Exploiting Social Context for Review Quality Prediction

Yue Lu
UIUC Computer Science
Urbana, IL, USA
yuelu2@uiuc.edu

Panayiotis Tsaparas
Microsoft Research
Mountain View, CA, USA
panats@microsoft.com

Alexandros Ntoulas
Microsoft Research
Mountain View, CA, USA
antoulas@microsoft.com

Livia Polanyi
Microsoft Corporation
San Francisco, CA, USA
lipolany@microsoft.com

ABSTRACT
Online reviews in which users publish detailed commentary about their experiences and opinions with products, services, or events are extremely valuable to users who rely on them to make informed decisions. However, reviews vary greatly in quality and are constantly increasing in number, therefore, automatic assessment of review helpfulness is of growing importance. Previous work has addressed the problem by treating a review as a stand-alone document, extracting features from the review text, and learning a function based on these features for predicting the review quality. In this work, we exploit contextual information about authors’ identities and social networks for improving review quality prediction. We propose a generic framework for incorporating social context information by adding regularization constraints to the text-based predictor. Our approach can effectively use the social context information available for large quantities of unlabeled reviews. It also has the advantage that the resulting predictor is usable even when social context is unavailable. We validate our framework within a real commerce portal and experimentally demonstrate that using social context information can help improve the accuracy of review quality prediction especially when the available training data is sparse.

Categories and Subject Descriptors
H.3.m [Information Storage and Retrieval]: Miscellaneous; I.2.6 [Artificial Intelligence]: Learning

General Terms
Algorithms, Experimentation

Keywords
review quality, review helpfulness, social network, graph regularization

1. INTRODUCTION
Web 2.0 has empowered users to actively interact with each other, forming social networks around mutually interesting information and publishing large amounts of useful user-generated content online. One popular and important type of such user-generated content is the review, where users post detailed commentary on online portals about their experiences and opinions on products, events, or services. Reviews play a central role in the decision-making process of online users for a variety of tasks including purchasing products, booking flights and hotels, selecting restaurants, and picking movies to watch. Sites like Yelp.com and Epinions.com have created a viable business as review portals, while part of the popularity and success of Amazon.com is attributed to their comprehensive user reviews. As online commerce activity continues to grow [9], the role of online reviews is expected to become increasingly important.

Unfortunately, the abundance of user-generated content comes at a price. For every interesting opinion, or helpful review, there are also large amounts of spam content, unhelpful opinions, as well as highly subjective and misleading information. Sifting through large quantities of reviews to identify high quality and useful information is a tedious, error-prone process. It is thus highly desirable to develop reliable methods to assess the quality of reviews automatically. Robust and reliable review quality prediction will enable sites to surface high-quality reviews to users while benefiting other popular applications such as sentiment extraction and review summarization [8, 7], by providing high-quality content on which to operate.

Automatic review quality prediction is useful even for sites providing a mechanism where users can evaluate or rate the helpfulness of a review (e.g., Amazon.com and Epinions.com). Not all reviews receive the same helpfulness evaluation [10]. There is a rich-get-richer effect [11] where the top reviews accumulate more and more ratings, while more recent reviews are rarely read and thus not rated. Furthermore, such helpfulness evaluation is available only within a specific Web site, and is not comparable across different sources. However, it would be more useful for users if reviews from different sources for the same item could be aggregated and rated automatically on the same scale. This need is addressed by a number of increasingly popular aggregation sites such as Wise.com. For these sites, automatic review rating is essential in order to meaningfully present the collected reviews. Moreover, previous work [17, 10, 11, 6, 12, 15] attempts to solve the problem of review evaluation by treating each review as a stand-alone text document, extracting features from the text and learning a function based on these features for predicting review quality. However, in addition to textual content, there is much more information available that is useful for this task. Online reviews are produced by identifiable authors (reviewers) who interact with one another to form social networks. The history of reviewers and their social network interactions provide a social context for the reviews. In our approach, we mine combined textual, and social context information to evaluate the quality of individual reviewers and to assess the quality of the reviews.

Work done while at Microsoft Research.

Copyright is held by the International World Wide Web Conference Committee (IW3C2). Distribution of these papers is limited to classroom use, and personal use by others.
WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.
In this paper, we investigate how the social context of reviews can help enhance the accuracy of a text-based quality predictor. To the best of our knowledge, this is the first time that textual, author and social network information are combined for assessing review quality. Expressed very generally, our idea is that social context reveals a lot about the quality of reviewers, which in turn affects the quality of the reviews. We formulate hypotheses that capture this intuition and then mathematically model these hypotheses by developing regularization constraints which augment text-based review quality prediction. The resulting quality predictor is formulated into a well-formed convex optimization problem with efficient solution. The proposed regularization framework falls under the category of semi-supervised learning, making use of a small amount of labeled data as well as a large amount of unlabeled data.

It also has the advantage that the learned predictor is applicable to any review, even reviews from different sources or reviews for which the reviewer’s social context is not available. Finally, we experiment with real review data from an online commerce portal. We test our hypotheses and show that they hold for all three categories of data we consider. We then experimentally demonstrate that our novel regularization methods that combine social context with text information can lead to improved accuracy of review quality prediction, especially when the available training data is sparse.

The remainder of our paper is structured as follows. We first formally define the problem in Section 2. In Section 3 we present a text-based quality predictor which we use as our baseline. In Section 4, we outline our proposed methods for exploiting social context, formulate our hypotheses, and provide the mathematical modeling. In Section 5 we experimentally validate our hypotheses, evaluate the prediction performance of our methods and compare against baselines. Finally, we go over the related work in Section 6 and conclude in Section 7.

2. PROBLEM DEFINITION

A review system consists of three sets of three different types of entities: a set \( I = \{i_1, \ldots, i_N\} \) of \( N \) items (products, events, or services); a set \( R = \{r_1, \ldots, r_N\} \) of \( N \) reviews over these items; and a set \( U = \{u_1, \ldots, u_m\} \) of \( m \) reviewers (or users) that have authored these reviews. Each entity has a set of attributes \( T \) associated with it. For an item \( i \) or a user \( u \), \( T_i \) and \( T_u \) are sets of attribute-value pairs describing the item and the user respectively while for a review \( r \), \( T_r \) is the text of the review. We also give relationships between these sets of entities. There is a function \( M : R \to I \) that maps each review \( r \) to a unique item \( i = M(r) \); an authorship function \( A : R \to U \), that maps each review \( r \) to a unique reviewer \( u = A(r) \); and a relation \( S \subset U \times U \) that defines the social network relationships between users.

Since each review is associated with a unique item, we omit the set \( I \), unless necessary, and assume all information about the item \( i \) (item identifier and attributes) is included as part of the attributes \( T_r \) of review \( r \). We also model the social network relation as a directed graph \( G_S = (U, S) \) with adjacency matrix \( S \), where \( S_{uv} = 1 \) if there is a link or edge from \( u \) to \( v \) and zero otherwise. We assume that the links between users in the social network capture semantics of trust and friendship: the meaning of user \( u \) linking to user \( v \) is that \( v \) values the opinions of user \( u \) as a reviewer.

The information about the authors of the reviews along with the social network of the reviewers places the reviews within a social context. More formally we have the following definition.

**Definition 1 (Social Context).** Given a set of reviews \( R \), we define the social context of the set \( R \) as the triple \( C(R) = (U, A, S) \), of the set of reviewers \( U \), the authorship function \( A \), and the social network relation \( S \).

The set of reviews \( R \) contains both labeled \((R_L)\) and unlabeled \((R_U)\) reviews. For each review \( r_i \in R_L \), in the labeled subset of reviews we observe a numeric value \( q_i \) that captures the true quality and helpfulness of the review. We use \( L = \{(r_i, q_i)\} \), to denote the set of review-quality pairs. Such quality values can be obtained through manual labeling or through feedback mechanisms in place for some online portals.

Given the input data \( \{R_U, U, R_L, C(R), L\} \), we want to learn a quality predictor \( Q \) that, for a review \( r \), predicts the quality of the review. A review \( r \) is represented as an \( f \)-dimensional real vector \( r \) over a feature space \( F \) constructed from the information in \( R \) and \( C(R) \). The quality predictor is a function \( Q : R^f \to \mathbb{R} \) that maps a review feature vector to a numerical quality value.

Previous work has used the information in \( \{R_L, L\} \) for learning a quality predictor, based mostly on different kinds of textual features. In this paper, we investigate how to enhance the quality predictor function \( Q \) using the social context \( C(R) \) of the reviews in addition to the information in \( \{R_L, L\} \). Our exploration for the prediction function \( Q \) takes the following steps. First we construct a text-based baseline predictor that makes use of only the information in \( \{R_L, L\} \). Then we enhance this predictor by adding social context features that we extract from \( C(R_L) \). In the last step, which is the focus of this paper, we propose a novel semi-supervised technique that makes use of the labeled data \( \{R_L, L\} \), the unlabeled data \( R_U \), and the social context information \( C(R) \) for both labeled and unlabeled data.

3. TEXT-BASED QUALITY PREDICTION

The text of a review provides rich information about its quality. In this section, we build a baseline supervised predictor that makes use of a variety of textual features as detailed in the top part of Table 1. We group the features into four different types.

- **Text-statistics features:** This category includes features that are based on aggregate statistics over the text, such as the length of the review, the average length of a sentence, or the richness of the vocabulary.
- **Syntactic Features:** This category includes features that take into account the Part-Of-Speech (POS) tagging of the words in the text. We collect statistics based on the POS tags to create features such as percentage of nouns, adjectives, punctuations, etc.
- **Conformity features:** This category compares a review \( r \) with other reviews by looking at the KL-divergence between the unigram language model \( T_i \) of the review \( r \) for item \( i \), and the unigram model \( T_i \) of an “average” review that contains the text of all reviews for item \( i \). This feature is used to measure how much the review conforms to the average and is defined as \( D_{KL}(T_i | T_\bar{i}) = \sum_w T_i(w) \log(T_i(w) / T_\bar{i}(w)) \) where \( w \) takes values over the tokens of the unigram models.
- **Sentiment features:** This category considers features that take into account the positive or negative sentiment of words in the review. The occurrence of such words is a good indication about the strength of the opinion of the reviewer.

With this feature set \( F \), we can now represent each review \( r \) as an \( f \)-dimensional vector \( r \). Given the labeled data in \( \{R_L, L\} \), we want to learn a function \( Q : R^f \to \mathbb{R} \) that for a review \( r \), it predicts a numerical value \( \hat{q} \) as its quality. We formulate the problem as a linear regression problem, where the function \( Q \) is defined as a linear combination of the features in \( F \). More formally, the function \( Q \) is fully defined by an \( f \)-dimensional column weight vector \( w \), such that \( Q(r) = w^T r \), where \( w^T \) denotes the transpose of the vector. In the following, since \( Q \) is uniquely determined by
4. INCORPORATING SOCIAL CONTEXT

In Section 3, our work is based on the following two premises:

1. The quality of a review depends on the quality of the reviewer. Estimating the quality of the reviewer can help in estimating the quality of the review.

2. The quality of a reviewer depends on the quality of their peers in the social network. We can obtain information about the quality of the reviewers using information from the quality of their friends in their social network.

We investigate two different ways of incorporating the social context information into the linear quality predictor. The first is a straightforward expansion of the feature space to include features extracted from the social context. The second approach is novel in that it defines constraints between reviews, and between reviewers, and adds regularizers to the linear regression formulation to enforce these constraints. We describe these two approaches in detail in the following sections.

4.1 Extracting features from social context

A straightforward use of the social context information is by extracting additional features for the quality predictor function. The social context features we consider are shown in the bottom part of Table 1. The features capture the engagement of the author (ReviewNum), the historical quality of the reviewer (AvgRating), and the status of the author in the social network (In/Out-Degree, PageRank).

This approach of using social context is simple and it fits directly into our existing linear regression formulation. We can still use Equation 2 for optimizing the function $Q$, which is now defined over the expanded feature set $F$. The disadvantage is that such information is not always available for all reviews. Consider for example, a review written anonymously, or a review by a new user with no history or social network information. Predicting using social network features is no longer applicable. Furthermore, as the dimension of features increases, the necessary amount of labeled training data to learn a good prediction function also increases.

4.2 Extracting constraints from social context

We now present a novel alternative use of the social context that does not rely on explicit features, but instead defines a set of constraints for the text-based predictor. These constraints define hypotheses about how reviewers behave individually or within the social network. We require that the quality predictor respects these constraints, forcing our objective function to take into account relationships between reviews, and between different reviewers.

4.2.1 Social Context Hypotheses

We now describe our hypotheses, and how these hypotheses can be used in enhancing the prediction of the review quality. In Section 5 we validate them experimentally on real-world data, and we demonstrate that they hold for all the three data sets we consider.

**Author Consistency Hypothesis**: The hypothesis is that reviews from the same author will be of similar quality. A reviewer that writes high quality reviews is likely to continue writing good reviews, while a reviewer with poor reviews is likely to continue writing poor reviews.

**Trust Consistency Hypothesis**: We make the assumption that a link from a user $u_1$ to a user $u_2$ is an explicit or implicit statement of trust. The hypothesis is that the reviewers trust other reviewers in a rational way. In this case, reviewer $u_1$ trusts reviewer $u_2$ only if the quality of reviewer $u_2$ is at least as high as that of reviewer $u_1$. Intuitively, we claim that it does not make sense for users in the social network to trust someone with quality lower than themselves.

**Co-Citation Consistency Hypothesis**: The hypothesis is that people are consistent in how they trust other people. So if two reviewers $u_1$ and $u_2$ are trusted by the same third reviewer $u_3$, then their quality should be similar.

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Type</th>
<th>Feature Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NumToken</td>
<td>Text-Stat</td>
<td>Total number of tokens.</td>
</tr>
<tr>
<td>NumSent</td>
<td>Text-Stat</td>
<td>Total number of sentences.</td>
</tr>
<tr>
<td>UniqWordRatio</td>
<td>Text-Stat</td>
<td>Ratio of unique words.</td>
</tr>
<tr>
<td>SentLen</td>
<td>Text-Stat</td>
<td>Average sentence length.</td>
</tr>
<tr>
<td>CapRatio</td>
<td>Syntactic</td>
<td>Ratio of capitalized sentences.</td>
</tr>
<tr>
<td>POS:ADJ</td>
<td>Syntactic</td>
<td>Ratio of adjectives.</td>
</tr>
<tr>
<td>POS:COMP</td>
<td>Syntactic</td>
<td>Ratio of comparatives.</td>
</tr>
<tr>
<td>POS:V</td>
<td>Syntactic</td>
<td>Ratio of verbs.</td>
</tr>
<tr>
<td>POS:RB</td>
<td>Syntactic</td>
<td>Ratio of adverbs.</td>
</tr>
<tr>
<td>POS:FW</td>
<td>Syntactic</td>
<td>Ratio of foreign words.</td>
</tr>
<tr>
<td>POS:SYM</td>
<td>Syntactic</td>
<td>Ratio of symbols.</td>
</tr>
<tr>
<td>POS:CD</td>
<td>Syntactic</td>
<td>Ratio of numbers.</td>
</tr>
<tr>
<td>KLall</td>
<td>Conformity</td>
<td>KL div $D_{KL}(P_{r_i}\parallel P_{r_j})$.</td>
</tr>
<tr>
<td>PosSEN</td>
<td>Sentiment</td>
<td>Ratio of positive sentiment words.</td>
</tr>
<tr>
<td>NegSEN</td>
<td>Sentiment</td>
<td>Ratio of negative sentiment words.</td>
</tr>
</tbody>
</table>

The solution we describe in Section 3 considers each review as a stand-alone text document. As we have discussed, in many cases we also have available the social context of the reviews, that is, additional information about the authors of the reviews, and their social network. In this section we discuss different ways of incorporating social context into the quality predictor we described in Section 3. Our work is based on the following two premises:

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Type</th>
<th>Feature Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReviewNum</td>
<td>Author</td>
<td>Num. of past reviews by the author.</td>
</tr>
<tr>
<td>AvgRating</td>
<td>Author</td>
<td>Past average rating for the author.</td>
</tr>
<tr>
<td>In-Degree</td>
<td>SocialNetwork</td>
<td>In-degree of the author.</td>
</tr>
<tr>
<td>Out-Degree</td>
<td>SocialNetwork</td>
<td>Out-degree of the author.</td>
</tr>
<tr>
<td>PageRank</td>
<td>SocialNetwork</td>
<td>PageRank score of the author.</td>
</tr>
</tbody>
</table>

Table 1: Textual Features and Social Context Features
**Link Consistency Hypothesis:** The hypothesis is that if two people are connected in the social network (u₁ trusts u₂, or u₂ trusts u₁, or both), then their quality should be similar. The intuition is that two users that are linked to each other in some way, are more likely to share similar characteristics than two random users. This is the weakest of the four hypotheses but we observed that it is still useful in practice.

### 4.2.2 Exploiting hypotheses for regularization

We now describe how we enforce the hypotheses defined above by designing regularizing constraints to add into the text-based linear regression defined in Section 3.

**Author Consistency:** We enforce this hypothesis by adding a regularization term into the regression model where we require that the quality of reviews from the same author is similar. Let \( A_{ij} \) be the feature-review matrix defined as \( A_{ij} = 1 \) if review \( r_i \) and review \( r_j \) are authored by the same reviewer, and zero otherwise. Then, Equation 3 becomes:

\[
\Omega_2(w) = \frac{1}{n_r} \sum_{i=1}^{n_r} (w^T r_i - q_i)^2 + \alpha w^T w + \beta \sum_{i<j} A_{ij} (w^T r_i - w^T r_j)^2
\]

Let \( R = [r_1, ..., r_n] \) be an \( f \times n \) feature-review matrix defined over all reviews (both labeled and unlabeled). Then the last regularization constraint of Equation 4 can be written as:

\[
\sum_{i<j} A_{ij} (w^T r_i - w^T r_j)^2 = w^T R_{\Delta A} R^T w
\]

**Co-Citation Consistency:** Let \( u \) be a reviewer. Given a review quality predictor function \( Q \), we define the reviewer quality \( Q(u) \) as the average quality of all the reviews authored by this reviewer as it is estimated by our quality predictor. That is,

\[
\hat{Q}(u) = \frac{\sum_{r \in R_u} Q(r)}{|R_u|} = \sum_{r \in R_u} w^T r_i
\]

We enforce the trust consistency hypothesis by adding a regularization constraint to Equation 2. Let \( N_u \) denote the set of reviewers that are connected to by reviewer \( u \). We have

\[
\Omega_2(Q) = \Omega_1(Q) + \beta \sum_{u \in U} \sum_{r_i \in R_u} (Q(r_i) - \hat{Q}(u))^2
\]

The regularization term is greater than zero for each pair of reviewers \( u_1 \) and \( u_2 \) where \( u_1 \) trusts \( u_2 \), but the estimated quality of \( u_1 \) is greater than that of \( u_2 \). Minimizing function \( \Omega_2 \) will push such cases closer to zero, forcing the quality of a reviewer \( u_1 \) to be no more than that of \( u_2 \), and thus enforcing the trust consistency hypothesis.

Formally, for a reviewer \( u \), let \( h_u \) be the \( n \)-dimensional normalized indicator vector where \( h_u(i) = 1/|R_u| \) if user \( u \) has written review \( r_i \), and zero otherwise. Then we have that \( Q(u) = w^T R_h u \). We can thus write the objective function as

\[
\Omega_3(w) = \frac{1}{n_r} \sum_{i=1}^{n_r} (w^T r_i - q_i)^2 + \alpha w^T w + \beta \sum_{u,v \in U} S_{uv} \max \left \{ 0, w^T R_h u - w^T R_h v \right \}^2
\]

where \( S \) is the social network matrix. The optimization problem is still convex, but due to the max function, no nice closed form solution exists. We can still solve it and find the global optimum by gradient descent, where the gradient of the objective function is

\[
\frac{\partial \Omega_3(w)}{\partial w} = \frac{1}{n_r} \sum_{i=1}^{n_r} r_i r_i^T w - \frac{1}{n_r} \sum_{i=1}^{n_r} r_i q_i + \alpha w + \beta \sum_{u,v \in U} S_{uv} R(h_u - h_v)(h_u - h_v)^T R^T w
\]

Let \( H = [h_1, ..., h_n] \) be an \( n \times n \) matrix defined over all reviewers and \( Z \) be a new matrix such that

\[
Z_{uv} = \left \{ \begin{array}{cc} S_{uv} & \text{if } \left [ \text{diag}(w^T RH) \right ]_u v > 0 \\ 0 & \text{otherwise} \end{array} \right \}
\]

Now we can rewrite the gradient as

\[
\frac{\partial \Omega_3(w)}{\partial w} = \frac{1}{n_r} \sum_{i=1}^{n_r} r_i r_i^T w - \frac{1}{n_r} \sum_{i=1}^{n_r} r_i q_i + \alpha w + \beta R H Z H^T R^T w
\]

where \( H = D_Z + D_Z^T - Z \) can be thought of the graph Laplacian generalized for directed graphs with \( D_Z \) and \( D_Z^T \) the diagonal matrices of the row, and column sums of \( Z \) respectively.

**Co-Citation Consistency:** We enforce this hypothesis by adding a regularization term into the regression model, where we require that the quality of reviews authored by two co-cited reviewers is similar. Then, the objective function (Equation 2) becomes:

\[
\Omega_4(Q) = \Omega_1(Q) + \beta \sum_{u \in U} \sum_{r_i \in R_u} (Q(r_i) - \hat{Q}(u))^2
\]

Minimizing function \( \Omega_4 \) will cause the quality difference of reviewers \( x \) and \( y \) to be pushed closer to zero, making them more similar. We can again formulate these constraints as a graph regularizer. Let \( C \) be the co-citation graph adjacency matrix, where \( C_{ij} = 1 \) if two reviewers \( u_i \) and \( u_j \) are both trusted by at least one other reviewer \( u \). Using the same definition of matrix \( R \) and vector \( h_u \) as for trust consistency, the objective function now becomes

\[
\Omega_4(w) = \frac{1}{n_r} \sum_{i=1}^{n_r} (w^T r_i - q_i)^2 + \alpha w^T w + \beta \sum_{i<j} C_{ij} (w^T R_h u_i - w^T R_h u_j)^2
\]

Let \( D_C \) be the Laplacian of graph \( C \). The closed form solution is

\[
w = \left ( \sum_{i=1}^{n_r} r_i r_i^T + \alpha n_t I + \beta n_t R H C e H^T R^T \right )^{-1} \sum_{i=1}^{n_r} r_i q_i
\]
5. EXPERIMENTS

5.1 Data Sets

Our experiments employ the data from Ciao UK\(^1\), a community review web site. In Ciao, people not only write critical reviews for all kinds of products and services, but also rate the reviews written by others. Furthermore, people can add members to their network of trusted members or “Circle of Trust”, if they find their reviews consistently interesting and helpful.

We collected reviews, reviewers, and ratings up to May, 2009 for all products in three categories: Cellphones, Beauty, and Digital Cameras (DC). We use the average rating of the reviews (a real value between 0 and 5) as our gold standard of review quality. In order for the gold standard to be robust and resistant to outlier raters, we use only reviews with at least five ratings from different raters. We then apply some further pruning by imposing the conditions shown in the top part of Table 2. The purpose of the pruning is to obtain a dataset that is both large enough and has sufficient social context information. Because we need some information about reviewers’ history in order to test our Reviewer Consistency hypothesis, we require reviewers for Cellphone and Beauty to have at least two reviews each. We also require reviewers to be part of the 

\(^1\)http://www.ciao.co.uk/

<table>
<thead>
<tr>
<th>PRUNING SETTINGS</th>
<th>Cellphone</th>
<th>Beauty</th>
<th>Digital Camera</th>
</tr>
</thead>
<tbody>
<tr>
<td>min # of ratings/review</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>min # of reviews/reviewer</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>min # of trust links/reviewer</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>min # of reviews/ product</td>
<td>5</td>
<td>10</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>STATISTICS</th>
<th>Cellphone</th>
<th>Beauty</th>
<th>Digital Camera</th>
</tr>
</thead>
<tbody>
<tr>
<td># of reviews</td>
<td>1943</td>
<td>4849</td>
<td>3697</td>
</tr>
<tr>
<td># of products</td>
<td>2881</td>
<td>1709</td>
<td>3465</td>
</tr>
<tr>
<td># of links in Trust</td>
<td>158</td>
<td>308</td>
<td>380</td>
</tr>
<tr>
<td># of links in Cytation</td>
<td>2905</td>
<td>20374</td>
<td>3894</td>
</tr>
<tr>
<td># of links in Cocation</td>
<td>13678</td>
<td>32104</td>
<td>6022</td>
</tr>
<tr>
<td>Trust graph density</td>
<td>0.0075</td>
<td>0.0140</td>
<td>0.0006</td>
</tr>
<tr>
<td>Link graph density</td>
<td>0.0120</td>
<td>0.0220</td>
<td>0.0010</td>
</tr>
<tr>
<td>Cytation graph density</td>
<td>0.0353</td>
<td>0.1292</td>
<td>0.0037</td>
</tr>
<tr>
<td>Avg # of reviews/reviewer</td>
<td>2.2054</td>
<td>2.8373</td>
<td>1.0670</td>
</tr>
<tr>
<td>Ratio of Reciprocal links</td>
<td>0.4014</td>
<td>0.4243</td>
<td>0.4555</td>
</tr>
<tr>
<td>Clustering coefficient</td>
<td>0.2286</td>
<td>0.3072</td>
<td>0.2523</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHARACTERISTICS</th>
<th>Cellphone</th>
<th>Beauty</th>
<th>Digital Camera</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social Context</td>
<td>rich</td>
<td>rich</td>
<td>sparse</td>
</tr>
<tr>
<td>Quality Distribution</td>
<td>balanced</td>
<td>skewed</td>
<td>balanced</td>
</tr>
</tbody>
</table>

![Figure 1: Density Estimate of Gold Standard Review Quality.](image_url)
5.2 Consistency Hypotheses Testing

Before evaluating the prediction performance of different algorithms, we first validate our four consistency hypotheses over our data sets.

5.2.1 Author Consistency Hypothesis

For each dataset, we consider all $n^2$ pairs of reviews $(r_i, r_j)$, and we divide them into two disjoint groups: Rel:DifferentReviewer if $r_i$ and $r_j$ are authored by different reviewers, i.e., $u_i \neq u_j$, and Rel:SameReviewer if $u_i = u_j$. In each group, for each pair $(r_i, r_j)$ we compute the difference in quality, $dq_{ij} = q_i - q_j$, of the two reviews. Since for each value $dq_{ij}$ we also include value $-dq_{ij}$, the mean value of $dq_{ij}$ for both groups is zero. We are interested in the standard deviation, $std(dq_{ij})$, that captures how much variability there is in the difference of quality between reviews for the two groups. Table 3 shows the results for the different datasets. For a visual comparison, in Figure 2 we also plot the Kernel-smoothing density estimates of the two groups.

We observe that the standard deviation of the quality difference of two reviews by the same author is much lower than that of two reviews from different authors. This indicates that reviewers, to some extent, are consistent in the quality of reviews they write. The figures also clearly indicate that the density curve for Rel:SameReviewer is more concentrated around zero than that of Rel:DifferentReviewer. This indicates that reviewers are, to some extent, consistent in the quality of reviews they write. The figures also clearly indicate that the density curve for Rel:SameReviewer is more concentrated around zero than that of Rel:DifferentReviewer. This indicates that reviewers are, to some extent, consistent in the quality of reviews they write.

5.2.2 Social Network Consistency Hypotheses

In order to test the three social network consistency hypotheses, namely Trust Consistency, Co-Citation Consistency and Link Consistency, we look at the empirical distribution of $Q_i = \sum_{j \in R_u} q_j / |R_u|$, the difference in quality of two reviewers, where, similar to Equation 5

$$Q_i(u) = \frac{\sum_{i \in R_u} q_i}{|R_u|} \tag{9}$$

is defined as the average quality of the reviews written by $u$ in our dataset, but using gold standard quality. Again, we group the pairs of reviewers $(u_i, u_j)$ into the the following sets depending on the relationship between the two reviewers.

Rel:None: User $u_i$ is not linked to user $u_j$, i.e., $B_{ij} = 0$.
Rel:Trust: User $u_i$ trusts user $u_j$, i.e., $S_{ij} = 1$.
Rel:Cocitation: Users $u_i$ and $u_j$ are trusted by at least one other reviewer $u_k$, i.e., $C_{ij} = 1$.
Rel:Link: User $u_i$ trusts user $u_j$, or $u_j$ trusts $u_i$, i.e., $B_{ij} = 1$.

In Figure 3, we plot the Kernel-smoothing density estimate of the $Q_{ij}$ values for the four different sets of pairs, for the three categories. We further show in Table 4 the moments (mean and variance) of the four density estimates and p-values of the KS-test between pairs of density estimates.

The first observation is that the distribution of Rel:Trust is skewed towards the negative with a negative mean. This supports the Trust Consistency Hypothesis that when $u_i$ trusts $u_j$, the quality of $u_i$ is usually lower than that of $u_j$, i.e., $Q(u_i) - Q(u_j) < 0$. The remaining three distributions are all symmetric with mean zero. However, Rel:Cocitation and Rel:Link have a much more concentrated peak around zero, i.e., smaller variance, compared with Rel:None. This supports the Co-Citation and Link Consistency Hypotheses that reviewers are more similar in quality (quality difference closer to zero) if they are co-trusted by others, or linked in a trust graph regardless of direction.

In the results of the KS-test, we have only one high p-value, for Rel:Link and Rel:Cocitation, while all the other pairs have p-values close to zero. This implies that Rel:Trust, Rel:Cocitation, or Rel:Link do not come from the same distribution as Rel:None. This observation directly connects the quality of reviewers with their relations in the social network. The correlation between Rel:Link and Rel:Cocitation could potentially be explained by the relatively high reciprocity ratio (the percentage of links in the Trust social network that are reciprocal), and the relatively high clustering coefficient [14] which measures the tendency of triples to form triangles.

In summary, our experiments indicate that there exists correlation between review quality, reviewer quality, and social context. For all the three data sets considered, the statistics support our hypotheses for designing the regularizers.

5.3 Prediction Performance

For all three datasets (Cellphones, Beauty, and Digital Cameras), we randomly split the data into training and testing sets: 50% of the products for training ($R_{\text{train}}$), and 50% for testing ($R_{\text{test}}$). We keep the test data fixed, while sub-sampling from the training data to generate training sets of different sizes (10%, 25%, 50% or 100% of the training data). Our goal is to study the effect of different amount of training data on the prediction performance. We draw 10 independent random splits, and we report test set mean and standard deviation for our evaluation metrics. A polynomial kernel is

<table>
<thead>
<tr>
<th>Cellphone</th>
<th>Beauty</th>
<th>Digital Camera</th>
</tr>
</thead>
<tbody>
<tr>
<td>STD Cellphone</td>
<td>Beauty</td>
<td>Digital Camera</td>
</tr>
<tr>
<td>Rel:SameReviewer 0.00E+00</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>p-value 1.37E-48* 1.57E-287* 3.12E-11*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Statistics of Review Quality Difference to Support Reviewer Consistency Hypothesis

<table>
<thead>
<tr>
<th>Cellphone</th>
<th>Beauty</th>
<th>Digital Camera</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value Rel:None Rel:Trust Rel:Link Rel:Cocitation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rel:None -3.20E-827 4.3E-445 6.12E-177*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rel:Trust - - 3.4E-16* 6.89E-22* 0.0657</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rel:Link - - - - 0.3003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moments Rel:None Rel:Trust Rel:Link Rel:Cocitation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean 0.0000 -0.1376 0.0000 0.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance 0.6727 0.3255 0.3485 0.2914</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Statistics of Reviewer Quality Difference to Support Social Network Consistency Hypotheses.
used to enrich the feature representation for the linear model. We fix the parameter $\alpha$ of Linear Regression to the value that gives the best performance for the text-based baseline. Then, we report the best prediction performance by tuning the regularization weight $\beta$. We will discuss the parameter sensitivity in Section 5.3.3, while leaving the automatic optimization of parameters as future work.

We evaluate the effectiveness of different prediction methods using Mean Squared Error (MSE) over the test set $R_{test}$ of size $n_t$,

$$MSE(R_{test}) = \frac{1}{n_t} \sum_{i=1}^{n_t} (Q(r_i) - q_i)^2$$

MSE measures how much our predicted quality deviates from the true quality. A smaller value indicates a more accurate prediction.

### 5.3.1 Simple Text-free Baselines

Since the graph statistics in Section 5.2 support our design of regularizers, we will examine a few text-free baselines (TBL) that are based solely on social context. These baselines also serve as a sanity check for the experiments we report in the following section. For the following, $r$ denotes a test review written by reviewer $u_r$, and $Q^*(u)$ is the quality of reviewer $u$ as defined in Equation 9, when computed over the training data. If reviewer $u$ has no reviews in the training data, $Q^*(u)$ is undefined. We consider the following baselines for predicting the quality of $r$.

**TBL:Mean:** Simply predict as the mean review quality in the training data $R_{train}$, i.e., $Q(r) = \frac{1}{n_r} \sum_{i=1}^{n_r} q_i$.

**TBL:Reviewer:** Predict as the quality $Q^*(u_r)$ of the author $u_r$ in the training data. If it is not defined, predict as TBL:Mean.

**TBL:Link:** Predict as the mean quality of all the reviewers connected to $u_r$ in the link graph; if no such reviewer exists in the training set, the value is undefined simply predict as TBL:Mean.

**TBL:CoCitation:** Similar to TBL:Link, predict as the mean quality of all reviewers connected to $u_r$ in the Co-Citation graph. If this is not defined predict as TBL:Mean.

We compare the four simple text-free baselines against **BL:Text**: the Linear Regression baseline that uses only text information. Figure 4 shows the MSE with standard deviation where the $x$-axis corresponds to the different percentages of the training data we used. We observe that none of the text-free baselines works as well as Linear Regression with textual features, suggesting that social context by itself cannot accurately predict the quality of a review. The MSE of the text-free baselines is lower for the Beauty category, where quality distribution is highly skewed at 4, but the text-based predictor is still significantly better. Out of the three social-context based baselines, TBL:Reviewer appears to provide more accurate prediction than the other two when there is rich social context (Cellphones and Beauty), but it offers marginal improvements over TBL:Mean in the case where the social context is sparse (Digital Cameras). TBL:CoCitation consistently outperforms TBL:Link, which is in line with our observation in Table 4 that the variance of Rel:Cocitation is smaller than that of Rel:Link.

### 5.3.2 Incorporating Social Context

We now compare the different techniques for review quality prediction that make use of text and social context of reviews. We consider the following methods.

**BL:Text:** Linear Regression described in Section 3 (Equation 2) using only textual features.

**BL:Text+Rvr:** Linear Regression described in Section 4.1 using both textual, and social context features.

**REG:Reviewer:** Linear Regression with a regularizer under Reviewer Consistency Hypothesis (Equation 4).
We expect a similar trend for larger training data available, the better the performance. BL:Text+Rvr gives the best improvement for training percentage of 50% and 100% for all three categories. We expect a similar trend for larger amounts of training data. On the other hand, when there is little training data, the social context features are too sparse to be helpful, and it may be the case that the MSE actually increases, e.g., when training with 10% and 25% of the training data for Cellphone, and training with 10% for Digital Cameras. There are techniques for dealing with sparse data, however, exploring such techniques is beyond the scope of this paper.

Using social context as regularization (method names starting with REG) consistently improves over the text-only baseline. The advantage of the regularization methods is most significant when the training size is small, e.g., using training percentage of 10% and 25% of the training data for Cellphone, and there is 25% in all three data sets. This is often the case in practice, where the training size is small, e.g. using training percentage of 10% and 25% of the training data for Cellphone, and training with 10% for Digital Cameras. There are techniques for dealing with sparse data, however, exploring such techniques is beyond the scope of this paper.

The first observation is that adding social context as additional features BL:Text+Rvr can improve significantly over the text-only baseline when there is sufficient amount of training data. The more training data available, the better the performance. BL:Text+Rvr gives the best improvement for training percentage of 50% and 100% for all three categories. We expect a similar trend for larger amounts of training data. On the other hand, when there is little training data, the social context features are too sparse to be helpful, and it may be the case that the MSE actually increases, e.g., when training with 10% and 25% of the training data for Cellphone, and training with 10% for Digital Cameras. There are techniques for dealing with sparse data, however, exploring such techniques is beyond the scope of this paper.

Using social context as regularization (method names starting with REG) consistently improves over the text-only baseline. The advantage of the regularization methods is most significant when the training size is small, e.g., using training percentage of 10% and 25% of the training data for Cellphone, and training with 10% for Digital Cameras. There are techniques for dealing with sparse data, however, exploring such techniques is beyond the scope of this paper.

REG:Link: Linear Regression with a regularizer under Link Consistency Hypothesis (Equation 8).

REG:Cocitation: Linear Regression with a regularizer under Co-citation Consistency Hypothesis (Equation 7).

REG:Trust: Linear Regression with a regularizer under Trust Consistency Hypothesis (Equation 6)

It is possible to consider combinations of the different regularizers. This would introduce multiple $\beta$ parameters (one for each regularizer), and careful tuning is required to make the technique work. We defer the exploration of this idea to future work.

Among the different regularization techniques, for both Cellphone and Beauty reviews, where there is relatively rich social context information, REG:Reviewer appears to be the most effective. For the Cellphone dataset, REG:Reviewer outperforms BL:Text+Rvr even with 50% of training data, indicating that social context regularization can be helpful when we have rich social context and balanced data. Among the regularization methods using the social network, REG:Trust, which is based on the most reasonable hypothesis, performs best in practice. This means that the direction of the trust social network carries more useful information than the simplified undirected link graphs and co-citation graphs.
Finally, for the Digital Camera reviews where the social context is very sparse there is still some improvement observed using regularization when the training data is small, but the improvement is not as significant as on the other two categories where the social context is richer; that is exactly what we expected.

In addition to the experiments on our test data, we are interested in testing our algorithms on data for which we have no social context information. Our premise is that using regularization can help to incorporate signals from the social network to the text-based predictor, thus improving accuracy prediction even if social context is not available. We now validate this premise. We use the Cellphone dataset, and we consider the case where we train on 10% of the training data. Within the test data of Cellphone, there is a subset of data (144 reviews on average across splits) that has no social context information, i.e., the author has only one review, and is not in the social network. Regularization methods only adjust weights on textual features and are thus applicable to those anonymous reviews too, even though these reviews do not contribute to the added regularization terms. In Table 6, we report the percentage of improvement of four regularization methods over BL:Text. We still observe some improvement on anonymous reviews with no social context, although as expected less than on reviews with social context. This indicates the generalizability of the proposed regularization methods.

To further support the generalizability claim, we try an extra set of experiments testing our regularization methods on a held-out set of reviews which are not used in the optimization process and for which we use only the textual features and hide their social context. More specifically, after learning a quality prediction function \( \hat{Q} \) using 10% of the training data, we apply it to the remaining 90% of the training data, by multiplying the learned weight vector \( w \) with the text feature vectors of the held-out reviews. From the last row in Table 6, we can clearly see that compared with the text-only baseline, all regularization methods can learn a better weight vector \( w \) that captures more accurately the importance of textual features for predicting the true quality on the held-out set.

In summary, we make the following observations.

- Adding social context as features is effective only when there is enough training data to learn the importance of those additional features.
- On the other hand, regularization methods work best when there is little training data by exploiting the constraints defined by the social context and the large amount of unlabeled data.
- Since regularization techniques incorporate the social context information into the text-based predictor, they provide improvements even when applied to data without any social context.

5.3.3 Parameter Sensitivity

Regularization methods have one parameter \( \beta \) to set: the trade-off weight for the regularization term. The value of the regularization weight defines our confidence in the regularizer: a higher value results in a higher penalty when violating the corresponding regularization hypothesis. In the objective functions (Equations 4, 6, 7, and 8), the contribution from the regularization term depends on \( \beta \) as well as the number of non-zero edges in the regularization graph.

\(^2\)Although we prune the data by requiring that each reviewer has at least two reviews and a link in the social network, due to multiple consecutive pruning conditions some reviewers end up with only one review and no links in the final pruned subset.

6. RELATED WORK

The problem of assessing the quality of user-generated content has recently attracted increasing attention. Most previous work [17, 10, 11, 6, 12, 15] has typically focused on automatically determining the quality (or helpfulness, or utility) of reviews by using textual features. The problem of determining review quality is formulated as a classification or regression problem with users’ votes serving as the ground-truth. In this context, Zhang and Varadarajan [17] found that shallow syntactic features from the text of reviews are most useful, while review length seems weakly correlated with review quality. In addition to textual features, Kim et al. [10] included metadata features including ratings given to an item under review and concluded that review length and the number of stars in product rating are most helpful within their SVM regression model. Ghasse and Ipeiritos [6] combined econometric models with textual subjectivity analysis and demonstrated evidence that extreme reviews are considered to be most helpful. In [12], the authors incorporated reviewers’ expertise and review timeliness in addition to the writing style of the review in a non-linear regression model.

In our work, we extend previous work by using author and social network information in order to assess review quality.

Although user votes can be helpful as ground-truth data, Liu et al. [11] identified a discrepancy between votes coming from Amazon.com and votes coming from an independent study. More specifically, they identified a “rich-get-richer” effect, where reviews accumulate votes more quickly depending on the number of votes they already have. This observation further enhances our motivation to
automatically determine the quality of reviews in order to avoid such biases. Danescu-Niculescu-Mizil et al. [5] showed that the perceived helpfulness of a review depends not only on its content but also on the other reviews of the same product. We include one of their hypotheses, i.e. conformity hypothesis, as a feature into our model. A recent paper [15] took an un-supervised approach to finding the most helpful book reviews. Although their method is shown to outperform users’ votes, it is evaluated on only 12 books and thus is not clear whether it is robust and generalizable.

The problem of assessing the quality of user-generated data is also critical in domains other than reviews. For example, previous works [2, 4] focused on assessing the quality of postings within the community question/answering domain. The work in [2] combines textual features with user and community meta-data features for assessing the quality of questions and answers. In [4], the authors propose a co-training idea that jointly models the quality of the author and the review. However, their work does not model user relationships, but rather uses all community information for exacting features.

Regularization using graphs has appeared as a type of effective method in the semi-supervised learning literature [19]. The interested reader may examine [18, 20, 3]. The resulting formulation is usually a well-formed convex optimization problem which has a unique and efficiently computable solution. These types of graph regularization methods have been successfully applied in Web-page categorization [16] and Web spam detection [1]. In both cases, the link structure among Web pages is nicely exploited by the regularization which, in most cases, has improved the predictive accuracy within the problem at hand. Recently, Mei et al. [13] propose to enhance topic models by regularizing on a contextual graph structure. In our scenario, the social network of the reviewers defines the context, and we exploit it to enhance review quality prediction.

7. CONCLUSION AND FUTURE WORK

In this paper we studied the problem of automatically determining review quality using social context information. We studied two methods for incorporating social context in the quality prediction: either as features, or as regularization constraints, based on a set of hypotheses that we validated experimentally. We have demonstrated that prediction accuracy of a text-based classifier can greatly improve, when working with little training data, by using regularization on social context. Importantly, our regularization techniques make the general approach applicable even when social context information is unavailable. The method we propose is quite generalizable and applicable for quality (or attribute) estimation of other types of user-generated content. This is a direction that we intend to explore further.

As further future work, social context can be enhanced with additional information about items and authors. Information about product attributes, for example, enables estimates of similarity between products, or categories of products which can be exploited as additional constraints. Furthermore, although a portal may lack an explicit trust network, we plan to construct an implicit network using the ratings reviewers attach to each others’ reviews and then apply our techniques to this case. Finally, rather than predicting the quality of each review, it would be interesting to adapt our techniques for computing a ranking of a set of reviews.

8. REFERENCES