Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches

Pimwadee Chaovalit
Department of Information Systems
University of Maryland, Baltimore County
pchao1@umbc.edu

Lina Zhou
Department of Information Systems
University of Maryland, Baltimore County
zhoul@umbc.edu

Abstract

Web content mining is intended to help people discover valuable information from large amount of unstructured data on the web. Movie review mining classifies movie reviews into two polarities: positive and negative. As a type of sentiment-based classification, movie review mining is different from other topic-based classifications. Few empirical studies have been conducted in this domain. This paper investigates movie review mining using two approaches: machine learning and semantic orientation. The approaches are adapted to movie review domain for comparison. The results show that our results are comparable to or even better than previous findings. We also find that movie review mining is a more challenging application than many other types of review mining. The challenges of movie review mining lie in that factual information is always mixed with real-life review data and ironic words are used in writing movie reviews. Future work for improving existing approaches is also suggested.

1. Introduction

The Semantic Web enables rich representation of information on the Web. Before the vision is turned into widely accessible reality; however, we have to deal with an enormous amount of unstructured and/or semi-structured data on the Web. The unstructuredness implies that data are in free format, mostly in text form, which are very difficult to manage. Therefore, there is a great need for a set of techniques to handle such unstructured data effectively. Text mining represents a set of intelligent techniques dealing with a large amount of unstructured text data. In view that unstructured data constitute a large proportion of data on the Web, mining content on the Web is a worthwhile effort since we might uncover invaluable information otherwise found nowhere else in any of the enterprise's databases.

How to mine unstructured data such as feedback surveys, e-mail complaints, bulletin boards, opinions, and product reviews that are available at websites is a challenging issue. In particular, online product reviews are often unstructured, subjective, and hard to digest within short timeframe. Product reviews exist in various forms across different websites [1]. They could appear on commercial product sites with complementary reviews (like Amazon), professional review sites (such as www.dpreview.com, www.imdb.com, www.cnet.com, www.zdnet.com), consumer opinion sites on broad topics and products (such as www.consumerreview.com, www.epinions.com, www.bizrate.com), news or magazines with feature reviews (www.rollingstone.com), or even on bulletin boards and Usenet groups which archive less formal reviews. Reviews for products or services are usually based on opinions expressed in very unstructured format, even though some of the above-mentioned review sites adopt objective measures (such as ratings, stars, scales) to provide quick information to website visitors. Opinion mining is an application domain that web content mining can play a role to fulfill the needs.

The main objective of this work is to classify a large number of opinions using web-mining techniques into bipolar orientation (i.e. either positive or negative opinion). Such kind of classification could help consumers in making their purchasing decisions. Research results along this line can lead to users' reducing the time on reading threads of text and focusing more on analyzing summarized information. Review mining can be potentially applied in constructing information presentation. For example, review classification could be integrated with search engines to provide statistics such as "500 hits found on Paris travel review, 80% of which are positive and 20% are negative" [2]. Such kind of summarization of product reviews would be even more valuable to customers if the summaries were available in various forms on the web, such as review bulletin boards.

The extant literature [1, 3, 4] shows that two types of techniques have been utilized in opinion mining applications: machine learning and semantic orientation. The machine learning approach applied to this problem mostly belongs to supervised classification in general and

text classification techniques in particular for opinion mining. This type of technique tends to be more accurate because each of the classifiers is trained on a collection of representative data known as corpus. Thus, it is called "supervised learning". In contrast, using semantic orientation approach to opinion mining is "unsupervised learning" because it does not require prior training in order to mine the data. Instead, it measures how far a word is inclined towards positive and negative.

Each of the above approaches has its pros and cons. Even though supervised machine learning is likely to provide more accurate classification result than unsupervised semantic orientation, a machine learning model is tuned to the training corpus, and thus needs retraining if it is to be applied elsewhere [2]. It is also subject to over-training and highly dependent upon the quality of training corpus. Thus, the selection of opinion mining techniques tends to be a trade-off between accuracy and generality. To our best knowledge, the two approaches have yet to be compared in the same domain. It is still an open question as to which approach is better for opinion mining.

To address the above question, we adopted both supervised and unsupervised approaches to opinion mining and compare their performances in the same domain. The findings will help us gain insight into the strengths and limitations of machine learning and semantic orientation techniques for opinion mining. In view that movie review mining is a more challenging domain for opinion mining [3], it is chosen as the subject of this study. Empirical evaluation results on movie review mining are very encouraging. The supervised approach achieved 84.49% accuracy in three-fold cross-validation and 66.27% accuracy on hold-out samples. The semantic orientation approach correctly predicted 77% of movie reviews. The result confirms our expectation that a supervised machine learning approach is more accurate but requires a significant amount of time to train the model. Conversely, the semantic orientation approach is slightly less accurate but is more efficient to use in real-time applications. Overall, the results show that it is practically feasible to mine opinions from unstructured data automatically. Automated movie review mining would be a desirable goal that can enhance the currently unstructured web content with semantic web content in the future.

The rest of the paper is organized as follows. We first provided background on opinion mining and approaches to opinion mining, including supervised machine learning and unsupervised semantic orientation in Section 2. Then, we designed methodology for mining movie reviews in Section 3, which is followed by empirical evaluations in Section 4. Next, we discussed the findings of the study, limitations, and challenges of movie review mining. We concluded the paper with future directions.

2. Background

In this section, we reviewed the scope and prior work on opinion mining, and introduce two techniques of interest: machine learning and semantic orientation.

2.1. Opinion mining

Opinion mining aims to mine reviews of various products (i.e. PDAs, cameras, electronics, cars, books, music records, movie reviews, etc.) by classifying them into positive or negative opinions [1, 2, 4]. Moreover, these opinions could be summarized in order to give users statistics information [2]. Morinaga et al. [4] further used classified opinions to analyze product reputations by applying reputation analysis process. Product reputation, which is derived from the reputation analysis, is the higher-level of knowledge than recommendations obtained from opinion mining.

Mining opinions from product reviews on the web is a complex process, which requires more than just text mining techniques. First, data of product reviews are to be crawled from websites, in which web spiders or search engines play an important role. Second, data is usually neither clean nor in the desired format, thus several techniques are utilized to automate the data preparation process. There are various options for data preparation, depending on the form of the raw data and the target format. For example, a set of collected web pages is very likely to contain other types of information other than reviews. Separating reviews from non-reviews data is an important task. A technique that can support this task is called objectivity classification [5] or subjectivity analysis [6]. Subjectivity analysis is the process of distinguishing subjective sentences expressing opinions and evaluations from objective sentences expressing factual information [6].

Automated opinion review mining is beneficial to both consumers and retailers/manufacturers. Consumers would know which products to buy or not to buy and retailers/manufacturers would know their competitors' performances. Given the advance in machine learning and computing resources, opinions and reviews on several genres of products and services can be semi-automatically classified into *recommended* or *not recommended*. Examples of past work include mining reviews of automobiles, banks, movies, travel destinations [3], electronics [1, 4] and mobile devices [4]. Potential applications include extracting opinions or reviews from discussion forums efficiently, and integrating automatic review mining with search engines to provide quick statistics of search results.

2.2. Movie review mining

Special challenges are associated with movie review mining. As it has been pointed out elsewhere [3], movie review mining is very domain specific and word

semantics in a particular review could contradict with overall semantic direction (good or bad) of that review. For example, an "unpredictable" camera gives negative meaning to that camera model, whereas a movie with "unpredictable" plot sounds positive to moviegoers. Therefore, we need to train the machine learning classifiers with movie review dataset as well as adapt the semantic orientation approach to movie review domain.

2.3. Machine learning vs. semantic orientation

Some research studies have employed a supervised training approach in opinion mining [1, 4]. It starts with collecting training dataset from certain websites. An ideal training sample should be representative in order to get good accuracy of prediction. These data, if not categorized properly, need to be manually labeled by human effort, so that the opinions are associated with objective ratings. Dave et al. [1] found reviews from websites (C|net and Amazon) that provide binary ratings along with opinions as a perfect training set. Such a set of data is called a corpus.

The next step is to train a classifier on the corpus. Once a supervised classification technique is selected, an important decision to make is feature selection. In text classification, features denote properties of textual data which are measured to classify the text, such as bag-of-words, n-grams (e.g. unigram, bi-grams, tri-grams), word position, header information, and ordered word list [7]. They can tell us how documents are represented. After appropriate features have been selected, the classifier is trained on the training dataset. The training process is usually an iterative process in order to produce a better model.

The performance of the classifier trained on the training data is finally evaluated on the test dataset based on chosen criteria.

The above processes are usually repeated in multiple iterations if the testing result is not satisfactory. In the following iterations, model parameters are adjusted according to the difference between predicted class and actual class labels. For example, when mining movie reviews, we may apply word stemming, remove stop words, filter input features, and so on, in an effort to produce a better model.

Another approach uses semantic orientation of a word in order to mine opinions. Semantic orientation from a word could be positive (i.e. praise) or negative (i.e. criticism). It indicates the direction that the word is in relative to the average [8]. There are several dimensions we could consider regarding semantic orientation: direction and intensity [2]. Direction indicates whether a word has positive or negative meaning. In opinion mining application, a word could indicate praise or criticism. Intensity designates how strong the word is. In opinion mining, a review could be found negatively milder than some other negative reviews. Another related work using semantic orientation included conjunctive words (i.e. and, but) to improve training a supervised learning algorithm [8], because we can understand the tone of the sentence from its conjunctions. "And" indicates that both adjectives have the same semantic orientation, whereas "but" indicates adjectives with opposite semantic orientations.

Turney's study [3] on review classification using word semantic orientation consisted of three steps. First, a part-of-speech tagger extracted two-word phrases containing at least one adjective or one adverb from the review. The adjective or adverb carries semantic orientation, while the other word in the phrase provides context. Second, a technique called SO-PMI (Semantic Orientation using Pointwise Mutual Information) was used to calculate semantic orientation for the selected phrases. The extracted phrases will be judged in terms of how inclined they are towards positive or negative edges. The overall semantic orientation of each review is determined by averaging the SO-PMI values of all the phrases in it. Finally, the entire piece of review is identified as either positive or negative by comparing the overall semantic orientation and a baseline value (Zero was the baseline used in Turney's experiment). The best result achieved by Turney's experiment was 84% accuracy for automobile reviews and 66% for movie reviews.

In sum, for a semantic orientation approach, good associations and bad associations account for positivity and negativity; whereas for machine learning technique, document features determine whether a review belongs to positive or negative classes.

3. Methodology

3.1. Machine learning approach

3.1.1. Corpus. We need a collection of movie reviews that include both positive reviews and negative reviews. Good corpus resources should have good review quality, available metadata, easy spidering, and reasonably large number of reviews and products [1].

We select a ready-to-use and clean dataset in movie reviews domain (http://www.cs.cornell.edu/people/pabo/movie-review-data/) as the corpus. The data have been used in Pang et al.'s experiment on classifying movie reviews [9]. There are 1,400 text files in total with 700 labeled as positive reviews and the rest 700 labeled as negative reviews. They were originally collected from IMDB (Internet Movie Database) archive of review newsgroups at http://reviews.imdb.com/Reviews. The ratings were removed. The rating decision was made in order to transform a 4-star or 5-star rating-system reviews into positive and negative reviews. Moreover, the data were examined manually to ensure quality. A few non-English and incomplete reviews were removed. Misclassified reviews based on sentimental judgment were also corrected.

3.1.2. N-gram Classifiers. In the light that n-gram models provide one of the best performance in text classification in general, we selected n-gram models as supervised approach, which represent text documents by word tuples [7], We employed a shareware Rubryx version 2.0 (http://www.sowsoft.com/rubryx) as our classification tool. The tool is implemented with classification algorithms based on n-gram (unigram, bi-grams, and tri-grams) features. Several options are given to adapt classification models, such as adding stop-word lists. The stop-word lists can be in unigram, bi-grams, and tri-grams forms. Classification models can also be adapted by incorporating domain-specific dictionaries into feature extract from documents. We did not make use of this option due to the lack of dictionary.

Classifying text data confronted us with a wide range of languages used in expressing opinions to product review. It is risky to over-filter data, for we may remove important information along with unwanted data. Therefore, we captured a large number of features at the beginning [1], and then tried multiple sets of features to select the one that best suits our problems.

3.2. Semantic Orientation (SO) approach

SO approach to review mining is a multi-step process. Minipar (http://www.cs.ualberta.ca/~lindek/minipar.htm) was used to tag and parse the review documents at first. Based on parts-of-speech in the parsed output, two-word phrases were then selectively extracted. Only two-word phrases conforming to certain patterns were extracted for further processing. We adopted phrase patterns from Turney's study [3]. Adjective or adverb in the patterns provides subjectivity, while the other word provides context. The following table summarizes five patterns used in the extraction of phrases.

Table 1: Two-word phrase patterns

	First word	Second word
a)	Adjective	Noun
b)	Adverb	Adjective
c)	Adjective	Adjective
d)	Noun	Adjective
e)	Adverb	Verb

The next step was to determine the semantic orientation of a phrase' SO(*phrase*) according to Formula (1) [3]. *hits*(·) denotes the number of pages returned for a query consisting of phrase · from a search engine. For example, *hits*('poor') represents the number of pages returned for a search query '*poor*'. When there are both *phrase* and 'excellent' (or 'poor') connected by NEAR operator in the parameter of *hits* function, it defines the similarity between *phrase* and 'excellent' (or 'poor'). In other words, the similarities were measured with co-occurrences of the phrases and 'excellent' (or 'poor'). We automated this step by sending search queries to Google search engine and crawling Web pages for the information.

$$SO(phrase) = \log_2 \left(\frac{hits(phrase\ NEAR\ \text{“excellent”})\ hits(\text{“poor”})}{hits(phrase\ NEAR\ \text{“poor”})\ hits(\text{“excellent”})} \right) \quad (1)$$

A phrase's semantic orientation would be positive if it is associated more strongly with "excellent" than "poor" and would be negative if it is associated more strongly with "poor" than "excellent".

Finally, a review's semantic orientation was calculated by averaging the SO values of all the extracted phrases in it. The movie is recommended to watch if its average semantic orientation exceeds a threshold and is not recommended if otherwise.

3.3. Test data

In an attempt to reduce bias inherent with the supervised learning approach, another movie review site www.moviejustice.com was used as testing data for this experiment. We collected 384 reviews of movies from its review section "Movie Vault". Each review was written by one reviewer and accompanied with MJ Rating ranging from A+, A, A-, B+, … to F. Those without a particular rating were excluded from the dataset later. Therefore, the number of test dataset is 378 reviews. The statistics of test dataset is shown in Table 2.

Table 2: Distribution of test data by class

Ratings	Number of reviews
A+	74
A	86
A-	29
B+	38
B	41
B-	17
C+	19
C	18
C-	9
D+	8
D	14
D-	10
F	15
Total	378
No rating (excluded)	6

3.4. Rating decisions

As shown in Table 2, a movie was rated as one of the five categories. If we are to group them into *recommended* and *not recommended*, we should set a dividing line to separate the data. It is obvious that the data are positively skewed. Therefore, it is a subjective

decision as to whether a particular rating should fall under positive or negative categories, especially for ratings with neutral tone (such as B and C).

We applied the rating decision of Pang et al. in this study. However, the ratings of original IMDB review dataset collected by Pang et al. are not complete [9]. Some ratings are in different forms and some are even missing, making them incomparable across different rating systems of different reviewers. With a five-star rating system, reviews given four stars and up were considered positive while reviews given two stars and below were negative. With four-star rating system, reviews given three stars and up were positive while reviews given one star and below were considered negative. Reviews that fall in neutral range are discarded. In the supervised classification experiment, Dave et al. [1] proved improvements in accuracy after removing irrelevant or indeterminate cases from the dataset. Therefore, we decided to ignore the neutral ratings, and group ratings into two main categories as follows:

The rating system of movie justice was comparable to a five-star rating system (A, B, C, D, and F comparing to 5-, 4-, 3-, 2-, and 1-star) used by Pang et al. Accordingly, movies rated in A and B ranges received positive reviews and those rated in D and F ranges received negative reviews, but movie reviews in C range were discarded. Finally, we obtained 285 positive opinions and 47 negative opinions.

3.5. Evaluation techniques

As with other classification problems, classification accuracy was selected as metrics in this study. The classified reviews will be compared with the actual class they fall in. Accuracy is measured as the ratio between the number of reviews that has been classified correctly to the total number of reviews being classified.

In addition, we also evaluate the mining performance from the information retrieval perspective. *Recall* was measured by the ratio of the number of reviews correctly classified into a category to the total number of reviews belonging to that category. This measurement indicates the ability to recall items in the category. *Precision* was measured as the ratio of the number of reviews classified correctly to the total number of reviews in that category. In conclusion, we used three types of measures to evaluate classifiers, including accuracy, precision, and recall.

4. Experiment result and analyses

4.1. Supervised machine learning approach

One of the problems we found during the preliminary experiment is the stop word. It was not surprising to see that stop words make frequent appearances in *n*-gram features and dominate the classifier models. Careful selection of stop words is crucial, so that researchers do not exclude important keywords or use their own biases in judging them.

Therefore, the first trial of classification involves removing stop words from the *n*-gram features. A stop-word list was obtained from http://snowball.tartarus.org/english/stop.txt and was included in the Rubryx software. We also removed bi-grams and tri-grams that all of the single-word components are stop words, but not for bi-grams and tri-grams that at least one single-word components are not listed as stop words. For example, "*on the*" and "*in the*" will be considered as bi-gram stop words, whereas "*carry on*" will be preserved as a legitimate feature.

Rubryx software creators recommend that the number of training documents need not to be large. Five or six representative files for each class would be enough. Therefore, we started from the training size of 5 per category. The first trial of classification without extensive preprocessing (only removing stop words) gave us a very poor result. The precision of *recommended* was 87.92% (131 out of 149 reviews were classified correctly as positive). The precision of *not recommended* was 14.70% (20 out of 136 reviews were classified correctly as negative). It could happen that a large number of negative reviews were classified into positive category because the data were positively skewed. Therefore, we also calculated the recall rates. The recall for negative reviews was 42.55% (20 out of 47 reviews) and that for positive reviews was 45.97% (131 out of 285 reviews). The performance on neither positive nor negative categories was good. The total accuracy of this trial was 45.48% (151 out of 332 reviews are classified correctly). The miss rate was actually larger than the hit rate and it was even less satisfactory when compared to the probability that an unbiased classifier randomly classifies documents (50%). Table 3 cross-tabulates the number of reviews in each category and their predicted categories.

Table 3: Confusion matrix of classifier-first trial

	Actual class		
Predicted class	Positive Reviews	Negative Reviews	Sum
Recommended	**131**	18	149
Not recommended	116	**20**	136
Unclassified	38	9	47
Sum	285	47	332

(Training size = 5 documents for both categories)

The possible explanation for the poor classification result might be the nature of movie review domain. Each review in the training corpus contains a large amount of text, with regularly about 600 words per review. Reviewers used a wide variety of words in their reviews, which easily led to sparsity in bag-of-words features. Rubryx software creators recommend that the number of training documents need not be large. The result of our first trial suggested that our five documents randomly

selected from the corpus might not be representative. Therefore, larger size of training data was considered in the next trials. What we would like to see later on was, whether more carefully selected training files would improve the classification.

We can also derive from the first trial of experiment that the selection of training documents is tricky since the documents should contain representative content for the particular category. Developing few training documents with representative content would be very beneficial and expectantly improving classifier accuracy tremendously. Incorporating a set of dictionary for movie review domain is another option to tune-up the classifier when we do not have very carefully selected training samples, for the dictionaries would limit words for feature selection.

In the next trials, we increased the size of training corpus to 10 for each category. The result was shown in Table 4. It suggested that corpus with better quality than the previous one should be utilized. When we examined the training corpus, we found that the reviews contain not only reviewers' opinions but also fact sentences. Some reviewers preferred to summarize the story along with their comments. In the light of the subjectivity of training corpus, a review file ranging from 300 to 800 words can be reduced to only those sentences that really expressed subjectivity.

As we mentioned in section 2.1, subjectivity analysis could have been our option to preprocess training corpus by separating subjective review sentences from factual non-review sentences. Wiebe [6] considered a sentence to be subjective if human readers perceive the significant expression of subjectivity or judgments, and be objective otherwise. However, given the limited timeframe for the project, the non-automatic process of subjectivity analysis might not be feasible at present.

Table 4: Classifications results of a supervised machine learning approach

	Same Corpus	Training size	Total accuracy
Preliminary Trial	N	5/category	45.48%
Second Trial	N	10/category	74.40%
Third Trial	N	15/category	18.67%
Large training dataset	N	200/ category	66.27%
3-fold validation	Y	221, 221, and 222 for 3 iterations respectively	***85.54%**

Table 4 showed that the classification results from the first three trials were not consistent and also quite poor (only 45.48%, 74.40%, and 18.67% respectively), we suspected that the training size might be too small and thus considered training the classifier with a bigger training dataset. This time the training size is 200 documents per category and the derived classification accuracy is 66.27%, with 70.88% positive recall and 38.30% negative recall.

In order to test the performance of classifiers on different movie reviews that are from the same web sites, we applied k-fold cross validation. Since our dataset were not very large, we divided them into three sub-sets. In the first iteration, the first subset of data was held out and the second and third subsets were used for training. Upon the training is finished, the first dataset that was previously held out was tested for performance. In the second iteration, the classifier was trained on the first and third subsets of data and tested on the second subset of data. In the last iteration, the classifier was trained in the first and second subsets of data and tested on the last subset of data. From our experiment, 3-fold cross validation derived an average accuracy of 85.54%, which was quite good. Still, we found the same problem of poor recall rate. Although the recall rate for positive reviews were satisfactory (98.95%, 100%, and 98.95% respectively for 3 iterations), the classifier did not perform well in recognizing negative reviews (6.25%, 0%, and 0% recall rate). The poor recall rate might be explained by our positively skewed dataset that caused the truth bias.

Furthermore, the accuracy may be hampered when the test dataset comes from a different source, jeopardized when being tested across different dataset.

4.2. Unsupervised learning approach

Semantic orientation approach requires extracting phrases containing adjectives or adverbs from review data. Following Turney's approach [3], five patterns of phrases will be extracted to find their semantic orientation values. Since we utilized different POS-tagger from Turney's experiment, the outputs of tags are different. Before extracting phrases, we created a mapping between Brill's POS-tagger and the tag set of Minipar to make sure that their POS-tags are, though not equivalent, comparable. Based on Minipar's output, a post-processor was developed to automatically extract phrases from the tagged text and generate Google search queries. Another application was developed to calculate hits returns of phrases, as shown in formula (1), by interfacing with the Google search engine. To reduce noise introduced by a POS tagger, we manually went through the extracted phrases to clean up misrecognized ones. Finally, the semantic orientation of a review is derived by comparing the SO value of the review with a pre-specified threshold.

Our first study of semantic orientation approach set zero as a threshold to separate positive and negative meanings according to the previous study [3]. The result of the trial on a small sample of movie review data was extremely unsatisfying. Data, both positive and negative reviews, seem to suspiciously lean towards the negative end. Sometimes, apparently positive phrases (by the authors' judgment) received very negative scores. For example, the phrase "beautiful day" (adjective noun)

sounds positive but got very negative ranking, almost as negative as the phrase "main failure". Table 5 showed the semantic orientation (SO) values of different positive and negative phrases, regardless of context. The first two rows indicate phrases with positive meanings but received negative SO. The middle two rows indicate negative phrases with seemingly suitable negative SO. The last two rows indicate positive phrases with seemingly suitable positive SO. In conclusion, phrases were categorized with negative bias if we use the baseline from Turney's study [3]. It is possible that during the time of this study, factors in calculating SO was different from the time of Turney's study, thereby leading to very different results. It may also be that different movie review source have different text genre.

Table 5: The example of phrases with their semantic orientation

Phrase	Hits (phrase NEAR "excellent")	Hits (phrase NEAR "poor")	SO
Beautiful day	28,143	33,453	-0.7816
Decent film	500	468	-0.4440
Main failure	95	117	-0.8399
Cheap looking	986	1,002	-0.5627
Special effects	131,415	63,565	0.5084
Handedly inspiring	5	1	1.7897

Therefore, we decided to setup our own baseline to separate positive SO from negative SO. We calculated SO values for six related phrases with review mining domain. The six phrases were selected by us using heuristic approach to overcome the misclassification problem. They were chosen based on the observation that people use them to judge the quality of movies in movie reviews. Theses phrases include "thumbs up", "thumbs down", "positive", "negative", "success", and "failure". Three of them convey positive meaning and the rest convey negative meaning.

Table 6: Calculation of SO baseline adjustment

Phrase	Hits (phrase NEAR "excellent")	Hits (phrase NEAR "poor")	SO
Thumbs up	47,300	38,700	-0.1179
Thumbs down	9,180	9,520	-0.4598
Positive	1,290,000	1,450,000	-0.5760
Negative	917,000	1,290,000	-0.8998
Success	1,430,000	1,440,000	-0.3700
Failure	829,000	1,250,000	-0.9999
Average SO	(our baseline)		**-0.57**

We calculated semantic orientation (SO) of phrases in the same way as we did with the extracted phrases from movie reviews. The average of their SO values are shown in table 6. We take the average SO as our new baseline to classify review into positive or negative. Therefore, a review will be classified as positive if its SO is more than –0.57 and as negative otherwise. The result after adjusting the baseline is shown in Table 7.

Table 7: Classification result of the Semantic Orientation approach (after baseline adjustment)

	Actual class		
Predicted class	Positive Reviews	Negative Reviews	Sum
Recommended	**67**	4	71
Not recommended	19	**10**	29
Sum	86	14	100

The accuracy of mining 100 reviews from using semantic orientation approach was 77% (77 out of 100 are classified correctly), which was quite good. The recall rate for positive reviews was 77.91%, and that for negative reviews was 71.43%.

5. Discussion

5.1. Summary

We obtained results that are comparable and even better than those from previous studies for both approaches to movie review mining. We improved the semantic orientation approach by adapting the threshold to movie review domain and automated the time-consuming process of collecting data from the Web.

Pang et al. mined movie reviews using various machine-learning techniques to examine whether it will be as effective as other classification problems for sentiment classification like movie review mining [9]. They obtained the best classification accuracies ranging from 77.4% to 82.9% by varying input features (i.e. unigrams, bigrams, unigrams + bigrams). Our results were 85.54% for 3-fold cross validation and 66.27% when tested on the test dataset.

As for semantic orientation approach, Turney obtained 65.83% accuracy in mining 120 movie reviews from epinions website [3]. We obtained 77% classification accuracy on 100 movie reviews from Movie Vault after adjusting the dividing baseline.

The result confirmed our expectation that the machine learning approach is more accurate but requires a significant amount of time to train the model. In comparison, the semantic orientation approach is slightly less accurate but is more efficient to use in real-time applications. The performance of semantic orientation also relies on the performance of the underlying POS tagger. Overall, they show that it is feasible to mine opinions from unstructured data automatically. Even though we did not mine movie reviews on the fly in this experiment, there is a high chance that this semantic orientation approach could be extended towards the Semantic Web in the future. The combination of Google

search queries with manual processing indicates that we took a step toward the ultimate goal of automatic semantic understanding of web content.

There are some difficulties inherited in movie review mining. For example, some factual information embedded in reviews skewed the semantic orientation of dataset. Frequently, good movies contain violent scenes and do not have happy endings. The fact that they convey stories of tragedy was picked up by both machine learning techniques and semantic orientation techniques. Thus, even though the movie itself is of high quality, it could be misclassified easily.

The second example from Table 8 provides a glimpse of the writing style of movie reviews. Sometimes, reviewers expressed ironic tone using sarcastic words. Those reviews were subject to misinterpretation by semantic orientation approach, which focuses on phrases' semantic orientation.

Table 8: Examples of problems causing misclassification

First Example	*Embedded factual information*
Sample Phrase:	Horrific sequence [adjective noun]
SO of Sample Phrase:	-1.2547 (negative)
Context of Sample Phrase:	In my opinion, in what is probably the films most **horrific sequence**, Francis travels to a nearby insane asylum to check if anyone by the name of Caligari has been there or escaped.
Author's rating:	A+ (positive)
Movie:	Cabinet of Dr. Caligari
Second Example	*Sarcastic review writing*
Sample Phrase:	Terrifically written [adverb adjective]
SO of Sample Phrase:	0.0455 (positive)
Context of Sample Phrase:	As with most highly depressing movies, the story in About Schmidt dragged very slowly for the first 100% of it (the last 0% was actually very well paced and **terrifically written**).
Author's rating:	D- (negative)
Movie:	About Schmidt

To overcome these difficulties, additional techniques should be involved in the movie review mining using semantic orientation. For reviews with factual information mixed with actual opinions, subjectivity analysis [6] mentioned before could be an alternative to solve this problem. However, it is difficult to address the problem reflected in the second example, which may need more complex knowledge from natural language processing to detect a sarcastic review style.

5.2. Limitations

The experiment of movie review mining depends largely on the preprocessing steps. Factual information that is mixed with actual reviews creates problems to both machine learning and semantic orientation classifications.

The performance of machine learning approaches depends largely on the careful feature selection, for *n*-grams features from movie reviews are sparse. It is clear that simply applying *n*-grams to the classification technique would hardly yield a promising result in movie review mining, unless data is preprocessed to reduce noise. We could enhance review classification by more carefully selecting features that conform to certain patterns as in the semantic orientation approach. Applying POS tagger to facilitate better feature selection could be another option. Consequently, we can take advantage of supervised learning method with text document features for movie review mining.

We should take caution in generalizing the result of semantic orientation approach obtained from a small sample size of data. There were some arbitrary parameters in the method that could change the results. Those parameters include the words selected to compare with phrases (in this case, words are "excellent" and "poor") and the phrase patterns of POS tags, as illustrated in Table 1. This study suggested that human language is very subtle and some meanings conveyed were not captured by the existing patterns. As previously discussed, good movies containing violent or unhappy scenes were often recognized incorrectly. If we could deal with mixed factual information in the reviews and sarcastic style of review writing, the classification result could be improved.

5.3. Future work

We should continue searching for the best features for movie review classification using machine learning approaches. First, we may use TFIDF weighting to reduce the number of features and to solve sparse features problem. Second, specific lexicon or dictionary for movie review domain could be employed to limit the words in classification. The lexicon could support removing factual information that is not relevant to the opinions themselves. Third, a POS tagger can be applied to limit features to words from certain categories such as objectives and adverbs. Fourth, we may improve the performance of classification by developing representative training documents.

For the semantic orientation approach, selecting words other than "excellent" and "poor" in order to compare their similarities with extracted phrases is an alternative to explore. There might be some other words that can better represent polarities for movie review mining and could produce more realistic results for semantic orientation. In addition, certain patterns of two-word phrases could be revisited. During the experiment, we noticed that some other patterns could be used to represent the tone of reviews. The effectiveness of those phrases needs to be verified empirically in future.

As we have discussed, factual information residing in movie reviews distorted the classification results. It would be great if we could employ effective preprocessing steps such as subjectivity analysis [6], which can not only help improve the quality of training corpus but also help feature selection for classification. With this step, we would be able to gain more quality from the training corpus in case of machine learning, and gain smaller and better review documents in case of semantic orientation.

5.4. Conclusion

Movie review mining is a challenging sentimental classification problem. Not only does it deal with classification of personal opinions, but diverse opinions from product reviews as well. Due to the sparsity of words in movie reviews, it is difficult for supervised learning approach to use bag-of-words features. Pang et al.'s experiment also confirmed such difficulty in using machine learning approach to classify movie reviews [9]. Moreover, some parts of a review may not express opinions. Some reviewers prefer to describe factual background information about a movie before expressing their opinions, which can be considered as noise to the classification. Lastly, movie review mining is a very challenging issue for semantic orientation techniques. The findings of this study not only advance the research on movie review mining, but also contribute to other text classification problems such as separate "flames" messages in bulletin boards as mentioned in [3, 9].

Acknowledgement

The authors would like to thank Li Ding for his help in developing tools for gathering response to search queries from a search engine.

References

[1] Kushal Dave, Steve Lawrence, and David M. Pennock, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," presented at the 12th international conference on World Wide Web, Budapest, Hungary, 2003.

[2] Peter D. Turney and Michael L. Littman, "Measuring Praise and Criticism: Inference of Semantic Orientation from Association," *ACM Transactions on Information Systems*, vol. 21, pp. 315-346, 2003.

[3] Peter D. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," presented at the Association for Computational Linguistics 40th Anniversary Meeting, New Brunswick, N.J., 2002.

[4] Satoshi Morinaga, Kenji Yamanishi, Kenji Tateishi, and Toshikazu Fukushima, "Mining Product Reputations on the web," presented at the 8th ACM SIGKDD international conference on Knowledge discovery and data mining, Edmonton, Alberta, Canada, 2002.

[5] Aidan Finn, Nicholas Kushmerick, and Barry Smyth, "Genre classification and domain transfer for information filtering," presented at the 24th European Colloquium on Information Retrieval Research (ECIR'02), Glasgow, UK, 2002.

[6] Janyce M. Wiebe, "Learning Subjective Adjectives from Corpora," presented at the 17th National Conference on Artificial Intelligence, Menlo Park, California, 2000.

[7] Dunja Mladenic, "Text-Learning and Related Intelligent Agents: A Survey," *Intelligent Information Retrieval*, vol. IEEE Intelligent Systems, pp. 44-54, 1999.

[8] V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," presented at Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the ACL, 1997.

[9] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," presented at the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP'2002), 2002