Relation between a Book Review Content and Its Rating

P. Tanawongsuwan
Graduate School of Applied Statistics
Information Technology Center
National Institute of Development Administration (NIDA)
Thailand

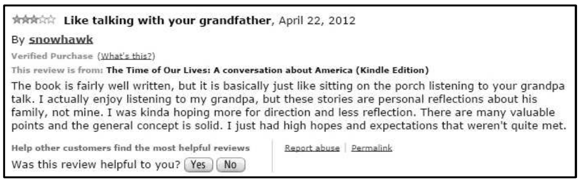
***Abstract*—While millions of products and services are consumed everyday, consumers may share their experiences by posting a review on the Web. A review usually consists of paragraphs of text. Readers wishing to learn details may choose to read the textual description. In many reviews, a product rating is also provided. Such rating might appear simply as a number of stars, a list of adjectives (excellent, good, poor, etc) to choose from. Readers may quickly look at those ratings, just to get a rough idea about the tone of the reviews. This research aims to discover the relation between a product review content and its rating. Specifically, different parts of speech in the review text are analyzed and used to predict the rating. Some classification algorithms are employed as discovery tools. Data used for this study in particular is a number of book reviews from amazon.com's Books department.**

Keywords-text analysis; text mining; opinion mining; book review; product review

I. INTRODUCTION

A. *Hypothesis*

When a person is considering buying a product, he often seeks related information from sources including a review. A review is usually in the form of a textual description, though many come with pictures or video clips. A number of review formats also contain some kind of reviewer's rating of the product, to summarize the reviewer's overall opinion, and for reader's convenience of a quick glance. Examples of such rating might be a star rating of 1-5, as seen in Fig. 1, a list of descriptive adjectives (excellent, good, fair, poor, etc). This research studies the relation between a review's textual content and its rating. Specifically, a hypothesis is that different parts of speech in the text can help predict the product's rating.



★★★☆☆ **Like talking with your grandfather**, April 22, 2012

By **snowhawk**

Verified Purchase (What's this?)

This review is from: **The Time of Our Lives: A conversation about America (Kindle Edition)**

The book is fairly well written, but it is basically just like sitting on the porch listening to your grandpa talk. I actually enjoy listening to my grandpa, but these stories are personal reflections about his family, not mine. I was kinda hoping more for direction and less reflection. There are many valuable points and the general concept is solid. I just had high hopes and expectations that weren't quite met.

Help other customers find the most helpful reviews

Was this review helpful to you?

FIGURE I. A BOOK REVIEW FROM AMAZON.COM.

B. *Related Works*

As for related work, quite a number have been found. Reference [1], [2], and [3] all look into determining whether a product review has a positive or negative tone. Reference [3] presents a prototype system for mining topics and sentiment orientation jointly from free text customer feedback, with application to a database of car reviews.

Though using different methods, [4], [5], and [6] also classify product reviews as being either positive or negative. Reference [6] presents an unsupervised learning algorithm for the classification task, in which a review is classified based on the average semantic orientation of the phrases in the review that contain adjectives or adverbs.

Through classification algorithms, [7] presents a system that evaluates the helpfulness of a book review, based mainly on its content. Ref [8] has a similar goal, though it uses a different approach.

Focused on online customer reviews of products, [9] proposes a framework for analyzing and comparing consumer opinions of competing products. A technique based on language pattern mining is proposed to extract product features from pros and cons in a particular type of reviews.

Other work related to mining opinion and blogs are, for example, [10] and [11]. Investigating existing technology and their utility for sentiment classification on web log posts, [11] reports the performance of a Naive Bayes classifier in the prediction of a posting's political category.

II. DATASET

As one of the largest online retailers, amazon.com carries a huge number of items in various categories. The dataset used in this research comes from amazon.com's Books department. Each book title for sale is listed on amazon.com website along with its relevant information, such as, title, authors, formats, editions, average star rating, customer reviews.

A book title may receive a number of reviews. As seen in Fig. 1, a book review has several components, which are, from top, star rating, review's title, date written, reviewer's name, book title, textual content, review's helpfulness poll, abuse report, and review's permanent link. Though not all reviews have exactly the same components as one in Fig. 1, every one of them has the three most relevant to this study, the star rating, the review's title and the review's textual content.

In this research, data of book reviews are collected during the year 2008-2013. All reviews belong to a book written in English language.In a book review, its title and textual content are quite free-form. Therefore, review texts can vary from one to another in format, length, vocabulary, formality, and readability.

© 2015. The authors - Published by Atlantis Press

III. Attribute preparation

A. Part-of-Speech (POS) tagging

A sentence may consist of words with different roles, orparts of speech (POS), such as, nouns, verbs, adjectives, etc. In this research, POS tagging, which is the process of labeling each word with such roles, is performed using Stanford Log-linear Part-Of-Speech Tagger, described in [12] and [13]. The Tagger uses the Penn Treebank tag set [14], which consists of 48 POS tags, with the last 12 being punctuation marks and special symbols. For this research, these 12 tags are combined into a single tag, making our tag set 37 members in total.

B. Attribute Preparation

Given a book review, as shown in Fig. 1, extracted are three items, namely, the number of stars it receives, the review's title, and the review's textual content. The number of stars is an integer of value between 1 through 5, inclusive. Texts in review's title and textual content are assigned POS tags.

For the review's textual content, the number of occurrences of each tag is counted, as well as the percentages calculated.

The review's title is processed slightly differently. Unlike review's textual content, most review's titles are made up of only few words. Instead of looking at the occurrence number and percentage of each tag, which might not be very meaningful, the occurrence itself is examined. In other words, each tag can either appear or not appear in a review's title.

C. Attribute Aggregation

The Penn Treebank tag set, with our slight modification, consists of 37 tags. A book review is considered a short text, as most range around a few hundred words, and very few extend over a thousand words. With these many tags performing on short texts, some tags might have an unusually high occurrence percentage in some reviews and a zero in others. To reduce such fluctuations, a set of aggregated tags are introduced. Each aggregated tag is formed by grouping together similar Penn Treebank tags. All aggregated tags are shown in Table 1.

Similarly with the original tags, each aggregated tag is assessed in terms of its occurrence in review's title and the occurrence number and percentage in review's textual content.

TABLE I. Aggregated tags and their members.

Aggregated Tags	Members	Aggregated Tags	Members
Noun	noun (singular or mass)	Verb	verb (base form)
	noun (plural)		verb (past tense)
	proper noun (singular)		verb (gerund/present participle)
	proper noun (plural)		verb (past participle)
Proper Noun	proper noun (singular)		verb (non-3rd person, singular, present)
	proper noun (plural)		verb (3rd person, singular, present)
Pronoun	personal pronoun	Adjective	adjective
	possessive pronoun		adjective (comparative)
Wh	*wh*-determiner		adjective (superlative)
	wh-pronoun	Adverb	adverb
	possessive *wh*-pronoun		adverb (comparative)
	wh-adverb		adverb (superlative)

D. Summary of Input Attributes

From all the steps for attribute preparation described earlier, Table 2 summarizes all the input attributes used in this study. For convenience in describing this paper, similar attributes are arranged in sets, which are named from A through E. The target attribute is the star rating of a book review.

These attributes will be used in experiments described in the next section.

TABLE II. Summary of input attributes in this study.

Sets	Description	Number of Attributes
A	The percentage of each tag in review's textual content	37
B	The percentage of each aggregated tag in review's textual content	7
C	The number of each aggregated tag in review's textual content	7
D	The appearance of each tag in review's title	37
E	The appearance of each aggregated tag in review's title	7
	TOTAL	95

IV. Experiments

A. Experiment 1: A Neural Network Prediction Model

Our hypothesis is that, parts of speech in a book review can help predict the book's rating. Therefore, our first experiment is to build a prediction model from reviews' parts of speech. Input is the attribute set in Table 2, while output is prediction of a book's rating. Neural network is chosen as the prediction model because it is known to handle continuous and numeric attributes fairly well.

B. Experiment 2: Bayesian network and decision tree classification models

In comparison with Experiment 1, two classification models, namely, Bayesian Network and Decision Tree, become our choices. Unlike Neural network, the two latter models only handle discrete attributes. Our continuous-valued attributes can be discretized, even though they will be rather cardinal than nominal. Nevertheless, it might still be interesting to see how results might come out from these models.

C. Experiment 3: 5-Class and Binary Models

As the two earlier experiments develop, results are not exactly outstanding. The book's rating, which is our target attribute, originally takes on values from 1 through 5. In this experiment, adjustment is made so that its value can only be high or low. In doing so, while it is almost certain to see better classification results, there might be other interesting insights yet to be discovered. Results are compared with those from experiment 2.

D. Experiment 4: Attribute Inclusion-Exclusion

Of attribute sets A through E listed in Table 2, if some should be given less weight than the others, they would be sets C, D and E. As for attributes in set C, which count the number of times an aggregated tag appears in a review's textual content, these numbers could vary quite considerably. Therefore, their ability as a classifier might not be as strong as that of attributes in set A and B.

As for sets D and E, attributes are based on review's title, which mostly contains no more than twenty words. Compared with review's textual content, whose words are much more ample, the review's title might not be as competitive for classification.

For this experiment, attribute sets C and D-E take turn being excluded from input attribute set. The purpose of such exclusion is to simplify the classification model. Results are compared among different exclusions.

E. *Performance Measurement*

In order to measure the performance of each prediction/classification model listed in the experiments above, a 10-fold stratified cross validation is performed. A number of quantities are reported, such as, accuracy (percentage of correctly classified instances), mean absolute error, root mean squared error, average precision and recall.

V. Results and discussion

A. *Result 1: a Neural Network Prediction Model*

In experiment 1, all attributes in Table 2 are used as input to build a Neural Network model that predicts the star rating of a book. As predicted values are numeric, errors are reported as follows. The mean absolute errors for the training and test sets are 0.0857 and 1.315, respectively. The low training set's error becomes almost meaningless when the test set's error turns out to be quite high. Also reported are the root mean squared errors, whose numbers are 0.1048 and 1.7848 for the training and test sets, respectively. The test set's errors are both over 0.5, meaning that, on average, the predictions are off from their correct values.

B. *Result 2: Bayesian network and Decision tree Classification Models*

In experiment 1, in which a Neural Network model is used for prediction of a book rating, results do not prove much promising. Our experiments now shift from prediction to classification. Experiment 2 compares two classification models, Bayesian Network and Decision Tree. Results are shown in Table 3. The accuracy measures the percentage of correctly classified instances. Even though the test set's accuracies, being slightly lower than 50%, do not show an interesting performance, the mean absolute errors reveal something more positive. Those error numbers for all training and test sets are significantly lower than 0.5, meaning that, on average, most instances are either correctly classified or misclassified to an adjacent category. Speaking in terms of our data, this means both classification models estimate the star rating of each book either correctly or off by no more than one star, on average. As for precision and recall, the numbers are just fine but nothing outstanding.

As shown in Table 3, Decision Tree's overall performance is slightly better than Bayesian Network's. In the next experiment, Decision Tree classification models are further investigated.

TABLE III. Classification performance by Bayesian network and decision tree.

	Accuracy	**Mean Absolute Error**	**Precisio n**	**Recall**
Bayesian Network				
Training Set	57.29%	0.2039	0.5163	0.5771
Test Set	49.25%	0.2403	0.4328	0.4962
Decision Tree				
Training Set	85.43%	0.0875	0.8598	0.8590
Test Set	47.24%	0.2182	0.4751	0.4757

C. *Result 3: 5-Class and Binary Models*

Given classification performance in the previous experiment, the Decision Tree model is explored further. In this experiment, the target attribute, which originally takes on values between 1 through 5, is modified to become a binary attribute whose value can be either high or low. Classification performance is compared between the original 5-class model and the new binary model, as shown in Table 4.

It is a delight but no surprise that all the performance numbers look quite satisfactory for both training sets. Accuracy, precision, and recall are all as high as 85-95%, while the mean absolute errors are both lower than 0.09. As for the test set's performance, the accuracy, precision, and recall all improve from the 5-class model to the binary model. Such improvement is expected to happen. However, the mean absolute error does not show similar progress as it increases, though only slightly. Overall, the binary model performs fairly well in classification.

TABLE IV. Classification performance by 5-class model and binary model.

	Accuracy	**Mean Absolute Error**	**Precisio n**	**Recall**
5-class Model				
Training Set	85.43%	0.0875	0.8598	0.8590
Test Set	47.24%	0.2182	0.4751	0.4757
Binary Model				
Training Set	94.97%	0.0892	0.9536	0.9547
Test Set	81.41%	0.2218	0.7968	0.8187

D. *Result 4: Attribute Inclusion-Exclusion*

Given the input attribute sets in Table 2, this experiment compares how different combinations of them may affect the classification performance. Our reasoning is that, some sets might play a stronger role than others, in classifying an instance. For the weaker ones, excluding them might help simplify the classification model. For the reasons described in the previous section, sets A and B are potentially stronger players, so they are always included as inputs. Sets C and D-E look weaker, so they first take turn and then do simultaneously being excluded as input attributes. Results for classification on test sets are shown in Table 5.

From Table 5, A-B-D-E combination performs better than A-B-C-D-E and A-B combination performs better than A-B-C. This seems to suggest that excluding attribute set C might yield better performance. In fact, all performance numbers (i.e., accuracy, mean absolute error, precision and recall) slightly improve with the exclusion. As for exclusion of attribute sets D-E, Table 5 shows that A-B-C combination is better off with D-E, while A-B combination is better off without. In this case, excluding attribute sets D and E does not always improve classification performance. Among the four combinations compared, A-B seems to be the winner, though by not much.

TABLE V. CLASSIFICATION PERFORMANCE, COMPARED AMONG DIFFERENT SETS OF INPUT ATTRIBUTES.

Input Attribute Sets	Accuracy	Mean Absolute Error	Precision	Recall
A-B-C-D-E	81.41 %	0.2218	0.7968	0.8187
A-B-C	80.90 %	0.2281	0.7826	0.8137
A-B-D-E	82.91 %	0.2093	0.8037	0.8339
A-B	83.92 %	0.2031	0.8173	0.8440

VI. CONCLUSIONS AND FUTURE WORK

In this study, the hypothesis is that, given a book review, different parts of speech in its content may have some relation with the book's rating given by the reviewer. Such hypothesis is tested out by first counting different parts of speech appearing in both review's title and textual content. Some parts of speech are combined into aggregated units, of which appearance counts are also determined. The appearance counts and their corresponding percentages are then used as input attributes for training prediction/classification models. These models are explored further in some dimensions. From all experiments performed, there are outcomes which suggest that there exist some relations between a book review's content and its rating, i.e. the former can help predict/classify the latter.

As for future work, in order for such relations to be understood better, individual parts of speech that are more predominant as a classifying factor may be identified and then seek for linguistic explanation.

REFERENCES

[1] Tanawongsuwan, P., "Product review sentiment classification using parts of speech: A case study of textbook reviews," in *Proc. 3rd IEEE International Conference on Computer Science and Information Technology*, Chengdu China, pp. 424-427, 2010.

[2] Pang, B., Lee, L., and Vaithyanathan, S., "Thumbs up? sentiment classification using machine learning techniques," in *Proc. of Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2002.

[3] Gamon, M., Aue, A., Corston-Oliver, S. and Ringger, S., "Pulse: mining customer opinions from free text," *Proc. International Symposium on Intelligent Data Analysis (IDA 2005)*, Madrid, Spain, Sep. 2005.

[4] Cui, H., Mittal, V. and Datar, M., "Comparative experiments on sentiment classification for online product reviews," in *Proc. 21st National Conference on Artificial Intelligence (AAAI 2006)*, Boston, Massachusetts, USA, Jul. 2006.

[5] Dave, K., Lawrence, S. and Pennock, D., "Mining the peanut gallery: opinion extraction and semantic classification of product reviews," in *Proc. 12th Int'l Conf. on World Wide Web (WWW 2003)*, Budapest, Hungary, May 2003.

[6] Turney, P., "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews," in *Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, Philadelphia, Pennsylvania, USA, Jul. 2002.

[7] Tanawongsuwan, P., "Content-based evaluation of textbook reviews," in *Proc. 6th International Joint Conference on Computer Science and Software Engineering*, Phuket Thailand, 2009.

[8] Tanawongsuwan, P., "Part-of-speech approach to evaluation of textbook reviews," in *Proc. International Conference on Industrial and Intelligent Information*, Bangkok, pp. 352-356, 2010.

[9] Liu, B., Hu, M. and Cheng, J.,"Opinion observer: analyzing and comparing opinions on the web," in *Proc. 14th Int'l Conf. on World Wide Web (WWW 2005)*, Chiba, Japan, May 2005.

[10] Attardi, G. and Simi, M., "Blog mining through opinionated words," in *Proc. 15th Text Retrieval Conference (TREC 2006)*, Gaithersburg, Maryland, USA, Nov. 2006.

[11] Durant, K. and Smith, M., "Mining sentiment classification from political web logs," in *Proc. KDD Workshop on Web Mining and Web Usage Analysis (WebKDD 2006) in conjunction with the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

[12] Toutanova, K. and Manning, C. D., "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger," in *Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, pp. 63-70, 2000.

[13] Toutanova, K., Klein, D., Manning, C. and Singer, Y., "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proc. of HLT-NAACL*, pp. 252-259, 2003.

[14] Marcus, M. P., Santorini, B. and Marcinkiewicz, M. A., "Building a large annotated corpus of English: The Penn Treebank," *Computational Linguistics*, vol. 19, no. 2, pp. 313-330, 1993.