Automatic Opinion Polarity Classification of Movie Reviews

Franco Salvetti
Department of Computer Science, University of Colorado at Boulder

Stephen Lewis
Department of Linguistics, University of Colorado at Boulder

Christoph Reichenbach
Department of Computer Science, University of Colorado at Boulder

One approach to assessing 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 filtering, applied to reviews, on the accuracy of two statistical classifiers (Naive Bayes and Markov Model) with respect to OvOP identification is observed. Two kinds of lexical filters, one based on hypernymy as provided by WordNet (FELLBAUM, 1998), and one hand-crafted filter based on part-of-speech (POS) tags, are evaluated. A ranking criterion based on a function of the probability of having positive or negative polarity is introduced and verified as being capable of achieving 100% accuracy with 10% recall. Movie reviews are used for training and evaluation of each statistical classifier, achieving 80% accuracy.

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 to 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 that is already interested in a certain product may want to read some negative reviews just to pinpoint possible drawbacks, but has 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, if requested. Such a classification is here defined as the overall opinion polarity (OvOP) of a review, or simply the polarity. The process of identifying OvOP of a review will be referred to as Overall Opinion Polarity Identification (OvOPI).

A system that is capable of labeling 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 first appears 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 filtering, applied to

Colorado Research in Linguistics. June 2004. Volume 17, Issue 1. Boulder: University of Colorado.

reviews, on the accuracy of two statistical classifiers trained on such filtered data. We have implemented two different kinds of lexical filters, one based on hypernymy as provided by WordNet (FELLBAUM, 1998), and one based on part-of-speech (POS) tags.

The results obtained by experiments based on movie reviews revealed that WordNet filters produce less improvement than do POS filters, and that for neither is there evidence of significantly improved performance over the system without filters, although the overall performance of our system is comparable to systems in current research, achieving an accuracy of 81%.

In the domain of OvOPI of reviews it is often acceptable to sacrifice recall for accuracy. Here we also present a system whereby the reviews are ranked based on a function of the probability of being positive/negative. Using this ranking method we achieve 100% accuracy 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 accuracy 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. combined a log-linear statistical model that examined the conjunctions between adjectives, (such as "and", "but", "or"), with a clustering algorithm that grouped the adjectives into two sets which were then labeled positive and negative (HATZIVASSILOGLOU et al, 1997). Their model predicted whether adjectives carried positive or negative polarity with 82% accuracy. However, because the model was unsupervised it required an immense, 21 million word corpus to function.

Turney extracted n-grams based on adjectives (TURNEY, 2002). 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. Turney achieved an average of 74% accuracy 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. concluded that the task of polarity classification was not the same as topic classification (PANG et al, 2002). 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 returned a 77.3% accuracy using bigrams. Their best results came using unigrams, calculated by the Support Vector Machine at 82.9% accuracy. Maximum Entropy performed best using both unigrams and bigrams at 80.8% accuracy, and Naïve Bayes performed best at 81.5% using unigrams with POS tags.

3. Statistical approaches to polarity identification

There are many possible approaches to identifying the actual polarity of a document. Our analysis uses statistical methods, namely supervised machine learning, to identify the likelihood of reviews having "positive" or "negative" polarity with respect to 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 Naïve Bayes classifier is a well-known supervised machine learning approach. 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

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

$$\widehat{c} = \mathtt{argmax}\ \Pr(c|\mathtt{rv})$$

where

- rv is the review under consideration,
- w is a word/POS-tag pair that appears in the given document,
- Pr(appw|class) is the probability that a word/POS-tag pair appears in a document of the given class in training data, and
- bc 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 are based on 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 Models

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 is a widely used probabilistic model that does capture connectivity among words.

This Markov Model classifier develops two language models: one on positive reviews and another on negative reviews. The classifier then generates two probabilities for an unseen review, one from the positive model and the other from the negative one. It compares the two probabilities and determines the classification. The following formula is the one used to compute the probability that a document could be generated using each language model.

$$\Pr(\mathtt{rv}) = \prod_{\mathtt{sn} \in \mathtt{Review}} \Pr(\mathtt{sn})$$

$$\Pr(\mathtt{sn}) = \Pr(w_1|\langle s \rangle) \left(\prod_i \Pr(w_{i+1}|w_i) \right) \Pr(\langle /s \rangle)$$

where

- rv is the review under consideration,
- sn is a sentence,
- ⟨s⟩ is the start of a sentence,
- ⟨/s⟩ is the end of a sentence.

4. Features for analysis

Statistical analysis depends on a sequence of tokens it uses as characteristic features of the objects 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 we find within reviews would be to use individual words from the review data as tokens. However, just using the words discards semantic information about the remainder of the sentence; as such, it may be desirable to first perform some sort of semantic analysis to enrich the tokens with useful information, or even discard misleading or irrelevant information (noise), in order to increase accuracy.

Three basic approaches for handling this kind of data preprocessing come to mind:

- Leave the data as-is: Each word will be represented by itself
- Parts-of-speech tagging: Each word is enriched by a POS tag, as determined by a standard tagging technique (such as the Brill Tagger (BRILL, 1995))
- Perform POS tagging and parse (using e.g. the Penn Treebank (MARCUS et al, 1994))

Unfortunately, the third approach not only had severe performance issues during our early experiments, but also raises conceptional questions of how such data would be incorporated into a statistical analysis. We thus focus our analysis in this paper on POS-tagged data (sentences consisting of words enriched with information about their parts of speech), which seems to be a good candidate for a worthwhile source of information, for the following reasons:

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

We thus make the following assumptions about our test and training data:

1. All words are transformed into upper case,
2. All words are stemmed,
3. All words are transformed into (word, POS) tuples by POS tagging (notation word / POS).

All of these are computationally easy to achieve (with a reasonable amount of accuracy) using the Brill Tagger.

5. Part of Speech Filters

Careful analysis of movie reviews has made it clear that even the most positive reviews have portions with negative polarity or no clear polarity at all. Since the training data used here consists of complete classified reviews, the presence of parts with conflicting polarities or lack of polarity within a review presents a major obstacle for accurate OvOPI. As illustration of this inconsistent polarity, the following were all taken from a single review[1].

"Special effects are first-rate"
(positive polarity)

"The character is written thinly"
(negative polarity)

"The scenes were shot in short segments"
(no clear polarity)

This observation can be taken to lower levels as well. Individual phrases and words vary in their contribution to opinion polarity. It may even be said that only some part of the meaning of a word contributes to opinion polarity (see WordNet filter section below). Any portion that does not contribute to the OvOP is noise. To reduce noise, filters were developed that use POS tags to do the following.

1. Introduce custom parts of speech when the tagger does not provide desired specificity (negation and copula).
2. Remove the words that are least likely to contribute to the polarity of a review (determiner, preposition, etc.)
3. Reduce parts of speech that introduce unnecessary variance to POS only. 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 filter rules, yielding several separate filters. An example of a specification of POS filter rules is shown in Figure 1.

[1] APOLLO 13, A film review by Mark R. Leeper, Copyright © 1995 Mark R. Leeper

(1a) Copula Conversion:

is/* → */COP

(1b) Negation conversion:

not/* → /NEG

(2) Noun generalization:

*/NN → /NN

(3) POS Tossing:

*/CC → Ø

Figure 1: Abbreviated filter rule specification (illustrative details only)

The POS filters are not designed to reduce the effects of conflicting polarity. They are only designed 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

6.1. Settings

- Data: taken from Cornell Data (PANG et al, 2002)
- Part-of-speech tagger: Brill tagger (BRILL, 1995)
- WordNet: WordNet version 1.7.13 (FELLBAUM, 1998)

Movie reviews are used for training and evaluation of each statistical 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 (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. . . F or 1. . . 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 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, 200, ..., 1500 (in increments of 100). HTML documents were converted to plain text,

tagged using the Brill tagger, and fed into filters and classifiers. The particular combinations of filters and classifiers and their results are described in the following sections.

The fact that as a training set we used data labeled 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 labeled 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"; see "WordNet filtering", below) 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 filter module and the generalization filter module. For each training data set, add-one smoothing was applied to the Naïve Bayes classifier. Table 1 shows the resulting accuracies of each data set type and size.

Size	All-POS	JJ	JJ+WN
100	0.615	0.64	0.65
200	0.74	0.67	0.665
300	0.745	0.7	0.69
400	0.74	0.7	0.73
500	0.74	0.705	0.705
600	0.76	0.71	0.67
700	0.775	0.715	0.71
800	0.765	0.715	0.70
900	0.785	0.725	0.71
1000	0.765	0.755	0.72
1100	0.785	0.75	0.76
1200	0.765	0.734	0.75
1300	0.775	0.73	0.71
1400	0.775	0.735	0.745
1500	0.795	0.73	0.735

Table 1: Accuracies of Naïve Bayes classifier. JJ means "adjectives only", WN indicates synset mapping using the generalization filter.

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 filtering does not work. Adjectives constitute 7.5% of the text in the data. The accuracy 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 accuracies are better in all-POS data, adjectives can still be considered good clues of opinion polarity.

6.3. Markov Model

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

1. Tagged data without any filtering,
2. Tagged data with POS filters,
3. Tagged data with both POS filters and generalization filters.

Witten-Bell smoothing is applied to this classifier.

6.3.1. POS filtering

One design principle of the filter rules is that they filter out parts of speech that do 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 "filter rules" that take Brill-tagged text as input and return less noisy, more concentrated sentences that have a combination of words and word/POS tag pairs removed from the original. A summary of the filter rules defined in this experiment is shown in Table 2.

POS[1]	rule1	rule2	rule3	rule4	rule5
JJ[2]	K	K	K	K	K
RB[3]	D	K	K	K	K[4]
VBG	K	K	K	K	D
VBN	K	K	K	K	D
NN[5]	G	G	G	G	G
VBZ	D	D	K	K	D
CC	D	D	D	K	K
COP[6]	K	K	K	K	K

K: Keep D: Drop G: Generalize

[1]Abbreviations of POSs are based on the tree bank's notation
[2]JJ includes JJ, JJR and JJS
[3]RB includes RB, RBR and RBS except "not"
[4]RBRs are dropped and RB and RBS stay
[5]NN and NNS are generalized to NN, and NNP and NNPS are generalized to NNP
[6]COPS are particular verbs: is, was, am, are, were, be, been, like, liked, dislike, disliked, hate, hated, seem and seemed

Table 2: Summary of POS filter rules

Wiebe et al., as well as other researchers, showed 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 filter did not keep the original sentence structure. The concrete POS filtering rules applied in this experiment are shown in Table 2. The following is an example of the 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 accuracies on POS filter rules and different sizes of data sets are listed in Table 3.

size	rule1	rule2	rule3	rule4	rule5	All-POS
100	0.555	0.625	0.625	0.63	0.63	0.575
200	0.675	0.71	0.71	0.7	0.7	0.655
300	0.66	0.635	0.635	0.655	0.655	0.675
400	0.7	0.66	0.66	0.685	0.685	0.71
500	0.64	0.665	0.665	0.68	0.68	0.72
600	0.685	0.75	0.75	0.765	0.765	0.745
700	0.705	0.7	0.7	0.69	0.69	0.735
800	0.7	0.74	0.74	0.715	0.715	0.69
900	0.7	0.74	0.74	0.765	0.765	0.76
1000	0.73	0.745	0.745	0.73	0.73	0.765
1100	0.75	0.745	0.745	0.715	0.715	0.775
1200	0.71	0.71	0.71	0.72	0.72	0.765
1300	0.715	0.695	0.695	0.705	0.705	0.77
1400	0.755	0.745	0.745	0.755	0.755	0.805
1500	0.725	0.73	0.73	0.75	0.75	0.77

Table 3: Accuracies on POS filtering

7. WordNet filtering

In non-technical written text it is uncommon to encounter repetitions of identical words; this is generally considered "bad style". As such, many authors attempt to use synonyms for words whose meanings they need often, propositions, or even generalizations. We attempted to address two of these perceived issues by identifying words with a set of likely synonyms, and by hypernymy generalization. For the implementation of these techniques, we took advantage of the WordNet (FELLBAUM, 1998) system, 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 only available in instances where the word under consideration falls in one of the four classes of words we outlined above. We determined the appropriate category for each word by examining the tag it was assigned by the Brill tagger, not touching words which fell outside of these classes.

7.2. Hypernyms

For verbs and nouns, WordNet provides a hypernymy relation, which can be informally described as follows: Let s_1, s_2 be synsets. Then s_1 is hypernym of s_2, notation $s_1 \sqsupset s_2$, 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, and $s_1 \neq s_2$. 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 $s \sqsubset a$. We say that a synset h is a *Level n hypernym* of a synset s if and only if $h \sqsubset^* s$ and one of the following holds for some $a \in A$:

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

For example, given the WordNet database, a hypernym generalization of level 4 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 hypernymization 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.

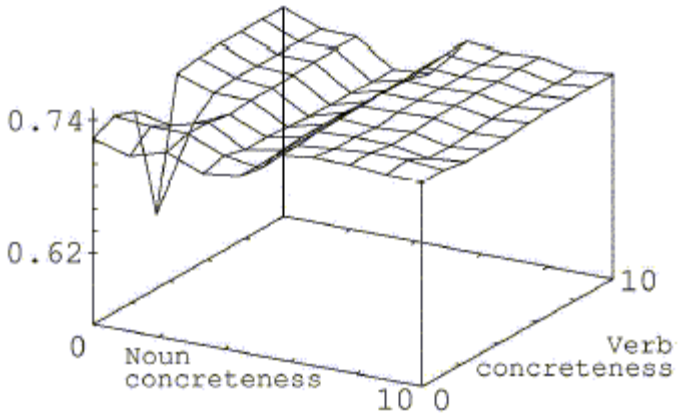

Figure 2: Hypernym generalization with 1500 reviews from each class. The x and y axis describe the level of hypernym generalization for nouns and verbs, z the accuracy we achieved. Maximum concreteness at level 10 indicates no generalization.

As we can see from Figure 2, applying hypernym generalization to information gathered from large data sets yielded little improvement; instead, we observed a degradation in the quality of our classification caused by the loss of information. 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, not shown here) we observed an improvement of 2.5% (absolute) in comparison to both full generalization and no generalization at all. Increasing the size of the set of observable events by using trigram models resulted in a small gain (around 1%). Interestingly, the effect of verb generalization was relatively small in comparison to noun generalization for similar hypernymy levels.

7.4. Discussion

Our results indicate that, except for very small data sets, the use of Word-Net hypernymy generalization is not significantly beneficial to the classification process. We assume that this is due to at least the following reasons:

- 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 here.
- Choosing the first synset out of the set of choices is unlikely to yield the correct result, given the lack of WordNet's specialization on 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.

Considering this, the use of a domain-specific database seems to be a promising approach to improving our performance for 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:

$$\mathbf{score = \log Pr(+|rv) - \log Pr(-|rv)}$$

where

- rv is the review under consideration,
- Pr(+|rv) is the probability of rv being a review of positive polarity,
- Pr(-|rv) analogously is the probability of the review 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 accuracies and recalls on different N are summarized in Table 4.

N	precision	recall
10	1.000	0.100
20	0.975	0.195
30	0.900	0.270
40	0.900	0.360
50	0.880	0.440
60	0.867	0.520
70	0.830	0.580
80	0.780	0.625
90	0.780	0.680

Table 4: 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 accuracy. While the number of reviews returned is specified in this particular example, it is also possible to use assurance as the cutoff criterion by giving log posterior odds.

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 filtering. For both Bigram and Trigram Markov Models, we observed a noticeable improvement with our generalization filter 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 filters 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 filter and/or the generalization filter climbs at higher rates than those without the filters and results in lower accuracy with larger data sets. One possible explanation of the higher climbing rates is that the POS filter and the generalization filter compact the possible events in language models while respecting the underlying model by reducing vocabulary. This also explains why the plateau effect is observed with smaller data set sizes. The degraded results with filters also indicate that by removing information from training and test data, the compacted language model loses resolution.

10. Conclusion

A framework of two-phased classification mechanism is introduced and implemented with a POS filter, a generalization filter, a Naïve Bayes classifier and a Markov Model classifier. Accuracies of combinations of filters and classifiers are evaluated by experiments. Although the results from classifications without filters are better than those with filters, the POS filters and generalization filters are observed to still have potential to improve overall opinion polarity identification. Generalization filtering using WordNet shows good accuracy for small data sets and warrants further research. Using the Naïve Bayes classifier with ranking on adjectives has confirmed that desired 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. Further acknowledgements go to Larry D. Blair, Assad Jaharria, Helen Johnson, Jim Martin, Jeff Rueppel and Philipp Wetzler for their valuable contributions.

References

ERIC BRILL. "Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging". Computational Linguistics, 21(4):543–565, 1995. Tagger available from http://www.cs.jhu.edu/˜brill/RBT1 14.tar.Z.

CHRISTIANE FELLBAUM. *Wordnet: An electronic lexical database*, 1998. WordNet is available from http://www.cogsci.princeton.edu/˜wn/.

VASILEIOS HATZIVASSILOGLOU. *Effects of adjective orientation and gradability on sentence subjectivity*, 2000.

VASILEIOS HATZIVASSILOGLOU AND KATHLEEN R. MCKEOWN. "Predicting the semantic orientation of adjectives". In Philip R. Cohen and Wolfgang Wahlster, editors, 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, pages 174–181, Somerset, New Jersey, 1997. Association for Computational Linguistics.

ROBERT M. LOSEE. "Natural language processing in support of decisionmaking: phrases and part-of-speech tagging". Information Processing and Management, 37(6):769–787, 2001.

MITCHELL P. MARCUS, BEATRICE SANTORINI, AND MARY ANN MARCINKIEWICZ. "Building a large annotated corpus of english: The penn Treebank". Computational Linguistics, 19(2):313–330, 1994.

BO PANG, LILLIAN LEE, AND SHIVAKUMAR VAITHYANATHAN. "Thumbs up? sentiment classification using machine learning techniques". In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002. Movie Review

data available at http://www.cs.cornell.edu/people/pabo/movie-review-data/movie.zip.

PETER TURNEY AND MICHAEL LITTMAN. "Measuring praise and criticism: Inference of semantic orientation from association". ACM Transactions on Information Systems (TOIS), 21(4):315–346, 2003.

PETER TURNEY. "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews". In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), pages 417–424, 2002.

JANYCE WIEBE, REBECCA F. BRUCE, AND THOMAS O'HARA. "Development and use of a gold-standard data set for subjectivity classifications". In Proc. 37th Annual Meeting of the Assoc. for Computational Linguistics (ACL-99), 1999.

JANYCE WIEBE. "Learning subjective adjectives from corpora". In AAAI/IAAI, 2000.