# When Specialists and Generalists Work Together: Overcoming Domain **Dependence in Sentiment Tagging**

Alina Andreevskaia Concordia University Montreal, Quebec andreev@cs.concordia.ca

**Sabine Bergler** Concordia University Montreal. Canada bergler@cs.concordia.ca

#### Abstract

This study presents a novel approach to the problem of system portability across different domains: a sentiment annotation system that integrates a corpus-based classifier trained on a small set of annotated in-domain data and a lexicon-based system trained on Word-Net. The paper explores the challenges of system portability across domains and text genres (movie reviews, news, blogs, and product reviews), highlights the factors affecting system performance on out-of-domain and smallset in-domain data, and presents a new system consisting of the ensemble of two classifiers with precision-based vote weighting, that provides significant gains in accuracy and recall over the corpus-based classifier and the lexicon-based system taken individually.

#### 1 Introduction

One of the emerging directions in NLP is the development of machine learning methods that perform well not only on the domain on which they were trained, but also on other domains, for which training data is not available or is not sufficient to ensure adequate machine learning. Many applications require reliable processing of heterogeneous corpora, such as the World Wide Web, where the diversity of genres and domains present in the Internet limits the feasibility of in-domain training. In this paper, sentiment annotation is defined as the assignment of positive, negative or neutral sentiment values to texts, sentences, and other linguistic units. Recent experiments assessing system portability across different domains, conducted by Aue and Gamon (2005), demonstrated that sentiment annotation classifiers trained in one domain do not perform well on other domains. A number of methods has been proposed in order to overcome this system portability limitation by using out-of-domain data, unlabelled in-domain corpora or a combination of in-domain and out-of-domain examples (Aue and Gamon, 2005; Bai et al., 2005; Drezde et al., 2007; Tan et al., 2007).

In this paper, we present a novel approach to the problem of system portability across different domains by developing a sentiment annotation system that integrates a corpus-based classifier with a lexicon-based system trained on WordNet. By adopting this approach, we sought to develop a system that relies on both general and domainspecific knowledge, as humans do when analyzing a text. The information contained in lexicographical sources, such as WordNet, reflects a lay person's general knowledge about the world, while domainspecific knowledge can be acquired through classifier training on a small set of in-domain data.

The first part of this paper reviews the extant literature on domain adaptation in sentiment analysis and highlights promising directions for research. The second part establishes a baseline for system evaluation by drawing comparisons of system performance across four different domains/genres movie reviews, news, blogs, and product reviews. The final, third part of the paper presents our system, composed of an ensemble of two classifiers one trained on WordNet glosses and synsets and the other trained on a small in-domain training set.

### 2 Domain Adaptation in Sentiment Research

Most text-level sentiment classifiers use standard machine learning techniques to learn and select features from labeled corpora. Such approaches work well in situations where large labeled corpora are available for training and validation (e.g., movie reviews), but they do not perform well when training data is scarce or when it comes from a different domain (Aue and Gamon, 2005; Read, 2005), topic (Read, 2005) or time period (Read, 2005). There are two alternatives to supervised machine learning that can be used to get around this problem: on the one hand, general lists of sentiment clues/features can be acquired from domain-independent sources such as dictionaries or the Internet, on the other hand, unsupervised and weakly-supervised approaches can be used to take advantage of a small number of annotated in-domain examples and/or of unlabelled indomain data.

The first approach, using general word lists automatically acquired from the Internet or from dictionaries, outperforms corpus-based classifiers when such classifiers use out-of-domain training data or when the training corpus is not sufficiently large to accumulate the necessary feature frequency information. But such general word lists were shown to perform worse than statistical models built on sufficiently large in-domain training sets of movie reviews (Pang et al., 2002). On other domains, such as product reviews, the performance of systems that use general word lists is comparable to the performance of supervised machine learning approaches (Gamon and Aue, 2005).

The recognition of major performance deficiencies of supervised machine learning methods with insufficient or out-of-domain training brought about an increased interest in unsupervised and weaklysupervised approaches to feature learning. For instance, Aue and Gamon (2005) proposed training on a samll number of labeled examples and large quantities of unlabelled in-domain data. This system performed well even when compared to systems trained on a large set of in-domain examples: on feedback messages from a web survey on knowledge bases, Aue and Gamon report 73.86% accuracy using unlabelled data compared to 77.34% for in-domain and 72.39% for the best out-of-domain training on a large training set.

Drezde et al. (2007) applied structural correspondence learning (Drezde et al., 2007) to the task of domain adaptation for sentiment classification of product reviews. They showed that, depending on the domain, a small number (e.g., 50) of labeled examples allows to adapt the model learned on another corpus to a new domain. However, they note that the success of such adaptation and the number of necessary in-domain examples depends on the similarity between the original domain and the new one. Similarly, Tan et al. (2007) suggested to combine out-of-domain labeled examples with unlabelled ones from the target domain in order to solve the domain-transfer problem. They applied an outof-domain-trained SVM classifier to label examples from the target domain and then retrained the classifier using these new examples. In order to maximize the utility of the examples from the target domain, these examples were selected using Similarity Ranking and Relative Similarity Ranking algorithms (Tan et al., 2007). Depending on the similarity between domains, this method brought up to 15% gain compared to the baseline SVM.

Overall, the development of semi-supervised approaches to sentiment tagging is a promising direction of the research in this area but so far, based on reported results, the performance of such methods is inferior to the supervised approaches with indomain training and to the methods that use general word lists. It also strongly depends on the similarity between the domains as has been shown by (Drezde et al., 2007; Tan et al., 2007).

#### **3** Factors Affecting System Performance

The comparison of system performance across different domains involves a number of factors that can significantly affect system performance – from training set size to level of analysis (sentence or entire document), document domain/genre and many other factors. In this section we present a series of experiments conducted to assess the effects of different external factors (i.e., factors unrelated to the merits of the system itself) on system performance in order to establish the baseline for performance comparisons across different domains/genres.

#### 3.1 Level of Analysis

Research on sentiment annotation is usually conducted at the text (Aue and Gamon, 2005; Pang et al., 2002; Pang and Lee, 2004; Riloff et al., 2006; Turney, 2002; Turney and Littman, 2003) or at the sentence levels (Gamon and Aue, 2005; Hu and Liu, 2004; Kim and Hovy, 2005; Riloff et al., 2006). It should be noted that each of these levels presents different challenges for sentiment annotation. For example, it has been observed that texts often contain multiple opinions on different topics (Turney, 2002; Wiebe et al., 2001), which makes assignment of the overall sentiment to the whole document problematic. On the other hand, each individual sentence contains a limited number of sentiment clues, which often negatively affects the accuracy and recall if that single sentiment clue encountered in the sentence was not learned by the system.

Since the comparison of sentiment annotation system performance on texts and on sentences has not been attempted to date, we also sought to close this gap in the literature by conducting the first set of our comparative experiments on data sets of 2,002 movie review texts and 10,662 movie review snippets (5331 with positive and 5331 with negative sentiment) provided by Bo Pang (http://www.cs.cornell.edu/People/pabo/moviereview-data/).

#### 3.2 Domain Effects

The second set of our experiments explores system performance on different domains at sentence level. For this we used four different data sets of sentences annotated with sentiment tags:

- A set of movie review snippets (further: movie) from (Pang and Lee, 2005). This dataset of 10,662 snippets was collected automatically from www.rottentomatoes.com website. All sentences in reviews marked "rotten" were considered negative and snippets from "fresh" reviews were deemed positive. In order to make the results obtained on this dataset comparable to other domains, a randomly selected subset of 1066 snippets was used in the experiments.
- A balanced corpus of 800 manually annotated sentences extracted from 83 newspaper texts

(further, news). The full set of sentences was annotated by one judge. 200 sentences from this corpus (100 positive and 100 negative) were also randomly selected from the corpus for an inter-annotator agreement study and were manually annotated by two independent annotators. The pairwise agreement between annotators was calculated as the percent of same tags divided by the number of sentences with this tag in the gold standard. The pair-wise agreement between the three annotators ranged from 92.5 to 95.9% ( $\kappa$ =0.74 and 0.75 respectively) on positive vs. negative tags.

- A set of sentences taken from personal weblogs (further, blogs) posted on Live-Journal (http://www.livejournal.com) and on http://www.cyberjournalist.com. This corpus is composed of 800 sentences (400 sentences with positive and 400 sentences with negative sentiment). In order to establish the interannotator agreement, two independent judges were asked to annotate 200 sentences from this corpus. The agreement between the two annotators on positive vs. negative tags reached 99% (κ=0.97).
- A set of 1200 product review (PR) sentences extracted from the annotated corpus made available by Bing Liu (Hu and Liu, 2004) (http://www.cs.uic.edu/ liub/FBS/FBS.html).

The data set sizes are summarized in Table 1.

	Movies	News	Blogs	PR
Text level	2002 texts	n/a	n/a	n/a
Sentence level	10662	800	800	1200
	snippets	sent.	sent.	sent.

Table 1: Datasets

# **3.3** Establishing a Baseline for a Corpus-based System (CBS)

Supervised statistical methods have been very successful in sentiment tagging of texts: on movie review texts they reach accuracies of 85-90% (Aue and Gamon, 2005; Pang and Lee, 2004). These methods perform particularly well when a large volume of labeled data from the same domain as the

test set is available for training (Aue and Gamon, 2005). For this reason, most of the research on sentiment tagging using statistical classifiers was limited to product and movie reviews, where review authors usually indicate their sentiment in a form of a standardized score that accompanies the texts of their reviews.

The lack of sufficient data for training appears to be the main reason for the virtual absence of experiments with statistical classifiers in sentiment tagging at the sentence level. To our knowledge, the only work that describes the application of statistical classifiers (SVM) to sentence-level sentiment classification is (Gamon and Aue, 2005)<sup>1</sup>. The average performance of the system on ternary classification (positive, negative, and neutral) was between 0.50 and 0.52 for both average precision and recall. The results reported by (Riloff et al., 2006) for binary classification of sentences in a related domain of subjectivity tagging (i.e., the separation of sentiment-laden from neutral sentences) suggest that statistical classifiers can perform well on this task: the authors have reached 74.9% accuracy on the MPQA corpus (Riloff et al., 2006).

In order to explore the performance of different approaches in sentiment annotation at the text and sentence levels, we used a basic Naïve Bayes classifier. It has been shown that both Naïve Bayes and SVMs perform with similar accuracy on different sentiment tagging tasks (Pang and Lee, 2004). These observations were confirmed with our own experiments with SVMs and Naïve Bayes (Table 3). We used the Weka package (http://www.cs.waikato.ac.nz/ml/weka/) with default settings.

In the sections that follow, we describe a set of comparative experiments with SVMs and Naïve Bayes classifiers (1) on texts and sentences and (2) on four different domains (movie reviews, news, blogs, and product reviews). System runs with unigrams, bigrams, and trigrams as features and with different training set sizes are presented.

### **4** Experiments

#### 4.1 System Performance on Texts vs. Sentences

The experiments comparing in-domain trained system performance on texts vs. sentences were conducted on 2,002 movie review texts and on 10,662 movie review snippets. The results with 10-fold cross-validation are reported in Table  $2^2$ .

	Trained	on Texts	Trained on Sent.	
	Tested on	Tested on	Tested on	Tested on
	Texts	Sent.	Texts	Sent.
1gram	81.1	69.0	66.8	77.4
2gram	83.7	68.6	71.2	73.9
3gram	82.5	64.1	70.0	65.4

Table 2: Accuracy of Naïve Bayes on movie reviews.

Consistent with findings in the literature (Cui et al., 2006; Dave et al., 2003; Gamon and Aue, 2005), on the large corpus of movie review texts, the indomain-trained system based solely on unigrams had lower accuracy than the similar system trained on bigrams. But the trigrams fared slightly worse than bigrams. On sentences, however, we have observed an inverse pattern: unigrams performed better than bigrams and trigrams. These results highlight a special property of sentence-level annotation: greater sensitivity to sparseness of the model: On texts, classifier error on one particular sentiment marker is often compensated by a number of correctly identified other sentiment clues. Since sentences usually contain a much smaller number of sentiment clues than texts, sentence-level annotation more readily yields errors when a single sentiment clue is incorrectly identified or missed by the system. Due to lower frequency of higher-order n-grams (as opposed to unigrams), higher-order ngram language models are more sparse, which increases the probability of missing a particular sentiment marker in a sentence (Table  $3^3$ ). Very large

<sup>&</sup>lt;sup>1</sup>Recently, a similar task has been addressed by the Affective Text Task at SemEval-1 where even shorter units – headlines – were classified into positive, negative and neutral categories using a variety of techniques (Strapparava and Mihalcea, 2007).

<sup>&</sup>lt;sup>2</sup>All results are statistically significant at  $\alpha = 0.01$  with two exceptions: the difference between trigrams and bigrams for the system trained and tested on texts is statistically significant at alpha=0.1 and for the system trained on sentences and tested on texts is not statistically significant at  $\alpha = 0.01$ .

<sup>&</sup>lt;sup>3</sup>The results for movie reviews are lower than those reported in Table 2 since the dataset is 10 times smaller, which results in less accurate classification. The statistical significance of the

Dataset	Movie	News	Blogs	PRs		
Dataset size	1066	800	800	1200		
	unigr	ams				
SVM	68.5	61.5	63.85	76.9		
NB	60.2	59.5	60.5	74.25		
nb features	5410	4544	3615	2832		
	bigrams					
SVM	59.9	63.2	61.5	75.9		
NB	57.0	58.4	59.5	67.8		
nb features	16286	14633	15182	12951		
trigrams						
SVM	54.3	55.4	52.7	64.4		
NB	53.3	57.0	56.0	69.7		
nb features	20837	18738	19847	19132		

training sets are required to overcome this higher ngram sparseness in sentence-level annotation.

Table 3: Accuracy of unigram, bigram and trigram models across domains.

#### 4.2 System Performance on Different Domains

In the second set of experiments we sought to compare system results on sentences using in-domain and out-of-domain training. Table 4 shows that indomain training, as expected, consistently yields superior accuracy than out-of-domain training across all four datasets: movie reviews (Movies), news, blogs, and product reviews (PRs). The numbers for in-domain trained runs are highlighted in bold.

	Test Data			
Training Data	Movies	News	Blogs	PRs
Movies	68.5	55.2	53.2	60.7
News	55.0	61.5	56.25	57.4
Blogs	53.7	49.9	63.85	58.8
PRs	55.8	55.9	56.25	76.9

Table 4: Accuracy of SVM with unigram model

It is interesting to note that *on sentences*, regardless of the domain used in system training and regardless of the domain used in system testing, unigrams tend to perform better than higher-order ngrams. This observation suggests that, given the constraints on the size of the available training sets, unigram-based systems may be better suited for sentence-level sentiment annotation.

#### 5 Lexicon-Based Approach

The search for a base-learner that can produce greatest synergies with a classifier trained on small-set in-domain data has turned our attention to lexiconbased systems. Since the benefits from combining classifiers that always make similar decisions is minimal, the two (or more) base-learners should complement each other (Alpaydin, 2004). Since a system based on a fairly different learning approach is more likely to produce a different decision under a given set of circumstances, the diversity of approaches integrated in the ensemble of classifiers was expected to have a beneficial effect on the overall system performance.

A lexicon-based approach capitalizes on the fact that dictionaries, such as WordNet (Fellbaum, 1998), contain a comprehensive and domainindependent set of sentiment clues that exist in general English. A system trained on such general data, therefore, should be less sensitive to domain changes. This robustness, however is expected to come at some cost, since some domain-specific sentiment clues may not be covered in the dictionary. Our hypothesis was, therefore, that a lexiconbased system will perform worse than an in-domain trained classifier but possibly better than a classifier trained on out-of domain data.

One of the limitations of general lexicons and dictionaries, such as WordNet (Fellbaum, 1998), as training sets for sentiment tagging systems is that they contain only definitions of individual words and, hence, only unigrams could be effectively learned from dictionary entries. Since the structure of WordNet glosses is fairly different from that of other types of corpora, we developed a system that used the list of human-annotated adjectives from (Hatzivassiloglou and McKeown, 1997) as a seed list and then learned additional unigrams

results depends on the genre and size of the n-gram: on product reviews, all results are statistically significant at  $\alpha = 0.025$ level; on movie reviews, the difference between Na $\ddot{v}$ e Bayes and SVM is statistically significant at  $\alpha = 0.01$  but the significance diminishes as the size of the n-gram increases; on news, only bi-grams produce a statistically significant ( $\alpha = 0.01$ ) difference between the two machine learning methods, while on blogs the difference between SVMs and Na $\ddot{v}$ e Bayes is most pronounced when unigrams are used ( $\alpha = 0.025$ ).

from WordNet synsets and glosses with up to 88% accuracy, when evaluated against General Inquirer (Stone et al., 1966) (GI) on the intersection of our automatically acquired list with GI. In order to expand the list coverage for our experiments at the text and sentence levels, we then augmented the list by adding to it all the words annotated with "Positiv" or "Negativ" tags in GI, that were not picked up by the system. The resulting list of features contained 11,000 unigrams with the degree of membership in the category of positive or negative sentiment assigned to each of them.

In order to assign the membership score to each word, we did 58 system runs on unique nonintersecting seed lists drawn from manually annotated list of positive and negative adjectives from (Hatzivassiloglou and McKeown, 1997). The 58 runs were then collapsed into a single set of 7,813 unique words. For each word we computed a score by subtracting the total number of runs assigning this word a negative sentiment from the total of the runs that consider it positive. The resulting measure, termed Net Overlap Score (NOS), reflected the number of ties linking a given word with other sentimentladen words in WordNet, and hence, could be used as a measure of the words' centrality in the fuzzy category of sentiment. The NOSs were then normalized into the interval from -1 to +1 using a sigmoid fuzzy membership function (Zadeh, 1975)<sup>4</sup>. Only words with fuzzy membership degree not equal to zero were retained in the list. The resulting list contained 10,809 sentiment-bearing words of different parts of speech. The sentiment determination at the sentence and text level was then done by summing up the scores of all identified positive unigrams (NOS>0) and all negative unigrams (NOS<0) (Andreevskaia and Bergler, 2006).

# 5.1 Establishing a Baseline for the Lexicon-Based System (LBS)

The baseline performance of the Lexicon-Based System (LBS) described above is presented in Table 5, along with the performance results of the indomain- and out-of-domain-trained SVM classifier.

Table 5 confirms the predicted pattern: the LBS performs with lower accuracy than *in-domain*-

	Movies	News	Blogs	PRs
LBS	57.5	62.3	63.3	59.3
SVM in-dom.	68.5	61.5	63.85	76.9
SVM out-of-dom.	55.8	55.9	56.25	60.7

Table 5: System accuracy on best runs on sentences

trained corpus-based classifiers, and with similar or better accuracy than the corpus-based classifiers trained on *out-of-domain* data. Thus, the lexiconbased approach is characterized by a bounded but stable performance when the system is ported across domains. These performance characteristics of corpus-based and lexicon-based approaches prompt further investigation into the possibility to combine the portability of dictionary-trained systems with the accuracy of in-domain trained systems.

## 6 Integrating the Corpus-based and Dictionary-based Approaches

The strategy of integration of two or more systems in a single ensemble of classifiers has been actively used on different tasks within NLP. In sentiment tagging and related areas, Aue and Gamon (2005) demonstrated that combining classifiers can be a valuable tool in domain adaptation for sentiment analysis. In the ensemble of classifiers, they used a combination of nine SVM-based classifiers deployed to learn unigrams, bigrams, and trigrams on three different domains, while the fourth domain was used as an evaluation set. Using then an SVM meta-classifier trained on a small number of target domain examples to combine the nine base classifiers, they obtained a statistically significant improvement on out-of-domain texts from book reviews, knowledge-base feedback, and product support services survey data. No improvement occurred on movie reviews.

Pang and Lee (2004) applied two different classifiers to perform sentiment annotation in two sequential steps: the first classifier separated subjective (sentiment-laden) texts from objective (neutral) ones and then they used the second classifier to classify the subjective texts into positive and negative. Das and Chen (2004) used five classifiers to determine market sentiment on Yahoo! postings. Simple majority vote was applied to make decisions within

<sup>&</sup>lt;sup>4</sup>With coefficients:  $\alpha = 1$ ,  $\gamma = 15$ .

the ensemble of classifiers and achieved accuracy of 62% on ternary in-domain classification.

In this study we describe a system that attempts to combine the portability of a dictionary-trained system (LBS) with the accuracy of an in-domain trained corpus-based system (CBS). The selection of these two classifiers for this system, thus, was theorybased. The section that follows describes the classifier integration and presents the performance results of the system consisting of an ensemble CBS and LBS classifier and a precision-based vote weighting procedure.

#### 6.1 The Classifier Integration Procedure and System Evaluation

The comparative analysis of the corpus-based and lexicon-based systems described above revealed that the errors produced by CBS and LBS were to a great extent complementary (i.e., where one classifier makes an error, the other tends to give the correct answer). This provided further justification to the integration of corpus-based and lexicon-based approaches in a single system.

Table 6 below illustrates the complementarity of the performance CBS and LBS classifiers on the positive and negative categories. In this experiment, the corpus-based classifier was trained on 400 annotated product review sentences<sup>5</sup>. The two systems were then evaluated on a test set of another 400 product review sentences. The results reported in Table 6 are statistically significant at  $\alpha = 0.01$ .

	CBS	LBS
Precision positives	89.3%	69.3%
Precision negatives	55.5%	81.5%
Pos/Neg Precision	58.0%	72.1%

Table 6: Base-learners' precision and recall on product reviews on test data.

Table 6 shows that the corpus-based system has a very good precision on those sentences that it classifies as positive but makes a lot of errors on those sentences that it deems negative. At the same time, the lexicon-based system has low precision on positives and high precision on negatives<sup>6</sup>. Such complementary distribution of errors produced by the two systems was observed on different data sets from different domains, which suggests that the observed distribution pattern reflects the properties of each of the classifiers, rather than the specifics of the domain/genre.

In order to take advantage of the observed complementarity of the two systems, the following procedure was used. First, a small set of in-domain data was used to train the CBS system. Then both CBS and LBS systems were run separately on the same training set, and for each classifier, the precision measures were calculated separately for those sentences that the classifier considered positive and those it considered negative. The chance-level performance (50%) was then subtracted from the precision figures to ensure that the final weights reflect by how much the classifier's precision exceeds the chance level. The resulting chance-adjusted precision numbers of the two classifiers were then normalized, so that the weights of CBS and LBS classifiers sum up to 100% on positive and to 100% on negative sentences. These weights were then used to adjust the contribution of each classifier to the decision of the ensemble system. The choice of the weight applied to the classifier decision, thus, varied depending on whether the classifier scored a given sentence as positive or as negative. The resulting system was then tested on a separate test set of sentences<sup>7</sup>. The small-set training and evaluation experiments with the system were performed on different domains using 3-fold validation.

The experiments conducted with the Ensemble system were designed to explore system performance under conditions of limited availability of annotated data for classifier training. For this reason, the numbers reported for the corpus-based classifier do not reflect the full potential of machine learning approaches when sufficient in-domain training data is available. Table 7 presents the results of these experiments by domain/genre. The results

<sup>&</sup>lt;sup>5</sup>The small training set explains relatively low overall performance of the CBS system.

<sup>&</sup>lt;sup>6</sup>These results are consistent with an observation in (Kennedy and Inkpen, 2006), where a lexicon-based system performed with a better precision on negative than on positive texts.

<sup>&</sup>lt;sup>7</sup>The size of the test set varied in different experiments due to the availability of annotated data for a particular domain.

are statistically significant at  $\alpha = 0.01$ , except the runs on movie reviews where the difference between the LBS and Ensemble classifiers was significant at  $\alpha = 0.05$ .

		LBS	CBS	Ensemble
News	Acc	67.8	53.2	73.3
	F	0.82	0.71	0.85
Movies	Acc	54.5	53.5	62.1
	F	0.73	0.72	0.77
Blogs	Acc	61.2	51.1	70.9
	F	0.78	0.69	0.83
PRs	Acc	59.5	58.9	78.0
	F	0.77	0.75	0.88
Average	Acc	60.7	54.2	71.1
	F	0.77	0.72	0.83

Table 7: Performance of the ensemble classifier

Table 7 shows that the combination of two classifiers into an ensemble using the weighting technique described above leads to consistent improvement in system performance across all domains/genres. In the ensemble system, the average gain in accuracy across the four domains was 16.9% relative to CBS and 10.3% relative to LBS. Moreover, the gain in accuracy and precision was not offset by decreases in recall: the net gain in recall was 7.4% relative to CBS and 13.5% vs. LBS. The ensemble system on average reached 99.1% recall. The F-measure has increased from 0.77 and 0.72 for LBS and CBS classifiers respectively to 0.83 for the whole ensemble system.

#### 7 Discussion

The development of domain-independent sentiment determination systems poses a substantial challenge for researchers in NLP and artificial intelligence. The results presented in this study suggest that the integration of two fairly different classifier learning approaches in a single ensemble of classifiers can yield substantial gains in system performance on all measures. The most substantial gains occurred in recall, accuracy, and F-measure.

This study permits to highlight a set of factors that enable substantial performance gains with the ensemble of classifiers approach. Such gains are most likely when (1) the errors made by the classifiers are complementary, i.e., where one classifier makes an error, the other tends to give the correct answer, (2) the classifier errors are not fully random and occur more often in a certain segment (or category) of classifier results, and (3) there is a way for a system to identify that low-precision segment and reduce the weights of that classifier's results on that segment accordingly. The two classifiers used in this study – corpus-based and lexicon-based – provided an interesting illustration of potential performance gains associated with these three conditions. The use of precision of classifier results on the positives and negatives proved to be an effective technique for classifier vote weighting within the ensemble.

#### 8 Conclusion

This study contributes to the research on sentiment tagging, domain adaptation, and the development of ensembles of classifiers (1) by proposing a novel approach for sentiment determination at sentence level and delineating the conditions under which greatest synergies among combined classifiers can be achieved, (2) by describing a precision-based technique for assigning differential weights to classifier results on different categories identified by the classifier (i.e., categories of positive vs. negative sentences), and (3) by proposing a new method for sentiment annotation in situations where the annotated in-domain data is scarce and insufficient to ensure adequate performance of the corpus-based classifier, which still remains the preferred choice when large volumes of annotated data are available for system training.

Among the most promising directions for future research in the direction laid out in this paper is the deployment of more advanced classifiers and feature selection techniques that can further enhance the performance of the ensemble of classifiers. The precision-based vote weighting technique may prove to be effective also in situations, where more than two classifiers are integrated into a single system. We expect that these more advanced ensemble-ofclassifiers systems would inherit the benefits of multiple complementary approaches to sentiment annotation and will be able to achieve better and more stable accuracy on in-domain, as well as on out-ofdomain data.

#### References

- Ethem Alpaydin. 2004. *Introduction to Machine Learning*. The MIT Press, Cambridge, MA.
- Alina Andreevskaia and Sabine Bergler. 2006. Mining WordNet for a fuzzy sentiment: Sentiment tag extraction from WordNet glosses. In *Proceedings the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, IT.
- Anthony Aue and Michael Gamon. 2005. Customizing sentiment classifiers to new domains: a case study. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, Borovets, BG.
- Xue Bai, Rema Padman, and Edoardo Airoldi. 2005. On learning parsimonious models for extracting consumer opinions. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, Washington, DC.
- Hang Cui, Vibhu Mittal, and Mayur Datar. 2006. Comparative experiments on sentiment classification for online product reviews. In *Proceedings of the 21st International Conference on Artificial Intelligence*, Boston, MA.
- Kushal Dave, Steve Lawrence, and David M. Pennock. 2003. Mining the Peanut gallery: opinion extraction and semantic classification of product reviews. In *Proceedings of WWW03*, Budapest, HU.
- Mark Drezde, John Blitzer, and Fernando Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, CZ.
- Christiane Fellbaum, editor. 1998. WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA.
- Michael Gamon and Anthony Aue. 2005. Automatic identification of sentiment vocabulary: exploiting low association with known sentiment terms. In *Proceedings of the ACL-05 Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, Ann Arbor, US.
- Vasileios Hatzivassiloglou and Kathleen B. McKeown. 1997. Predicting the Semantic Orientation of Adjectives. In Proceedings of the the 40th Annual Meeting of the Association of Computational Linguistics.
- Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *KDD-04*, pages 168–177.
- Alistair Kennedy and Diana Inkpen. 2006. Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. *Computational Intelligence*, 22(2):110–125.
- Soo-Min Kim and Eduard Hovy. 2005. Automatic detection of opinion bearing words and sentences. In *Proceedings of the Second International Joint Conference*

on Natural Language Processing, Companion Volume, Jeju Island, KR.

- Bo Pang and Lilian Lee. 2004. A sentiment education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics*.
- Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43nd Meeting of the Association for Computational Linguistics*, Ann Arbor, US.
- Bo Pang, Lilian Lee, and Shrivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Conference on Empirical Methods in Natural Language Processing*.
- Jonathon Read. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL-2005 Student Research Workshop*, Ann Arbor, MI.
- Ellen Riloff, Siddharth Patwardhan, and Janyce Wiebe. 2006. Feature subsumption for opinion analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Sydney, AUS.
- P.J. Stone, D.C. Dumphy, M.S. Smith, and D.M. Ogilvie. 1966. *The General Inquirer: a computer approach to content analysis*. M.I.T. studies in comparative politics. M.I.T. Press, Cambridge, MA.
- Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 Task 14: Affective Text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, Prague, CZ.
- Songbo Tan, Gaowei Wu, Huifeng Tang, and Zueqi Cheng. 2007. A Novel Scheme for Domain-transfer Problem in the context of Sentiment Analysis. In *Proceedings of CIKM 2007*.
- Peter Turney and Michael Littman. 2003. Measuring praise and criticism: inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21:315–346.
- Peter Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics*.
- Janyce Wiebe, Rebecca Bruce, Matthew Bell, Melanie Martin, and Theresa Wilson. 2001. A corpus study of Evaluative and Speculative Language. In *Proceedings* of the 2nd ACL SIGDial Workshop on Discourse and Dialogue, Aalberg, DK.
- Lotfy A. Zadeh. 1975. Calculus of Fuzzy Restrictions. In L.A. Zadeh et al., editor, *Fuzzy Sets and their Applications to cognitive and decision processes*, pages 1–40. Academic Press Inc., New-York.