Utility Scoring of Product Reviews

Zhu Zhang
Department of Management Information Systems
University of Arizona
Tucson, AZ 85721
zhuzhang@u.arizona.edu

Balaji Varadarajan
Department of Computer Science
University of Arizona
Tucson, AZ 85721
vbalaji@cs.arizona.edu

ABSTRACT

We identify a new task in the ongoing research in text sentiment analysis: predicting utility of product reviews, which is orthogonal to polarity classification and opinion extraction. We build regression models by incorporating a diverse set of features, and achieve highly competitive performance for utility scoring on three real-world data sets.

Categories and Subject Descriptors

H.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Miscellaneous; I.2.7 [**ARTIFICIAL INTELLIGENCE**]: Natural Language Processing—*Language parsing and understanding*; I.2.6 [**ARTIFICIAL INTELLIGENCE**]: Learning—*Concept learning, Induction*

General Terms

Algorithms, Experimentation

Keywords

text sentiment analysis, utility, regression

1. INTRODUCTION

Recently, there has been a swell of interest in text subjectivity and sentiment analysis. In particular, product review data have been heavily used for subjectivity classification, polarity prediction, and opinion extraction, due to their linguistic properties and their implications for E-commerce applications.

Online shoppers often wade through other people's reviews of a certain product to gauge their shopping decision; manufacturers may also examine product reviews to monitor customer opinions and predict market trends. In both scenarios, the review readers seek (relatively) unbiased evaluation of a given product, by leveraging information from multiple reviews, although each individual review can be subjective in nature.

CIKM'06, November 5–11, 2006, Arlington, Virginia, USA.

When thinking along this line, product reviews are not equally "useful". Below is a customer review[1] for a digital camera, Canon Powershot SD300. While the review encompasses mixed feeling about the product, it is found helpful by 170 out of 178 Amazon shoppers.

> *Handling: The interface is similar to that in other Canon digital compacts, which helps your learning curve. The case is in metal, except the USB and the battery cover. They are both made of plastic and feel very fragile. The metal tripod mount is located closely below the lens. The LCD screen is reasonably viewable under daylight conditions.*
>
> *Canon celebrated that SD300 is the first compact that uses the DIGIC II processing, and with my experience so far, the camera does respond faster when compared with DIGIC based ones. It however appears to be on par with recent Sony and Olympics models. I did not measured the various response times scientifically however.*
>
> *Picture quality: Contrary to other comments, I was not "blown away" by quality of the pictures. The lens produces not serious, but significant purple edges in bright sunlight, and shows problems with dark corners like other compacts. Color production is rich with high contrast, a big plus. Sadly I find the pictures appear noisy when taking under low light conditions. I suspect either I have a faulty unit, or there are some design issues?*
>
> *Complaints:*
>
> - *Can't review picture histogram easily (two to three steps)*
> - *Noisy operations*
> - *No battery level indicator*
>
> *So far I find the SD300 a good and decent pocket camera. However in the same market (similar specs and price) there are many other choices, and the SD300 does not excel specifically in any area.*

On the other hand, imagine product reviews as simple as

> *X is the greatest product I've ever seen.*

[1]From `http://www.amazon.com`

or

Product Y sucks.

They certainly encode strong polarity and reflect strong opinion of their authors. However, they are not particularly reliable or useful in informing their readers' shopping decisions.

When presented with a mixed bag of positive and negative, useful and not-so-useful reviews, how should one leverage the diverse evidence? We envision the following "weighted average" framework, which should be considered as a motivating thought at a very general level:

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

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

Equation (1) implies two orthogonal characteristics of product reviews. While much previous research has been focusing on predicting the the polarity of text, $Polarity(T_i(P))$, we try to approach the the utility of reviews, $u(T_i(P))$, which is a new and important research problem. Polarity and utility of reviews are orthogonal only in the sense that they can be modeled independently; they are certainly relevant and to be integrated in a framework like Equation (1).

The paper is organized as follows:

After reviewing related work in text subjectivity and polarity analysis, we formally define utility prediction as a regression problem. We present and analyze experimental results on three Amazon customer review collections before concluding the paper with remarks on future work.

2. RELATED WORK

2.1 Subjectivity in Text

Subjectivity in natural language refers to aspects of language used to express opinions, evaluations, and speculations. [19] provided a good overview of work in learning subjective language from corpora. Clues of subjectivity are generated and tested, including low-frequency words, collocations, and adjectives and verbs identified using distributional similarity. The features are also examined working together in concert. In addition, the authors showed that the density of subjectivity clues in the surrounding context strongly affects how likely it is that a word is subjective. Finally, the subjectivity clues are used to perform opinion piece recognition (a type of text categorization and genre detection). In a somewhat different perspective, [21] presented experimental results in classifying the strength of opinions in text. A wide range of features were used, including new syntactic features developed for opinion recognition.

Yu and Hatzivassiloglou [23] approached the problem of separating opinions from fact, at both the document and sentence level. They presented a Bayesian classifier for discriminating between documents with a preponderance of opinions such as editorials from regular (fact-based) news stories, and described three unsupervised statistical techniques for the task of detecting opinions at the sentence level. A model for sentence-level polarity classification was also discussed.

The importance of acquiring lexical clues for subjectivity analysis has been recognized. [18] tried to learn a collection of subjective adjectives using word clustering according to distributional similarity [5], seeded by a small amount of detailed manual annotation. With a finer taxonomy of subjective adjectives in mind, [4] learned sets of dynamic adjectives, semantically oriented adjectives, and gradable adjectives from corpora using a simple log-linear model, and established that these adjectives are strong predictors of subjectivity. Not only adjectives, but also nouns can be subjective. In [13], a bootstrapping technique is used to learn subjective nouns from corpora. At the phrase level, [11] presented a bootstrapping process that learns linguistically rich extraction patterns for subjective (opinionated) expressions. High-precision classifiers are used to label unannotated data and automatically create a large training set, which is then given to an extraction pattern learning algorithm. The learned patterns are then used to identify more subjective sentences. The bootstrapping process learns many subjective patterns and increases recall in subjectivity identification while maintaining high precision.

Instead of viewing subjectivity as a sentence- or passage-level property and studying the "overall" polarity of text, [20] focused on phrase-level sentiment analysis. They first determined whether an expression is neutral or polar and then disambiguated the polarity of the polar expressions. With this approach, the system was able to automatically identify the contextual polarity for a large subset of sentiment expressions, achieving results that are significantly better than baseline.

In [2], another aspect of opinion analysis was pursued: identifying the sources of opinions, emotions, and sentiments. The problem was viewed as an information extraction task; a hybrid approach was adopted, which combines Conditional Random Fields and extraction pattern learning.

Subjectivity analysis has also been shown to be useful for other NLP applications such as information extraction [12] and question answering [15].

2.2 Mining Polarity and Opinions in Product Reviews

Due to their inherent subjective nature and their apparent implications for business activities, product review data have been a popular target for research on polarity analysis and opinion extraction, as instances of text subjectivity research.

One of the early papers in this domain is [17], which presented a simple unsupervised learning algorithm for classifying reviews as "recommended" (thumbs up) or "not recommended" (thumbs down). Specifically, the semantic orientation of a phrase is calculated as the mutual information between the given phrase and the word "excellent" minus the mutual information between the given phrase and the word "poor". In turn, the polarity of a review is predicted by the average semantic orientation of the phrases in the review that contain adjectives or adverbs.

[3] identified a set of features for automatically distinguishing between positive and negative reviews, and empirically compared a number of classifiers on CNET and AMAZON review data.

[9] also considered the problem of classifying documents by overall sentiment, e.g., determining whether a review is

positive or negative. Using the IMDB movie review data[2], the authors showed that standard machine learning techniques outperform human-produced baselines. However, the three machine learning methods employed (Naive Bayes, maximum entropy classification, and support vector machines) do not perform as well on sentiment classification as on traditional topic-based categorization, due to challenges related to non-compositional semantics and discourse structures. [7] studied the same problem with an additional tweak. Subjective portions of text are first extracted using a graph min-cut algorithm, and then fed into text categorization algorithms to approach sentiment polarity classification.

Pushing further along the same line, [8] addressed the rating-inference problem, wherein rather than simply decide whether a review is "thumbs up" or "thumbs down", one must determine an author's evaluation with respect to a multi-point scale (e.g., one to five stars). A metric labeling approach was compared with both multi-class and regression versions of SVMs.

Shifting from classification to extraction, [10] introduced OPINE, an unsupervised information extraction system which mines reviews in order to build a model of important product features, their evaluation by reviewers, and their relative quality across products.

In a similar effort, [6] proposed a novel framework for analyzing and comparing consumer opinions of competing products. A prototype system called Opinion Observer is also implemented. Through the visualization offered by the system, the user is able to clearly see the strengths and weaknesses of each product in the minds of consumers in terms of various product features. Supervised rule discovery techniques are used to extract product features and corresponding Pros and Cons.

3. PROBLEM DEFINITION

Utility (or, reliability, usefulness, informativeness) is not the same as indifference. Totally indifferent or neutral reviews are useless; well-grounded subjective opinions can be convincing and illuminating. In other words, the utility of a product review is a property orthogonal to its polarity or embedded opinions. Our goal in this research is to build 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 $f_1(T),\ldots,f_j(T),\ldots,f_P(T)$ can be computed. Our task is to approximate a function

$$u(T) = F(f_1, \ldots, f_j, \ldots, f_P)$$

The output $u \in [0, 1]$ should reflect the real utility of T as accurately as possible.

Given an estimated function F, we can use the following metrics to evaluate its quality, both of which are standard in regression analysis:

- Squared correlation coefficient

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

- Mean squared error

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

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.

4. DATA COLLECTION

Although a number of online shopping sites such as CNET[3] can be potential sources for customer review data, Amazon offers a relatively convenient set of APIs, Amazon Web Services (AWS)[4], through which various product data including customer reviews can be accessed.

We downloaded reviews and related data in three different domains: Electronics, Video, and Books. For electronics and books, we used keyword-based search to identify Canon products (*Canon*) and engineering books (*Engineering*); for videos, we used the "AudienceRating" field to retrieve PG-13 movies (*PG-13*).

We developed a set of Perl scripts to automate the data collection process. Relevant to our study, we downloaded the following data:

1. Customer Reviews for product items whose ASIN (Amazon Standard Identification Number) is seen in the search response to the ItemSearch queries corresponding to the three domains above.

2. Product meta data (including both product description and Amazon Editorial Reviews) corresponding to the product items.

3. All reviews written by a customer whose review appears in the data collected in 1.

For all customer-written reviews, we also collected the votes they get regarding there usefulness ("x out of y people found the following review helpful"). These data are used to approximate the target value of the regression model (see Section 6.1).

The basic statistics of the three collections are summarized in Table 1.

Collection	# Reviews	# Authors
Canon	2,394	2,384
Engineering	6,120	6,032
PG-13	11,543	11,445

Table 1: Statistics of Amazon Customer Review Data

It should be noted that AWS restricts the number of search results to a maximum of 2500 records per query. Therefore our data collection is constrained by this limit.

[2] http://reviews.imdb.com/Reviews/

[3] http://www.cnet.com

[4] http://aws.amazon.com/ecs

5. UTILITY SCORING USING STATISTICAL REGRESSION MODELS

In this section, we discuss two aspects of the statistical learning framework: the learning (regression) algorithms and the features.

5.1 Learning Algorithms

We experiment with two types of regression algorithms:

- ϵ-Support Vector Regression (ϵ-SVR) implemented in LIBSVM [1]. Just as SVM represents the state of the art in machine learning, SVR is attractive due to the power of different kernel functions. In this study, we use the radial basis kernel function (RBF), which handles the potentially non-linear relationship between the target value and features.
- Simple linear regression (SLR) implemented in WEKA [22]. SLR is a classical regression tool and has been widely used in numerous applications. It serves as a reasonable baseline.

In both cases, we apply the original algorithms as they are implemented in the machine learning packages. It is not our intention to make contribution to the learning algorithms in this study.

5.2 Features

Generally speaking, a good product review is a "reasonable" mixture of subjective valuation and objective information. The feature space in the statistical learning framework ought to capture this linguistic phenomenon.

Given a product review text T, we compute the following features and feed them into the regression algorithms.

5.2.1 Lexical Similarity Features (*LexSim*)

Clearly an informative customer review should not be a literal copy or loyal rephrase of the product specification S, because it should reflect the customer's own experience with the product, not the manufacturer's description or expectation. On the other hand, a good review is supposed to base subjective judgement on objective observation, therefore it should echo the product specification to a reasonable extent, in a positive or negative tone.

Similar situation is conceivable between a customer review and an editorial review E, the latter of which approximates a relatively objective and authoritative view of the product.

With these motivations in mind, 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.

We use the standard cosine similarity in vector space model, with TF*IDF term weighting, as defined in information retrieval literature [14].

5.2.2 Shallow Syntactic Features (*ShallowSyn*)

We compute counts of words with the following part-of-speech tags in T, in order to characterize the subjectivity-objectivity mixture of the text at a shallow syntactic level:

- **Proper nouns**: reference to existing, maybe technical, concepts.
- **Numbers**: the tendency of quantification.
- **Modal verbs**: reflection of certainty, confidence, mood, etc., which are all instances of modality.
- **Interjections**: signals for emotion
- **Comparative and superlative adjectives**: indicators of comparison
- **Comparative and superlative adverbs**: indicators of comparison, again
- **Wh-determiners, wh-pronouns, possessive wh-pronouns, wh-adverbs**: wh-words that signify either questions or other interesting linguistic constructs such as relative clauses.

We also compute simple counts such as number of words and number of sentences in the review text T.

5.2.3 Lexical Subjectivity Clues (*LexSubj*)

This is an interesting set of features, which we use to capture the subjectivity-objectivity-mixture at a lexical semantic level, by taking advantage of lexical resources created by other researchers.

Specifically, we calculate counts of words in the following clue lists respectively:

1. The list of subjective adjectives learned in [18]. More precisely, the list of adjectives learned using the process presented in [18], but from a larger corpus than that in the original paper[5].
2. The list of subjective adjectives learned in [4]. More specifically, it contains the following categories:
   - Dynamic adjectives (e.g., "careful", "serious").
   - Polarity *Plus* adjectives (e.g., "amusing"), manually or automatically identified.
   - Polarity *Minus* adjectives (e.g., "awful"), manually or automatically identified.
   - Gradability *Plus* adjectives (e.g., "appropriate"), manually or automatically identified.
   - Gradability *Minus* adjectives (e.g., "potential"), manually or automatically identified.
3. The list of strong subjective nouns and weak subjective nouns generated by Basilisk [16].
4. The list of strong subjective nouns (e.g., "domination", "evil") and weak subjective nouns (e.g., "reaction", "security") generated by MetaBoot [13].

In order for the learned model to generalize well, we do not count the frequency of each individual word. Instead, we only count the total occurrences of words in each list. For example, if the review text T contains 5 weak subjective nouns, we give a value "5" to the feature "WeakSubjNoun" instead of assigning the value "1" to 5 binary features.

In total, we have 14 word lists, therefore 14 features in this category.

[5]see `http://www.cs.pitt.edu/~wiebe/pubs/aaai00/`

6. EXPERIMENTS AND ANALYSIS

6.1 Data Treatment

The only thing we have not specified so far is how to acquire the target value of the regression model, given a review text T_i. In other words, we have to operationalize the gold-standard definition of $u(T_i)$.

Notice that almost every Amazon customer review comes with a vote regarding its usefulness ("x out of y people found the following review helpful"). This actually provides a direct and convenient way to approximate the gold-stand utility value of a given review. Formally, we define the utility as

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

In our experiments, we only use the reviews with at least 10 votes (i.e., $y > 10$), in order to ensure the robustness of the regression model.

Table 2 summarizes the distribution of review utility scores in the four Amazon collections, as well as their total number of distinct reviews after filtering out duplicates and those with no more than 10 votes.

Collection	Mean	Std. Dev.	# Reviews
Canon	0.7914	0.2839	624
Engineering	0.7531	0.2992	1,255
PG-13	0.5605	0.3756	654

Table 2: Statistics of Data Sets after Treatment

6.2 Experimental Results

Before we present the regression results, one might intuitively expect that the utility of a review strongly correlates with its length (in a positive way). However, as we can see from Table 3, the correlation between the two variables is in fact very weak. Therefore it is necessary to build non-trivial regression models.

Collection	r^2
Canon	0.0042
Engineering	0.0997
PG-13	0.0853

Table 3: Correlation between Utility Score and Review Length

The regression performance on the three review collections (*Canon*, *PG-13*, and *Engineering*) are summarized in Table 4, 5, and 6 respectively. All results presented in this section are based on 10-fold cross validation.

In all three tables, the rows represent different combinations of features; and the columns correspond to the performance of SVR and SLR, measured by squared correlation coefficient r^2 and mean squared error σ^2 respectively. The most competitive results are marked by bold fonts.

6.3 Discussion

We have the following observations, based on the experimental results.

- Across all three collections, the results are relatively similar qualitatively. The strongest model for each collection always achieves $r^2 > 0.30$ and $\sigma^2 < 0.10$, and apparently outperforms the length-based baseline.
- Generally speaking, SVR significantly outperforms SLR, mostly due to its regularized model and the non-linear power of its kernel function. Further more, since the prediction model only depends on a subset of the training data, SVR is more resistant to outliers, from which SLR sometimes suffer (e.g., producing very large mean squared error).
- Looking at the feature vector for the regression models, different groups of features have different effect on the final output:
  - The set of lexical similarity features play a very minor role in the regression model. Intuitively, how useful a review is does not necessarily correlate with how "similar" it is to the corresponding product specification or authoritative review. Instead, the utility is based on inherent properties of the review itself.
  - 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. It also justifies the "orthogonal" view presented at the beginning of the paper.
  - 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.

7. CONCLUSION AND FUTURE WORK

In this study, we identifed a new task in the ongoing research in text sentiment analysis: predicting utility scores of product reviews, which is orthogonal to polarity classification and opinion extraction. The motivation is to leverage information from multiple sources (reviewers). We viewed the problem as one of regression, and built regression models by incorporating a diverse set of features, which achieved highly competitive performance on three Amazon product review collections. In particular, the shallow syntactic features turned out to be the most influential predictors, which indicates that the perceived utility of a product review highly depends on its linguistic style.

The following directions are identified for future work:

- Utility prediction is only one aspect of text sentiment analysis. It is certainly desirable to consider it in concert with polarity classification, opinion extraction, strength scoring, and other aspects, in an integrated framework. An immediate first step is to, based on the idea laid out in Equation (1), model the expected valuation of products by leveraging opinions from multiple customers.
- In the context of online shopping, it makes sense to incorporate stronger user profiling (e.g., demographic

Feature Set	ϵ-SVR		SLR	
	r^2	σ^2	r^2	σ^2
LexSim	0.0049	0.0957	0.0064	0.0800
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 4: Regression Performance on Canon Product 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

Table 5: Regression Performance on PG-13 Movie Reviews

info, shopping history, etc.) into text sentiment analysis. On the other hand, text sentiment analysis is a useful tool for businesses. Various interesting applications, such as market trend tracking and customer relation management, are foreseeable.

8. REFERENCES

[1] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[2] Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics.

[3] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 519–528, New York, NY, USA, 2003. ACM Press.

[4] V. Hatzivassiloglou and J. M. Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th conference on Computational linguistics*, pages 299–305, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

[5] D. Lin. Automatic retrieval and clustering of similar words. In *COLING-ACL*, pages 768–774, 1998.

[6] B. Liu, M. Hu, and J. Cheng. Opinion observer: analyzing and comparing opinions on the web. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 342–351, New York, NY, USA, 2005. ACM Press.

[7] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 271–278, Barcelona, Spain, July 2004.

[8] B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[9] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, Philadelphia, July 2002. Association for Computational Linguistics.

[10] A.-M. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 339–346, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics.

[11] E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. In M. Collins and M. Steedman, editors, *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 105–112, 2003.

[12] E. Riloff, J. Wiebe, and W. Phillips. Exploiting subjectivity classification to improve information extraction. In *AAAI*, pages 1106–1111, 2005.

[13] E. Riloff, J. Wiebe, and T. Wilson. Learning subjective nouns using extraction pattern bootstrapping. In W. Daelemans and M. Osborne, editors, *Proceedings of CoNLL-2003*, pages 25–32. Edmonton, Canada, 2003.

[14] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.

[15] V. Stoyanov, C. Cardie, and J. Wiebe. Multi-perspective question answering using the opqa corpus. In *Proceedings of Human Language Technology*

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 6: Regression Performance on Engineering Book Reviews

Conference and Conference on Empirical Methods in Natural Language Processing, pages 923–930, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics.

[16] M. Thelen and E. Riloff. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 214–221, Philadelphia, July 2002. Association for Computational Linguistics.

[17] P. D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 417–424, Morristown, NJ, USA, 2001. Association for Computational Linguistics.

[18] J. Wiebe. Learning subjective adjectives from corpora. In *AAAI/IAAI*, pages 735–740, 2000.

[19] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. Learning subjective language. *Computational Linguistics*, 30(3):277–308, 2004.

[20] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics.

[21] T. Wilson, J. Wiebe, and R. Hwa. Just how mad are you? finding strong and weak opinion clauses. In *AAAI*, pages 761–769, 2004.

[22] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2005.

[23] H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136, Morristown, NJ, USA, 2003. Association for Computational Linguistics.