Chapter X

Opinion Polarity Identification of Movie Reviews

Franco Salvetti and **Christoph Reichenbach**
University of Colorado at Boulder
[1]Dept. of Computer Science, University of Colorado at Boulder,
Campus Box 430
Boulder, CO 80309-430, U.S.A.
Email: {franco.salvetti, christoph.reichenbach}@colorado.edu

Stephen Lewis
University of Colorado at Boulder
[2]Dept. of Linguistics, University of Colorado at Boulder,
Campus Box 295
Boulder, CO 80309-295, U.S.A.
Email: stephen.lewis@colorado.edu

Abstract

One approach to the assessment of overall opinion polarity (OvOP) of reviews, a concept defined in this paper, is the use of supervised machine learning mechanisms. In this paper, the impact of lexical feature selection and feature generalization, applied to reviews, on the precision of two probabilistic classifiers (Naïve Bayes and Markov Model) with respect to OvOP identification is observed. Feature generalization based on hypernymy as provided by WordNet, and feature selection based on part-of-speech (POS) tags are evaluated. A ranking criterion is introduced, based on a function of the probability of having positive or negative polarity, which makes it possible to achieve 100% precision with 10% recall. Movie reviews are used for training and testing the probabilistic classifiers, which achieve 80% precision.

Keywords: opinion polarity, sentiment identification, synonymy feature generalization, hypernymy feature generalization, POS feature selection, probabilistic classification.

1 Introduction

The dramatic increase in use of the Internet as a means of communication has been accompanied by an increase in freely available online reviews of products and services. Although such reviews

are a valuable resource for customers who want to make well-informed shopping decisions, their abundance and the fact that they are mixed in terms of positive and negative overall opinion polarity are often obstacles. For instance, a customer who is already interested in a certain product may want to read some negative reviews just to pinpoint possible drawbacks, but may have no interest in spending time reading positive reviews. In contrast, customers interested in watching a good movie may want to read reviews that express a positive overall opinion polarity.

The overall opinion polarity of a review, with values expressed as positive or negative, can be represented through the classification that the author of a review would assign to it. Such a classification is here defined as the overall opinion polarity (OvOP) of a review, or simply, the polarity. The process of identifying the OvOP of a review will be referred to as Overall Opinion Polarity Identification (OvOPI).

A system that is capable of labelling a review with its polarity is valuable for at least two reasons. First, it allows the reader interested exclusively in positive (or negative) reviews to save time by reducing the number of reviews to be read. Second, since it is not uncommon for a review that starts with positive polarity to turn out to be negative, or vice versa, it avoids the risk of a reader erroneously discarding a review just because it appears at first to have the wrong polarity.

In this paper we frame a solution to OvOPI based on a supervised machine learning approach. In such a framework we observe the effects of lexical feature selection and generalization, applied to reviews, on the precision of two probabilistic classifiers. Feature generalization is based on hypernymy as provided by WordNet (Fellbaum, 1998), and feature selection is based on part-of-speech (POS) tags.

The results obtained by experiments based on movie reviews revealed that feature generalization based on synonymy and hypernymy produces less improvement than feature selection based on POS, and that for neither is there evidence of significantly improved performance over the system without neither such a selection nor such a generalization, although the overall performance of our system is comparable to that of systems in current research, achieving an precision of 80%.

In the domain of OvOPI of reviews it is often acceptable to sacrifice recall for precision. Here we also present a system whereby the reviews are ranked based on a function of the probability of being positive/negative. This ranking method achieves 100% precision when we accept a recall of 10%. This result is particularly interesting for applications that rely on web data, because the customer is not always interested in having all the possible reviews, but many times is interested in having just a few positive and a few negative. From this perspective precision is more important than recall.

2 Related Research

Research has demonstrated that there is a strong positive correlation between the presence of adjectives in a sentence and the presence of opinion (Wiebe, *et al.*, 1999). Hatzivassiloglou, *et al.*, (1997) combined a log-linear probabilistic model that examined the conjunctions between adjectives ("and", "but", "or") with a clustering algorithm that grouped the adjectives into two sets which were then labelled positive and negative. Their model predicted whether adjectives carried positive or negative polarity with 82% precision. However, because the model was unsupervised it required an immense, 21 million word corpus to function.

Turney (2002) extracted n-grams based on adjectives. In order to determine if an adjective had a positive/negative polarity he used AltaVista and its function, NEAR. He combined the number of co-occurrences of the adjective under investigation near the adjective "excellent" and near the adjective "poor", thinking that high occurrence near "poor" implies negative polarity and high occurrence near "excellent" implies positive polarity. He achieved an average of 74% precision in OvOPI across all domains. The performance on movie reviews, however, was especially poor at only 65.8%, indicating that OvOPI for movie reviews is a more difficult task than for other product reviews.

Pang, *et al*., (2002) note that the task of polarity classification is not the same as that of topic classification; they point out that topic classification can often be performed by keyword identification, whereas sentiments tend to be expressed in more subtle ways. They applied Naïve Bayes, Maximum Entropy and Support Vector Machine classification techniques to the identification of the polarity of movie reviews. They reported that the Naïve Bayes method achieved 77.3% precision using bigrams. Their best results came using unigrams, calculated by the Support Vector Machine at 82.9% precision. Maximum Entropy performed best using both unigrams and bigrams at 80.8% precision, and Naïve Bayes performed best at 81.5% using unigrams with POS tags.

3 Probabilistic Approaches to Polarity Identification

There are many possible approaches to identifying the actual polarity of a document. Our analysis applies probabilistic methods, namely supervised machine learning, to identify the likelihood of reviews having "positive" or "negative" polarity using previously hand-classified training data. These methods are fairly standard and well understood; we list them below for the sake of completeness.

3.1 Naïve Bayes Classifier

The use of Naïve Bayes classifiers (Duda, *et al*., 1973) is a well-known supervised machine learning technique. In this paper the "features" used to develop Naïve Bayes are referred to as "attributes" to avoid confusion with text "features".

In our approach, all word/POS-tag pairs that appear in the training data are collected and used as attributes. The formula of our Naïve Bayes classifier is defined as:

$$P(c|rv) = \prod_{w \in W} P(app_w|c) \cdot \prod_{w \notin W} P(\neg app_w|c)$$

$$\hat{c} = \arg\max P(c|rv)$$

where:

- *rv* is the review under consideration,
- *w* is a word/POS-tag pair that appears in the given document,
- $P(app_w|class)$ is the probability that a word/POS-tag pair appears in a document of the given class in training data, and

- $\hat{c}$ is an estimated class.

One interesting aspect of this particular application of Naïve Bayes is that most attributes do not appear in a test review, which means most factors in the product probability represent what is not written in a review. This is one major difference from the Markov Model classifier described in the next section.

3.2 Classifier Based on Markov Model

Because the Naïve Bayes classifier defined in the previous section builds probabilistic models based on individual occurrences of words, it is provided with relatively little information regarding the phrasal structure. Markov Model (Jurafsky, *et al.*, 2000) is a widely used probabilistic model that captures connectivity among words.

This Markov Model classifier develops two language models, one on positive reviews and the other on negative reviews. Any given unseen review *rv* is classified by computing the probability P(+|*rv*) of this review having been generated with the language model for positive reviews, and a corresponding probability P(-|*rv*) for the language model for negative reviews. If P(+|*rv*) > P(-|*rv*), we consider *rv* to have been classified as a positive review, and classified as a negative review if P(+|*rv*) < P(-|*rv*) (we did not observe reviews with equal probabilities arising in our tests). The following formula is used to compute the probability that a document could be generated with each language model.

$$P(rv) = \prod_{sn \in Reviews} P(sn)$$

$$P(sn) = P(w|< s >) \cdot \left(\prod_i P(w_{i+1}|w_i) \right) \cdot P(< / s >)$$

where:

- *rv* is the review under consideration,
- *sn* is a sentence,
- <*s*> is the start of a sentence, and
- </s> is the end of a sentence.

4 Features for Analysis

Statistical analysis depends on a sequence of tokens that it uses as characteristic features of the objects which it attempts to analyze; the only necessary property of these features is that it must be possible to identify whether two features are equal.

The most straightforward way of dealing with the information found in reviews would be to use individual words from the review data as tokens. However, using just words discards semantic information about the remainder of the sentence; as such, it may be desirable first to perform some sort of semantic analysis to enrich the tokens with useful information, or even discard misleading or irrelevant information, in order to increase precision.

The following are three possible approaches to this kind of pre-processing:

- Leave the data as is; each word is represented by itself.
- Tag data as parts of speech (POS); each word is enriched by a POS tag, as determined by a standard tagging technique, such as Brill Tagger (Brill, 1995).
- Tag as POS and parse using, for instance, Penn Treebank (Marcus, *et al*., 1994).

The third approach had severe performance issues during our early experiments, raising conceptual questions of how such data would be incorporated into a statistical analysis, whereas concentrating on POS-tagged data (sentences consisting of words enriched with their POS) was more promising because of the following:

1. As discussed by Losee (2001), information retrieval with POS-tagged data improves the quality of analysis in many cases.
2. It is a computationally efficient way of increasing the amount of (potentially) relevant information.
3. It gives rise to POS-based feature selection techniques for further refinement.

The following are assumptions about our test and training data:

1. All words are in upper case.
2. All words are stemmed.
3. All words are POS tagged; we denote (word, POS) pairs as "word/POS".

5 Part of Speech Feature Selection

Even the most positive reviews have portions with negative polarity or no clear polarity at all. Since the training data consists of complete classified reviews, the presence of parts with conflicting polarities or lack of polarity presents a major obstacle to accurate OvOPI. To illustrate this inconsistent polarity, the following were all taken from a single review of *Apollo 13* (Leeper, 1995):

- Positive polarity: "*Special effects are first-rate*."
- Negative polarity: "*The character is written thinly*."
- No clear polarity: "*The scenes were shot in short segments*."

Note that at different levels of granularity, individual phrases and words vary in their contribution to opinion polarity. Sometimes only part of the meaning of a word contributes to opinion polarity (section 7). Any portion that does not contribute to the OvOP is considered noise. To reduce such noise, feature selection was introduced by using POS tags to do the following:

1. Introduce custom parts of speech, e.g. NEG and COP, when the tagger does not provide desired specificity (Brill Tagger does not provide POS for "negation" and "copula").
2. Remove the words that are least likely to contribute to the polarity of a review (determiner, preposition, etc.).

3. Reduce only parts of speech that introduce unnecessary variance to POS. It may be useful, for instance, for the classifier to record the presence of a proper noun. However, to include individual proper nouns would unnecessarily decrease the probability of finding the same n-grams in the test data.

Experimentation involved multiple combinations of such feature selection rules, yielding several separate results. An example of a specification of POS feature selection rules is shown in Figure 1.

	Rule	**Example**
Copula Conversion	is/* → */COP	be/VB → be/COP
Negation Conversion	not/* → /NEG	not/RB → /NEG
Noun Generalization	*/NN → /NN	food/NN → /NN
POS Tossing	*/CC → ∅	nor/CC → ∅

Figure 1. Abbreviated feature selection rule specification.

POS feature selection rules are not designed to reduce the effects of conflicting polarity, but to reduce the effect of lack of polarity. The effects of conflicting polarity have instead been addressed by careful preparation of the training data, as will be seen in the following section.

6 Experiments

For evaluation of their effects, feature selection rules were applied to a POS tagged corpus of movie reviews prior to training and classification. The experimental settings and results are given below.

6.1 Settings

- Data: Cornell Movie Reviews (Pang, *et al.*, 2002).
- Part-of-speech tagger: Brill (1995).
- WordNet: version 1.7.13 (Fellbaum, 1998).

Movie reviews were used for training and evaluation of each probabilistic classifier. The decision to use only movie reviews for training and test data was based on the fact that OvOPI of movie reviews is particularly challenging as shown by Turney (2002), and therefore can be considered a good environment for testing any system designed for OvOPI. The other reason for using movie reviews is the availability of large bodies of free data on the web. Specifically we used the data available through Cornell University from the Internet Movie Database.

The Cornell data consists of 27,000 movie reviews in HTML form, using 35 different rating scales such as A to F or 1 to 10 in addition to the common 5 star system. We divided them into two classes (positive and negative) and took 100 reviews from each class as the test set. For training sets, we first identified the reviews most likely to be positive or negative. For instance, when reviews contained letter grade ratings, only the A and F reviews were selected. This was done in an attempt to minimize the effects of conflicting polarities and to maximize the likelihood that our positive and negative labels would match those that the authors would have assigned to the reviews. From these reviews, we took random samples from each class in set sizes ranging from 50 to 750 reviews (in increments of 50). These sets consisted of the reviews that remained after the test sets had been removed. This resulted in training set sizes of 100 to 1500 in increments of 100. HTML documents were converted to plain text, tagged using the Brill Tagger, and fed into

feature selection modules and classifiers. The particular combinations of feature selection rules and classifiers and their results are described in the following sections.

The fact that as a training set we used data labelled by a reader and not directly by the writer poses a potential problem. We are learning a function that has to mimic the label identified by the writer, but we are using data labelled by the reader. We assume that this is an acceptable approximation because there is a strong practical relation between the label identified by the original writer and the reader. The authors themselves may not have made the polarity classifications, but we assume that language is an efficient form of communication. As such, variances between author and reader classification should be minimal.

6.2 Naïve Bayes

According to linguistic research, adjectives alone are good indicators of subjective expressions (Wiebe, 2000). Therefore, determining opinion polarity by analyzing occurrences of individual adjectives in a text should be an effective method. To identify the opinion polarity of movie reviews, a Naïve Bayes classifier using adjectives is a promising model. The effectiveness of adjectives compared to other parts of speech is evaluated by applying and comparing the results on data with only adjectives against data with all parts of speech. The impact of at-level generalization from adjectives to synsets, or "Sets of Synonyms" (section 7) is also measured. The Naïve Bayes classifier described above was applied to:

1. Tagged data.
2. Data containing only the adjectives.
3. Data containing only the synsets of the adjectives.

The adjectives in 3 were generalized to at-level synsets using a combination of the POS feature selection module and the feature generalization module. In Table 1 some of the most important POS tags used in this paper.

Tag	**Description**	**Example**
CC	Coordin. Conjunction	and, but, or
JJ	Adjective	Yellow
JJR	Adj., comparative	Bigger
JJS	Adj., superlative	Wildest
NN	Noun, sing. or mass	Llama
NNS	Noun, plural	Llamas
NNP	Proper noun, singular	IBM
NNPS	Proper noun, plural	Carolinas
RB	Adverb	quickly, never
RBR	Adverb, comparative	Faster
RBS	Adverb, superlative	Fastest
VB	Verb, base form	Eat
VBG	Verb, gerund	Eating
VBN	Verb, past participle	Eaten
VBZ	Verb, 3sg pres	Eats

Table 1. Some of the most important Penn Treebank part-of-speech tags (Jurafsky, et. al., 2000).

For each training data set, add-one smoothing (commonly known as Laplace smoothing) was applied to the Naïve Bayes classifier. Table 2 shows the resulting precisions of each data set type and size. The All-POS column describes the results when training and classifying on POS tagged data. The column tagged as JJ contains the results obtained after stripping away all words not tagged as adjectives, while the results listed in the JJ+WN column were generated after further generalization of all adjectives to their respective WordNet synsets.

Size	All-POS	JJ	JJ+WN
100	0.615	0.640	0.650
200	0.740	0.670	0.665
300	0.745	0.700	0.690
400	0.740	0.700	0.730
500	0.740	0.705	0.705
600	0.760	0.710	0.670
700	0.775	0.715	0.710
800	0.765	0.715	0.700
900	0.785	0.725	0.710
1000	0.765	0.755	0.720
1100	0.785	0.750	0.760
1200	0.765	0.735	0.750
1300	0.775	0.730	0.710
1400	0.775	0.735	0.745
1500	0.795	0.730	0.735

Table 2. Precisions of Naïve Bayes classifier trained on the results of different feature selections.

The results indicate that at-level generalization of adjectives is not effective and that extracting only adjectives degrades the classifier. However, this does not imply that feature selection does not work. Adjectives constitute 7.5% of the text in the data. The precision achieved on such a small portion of the data indicates that a significant portion of the opinion polarity information is carried in the adjectives alone. Although the resulting precisions are better in all-POS data, adjectives can still be considered good clues to opinion polarity.

6.3 Markov Model

Three types of data are applied to the Markov Model classifiers:

1. Tagged data without feature selection.
2. Tagged data with POS feature selection.
3. Tagged data with both POS feature selection and feature generalization.

Witten-Bell (Jurafsky, *et al.*, 2000) smoothing is applied to these classifiers.

6.3.1 POS Feature Selection

One design principle of the feature selection rules is that they filter out parts of speech that should not contribute to the opinion polarity, and keep the parts of speech that do contribute such meaning. Based on analysis of movie review texts, we devised feature selection rules that take POS-tagged text as input and return less noisy, more concentrated sentences that have

combinations of words and word/POS tag pairs removed from the originals. Table 3 is a summary of the feature selection rules defined in this experiment.

A new Part of Speech, "COP", was introduced to capture special verbs – is, was, am, are, were, be, been, like, liked, dislike, disliked, hate, hated, seem and seemed – which are here considered particularly relevant for capturing opinion polarities.

Parts of Speech	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
JJ/JJR/JJS	Keep	Keep	Keep	Keep	Keep
RB/RBS (without "not")	Drop	Keep	Keep	Keep	Keep
RBR (without "not")	Drop	Keep	Keep	Keep	Drop
VBG	Keep	Keep	Keep	Keep	Drop
NN/NNS (generalized to NN)	Gener.	Gener.	Gener.	Gener.	Gener.
NNP/NNPS (generalized to NNP)	Gener.	Gener.	Gener.	Gener.	Gener.
VBZ	Drop	Drop	Keep	Keep	Drop
CC	Drop	Drop	Drop	Keep	Keep
COP	Keep	Keep	Keep	Keep	Keep

Table 3. Summary of POS feature selection rules – "Gener." stands for "Generalize".

Wiebe et al., as well as other researchers, have shown that subjectivity is especially concentrated in adjectives (Wiebe, *et al.*, 1999; Hatzivassiloglou, 2000; Turney, *et al.*, 2003). Therefore, no adjectives or their tags were removed, nor were copula verbs or negative markers. However, noisy information, such as determiners, foreign words, prepositions, modal verbs, possessives, particles, interjections, etc., were removed from the text stream. Other parts of speech, such as nouns and verbs, were removed, but their POS-tags were retained.

The output returned from the feature selection module did not keep the original sentence structure. The concrete POS feature selection rules applied in this experiment are shown in Table 3. The following is an example of sentence preprocessing:

- "All Steve Martin fans should be impressed with this wonderful new comedy."
- /NNP /NNP /NN be/COP /VBN wonderful/JJ new/JJ /NN.

The resulting precisions for POS feature selection rules and different sizes of data sets are listed in Table 4.

Size	**Rule 1**	**Rule 2**	**Rule 3**	**Rule 4**	**Rule 5**	**All-POS**
100	0.555	0.625	0.625	0.630	0.630	0.575
300	0.660	0.635	0.655	0.655	0.655	0.675
500	0.640	0.665	0.665	0.680	0.680	0.720
700	0.705	0.700	0.700	0.690	0.690	0.735
900	0.700	0.740	0.740	0.765	0.765	0.760
1100	0.750	0.745	0.745	0.715	0.715	0.775
1300	0.715	0.695	0.695	0.705	0.705	0.805
1500	0.725	0.730	0.730	0.750	0.750	0.770

Table 4. Precisions on POS feature selection.

7 Synonymy and Hypernymy Feature Generalization

In non-technical written text repetition of identical words is not common, and is generally considered "bad style". As such, many authors attempt to use synonyms for words whose meanings they need often, propositions, and even generalizations. We attempted to address two of these perceived issues by identification of words with a set of likely synonyms, and by hypernymy generalization. For the implementation of these techniques, we took advantage of the WordNet system (Fellbaum, 1998), which provides the former by means of synsets for four separate classes of words (verbs, nouns, adjectives and adverbs), and the latter through hypernymy relations between synsets of the same class.

7.1 Synonyms

WordNet maps each of the words it supports into a synset, which is an abstract entity encompassing all words with a "reasonably" similar meaning. In the case of ambiguous words, multiple synsets may exist for a word; in these instances, we picked the first one. Note that synonyms (and general WordNet processing) are available only in instances where the word under consideration falls into one of the four classes of words outlined above. We determined the appropriate category for each word by using the assigned tag, and did not consider words which fell outside the classes supported by WordNet.

7.2 Hypernyms

For verbs and nouns WordNet provides a hypernymy relation which can be informally described as follows. If s_1 and s_2 are synsets, then s_1 is hypernym of s_2, notation $s_1 \succ s_2$, if and only if anything that can be described by a word in s_2 can also be described by a word in s_1, where $s_1 \neq s_2$. For each of the hypernym categories, we determine a set of abstract synsets A such that, for any $a \in A$, there does not exist any s such that $s \succ a$. We say that a synset h is a *level n hypernym* of a synset s if and only if $h \succ^* s$ and one of the following holds for some $a \in A$:

1. $a \succ^n h$
2. $s = h$ and $a \succ^l s$, with $l < n$

For example, given the WordNet database, a generalization to a level 4 hypernym for the nouns "movie" and "performance" will generalize both of them to one common synset which can be characterized by the word "communication".

7.3 Analysis

In order to determine the effects of translating words to synsets and performing hypernym generalization on them, we ran a series of tests which quickly determined that the effects of pure synset translation were negligible. We thus experimented with the computation of level n hypernyms with $n \in \{0,\ldots,10\}$, separately for nouns and verbs.

As we can see from Figure 2, applying hypernym generalization to information gathered from large data sets yielded little improvement; instead, we observed a decline in the quality of our classification caused by the loss of information.

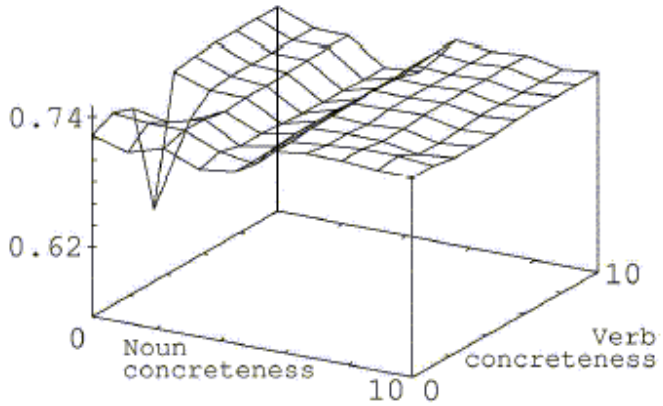

Figure 2. Hypernym generalization with 1500 reviews from each class.

The *x* and *y* axes describe the level of hypernym generalization for nouns and verbs, *z* the precision achieved. Maximum concreteness is reached at level 10, which indicates practically no generalization.

We assume that for larger data sets bigram classification is already able to make use of the more fine-grained data present. Shrinking the size of our training data, however, increased the impact of Wordnet simplification; for very small data sets (50 reviews and less) we observed an improvement of 2.5% (absolute) in comparison to both full generalization and no generalization. Increasing the size of the set of observable events by using trigram models resulted in a small gain (around 1%). The effect of verb generalization was relatively small in comparison to noun generalization for similar levels of hypernymy.

7.4 Discussion

Our results indicate that, except for very small data sets, the use of Word-Net hypernym generalization is not significantly beneficial to the classification process. We conjecture that the following reasons could explain such a phenomenon:

- WordNet is too general for our purposes. It considers many meanings and hypernymy relations which are rarely relevant to the field of Movie Reviews, but which potentially take precedence over other relations which might be more appropriate to our task.
- Choosing the first synset out of the set of choices is unlikely to yield the correct result, given the lack of WordNet specialization in our domain of focus.
- For reasonably large data sets supervised learning mechanisms gain sufficient confidence with related words to make this particular auxiliary technique less useful.

In light of these reasons, the use of a domain-specific database might improve the performance of this technique.

8 Selection by Ranking

The probabilistic models computed by the Naïve Bayes classifiers were sorted by log posterior odds on positive and negative orientations for the purpose of ranking, i.e. by a score computed as follows:

$$score = \log P(+|rv) - \log P(-|rv)$$

where:

- *rv* is the review under consideration.
- *P*(+|*rv*) is the probability of *rv* being of positive polarity.
- *P*(-|*rv*) is the probability of *rv* being of negative polarity.

We modified the classifier so that it:

1. sorts the reviews in the test data by log posterior odds,
2. returns the first *N* reviews from the sorted list as positive reviews,
3. returns the last *N* reviews from the sorted list as negative reviews.

The resulting precisions and recalls on different *N* are summarized in Table 5.

N	Precision	Recall
10	1.00	0.10
30	0.90	0.27
50	0.88	0.44
70	0.83	0.58
90	0.78	0.68

Table 5. Precisions and Recalls by Number of Inputs.

The classifier was trained on the same 1500-review data set and was used with ranking on a repository of 200 reviews which were identical to the test data set. The result is very positive and indicates that adjectives provide enough sentiment to detect extremely positive or negative reviews with good precision. While the number of reviews returned is specified in this particular example, it is also possible to use assurance as the cut-off criterion by giving log posterior odds.

This idea has already been applied in information retrieval tasks. Zhai, *et al.*, (1999) refer to it as *adaptive filtering* and identify two basic problems: *threshold setting*, which assigns initial threshold values, and *threshold updating*, which updates these thresholds based on feedback. Shanahan, *et al.*, (2003) apply Zhai's beta-gamma algorithm to SVM classification in order to improve recall.

9 Discussion

Taking all results into consideration, both the Naïve Bayes classifier and Bigram Markov Model classifier performed best when trained on sufficiently large data sets without feature selection. For both Bigram and Trigram Markov Models, we observed a noticeable improvement with our feature generalization when training on very small data sets; for trigram models, this improvement even extended to fairly large data sets (1500 reviews).

One explanation for this result is that the feature selection and generalization are unable to make use of the more fine-grained information provided to them. A likely reason for this is that the ratio

between the size of the set of observable events and the size of the training data set is comparatively large in both cases. However, further research and testing will be required in order to establish a more concrete understanding of the usefulness of this technique. The learning curve of classifiers with the POS features selection and/or the feature generalization climbs at higher rates than those without, and results in lower precision with larger data sets. One possible explanation of the higher climbing rates is that the POS feature selection and the feature generalization compact the possible events in language models while respecting the underlying model by reducing the size of the vocabulary. This also explains why the plateau effect is observed with data sets of smaller size. The degraded results with feature selection and generalization also indicate that when information is removed from training and test data, the compacted language model loses resolution.

10 Conclusion

In a supervised machine learning framework a two-phased classification mechanism is introduced and implemented with a POS feature selection, a feature generalization, a Naïve Bayes classifier and a Markov Model classifier. Precisions of combinations of feature selection and generalization and classifiers are evaluated by experiments. Although the results from classifications without feature selection and generalization are generally better than the results from those with, the POS feature selection and feature generalization still have potential to improve overall opinion-polarity identification. Feature generalization using synonymy and hypernymy shows good precision for small data sets and warrants further research. Using the Naïve Bayes classifier with ranking on adjectives has confirmed that high precision can be achieved by dropping recalls. For the task of finding reviews of strong positive or negative polarity within a given data set, very high precision was observed for adequate recall.

Acknowledgements: The authors would like to thank Tomohiro Oda for his extensive help and support during the course of all stages of the project. We also wish to acknowledge Larry D. Blair, Francis I. Dunahue, Assad Jarrahian, Helen Johnson, James H. Martin, Jim Glasscock, Jeff Rueppel and Philipp Wetzler for their valuable contributions.

11 Bibliography

Brill, E. (1995) *Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging*. Computational Linguistics, 21(4):543-565.

Duda, R. O. and Hart, P. E. (1973) *Pattern Classification and Scene Anaylsis*. A Wiley Interscience Publication, New York.

Fellbaum, C. (1998) *Wordnet: An Electronic Lexical Database*. The MIT Press.

Hatzivassiloglou, V. and McKeown, K. R. (1997) *Predicting the Semantic Orientation of Adjectives*. In Cohen, P. R. and Wahlster. W. (Ed.) *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*. 174-181. Association for Computational Linguistics.

Hatzivassiloglou, V., and Wiebe, J. (2000) *Effects of Adjective Orientation and Gradability on Sentence Subjectivity*. Proceedings of the 18th International Conference in Computational Linguistics.

Jurafsky, D. and Martin, J. H. (2000) *Speech and Language Processing*. Prenctice Hall.

Leeper, M. R. (1995) *Review of Apollo 13*, Usenet rec.arts.movies.reviews.

Losee, R. M. (2001) *Natural Language Processing in Support of Decision-making: Phrases and Part-of-Speech Tagging*. Information Processing and Management, 37(6):769-787.

Marcus, M. P., Santorini, B. and Marcinkiewicz, M. A. (1994) *Building a Large Annotated Corpus of English: The Penn Treebank*. Computational Linguistics, 19(2):313-330.

Pang, B., Lee, L. and Vaithyanathan, S. (2002) *Thumbs up? Sentiment Classification using Machine Learning Techniques*. Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Shanahan, J. G., Roma, N. (2003) *Improving SVM Text Classification Performance through Threshold Adjustment*. European Conference on Machine Learning (ECML) 2003, 361-372.

Turney, P. (2002) *Thumbs up or Thumbs down? Semantic Orientation applied to Unsupervised Classification of Reviews*. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL02), 417-424.

Turney, P. and Littman, M. (2003) *Measuring Praise and Criticism: Inference of Semantic Orientation from Association*. ACM Transactions on Information Systems (TOIS), 21(4):315-346.

Wiebe, J., Bruce, R. F. and O'Hara, T. (1999) *Development and Use of a Gold-Standard Data Set for Subjectivity Classifications*. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL99), 223-287.

Wiebe, J. (2000) *Learning Subjective Adjectives from Corpora*. Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Application of Artificial Intelligence, 735-740. AAAI Press / The MIT Press.

Zhai, C., Jansen, P., Stoica, E., Grot, N., Evans, D.A. (1999) *Threshold Calibration in CLARIT Adaptive Filtering*. Seventh Text Retrieval Conference (TREC-7), 149-156.