Estimating Customer Reviews in Recommender Systems Using Sentiment Analysis Methods

Konstantin Bauman,1 Bing Liu,2 Alexander Tuzhilin1
1Stern School of Business, New York University
2University of Illinois at Chicago (UIC)

Abstract

The paper presents a method for estimating unknown user reviews in terms of which specific aspects of a particular item, such as a restaurant, a user would mention in a review that he/she would write about the item and also which sentiments the user would express about these aspects. Unlike the traditional rating-based recommendation methods, the proposed approach estimates user experiences of an item in terms of the most crucial aspects of the item for the user. Therefore, this approach enables more detailed item recommendations to the user. We apply this method to two real-life review datasets from Yelp to evaluate its performance.

1 Introduction

The use of recommender systems (RSes) has exploded over the last several years to the effect that most of the major companies, including Amazon, Netflix, Google, Facebook, Microsoft, Twitter, LinkedIn, Yahoo!, eBay, Pandora and others, extensively use recommendations as a part of their products or services. Furthermore, RSes constitute mission-critical technologies in some of these companies. For example, at least 75% of Netflix movie downloads come from its recommendation engine, making it of strategic importance to Netflix1 2. Similarly, the whole business model of Stitch Fix in its entirety (100%) relies on recommender systems3. Due to the importance of the recommendation

---

2Hunt N. Quantifying the Value of Better Recommendations, Keynote RecSys 2014. recsys.acm.org/recsys14/keynotes
problem, there has been extensive research conducted on recommender systems in the industry and academia, both in computer science [19] and information systems [6, 23, 24]. Although early paradigm of RSes was based on a two dimensional (2D) matrix of user ratings of items, such as restaurants or hotels, and on how to estimate the unknown ratings in that matrix (the so-called, matrix completion problem of collaborative filtering [11]), there has been extensive effort in the RS community to go beyond this 2D paradigm and to study numerous other aspects of the multifaceted recommendation problem [2].

One such direction is an attempt to use user-generated reviews to improve recommendations. In particular, several papers tried to improve estimation of unknown ratings by using user reviews [7, 8, 17, 21, 22]. The common theme of these papers is how to extract useful information from the user reviews to better predict unknown ratings, e.g. how to do it for Yelp ratings using Yelp reviews. For example, [7] finds six aspects in restaurant reviews, trains classifiers to identify them in the text, and shows that this information improves rating prediction quality. In [9] authors trained a model for extracting the “trip type” contextual variable from the user review and showed how to improve rating predictions with this variable. As another example, [16] uses the LDA-based approach combined with Matrix Factorization for better predicting the unknown ratings. In particular, [16] obtains highly interpretable textual labels for latent rating dimensions, which helps justifying particular rating values using texts of the reviews. The more recent papers [5] and [13] go beyond [16] and use a more complicated graphical models to predict unknown ratings based on collaborative filtering and topic modeling of user reviews. In [3], a consumer choice model is presented that learns consumers relative preferences for different product features not only in terms of the characteristics of products and users but also in terms of user generated reviews. [3] uses text mining to extract important features and consumer sentiments about these features from the reviews and use this information in their consumer choice model. Further, [8] recommend hotels to travelers by ranking them based on their utility that depends not only on the hotel and consumer features
but also on the hotel reviews. In particular, [8] mines the user reviews about the hotels to extract hotel’s most important features and user sentiments about these features, and incorporates this information into the utility estimation model.

Most of this work focuses on how to use reviews for better estimation of unknown ratings. In this paper, we focus on a review-based recommendation method that suggests items to users based on the entire user reviews of items, as opposed to the ratings or ranking based methods. By analyzing the entire review using text mining and sentiment analysis methods and estimating future reviews that the user can write about an item, we can rely on significantly richer information, as opposed to using a single or even multiple ratings when deciding what to recommend to the user. This idea has been explored in [1] where the authors constructed aspect ontology for the Digital Camera application, developed a set of rules for identifying the aspects from the ontology in text and also their sentiments. Based on the collected data, they aggregate item’s (i.e. camera’s) profiles and present simple recommendations using knowledge based recommendation techniques. In contrast to the knowledge-based approach of [1], we estimate the unknown review using text mining and sentiment analysis methods. Further, the RecSys poster paper [20] proposes a method of extracting aspect-specific ratings from the reviews and recommending those existing reviews to the users which they have not seen before. In contrast to [20], we focus on estimating the future reviews that do not exist yet and that the user may want to write.

In this paper we present a new approach to predicting a review that a user may write about a particular item. When processing the reviews, we focus on the set of salient aspects of these reviews identified by our system. In particular, we predict which aspects of an item will be important to the user in a review and also estimate the sentiments that the user will express about these aspects. This allows us to construct new (previously not existing) reviews by estimating the set of the most salient aspects and their sentiments. The contributions of this paper lie in
• Proposing a novel review estimation method based on the sentiment analysis and 
the machine learning techniques that predict the set aspects and sentiments about 
these aspects that the user would express in a review. Note that this entire approach 
does not depend on or involves any rating data, which makes our method useful in 
those applications that do not naturally have ratings.

• Developing simple and powerful explanations of why particular items are recom-
  mended to the users. These explanations can be constructed based on the estimated 
  aspects of the reviews and user’s sentiments about these aspects. For example, the 
  Lupulo restaurant in New York City may be recommended to Jane Doe because she 
  will love the duck as the main course, appetizers and the wine list there but she 
  may not be entirely happy with the desert menu and the service in that restaurant.

• Testing the proposed review estimation method on the actual “real world” reviews 
  and showing that our method can predict aspects and sentiments of the unknown 
  reviews well in comparison to the baselines.

2 Overview of the Proposed Method

In this section we present a method of estimating unknown reviews in terms of predicting 
which key aspects of the item the user will mention in review and what sentiments about 
these aspects the user would express. More specifically, in this paper we follow the aspect-
based sentiment analysis approach [15], assume that each review contains a set of item’s 
most salient characteristics, called aspects, and that the reviewer expresses opinions with 
the corresponding sentiments about these aspects. For example, consider Yelp review 
presented in Figure 1. It has the following aspects and the sentiments about them: 
(smell, positive), (sandwich, positive), (sauce, positive). More formally, we follow [10, 14] 
and define an opinion as follows.

Definition: An opinion is a quintuple, \((e, a, so, h, t)\), where \(e\) is the name of an entity, 
a is an aspect of \(e\), so is the orientation of the opinion about aspect \(a\) of entity \(e\), \(h\) is
Given a collection of documents $D$ with opinions about them, the goal of sentiment analysis is to discover all the opinion quintuples $(e, a, so, h, t)$ in $D$.

We use the following review about the Taqueria restaurant as an example to show what sentiment analysis does (an id number is associated with each sentence):

Posted by: John, Date: 3/9/2015,
Text: “(1) Had lunch in Taqueria today. (2) Ordered the taco with rice and beans and it was great. (3) The service was quick. (4) The atmosphere was dark and soothing.”

In this review, sentence (2) expresses a positive opinion about the food in the Taqueria restaurant. Sentence (3) expresses a positive opinion on the aspect of “service” in that restaurant. Overall, the sentiment analysis system should produce the following three opinion tuples: (Taqueria, food, positive, John, 3/9/2015), (Taqueria, service, positive, John, 3/9/2015), (Taqueria, atmosphere, positive, John, 3/9/2015)

Since we know the opinion holder, the item being reviewed and the time when a review is posted, the sentiment analysis system only needs to discover aspects and also the sentiment orientations about the aspects commented by the reviewer of each review. To accomplish this task, we used a state-of-the-art sentiment analysis system, called Opinion Parser [15], which is also used by two commercial companies. The Opinion Parser aspect extraction algorithm uses Double Propagation (DP) method from [18]. The sentiment classification algorithm is the lexicon-based method [15]. The DP algorithm is
based on the idea that an opinion must have target(s), and a sentiment expression and its targets often have some grammar dependency relation. This observation can be exploited for aspect extraction. For example, consider the sentence “The restaurant has very tasty fish.” If we know that tasty is a sentiment expression, we can extract fish as an aspect because of a grammar modification relation between tasty and fish. The DP method has many sophisticated grammar rules and pruning methods for accurate extraction of aspects. The lexicon-based sentiment classification algorithm uses a set of sentiment expressions (such as good, amazing, bad, cost an arm and leg, etc), a set of sentiment composition rules, and grammar analysis to determine the sentiment about each aspect in a sentence. For example, from the sentence “The Burger King is doing very well in this poor economy,” the system finds the opinion about Burger King is positive and about economy is negative. The detailed algorithms used in Opinion Parser are quite involved and are presented in [15].

In this paper we use Opinion Parser to build a set of aspects $A_0$ occurring in the set of reviews $R$ for a given application (e.g. Restaurants). Furthermore, for each review $r$ we identify a set of aspects $A_r$ occurring in the review with corresponding sentiments expressing user’s opinions about aspects from $A_r$.

Given a set of users, items and reviews, our goal is to estimate unknown reviews that users would produce about items in terms of estimating the aspects appearing in the reviews and potential sentiments that the users would express about these aspects. For the case of the review-based recommendations, this step is analogous to the problem of estimating unknown ratings in RS. To solve this problem, we propose the following method consisting of 8 steps presented in Figure 2, which are described below.

(1) **Extract the set of aspects**

In this step we use Opinion Parser to build a set of aspects $A_0$, as explained above.

(2) **Identification of specific reviews**

In this step we classify all the reviews into generic and specific. We follow [4] and
define *specific* reviews as those that describe a particular experience of an item by a user, such as a particular visit to a restaurant. In contrast, *generic* reviews refer to the overall impressions about a particular item. For example, a generic review of a restaurant may say that a person is a regular visitor of a certain restaurant and that she likes food there. Generic reviews tend to be short and contain only a small number of aspects [4] in contrast to the specific reviews that cover many more details about various aspects of user experiences with the item being reviewed. Since we try to predict a set of aspects describing future experiences of a user with a certain item, generic reviews tend to be less relevant for this task. Therefore, we focus on the specific reviews in the rest of this paper and filter out generic reviews as being irrelevant. We identify specific reviews using the supervised learning approach as follows. First, we label a small set of reviews to be either specific or generic. Then we train a classification model on a labeled set of reviews. Finally, we identify a new review as being generic or specific using that prediction model. We use the same set of features in this classification task as in [4], such as the numbers of sentences, words, verbs, verbs in the past tense, and the ratio of the number of verbs in past tense to the whole number of verbs in the review.

Additional details of this learning process will be presented in Section 3.2.

(3) Aspect identification and sentiment aggregation

In this step we apply *Opinion Parser* to each specific review \( r \) identified in Step 2 in order to determine a set of aspects \( A_r \) appearing in the review with corresponding
sentiments expressing users opinion about aspects from $A_r$. If an aspect appears in more than one sentence of a review, we compute an aggregate sentiment for that aspect as follows. First, we calculate the average ($avg$) of all the sentiments, assuming that positive sentiment is +1, negative is −1 and neutral is 0. Then the final sentiment about the aspect is $\text{sign}(avg)$. If $avg = 0$, then the aggregate sentiment is neutral. At the end of this step a review is reduced to a set of aspects and sentiments about these aspects.

(4) Building user and item profiles

Next, we build user and item profiles based on the set of identified specific reviews. For each user $u$ and each item $i$ (e.g. Restaurant or Spa Salon), we build profiles $P_u$ and $P_i$ based on the set of historical reviews $H_u$ and $H_i$ corresponding to user $u$ and to item $i$. In particular, for each aspect $x$ from $A_0$ we compute:

- $F_x$ – Fraction of reviews from $H_u$ ($H_i$) containing aspect $x$, i.e. number of reviews from $H_u$ ($H_i$) containing aspect $x$ divided by size of set $H_u$ ($H_i$).
- $\text{TFIDF}_x$ – Number of reviews from $H_u$ ($H_i$) containing aspect $x$ divided by logarithm from fraction of users (items) having aspect $x$ at least in one review. This is the same measure as TF-IDF in text mining, showing the importance of a particular aspect for user $u$ (item $i$) in comparison to other users (items).
- Numbers and fractions of reviews from $H_u$ ($H_i$) containing aspect $x$ with positive / neutral / negative sentiment.
- $S_x$ – Average sentiment of aspect $x$ in set $H_u$ ($H_i$).

The constructed profile reflects the importance of various aspects from $A_0$ for particular users and items based on their reviews. Among other things, these profiles contain information about frequencies of aspects in user’s and item’s reviews. This information should help us to predict if a particular aspect would appear in a new review.

(5) Aspect selection

In order to simplify our model we eliminate the “unimportant” aspects, i.e., those that appear infrequently in the reviews and therefore do not affect the overall performance of
the system. In particular, we select a subset of those aspects $A_1$ from set $A_0$ that have:
(a) relatively high $F_x$ for a sufficient number of items’ profiles; and (b) relatively high $TFIDF_x$ for a sufficient number of items’ profiles, where “sufficient” assumes that the number is above a certain threshold. Thus, we construct a set of important aspects $A_1$ and focus subsequently only on this set in the next steps of our method. For example, aspect service is important because it is frequent for many items, while aspect internet is unimportant because its pretty rare in restaurant application. Therefore, we use aspect service and drop aspect internet in the subsequent steps of our method.

(6) Training the Aspect Presence model

In order to predict if a certain aspect $x$ would appear in the future review of user $u$ and item $i$, we train a classification model based on the historical reviews. Note, that we encode “presence” of aspect in a review as 1 and “absence” as 0. In this paper we study two approaches to this prediction problem. The first approach is based on the information collected in user’s ($P_u$) and item’s ($P_i$) profiles. We train a separate classification model for each aspect using standard machine learning algorithms (e.g. SVM, Random Forest) based on features from $P_u$, $P_i$ and their interaction.

An alternative approach is based on the well-known Matrix Factorization (MF) method [12], where we use the “presence/absence of an aspect in a review” measure as the “rating” for the MF model. The resulting MF prediction is mapped into the aspect presence/absence classification using a threshold value. For this threshold we use the average “presence” of an aspect in the train set of reviews.

(7) Training the Aspect Sentiment model

In this step we train a separate “Aspect Sentiment” model for each aspect $x$ in order to predict the sentiment that user $u$ would have about aspect $x$ of item $i$. For this purpose we use only non-neutral sentiments and encode them as 0 for negative and 1 for positive sentiments. We address this prediction problem with the same two approaches. First, we build a classification model using standard machine learning techniques (e.g., SVM,
Random Forests) based on the features from user’s profile $P_u$, item’s profile $P_i$ and their interaction. The second approach is to train the standard Matrix Factorization model on sentiments as ratings for a particular aspect. Similarly to the previous step, we map the MF prediction to the positive or the negative class using a threshold value. This threshold is defined as the average sentiment of an aspect in the training set of the reviews.

In this paper we predict the aspects and the sentiments of the review using binary classification methods, as opposed to more complicated classification or even regression schemes, because we want to provide simple “like/dislike” predictions of relevant aspects of the review rather than more complicated estimations of how much the user would like various aspects of an item.

(8) Predict the set of important aspects & their sentiments for a review

Once all models are built, we apply them to predict a new review $r$ that user $u$ may write about item $i$. First of all, we apply all the aspect presence models in order to identify a set of aspects $A_r$ that would appear in review $r$. Secondly, we apply the aspect sentiment models to set $A_r$ and predict the sentiments for those aspects. And finally, we provide an explanation of what is “special” about item $i$ to user $u$ by presenting the estimated set of aspects $A_r$ with the set of predicted sentiments.

In summary, we proposed a method for predicting a new review of an item by a user by identifying a set of aspects that the user would mention in that review and predicting the sentiments that she would express about those aspects. In Section 3, we empirically validate our method on data from two applications and will show the results in Section 4.

3 Empirical Study

To demonstrate how well our method works in practice, we tested it on the Yelp dataset\textsuperscript{4} with the goal of predicting sets of aspects and their sentiments for the unknown test set of reviews for restaurants and beauty & spas applications. We describe the Yelp data in

\textsuperscript{4}www.yelp.com/dataset\textunderscore challenge/dataset
Section 3.1 and the specifics of our experiments in Section 3.2.

3.1 Dataset Descriptions

The Yelp dataset contains reviews of various businesses, including restaurants, beauty & spas and others, provided by various users of Yelp describing their experiences visiting these businesses. In our case, these reviews were collected in the Phoenix metropolitan area in Arizona over the period of 6 years. In this study we used all the reviews in the dataset for the 4503 restaurants produced by 36,473 users (158,430 reviews in total) and for the 764 beauty & spas produced by 4272 users (5,579 reviews in total). We selected these two categories of businesses (out of 22 categories) because they contained some of the largest numbers of reviews and also differed significantly from each other. A review of a business by a user is defined by its text, the date of the review and its rating.

3.2 Applying the Proposed Method

We applied the 8-step method presented in Section 2 to the Yelp data. As a result, we managed to extract 69 aspects for Restaurants and 45 aspects for Beauty&Spas in Step 1 of our method using Opinion Parser. Table 1 presents several aspects pertaining to Restaurant application with examples of corresponding words. In Step 2, we labeled 300 reviews to be ether specific or generic and trained a classification model on this labels. We tried Naive Bayes (NB), SVM, Logistic Regression (LR) and Random Forests (RF) classification models and selected NB model as the method of choice based on its performance. The cross validation accuracy was 0.87 and 0.85 for the restaurant and the beauty&spa applications respectively for NB. Consequently, we have identified 80,556 specific reviews for the restaurants and 3,419 specific reviews for the beauty&spas cases.

Further, the set of selected specific reviews is partitioned into three sets: stat, train and test in the ratio of 40/40/20. These sets of reviews are subsequently used for building profiles, training the aspect presence and aspect sentiment classification models, and testing the performance of the overall method.
After identifying the sets of aspects in the reviews and aggregating their sentiments in Step 3, we built the user and item profiles in Step 4. In order to avoid cold-start problem we use only those users and items that have more than a certain number of reviews in their profiles. We set these threshold numbers to 5 for the restaurant application and to 1 for the beauty&spa application. After selecting all the users and businesses satisfying these threshold values, we obtained the final numbers of users, items and reviews in restaurant and beauty&spa applications that are presented in Tables 2.

After we built the profiles of users and items (restaurants and beauty&spa salons), we select subset $A_1$ of the most important aspects from set $A_0$, as described in Step 5 of our method, using the threshold values of 100 and 20 (for restaurants and beauty&spa respectively) for a number of items having $F_x$ at least 0.1. And we use the same threshold values for the number of items having $TFIDF_x$ more than 1. As a result, we reduced the number of important aspects to 32 for restaurants and to 21 for beauty&spa.

In Step 6 of our method, we build the Aspect Presence models for each aspect from $A_1$. According to the profile-based approach we use the train set of reviews to train various standard ML classification models, including Logistic Regression, SVM and Random Forests (RF), based on user’s and item’s profiles constructed in the previous step.
We selected RF as the best-performing one from all these models and use only it subsequently when comparing the two approaches to building the Aspect Presence model. As the second approach, we use the Matrix Factorization (MF) model for predicting the presence/absence of an aspect in a review.

We have also built the Aspect Sentiment model in Step 7 using similar principles as explained in the previous step and also described in Section 2. In addition, we compared several Machine Learning classification techniques and selected the Random Forest model to present the best results for the profile-based approach. Also we trained MF model as the second approach to predicting aspect sentiment.

Finally, in Step 8, we predict the set of aspects and the sentiments about these aspects for each review in the test set. The results of these predictions are reported in Section 4. Before reporting these results, however, we describe the performance measures that we use in our study in Section 3.3.

### 3.3 Performance Measures

In this work, we compare the results of our proposed method with three baselines in terms of various classification measures to see how well our method works in practice vis-a-vis other alternative approaches. As the first baseline, we use the “All Aspects Included” method, which always predicts that all the important aspects selected in Step 5 of our method would appear in all the reviews. This simple method is included in our study because it represents the standard multi-criteria approach where the system tries to predict ratings for a fixed set of aspects across all the users and items in the application. We also include the method “All Aspects Positive” as the sentiment prediction baseline and define it in a similar and obvious manner.

The second baseline that we use in this study is the random predictions method. We included it in our study to demonstrate that our method outperforms random predictions of aspects and user sentiments about them. As a third baseline we use the method predicting that aspect x would occur in a review of item i if x appears in more than
50% of item i’s historical reviews. In other words, this “Item Average” aspect presence predictor uses statistic $F_x$ from item’s profile $P_i$ with a threshold level of $F_x = 0.5$. We also define the “Item Average” baseline for the aspect sentiment prediction in a similar way based on the average sentiment of aspect $x$ in item i’s historical reviews (statistic $S_x$ in the item’s profile $P_i$).

In order to show that the overall method works good we have to compare it with a certain baseline that constitutes the whole process of prediction unknown reviews starting with a set of historical reviews. However, nobody addressed this particular problem before. There are some close works [5, 13] where authors built probabilistic models in order to predict ratings based on estimated aspects and sentiments. Their models could be transformed for somehow to produce predictions of a review, but it’s hard to say anything about the accuracy of such transformation, since their main goal is rating predictions and not the reviews. Therefore, we focus only three baselines described above for aspect presence and aspect sentiment prediction steps of our method.

We use the following measures in our comparison study:

- **Jaccard similarity coefficient** computes the standard similarity measure between the set of predicted aspects and the set of real aspects presented in a review. The Jaccard measure for the particular predictor is the average Jaccard similarity coefficient computed over all the reviews in the test set.

- $F_1$ and $F_0$ compute the standard $F_1$ score as the harmonic mean of precision and recall measures predicting the “presence” and “absence” classes

- $A(F1)$ and $H(F1)$ – compute the average and the harmonic means for the $F_1$ and the $F_0$ measures

- **Receiver Operating Characteristic (ROC)** - the standard ROC curve measure.

We use the same set of measures in case of the aspect sentiment prediction, where $F_1$ and $F_0$ stand for $F_1$ predicting score of “positive” and “negative” classes respectively.

In the next section we present the obtained results.
4 Results

We applied the method described in Section 2 on the Yelp’s restaurant and beauty&spas applications. The results of different aspect presence predictions for the restaurants application are presented in Table 3. Our findings show that Random Forest (RF), Matrix Factorization (MF) and All-Aspects-Included (AAI) predictors statistically outperform Random and Item-Average (IA) predictors in terms of the Jaccard measure. Further, there are no statistically significant differences between these three predictors. In addition, Figure 3a presents the distributions of Jaccard coefficient for RF and IA predictors over the test set of reviews. It also shows that the RF prediction tends to get higher Jaccard similarity coefficient than IA prediction.

Further, AAI, RF and MF are also comparable in terms of the $F_{1,\text{presence}}$ ($F_{1}$) mea-
sure, but AAI does not predict the “absence” class at all, which is actually very important in our study. Therefore, the RF predictor outperforms AAI in terms of the $Avg(F1)$ and $Harmonic(F1)$ measures. Moreover, RF outperforms MF in terms of these measures and, therefore, constitutes the best predictor for the Aspect Presence model. In addition, the ROC curves presented on Figure 3b, also show that RF prediction based on user’s and item’s profiles outperforms other approaches.

The results of the aspect presence prediction in the beauty&spa application are presented in Table 3 and Figure 4. In particular, they show that RF model is comparable with others in terms of $Jaccard$ measure and outperforms other methods in terms of $Avg(F1)$, $Harmonic(F1)$ and ROC curve. Therefore, this results confirm the advantage of using our profile-based method with the RF classification model.

The results of applying aspect sentiment models to predicting restaurant sentiments are presented in Table 4. The first column of Table 6 shows that the All-Positive (AP) predictor outperforms others for this measure. Although the 0.857 performance level of the AP predictor is high in comparison to others, it is quite natural because more than 80% of sentiments are indeed positive in the test reviews. Note, however, that the prediction quality of the AP classifier on the negative class is extremely poor, i.e. $F1_{negative} = 0$, as second column of Table 4 shows. In fact, the MF approach outperforms other methods
Table 4: Restaurants: “Aspect Sentiment” prediction quality

<table>
<thead>
<tr>
<th>Predictor</th>
<th>$F_{1_1}$</th>
<th>$F_{1_0}$</th>
<th>$avg(F1)$</th>
<th>$Harm(F1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>0.857</td>
<td>0.000</td>
<td>0.428</td>
<td>0.000</td>
</tr>
<tr>
<td>Random</td>
<td>0.561</td>
<td>0.413</td>
<td>0.487</td>
<td>0.475</td>
</tr>
<tr>
<td>IA</td>
<td>0.794</td>
<td>0.162</td>
<td>0.478</td>
<td>0.269</td>
</tr>
<tr>
<td>RF</td>
<td>0.627</td>
<td>0.403</td>
<td>0.515</td>
<td>0.490</td>
</tr>
<tr>
<td>MF</td>
<td>0.664</td>
<td>0.434</td>
<td><strong>0.549</strong></td>
<td><strong>0.524</strong></td>
</tr>
</tbody>
</table>

Note: The worst performer is labeled with boldface.

in predicting negative sentiments, as the second column of Table 4 shows. Moreover, MF approach outperforms other methods in terms of $Avg(F1)$ and $Harmonic(F1)$. This means that the MF approach outperforms others in predicting sentiments. This is confirmed further in Figure 5, where the ROC curve of the MF approach performs better than other methods for the aspect sentiment prediction problem for the restaurants application. Note that the situation is different for the beauty&spa application where the IA method slightly outperforms others. This is the case because of the small sizes of training and testing sets, which makes it harder to train other models, whereas the IA does not require large volumes of data to achieve good prediction results.

5 Conclusions

In this paper, we present a new method of estimating unknown reviews of items that a user may produce that is based on the sentiment analysis and machine learning techniques. The proposed method estimates which aspects of an item the user would mention in a new review and what sentiments he or she would express about these aspects. One of the distinguishing features of the proposed method is that it relies exclusively on the reviews and does not use any ratings or rankings data. The proposed method can also be used for providing explanations of why particular items would be of interest to the users.

We tested the proposed method on the Yelp reviews of restaurants and beauty&spas and showed that our method compares favorably with three baseline approaches. In particular, we have shown that for the aspect prediction problem on large datasets (such as Restaurants), the profile-based approach with Random Forest classification works the
best in terms of various classification measures. For the sentiment prediction problems on large datasets, the Matrix Factorization method works the best in terms of various classification measures. For the smaller datasets, such as beauty&spas, simpler and more robust methods, such as Item Average work the best because they are less sensitive to the problem of training the machine learning models on smaller datasets.

The contributions of this paper lie in proposing a novel review estimation method, developing simple and powerful explanations of why users may be interested (or disinterested) in particular items, and testing the proposed method on the actual reviews.

These tests produced reasonably good performance results, e.g., $F_1$ performance measure being in the $0.6 - 0.75$ range and the Jaccard coefficient in the $0.4 - 0.55$ range. Although not “spectacular” in comparison to other predictive modeling applications, these results are “reasonable” because the problem of aspect and sentiment prediction is a difficult one for the following reason. Most of the users do not visit many restaurants that often. Therefore, a user does not really know all the aspects of a restaurant well and thus cannot produce a comprehensive review of an average restaurant covering all the relevant aspects of the establishment, including those that can be of interest to him/her. For example, if a user likes a certain fish preparation and this fish is served in the restaurant that the user visits, it does not mean that the user would order that fish in that restaurant and, therefore, mention it in the review. This is one of the reasons why comprehensive predictions of all the right aspects in an average review are difficult, and our results of $0.6 - 0.7$ for $F_1$ and $0.4 - 0.55$ for Jaccard measures are “reasonable,” as compared to the baselines used in our study.

Although we focus only on the estimation of unknown reviews in this paper, we are planning to use the proposed method in developing the review-based recommendations as a part of our future work. In particular, we plan to develop techniques of ranking items based on the estimated reviews. We also plan to compare the proposed recommendation methods with the rating-based approaches and develop novel methods that combine
estimated reviews and ratings into one recommendation model. Finally, we would like to compare the performance of the considered review-based recommendations with the rating based approaches. Unfortunately, the only good method to do this comparison is via A/B testing, and we are currently exploring ways to accomplish this task.

References


[20] Suresh, V., Roohi, S., and Eirinaki, M. Aspect-based opinion mining and recommendationsystem for restaurant reviews. RecSys ’14, ACM.


