**Latent Aspect Rating Analysis on Review Text Data:**

**A Rating Regression Approach**

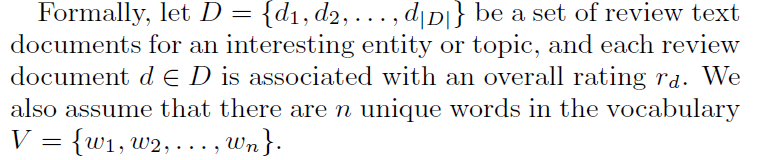
1. INTRODUCTION

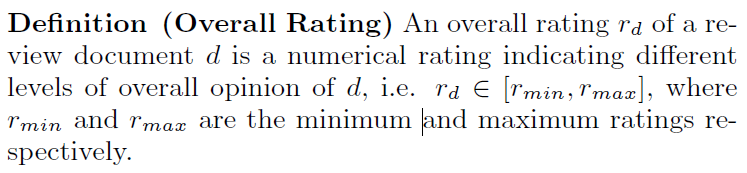
To understand such subtle differences, it is necessary to further reveal the relative importance weight that a reviewer placed on each aspect when assigning the overall rating.

To solve this new mining problem, we propose a two-stage approach based on a novel latent rating regression model. In the first stage, we employ a bootstrapping-based algorithm to identify the major aspects (guided by a few seed words describing the aspects) and segment reviews. In the second stage, we propose a generative Latent Rating Regression (LRR) model which aims at inferring aspect ratings and weights for each individual review based only on the review content and the associated overall rating.

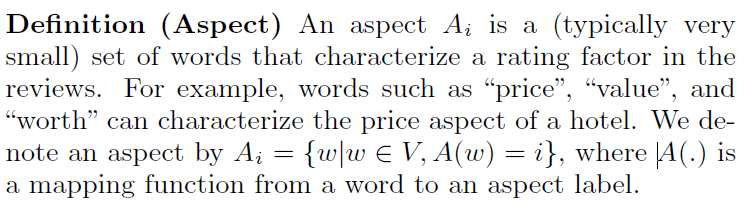
2. RELATEDWORK

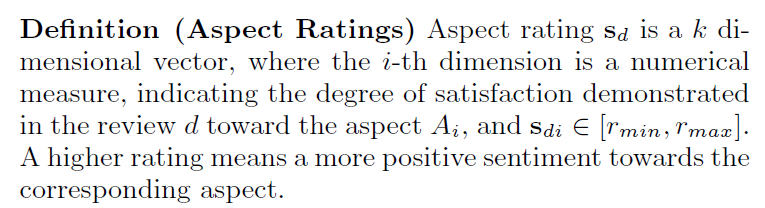
3. PROBLEM DEFINITION

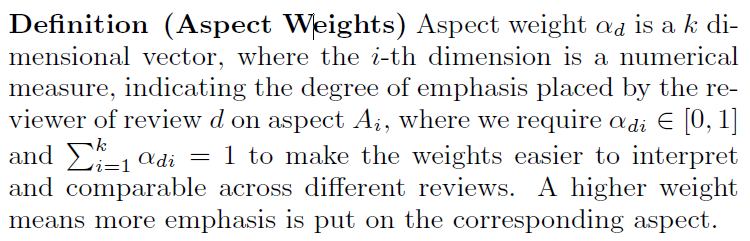


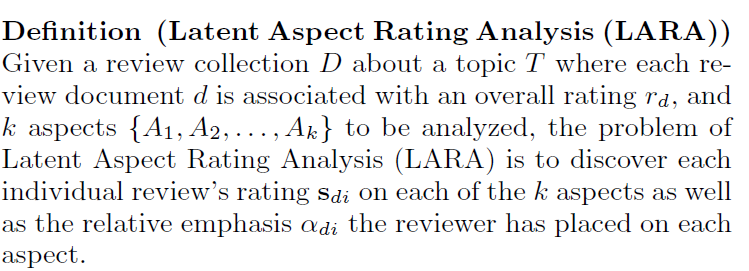


We further assume that we are given *k* aspects, which are rating factors that potentially affect the overall rating of the given topic. For example, for hotel reviews, possible aspects may include \price" and \location." An aspect is specified through a few keywords, and provides a basis for latent aspect rating analysis.









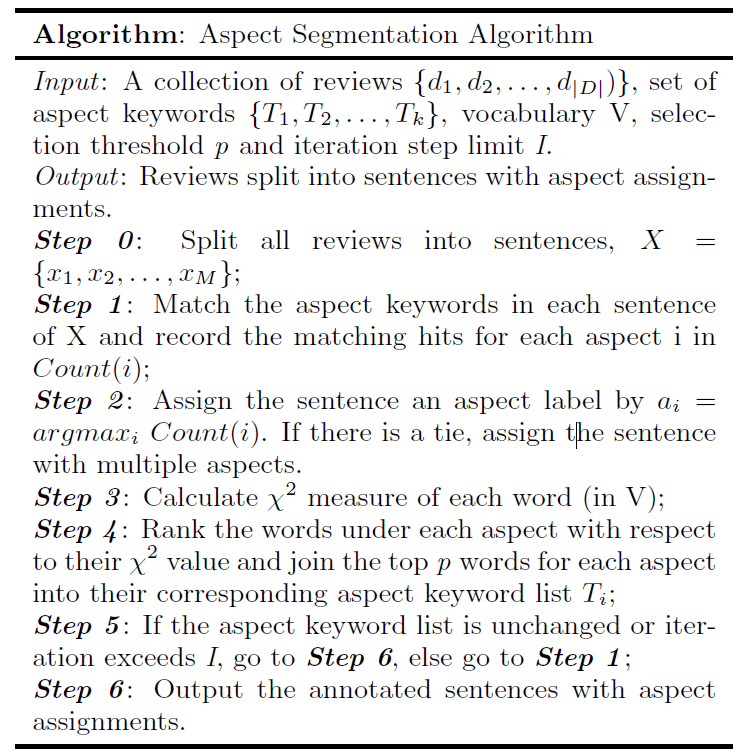
4. METHODS

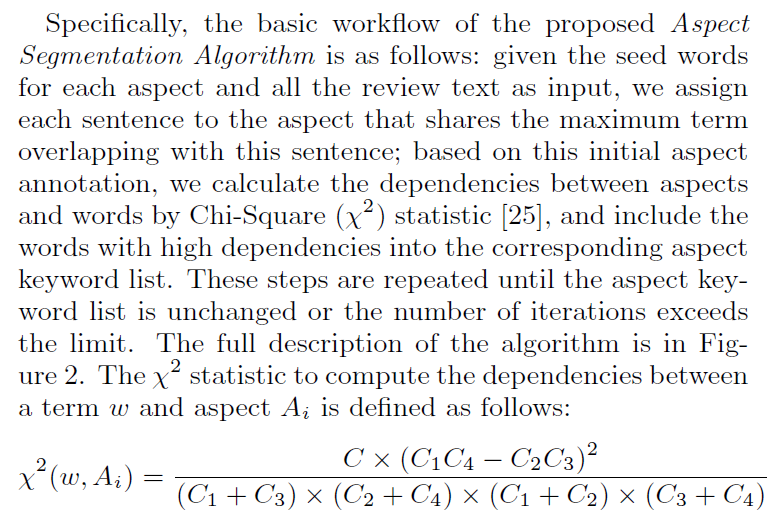
Specifically, we assume that the reviewer generates the overall rating of a review based on a weighted combination of his/her ratings on all aspects, and the rating on each aspect is generated based on another weighted combination of the words in the review that discusses the corresponding aspect. After fitting such a two-fold regression model to all the review data, we would be able to obtain the latent aspect ratings and weights, thus solving the problem of LARA.

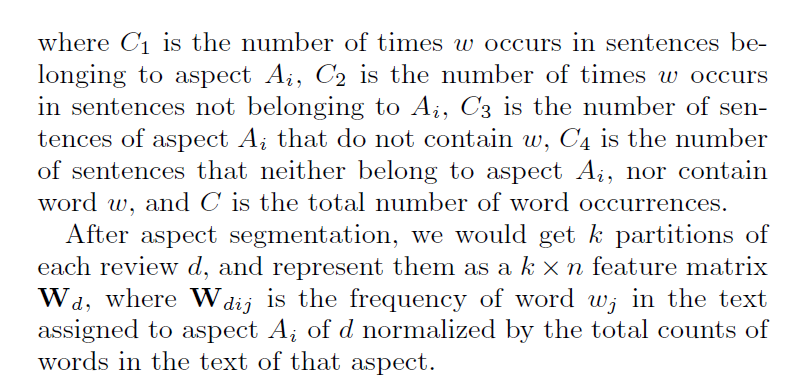
4.1 Aspect Segmentation

The goal of this first step is to map the sentences in a review into subsets corresponding to each aspect. Since we

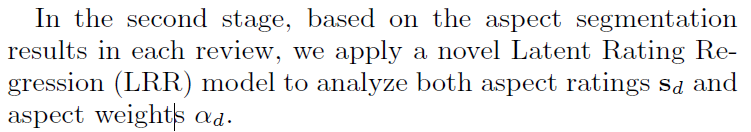
assume that only a few keywords are specified to describe each aspect, we design a boot-strapping algorithm to obtain more related words for each aspect.







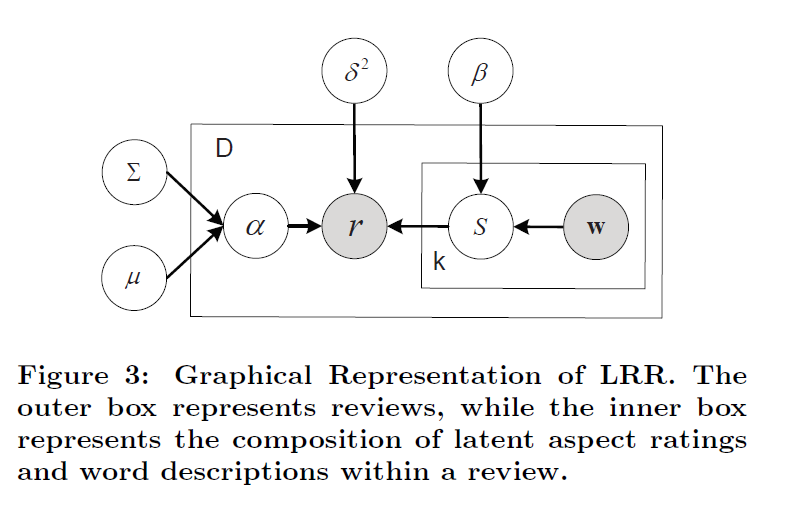
4.2 Latent Rating Regression Model (LRR)



*4.2.1 The Generation Assumption*

Our assumption of reviewer's rating behavior is as follows: to generate an opinionated review, the reviewer first decides the aspects she wants to comment on; and then for each aspect, the reviewer carefully chooses the words to express her opinions. The reviewer then forms a rating on each aspect based on the sentiments of words she used to discuss that aspect. Finally the reviewer assigns an overall rating depending on a weighted sum of all the aspect ratings, where the weights reflect the relative emphasis she has placed on each aspect.

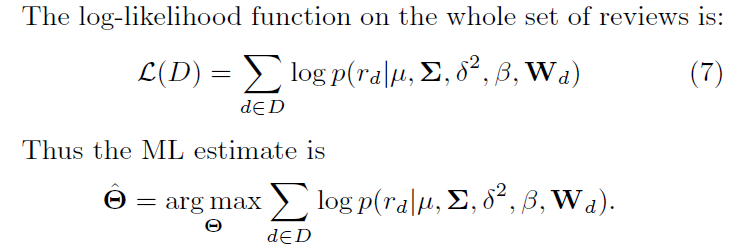
*4.2.2 The LRR Model*



*4.2.3 Discussion*

LDA is fully unsupervised, but LRR is partially supervised: although we do not have direct supervision on each aspect rating, the overall rating imposes constraints on aspect ratings and thus provides indirect supervision.

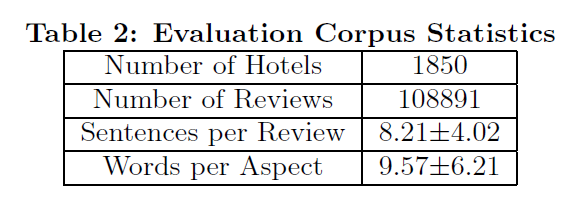
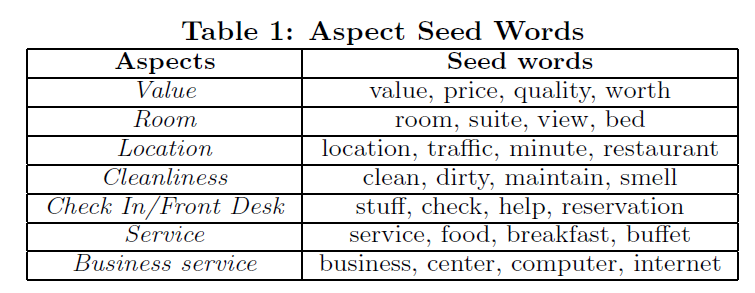
4.3 LRR Model Estimation



5. EXPERIMENT RESULTS

5.1 Data Set and Preprocessing

We first perform simple pre-processing on these reviews: 1) converting words into lower cases; 2) removing punctuations, stop words defined in [1], and the terms occurring less than 5 times in the corpus; 3) stemming each word to its root with *Porter Stemmer* [22].



5.2 Qualitative evaluation

6. CONCLUSIONS

In this paper, we defined a novel text mining problem named Latent Aspect Rating Analysis (LARA) to analyze

opinions expressed in online reviews at the level of topical aspects. LARA takes a set of review texts with overall ratings and a specification of aspects as input, and discovers each individual reviewer's latent ratings on the given aspects and the relative emphasis a reviewer has placed on different aspects.