Revisiting The Old Kitchen Sink: Do We Need Sentiment Domain Adaptation?

Riham Hassan Mansour[1]
rihamma@microsoft.com

Nesma Refaei[3]
nesma.a.refaei@eng.cu.edu.eg

Michael Gamon[2]
mgamon@microsoft.com

Khaled Sami[1]
t-ksami@microsoft.com

Ahmed Abdel-Hamid[1]
ahmedab@microsoft.com

[1]Microsoft Research, 306 Maadi Courniche, Basatin, Cairo, Egypt
[2]Microsoft Research, One Microsoft Way, Redmond, WA 98051, USA
[3]Computer Engineering, Cairo University/ Gamaet El Qahera St., Giza, Egypt

Abstract

In this paper we undertake a large cross-domain investigation of sentiment domain adaptation, challenging the practical necessity of sentiment domain adaptation algorithms. We first show that across a wide set of domains, a simple "all-in-one" classifier that utilizes all available training data from all but the target domain tends to outperform published domain adaptation methods. A very simple ensemble classifier also performs well in these scenarios. Combined with the fact that labeled data nowadays is inexpensive to come by, the "kitchen sink" approach, while technically non-glamorous, might be perfectly adequate in practice. We also show that the common anecdotal evidence for sentiment terms that "flip" polarity across domains is not borne out empirically.

1 Introduction

Automatic detection and analysis of sentiment around products, brands, political issues etc. has triggered a large amount of research in the past 15 – 20 years (for a recent overview see Pang & Lee 2008 and Liu 2012). Early work focused on algorithms for mining sentiment dictionaries (Hatzivassiloglou and McKeown 1997, Turney 2002); this was followed by the exploration of supervised techniques (Pang et al. 2002) and, somewhat more recently, by investigations of domain adaptation techniques. Also more recently, the focus has broadened from the detection of polarity (negative/positive sentiment) to more nuanced approaches that try to identify targets and holders of sentiment, sentiment strength, or finer-grained mood distinctions (e.g. Wilson et al. 2006, Kim and Hovy 2006). Within the polarity detection paradigm, a number of common assumptions have been shared in the community and are frequently repeated in the literature. Two of these fundamental assumptions are:

1. Obtaining sufficient labeled data for supervised training is expensive
2. Sentiment models trained on one domain tend to perform poorly on new, unseen domains

A conclusion that is often drawn from these assumptions is that domain adaptation of sentiment models from a domain with sufficient labeled data to a new domain with little labeled data is an important problem and requires new and sophisticated algorithms.

In this paper, we empirically re-examine the assumptions above. Based on a wide range of experiments on 27 different domains, we challenge the conclusion that domain adaptation for polarity detection necessarily requires novel and sophisticated machinery. It is important to keep in mind, however, that our claims are strictly limited to the problem under investigation, namely polarity detection. We do not make any claims whatsoever about domain adaptation for other sentiment-related problems or general problems in machine learning. Based on readily available data from 27 domains, we show that a "kitchen sink" approach where all source domain data are combined to train a single classifier sets a surprisingly high baseline for polarity identification accuracy across domains. We also show on a previously released data set of four domains that the result is competitive with a state-of-the-art domain adaptation approach using Structural

Proceedings of Recent Advances in Natural Language Processing, pages 420–427,
Hissar, Bulgaria, 7-13 September 2013.

Correspondence Learning. We then show that a straightforward ensemble learner can, for some domains, improve results further, without any need for specialized learning algorithms. Since most work in domain-adaptation only provides published results on pairwise adaptation between domains and not on multi-domain adaptation, we hope to establish a new baseline for future adaptation techniques to compare against.

2 Related Work

Of direct importance to the discussion in this paper are results from domain adaptation in polarity detection. One of the earlier successful approaches (Blitzer et al. 2006, 2007) involved Structural Correspondence Learning (SCL). SCL identifies "pivot" features that are both highly discriminative in the labeled source domain data and also frequent in the unlabeled target domain data. In a subsequent step, linear predictors for the pivot terms are learned from the unlabeled target data and from the source data.

Daumé (2007) approached domain adaptation from a fully labeled source domain to a partially labeled target domain by augmenting the feature space. Instead of using a single, general, feature set for source and target, three distinct feature sets are created: the general set of features, a source-domain specific version of the feature set, and a target-specific version of the feature set.

Li and Zong (NLP-KE 2008) explore a classifier combination technique they call "Multiple-Label Consensus Training" which results in better accuracy than non-adapted models on the data sets used in Blitzer et al. (2007). They also addressed the multi-domain sentiment analysis problem using feature –level fusion and classifier-level fusion approaches in Li and Zong (ACL 2008).

Dredze and Crammer (2008) have proposed a multi-domain online learning framework based on parameter combination from multiple Confidence Weighted (CW) classifiers. Their Multi-Domain Regularization (MDR) framework seeks to learn domain specific parameters guided by the shared parameter across domains.

Samdani and Yih (2011) propose an ensemble learner that consists of classifiers trained on different feature groups. The feature groups are identified based on how stable the feature distribution is across domains, which can either be estimated from the data directly or can be hypothesized based on domain knowledge.

Chen et al. (2011) use a specific co-training algorithm for domain adaptation on the Blitzer et al. (2007) data set. In averaged pair-wise comparisons they establish gains over a source-plus-target logistic regression baseline.

Glorot et al. (2011) investigate a deep learning approach to domain adaptation and report increased accuracy across domains both on the Blitzer et al. (2007) 4-domain data set and the larger Amazon review data set (25 domains) also made available in that release. They also introduce a new metric for transfer learning: *Transfer Ratio*.

3 Datasets & Experimental Setup

This section illustrates the datasets, the methods and the setup of our experiments.

3.1 Datasets

The datasets we used in our experiments have been obtained from three sources:

1. Amazon reviews[1]: this dataset contains more than 5.8 million reviews. It has been used in previous work on sentiment analysis (see Glorot et al. (2011)). The Amazon reviews include 25 domains as shown in Table 1.
2. Hotel reviews[2]: this dataset includes full reviews of hotels in 10 different cities (Dubai, Beijing, London, New York City, New Delhi, San Francisco, Shanghai, Montreal, Las Vegas, Chicago). There are about 80-700 hotels in each city. The extracted fields include date, review title and the full review. The total number of reviews is 259,000.
3. Twitter: this dataset has been obtained and annotated in Choudhury et al. (AAAI 2012) over a 1 year period of time from Nov. 1, 2010 to Oct. 31, 2011. The dataset has been originally annotated for affects. We mapped the positive affects "joviality" and "serenity" to positive sentiment and the negative affects "fatigue", "hostility", and "sadness" to negative sentiment. We selected a balanced dataset of 2,000 tweets from the various months of the collected tweets.

The average review length for the Amazon and hotel reviews is 437 characters and 97 words. In total, we used 27 domains namely the

[1] Amazon reviews could be obtained at http://liu.cs.uic.edu/download/data/

[2] Hotel reviews could be obtained at http://mlr.cs.umass.edu/ml/datasets/OpinRank+Revie w+Dataset

25 Amazon domains, the hotel domain and the Twitter domain. We considered Twitter as a domain though the content of tweets spans multiple domains since it has different characteristics from the product reviews. Tweets are constrained to 140 characters each and lack context.

Domain	Dataset Size	Labeled Data Size	Domain	Dataset Size	Labeled Data Size	Domain	Dataset Size	Labeled Data Size
Apparel	9252	2000	Kitchen & housewares	19856	2000	Electronics	23009	2000
Automotive	736	304	Magazines	4191	1940	Gourmet food	1575	416
Baby	4256	1800	Music	174180	2000	Grocery	2632	704
Beauty	2884	986	Musical instruments	332	96	Health & personal care	7225	2000
Books	975194	2000	Office products	431	128	Jewelry & watches	1981	584
Camera & photo	7408	1998	Outdoor living	1599	654	Toys & games	13147	2000
Cell phones & service	1023	768	Software	2390	1830	Video	36180	2000
Computer & video games	2771	916	Sports & outdoors	5728	2000	Hotel	259,000	2000
Dvd	124438	2000	Tools & hardware	112	28	Tweets	1,107,282	2000

Table 1: Dataset sizes of the 27 Domains.

The Amazon reviews and the hotel reviews are rated between 1 and 5 on a 5 point scale where 1 is the most negative and 5 is the most positive. We have extracted only the reviews that are rated 5 and 1 to represent the positive and negative reviews respectively. Further, we ensured that the datasets we extracted and used for training are balanced between positive and negative reviews. Table 1 summarizes the 27 domains and their dataset sizes including the balanced datasets we used for training.

3.2 Experimental Setup

In our experiments, we employed the datasets of the 27 domains mentioned in section 3.1. In each experiment, we have employed one domain for testing while the other 26 domains have been used for training. We compared four domain adaptation techniques:

1. One classifier trained in all source domains.
2. An ensemble of classifiers, each trained on a source domain, combined into an ensemble.
3. The domain adaptation approach proposed in Daumé (2007).
4. We also compared the results of approaches 1 and 2 to published results on Structural Correspondence Learning (SCL) by using the same datasets as in Blitzer et al. (2007).

In all our experiments, we employed Maximum Entropy-based classification with vanilla parameter settings and feature reduction using log-likelihood ratio (LLR). Further, we used the accuracy metric to indicate the performance of each of the above four domain adaptation techniques. We also employed the Transfer Ratio metric proposed by Glorot et al. (2011) to measure the performance of the all-in-one and ensemble classifiers. The rest of the subsection illustrates the experimental setup for each of the above four approaches.

In-domain Classifiers

To establish a "ceiling" performance we built an in-domain classifier for each of the 27 domains. The in-domain classifier is trained with a dataset of that one domain and tested on the same domain (using cross-validation). This standard in-domain supervised setup establishes an upper bound for classification performance (although in some cases we will see that other techniques can outperform this upper bound). Features consist of binary unigram and bigram features. On average, the total number of features in each domain is 52,039. Feature reduction was performed using LLR, retaining only the top 20,000 most predictive features as established on the training set.

We compare the results obtained from testing each domain with the three approaches to its in-domain classifier results.

All-in-one Classifier

The all-in-one classifier is a maximum entropy classifier trained with the source domain datasets merged together. In this setting, the classifier is trained with data from multiple domains, which exposes it to multiple sentiment vocabularies at training time, creating a somewhat domain-

independent and general model. The all-in-one classifier is trained with 26 domains datasets while being tested on the held-out 27th domain.

Ensemble Classifiers

One approach to address the problem of domain adaptation is to construct an ensemble of classifiers, all of which contribute partially to the final result (see Dietterich (1997) for an overview). We constructed an ensemble of in-domain sentiment classifiers, one for each source domain. There are various techniques to combine the contribution of each classifier in the ensemble. We employed three techniques in our experiment settings:

1. Majority vote: the results are obtained by taking the majority of votes from the multiple classifiers in the ensemble. For example, if 20 classifiers vote positive and only 6 classifiers vote negative, the final result is positive
2. Sum of weights: the results are obtained by summing up the class probabilities from each classifier.
3. Meta-classification: the results are obtained by combining the weight of each classifier's vote in a meta-classifier. The meta-classifier weights are learned through a machine learning model trained on a small labeled set of data from the target domain. We used both logistic regression and SVM to train the meta-classifier. We have experimented with multiple sizes of labeled target data ranging from 5 positive and 5 negative meta-training examples to 50 positive and 50 negative examples. The following steps are used to train the meta-classifier.
    a) For each review *r* in the set of labeled data in target domain D that is used to train the meta-classifier; we create a vector V consisting of the vote of each source-domain classifier on *r* and the label of r.
    b) We construct a matrix M of the set of vectors Vs created in step 1.
    c) We employ either logistic regression or SVM. We have used SVM^{light} implementation[3] to train the ensemble using SVM with the matrix M and a radial basis kernel function.

Hal Daumé's Domain-Adaptation Approach

Daumé (2007) addresses domain adaptation where a large, annotated corpus of data from the source domain is available with only a small, annotated corpus of the target domain. Daumé's work leverages both annotated datasets to obtain a model that performs well on the target domain. For K source domains, the augmented feature space consists of K+1 copies of the original feature space. However, creating three versions of each feature in both the source and the target domains grows the feature space exponentially, which is prohibitive in a many-domain adaptation scenario such as ours which consists of a total of 27 domains.

We addressed this challenge by considering the 26 source domains as a single source domain being adapted to the target domain. This setup along with feature reduction enabled us to apply Daumé's approach without too much of an inflation of the feature space. However, we also recognize that this likely compromises the power of the feature augmentation approach.

Blitzer's Structural Correspondence Learning

Blitzer et al. (2007) employ the Structural Correspondence Learning (SCL) algorithm for sentiment domain adaptation. Blitzer et al. evaluate the SCL domain adaptation on four publicly released datasets from Amazon product reviews: books, DVDs, electronics and kitchen appliances. In these four datasets, reviews with rating > 3 were labeled positive, those with rating < 3 were labeled negative, and the rest discarded because their polarity was ambiguous. 1000 positive and 1000 negative labeled examples were used for each domain. Some unlabeled data were additionally used including 3685 (DVDs) and 5945 (kitchen). Each labeled dataset was split into 1600 instances for training and 400 instances for testing. The baseline in Blitzer et al. (2007) is a linear classifier trained without adaptation, while their ceiling reference is the same as ours, which is the in-domain classifier trained and tested on the same domain.

We conducted a set of experiments employing the four datasets used for SCL domain adaptation. In these experiments, we compare the results of our all-in-one classifier and the ensemble classifier trained and tested on the four datasets to the results of SCL and its variation SCL-MI domain adaptation as reported by Blitzer et al. (2007) on the same datasets. We employ the same training and test split size for cross-validation of the SCL domain adaptation approach. Further, we replicated both the approach

[3] Implementation of SVM^{light} : http://svmlight.joachims.org/

baseline and ceiling in-domain classifiers for the four domains.

4 Results & Discussion

This section summarizes the results of the experiments described in section 3.2 while further scrutinizing the comparison between the four domain adaptation sentiment analysis techniques. We also report the Transfer Ratio results of the all-in-one and ensemble classifiers. Generally, the all-in-one classifier is closely comparable to the in-domain classifier of each domain

4.1 Results

In this section, we summarize the various results obtained from the set of experiments described in section 3.2. In the summary of each experiment results, we also plot the in-domain classifier results of each domain as the ceiling of comparison.

All-in-one Classifier Experiments

In the all-in-one classifier experiments, the sentiment classifier is trained with 26 domain datasets while testing it with the 27th domain. Table 3 summarizes the results. The results of the all-in-one classifier are very close to the in-domain classifiers in most domains except for the apparel, beauty, magazines, outdoor living, office products and software.

Ensemble Classifier Experiments

We produced the results of the ensemble of classifiers using the three settings: majority votes, sum of weights, and meta-classification using both logistic regression and SVM. Table 3 summarizes the results of the three settings used in the ensemble.

Table 3 shows that the ensembles with sum of weights and meta-training (SVM sigmoid kernel) are the most comparable to the in-domain classifier of each domain. We also experimented with variations of logistic regression and SVM for meta-training. The non-linear (RBF kernel) SVM meta-classifier outperforms the linear logistic regression model. We have employed two variations of SVM, namely, a radial basis function with gamma 0.01 and sigmoid kernel. In most domains, the SVM model trained with 50 positive and 50 negative feedback examples is not far off the one trained with 5 positive and 5 negative feedback examples. This shows that even with little labeled data in the target domain, the ensemble could effectively combine the weights of the classifier votes. We expect the ensemble to achieve steady but slow performance gains over time while collecting more feedback examples.

Hal Daumé's Domain-Adaptation Approach

We compared the performance of the all-in-one and ensemble classifiers to Daumé's feature augmentation algorithm. Table 3 shows that the all-in-one classifier exceeds Daumé's approach in all 27 domains given our current implementation of Daumé's approach. The ensemble exceeds Daumé's approach on all domains except office, kitchen & housewares, magazines, office products, and tweets.

Structural Correspondence Learning (SCL)

We employed the four domains datasets used in Blitzer et al. (2007) to train and test the all-in one and the ensemble classifiers. We also replicated the in-domain results of these four datasets using our maximum entropy classifier. We compare the results of the all-in-one and the ensemble classifier to the SCL and its variation SCL-MI adaptation techniques using the four datasets used to evaluate SCL and SCL-MI in Blitzer et al. (2007).

Note that the results published in Blitzer's work represent pairwise domain-adaptation, while our ensemble and all-in-one results are based on training on three of Blitzer's domains and testing on the held-out fourth domain. This makes it impossible to draw a direct comparison, but we can still observe that in general, it is best to simply combine as many domains as possible in an all-in-one or ensemble approach as compared to carefully adapting a single domain. Table 2 summarizes the results of the comparison.

Classifier	Books	DVD	Electronics	Kitchen
In-Domain	81.50%	83.00%	84.50%	83.50%
SCL Adaptation	72.80%	74.60%	78.40%	80.80%
SCL-MI Adaptation	74.60%	76.30%	78.90%	82.10%
All-in-one Classifier	79.00%	82.50%	79.50%	80.00%
Ensemble	79.00%	77.50%	80.00%	85.50%

Table2: Comparison of SCL, All-in-one, and Ensemble Classifiers

Reporting Transfer Ratio

Glorot et al. (2011) introduced a definition for the *transfer loss* t for a source domain S and a target domain T. It represents loss of accuracy using a transfer model compared to an in-domain model:

Where is the transfer error defined as the test error obtained by a method trained on the source domain S and tested on the target domain T. is the test error obtained by the base-line method.

The *transfer ratio* Q also characterizes the transfer but is defined by replacing the difference by a quotient in t:

Domain	In-Domain	All-in-one	Ensemble-sum of weights	Ensemble-majority votes	Ensemble (logistic re-gression)	Ensemble (sigmoid kernel)	Hal-Daume
Apparel	90.87%	92.81%	96.40%	90.30%	90.65%	97.12%	92.09%
Automotive	83.85%	92.31%	92.31%	86.76%	96.15%	96.15%	76.92%
Baby	91.94%	89.15%	89.15%	89.72%	77.52%	83.72%	82.95%
Beauty	90.00%	89.87%	84.81%	87.88%	87.34%	83.54%	75.95%
Books	87.19%	87.50%	82.03%	83.16%	74.22%	80.47%	75.78%
Camera & photo	94.33%	94.03%	92.54%	90.35%	89.55%	88.81%	87.31%
Cell-phones & service	93.13%	95.31%	89.06%	90.45%	95.31%	95.31%	75.00%
Computer & video-games	95.77%	90.14%	87.32%	87.87%	80.28%	77.46%	71.83%
DVD	91.11%	89.68%	85.71%	83.65%	78.57%	86.51%	82.54%
Electronics	92.35%	92.65%	90.44%	87.22%	91.91%	85.29%	80.15%
Gourmet-food	89.68%	94.12%	82.35%	83.89%	82.35%	85.29%	79.41%
Grocery	92.41%	90.74%	92.59%	88.18%	85.19%	88.89%	79.63%
Health & personal-care	93.55%	95.65%	92.75%	89.78%	83.33%	87.68%	86.23%
Hotel	95.15%	96.00%	93.00%	90.36%	87.50%	88.50%	85.00%
Jewelry & watches	94.78%	97.83%	97.83%	89.90%	93.48%	93.48%	80.43%
Kitchen & housewares	93.33%	92.03%	92.03%	90.30%	86.23%	89.86%	93.07%
Magazines	96.38%	90.58%	89.86%	83.81%	76.81%	85.51%	89.13%
Music	90.39%	89.15%	88.37%	81.61%	79.07%	80.62%	72.87%
Musical instruments	95.71%	100.00%	100.00%	91.18%	100.00%	100.00%	85.71%
Office products	95.56%	100.00%	100.00%	92.00%	100.00%	88.89%	100.00%
Outdoor living	97.27%	89.09%	90.91%	89.37%	85.45%	89.09%	83.64%
Software	94.81%	90.70%	94.57%	89.08%	93.80%	89.92%	87.60%
Sports & outdoors	94.62%	89.23%	88.46%	88.76%	83.08%	86.92%	86.15%
Tools & hardware	100.00%	100.00%	100.00%	92.86%	100.00%	100.00%	100.00%
Toys & games	93.80%	96.27%	94.78%	89.47%	91.79%	94.03%	91.79%
Video	91.93%	90.30%	81.34%	85.22%	73.13%	82.09%	80.60%
Tweets	72.82%	68.50%	63.50%	62.82%	60.00%	57.50%	61.50%

Table 3: Performance of the All-in-One, Ensemble and Hal Daume's Classifiers

$$- \quad Q = \frac{1}{n} \sum_{(S,T) S \neq T} \frac{e(S,T)}{e_b(T,T)}$$

Where n is the number of couples (S, T) with S≠T.

The all-in-one classifier had a 1.12 transfer ratio across domains, which is very close to the best result of ~1.07 in Glorot et al. The ensemble with Sigmoid kernel of SVM trained on 50 positive and 50 negative feedback examples from the target domain had 1.81 transfer ratio. The ensemble with radial basis function (gamma=0.01) trained on 5 positive and 5 negative feedback examples from the target domain had 1.85 transfer ratio. Note that the transfer ratio of the in-domain classifier, which is used a base-line for calculating the transfer ratio is 1. The transfer ratio of the all-in-one classifier is better than the transfer ratio of the ensemble with its two variations.

4.2 Discussion

The results in the previous section indicate that both the all-in-one and the ensemble approaches exceed both Daumé's domain adaptation technique on the 27 datasets (given our current implementation of Daumé's approach) and SCL on the four datasets in Blitzer et al. (2007) and that the all-in-one approach achieves comparable results in terms of transfer ratio to Glorot et al. (2011).

The ensemble approach exceeds the all-in-one in some domains like apparel and automotive. They both are very close in some domains like

cell phones & services, musical instruments, tools & hardware and outdoor living. For the rest of the 27 domains, the all-in-one exceeds the ensemble classifier. The all-in-one classifier exceeds the ensemble in using the transfer ratio metric.

When comparing the all-in-one and the ensemble approaches on the four datasets in Blitzer et al. (2007), the all-in-one exceeds the ensemble only in the DVD domain. The ensemble exceeds the all-in-one in electronics and kitchen & housewares. They both perform at the same accuracy level on the books domain.

We have also employed NcNemar significance test between pairs of the all-in-one, the ensemble and Daumé's approaches on the 27 domains. Table 4 shows the significance difference between the approaches' combinations.

Pair of Approaches	Average NcNemar Test	p-value
All-in-one & In-domain	2.066976595	No significant difference p = 0.20
All-in-one & Ensemble	2.736901971	No significant difference p = 0.10
All-in-one & Daumé's	8.976122	Significant at p = 0.01
Ensemble & In-domain	4.077642586	Significant at p = 0.05
Ensemble & Daumé's	11.47808047	Significant at p = 0.001
Daumé's & In-domain	10.46852763	Significant at p = 0.01

Table 4: NcNemar Significance Test Results

Finally, we would like to do some initial exploration of the role of features across domains. The commonly held belief is that sentiment indicators such as "hot" can change their polarity from domain to domain (e.g. it is positive in the food domain while it is negative in the negative domain), contributing to the need for domain adaptation. On the other hand, the success of the all-in-one classifier indicates that a greater number of observed sentiment features and more solid statistics on those features are more important than capturing domain-specific polarity changes.

In order to gather evidence for or against these hypotheses, we first calculated the number of overlapping features between each pair of domains within the 27 domains. The average percentage of features that overlap between pairs of domain is only 12.48%. Furthermore, only a very small set of the highly sentiment-correlated features overlap. 16 features overlap among the 27 domains which accounts for only 0.08% of the features. Examples of positive overlapping feature are "highly", "excellent", and "great". Negative overlapping features are "waste", "terrible", and "worst". This low feature overlap of sentiment-bearing features lends some support to the hypothesis that in order to capture a general, large-scale sentiment vocabulary nothing beats diverse and plentiful training data. The low feature overlap also justifies why the all-in-one classifier exceeds the ensemble though the latter has access to some labeled data in the target

Second, we examined the question of polarity-changing sentiment features. Among the top 1000 features in each domain ranked by LLR, we counted the common features among multiple domains. The number of common features among 15 domains is 42 features. Only 13 features are common among 20 domains while there are no common features from the highest 1000 likelihood ratio features among the 27 domains. Most features do not flip polarity across domains. For example the word "waste" is common among 20 domains and maintains a negative polarity across the domains. Very few features flip polarity across domains. The word "highly" is shared across 23 domains. It maintains a positive polarity in all domains while it flips in Tools & Hardware. The word "refund" is shared in 20 domains. It maintains a negative polarity in almost all domains except Gourmet Food.

5 Conclusion

In this paper, we empirically re-examine the assumption that adapting one or multiple domains with plenty of labeled sentiment polarity data to one domain with little labeled data requires new and sophisticated algorithms. We evaluate four domain adaptation techniques on a wide variety of domains in two major groups of state-of-the-art datasets. Our experiments show that overall, simple domain adaptation techniques like the all-in-one classifier do comparably well if not better than more sophisticated domain adaptation techniques. Combined with the fact that labeled sentiment data tends to be cheap to come by through either the collection of product reviews from the web or inexpensive crowd-sourced labeling, this indicates that in practice, domain-adaptation for sentiment detection might be of less importance than previously claimed.

We also show that the often anecdotally observed "polarity-flip" of sentiment terms from one domain to another in practice is a rather rare occurrence and might not be as detrimental to sentiment domain adaptation as assumed in much of the literature.

References

John Blitzer, Ryan McDonald and Fernando Pereira. 2006. Domain Adaptation with Structural Correspondence Learning. In *Proceedings of EMNLP*.

John Blitzer, Mark Dredze and Fernando Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain-Adaptation for Sentiment Classification. In *Proceedings of ACL*.

Minmin Chen, Kilian Q. Weinberger and John C. Blitzer. 2011. Co-Training for Domain Adaptation. In *Proceedings of Advances in Neural Information Processing (NIPS).*

Hal Daumé III. 2007. Frustratingly Easy Domain Adaptation. In *Proceedings of ACL*.

Thomas G. Dietterich.1997. Machine Learning Research: Four Current Directions. *In: AI Magazine. 18 (4), 97-136.*

Xavier Glorot, Antoine Bordes and Yoshua Bengio. 2011. Domain-Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. In *Proceedings of ICML*.

Soo-Min Kim and Edward Hovy. 2006. Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In: *Proceedings of the ACL Workshop on Sentiment and Subjectivity in Text, 1-8.*

Shoushan Li and Chengqing Zong. 2008. Multi-Domain Adaptation for Sentiment Classification: Using Multiple Classifier Combining Methods. In *Proceedings of Natural Language Processing and Knowledge Engineering*.

Shoushan Li and Chengqing Zong. 2008. Multi-Domain Sentiment Classification. In *Proceedings of Association of Computing Linguistics.*.

Mark Dredze and Koby Crammer. 208. Online Methods for Multi-Domain Learning and Adaptation. In *proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing.*

Bo Pang and Lilian Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*: Vol. 2: No 1–2, pp 1-135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2002).*

Bing Liu. 2012. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-1997).*

Rajhans Samdani and Wen-Tau Yih. 2011. Domain Adaptation with Ensemble of Feature Groups. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*.

Peter D. Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2002).*

Theresa Wilson , Janyce Wiebe, & Rebecca Hwa. 2006. Recognizing strong and weak opinion clauses. *Computational Intelligence 22 (2)*: 73-99.

Munmun De Choudhury, Michael Gamon, and Scott Counts. Happy, Nervous or Surprised? Classification of Human Affective States in Social Media, *Association for the Advancement of Artificial Intelligence, June 2012*

Munmun De Choudhury, Scott Counts, and Michael Gamon. Not All Moods are Created Equal! Exploring Human Emotional States in Social Media. *Association for the Advancement of Artificial Intelligence, June 2012*