Predicting how many ”Useful” votes a yelp review will receive?

Harsh Dani
hdani@asu.edu

Harsh Kumar
hkumar12@asu.edu

Sriram Ganapathyraman
sganapa5@asu.edu

Abstract

With Social Media taking the field of user interactivity and web usability to new dimensions, problems associated with how to use the vast content in a useful and an effective manner still remains. With consumers buying myriad of products on the Internet, lots of content is bound to get generated in the form of user reviews or recommendations. An open question that intrigues both the producers and consumers is how to find the useful content that will interest the user base or satisfy an individual in terms of underlying quality. We attempted to resolve this problem on Kaggle data set by predicting the number of useful votes that a user review will receive. The prediction mechanism makes use of regression techniques to train the extracted feature data. We analyze the use of various regressor and their accuracy value to compare their performance in predicting the votes.

1 Introduction

The problem domain was identified by considering the common challenges that a user faces while browsing through various e-commerce or social networking domains. Most common of those are the lack of useful review comments or votes regarding the product purchase that could motivate them to form an unbiased opinion regarding the trustfulness of the vendor. With the Internet becoming accessible to the common population there is always an inherent ambiguity and lack of significance level in every product, business or service decision that we take a while finalizing our purchase. Consider for example the case of e-commerce giants Amazon or Ebay where each user review being attached with a helpfulness count can greatly influence their business and financial strategy. For this reason only, the practical implementation of this problem domain seems viable to speed up the overall decision process.

There is always a lot of overhead involved in crawling through each set of review comments and determining it’s usefulness value as the density of text at disposal often sways them from the actual worth and price of the purchase. For alleviating these issues websites like yelp include an attribute level property for each review regarding it’s helpfulness count that has been assigned by the potential buyer. This might actually serve the purpose but still issues like the spam intrusion or the worthlessness and negativity of the review degrades the effectiveness of the process. In addition to these,

- Some of the reviews that were recently added to the site will not have a count associated with them and predicting their usefulness value becomes difficult even if they actually serve the purpose.
- Buyers out of natural tendency usually consider reading only a few out of the review comments as the effort involved in going through each of the review comment and predicting it’s positive or negative aspect becomes painstaking sometimes.

In this paper we experiment on the given dataset and understand the task of automating the process of assigning a score to the review comment based upon the training set provided by yelp comprising of features considered over a 8 year period. The regression problem that we solve here proceeds by the selection of a fixed set of attributes from the given User, Business and Checkin and review data and identifying the accuracy count that each combination outputs in predicting the usefulness value. The model developed out of this is trained on significant attributes including the freshness, useful count, length and the number of lines present to name some. So our goal is served by the identification of a unique combination of feature set and regressor that minimize the error percentage between the predicted and the actual rating count.

The rest of the paper is structured as follows. Section 2 includes prior work related to the same problem that we are addressing. Section 3 describes problem definition and observation. Section 4 Methods for predicting helpfulness and Section 5 describes modeling the response. Section 6 is evaluation. In Section 7 we describe conclusion and future work.

1.1 Yelp

Yelp is a networking site that was setup with core focus on bringing local businesses together. Yelps main focus is to help people recommend most useful products among the millions that they usually get. Business owners get a chance to update their product statement and purchase information and with their current recommendation software considering measures of quality, reliability and activity makes them popular among other providers in the market.

1.2 Data

The data set downloaded from kaggle comprises of:

- 229,907 reviews of 11537 businesses in phoenix by 43,873 users
- It also consists of 8282 business check-in's.
- It was collected over a time frame from March 2005 to January 2013.

The link to data-set is `https://www.kaggle.com/c/yelp-recruiting/data`

2 Prior Work

With the increase in the number of online reviews, automatic review mining has generated a lot of research attention. Some of the earlier studies focused solely on determining the sentiment of the review classifying it either as positive or negative. Lina Zhou et al., [1] investigated movie review mining using machine learning and semantic orientation. Supervised classification and text classification techniques were used in their proposed machine learning approach to classify the movie review.

Bo Pang et al., [2] used machine learning techniques to investigate the effectiveness of classification of documents by overall sentiment. Their experiments demonstrated that the machine learning techniques are better than human produced baseline for sentiment analysis on movie review data. However these methods did not consider the helpfulness of each review and were concerned mainly on classifying them into one of the two categories. Our study departs from those review mining approaches as we are trying to predict the importance of each review so that websites could list these reviews based on their importance.

One recent work that is closely related to our study attempts to examine the economic impact of the online reviews [3].The approach followed in this study takes into account the subjectivity of each review by doing a semantic analysis and predicts the reviews that affect the sales as well as those which are helpful to the users. But this method does not take into account other important features related to the review such as reviewers features or the product features or the freshness of the review.

Yang Liu et al [4] have proposed another method which predicts the helpfulness of each review by considering features such as reviewers expertise, reviewers writing style as well as freshness of the review. The freshness of the review is measured using the timestamp of the review. Helpfulness score of a review can significantly depend on the timestamp of the review as users might be interested in getting the views of recent reviewers. In their method they had predicted the helpfulness score of a by constructing a radial basis function by representing the features in the form of a vector. In this method the reviewers expertise is modelled by classifying the reviewer as either prolific or Non-prolific based on the number of reviews the reviewers has posted in that product or business category. But this model might fail to hold since this requires sufficient past reviews of the reviewer in order to achieve meaningful results.

Another relevant study in this area analyzes the trustworthiness of a review and tries to predict the spams that exist in online reviews [5].In this method each review is classified either as untruthful opinion, brand only review or non-review meaning the review text is not even related to product or service. This study provides a novel method to identify the spammers and the results after running this model can be used as a preprocessing step in our approach.

Soo-Min Kim et al [6] had proposed a method to automatically assess review helpfulness which take into account various features which fall under one of these five classes. The five classes of features are syntactical features (e.g., percentage of verbs and nouns), lexical features (e.g., ngrams), structural features (html tags, punctuation, review length), semantic features (e.g., product feature mentions) and metadata (e.g., star rating).The helpfulness score is then predicted by running an SVM regression on these extracted features.

Jiliang Tang et al [7] have proposed a review helpfulness rating prediction method which is context-aware.It takes into account relation between author of the review and the rater of the review as well as the raters trust network.

3 Problem Definition and Observation

Here in this section, we formally define the problem that we are trying to solve and the background research and study that we have done to achieve the same.

3.1 Problem Definition

The formal definition of a helpful review in business terms start with the number of votes that a review comment has actually received by other users either using or buying the same item. However, this definition might seem to be a bit vague in terms of the actual purpose that we are out to set in helping the consumers in getting a product that best suites their needs. Our model being trained can very well be represented as $M \in (u, b, c)$where $u \in U$ denotes the number of users present during the collection,$b \in B$ represents the businesses for which the check-in's and ratings are provided and $c \in C$represent the number of check-in's that the business has received.

The helpfulness of the review was measured by the root mean square logarithm value that we observe in the training and the testing data.

3.2 Observation

Observing from the common facts and trends that we see on major ecommerce websites the top-2 factors that influence the decision of a buyer based on the review comments can be enlisted as follows:

- Freshness: This is the single most important factor that affects the usefulness of a review. Considering the vast business domain on sites like yelp, the trustiness is usually defined by the well established nature of review with respect to time. Reviews written early in the time frame, that starts from the time when the review got published and the time when the actual observations started on the given data set, usually draw more attention from the potential buyers.

- No. of Reviews: Number of reviews that each business has received can be an important feature in predicting the review helpfulness count, as more famous a business more number of users will review it.

Table 1: Extracted Features

DATA SET	FEATURES
Review	Length of review
	No.of words
	Stemmed Review Length
	Freshness
	Stars
Business	Top40Categories
	ReviewCount
	Stars
	City
Check-In	NoofCheck-In's
User	AverageStars
	ReviewCount
	Votes_Cool
	Votes_Funny
	Votes_Useful
Derived	RatioOfVotesUseful over NumberOfReviews
	Delta Of user stars and review stars
	Delta Of business stars and review stars

4 Methods for Predicting Helpfulness

4.1 Data Processing

4.1.1 Processing the Review Data

Before any feature extraction technique to be performed on the text data, cleansing needs to be performed to remove unwanted noise that effectively won't contribute to the final predicted value. The set of tasks that we executed on the review comments were restricted but including the removal of the punctuation words or symbols using `WordPunctTokenizer`, from the text as they effectively won't contribute to the final score count. Then we removed the set of stop words using `nltk` from the input text as due to lack of information gap that they actually convey to the sentence doesn't have any effect on the usefulness value. Lastly we followed a procedure called stemming, using `PorterStemmer`, a famous stemmer algorithm used in `nltk`, where without any significant loss to the meaning of the word we retrieved only the stem or root of the words thus avoiding frequency and processing overhead.

4.1.2 Other Features

In addition to the review text we considered additional features in evaluating our model for finding the final prediction value. The common features that we trained our model upon are listed in Table1 below.

5 Modelling the response

Here we try to model the response parameter for the number of useful votes using various regression models and analyze the results.

Figure 1 describes the Top-10 features identified by Gradient Boosting regressor.

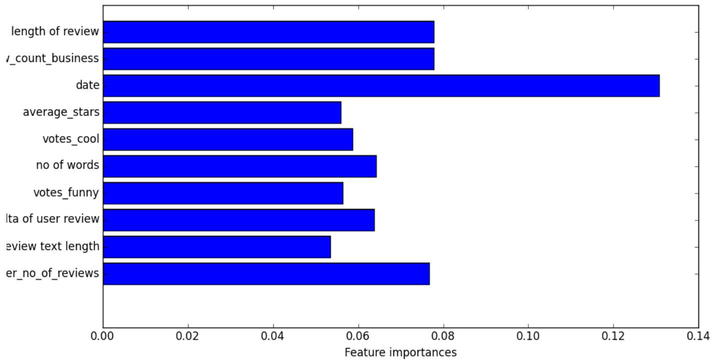

Figure 1: Top 10 Features identified by Gradient Boosting Regressor

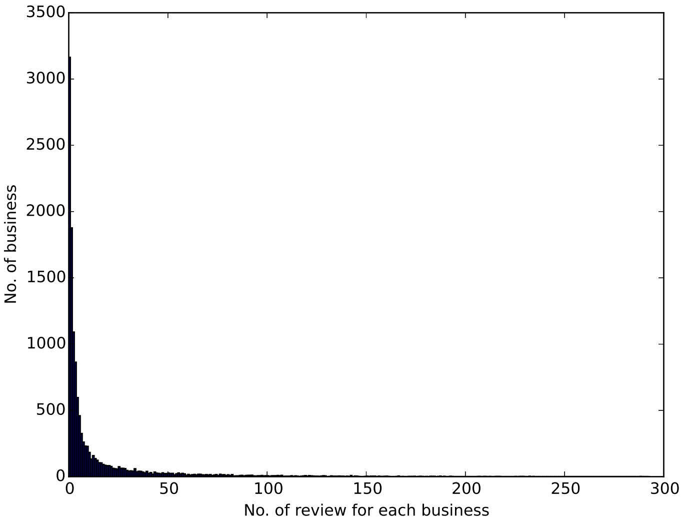

Figure 2: Number of Business Reviews

The distribution for the number of useful reviews for each business follows a power law distribution as illustrated in Figure 2.

With number of businesses in a city serving as a feature value we tried visualizing the most frequent cities occurring in the business data in Figure 3.

For the review length as a feature we have the distribution as shown in Figure 4.

Since, the distribution is mostly non-linear for each of the training features, so we tried fitting a non parametric Gradient-Boosting, Random-Forest, Lasso Regression and AdaBoost regression model on the response parameter.

The result for the predicted number of helpful votes for reviews after fitting the regression models is shown in Figure 5.

Above result can be compared with the actual useful votes at our disposal using the training test as shown in Figure 6.

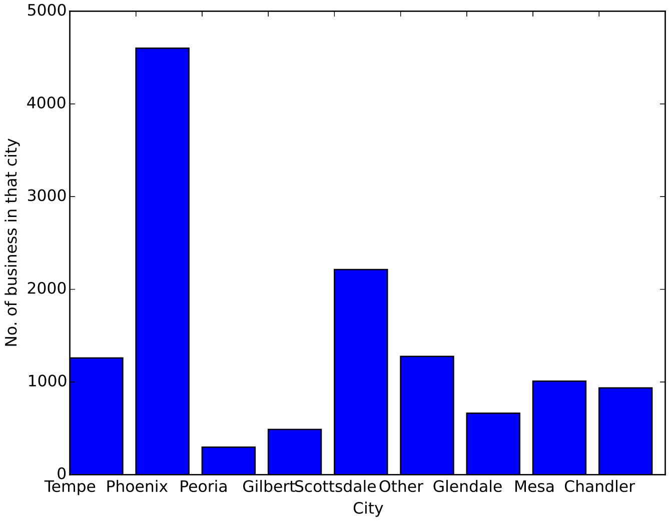

Figure 3: Number of Businesses in City

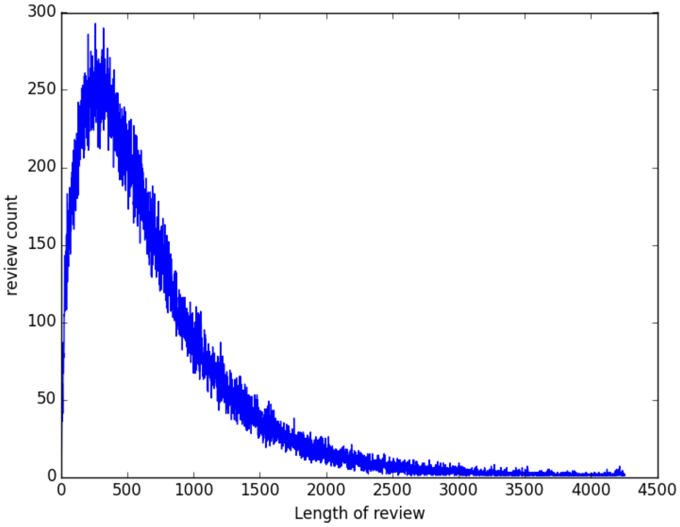

Figure 4: Review Length

6 Evaluation

For evaluating the result that we got after training the model, we employed the root mean square logarithmic error (RMSLE).

$\texttt{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_{ij}+1)-\log(a_{ij}+1))^2}$

The rmsle count prediction values that we got using different models is listed in Table 2.

The plot for the predicted values for each of the regression models is shown in Figure 7.

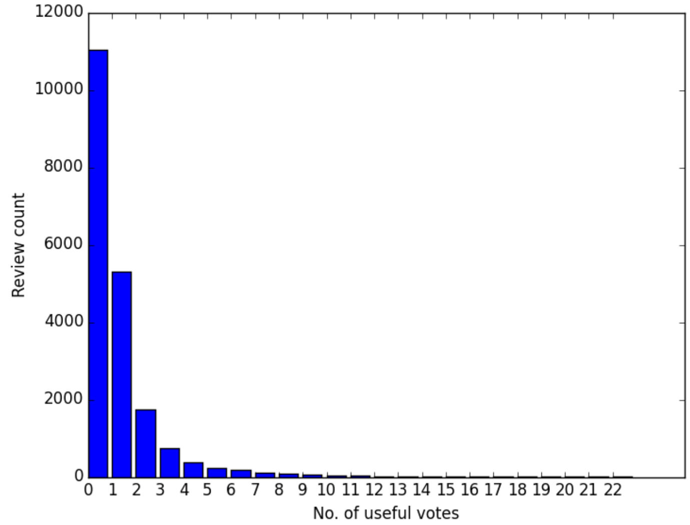

Figure 5: Predicted Useful Votes

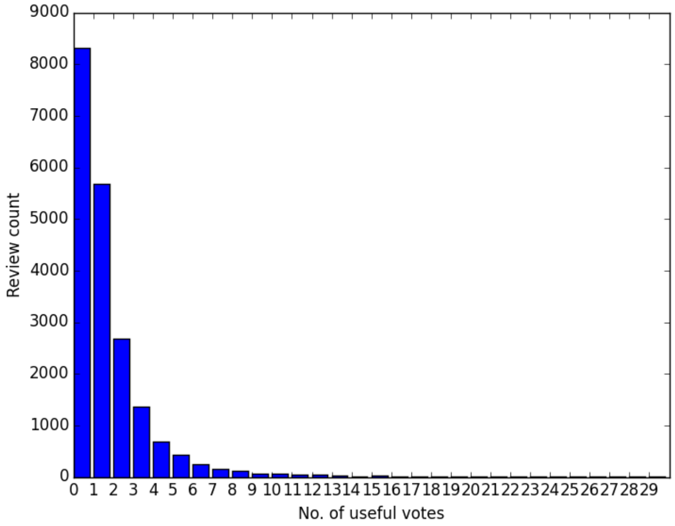

Figure 6: Actual Useful Votes

7 Conclusion and Future Work

In this paper we analyzed the implementation of the regression models in predicting the usefulness of the reviews and bettering the results using different combinations of feature vectors. Various feature vectors projections confirmed our understanding of proceeding with the respective models. We can confirm that Gradient-Boosting model almost always guarantees a better accuracy in predicting values in comparison with other regression models. Our study focused on the analysis of yelp data set provided to be kaggle and proceeded in an effective manner starting from the pre-processing of the training text and then, the training of the various regression models and reporting the observed values. In future work, we would like to work upon the trustiness model assignment to each of the users based upon the useful votes count that they have received and then following an iterative

Table 2: Comparision between various regressor

Model	TrainingData	ValidationData	RMSLE
Gradient-Boosting	45,000	5000	0.477
	90,000	10,000	0.471
	1,80,000	20,000	0.464
Random-Forest	45,000	5000	0.49
	90,000	10,000	0.498
	1,80,000	10,000	0.489
Lasso-regression	45,000	5000	0.5227
	90,000	10,000	0.5241
	1,80,000	20,000	0.5211
Adaboost	45,000	5000	0.73
	90,000	10,000	0.76
	1,80,000	20,000	0.90

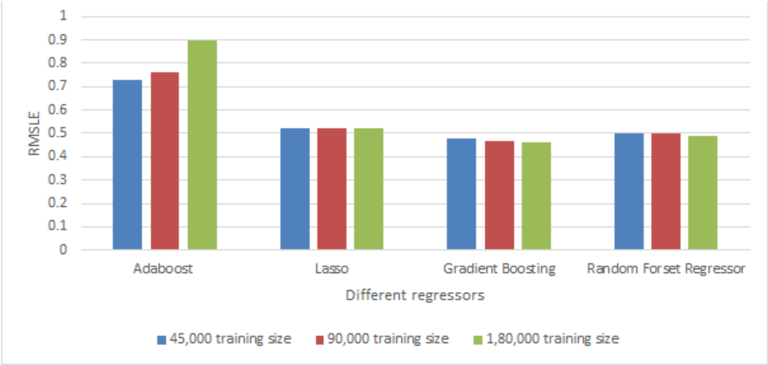

Figure 7: Predicted Values

updating algorithm for the usefulness and trustfulness score assignment. [8] We would also like to work upon the extraction of bigram features from the review text and use them as a feature. The idea for extracting emotions out of the review text and score assignment can be used as a feature, and we estimate an improved prediction value.

References

[1]Lina Zhou,Pimwadee Chaovalit, M̈ovie review mining:a comparison between Supervised and Unsupervised Classification Approaches", Proceedings of the 38th Hawaii International Conference on system sciences, 2005.

[2] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, In Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP), pages 7986, 2002

[3] A. Ghose, P. G. Ipeirotis, D̈esigning novel review ranking systems: predicting the usefulness and impact of reviews", In ICEC, pages 303310, 2007.

[4] Yang Liu, Xiangji Huang, Aijun An, Xiaohui Yu, M̈odeling and Predicting the Helpfulness of Online Reviews", IEEE International Conference on Data Mining (ICDM), 2008.

[5] N. Jindal, B. Liu, Öpinion spam and analysis", Web Search and Data mining (WSDM), pages 219230, 2008.

[6] S. Kim, P. Pantel, T. Chklovski, M. Pennacchiotti, Äutomatically assessing review Helpfulness", In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2006

[7] Jiliang Tang, Huiji Gao, Xia Hu, Huan Liu, ”Context-Aware Review Helpfulness Rating”, 7th ACM Confereence on Recommendation System (RecSys), 2013.

[8] D. Shizanki, K. Stuckman, R. Yates, T̈rust and helpfulness in Amazon Reviews: Final Report”, Stanford University, 2013.