Predicting Useful Restaurant Reviews by Subtopics Using Yelp Data

Yu-Hua Cheng
Statistics Department
Columbia University
yc2911@columbia.edu

Jingchi Wang
Quantitative Methods in the Social Sciences
Columbia University
jw3153@columbia.edu

Abstract—**The objective of this project is to first categorize the Yelp restaurant reviews into different subtopics, and then predict the usefulness of reviews in certain subtopics. Methodologies include Latent Dirichlet allocation, Generalized Linear Regression, Support Vector Machine, Maximum Entropy, Random Forests, and Boosting. We expected to provide Yelp users with targeted reviews of interest as well as a high quality of them.**

Keywords-restaurant review; online community; usefulness prediction; topic modeling

1. INTRODUCTION

Yelp reviews play an important role for users' evaluation towards one business. When one tends to visit a new place, reviews can be served as one way of recommendation that carries comprehensive information of different aspects. But the aspects of interest may differ among users. Take restaurant business for example, while one may like to understand the food quality, the other may be interested in the staff service. Therefore, if we could categorize the reviews into subtopics of different aspects, we would help users target their desired reviews in a certain aspect easier. We focused on reviews in restaurant business only along the project, and hence our expected subtopics would separately represent the aspects such as food, staff service, waiting time, price, or place.

After categorizing the restaurant reviews into subtopics, we then would predict the usefulness of reviews, using the "useful votes" as the training labels. But what is the necessity of the usefulness prediction, and why not just use the true labels of "useful votes" as our evaluation of reviews quality? The reasons are that first, for some less popular restaurants, all the reviews may have zero useful vote because they are barely viewed. Second, a newly posted review is more likely to have fewer votes, but it does not necessarily mean it has lower quality of usefulness. Ideally, our prediction model could assure to provide users with high quality of reviews.

In summary, we expected our project could contribute a more functional review service to Yelp users, by means of targeted subtopics of users' interests and useful reviews.

2. RELATED WORKS

As online communities like Yelp becoming increasingly helpful and popular in everyday life, many researchers are interested in predicting the quality, or usefulness, of the content posted by users. In general, online communities fall into two types: product/business review sites, and question answering sites. Although the two types serve different purpose, they share many common features, and usually have similar analysis results. In terms of research methods, some make use of classic regression, visualizations, or time series analysis to discover the correlations between various numerical or categorical features from online community. Some attempt to analyze users' behavior and context in social network perspectives. Others use more advanced machine learning techniques to automate the prediction process.

In a case study of Stack Overflow (a Q&A site for programmer), Ashton Anderson et al. (2008) emphasize on the value of community activity, and use question and all corresponding answers as fundamental unit of analysis. In order to understand the dynamic influence of community activity, they use time-series models to analyze the temporal distribution of answers and votes. Three most important findings are high assortativity in the reputations of co-authors, strong correlation between reputation and answer speed, and the importance of answering time to be chosen. These findings may prove intuitive in predicting the long-term value of community activity.

Jiwoon Jeon et al. (2006) develop a framework to predict content quality by only using a variety of non-text features, such as click/copy count and user activity. They proposed the use of kernel density estimation (KDE) for feature conversion, and maximum entropy for answer quality estimation. In a retrieval experiment, the results show that when their prediction score is incorporated into natural language retrieval model, the query performance can be significantly improved. Thus, they conclude that the non-text features are very important in predicting usefulness.

Vicenç Gómez et al. (2008) analyzed the network structure of a tech-news website called Slashdot. They demonstrate that the network in this website shares many common features with traditional social networks, such as large component, high clustering and short average path length, though the reciprocity and assortativity are

relatively smaller. They used log-normal distribution to explain the degree distributions, and also prove that this method is better. Apart from it, they also find the discussion threads to be highly heterogeneous and self-similar.

In order to make use of both textual content and social context in predicting post quality, Yue Lu et al. (2010) combine the methods of text mining and social network analysis. Their textual analysis includes features like text-statistics, syntactic, conformity and sentiment. In incorporating social context, they use it either as features or as regularization constraints, based on a set of hypotheses (author/trust/co-citation consistency). They argue that the regularization techniques is superior because of its generalizability even in the absence of social context data. The results form experiment show that prediction accuracy can be greatly improved once social context regularization is in use.

There are also many studies on product review sites. Soo-Min Kim et al. (2006) select multiple features of product reviews from Amazon.com, in an attempt to develop a system for automatically predicting helpfulness of reviews. They use support vector machine (SVM) as the only statistical learning algorithm for modeling. The results with a 0.66 rank correlation, though not very high, show good prediction performance. They also find some features to be very important in usefulness, such as review length, unigrams, and rating score.

Using Yahoo Answer dataset, Yandong Liu et al. (2008) develop models to predict the asker's satisfaction (usefulness of answers) in Q&A sites. They define the satisfaction of asker as closing the question, selecting best answer and giving high rating. Apart from questions and answers, other relevant features are also taken into account, such as asker history, answerer history and question category. In terms of prediction models, various statistical learning algorithms are used, including decision trees, support vector machines, boosting and naïve bayes. To estimate the accuracy of their models, human judgment (from Amazon's Mechanical Turk) is used for comparison. As a conclusion, they show that their models, by making use of contextual information such as asker's history, outperform human raters in predicting the usefulness of answers. Thus, they conclude that background information such as user's history play an important role in predicting usefulness of his/her post.

3. SYSTEM OVERVIEW

Our dataset came from the Yelp Dataset Challenge, which includes data from Phoenix, Las Vegas, Madison, Waterloo and Edinburgh. Since we only focused on the reviews in restaurant business, the first task was to filter out the restaurant reviews. We then set a threshold that only considered the restaurants with more than 10 reviews, to reduce the bias of useful votes labels, because if restaurants have few reviews, it could indicate the reviews were posted lately, and consequently zero useful vote may be the reason that it was barely viewed, yet not necessary the cause of its quality.

The final data contained 706,646 reviews, from 9,569 restaurants and 185,469 users, with 20 attributes reflecting different aspect about reviews, users, and restaurants.

The workflow of our project is described below:

1. Identifying subtopics by the methodology of Latent Dirichlet allocation
2. Predicting review usefulness using two different approach
   I. Non-text prediction: Regression using numerical features
   II. Textual prediction: Classification using content of reviews
3. Evaluating model performance by different metrics such as precision, recall, or F score
4. Conclusion

4. ALGORITHM

4.1 Topic Modeling - Latent Dirichlet allocation

Latent Dirichlet allocation (LDA) is a widely used topic model for generating topics from documents. It assumes each document can be explained as a mixture of topics, with mixture weights c1 to ck, where k is the total number of topics. And each document is generated in the following process:

Sample topic proportions c1:k ~ Dirichlet (ϕ)
For i = 1, ..., M:
Sample topic for word i as ki ~ Multinomial (c1:k)
Sample word i ~ Multinomial (Θk)

We expected LDA could generate the topics of food quality, staff service, waiting time, and environments. Therefore, in order to achieve the desired results, we first conducted the following data preprocessing.

First, if we implemented LDA on the entire restaurant reviews, the output will consist of subtopics from restaurant categories. For instance, one subtopic may contain terms like fried rice, dumplings, or chow mien, which indicates a Chinese category. While another subtopic is likely to contain terms like burgers, fries, or steaks, which indicates an American category. Obviously, this was not the expected output. Therefore, we have decided to implement LDA for reviews in different restaurant categories, which includes American, Nightlife, Mexican, Italian, Bars, Breakfast & Brunch, Chinese, Japanese, etc.

Second, we only extracted the nouns from each review and treated it as the input of documents in LDA, since nouns are most related to the terms of expected subtopics.

4.2 Prediction (Classification) of Useful Reviews

In general, the purpose of prediction is to categorize restaurants reviews into useful and not useful. Since we want to fully make use of all relevant information, two separate groups of classification models will be used: one based on non-text (numerical) features, and the other based on textual content.

4.2.1 Non-text Prediction Models

4.2.1.1 Features

In the non-text prediction, ten features are selected, as shown in Table 1. They can be categorized into three groups: Review, User and Restaurants. The group names reflect the different aspects of features.

Since we are predicting useful reviews, the dependent variable is useful vote number received by a review. It is a direct measure of usefulness in the community's (users') opinion, which indicate how many users consider a review as useful. This feature is used as a numerical variable in regression modeling. In classification models, as well as textual prediction, it will be transformed into a categorical (or dummy) variable. For the purpose of categorization, the reviews with vote number of one or higher will be treated as useful, whereas lower than one vote will be treated as not useful. In other words, we consider a review to be useful, if at least one user thinks it useful. This is a way to control for popularity of a restaurant. The reason why a review received fewer votes may not be that it is not as useful. Instead, it could actually because fewer people visit the restaurant!

The explanatory variables, as mentioned, are grouped into three classes, representing different aspects. The Review group has two variables that are directly from review: length and stars. In assumption, length has a positive effect on helpfulness, because longer content usually means more details. Review stars should have a negative one, because we are often more careful about negative facts that may ruin our experience. The User group has more features, which reveal the background information about the author of a review. They are seniority, past review count, useful votes, fans and average stars. It is assumed that more experienced and popular users usually post more useful reviews. Therefore, except for average stars, all other features in this group should be positively correlated with usefulness. The Restaurant group has two features, which reflects the popularity of a restaurants. It is assumed that both review count and average review star are negatively correlated with helpfulness, because higher popularity would generally make a review harder to be considered as useful.

Table 1: Numerical Features

	Feature Group	Feature Name	Description	Expected Direction
Dependent Variable	Review	review_usefulvotes	Number of useful votes received by a review. Selected as a direct measure of usefulness.	N/A
Explanatory Variables	Review	review_length	Word count of textual content in a review.	Positive
		review_stars	Number of stars given by a review.	Negative
	User	user_seni	Seniority of user. Length of user's membership by month.	Positive
		user_usefulvotes	Number of useful votes voted by a user.	Positive
		user_reviewcount	Number of reviews posted by a user.	Positive
		user_fans	Number of fans a user has. Measures user's popularity in the community	Positive
		user_avgstars	Average review stars from user's posted reviews.	Negative
	Restaurant	rest_reviewcount	Number of reviews received by a restaurant.	Negative
		rest_stars	Average stars of reviews received by a restaurant.	Negative

4.2.1.2 Non-text Algorithms

Classic Regression (OLS)

A classic linear regression model using ordinary least square will be used. Although this model is the simplest one, it may turn out to be very helpful, because it fully takes into account the numerical information. Different from all other classification model, the dependent variable here is numerical, and can be transformed into categorical (dummy) for classification purpose.

Logistics Regression

The logistics classifier is a popular probabilistic model used for predicting categorical dependent variable. It

models the probability that the response belongs to a particular category, and it has an advantage that the predicted probability always falls between 0 and 1. Different from OLS, the coefficients here are estimated using maximum likelihood method. Also, the dependent variable here is binomial (dummy) usefulness.

4.2.2 Textual Prediction

In the textual models, the dependent variable is still binomial usefulness defined previously. The independent variables now are the document term matrix of the textual content of reviews.

Because of complexity and uncertainty of text classification, more variety of algorithms are used here. There are seven algorithms: Support Vector Machines, MaxEnt, GLMNET, Supervised LDA, Boosting, Bagging, and Random Forests.

SVM (Support Vector Machines). Almost the most widely used supervised learning algorithm for classification. It is very different from logistic model in that non-probabilistic binary linear classifier is used. Moreover, non-linear classification can also be performed with SVM using a specific kernel.

MaxEnt (Maximum Entropy). Also named as multinomial logistic regression. It generalize logistic regression to multi-class problem.

GLMNET (Lasso and elastic-net regularized generalized linear models). Lasso and ridge are regularization methods that prevent overfitting of models by using penalties on extreme values. Elastic-net is a regularized method that linearly combine the penalties of lasso and ridge.

SLDA (Supervised latent Dirichlet allocation). A supervised topic model for labeled documents. Uses maximum-likelihood method for estimation.

Boosting. An algorithm for reducing bias in supervised learning. It seeks to create strong learner using a set of weak learner.

Bagging. Full name is bootstrap aggregating. An algorithm for improving stability and accuracy, as well as reducing overfitting and variance of machine learning algorithms. Often used on decision trees.

Random Forests. An ensemble classification method that create multiple decision trees in training. It outputs the most frequent categories that given by all individual trees.

5. EXPERIMENT RESULTS

5.1 Topic Modeling (LDA) Results

Once the subtopics were generated, we manually tagged the names of them, based on their frequent terms. There are overlapping names for different subtopics, but this case was allowed because if we constrained the total number of subtopics k to a small value such as 3 or 4, we were not able to explore the enough aspects in subtopics. Instead, setting k to an appropriately large value would allow spreading out the aspects in subtopics.

We have set k, the total number of subtopics, equal to 8, and below are the LDA results in different restaurant categories.

Table 2: LDA Result 1

Thai Food		Food & Service & Time		Place & Food & Service		Service & Time & Food	
pork	0.034	food	0.104	place	0.058	order	0.060
thai	0.032	restaurant	0.022	food	0.045	time	0.024
tea	0.022	service	0.022	sum	0.033	food	0.022
pho	0.022	time	0.020	service	0.033	restaurant	0.022
time	0.020	panda	0.014	vegas	0.023	place	0.019
duck	0.018	love	0.012	restaurant	0.016	times	0.016
place	0.016	rice	0.011	table	0.014	menu	0.016
pad	0.010	beans	0.009	night	0.014	dishes	0.016
food	0.009	express	0.009	people	0.012	dish	0.015
town	0.009	stars	0.008	time	0.01	service	0.015

First 4 subtopics with top 10 most frequent terms, in Chinese category

Table 3: LDA Result 2

Chinese Food		Japanese Food & Price		Chinese Food		Food	
dish	0.029	food	0.064	chicken	0.102	soup	0.051
beef	0.026	place	0.051	rice	0.052	noodles	0.049
sauce	0.021	lunch	0.048	food	0.030	noodle	0.040
tofu	0.016	sushi	0.027	egg	0.029	dishes	0.028
menu	0.012	buffet	0.021	shrimp	0.027	food	0.023
dumplings	0.012	prices	0.020	sauce	0.023	place	0.021
salad	0.012	dinner	0.016	crab	0.018	beef	0.020
flavor	0.011	price	0.015	rolls	0.014	restaurant	0.018
everything	0.010	service	0.014	soup	0.013	bowl	0.014
shrimp	0.010	great	0.013	place	0.012	pot	0.014

Last 4 subtopics with top 10 most frequent terms, in Chinese category

Table 4: LDA Result 3

Time & Service		Food		Food		Food	
table	0.031	food	0.052	steak	0.022	menu	0.033
time	0.028	place	0.038	vegas	0.019	salad	0.022
order	0.020	wings	0.025	restaurant	0.014	chicken	0.019
server	0.019	restaurant	0.023	dinner	0.012	lunch	0.016
service	0.019	service	0.019	dessert	0.012	food	0.016
minutes	0.018	tacos	0.017	service	0.012	dish	0.011
waitress	0.016	great	0.014	coffee	0.009	time	0.010
night	0.013	chicken	0.012	cheese	0.009	shrimp	0.010
people	0.013	prices	0.012	meal	0.008	sauce	0.009
place	0.012	cheese	0.012	salad	0.008	love	0.009

First 4 subtopics with top 10 most frequent terms, in Nightlife category

Table 5: LDA Result 4

Bar Service		Drinks Food		Place & Service		Fast Food	
bar	0.034	wine	0.031	place	0.073	beer	0.043
hour	0.032	place	0.021	food	0.052	burger	0.038
food	0.022	food	0.021	bar	0.041	fries	0.030
fish	0.016	drinks	0.012	drinks	0.028	pizza	0.020
chips	0.014	glass	0.011	service	0.025	place	0.020
pub	0.014	menu	0.010	time	0.023	cheese	0.019
people	0.014	hour	0.010	night	0.020	food	0.017
service	0.014	meal	0.009	atmosphere	0.019	bar	0.015
drink	0.013	bread	0.009	love	0.016	burgers	0.013
bartender	0.010	brunch	0.008	music	0.016	game	0.009

Last 4 subtopics with top 10 most frequent terms, in Nightlife category

The output of LDA also included the estimated topics proportions for each document (review), and we are going to demonstrate how to use this output practically as one new Yelp review feature. Suppose we offer the four options of subtopics: 1. Food quality 2. Service 3.Waiting time 4. Place. Now one user is interested in knowing the service of one Nightlife category's restaurant. If LDA assigned the following estimated topics proportions to a given review: {Time & Service : 50%, Fast Food : 30%, Place & Service : 20%}. We will compute the total proportions related to the targeted user's interest, service, as 50% + 20% = 70%, which is called T. For every review i in this restaurant, we computed Ti, and present the reviews to the user in a descending order of their T values.

5.2 Prediction Results

Consistent with methodology, the result of the prediction experiment is also divided into two parts. Both non-text and textual experiments use the same training set and test set data. They are randomly sample from more than 80,000 review ids, with a ratio of 0.8 and 0.2.

5.2.1 Evaluation Metrics

In order to evaluate the performance of our prediction models, we use three measures that are typical for classification tasks: precision, recall, and F score. Although the three metrics are all about the accuracy of prediction, they reflect different aspects of performance.

Precision: Proportion of "useful" predictions that truly belong to "useful". It is defined as

P = True Positives/ (True Positives + False Positives)

This metric indicates the degree to which the predicted categories are correct. However, a high precision can be misleading. An extreme case of high value is when there are very few "positive" in the data and all predictions are "negative".

Recall: Proportion of "useful" reviews that are correctly identified from prediction. It is defined as

R = True Positives/ (True Positives + False Negatives)

As is named, this metric reveals the degree to which a prediction model can recall (or discover) positive categories from test data. Similarly, in some condition, recall can also be misleading. An extreme case is when there are high proportion of "positive" categories in the data and all predictions are also "positive".

F Score: A Harmonic mean of precision and recall. Can be defined as

F = 2 * P * R / (P + R)

To put it simply, this metric is a type of average of precision and recall. As mentioned above, both precision and recall have weakness and can be unreliable in some condition. Therefore, using a harmonic mean of the two, an F score can provide more stable measure of accuracy. When either precision or recall is misleadingly high, however, the F score will also be unreliable and become unreliable.

5.2.2 Non-text Prediction Results

The regression results of the non-text experiment is shown in Table 6. We can see that for in both OLS and logistic regression, all variables are statistically significant. Most coefficients have the expected direction. In Review group, the length of a review has a positive effect on usefulness, because it reflect the detail of content. Review stars is negatively correlated, suggesting that a more negative review is often more useful. In User group, except for review count and average stars, all other variables are positively correlated with usefulness. This indicates that useful reviews are mostly from those more experienced and popular users. However, it is surprising that review count has a negative coefficient, which means those who write a lot do not necessarily write useful reviews. Another expected one is the negative coefficient on the average stars of past reviews, even though the stars of review has a positive one. One reason may be that although occasional negative reviews are more useful, in general the useful one actually come from the users with proper and positive attitude. In Restaurant group, review count is negatively correlated with usefulness. As expected, higher popularity means it is more difficult to write a useful review, because there are many competitors! Lastly, restaurant stars turns out to have a positive effect. One possible reason is that customers in higher rated restaurants usually have better experience and would more probably find the reviews useful.

Table 6: Non-text Model Results

Feature Group	Feature Name	OLS Coefficient	OLS p-value	Logit Coefficient	Logit p-value
Review	review_length	4.080e-03	< 2e-16	4.320e-03	< 2e-16
	review_stars	-4.685e-02	9.18e-13	-9.179e-02	< 2e-16
User	user_seni	6.416e-03	< 2e-16	9.919e-03	< 2e-16
	user_usefulvotes	9.234e-04	< 2e-16	1.315e-03	< 2e-16
	user_reviewcount	-1.447e-03	< 2e-16	-1.667e-03	< 2e-16
	user_fans	2.158e-03	1.11e-06	1.447e-02	7.94e-13
	user_avgstars	7.958e-02	8.26e-11	7.188e-02	4.64e-06
Restaurant	rest_reviewcount	-1.310e-04	< 2e-16	-2.653e-04	< 2e-16
	rest_stars	1.807e-01	< 2e-16	2.044e-01	< 2e-16

Apart from discussion of regressions, it is equally important to evaluate the prediction results. The evaluation is shown in Table 7. We can see that both OLS and Logit model has good precision. Recall and F score are not bad, too. For both algorithms, about 70% of the predictions are correct, and more than 50% of the useful reviews are identified.

Furthermore, the real performance of prediction could be even better than the evaluation metric look to be. As defined previously, the evaluation is based on the original (test) data. However, the original data itself may not be perfectly reliable in some cases. For instance, some useful reviews may not be discovered by other users, either because they are ignored from tons of reviews, or because the restaurants are rarely visited. For another example, because of occasional unreliable votes, some reviews may be treated as "useful" even though they are not useful. In either case, the evaluation metrics would be lower than the actual performance of prediction.

Table 7: Non-text Evaluations

Evaluation Metrics		Precision	Recall	F Score
Non-text Prediction Algorithms	OLS Regression	69.37%	53.52%	60.42%
	Logit Regression	70.96%	51.16%	59.45%

5.2.3 Textual Prediction Results

Table 8 shows the evaluation results from the textual experiment. We can obviously see that the performance is lower for all algorithms in this part. Although the recall of MaxEnt algorithm is 100%, it is misleading because this algorithm actually treats most reviews as useful. In summary, it shows that textual content itself cannot provide much hint on the usefulness of the reviews. One possible reason is the complexity nature of restaurant reviews. The terms used in these reviews could be too various to categorize into useful and not useful. Therefore, we choose non-text algorithm for prediction.

Table 8: Textual Prediction Evaluation

Evaluation Metrics		Precision	Recall	F Score
Textual Prediction Algorithms	SVM	56.71%	35.02%	43.30%
	MaxEnt	47.31%	100%	64.23%
	SLDA	56.32%	30.49%	39.56%
	Bagging	57.30%	31.78%	40.88%
	Boosting	58.20%	8.26%	14.46%
	GLMNET	57.78%	26.82%	36.64%
	Random Forest	57.36%	33.67%	42.43%

6. CONCLUSION

This study shows that, using various features, it is possible to efficiently predict useful restaurant reviews. From the results of non-text models, we find a number of significant features that are important in prediction. Specifically, the detail of reviews, user's experience, popularity, and restaurant stars all have positive and substantive influence on usefulness. Some other features have negative effects. For example, popularity of restaurant may prove a challenge for users to write a truly useful review. Also, as many users are more concerned with the cons of the place they eat, it is often helpful to reveal some negative facts in a review. These findings may prove important not only for prediction of useful reviews, but also for review writing suggestions. In terms of the performance of prediction, we find the non-text algorithms to be very accurate, according to three traditional evaluation measures. Moreover, by reviewing some "false" predictions, we find out that the actual performance of our model could be even better. Despite some flaws in the original data, our model is still able to discover some hidden useful reviews, as well as distinguish not useful ones. As for textual prediction, the result is not satisfiable, even though more algorithms are used. An interesting conclusion is that although we are predicting the usefulness of the review content, the content (text) itself turn out to be not as important as those numerical features. This may be partly due to the complexity and variety of terms used review content.

Furthermore, from topic modeling, we discover some well-defined subtopics. They reveals different aspects of the reviews, such as service, food, hours, and environment. With subtopics included as a further categorization of the predicted useful reviews, our system can become even more helpful, especially for users with specific purpose.

REFERENCES

[1] Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon. com. Management Information Systems Quarterly, 34(1), 11.

[2] Luca, M. (2011). Reviews, reputation, and revenue: The case of Yelp. com (No. 12-016). Harvard Business School.

[3] Tucker, T. (2011). Online word of mouth: Characteristics of Yelp. com reviews.The Elon journal of undergraduate research in communications, 2(1), 37-42.

[4] "Personalizing Yelp Star Ratings: a Semantic Topic Modeling Approach." Jack Linshi. Yale University.

[5] E. Agichtein, C. Castillo, D. Donato, A. Gionis, andG. Mishne. Finding high-quality content in social media with an application to community-based question answering.

[6] J. Jeon, W. Croft, J. Lee, and S. Park. A framework to predict the quality of answers with non-textual features. In Proceedings of SIGIR, 2006.

[7] Chen, P., Dhanasobhon, S., and Smith, M. 2008. "All Reviews Are Not Created Equal: The Disaggregate Impact of Reviews on Sales on Amazon.com," working paper, Carnegie Mellon University

[8] Clemons, E., Gao, G., and Hitt, L. 2006. "When Online Reviews Meet Hyperdifferentiation: A Study of the Craft Beer Industry," Journal of Management Information Systems (23:2), pp. 149-171.

[9] Ghose, A., and Ipeirotis, P. 2006. "Designing Ranking Systems for Consumer Reviews: The Impact of Review Subjectivity on Product Sales and Review Quality," in Proceedings of the 16th Annual Workshop on Information Technology and Systems.

[10] Ghose, A. and Ipeirotis, P.G. Designing novel review ranking systems: Predicting usefulness and impact of reviews. In Proc. International Conference on Electronic Commerce, ACM Press (2007).