Utility Scoring of Product Reviews

Zhu Zhang
zhuzhang@u.arizona.edu
Department of Management Information Systems,
University of Arizona

Balaji Varadarajan
vbalaji@cs.arizona.edu
Department of Computer Science,
University of Arizona

Online shoppers generally go through other people's reviews of a specific product before actually buying it; manufacturers can also examine product reviews to observe customer opinions and make improvements or predict market trends. Now it's undeniable that product reviews aren't equally "useful". In this paper, we identify a new task in text sentiment analysis: predicting the utility of product reviews, which is orthogonal to polarity classification and opinion extraction. We build regression models and classification models by incorporating a diverse set of features and achieve highly competitive performance for utility scoring and review classification on real-world data sets.

When provided with a variety of positive and negative, useful and not-so-useful reviews, how should one utilize such a diverse set of reviews and come to a decision? We conceive the following "weighted average" framework, which can be considered as a motivating thought at a very general level:

$$E(P) = \frac{\sum_{i=1}^{n} u(T_i(P)) * Polarity(T_i(P))}{\sum_{i=1}^{n} u(T_i(P))} \tag{1}$$

in which the evaluation $E(P)$ of a product P is a weighted average of the polarity of each individual review $T_i(P)$.

While much previous work has been done on predicting the polarity of text, $Polarity(T_i(P))$, we attempt to approach the utility of reviews, $u(T_i(P))$, which is a new and important research problem. The utility of a product review is a property orthogonal to its polarity or embedded opinions. Our goal in this research is to make a computational model to predict the utility of reviews.

We view the problem as one of regression. Formally, given a product review T, a number of features $f_1(T), ..., f_j(T), ..., f_p(T)$ can be computed. Our task is to approximate a function

$$u(T) = F(f_1, ..., f_j, ..., f_p) \tag{2}$$

The output $u \in [0,1]$ should reflect the real utility of T as accurately as possible. Given an estimated function F, we are able to use the subsequent metrics to judge its quality, both of which are standard in regression analysis:

(a) Squared correlation coefficient

$$r^2 = \frac{(\sum_{i=1}^{n}(u_i - \bar{u})(\hat{u}_i - \bar{\hat{u}}))^2}{(\sum_{i=1}^{n}(u_i - \bar{u}))^2(\sum_{i=1}^{n}(\hat{u}_i - \bar{\hat{u}}))^2} \tag{3}$$

(b) Mean squared error

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(u_i - \hat{u}_i) \tag{4}$$

In both equations above, u_i and $\hat{u}_i$ are the real and predicted utility scores respectively; $\bar{u}$ and $\bar{\hat{u}}$ represent the mean of the corresponding sample respectively.

We collected data of three different domains, i.e. *Canon* Products, *Engineering* Books and $PG-13$ movies from AWS and CNET.

Now coming to experiment we do it with two types of regression algorithms:

(a) ε-Support Vector Regression (ε-SVR) implemented in LIBSVM

(b) Simple linear regression (SLR) implemented in WEKA.

In both cases, we apply the original algorithms as they are implemented in the machine learning packages.

In general, a good product review is a "reasonable" mixture of subjective valuation and objective information. Given a product review text T, we compute the following features and feed them into the regression algorithms:

(a) Lexical Similarity Features (*LexSim*): Clearly a review should not be literal copy or loyal rephrase of the product specification S and an editorial review E, through this we measure the similarity between customer review and product specification, $sim(T,S)$, and that between customer review and editorial review, $sim(T,E)$, respectively.

Table 1: Regression Performance on Canon Product Reviews

Feature Set	ε-SVR		SLR	
	r^2	σ^2	r^2	σ^2
LexSim	0.0049	0.0957	0.0064	0.08
ShallowSync	0.2726	0.0601	0.0433	0.0772
LexSubj	0.0448	0.0902	0.0081	0.0806
ALL	0.3028	0.0565	0.0892	0.0736

Table 2: Regression Performance on Engineering Book Reviews

Feature Set	ε-SVR		SLR	
	r^2	σ^2	r^2	σ^2
LexSim	0.0216	0.0947	0.0232	0.0874
ShallowSync	0.31276	0.0615	0.0895	0.0816
LexSubj	0.0674	0.0907	0.0424	0.0857
ALL	0.3514	0.0581	0.1244	0.0786

Table 3: Regression Performance on PG-13 Movies Reviews

Feature Set	ε-SVR		SLR	
	r^2	σ^2	r^2	σ^2
LexSim	0.0014	0.1484	0.0467	0.1347
ShallowSync	0.4176	0.0829	0.0905	0.1285
LexSubj	0.0412	0.1479	0.0244	0.1376
ALL	0.4145	0.0826	0.1571	0.1193

(b) Shallow Syntactic Features (*ShallowSyn*): We compute counts of words with the part-of-speech tags, sentences in T, in order to characterize the subjectivity- objectivity mixture of the text at a shallow syntactic level.T.

(c) Lexical Subjectivity Clues (*LexSubj*): In this set of features, we capture the subjectivity-objectivity-mixture at a lexical semantic level.

To acquire the target value of the regression model/to enlist the gold-standard definition of $u(T_i)$. we make use of review's usefulness ("x out of y people found the following review helpful"). This actually provides a direct and convenient way to approximate the gold-standard utility value of a given review. Formally, we define the utility as:

$$u = \frac{x}{y} \tag{5}$$

One might intuitively expect that the utility of a review strongly correlates with its length (in a positive way). However, results show that the correlation between the two variables is in fact very weak. Therefore it is necessary to build non-trivial regression models.

Here results presented in Table 1,2,3 are based on 10-fold cross validation, from this we can observe that:

(a) Across all three collections, the results are relatively similar qualitatively. The strongest model for each collection always achieves $r^2 >$ 0.30 and σ^2< 0.10, and apparently outperforms the length-based baseline.

b) Generally speaking, SVR significantly outperforms SLR, SVR is also more resistant to outliers, from which SLR sometimes suffer .

(c) The set of lexical similarity features play a very minor role in the regression model, this implies instead, the utility is based on inherent properties of the review itself.

(d) The lexical subjectivity clues have very limited influence on the utility scoring. This means that the perceived "usefulness" of a product review barely correlates with the subjectivity or polarity embedded in the text.

(e) The shallow syntactic features account for most predicting power of the regression model. This phenomenon demonstrates that high-utility reviews do stand out due to the linguistic styles in which they are written.