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 important 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.

Most previous work [17, 10, 11, 6, 12, 15] attempts to solve the problem of review evaluation by treating each review as a standalone 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.

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 $I = \{i_1, ..., i_N\}$ of N *items* (products, events, or services); a set $R = \{r_1, ..., r_n\}$ of n *reviews* over these items; and a set $U = \{u_1, ..., u_m\}$ 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 are also given relationships between these sets of entities. There is a function $M : R \to I$ that maps each review r to a unique item $i_r = M(r)$; an authorship function $A : R \to U$, that maps each review r to a unique reviewer $u_r = A(r)$; 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_r (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 $\mathbf{S}$, where $\mathbf{S}_{uv} = 1$ 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 u values the opinions of user v 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 $C(R) = \langle U, A, S\rangle$, 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 $\{R_L \cup R_U, C(R), L\}$, 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 $\mathbf{r}$ over a feature space F constructed from the information in R and $C(R)$. So the quality predictor is a function $Q : \mathbb{R}^f \to \mathbb{R}$ 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_r of the review r for item i, and the unigram model $\overline{T}_i$ 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 $D_{KL}(T_r||\overline{T}_i) = \sum_w T_r(w) \log(T_r(w)/\overline{T}_i(w))$ 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 $\mathbf{r}$. Given the labeled data in $\{R_L, L\}$, we want to learn a function $Q : \mathbb{R}^f \to \mathbb{R}$ that for a review $\mathbf{r}_i$ it predicts a numerical value $\hat{q}_i$ 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 $\mathbf{w}$, such that $Q(\mathbf{r}) = \mathbf{w}^T\mathbf{r}$, where $\mathbf{w}^T$ denotes the transpose of the vector. In the following, since Q is uniquely determined the by

Feature Name	Type	Feature Description
TEXT FEATURES		
NumToken	Text-Stat	Total number of tokens.
NumSent	Text-Stat	Total number of sentences.
UniqWordRatio	Text-Stat	Ratio of unique words
SentLen	Text-Stat	Average sentence length.
CapRatio	Text-Stat	Ratio of capitalized sentences.
POS:NN	Syntactic	Ratio of nouns.
POS:ADJ	Syntactic	Ratio of adjectives.
POS:COMP	Syntactic	Ratio of comparatives.
POS:V:	Syntactic	Ratio of verbs.
POS:RB	Syntactic	Ratio of adverbs.
POS:FW	Syntactic	Ratio of foreign words.
POS:SYM	Syntactic	Ratio of symbols.
POS:CD	Syntactic	Ratio of numbers.
POS:PP	Syntactic	Ratio of punctuation symbols.
KLall	Conformity	KL div $D_{KL}(T_r \| \overline{T}_i)$
PosSEN	Sentiment	Ratio of positive sentiment words.
NegSEN	Sentiment	Ratio of negative sentiment words.
SOCIAL NETWORK FEATURES		
ReviewNum	Author	Num. of past reviews by the author.
AvgRating	Author	Past average rating for the author.
In-Degree	SocialNetwork	In-degree of the author.
Out-Degree	SocialNetwork	Out-degree of the author.
PageRank	SocialNetwork	PageRank score of the author.

Table 1: Textual Features and Social Context Features

weight vector $\mathbf{w}$ and vice versa, we will use Q and $\mathbf{w}$ interchangeably. Our goal is to find the f-dimensional weight vector $\hat{\mathbf{w}}$ that minimizes the objective function:

$$\Omega(\mathbf{w}) = \frac{1}{n_\ell}\sum_{i=1}^{n_\ell}\mathcal{L}(\mathbf{w}^T\mathbf{r}_i, q_i) + \alpha\mathbf{w}^T\mathbf{w} \qquad (1)$$

where $\mathcal{L}$ is the loss function that measures distance of the predicted quality $Q(\mathbf{r}_i) = \mathbf{w}^T\mathbf{r}_i$ of review $r_i \in R_L$ with the true quality value q_i, n_ℓ is the number of training examples, and $\alpha \geq 0$ is regularization parameter for $\mathbf{w}$. In our work, we use squared error loss (or quadratic loss), and we minimize the function

$$\Omega_1(\mathbf{w}) = \frac{1}{n_\ell}\sum_{i=1}^{n_\ell}(\mathbf{w}^T\mathbf{r}_i - q_i)^2 + \alpha\mathbf{w}^T\mathbf{w} \qquad (2)$$

The closed form solution for $\hat{\mathbf{w}}$ is given by

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}}\Omega_1(\mathbf{w}) = (\sum_{i=1}^{n_\ell}\mathbf{r}_i\mathbf{r}_i^T + \alpha n_\ell\mathcal{I})^{-1}\sum_{i=1}^{n_\ell}q_i\mathbf{r}_i$$

where $\mathcal{I}$ is the identity matrix of size f.

Once we have learned the weight vector $\mathbf{w}$, we can apply it to any review feature vector and predict the quality of unlabeled reviews.

4. INCORPORATING SOCIAL CONTEXT

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:

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.

Link Consistency Hypothesis: The hypothesis is that if two people are connected in the social network (u_1 trusts u_2, or u_2 trusts u_1, 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 R_u denote the set of reviews authored by reviewer u, including both labeled and unlabeled reviews. Then the objective function becomes:

$$\Omega_2(Q) = \Omega_1(Q) + \beta \sum_{u \in U} \sum_{r_i, r_j \in R_u} (Q(\mathbf{r}_i) - Q(\mathbf{r}_j))^2 \qquad (3)$$

Minimizing the regularization constraint will force reviews of the same author u to receive similar quality values. We can formulate this as a graph regularization. The graph adjacency matrix $\mathbf{A}$ is defined as $\mathbf{A}_{ij} = 1$ if review r_i and review r_j are authored by the same reviewer, and zero otherwise. Then, Equation 3 becomes:

$$\begin{aligned} \Omega_2(\mathbf{w}) = \quad & \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} \left(\mathbf{w}^T \mathbf{r}_i - q_i \right)^2 + \alpha \mathbf{w}^T \mathbf{w} \\ + \quad & \beta \sum_{i<j} \mathbf{A}_{ij} \left(\mathbf{w}^T \mathbf{r}_i - \mathbf{w}^T \mathbf{r}_j \right)^2 \qquad (4) \end{aligned}$$

Let $\mathbf{R} = [\mathbf{r}_1, ..., \mathbf{r}_n]$ 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

$$\sum_{i<j} \mathbf{A}_{ij} \left(\mathbf{w}^T \mathbf{r}_i - \mathbf{w}^T \mathbf{r}_j \right)^2 = \mathbf{w}^T \mathbf{R} \Delta_A \mathbf{R}^T \mathbf{w}$$

$\Delta_\mathbf{A} = \mathbf{D}_\mathbf{A} - \mathbf{A}$ is the graph Laplacian, and $\mathbf{D}_\mathbf{A}$ is a diagonal matrix with $\mathbf{D}_{\mathbf{A}ii} = \sum_j \mathbf{A}_{ij}$. The new optimization problem is still convex with the closed form solution [21]:

$$\hat{\mathbf{w}} = (\sum_{i=1}^{n_\ell} \mathbf{r}_i \mathbf{r}_i^T + \alpha n_\ell \mathcal{I} + \beta n_\ell \mathbf{R} \Delta_\mathbf{A} \mathbf{R}^T)^{-1} \sum_{i=1}^{n_\ell} q_i \mathbf{r}_i$$

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

$$\bar{Q}(u) = \frac{\sum_{r \in R_u} Q(\mathbf{r})}{|R_u|} = \frac{\sum_{r \in R_u} \mathbf{w}^T \mathbf{r}_i}{|R_u|} \qquad (5)$$

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

$$\Omega_3(Q) = \Omega_1(Q) + \beta \sum_{u_1} \sum_{u_2 \in N_{u_1}} \left(\max \left\{ 0, \bar{Q}(u_1) - \bar{Q}(u_2) \right\} \right)^2$$

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 Ω_3 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 $\mathbf{h}_u$ be the n-dimensional normalized indicator vector where $\mathbf{h}_u(i) = 1/|R_u|$ if user u has written review r_i, and zero otherwise. Then we have that $\bar{Q}(u) = \mathbf{w}^T \mathbf{R} \mathbf{h}_u$. We can thus write the objective function as

$$\begin{aligned} \Omega_3(\mathbf{w}) = \quad & \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} (\mathbf{w}^T \mathbf{r}_i - q_i)^2 + \alpha \mathbf{w}^T \mathbf{w} \qquad (6) \\ + \quad & \beta \sum_{u,v \in U} \mathbf{S}_{uv} \left(\max \left\{ 0, \mathbf{w}^T \mathbf{R} \mathbf{h}_u - \mathbf{w}^T \mathbf{R} \mathbf{h}_v \right\} \right)^2 \end{aligned}$$

where $\mathbf{S}$ 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

$$\begin{aligned} \frac{\partial \Omega_3(\mathbf{w})}{2 \partial \mathbf{w}} = \quad & \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} \mathbf{r}_i \mathbf{r}_i^T \mathbf{w} - \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} \mathbf{r}_i q_i + \alpha \mathbf{w} \\ + \quad & \beta \sum_{\substack{u,v, \\ \mathbf{w}^T \mathbf{R}(\mathbf{h}_u - \mathbf{h}_v) > 0}} \mathbf{S}_{uv} \mathbf{R} (\mathbf{h}_u - \mathbf{h}_v)(\mathbf{h}_u - \mathbf{h}_v)^T \mathbf{R}^T \mathbf{w} \end{aligned}$$

Let $\mathbf{H} = [\mathbf{h}_1, ..., \mathbf{h}_m]$ be an $n \times m$ matrix defined over all reviewers and $\mathbf{Z}$ be a new matrix such that

$$\mathbf{Z}_{uv} = \begin{cases} S_{uv} & \text{if } \left[\operatorname{diag}(\mathbf{w}^T \mathbf{R} \mathbf{H}) \mathbf{S} - \mathbf{S} \operatorname{diag}(\mathbf{w}^T \mathbf{R} \mathbf{H}) \right]_{uv} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Now we can rewrite the gradient as

$$\begin{aligned} \frac{\partial \Omega_3(\mathbf{w})}{2 \partial \mathbf{w}} = \quad & \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} \mathbf{r}_i \mathbf{r}_i^T \mathbf{w} - \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} \mathbf{r}_i q_i + \alpha \mathbf{w} \\ + \quad & \beta \mathbf{R} \mathbf{H} \Delta_\mathbf{Z} \mathbf{H}^T \mathbf{R}^T \mathbf{w} \end{aligned}$$

where $\Delta_\mathbf{Z} = D_\mathbf{Z} + D_{\mathbf{Z}^T} - \mathbf{Z} - \mathbf{Z}^T$ can be thought of the graph Laplacian generalized for directed graphs with $D_\mathbf{Z}$ and $D_{\mathbf{Z}^T}$ the diagonal matrices of the row, and column sums of $\mathbf{Z}$ 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:

$$\Omega_4(Q) = \Omega_1(Q) + \beta \sum_{u \in U} \sum_{x,y \in N_u} \left(\bar{Q}(x) - \bar{Q}(y) \right)^2$$

Minimizing function Ω_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 regularizaton. Let $\mathbf{C}$ be the co-citation graph adjacency matrix, where $\mathbf{C}_{ij} = 1$ if two reviewers u_i and u_j are both trusted by at least one other reviewer u. Using the same definition of matrix $\mathbf{R}$ and vector $\mathbf{h}_u$ as for trust consistency, the objective function now becomes

$$\begin{aligned} \Omega_4(\mathbf{w}) = \quad & \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} \left(\mathbf{w}^T \mathbf{r}_i - q_i \right)^2 + \alpha \mathbf{w}^T \mathbf{w} \\ + \quad & \beta \sum_{i<j} \mathbf{C}_{ij} \left(\mathbf{w}^T \mathbf{R} \mathbf{h}_i - \mathbf{w}^T \mathbf{R} \mathbf{h}_j \right)^2 \qquad (7) \end{aligned}$$

Let $\Delta_\mathbf{C}$ be the Laplacian of graph $\mathbf{C}$. The closed form solution is

$$\hat{\mathbf{w}} = \left(\sum_{i=1}^{n_\ell} \mathbf{r}_i \mathbf{r}_i^T + \alpha n_\ell \mathcal{I} + \beta n_\ell \mathbf{R} \mathbf{H} \Delta_\mathbf{C} \mathbf{H}^T \mathbf{R}^T \right)^{-1} \sum_{i=1}^{n_\ell} \mathbf{r}_i q_i$$

Link Consistency: The regularization for this hypothesis is very similar to the one for the co-citation consistency. We treat the trust network as an undirected graph. Let $\mathbf{B}$ be the corresponding matrix, where $\mathbf{B}_{ij} = 1$ if $\mathbf{S}_{ij} = 1$ or $\mathbf{S}_{ji} = 1$. Our objective function now becomes

$$\begin{aligned}\Omega_5(\mathbf{w}) = \quad & \frac{1}{n_\ell}\sum_{i=1}^{n_\ell}\left(\mathbf{w}^T\mathbf{r}_i - q_i\right)^2 + \alpha\mathbf{w}^T\mathbf{w} \\ + \quad & \beta\sum_{i<j}\mathbf{B}_{ij}\left(\mathbf{w}^T\mathbf{R}\mathbf{h}_i - \mathbf{w}^T\mathbf{R}\mathbf{h}_j\right)^2 \end{aligned} \tag{8}$$

with a similar closed form solution

$$\hat{\mathbf{w}} = \left(\sum_{i=1}^{n_\ell}\mathbf{r}_i\mathbf{r}_i^T + \alpha n_\ell\mathcal{I} + \beta n_\ell\mathbf{R}\mathbf{H}\Delta_{\mathbf{B}}\mathbf{H}^T\mathbf{R}^T\right)^{-1}\sum_{i=1}^{n_\ell}\mathbf{r}_i q_i$$

In all these cases, β is a weight on the added regularization term which defines a trade-off between the mean squared error loss and the regularization constraint in the final objective function.

Adding the regularization makes our problem a *semi-supervised* learning problem. That is, our algorithms operate on both the labeled and the unlabeled data. Although, only the labels of the labeled data are known to the algorithm, the unlabeled data are also used for optimizing the regularized regression functions. This gives considerable more flexibility to the algorithm, since it is able to operate even with little labeled data by making use of the unlabeled data and the constraints defined by the social context. Furthermore, through regularization the signal from the social context is incorporated into the textual features. The resulting predictor function operates only on textual features, so it can be applied even in the case where there is no social context.

5. EXPERIMENTS

In this section, we present the experimental evaluation of our techniques. For our experiments we use product reviews obtained from a real online commerce portal. We begin by describing the characteristics and preprocessing of our data sets. Then, we test the hypotheses we proposed in Section 4.2.2 on these real-world datasets. Finally, we evaluate the prediction performance of different methods and conduct some analysis.

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 trust social network (with at least one link in the social network), in order to test our hypotheses and methods based on social networks. Finally, we require for each product to have some representation in the dataset, that is, a sufficiently large number of reviews. The pruning thresholds are selected per category, so as to obtain sufficient volume of data. For the Digital Cameras category, this results in a minimum amount of pruning. Although DC reviews do not contain much social context information, we still include them here for comparison and generality purposes.

[1] http://www.ciao.co.uk/

	Cellphone	Beauty	Digital Camera
PRUNING SETTINGS			
min # of ratings/ review	5	5	5
min # of reviews/reviewer	2	2	1
min # of trust links/reviewer	1	1	0
min # of reviews/ product	5	10	5
STATISTICS			
# of reviews	1943	4849	3697
# of reviewers	881	1709	3465
# of products	158	308	380
# of links in Trust	2905	20374	3894
# of links in Link	4644	32104	6022
# of links in Cocitation	13678	188610	22136
Trust graph density	0.0075	0.0140	0.0006
Link graph density	0.0120	0.0220	0.0010
Cociation graph density	0.0353	0.1292	0.0037
Avg # of reviews/reviewer	2.2054	2.8373	1.0670
Ratio of Reciprocal links	0.4014	0.4243	0.4535
Clustering coefficient	0.2286	0.3072	0.2523
CHARACTERISTICS			
Social Context	rich	rich	sparse
Quality Distribution	balanced	skewed	balanced

Table 2: Data Pruning Settings, Statistics, and Characteristics

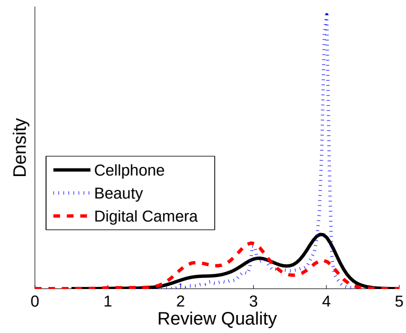


Figure 1: Density Estimate of Gold Standard Review Quality.

From the statistics in Table 2, we can see that Cellphone and Beauty reviews contain more rich social context information than DC reviews in the sense that the average number of reviews per reviewer is more than twice that for Digital Cameras, and the link density (defined as $D = \frac{2|E|}{|V|(|V|-1)}$ for a graph with vertices V and edges E) is more than 10 times that of Digital Cameras. We also plot the Kernel-smoothing density estimate (pdf) of the samples q_i (the gold standard review quality) in Figure 1. The distributions of q_i for the three categories are quite different. Beauty reviews are highly concentrated at rating 4, while Cellphone and DC reviews have a more balanced distribution of quality. We summarize the characteristics of the three data sets in the bottom of Table 2.

STD	Cellphone	Beauty	Digital Camera
Rel:DifferentReviewer	0.9187	0.7017	0.9571
Rel:SameReviewer	0.5937	0.4518	0.6176
p-value	1.37E-48*	1.57E-287*	3.12E-11*

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

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 $dq_{ji} = -dq_{ij}$ the mean value of dq_{ij} for both groups is zero. We are interested in the standard deviation, $\mathrm{std}(dq_{ij})$, 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 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 Rel:DifferentReviewer for all three categories. Moreover, two-sample Kolmogorov-Smirnov (KS) test of the samples in the two groups indicates that the difference of the two groups is statistically significant. The p-values are shown in the last row of Table 3. The star next to the p-value means there is strong evidence ($p < 0.01$) that the two samples come from different distributions.

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 $d\bar{Q}^*_{ij} = \bar{Q}^*(u_i) - \bar{Q}^*(u_j)$, i.e., the difference in quality of two reviewers, where, similar to Equation 5

$$\bar{Q}^*(u) = \frac{\sum_{r_i \in R_u} q_i}{|R_u|} \tag{9}$$

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., $\mathbf{B}_{ij} = 0$.

Rel:Trust: User u_i trusts user u_j, i.e., $\mathbf{S}_{ij} = 1$.

Rel:Cocitation: Users u_i and u_j are trusted by at least one other reviewer u_3, i.e., $\mathbf{C}_{ij} = 1$.

Rel:Link: User u_i trusts user u_j, or u_j trusts u_i, i.e., $\mathbf{B}_{ij} = 1$.

In Figure 3, we plot the Kernel-smoothing density estimate of the $d\bar{Q}^*_{ij}$ 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.

Cellphone				
p-value	Rel:None	Rel:Trust	Rel:Link	Rel:Cocitation
Rel:None	-	3.20E-82*	4.53E-44*	6.12E-177*
Rel:Trust	-	-	3.44E-16*	6.89E-22*
Rel:Link	-	-	-	0.0657
Moments	Rel:None	Rel:Trust	Rel:Link	Rel:Cocitation
Mean	0.0000	-0.1376	0.0000	0.0000
Variance	0.6727	0.3255	0.3485	0.2914
Beauty				
p-value	Rel:None	Rel:Trust	Rel:Link	Rel:Cocitation
Rel:None	-	0.00E+00*	0.00E+00*	0.00E+00*
Rel:Trust	-	-	3.83E-59*	3.75E-101*
Rel:Link	-	-	-	0.3003
Moments	Rel:None	Rel:Trust	Rel:Link	Rel:Cocitation
Mean	0.0000	-0.0824	0.0000	0.0000
Variance	0.4331	0.1806	0.1907	0.1903
Digital Camera				
p-value	Rel:None	Rel:Trust	Rel:Link	Rel:Cocitation
Rel:None	-	1.76E-135*	2.14E-87*	0.00E+00*
Rel:Trust	-	-	1.46E-21*	2.10E-34*
Rel:Link	-	-	-	0.3052
Moments	Rel:None	Rel:Trust	Rel:Link	Rel:Cocitation
Mean	0.0000	-0.1481	0.0000	0.0000
Variance	0.8763	0.4068	0.4471	0.4059

Table 4: Statistics of Reviewer Quality Difference to Support Social Network Consistency Hypotheses.

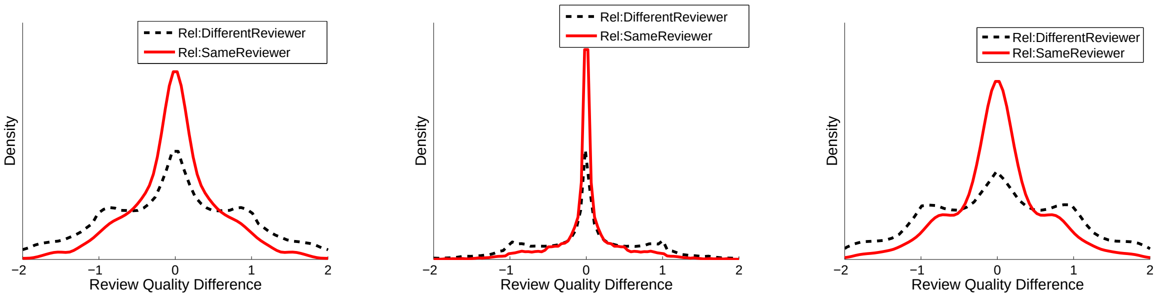
The first observation is that the distribution of Rel:Trust is skewed towards the negative with a negative mean. This supports the Trust Consistency Hypothesis that when u_i trusts u_j, the quality of u_i is usually lower than that of u_j, i.e., $\bar{Q}^*(u_i) - \bar{Q}^*(u_j) < 0$. The remaining three 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.

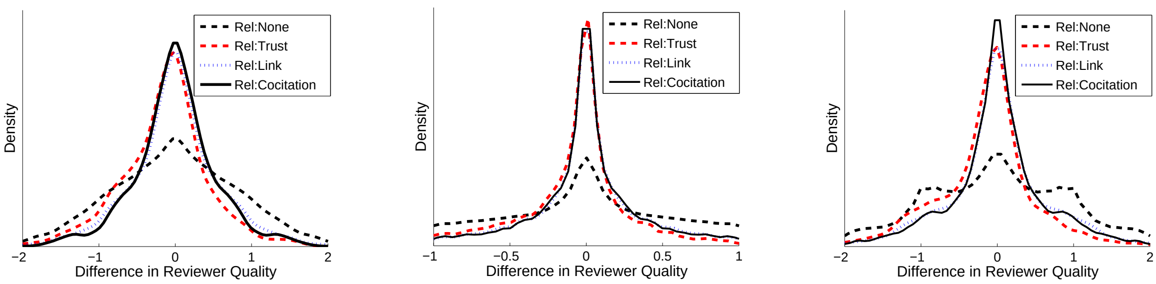
5.3 Prediction Performance

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_{train}), and 50% for testing (R_{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



(a) Cellphone (b) Beauty (c) Digital Cameras

Figure 2: Density Estimates of Review Quality Difference.



(a) Cellphone (b) Beauty (c) Digital Cameras

Figure 3: Density Estimates of Reviewer Quality Difference.

used to enrich the feature representation for the linear model. We fix the parameter α 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 β. 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,

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

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 $\bar{Q}^*(u)$ 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, $\bar{Q}^*(u)$ 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 $R_{\mathbf{train}}$, i.e., $Q(r) = \frac{1}{n_t}\sum_{i=1}^{n_t} q_i$.

TBL:Reviewer: Predict as the quality $\bar{Q}^*(u_r)$ 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, or 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.

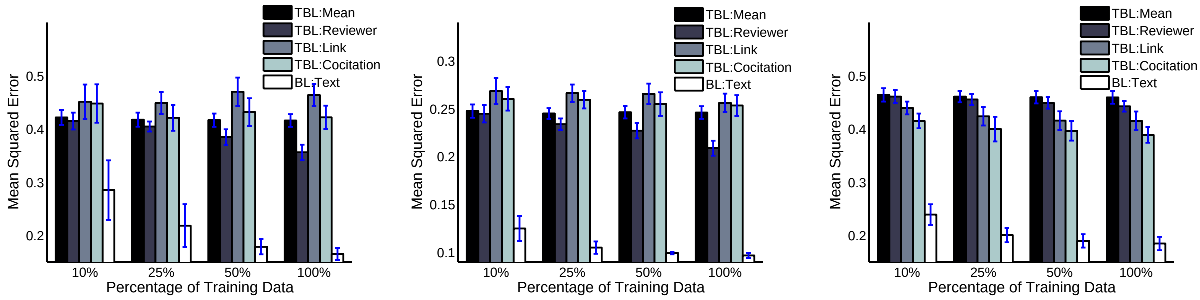
5.3.2 Incorporating Social Context

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).



(a) Cellphone (b) Beauty (c) Digital Cameras

Figure 4: MSE of Simple Text-free Baselines V.S. Text-only Baseline.

TRAINING SUBSET	10%	25%	50%	100%
Cellphone				
BL:Text	0.2852±0.0558	0.2183±0.0402	0.1787±0.0143	0.1654±0.0112
BL:Text+Rvr	0.3137±0.1079(9.99%)	0.2249±0.0518(3.02%)	0.1728±0.0116(-3.30%)	0.1552±0.0095(**-6.17%**)
REG:Link	0.2642±0.0292(-7.36%)	0.2113±0.0294(-3.21%)	0.1781±0.014(-0.34%)	0.1652±0.0111(-0.12%)
REG:CoCitation	0.2635±0.0359(-7.61%)	0.2064±0.0226(-5.45%)	0.1771±0.0133(-0.90%)	0.1647±0.0107(-0.42%)
REG:Trust	0.2563±0.0317(-10.13%)	0.2035±0.0205(-6.78%)	0.1768±0.0134(-1.06%)	0.1647±0.0108(-0.42%)
REG:Reviewer	0.2468±0.0223(**-13.46%**)	0.1958±0.0116(**-10.31%**)	0.1728±0.01(**-3.30%**)	0.1635±0.0089(-1.15%)
Beauty				
BL:Text	0.125±0.0132	0.1051±0.0064	0.0994±0.0014	0.0969±0.0028
BL:Text+Rvr	0.122±0.0123(-2.40%)	0.0973±0.0062(**-7.42%**)	0.089±0.002(**-10.46%**)	0.0857±0.0027(**-11.56%**)
REG:Link	0.1174±0.0073(-6.08%)	0.1036±0.0054(-1.43%)	0.0991±0.0016(-0.30%)	0.0968±0.0028(-0.10%)
REG:CoCitation	0.1166±0.007(-6.72%)	0.1036±0.0054(-1.43%)	0.099±0.0016(-0.40%)	0.0968±0.003(-0.10%)
REG:Trust	0.1157±0.0058(-7.44%)	0.1022±0.0044(-2.76%)	0.0986±0.0021(-0.80%)	0.0966±0.0029(-0.31%)
REG:Reviewer	0.112±0.0063(**-10.40%**)	0.1021±0.0049(-2.85%)	0.0984±0.0018(-1.01%)	0.0964±0.0028(-0.52%)
Digital Camera				
BL:Text	0.2392±0.0192	0.2007±0.0136	0.1897±0.0125	0.1848±0.0127
BL:Text+Rvr	0.2541±0.0239(6.23%)	0.2011±0.0106(0.20%)	0.1869±0.0096(**-1.48%**)	0.1801±0.0115(**-2.54%**)
REG:Link	0.2355±0.0211(-1.55%)	0.2002±0.0125(-0.25%)	0.1894±0.0124(-0.16%)	0.1848±0.0127(0.00%)
REG:CoCitation	0.2346±0.0204(-1.92%)	0.1994±0.0132(-0.65%)	0.1893±0.0126(-0.21%)	0.1848±0.0126(0.00%)
REG:Trust	0.2302±0.0183(**-3.76%**)	0.1984±0.0127(**-1.15%**)	0.189±0.0124(-0.37%)	0.1846±0.0127(-0.11%)
REG:Reviewer	0.2373±0.0189(-0.79%)	0.2005±0.0135(-0.10%)	0.1896±0.0124(-0.05%)	0.1848±0.0127(0.00%)

Table 5: MSE of Using Social Context as Features and as Regularization vs. Text-based Baseline

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

REG:Cocitation: Linear Regression with a regularizer under Cociation 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 β 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.

The results of our experiments are summarized in Table 5 where we show the mean MSE and the standard deviation for all techniques, over all categories, for different training data sizes. In the parentheses we have the percentage of reduction over MSE of the text-based baseline BL:Text. The best result (largest decrease of MSE) for each data set and each training size is emphasized in bold.

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% in all three data sets. This is often the case in practice, where we have limited resources for obtaining labeled training data, while there are large amounts of unlabeled data available.

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.[2] 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 the generalizablity of the proposed regularization methods.

To further support the generalizablity 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 optmization process and for which we use only the textual features and hide their social context. More specifically, after learning a quality prediction function Q using 10% of the training data, we apply it to the remaining 90% of the training data, by multiplying the learned weight vector $\mathbf{w}$ 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 $\mathbf{w}$ that captures more accurately the importance of textual features for predicting the true quality on the held-out set.

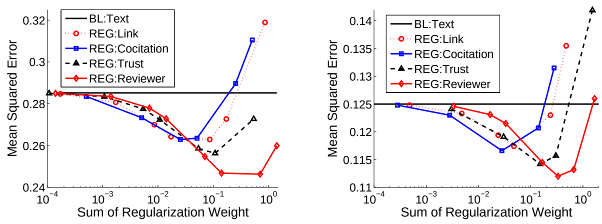
In summary, we make the following observations.

- 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 β to set: the tradeoff 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 β 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.



(a) Cellphone (b) Beauty

Figure 5: Parameter Sensitivity.

We define the sum of regularization weight as $\sigma = \beta \sum_{ij} \mathbf{M}_{ij}$, where $\mathbf{M}$ can be the co-author matrix $\mathbf{A}$, the directed trust matrix $\mathbf{S}$, the co-citation matrix $\mathbf{C}$, or the undirected link matrix $\mathbf{B}$.

Figure 5 shows how the prediction performance of regularization methods varies as we use different values of σ. We only show the parameter sensitivity for Cellphone and Beauty reviews where the social context is relatively rich. The training data size is fixed to be 10%. As we can see, even though Cellphone and Beauty reviews carry different characteristics, the curves follow a very similar trend: as long as we set $\sigma \leq 0.1$, all regularization methods achieve consistently better performance than the baseline. As σ goes to zero, the performance converges to the text-based baseline. In addition, the shape of the performance curve depends on the corresponding hypothesis. For example, the optimum σ for REG:Trust is larger than that of REG:Link and REG:Cociation. Also, even with a value of σ higher than the optimum, the error of the REG:Reviewer does not increase as quickly as for the other methods. These observations are in line with the previous observations that the history of the reviewer (REG:Reviewer) and the Trust graph (REG:Trust) provide a better signal than the Co-Citation graph, or the Link graph.

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. Ghose and Ipeirotis [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

Test on	# of Reviews	REG:Link	REG:CoCitation	REG:Trust	REG:Reviewer
All	1066	7.36%	7.61%	10.13%	13.46%
Reviews with no social context	144	3.33%	1.08%	3.15%	6.63%
Reviews with social context	922	8.11%	8.84%	11.47%	14.75%
Held-out reviews with hidden social context	893	10.38%	9.64%	11.73%	11.34%

Table 6: Improvement of Regularization Methods over BL:Text (Cellphone)

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, bur 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

[1] J. Abernethy, O. Chapelle, and C. Castillo. Web spam identification through content and hyperlinks. In *AIRWeb '08*, pages 41–44, 2008.

[2] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *WSDM*, pages 183–194. ACM, 2008.

[3] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

[4] J. Bian, Y. Liu, D. Zhou, E. Agichtein, and H. Zha. Learning to recognize reliable users and content in social media with coupled mutual reinforcement. In *WWW*, pages 51–60. ACM, 2009.

[5] C. Danescu-Niculescu-Mizil, G. Kossinets, J. Kleinberg, and L. Lee. How opinions are received by online communities: a case study on amazon.com helpfulness votes. In *WWW '09*, pages 141–150, 2009.

[6] A. Ghose and P. G. Ipeirotis. Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. (January 24, 2010), Available at SSRN: http://ssrn.com/abstract=1261751.

[7] M. Hu and B. Liu. Mining and summarizing customer reviews. In *KDD*, pages 168–177, 2004.

[8] M. Hu and B. Liu. Mining opinion features in customer reviews. In *AAAI*, pages 755–760, 2004.

[9] C. Johnson. Us ecommerce forcast: 2008 to 2012. `http://www.forrester.com/Research/Document/Excerpt/0,7211,41592,00.html`.

[10] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti. Automatically assessing review helpfulness. In *EMNLP*, pages 423–430, Sydney, Australia, July 2006.

[11] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou. Low-quality product review detection in opinion summarization. In *EMNLP-CoNLL*, pages 334–342, 2007. Poster paper.

[12] Y. Liu, X. Huang, A. An, and X. Yu. Modeling and predicting the helpfulness of online reviews. In *ICDM*, pages 443–452, 2008.

[13] Q. Mei, D. Cai, D. Zhang, and C. Zhai. Topic modeling with network regularization. In *WWW '08*, pages 101–110, 2008.

[14] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[15] O. Tsur and A. Rappoport. Revrank: a fully unsupervised algorithm for selecting the most helpful book reviews. In *ICWSM*, 2009.

[16] T. Zhang, A. Popescul, and B. Dom. Linear prediction models with graph regularization for web-page categorization. In *KDD*, pages 821–826. ACM, 2006.

[17] Z. Zhang and B. Varadarajan. Utility scoring of product reviews. In *CIKM '06*, pages 51–57, New York, NY, USA, 2006. ACM.

[18] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *NIPS*, 2003.

[19] X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

[20] X. Zhu, Z. Ghahramani, and J. D. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, pages 912–919. AAAI Press, 2003.

[21] X. Zhu and A. B. Goldberg. *Introduction to Semi-Supervised Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2009.