Originally published in: Jokinen, Kristiina/Bick, Eckhard (Eds.): Proceedings of the
17th Nordic Conference of Computational Linguistics (NODALIDA 2009), May 14-16, 2009, Odense, Denmark. -
Uppsala: NEALT, 2009. Pp. 198-205. (NEALT Proceedings Series 4)

Predictive Features in Semi-Supervised Learning for Polarity Classif cation and the Role of Adjectives

Michael Wiegand and **Dietrich Klakow**
Spoken Language Systems
Saarland University
D-66123 Saarbrücken, Germany
{Michael.Wiegand|Dietrich.Klakow}@lsv.uni-saarland.de

Abstract

In opinion mining, there has been only very little work investigating semi-supervised machine learning on document-level polarity classif cation. We show that semi-supervised learning performs signif cantly better than supervised learning when only few labeled data are available. Semi-supervised polarity classif ers rely on a predictive feature set. (Semi-)Manually built polarity lexicons are one option but they are expensive to obtain and do not necessarily work in an unknown domain. We show that extracting frequently occurring adjectives & adverbs of an unlabeled set of in-domain documents is an inexpensive alternative which works equally well throughout different domains.

1 Introduction

There has been an increasing interest in *opinion mining* in *natural language processing* in recent years. The highly interactive *Web 2.0* contains a huge amount of opinionated content. Advanced search engines and question answering systems should, therefore, be able to distinguish between factoid and opinionated content. Moreover, the classif cation of polarity in opinionated utterances or entire documents into positive and negative content, known as *polarity classif cation*, is another important functionality. This classif cation task, in particular, relies very much on *polar expressions*, i.e. key words indicating a specif c polarity.

In this paper we investigate *whether semi-supervised learning for document-level polarity classif cation works*, *what the best possible classif er is*, *what kind of feature set is most appropriate*, and, in particular, *how adjectives & adverbs perform as features*.

Semi-supervised learning is a class of machine learning methods that makes use of both labeled and unlabeled data for training, usually a small amount of labeled data and a large amount of unlabeled data. A classif er using unlabeled and labeled data can produce better performance than a classif er trained on the labeled data alone. Since labeled data are expensive to produce, semi-supervised learning is an inexpensive alternative to supervised learning.

The primary objective of our work is not to exceed the performance of supervised classif ers given a suff cient amount of labeled data as reported in previous research. Instead, we want to f nd out whether and how semi-supervised learning can produce better performance than supervised classif ers when only minimal amounts of labeled training data are available. Discriminative feature sets are far more important in this classif cation task than in supervised learning since there is less reliable information contained in small labeled datasets. We provide evidence that standard feature selection methods from semi-supervised topic classif cation (i.e. just using frequently occurring words) are not optimal for polarity classif cation. Polarity lexicons are an alternative option, however, they are expensive to create and their individual effectiveness may vary across different domains. We show that a small list of frequently occurring adjectives & adverbs cheaply extracted from an unlabeled in-domain dataset usually has competitive performance.

We consider polarity classif cation as a binary classif cation problem. That is, we assume that each document to be classif ed is subjective. We neglect the distinction between objective and subjective content since this classif cation is usually solved independently (Pang and Lee, 2004; Ng et al., 2006). Besides Ng et al. (2006) report that document-level subjectivity detection is a rather easy task compared to (binary) document-level po-

larity classif cation.

In our experiments, we primarily use the standard dataset from Pang et al. (2002) comprising movie reviews. To substantiate that our insights carry over to other domains, we also use a multi-domain dataset we created from *Rate-It-All*[1].

To the best of our knowledge, this is the f rst time that several semi-supervised classif ers are evaluated on this learning task in depth, in particular, in combination with various feature sets.

2 Related Work

Fully supervised polarity classif cation has been extensively explored. Both discriminative methods, such as *support vector machines (SVMs)*, and generative methods have been applied (Pang et al., 2002; Salvetti et al., 2006). Discriminative methods usually perform signif cantly better. If suff cient labeled data are available, supervised classif ers offer a reasonable performance even without dedicated feature selection. Various linguistic features, such as part-of-speech information, syntactic dependency information and semantic relations have been shown to increase performance of standard bag-of-words feature sets, (Ng et al., 2006; Gamon, 2004). However, Ng et al. (2006) report that the same improvement can be obtained by using higher order n-grams. We omit advanced linguistic features in this work, since, usually, the gain in performance hardly justif es the computational overhead of these methods (Gamon, 2004).

There are several *domain-independent* polarity lexicons containing important *polar expressions*. The most prominent manual lexicons are *General Inquirer*[2], the subjectivity lexicon from the *MPQA-project* (Wilson et al., 2005), and *Appraisal Groups* (Whitelaw et al., 2005). They have been successfully applied to polarity classif cation (Kennedy and Inkpen, 2005; Wilson et al., 2005; Whitelaw et al., 2005).

Moreover, several methods have been proposed to automatically induce polarity lexicons. Turney (2002) applies *Pointwise Mutual Information* in order to f nd similar words to a given list of polar seed words on web data. The polarity scores which are thus computed for each word can be used for a completely unsupervised classif cation algorithm of documents. A document is assigned the polarity derived from the average of the polarity scores of the words occurring within the document. The most recent semi-automatic lexicon is *SentiWordNet* (Esuli and Sebastiani, 2006) which assigns polarity to word senses in WordNet[3] known as *synsets*. The polarity of manually annotated seed synsets is expanded onto the remaining synsets of the WordNet ontology by measuring the overlap between their respective glosses.

The only works dealing with semi-supervised learning on this classif cation task we know of are Beineke et al. (2004) who combine Turney's web mining approach with evidence from labeled training data, and Aue and Gamon (2005) who focus on domain adaptation. Neither different algorithms nor feature sets are compared in these works.

In this paper, we look into adjectives & adverbs as features in detail. Pang et al. (2002) use feature sets exclusively comprising adjectives for supervised polarity classif cation but report performance to be worse than a standard bag-of-words representation. However, Ng et al. (2006) increase performance signif cantly by adding to a standard feature set higher order n-grams in which adjectives are replaced by their in-domain polarity which has been established via manual annotation.

3 Semi-Supervised Methods

Throughout the next sections, we adhere to the following notation: A document is denoted by $\vec{x}_i$. In total, there are N documents encompassing L labeled and U unlabeled documents. The label of an individual document $\vec{x}_i$ is $y_i \in \{-1, 1\}$. We tested three popular state-of-the-art semi-supervised classif ers in our experiments: *expectation maximization algorithm (EM)*, *transductive support vector machines (TSVMs)*, and *spectral graph transduction (SGT)*.

We use EM for a multinomial Naive Bayes classif er, similar to EM-λ proposed in Nigam et al. (2000). Since in all datasets we use the distribution of the classes is uniform, we omit the estimation of the class prior.

TSVMs use an extended objective function of SVMs: $OF_{tsvm} = \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=0}^{L}\xi_i + C^*\sum_{j=0}^{U}\xi_j^*$ which includes in addition to a weight vector $\vec{w}$, a regularizer C and a set of slack variables ξ_i for all labeled instances, an extra regularizer C^* and an extra set of *slack variables* ξ_j^*

[1] http://www.rateitall.com

[2] http://www.wjh.harvard.edu/~inquirer

[3] http://wordnet.princeton.edu

for unlabeled instances. A full account of the optimization is given in Joachims (1999).

In SGT (Joachims, 2003), all documents $\vec{x_i}$ of a collection (i.e. labeled and unlabeled) are represented as a symmetrized and similarity-weighted k nearest-neighbor (knn) graph G. Its adjacency matrix is defined as $A = A' + A'^T$ where

$$A'_{ij} = \begin{cases} \frac{sim(\vec{x_i},\vec{x_j})}{\sum_{\vec{x_k} \in knn(\vec{x_i})} sim(\vec{x_i},\vec{x_k})} & \text{if } \vec{x_j} \in knn(\vec{x_i}) \\ 0 & \text{else} \end{cases} \quad (1)$$

and $sim(\cdot,\cdot)$ is any common similarity function. The graph G is decomposed into its spectrum. For this, the smallest 2 to $d+1$ eigenvalues and eigenvectors of the normalized Laplacian $L = B^{-1}(B - A)$ where B is the diagonal degree matrix with $B_{ii} = \sum_j A_{ij}$ are computed. The spectrum is used for minimizing the normalized graph cut: $\min_{\forall y_i} \frac{cut(G^+,G^-)}{|\{i:y_i=1\}||\{i:y_i=-1\}|}$ where G^+ and G^- denote the set of positive and negative classified vertices in the graph. The cut-value $cut(G^+, G^-) = \sum_{i \in G^+} \sum_{j \in G^-} A_{ij}$ is the sum of the edge-weights of a cut partitioning the graph into two clusters.

4 The Different Feature Sets

The task of feature selection is to remove features that are irrelevant or noisy for a particular classification task. The reduction of these features does not only result in an increase in efficiency but may also improve the accuracy of a classifier.

4.1 Term Frequency Cut-off

The simplest feature selection method is using a term-frequency cut-off. The rationale behind this is that rarely observed terms do not contribute to a good classifier. Usually, this selection method is combined with stop-word removal[4]. Very frequently occurring terms, in particular function words, are not considered to be predictive for a particular class label, since they are uniformly distributed throughout all classes.

4.2 Polarity Lexicons

In our experiments we use Appraisal Groups (AG), General Inquirer (GI), the subjectivity lexicon from the *MPQA project* (MPQA), and SentiWordNet (SWN). From GI we use all polar expressions and from AG we only consider *orientation words* that are not neutral (Whitelaw et al., 2005). From MPQA, we use both *weak* and *strong* subjective words (Wilson et al., 2005) with either positive or negative prior polarity[5].

SentiWordNet (SWN) does not specify the polarity of individual words but synsets (i.e. senses of words). The database provides a non-negative polarity score $senseScore(s,p)$ for each synset s and polarity $p \in \{+,-\}$. Neutral polarity strength is denoted by 0. Usually, words have different senses associated with them. There are even words which have both senses with positive and negative polarity. Therefore, most words have various polarity scores associated with them. Our goal is to derive a unique polarity for each word with a corresponding score denoting its strength. We use the unique scores in order to find a subset of SWN with highly polar expressions. We estimate the strength of a word w and a polarity p, i.e. $wordScore(w,p)$, by: $wordScore(w,p) = \max_s [senseScore(s,p)]$ where $s \in synsets(w)$. The final polarity of the word, i.e. $pol(w)$, is the polarity with the maximum polarity score: $pol(w) = \arg\max_p [wordScore(w,p)]$. The unique score denoting the polarity strength is defined as: $strength(w) = \max_p [wordScore(w,p)]$. By using only the subset of SWN instead of the total (we chose all words with $strength(w) \geq 0.5$), we increased the accuracy of the semi-supervised classifiers by approximately 1.5% on average. We reduced the size of the initial version by 70% which substantially increased the efficiency of model learning. A subset of SWN based on taking the average rather than taking the maximum produced slightly worse results.

4.3 Adjectives & Adverbs

Adjectives, such as *superb* or *poor*, are usually regarded as very predictive words for polarity classification. The impact on semi-supervised learning has not yet been examined. Even if this feature set is too small for supervised learning (Pang et al., 2002; Salvetti et al., 2006), it might still be effective in semi-supervised learning. In contrast to supervised learning, large feature sets which are noisy cannot be compensated by the information contained in many labeled documents. Smaller

[4] We use a publicly available list of stopwords: `http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words`

[5] Note that just focusing on the strong entries resulted in a decrease in performance.

Feature Set	Type	#Words
Top n words	statistical selection	3000
Top n non-stopwords	statistical selection	2000
Top n adjectives & adverbs	stat. & linguistic select.	**600**
Appraisal Groups (AG)	manual polarity lexicon	2014
General Inquirer (GI)	manual polarity lexicon	2882
Subjectivity Lexicon (MPQA)	manual polarity lexicon	4615
SentiWordNet (SWN)	semi-automatic pol. lex.	11366

Table 1: Optimal size of the different feature sets.

but more predictive feature sets are preferable. We use feature sets of frequently occurring adjectives & adverbs in our document collection. The feature sets are extracted using C&C part-of-speech tagger[6]. After manually annotating the 600 most frequent stemmed adjectives & adverbs from the movie domain dataset (Pang et al., 2002), we estimate that more than 20% of the expressions are ambiguous with regard to part of speech[7]. Thus, our selection method if combined with stemming also captures some polar verbs and nouns. By looking at the list of extracted adjectives & adverbs from other domains, we observed that unlike current polarity lexicons this method allows both some colloquial expressions, such as *crappy*, and highly domain-dependent polar expressions, such as *creamy* or *crunchy* from the food domain, to be detected.

4.4 Optimal Feature Size

Table 1 lists the optimal size[8] of the different feature sets we used in our experiments[9]. Note that the subset selection for the polarity lexicons has been explained in Section 4.2. By far, the smallest feature set are adjectives & adverbs; the largest feature set is SWN.

5 Experiments

The results of *all* our experiments below are reported on the basis of 20 randomized partitionings. Each partitioning comprises a labeled dataset of varying length for training, and another dataset comprising 1000 documents used as unlabeled training data and test data[10]. We also experimented with larger amounts of unlabeled data but did not measure any improvement in performance. The labeled training data and the test data are always mutually exclusive. We report the results of experiments carried out on the movie review database (Pang et al., 2002) (benchmark dataset) and the results of cross-domain experiments using reviews from *Rate-It-All*. The movie dataset comprises 2000 reviews whereas for the other domains we could only acquire 1800 documents per domain. All datasets are balanced. We report statistical signif cance on the basis of a paired t-test using 0.05 as the signif cance level. We only state the results of the optimally sized feature sets (see Section 4.4). Since there is no difference in performance between the optimally sized feature set with the most frequent words and the most frequent non-stopwords, we only evaluated the latter feature set. We used *SVMLight*[11] for SVMs and TSVMs and *SGTLight*[12] for SGT. Feature vectors consist of tf-idf weighted words appearing in the pre-def ned feature set normalized by document length. This produced best results throughout our experiments. Further modif cations of the standard conf guration of *SVMLight* (e.g. changing regularization parameters) did not improve performance. We also conf rm the results from Aue and Gamon (2005) where further modif cations on EM, i.e. by weighting the unlabeled data[13], did not improve performance. For *SGTLight* we mainly adhered to the standard conf guration (as discussed in Joachims (2003)). Since we had no development data for optimizing the only task-sensitive parameter k we simply took the optimized value for the only text classif cation corpus tested in Joachims (2003) (i.e. *Reuters collection*). The current choice (i.e. $k = 800$) should thus guarantee a fairly unbiased setting. EM is smoothed by absolute discounting (Zhai and Lafferty, 2001). All classif ers are run with a reasonable parameter setting but we did not attempt to tune the parameters to the current task. We also stem the entire text since some polarity lexicons we use also include lemmas of inf ectional words,

[6] `http://svn.ask.it.usyd.edu.au/trac/candc`

[7] e.g. *interesting* (adj) and *interests* (noun) are both reduced to *interest*

[8] The optimal size was determined by testing all semi-supervised algorithms trained on various amounts of labeled documents and 1000 unlabeled documents.

[9] Due to the stemming we applied some of the entries in the original polarity lexicons were conf ated.

[10] It is not uncommon to use test data as unlabeled training data in semi-supervised learning (Aue and Gamon, 2005; Joachims, 1999; Joachims, 2003).

[11] `http://svmlight.joachims.org`

[12] `http://sgt.joachims.org`

[13] Note that this is similar to regularization in TSVMs.

SWN	AG	GI	MPQA	GI+Turney
54.20	54.45	59.90	61.95	63.30

Table 2: Accuracy of unsupervised algorithm using different polarity lexicons (movie domain): *best classif er is GI+Turney.*

such as nouns and verbs. Moreover, stemming has considerable advantages for the feature set comprising adjectives & adverbs (see discussion in Section 4.3). In-domain feature sets (i.e. frequent non-stopwords and frequent adjectives & adverbs) are obtained by considering the entire dataset of a particular domain.

5.1 Experiments on the Movie Domain

5.1.1 Unsupervised Algorithms using Different Polarity Lexicons

Before comparing the different polarity lexicons in the context of semi-supervised learning, we shortly display their performance using a completely unsupervised algorithm. A test document is assigned the polarity with the majority of polar expressions in that document. This experiment should give an idea of the intrinsic predictiveness of the polarity lexicons. Table 2 lists the results. Though all lexicons perform signif cantly better than the random baseline (i.e. 50%), the best performance of MPQA with 61.95 is still very low.

We also evaluated an extension GI+Turney which weights the polar expressions in GI according to the association scores to a very small number of manually selected highly polar seed words, such as *excellent* or *poor* (Turney and Littman, 2003)[14]. The scores for entries in GI are calculated in the same way as the scores for words in the web-based lexicon induction method using *Pointwise Mutual Information* (Turney, 2002). The improvement is signif cant, even though the scores have been gained by domain-independent web-data.

In the following, we show that very small amounts of labeled in-domain documents can produce signif cantly better results using semi-supervised learning.

5.1.2 Comparison of the Different Polarity Lexicons with Other Feature Sets

Table 3 displays the performance of different classif ers on different feature sets. On average, polarity lexicons perform signif cantly better than the top 2000 non-stopwords. The same holds for an inexpensive small feature set of in-domain adjectives & adverbs. On EM, we achieved even the best performance with the latter feature set. The best performing feature set for the movie dataset is AG. With the exception of EM, it is signif - cantly better than any other feature set using semi-supervised learning.

5.1.3 Complex Feature Sets that Do Not Improve Performance

Contrary to our expectations, adding explicit polarity information to the feature set by including the number of positive and negative polar expressions according to the pertaining polarity lexicon did not improve performance. We assume that the meaning of these polar expressions, occasionally even their polarity, varies across different contexts, therefore a unique polarity in the polarity lexicons may not always be correct.

We also experimented with more expressive features by adding bigrams with one token being either a polar expression, an adjective or an adverb. On semi-supervised learning we did not measure any increase in performance. We assume that this is due to data-sparseness. Similar to Ng et al. (2006), we observed an increase in performance by approximately 2% on supervised classif ers (when more than 400 labeled documents are used).

5.1.4 Semi-Supervised Classif ers

We compared all different learning algorithms using their respective best feature sets. Figure 1 displays the results. All semi-supervised algorithms are better than the strict supervised baseline (i.e. SVMs trained on AG) on small amounts of labeled data. EM gets worse than SVMs trained on AG when more than 400 labeled documents are used, but still outperforms SVMs trained on top 2000 non-stopwords when less than 700 labeled documents are used. TSVMs and SGT, on the other hand, constantly perform better than SVMs. Clearly, the best classif er is SGT which, with the exception of 1000 labeled data, is always signif- icantly better than any other classif er tested. At approximately 200 labeled documents, SGT already performs as well as SVMs trained on a standard feature set (i.e. top 2000 non-stopwords) using 1000 labeled documents. The best supervised performance at 80.6% is similar to the one pre-

[14]Unfortunately, currently only the weights for entries of GI were available to us.

	20 Labeled Documents						200 Labeled Documents					
	Top 2000	**SWN**	**MPQA**	**GI**	**AG**	**Adj**	**Top 2000**	**SWN**	**MPQA**	**GI**	**AG**	**Adj**
SVM	59.81	61.24	*63.07*	61.48	62.22	61.44	72.05	74.93	74.35	72.72	*75.88*	73.14
EM	67.50	67.31	68.73	66.63	69.44	*69.54*	73.44	76.46	75.02	73.80	75.46	*77.32*
TSVM	64.57	67.04	66.58	65.53	*68.87*	68.37	73.48	76.80	75.73	74.72	*77.89*	75.12
SGT	62.60	67.39	67.10	66.14	***70.28***	66.58	70.91	77.55	77.78	75.12	***80.21***	76.90

Table 3: Accuracy of different classif ers on different feature sets using 20 and 200 labeled documents (movie domain): *best conf guration is SGT+AG.*

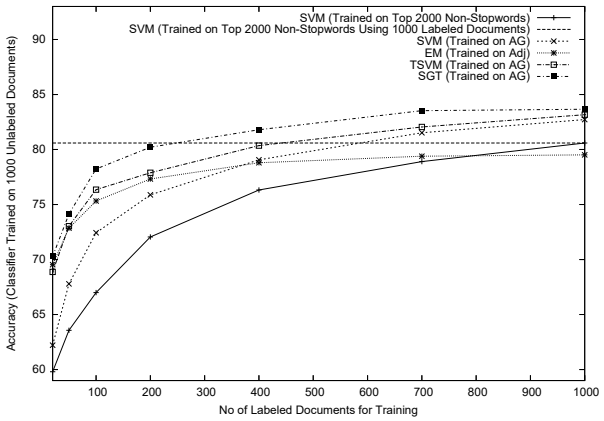

Figure 1: Performance of different learning algorithms on the best respective feature set (movie domain): *SGT+AG save 800 labeled documents in comparison to SVM+Top 2000 trained on 1000 labeled documents.*

sented in Pang et al. (2002). They report 81.4% with their most similar conf guration using 1400 labeled documents and training on 2633 words. Just using 20 labeled documents offers an increase by 7% in performance in comparison to the best unsupervised classif er (i.e. GI+Turney displayed in Table 2).

5.2 Cross-Domain Experiments

In order to validate our f ndings from Section 5.1, we extracted reviews from *Rate-It-All*. In particular, we want to know whether semi-supervised learning works there as well, whether SGT outperforms other classif ers, whether polarity lexicons improve performance, and whether adjectives and adverbs produce classif ers competitive to average polarity lexicons. We do not attempt to carry out detailed domain studies which would be beyond the scope of this section. We chose four domains from the list of *Topic Categories* of the website which we thought are very different from the movie domain and for which we could extract suff cient training data. We took *Computer & Internet (computer)*, *Products (products)*, *Sports & Recreation (sports)* and *Travel, Food, & Culture (travel)*. We follow the method from Blitzer et al. (2007) to infer the polarity of the reviews. Ratings with less than 3 stars are considered negative reviews whereas ratings with more than 3 stars are positive reviews. We decided not to consider *mixed* reviews, i.e. reviews rated with 3 stars. In general, we found far fewer mixed reviews[15]. On those domains which provided a reasonable amount of data, our initial supervised learning experiments showed that mixed polarity can only be poorly distinguished from def nite polarity[16]. Manual inspection of a random sample of reviews also showed that a great part of these documents are actually negative reviews. We only extracted reviews having at least 3 sentences in order to rule out too fragmentary instances. We did not f lter out mislabeled entries though we are aware of their presence in our set.

Table 4 lists the average performance of all classif ers on different feature sets using 20 labeled documents. For the sake of completeness we also include the results from the movie domain. There is no signif cant difference among the feature sets using SVMs, but there is a difference between top 2000 non-stopwords and the remaining feature sets on semi-supervised classif cation (with the exception of EM). All polarity lexicons and adjectives & adverbs perform signif cantly better than top 2000 non-stopwords using TSVMs and SGT. On average, the performance of EM is signif cantly worse than any of the other semi-supervised classif ers. The results of TSVMs

[15] In the *computer* domain, for example, there were only approximately 200 reviews.

[16] A binary classif er trained on 900 mixed and 900 def nite polar reviews from the *travel* domain only produced an accuracy of 63.1% on a three fold crossvalidation and the best feature set.

are similar with our previous observations on the benchmark dataset. SGT is the best performing classifier (in particular in combination with adjectives).

Table 5 shows the performance on the individual domains and feature sets using 20 labeled documents on SGT. On average, semi-supervised learning improves performance significantly over supervised learning. On some domains (e.g. *computer*) using a standard feature set (i.e. using top 2000 non-stopwords in the collection) produces good results. However, in some other domains, such as *travel*, there is no improvement whatsoever. Polarity lexicons can perform significantly better than top 2000 non-stopwords (e.g. GI on *travel* or, most notably, AG on *movie*) but there can also be a domain where they are actually worse than the standard feature set (e.g. the *sports* domain). There is no polarity lexicon which consistently outperforms all other polarity lexicons on all domains. A feature set comprising in-domain adjectives & adverbs, however, is more robust: Firstly, it never performs worse than the standard feature set. Secondly, it is never significantly worse than the average performance of polarity lexicons and, thirdly, there might be some domain, such as *sports*, where it significantly outperforms any other feature set. Considering the low effort to generate such a feature set should make it particularly attractive.

Figure 2 displays the performance of SGT on various feature sets averaged over all domains using various amounts of labeled training data. SGT only significantly outperforms SVMs when less than 200 labeled documents are used. Therefore, we restricted the figure to the range ending at that size. The lower performance of the averaged results must be due to some properties of the *Rate-It-All* data (either noise or the dataset is more difficult) since the individual performance of the semi-supervised classifiers on the movie domain was significantly better. Despite the lower performance, we can still use the averaged results to characterize the relation between the different feature sets in semi-supervised learning. Both polarity lexicons and adjectives & adverbs are significantly better than top 2000 non-stopwords and there is no significant difference between polarity lexicons and adjectives & adverbs.

All these results support both the competitiveness of adjective & adverbs and the robustness of SGT. Given the best feature set in a particular domain, the average gain in improvement compared to SVMs only trained on 20 labeled documents using top 2000 non-stopwords is approx. 8.5% when SGT is used. This is a clear indication that semi-supervised learning for polarity classification works across all domains when only tiny amounts of labeled data are used.

	Top 2000	SWN	MPQA	GI	AG	Adj
SVM	**61.17**	61.13	60.81	**61.17**	60.77	60.68
EM	64.41	65.09	64.08	63.88	65.10	**65.22**
TSVM	63.87	66.79	66.51	66.26	65.98	**67.20**
SGT	64.60	66.92	67.69	67.83	67.22	**68.30**

Table 4: Average accuracy of different semi-supervised classifiers across all domains using different feature sets (trained on 20 labeled documents & 1000 unlabeled documents): *best configuration is SGT+Adj.*

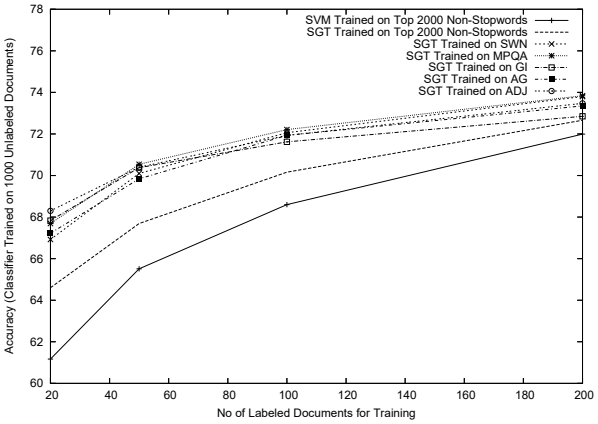

Figure 2: SGT trained on different amounts of labeled data and different feature sets averaged over all domains (1000 unlabeled documents): *polarity lexicons and Adj are very similar among each other and significantly better than top 2000 non-stopwords.*

6 Conclusion

In this paper we have shown that semi-supervised learning can be successfully applied to document-level polarity classification. Significant improvement over supervised classification can be achieved across all domains when less than 200 labeled documents are available. On the movie domain we even achieved improved performance

	SVM	SGT					
Domain	**Top 2000**	**Top 2000**	**SWN**	**MPQA**	**GI**	**AG**	**Adj**
computer	67.75	73.88	*75.77*	74.77	73.95	73.74	74.51
products	62.38	67.20	68.45	68.40	*69.84*	68.44	68.79
sports	57.96	61.83	57.57	59.80	60.62	58.53	*63.55*
travel	57.95	57.48	65.44	68.37	*68.62*	65.09	68.05
movies	59.81	62.60	67.39	67.10	66.14	*70.28*	66.58
average	61.17	64.60	66.92	67.69	67.83	67.22	**68.30**

Table 5: Accuracy of SGT on different domains using different feature sets (trained on 20 labeled documents & 1000 unlabeled documents): *on an individual domain either some polarity lexicon or Adj is the best feature set; on average Adj is the best feature set.*

across all amounts of labeled training data. SGT is the classif er which produces signif cantly better results than all other semi-supervised classif ers used in our experiments. On average, polarity lexicons and adjectives & adverbs perform better than just using frequent in-domain non-stopwords. Adjectives & adverbs are less expensive to obtain and more robust throughout different domains.

Acknowledgements

The authors would like to thank Grzegorz Chrupała, Sabrina Wilske and Theresa Wilson for interesting discussions. We, in particular, thank Stefan Kazalski for pre-processing the web documents from *Rate-It-All*. Michael Wiegand was funded by the German research council DFG through the International Research Training Group between Saarland University and University of Edinburgh.

References

A. Aue and M. Gamon. 2005. Customizing Sentiment Classif ers to New Domains: a Case Study. In *Proc. of RANLP*.

P. Beineke, T. Hastie, and S. Vaithyanathan. 2004. The Sentimental Factor: Improving Review Classif cation via Human-Provided Information. In *Proc. of ACL*.

J. Blitzer, M. Dredze, and F. Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classif cation. In *Proc. of ACL*.

A. Esuli and F. Sebastiani. 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In *Proc. of LREC*.

M. Gamon. 2004. Sentiment Classif cation on Customer Feedback Data: Noisy Data, Large Feature Vectors, and the Role of Linguistic Analysis. In *Proc. of COLING*.

T. Joachims. 1999. Transductive Inference for Text Classif cation Using Support Vector Machines. In *Proc. of ICML*.

T. Joachims. 2003. Transductive Learning via Spectral Graph Partitioning. In *Proc. of ICML*.

A. Kennedy and D. Inkpen. 2005. Sentiment Classif cation of Movie Reviews Using Contextual Valence Shifters. In *Workshop on the Analysis of Formal and Informal Information Exchang during Negotiations*.

V. Ng, S. Dasgupta, and S. M. Niaz Arif n. 2006. Examining the Role of Linguistic Knowledge Sources in the Automatic Identif cation and Classif cation of Reviews. In *Proc. of ACL*.

K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. 2000. Text Classif cation from Labeled and Unlabeled Documents Using EM. *Machine Learning*.

B. Pang and L. Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Miminum Cuts. In *Proc. of ACL*.

B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up? Sentiment Classif cation Using Machine Learning Techniques. In *Proc. of EMNLP*.

F. Salvetti, C. Reichenbach, and S. Lewis. 2006. Opinion Polarity Identif cation of Movie Reviews. In J. Shanahan, Y. Qu, and J. Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Applications*. Springer-Verlag.

P. Turney and M. Littman. 2003. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. In *Proc. of TOIS*.

P. Turney. 2002. Thumbs up or Thumbs down?: Semantic Orientation Applied to Unsupervised Classif cation of Reviews. In *Proc. of ACL*.

C. Whitelaw, N. Garg, and S. Argamon. 2005. Using Appraisal Groups for Sentiment Analysis. In *Proc. of CIKM*.

T. Wilson, J. Wiebe, and P. Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proc. of HLT/EMNLP*.

C. Zhai and J. Lafferty. 2001. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. In *Proc. of SIGIR*.