HelpMeter: A Nonlinear Model for Predicting the Helpfulness of Online Reviews

Yang Liu, Xiangji Huang, Aijun An, Xiaohui Yu
Department of Computer Science and Engineering
York University
{yliu, jhuang, aan, xhyu}@cse.yorku.ca

Abstract

With the flourish of the Internet, online review mining has attracted a lot of attention from the research community. However, compared to various well-studied sentiment analysis and opinion summarization problems, less effort has been made to analyze the quality of online reviews. The objective of this paper is to fill in this gap by automatically evaluating the “helpfulness” of reviews and consequently developing novel models to identify the most helpful reviews for a particular product. In particular, based on a thorough analysis of various factors that may affect the review quality, we propose HelpMeter, a nonlinear regression model for helpfulness prediction. Some preliminary experiments were conducted on a movie review data set, and the performance results confirm the superiority of the proposed method.

1 Introduction

The rapid growth of Web 2.0 has dramatically changed the way that customers express their opinions and interact with others. They are now encouraged to post comments of products at merchant websites and share their experience and knowledge with others via blogs and forums. In contrast to product descriptions provided by vendors, these customer reviews are naturally more user-oriented: in a review, customers describe a product in terms of purchasing scenarios and evaluate the product from a user’s point of view. Despite the fact that consumer’s evaluations can be very subjective, these comments are often considered more trustworthy than other traditional information sources. However, due to the large number of online reviews, sifting through the reviews to formulate an unbiased decision can be daunting and time-consuming.

Academics have recognized the importance of online product reviews and have produced some important results in review mining. However, most early work in this area was primarily focused on determining the semantic orientation of reviews [3, 7, 9, 10]. Compared to sentiment mining, identifying the quality of online reviews has received relatively less attention. A few recent studies along this direction attempt to detect the spams or low-quality posts that exist in online reviews. Jindal et al. [4] present a categorization of spam reviews, and propose some novel strategies to detect different types of spams. Liu et al. [6] propose a classification-based approach to discriminate the low-quality reviews from others, in the hope that such a filtering strategy can be incorporated to enhance the task of opinion summarization. Our work can be considered complementary to those studies in that the spam filtering model can be used as a preprocessing step in our approach.

Besides, there are also studies focusing on investigating how the different content features may affect the quality of reviews [2, 5, 11]. In those studies, a variety of the factors that may affect the helpfulness of reviews are explored, but most of them are related to the contents of the reviews only. Some important factors which are essential to the helpfulness prediction, such as the expertise of the reviewers and the timeliness of the reviews, are still missing. Furthermore, most of those studies rely on off-the-shelf solutions, such as SVM and logistic regression, to model the factors, which may not account for the unique characteristics of each individual factor.

In this paper, we address those problems, and develop a novel model called HelpMeter to measure the helpfulness of reviews. In this model, we not only investigate the impact of the basic semantic content, such as syntactical features, but introduce other major factors that may affect the review helpfulness rating, e.g., the reviewer’s expertise and the timeliness. In addition, we explore the possibility of developing a prediction model that can capture the distinctive characteristics of various factors. To achieve that, we propose a non-linear regression model based on radial basis functions that accommodate all these major factors. We finally compare our model with some of the state-of-the-art methods on a real IMDB dataset, and the experimental results confirm the effectiveness of the proposed approach.
2 Problem Formulation and Observations

For a given review, our objective is to find “helpfulness” defined as the expected fraction of people who will find the review helpful, where $H$ is a number falling in the range $[0, 1]$, and greater values of $H$ imply higher helpfulness. In the training data, this value can be approximated by the tally attached to that review, which takes the form of “$x$ out of $y$ people found the following review helpful”, i.e., $H = \frac{x}{y}$. As an effective indicator of the public opinions, this evaluation metric has also been widely adopted in previous product review helpfulness studies [5, 11].

To develop the helpfulness prediction model, we have examined the reviews on multiple popular Websites, and our efforts reveal that the following are among the most important factors that may affect the value of $H$.

1. **Reviewer Expertise**: Product reviews often involve personal experience, thoughts, and concerns. Also, it is common that different reviewers demonstrate expertise on different types of products. Those preferences and expertise might be well reflected through reviews they compose, and should be considered when building the prediction model.

2. **Writing Style**: Due to the large variation of the reviewers’ background and language skills, the online reviews are of dramatically different qualities. Some reviews are highly readable and therefore tend to be more helpful, whereas some reviews are either lengthy but with few sentences containing author’s opinions, or snappy but filled with insulting remarks. A proper representation of such difference must be identified and factored into the prediction model.

3. **Timeliness**: In general, the helpfulness of a review may significantly depend on when it is published. For instance, research shows that a quarter of a motion picture’s total revenue comes from the first two weeks [1], which means a timely review might be especially valuable for users seeking opinions about the movie.

We have also considered other possible factors that may affect the helpfulness values, however, none of them shows clear correlation with the value of helpfulness, and the detailed examination is available elsewhere [8].

3 HelpMeter: A Model for Helpfulness Prediction

In this section, we propose a model that accounts for the above three important factors. Once trained, it can be used for predicting the helpfulness of a given review.

3.1 Modeling major factors

As discussed above, the helpfulness of a review can largely depend on three factors, i.e., reviewer expertise, writing style, and timeliness. In order to estimate the expertise, we first use the genres provided by IMDB to represent each movie. As an example, the movie Casino Royal is labeled by IMDB as "Action", "Adventure", and "Thriller", which can be used to represent the movie for our purpose. Formally, each movie is represented by a $m$-dimensional vector $x = (x_1, x_2, \ldots, x_m)$, where $m$ is the number of different genres available for all movies. The next step is to measure the similarity of a given movie to movies that have been reviewed by the same reviewer, and relate this measure to the helpfulness score. We choose to approximate the relationship using radial basis functions (RBFs)

$$
\phi(x|\mu, \Sigma) = f \left( \frac{(x - \mu)^T(x - \mu)}{\sigma^2} \right),
$$

where $f$ is the Gaussian function, $\Sigma$ is the metric, term $(x - \mu)^T \Sigma^{-1} (x - \mu)$ represents the Euclidean distance between the input $x$ and the center $\mu$ in the metric defined by $\Sigma$, and $\sigma$ is called the spread of the RBF. RBFs are a particular good choice for modeling expertise in that when we represent each movie using a feature vector based on its genres, each center can be considered as corresponding to one “cluster” of movies that are similar to each other in terms of their genres. The helpfulness of a given movie is thus the weighted sum of the distance between the movie to those centers. The reviewer’s expertise on different clusters of movies can thus be naturally captured in that similar movies will have similar distances to the centers with the similar helpfulness scores. If we were to predict the helpfulness of a review based solely on the reviewer expertise factor, then we would fit the following regression model on the training data.

$$
\hat{H}_1 = \sum_{i=1}^{k_1} u_i \phi(x|\mu_i, \sigma_i),
$$

where $\hat{H}_1$ is the estimated helpfulness score, $x$ is the feature vector representing the movie, $k_1$ is the number of centers in the RBF network, and $\mu_i, \sigma_i$ and $u_i$ are the center, spread, and the weight of the $i$-th RBF respectively.

As noted by previous study, shallow syntactical features like part-of-speech provide more predicting powers than deeper features at the lexical level [11]. Thus, we choose to label the part-of-speech of the words contained in the reviews with a fixed set of tags using LingPipe\(^1\), a suite of Java libraries for the linguistic analysis of natural language. For each review, we parse it using the LingPipe tagger, and

\[^1\text{http://alias-i.com/lingpipe/}\]
count the number of words with each tag. Those counts are further normalized by dividing them with the word count of the review. The resulting numbers form a vector \( \mathbf{y} \) to represent the review for modeling writing style, with each number corresponding to one dimension. Again, if we were to use a radial basis function network to model the relationship between the feature vector \( \mathbf{y} \) and the helpfulness of the review and predict the helpfulness solely based on the writing style, we have

\[
\hat{H}_2 = \sum_{i=1}^{k_2} \nu_i \psi(y|\nu_i, \xi_i),
\]

where \( \hat{H}_2 \) is the estimated helpfulness, \( \nu_i, \xi_i \) are the weight, center, and the spread of the \( i \)-th RBF respectively, and \( k_2 \) is the number of RBFs.

Based on our observation, we hypothesize that the helpfulness of a movie review is subject to exponential decay with respect to time. Therefore, we propose the following model for movie reviews if the prediction of helpfulness were to be done only based on the timeliness:

\[
\hat{H}_3 = e^{-\beta(t-t_0)+d},
\]

where \( \hat{H}_3 \) is the estimated helpfulness, \( t_0 \) is the release time of the movie, \( t \) is the time when the review is published, and \( \beta \) and \( d \) are parameters in the model to be estimated. Intuitively, \( \beta \) controls the rate of decay in the helpfulness as we move further away in time from the movie release.

### 3.2 The complete model

Now that we have built the regression model for each individual factor, we are ready to propose the complete model, which we call the HelpMeter model, to incorporate all of the above factors. The idea is to consider the helpfulness score a weighted sum of the three individual models as:

\[
\hat{H} = \sum_{i=1}^{k_1} u_i \phi(x|\mu_i, \sigma_i) + \sum_{i=1}^{k_2} \nu_i \psi(y|\nu_i, \xi_i) + w \cdot e^{-\beta(t-t_0)}
\]

where \( p, q, \) and \( r \) are the weights of the three components. The HelpMeter model given in Equation 5 makes it possible to capture all of the factors discussed in this section, with the weights \( \{u_i\}_{i=1}^{k_1}, \{\nu_i\}_{i=1}^{k_2} \) and \( w \) controlling the “contribution” of each factor to the helpfulness score.

In HelpMeter, the set of parameters include the weights, centers, spreads, and the decay rate. Values of \( k_1 \) and \( k_2 \) are supplied by the user. To obtain the values of these parameters, we can estimate them from the training data, such that the sum of squared error (SSE) between the true values and the model output values is minimized. This optimization can be done with the steepest descent method.

### 4 Empirical Study

#### 4.1 Experiment settings

We collected our movie review data from the publicly accessible IMDB Website. In this dataset, 94,919 reviews by 56,588 reviewers for 504 movies released in the United States during the period from January 6, 2006 to November 21, 2007, were considered. To model reviewer expertise, we collected the genre labels for each movie, and the total number of genres is 27. Note that we only collected reviews posted by reviewers from the US as it helps to ensure the consistency in the release time (it is common for a movie to be released on different dates in different countries). To ensure the robustness of the prediction model, we only use those reviews with at least 10 votes, and reviewers with at least 25 posts. The number of such movie reviews is 22,819.

We use 10-fold cross validation to evaluate our approach. We evaluate the effectiveness of the proposed model using the Mean Squared Error (MSE), which was used in previous literatures [11] to measure the prediction accuracy of product reviews.

#### 4.2 Parameter selections

In the HelpMeter model, there are two user-chosen parameters that provide the flexibility to fine tune the model for optimal performance, i.e., the number of RBFs in the RBF network \( k_1 \) and \( k_2 \). We now study how the choice of these parameter values affects the prediction accuracy.

We first vary the value of \( k_1 \), and observe that there is a large improvement in accuracy when \( k_1 \) increases from 1 to 2, and the model achieves its best performance of \( MSE = 0.0332 \) at \( k_1 = 3 \). This implies that introducing multiple components to analyze the reviewer expertise can greatly improve the prediction accuracy. However, after \( k_1 \) past a threshold, the accuracy tends to decrease. This might be due to over-fitting the training data with more RBFs. Nonetheless, the accuracy remains stable for a wide range of \( k_1 \) values, indicating the insensitivity of the model with respect to the choice of \( k_1 \) values. It is also worth noting that the trend in accuracy remains the same regardless of the choice of \( k_2 \).

Similarly, we fix the values of \( k_1 \), and vary \( k_2 \) from 1 to 12. As shown in Figure 1 (b) there is also an optimal choice of \( k_2 \), which is 10. Similar to the case of \( k_1 \), the accuracy remains quite stable over a wide range of \( k_2 \), which again demonstrates that the model is not sensitive to the choice of parameter values.

#### 4.3 Comparing with alternative method

To show the effectiveness of the HelpMeter model, we compare the performance of our model with that of the
state-of-the-art Support Vector Machine regression method, which has been widely applied in previous studies [5, 11, 6]. For each review, we build a feature vector \((x, y, t)\) corresponding to each factor in the same way as described in Section 3 and concatenate them together to form one vector \(r\), and adopt the same settings as presented in [11].

<table>
<thead>
<tr>
<th>Method</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM Regression</td>
<td>0.1295</td>
</tr>
<tr>
<td>HelpMeter</td>
<td>0.0332</td>
</tr>
</tbody>
</table>

**Table 1. Comparison with other method**

As shown in Table 1, our proposed HelpMeter model shows clear advantage over the popular SVM regression model. We believe that the difference in accuracy is due to the fact that the various factors that affect the helpfulness rating have their own unique characteristics; therefore should be modeled separately. HelpMeter is more capable of such a task.

5 Conclusions

In this work, we’ve investigated automatically evaluating the quality of online reviews, which is orthogonal to the well-studies sentiment mining problem. We conducted a thorough analysis on various factors that may affect the helpfulness of a product review. Our observations revealed that there existed three major factors, namely, reviewer expertise, writing style, and timeliness, that could directly influence the performance of helpfulness prediction in the movie domain. We thus proposed a nonlinear model, HelpMeter, based on radial basis functions that took all those factors into account. Our empirical study on the IMDB dataset demonstrated the superiority of the proposed method.

In the future, we plan to investigate the possibility of incorporating other timeliness models to accommodate the situation that some product reviews, such as those on hotels, may evolve with the changes of product quality over time. In that scenario, a more recent review (i.e., reviews closer to the current time) may be more helpful than earlier reviews since certain aspects of the subject may change over time, which makes the earlier reviews outdated; we thus need to model the factor of timeliness differently.

**Acknowledgments**

This research is supported in part by research grant from NSERC of Canada and the Early Researcher Award from the Ministry of Research Innovation of Ontario.

**References**


