Understanding Rating Behaviour and Predicting Ratings by Identifying Representative Users

Rahul Kamath
University of Tokyo
7-3-1 Hongo, Bunkyo-ku
Tokyo, Japan

Masanao Ochi
University of Tokyo
7-3-1 Hongo, Bunkyo-ku
Tokyo, Japan

Yutaka Matsuo
University of Tokyo
7-3-1 Hongo, Bunkyo-ku
Tokyo, Japan

Abstract

Online user reviews describing various products and services are now abundant on the web. While the information conveyed through review texts and ratings is easily comprehensible, there is a wealth of hidden information in them that is not immediately obvious. In this study, we unlock this hidden value behind user reviews to understand the various dimensions along which users rate products. We learn a set of users that represent each of these dimensions and use their ratings to predict product ratings. Specifically, we work with restaurant reviews to identify users whose ratings are influenced by dimensions like 'Service', 'Atmosphere' etc. in order to predict restaurant ratings and understand the variation in rating behaviour across different cuisines. While previous approaches to obtaining product ratings require either a large number of user ratings or a few review texts, we show that it is possible to predict ratings with few user ratings and no review text. Our experiments show that our approach outperforms other conventional methods by 16-27% in terms of RMSE.

1 Introduction

With the advent of Web 2.0, a large number of platforms including e-commerce sites, discussion forums, blogs etc. have emerged that allow users to express their opinions regarding various businesses, products and services. These opinions are usually in the form of reviews, each consisting of text feedback describing various aspects of the product along with a single numeric rating representing the users' overall sentiment about the same (McAuley et al., 2012). Such user review ratings are normally aggregated to provide an overall product rating, which help other people form their own opinion and help them make an informed decision during purchase. However, in case of new products, there is a time delay till a sufficient number of ratings that give a 'complete picture' of the product can be obtained. In such a scenario, the seller of the product may find it useful to identify a few people whose ratings, when combined together, reflect this 'complete picture'. The seller may then invite these people to review the product and, as a result, reduce the time delay involved in getting the 'true' product rating.

Review text is unstructured and inherently noisy. But it can be a valuable source of information since users justify their ratings through such text (McAuley and Leskovec, 2013). Users tend to express their sentiments about different aspects of a product in the review text and provide a rating based on some combination of these sentiments (Ganu et al., 2009). However, some users are influenced heavily by one particular aspect of the product and this is reflected in their ratings. For example: While reviewing smartphones, the ratings provided by a user may be influenced heavily by just the battery-life, irrespective of the quality of other aspects of the phone. Similarly, while reviewing restaurants, some users' ratings may correlate with the ambience of the restaurant or the level of service provided. We call such users as 'representative users' since their ratings tend to 'represent' one particular dimension of the product.

Although latent factors obtained from ratings data have been used extensively for rating prediction,

29th Pacific Asia Conference on Language, Information and Computation pages 522 - 528
Shanghai, China, October 30 - November 1, 2015

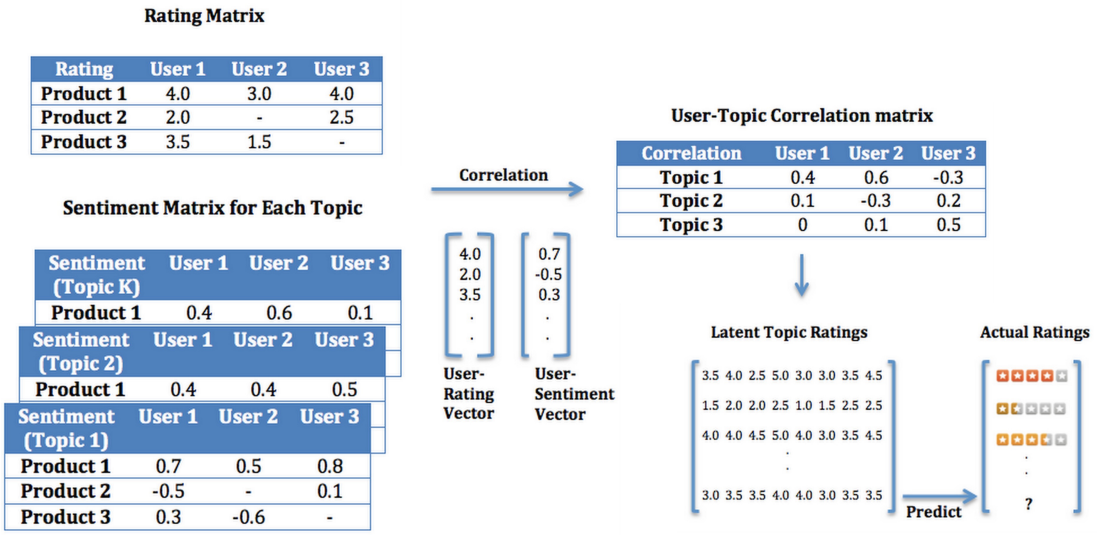

Figure 1: An overview of our proposed method

very few previous works have attempted to combine both review text and ratings. Our approach combines latent topics obtained from review text with users' rating data to learn representative users for each product. This enables us to predict ratings for new products by just looking at the ratings of a small set of users, even when no review text is available. In traditional methods, product ratings are obtained by modelling the product factors from ratings data. However, (McAuley and Leskovec, 2013) suggest that this approach is not accurate in case of new products due to the lack of sufficient number of ratings. They, in turn, propose a model which fits product factors from a few review texts. Our approach is free from both these constraints.

In this study, we use the topic model Multi-Grain Latent Dirchlet Allocation (MG-LDA) described in (Titov and McDonald, 2008a) on restaurant reviews obtained from Yelp[1] to obtain latent topics that correspond to ratable aspects of the restaurants. Since we segregate the reviews on the basis of restaurant category, we notice some interesting variations across different cuisines. The words associated with the extracted topics are then used to perform review segmentation where we identify the sentences that describe each topic. This also enables us to analyse the sentiment expressed regarding each topic in a review. We then capture the intuition of representative users to learn a set of users who best represent each topic. Latent topic ratings for restaurants are then obtained by aggregating the ratings of those users who represent that topic. The overall ratings of new restaurants are then predicted using a regression model. An overview of the proposed method is shown in Figure 1.

We also show how this concept could be used to better understand rating behaviour across different cuisines. For example: What do people who visit French restaurants care most about - food, service or value for money? How is this different from people who visit Italian restaurants?

The rest of the paper is structured as follows. Section 2 provides a review of related work. Section 3 describes our proposed method. In Section 4, we describe the experiments performed and report the results of our evaluation. Section 5 concludes the paper with a summary of the work and the scope for future work.

2 Related Work

One of the earliest attempts at rating prediction that combines both review text and ratings is (Ganu et al., 2009). However, their review segmentation method differs from ours in that their work depends on manual annotation of each review sentence into predetermined domain-specific aspects and the training of separate classifiers for each aspect. Furthermore,

[1] http://www.yelp.com

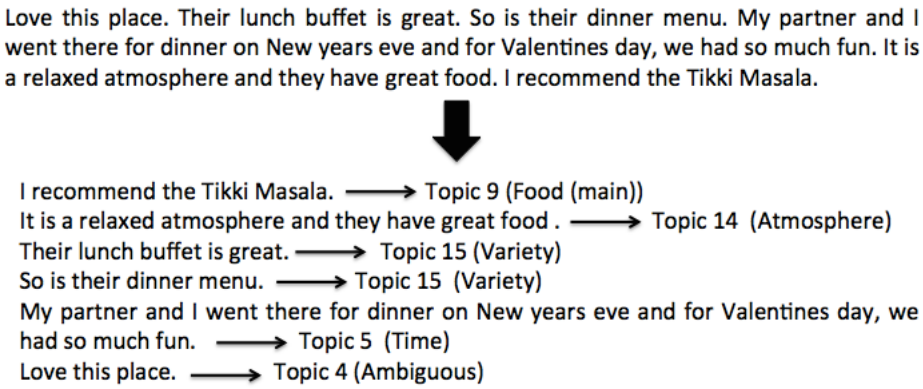

Figure 2: Review Segmentation

it does not capture the variation that may exist within the domain. For example: The aspects that affect ratings for French restaurants (e.g. 'Drinks (wine)', 'Deserts' etc.) may be different from those of Indian restaurants (e.g. 'Flavour (spiciness)', 'Variety' etc.). (Wang et al., 2010) approach the problem of segmentation by measuring the overlap between each sentence of the review and the seed words describing each aspect. However, these aspect seed words are chosen manually which are, again, domain-specific.

Topic models are normally used to make the segmentation task transferable across different domains. The problem of mapping such topics into aspects is studied in (Titov and McDonald, 2008b; Lu et al., 2011; Brody and Elhadad, 2010; McAuley et al., 2012; Jo and Oh, 2011). (Titov and McDonald, 2008b; McAuley et al., 2012) use explicit aspect ratings as a form of weak supervision to identify rated aspects while (Lu et al., 2011) use manually selected aspect seed words as a form of weak supervision. To remove the dependence on aspect ratings and aspect seed words, (Jo and Oh, 2011) develop a model that captures aspects using a set of sentiment seed words while (Brody and Elhadad, 2010) present an unsupervised method for extracting aspects by automatically deriving the sentiment seed words from review text. It is important to note that we do not map the latent topics we obtain into explicit aspects since it is not necessary for our final goal.

Rating prediction is also studied in (Gupta et al., 2010; Moghaddam and Ester, 2011; Baccianella et al., 2009) where the authors focus on multi-aspect rating prediction and in (McAuley and Leskovec, 2013) where the authors build a recommendation system using a combination of latent dimensions obtained from rating data and latent topics obtained from review text.

3 Methodology

3.1 Dataset and Preprocessing

We use the Yelp Challenge Dataset[2] consisting of around 1.12 million reviews of more than 42000 restaurants across 4 countries. These reviews are provided by more than 250000 users. Reviews contain a single star rating, text, author etc. Details of restaurants like average star rating, categories (cuisine) etc. are also available. We segment the restaurants according to its category since we would like to better understand the variation that exists across different cuisines. Note that we ignore the fact that certain restaurants may have multiple categories. For example: Some Indian restaurants may also serve Thai food.

We tokenize the review text along whitespaces, remove all punctuation and stop-words, and lemmatize the words using the NLTK Wordnet lemmatizer described in (Bird et al., 2009).

3.2 Topic Extraction

We run the topic model multi-grain LDA described in (Titov and McDonald, 2008a) on a corpus of restaurant reviews obtained from a single cuisine to extract K latent topics. Unlike standard topic

[2] http://www.yelp.com/dataset_challenge

Cuisine	Interpreted Topic	Top Words
Indian	Variety	buffet,lunch,dish,vegetarian,menu,selection,option,good,item,great
	Food	chicken,masala,tikka,curry,naan,lamb,dish,paneer,tandoori,ordered
	Flavour	spicy,spice,flavour,dish,hot,curry,food,like,sauce,taste
	Value	price,portion,food,meal,get,two,small,little,rice,bit
	Atmosphere	restaurant,place,nice,decor,inside,strip,little,clean,like,table,look
Italian	Food (Pizza)	pizza,crust,good,cheese,sauce,slice,thin,like,wing,great,topping
	Food (Salad)	salad,bread,cheese,garlic,tomato,fresh,sauce,olive,delicious,oil
	Service	service,staff,friendly,server,owner,customer,waiter,always,attentive
	Location	place,restaurant,location,strip,little,find,italian,away,parking,right
	Value	food,good,price,better,much,pretty,like,quality,portion,worth,nothing
French	Drinks	menu,wine,course,tasting,glass,bottle,ordered,selection,meal,two
	Dessert	dessert,chocolate,cream,cake,ice,coffee,sweet,creme,tart,also,good,souffle
	Food (Bread)	bread,egg,french,butter,good,toast,delicious,fry,cheese,fresh,croque
	Food	cheese,salad,soup,onion,ordered,good,french,delicious,appetizer,lobster
	Service Time	table,minute,time,wait,reservation,waiter,get,seated,server,took,order,got

Table 1: Local topics for Indian, Italian and French restaurants obtained using MG-LDA

modeling methods such as LDA and PLSA, which extract topics that correspond to global properties of a product, MG-LDA extracts much finer topics that correspond to ratable aspects of the product. To extract topics at such granular level, the model generates terms which are either chosen at the document level or chosen from a sliding window[3]. The terms chosen from the sliding window correspond to the fine topics.

3.3 Review Segmentation and Sentiment Analysis

Once cuisine-specific latent topics are obtained, the review segmentation task is performed where each review sentence s_i is assigned to one of the latent topics t_k. The purpose of this task is to understand which sentences of the review discuss which of the topics. The topic assignment is made as follows:

$$Topic(s_i) = \arg\max_{k} \sum_{w \in t_k} count(w, s_i) * P(w|t_k) \quad (1)$$

where w is the word associated with each topic, count(w, s_i) is the count of word w in sentence s_i and $P(w|t_k)$ is the probability as determined from the word distributions obtained using the MG-LDA model.

[3]A sliding window is a set of fixed number of adjacent sentences.

For every review, the sentences that discuss each topic are identified as shown in Figure 2. It is therefore possible to determine the sentiment expressed by the review author regarding each latent topic by averaging over the sentiments of its constituent sentences. We use the implementation TextBlob[4], which is based on the Pattern[5] library, to determine the polarity of each sentence. The polarity is obtained in the range of [-1, 1].

3.4 User Segmentation

We then proceed to learn the representative users for each latent topic. First, the feature vector $\theta_u^{overall}$ is obtained for each user u where each feature represents the users' review rating for a restaurant. We assume that each user writes only one review per restaurant. Similarly, $\theta_u^{t_k}$ is obtained where each feature represents the users' sentiment regarding topic t_k.

The influence of a topic on a users' rating is determined by calculating the Pearson's correlation between $\theta_u^{overall}$ and $\theta_u^{t_k}$. Only users who have provided a minimum of 5 reviews are considered. A user-topic correlation matrix C is thus obtained which indicates the dimensions along which each

[4]http://www.textblob.readthedocs.org/en/dev/

[5]http://www.clips.ua.ac.be/pattern

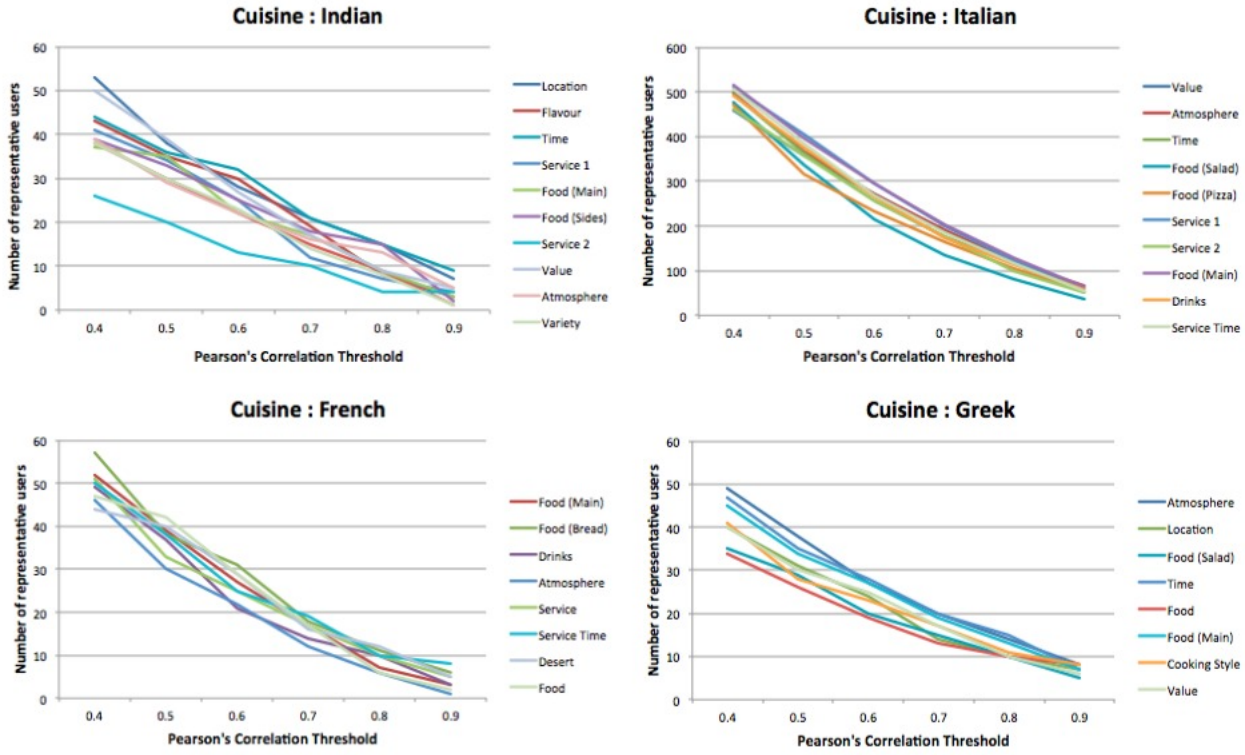

Figure 3: Number of Representative Users for various cuisines

user tends to rate restaurants. Simply put,

$$C(u, t_k) = PearsonCorr(\theta_u^{overall}, \theta_u^{t_k}) \quad (2)$$

The representative users for a topic are those users whose $C(u, t_k)$ value is above a certain threshold T for that particular topic. It is important to note that $C(u, t_k)$ value may not be available for all user-topic pairs since every user may not express sentiments regarding every topic.

3.5 Rating Prediction

We calculate the topic ratings of restaurants once we obtain a list of representative users for each latent topic. This rating is calculated as the average of the review ratings that are provided by the representative users of that particular topic. In case there are no representative users for a particular topic for that particular restaurant, this rating is calculated as the average of the other latent topic ratings. Such topic ratings provide some indication of the quality of various aspects of the restaurant (like food, service etc.), although we do not explicitly calculate the aspect ratings or map the topics to aspects.

Since the overall restaurant rating can be thought of as some combination of the ratings for food, service, atmosphere etc., we try to combine the latent topic ratings in some way. For this purpose, we fit a Support Vector Regression (SVR) model with radial basis function kernel on the latent topic ratings and use it to predict the overall rating of restaurants. During test time, just the ratings provided by a few representative users would be enough to obtain the overall restaurant rating. Such a rating takes into account the different dimensions of the restaurant and provides a 'complete picture' of the restaurant.

4 Experiments and Analysis

We use the topic model MG-LDA on a set of 8000 reviews each of Indian, Italian, French and Greek restaurants. The number of global topics is set at $K_{glo} = 40$ and local topics at $K_{loc} = 15$ (After trying various combinations, we found that this combination provides the best results. Previous works have also used a similar number of topics). The length of the sliding window is set at 2 and all the other parameters for the model is set at 0.1. We run

the chain for 1000 iterations. While the global topics are ignored, some select local topics as determined by the model are shown in Table 1. We try to interpret the topics manually by looking at the constituent words. Usually, around 5-6 local topics are ambiguous and difficult to interpret.

A quick look at the topics obtained shows us the variation that exists among different cuisines. For example: While Indian restaurants have 'Flavour' and 'Variety' as topics; Italian restaurants have 'Drinks'; French restaurants have 'Drinks' and 'Dessert' as topics. Greek restaurants have 'Cooking Style' as a topic with words like dry, fry, fresh, cooked, soft, tender etc. Also, certain words like table, minute, time, wait, hour, bar, seated etc. appear together in case of French and Italian restaurants signaling, perhaps, a long wait to get seated at such restaurants.

Review segmentation is then performed on around 8500 reviews of Indian restaurants, 61000 reviews of Italian restaurants and 17000 reviews of French restaurants, where each sentence is assigned to one of the 15 latent topics. Sentiment analysis is conducted and the user-topic correlation matrix is obtained for each restaurant category.

Using the user-topic correlation matrix, we segment the users according to each latent topic. Figure 3 shows the number of representative users for each topic for different correlation thresholds T. For the sake of clarity, we only show those latent topics that could be interpreted by us. It is interesting to observe that people who visit Indian restaurants tend to care the most about 'Location' and 'Value (Pricing)' and the least about 'Service' and 'Atmosphere'. On the other hand, people who visit French restaurants care the most about 'Food (Bread)' and 'Food (Main)' and the least about 'Atmosphere'. Similarly, while providing ratings, more number of users are influenced by the 'Atmosphere' at Greek restaurants than 'Food'. We then proceed to obtain the latent topic ratings for each restaurant. For this purpose, we only select those users whose correlation threshold, $T >= 0.4$ as representative users. For each latent topic, we average over the ratings provided by such users to obtain the topic ratings (out of 5). It is therefore possible to obtain crude ratings for aspects like 'Food', 'Service' etc. which give an indication of the quality of the aspects. We then fit an SVR model, the performance of which is described below.

4.1 Evaluation

To evaluate the performance of rating prediction, we determine the RMSE between the actual and predicted ratings for Italian restaurants. We compare the RMSE for MG-LDA and online LDA described in (Hoffman et al., 2010). In case of LDA, we detect $K = 50$ topics as in previous works. We use the latent topic ratings of 640 restaurants for training and 215 restaurants for test. The results are shown in Table 2.

Models	RMSE
(a) MG-LDA, SVR with rbf kernel (Proposed Model)	0.4909
(b) MG-LDA, SVR with linear kernel	0.5377
(c) LDA, SVR with rbf kernel	0.5812
(d) LDA, SVR with linear kernel	0.6277
(e) Baseline 1	0.6737
(f) Baseline 2	0.5831
Improvement	
(a) vs. (e)	27%
(a) vs. (f)	16%

Table 2: Evaluation (Italian Restaurants)

An RMSE of 0.4909 is obtained when using MG-LDA and SVR with rbf kernel. Each restaurant has an average of 22 representative users. Inviting these users to rate new restaurants would help in predicting the 'true' restaurant rating (which is the rating obtained once a considerable number of users have rated the restaurant over a period of time). However, conventional methods just average over their ratings, without taking into account the different topics that they represent. Such an approach gives an RMSE of 0.6737 (Baseline 1). Our approach outperforms this method by 27%. Also, since most people provide a rating of 3, 3.5 or 4 when rating restaurants, predicting a constant rating every time may also give a reasonable result. We find that predicting a rating of 3.64 (average over the test set) every time results in an RMSE of 0.5831 (Baseline 2). Our approach outperforms such a constant classifier by 16%.

We repeat the same procedure for Indian restaurants by using the latent topic ratings of 120 restau-

rants for training and 40 restaurants for test. The results are shown in Table 3.

Models	RMSE
MG-LDA, SVR with rbf kernel	0.4635
MG-LDA, SVR with linear kernel	0.5795
LDA, SVR with rbf kernel	0.5734
LDA, SVR with linear kernel	0.6997

Table 3: Evaluation (Indian Restaurants)

5 Conclusion and Future Work

In summary, we show how latent topics in review text could be used to unlock hidden value in user reviews. We utilise the intuition that, while rating products, certain users are influenced heavily by one particular aspect of the product. We learn such users by detecting the sentiments expressed by them with regard to each latent topic and then by comparing these sentiments with the actual ratings provided. We also use this to draw some interesting insights regarding users' rating behaviour across different cuisines and obtain latent topic ratings for restaurants. Overall ratings, which take into account the different dimensions of the restaurant, are then obtained using a regression model.

In the future, we would like to show that this approach is transferable to other domains like e-commerce. Also, it would be interesting to segregate the reviews by star ratings as this would help us understand the factors that a restaurant is getting right and those they are getting wrong. For example: The dimensions corresponding to review text having 5-star ratings would be different from those having 1-star ratings.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2009. Multi-facet rating of product reviews. In *Advances in Information Retrieval*, pages 461–472. Springer.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. " O'Reilly Media, Inc.".

Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 804–812. Association for Computational Linguistics.

Gayatree Ganu, Noemie Elhadad, and Amélie Marian. 2009. Beyond the stars: Improving rating predictions using review text content. In *WebDB*, volume 9, pages 1–6.

Narendra Gupta, Giuseppe Di Fabbrizio, and Patrick Haffner. 2010. Capturing the stars: predicting ratings for service and product reviews. In *Proceedings of the NAACL HLT 2010 Workshop on Semantic Search*, pages 36–43. Association for Computational Linguistics.

Matthew Hoffman, Francis R Bach, and David M Blei. 2010. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864.

Yohan Jo and Alice H Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 815–824. ACM.

Bin Lu, Myle Ott, Claire Cardie, and Benjamin K Tsou. 2011. Multi-aspect sentiment analysis with topic models. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 81–88. IEEE.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172. ACM.

Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 1020–1025. IEEE.

Samaneh Moghaddam and Martin Ester. 2011. Ilda: interdependent lda model for learning latent aspects and their ratings from online product reviews. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 665–674. ACM.

Ivan Titov and Ryan McDonald. 2008a. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, pages 111–120. ACM.

Ivan Titov and Ryan T McDonald. 2008b. A joint model of text and aspect ratings for sentiment summarization. In *ACL*, volume 8, pages 308–316. Citeseer.

Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 783–792. ACM.