that 25 topics indeed yield the best performance. In our subsequent experiments, we fix the number of topics at 25.

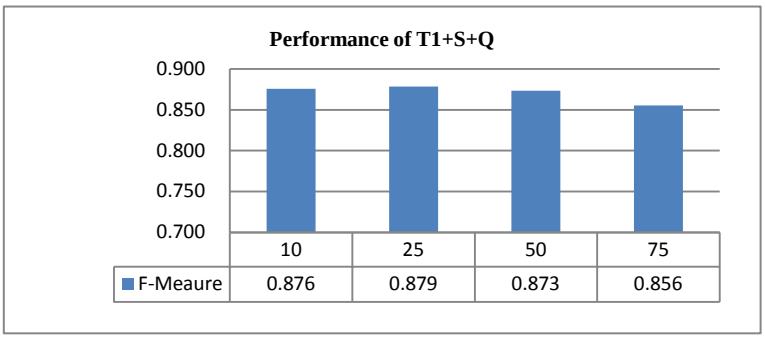


Figure5. F-measure of T1+S+Q under different numbers of topics

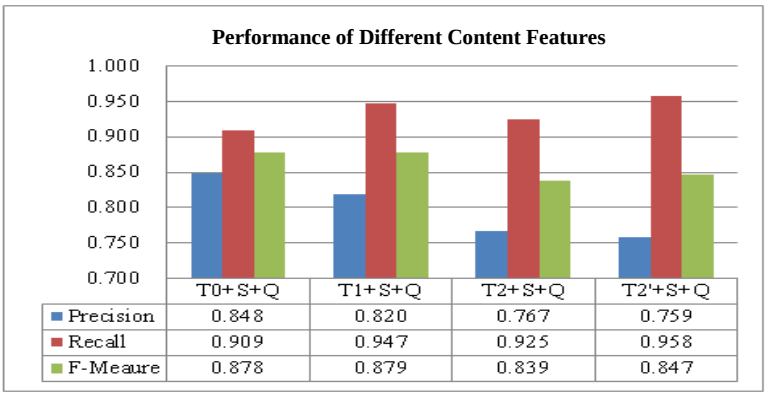


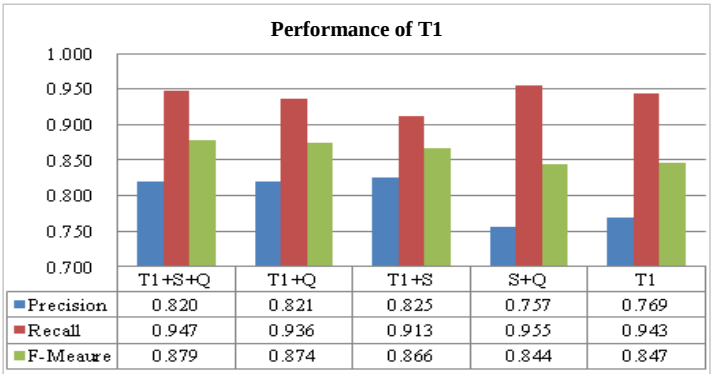
Figure 6. Performance of different methods for representing content features

4.5 Performance Result

In Section 3.1, we have described three methods for representing content features, namely T0, T1, and T2. T2 is a semantic-based LDA method, where the identified senses may be extended by hypernyms, hyponyms, or similar senses. To evaluate the effect of the sense extension, we name the methods with and without the sense extension as T2 and T2' respectively.

Our next experiment compares the performance of different methods for specifying content features, namely T0, T1, T2, and T2', and the result is shown in Figure 6. It can be seen that T0 and T1 have higher precision and F-measure values. In contrast, the semantic-based methods (T2 and T2') have high recall values but lower precision values. We observed that the semantics-based methods tend to mistakenly predict not-noteworthy reviews as "noteworthy." After looking closely at the high frequency words for topics in T1, we find that quite a few proper names do not show up in WordNet, a general-purpose ontology. For example "Kaohsiung" (place name) or "Hi-Lai"(hotel name) do not appear in WordNet. We thus attribute the poor precision values of the semantic-based methods to the lack to tourism-specific concepts in the general ontology such as WordNet. In addition, T2' is slightly better than T2, which shows

that sense extension improves the performance of the semantic-based method, though the extent of improvement is small. Comparing T0 and T1, T1 has higher recall yet lower precision, and their F-measure values are comparable. For the identification of noteworthy reviews, however, recall is deemed more important than precision as missing a noteworthy review could cause drastic damage to the hotel. Besides, T1 utilizes only 25 content features, in comparison with 4000 TF-IDF features used in T0. Thus, we conclude that T1 is a promising method for representing content features of hotel reviews.



	T1+S+Q	T1+Q	T1+S	S+Q	T1
Precision	0.820	0.821	0.825	0.757	0.769
Recall	0.947	0.936	0.913	0.955	0.943
F-Meaure	0.879	0.874	0.866	0.844	0.847

Figure7. Performance of different combinations of T1, S, and Q

5 CONCLUSIONS

Figure 7 displays different combinations of T1, the sentiment features (denoted S), and the quality features (denoted Q). As can be seen, the full combination, namely T1+S+Q achieves the best performance. Comparing T1, S, and Q, we find that the content feature (T1) is most important and achieves the performance comparable to the combination of sentiment features and quality features (S + Q). S and Q are both important because by excluding either one, the performance deteriorates to some degree.

In this paper, we have proposed an effective approach to identifying hotel reviews that are noteworthy for hotel staff. Our approach makes use of three types of features, namely content features, sentiment features, and quality features to discover the noteworthy reviews, and develop several methods for deriving these features. Through the experiments using tripadvisor.com data, we found that all the three types of features are important in predicting noteworthy hotel reviews. Specifically, content features have been shown to have most impact, followed by sentiment and quality. For deriving content features, we have proposed three methods. It has been shown that the LDA method achieves comparable performance to TF-IDF method with higher recall and much fewer features. The semantic-based LDA method achieves comparable recall as the (word-based) LDA method yet has lower precision. We attribute the lower precision of the semantic-based method to the general-purpose ontology, namely WordNet, used by our method, which excludes quite a few proper names in tourism domain. In the future, we plan to exercise the semantic-based method by incorporating more domain-specific ontology.

References

Adomavicius, G and Kwon, Y. (2007) New recommendation techniques for multicriteria rating systems. *Intelligent Systems, IEEE* 22(3):48-55.

Agirre, E and Soroa, A, (2009) Personalizing pagerank for word sense disambiguation, In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, p. 33-41.

Baccianella, S, Esuli, A and Sebastiani, F, (2010) Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, In: *LREC*, p. 2200-2204.

Blei, DM, Griffiths, TL and Jordan, MI. (2010) The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)* 57(2):7.

Blei, DM, Ng, AY and Jordan, MI. (2003) Latent dirichlet allocation. *the Journal of machine Learning research* 3):993-1022.

Carrillo-de-Albornoz, J, Plaza, L, Díaz, A and Ballesteros, M, (2012) Ucm-i: A rule-based syntactic approach for resolving the scope of negation, In: *In proceedings of the *SEM 2012 Shared Task: Resolving the Scope and Focus of Negation*, Association for Computational Linguistics, p. 282-287.

Carrillo-de-Albornoz, J, Plaza, L and Gervás, P, (2010) A hybrid approach to emotional sentence polarity and intensity classification, In: *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, Association for Computational Linguistics, p. 153-161.

Cheung, MY, Luo, C, Sia, CL and Chen, H. (2009) Credibility of electronic word-of-mouth: Informational and normative determinants of on-line consumer recommendations. *International Journal of Electronic Commerce* 13(4):9-38.

Chowdhury, G, (2010) Introduction to modern information retrieval, Facet publishing,

Dave, K, Lawrence, S and Pennock, DM, (2003) Mining the peanut gallery: Opinion extraction and semantic classification of product reviews, In: *Proceedings of the 12th international conference on World Wide Web*, ACM, p. 519-528.

Dong, R, Schaal, M, O'Mahony, MP and Smyth, B, (2013) Topic extraction from online reviews for classification and recommendation, In: *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, AAAI Press, p. 1310-1316.

Duan, W, Gu, B and Whinston, AB. (2008) Do online reviews matter?—an empirical investigation of panel data. *Decision Support Systems* 45(4):1007-1016.

Fuchs, M and Zanker, M (2012) Multi-criteria ratings for recommender systems: An empirical analysis in the tourism domain, In: *E-commerce and web technologies*, p. 100-111, Springer.

Ghose, A and Ipeirotis, PG. (2011) Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *Knowledge and Data Engineering, IEEE Transactions on* 23(10):1498-1512.

Ghose, A, Ipeirotis, PG and Li, B. (2012) Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Science* 31(3):493-520.

Griffiths, T. (2002) Gibbs sampling in the generative model of latent dirichlet allocation. *Standford University* 518(11):1-3.

Griffiths, TL and Steyvers, M. (2004) Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America* 101(Suppl 1):5228-5235.

Hofmann, T. (2001) Unsupervised learning by probabilistic latent semantic analysis. *Machine learning* 42(1-2):177-196.

Jannach, D, Gedikli, F, Karakaya, Z and Juwig, O (2012) Recommending hotels based on multi-dimensional customer ratings, In: *Information and communication technologies in tourism 2012*, p. 320-331, Springer.

Klein, D and Manning, CD, (2003) Accurate unlexicalized parsing, In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, Association for Computational Linguistics, p. 423-430.

Klein, LR. (1998) Evaluating the potential of interactive media through a new lens: Search versus experience goods. *Journal of business research* 41(3):195-203.

Lesk, M, (1986) Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone, In: *Proceedings of the 5th annual international conference on Systems documentation*, ACM, p. 24-26.

Levi, A, Mokryn, O, Diot, C and Taft, N, (2012) Finding a needle in a haystack of reviews: Cold start context-based hotel recommender system, In: *Proceedings of the sixth ACM conference on Recommender systems*, ACM, p. 115-122.

Liu, J, Cao, Y, Lin, C-Y, Huang, Y and Zhou, M, (2007) Low-quality product review detection in opinion summarization, In: *EMNLP-CoNLL*, p. 334-342.

Moghaddam, S, Jamali, M and Ester, M, (2011) Review recommendation: Personalized prediction of the quality of online reviews, In: *Proceedings of the 20th ACM international conference on Information and knowledge management*, ACM, p. 2249-2252.

Musat, C-C, Liang, Y and Faltings, B, (2013) Recommendation using textual opinions, In: *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, AAAI Press, p. 2684-2690.

Navigli, R. (2009) Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)* 41(2):10.

O'Mahony, MP and Smyth, B, (2009) Learning to recommend helpful hotel reviews, In: *Proceedings of the third ACM conference on Recommender systems*, ACM, p. 305-308.

O'Mahony, MP and Smyth, B. (2010) A classification-based review recommender. *Knowledge-Based Systems* 23(4):323-329.

Pang, B and Lee, L. (2008) Opinion mining and sentiment analysis. *Foundations and trends in information retrieval* 2(1-2):1-135.

Raman, T. (2009) Toward 2 w, beyond web 2.0. *Communications of the ACM* 52(2):52-59.

Sigala, M. (2008) Web 2.0, social marketing strategies and distribution channels for city destinations: Enhancing the participatory role of travelers and exploiting their collective intelligence. *Information communication technologies and city marketing: Digital opportunities for cities around the world*):220-244.

Sparks, BA, Perkins, HE and Buckley, R. (2013) Online travel reviews as persuasive communication: The effects of content type, source, and certification logos on consumer behavior. *Tourism Management* 39):1-9.

Steyvers, M and Griffiths, T. (2007) Probabilistic topic models. *Handbook of latent semantic analysis* 427(7):424-440.

Taboada, M, Brooke, J, Tofiloski, M, Voll, K and Stede, M. (2011) Lexicon-based methods for sentiment analysis. *Computational linguistics* 37(2):267-307.

Turney, PD, (2002) Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews, In: *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, p. 417-424.

Ye, Q, Law, R and Gu, B. (2009) The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management* 28(1):180-182.

Zhu, F and Zhang, X, (2006) The influence of online consumer reviews on the demand for experience goods: The case of video games, In: *International Conference on Information Systems*, Milwaukee, p. 367-382.