Multi-Domain Sentiment Relevance Classification with Automatic Representation Learning

Christian Scheible
Institut für Maschinelle Sprachverarbeitung
University of Stuttgart
scheibcn@ims.uni-stuttgart.de

Hinrich Schütze
Center for Information
and Language Processing
University of Munich

Abstract

Sentiment relevance (SR) aims at identifying content that does not contribute to sentiment analysis. Previously, automatic SR classification has been studied in a limited scope, using a single domain and feature augmentation techniques that require large hand-crafted databases. In this paper, we present experiments on SR classification with automatically learned feature representations on multiple domains. We show that a combination of transfer learning and in-task supervision using features learned unsupervisedly by the stacked denoising autoencoder significantly outperforms a bag-of-words baseline for in-domain and cross-domain classification.

1 Introduction

Many approaches to sentiment analysis rely on term-based clues to detect the polarity of sentences or documents, using the bag-of-words (BoW) model (Wang and Manning, 2012). One drawback of this approach is that the polarity of a clue is often treated as fixed, which can be problematic when content is not intended to contribute to the polarity of the entity but contains a term with a known lexical non-neutral polarity.

For example, movie reviews often have plot summaries which contain subjective descriptions, e.g., "April loves her new home and friends.", containing "loves", commonly a subjective positive term. Other domains contain different types of nonrelevant content: Music reviews may contain track listings, product reviews on retail platforms contain complaints that do not concern the product, e.g., about shipping and handling. Filtering such nonrelevant content can help to improve sentiment analysis (Pang and Lee, 2004). *Sentiment relevance* (Scheible and Schütze, 2013; Taboada et al., 2009; Täckström and McDonald, 2011) formalizes this distinction: Content that contributes to the overall sentiment of a document is said to be *sentiment relevant* (*SR*), other content is *sentiment nonrelevant* (*SNR*).

The main bottleneck in automatic SR classification is the lack of annotated data. On the sentence level, it has been attempted for the movie review domain (Scheible and Schütze, 2013) on a manually annotated dataset that covers around 3,500 sentences. The sentiment analysis data by Täckström and McDonald (2011) contains SR annotations for five product review domains, four of which have fewer than 1,000 annotated examples.

As the amount of labeled data is low, we adopt *transfer learning* (*TL*, (Thrun, 1995)), which has been used before for SR classification. In this setup, we train a classifier on a different task, using subjectivity-labeled data – for which a large number of annotated examples is available – and apply it for SR classification. To enable knowledge transfer between the tasks, feature space augmentation has been proposed. For this purpose, we employ automatic representation learning, using the stacked denoising autoencoder (SDA, (Vincent et al., 2010)) which has been applied successfully to other domain adaptation problems such as cross-domain sentiment analysis (Glorot et al., 2011).

In this paper, we present experiments on both multi-domain and cross-domain SR classification. We show that compared to the in-domain baseline, TL with SDA features increases F_1 by 6.8% on average. We find that domain adaptation using TL with the SDA compensates for strong domain shifts, reducing the average classification transfer loss by 12.7%.

2 Stacked Denoising Autoencoders

The *stacked denoising autoencoder* (*SDA*, (Vincent et al., 2010)) is a neural network (NN) model for unsupervised feature representation learning.

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 200–204,
Gothenburg, Sweden, April 26-30 2014.

An *autoencoder* takes an input vector $\mathbf{x}$, uses an NN layer with a (possibly) nonlinear activation function to generate a hidden feature representation $\mathbf{h}$. A second NN layer reconstructs $\mathbf{x}$ at the output, minimizing the error.

Denoising autoencoders reconstruct $\mathbf{x}$ from a corrupted version of the input, $\tilde{\mathbf{x}}$. As the model learns to be robust to noise, the representations are expected to generalize better. For discrete data, masking noise is a natural choice, where each input unit is randomly set to 0 with probability p.

Autoencoders can be *stacked* by using the $\mathbf{h_i}$ produced by the i^{th} autoencoder as the input to the $(i+1)^{\text{th}}$ one, yielding the representation $\mathbf{h_{i+1}}$. The $\mathbf{h}$ of the topmost autoencoder is the final representation output by the SDA. We let k-SDA denote a stack of k denoising autoencoders.

Chen et al. (2012) introduced a marginalized closed-form version, the mSDA. We opt for this version as it is faster to train and allows us to use the full feature space which would be inefficient with iterative backpropagation training.

3 Task and Experimental Setup

The task in this paper is multi- and cross-domain SR classification. Two aspects motivate our work: First, we need to address the sparse data situation. Second, we are interested in how cross-domain effects influence SR classification. We classify SR in three different setups: in-domain (ID), in which we take the training and test data from the same domain; domain adaptation (DA), where training and test data are from different domains; and transfer learning (TL), where we use a much larger amount of data from a different but related task. To improve the generalization capabilities of the models, we use representations learned by the SDA. We will next describe our classification setup in more detail.

Data We use the following datasets for our experiments. Table 1 shows statistics on the datasets.

CINEMA: The movie SR data (CINEMA) by Scheible and Schütze (2013) contains SR-annotated sentences for the movie review domain. Ambiguous sentences are marked as *unknown*; we exclude them.

PRODUCTS: The multi-domain product data (PRODUCTS) by Täckström and McDonald (2011) contains labeled sentences from five Amazon.com product review domains: BOOKS, DVDS, electronics (EL), MUSIC, and video games (VG). This dataset differs from CINEMA firstly in the product domains (except obviously for DVDS which also covers movies). Secondly, the data was collected from a retail site, which introduces further facets of sentiment nonrelevance, as discussed above. Thirdly, the annotation style has no *unknown* category: ambiguous examples are marked as SR.

Dataset	#doc	#sent	#SR	#SNR
CINEMA	125	3,487	2,759	728
PRODUCTS	294	3,836	2,689	1,147
–BOOKS	59	739	424	315
–DVDS	59	799	524	275
–ELECTRONICS	57	628	491	137
–MUSIC	59	638	448	190
–VIDEOGAMES	60	1032	802	230
P&L	–	10,000	5,000	5,000
UNLAB	7,500	68,927	–	–

Table 1: Dataset statistics

P&L: The subjectivity data (P&L) by Pang and Lee (2004) serves as our cross-task training data for transfer learning. The dataset was heuristically created for subjectivity detection on the movie domain by sampling snippets from Rotten Tomatoes as subjective and sentences from IMDb plot summaries as objective examples.

UNLAB: To improve generalization on PRODUCTS, we use additional unlabeled sentences (UNLAB) for SDA training. We extract the sentences of 1,500 randomly selected documents for each of the five domains from the Amazon.com review data by Jindal and Liu (2008).

SDA setup We train the SDA with 10,000 hidden units and tanh nonlinearity on the BoW features of all available data as the input. We optimize the noise level p with 2-fold cross-validation on the in-domain training folds.

Classification setup We perform SR classification with a linear support vector machine (SVM) using LIBLINEAR (Chang and Lin, 2011). We perform 2-fold cross-validation for all training data but P&L. We report overall macro-averaged F_1 over both folds. The feature representation for the SVM is either bag of words (BoW) or the k-SDA output. Unlike Chen et al. (2012), we do not use concatenations of BoW and SDA vectors as we found them to perform worse.

Evaluation As class distributions are heavily skewed, we use *macro-averaged* $F_1(s,t)$ (training on s and evaluating on t) as the basic evaluation measure. We evaluate DA with *transfer loss*, the difference in F_1 of a classifier CL with

	Features	Setup	CINEMA	BOOKS	DVDS	EL	MUSIC	VG	∅
1	Majority BL	–	39.6	28.9	32.6	39.2	35.1	39.0	35.7
2	BoW	ID	74.0	57.5	49.8	55.1	55.5	55.0	58.4
3	1-SDA	ID	73.6	55.3	48.4	43.8	41.8	44.1	52.6
4	2-SDA	ID	76.0	54.5	52.5	43.9	41.2	46.7	53.6
5	BoW	TL	71.5	60.7	60.2	50.3	55.1	53.2	59.6
6	1-SDA	TL	73.3	**62.9**	60.6	**59.0**	**59.9**	57.0	63.1
7	2-SDA	TL	76.2	**62.9**	**65.8**	**59.7**	**59.9**	**60.5**	**64.9**
8	BoW	ID+TL	76.6	**63.5**	61.7	52.4	56.7	57.0	62.3
9	1-SDA	ID+TL	**79.0**	**62.7**	62.1	**57.7**	**57.8**	**57.4**	63.9
10	2-SDA	ID+TL	**80.4**	**62.7**	**65.2**	**59.0**	**58.7**	**58.9**	**65.2**

Table 2: Macro-averaged F_1 (%) evaluating on each test domain on both folds. ∅ = row mean. **Bold:** best result in each column and results in that column not significantly different from it.

respect to the in-domain baseline BL: $L(s,t) = F_1^{(\mathrm{BL})}(t,t) - F_1^{(\mathrm{CL})}(s,t)$. L is negative if the classifier surpasses the baseline. As a statistical significance test (indicated by † in the text), we use *approximate randomization* (Noreen, 1989) with 10,000 iterations at $p < 0.05$.

4 Experiments

In-Domain Classification (ID) Table 2 shows macro-averaged F_1 for different SR models. We first turn to fully supervised SR classification with bag-of-words (BoW) features using ID training (line 2). While the results for CINEMA are high, on par with the reported results in related work, they are low for the PRODUCTS data. This is not surprising as the SVM is trained with fewer than 600 examples on each domain. Also, no *unknown* category exists in the latter dataset. While ambiguous examples on CINEMA are annotated as *unknown*, they receive an SR label on PRODUCTS. Thus, many examples are ambiguous and thus difficult to classify. SDA features worsen results significantly† (lines 3–4) on all domains except CINEMA and DVDS due to data sparsity. They are the two most homogeneous domains where plot descriptions make up a large part of the SNR content. On many domains, there is no single prototypical type of SNR which could be learned from a small amount of training data.

Transfer Learning (TL) TL with training on P&L and evaluation on CINEMA/PRODUCTS with BoW features (line 5) performs slightly worse than ID classification, except on BOOKS and DVDS where we see strong improvements. This result is easy to explain: Both BOOKS and DVDS contain SNR descriptions of narratives, which are covered well in P&L. This distinction is less helpful on domains like EL where SNR content is different, so we achieve worse results even with the much larger P&L data.

We find that 1-SDA (line 6) already performs significantly† better than the ID baseline on all domains except CINEMA which has a much larger amount of ID training data available than the other domains (approx. 1700 sentences vs. fewer than 600). Using stacking, 2-SDA (line 7) improves the results on three domains significantly,† and performs on par with the ID classifier on CINEMA. We found that stack depths of $k > 2$ do not significantly† increase performance.

Finally, we try a combination of ID and TL (ID+TL), training on both P&L and the respective ID training fold of CINEMA/PRODUCTS. The results for this experiment are shown in lines 8–10 in Table 2. Comparing BoW models, we beat both ID and TL across all domains (lines 2 and 5). With SDA features, we are able to beat ID for CINEMA. The results on the other domains are comparable to plain TL. This is a promising result, showing that with SDA features, ID+TL performs as well as or better than plain TL. This property could be exploited for domains where labeled data is not available. We will show below that SDA features become important when we apply ID+TL to domain adaptation.

We also conducted experiments using only the 5,000 most frequent features but found that the SDA does not generalize well from this input representation, particularly on EL and MUSIC. This confirms that in SR, rare features make an important contribution (such as named entities in the movie domain).

Domain Adaptation (DA) We now evaluate the task in a DA setting, comparing the ID and ID+TL

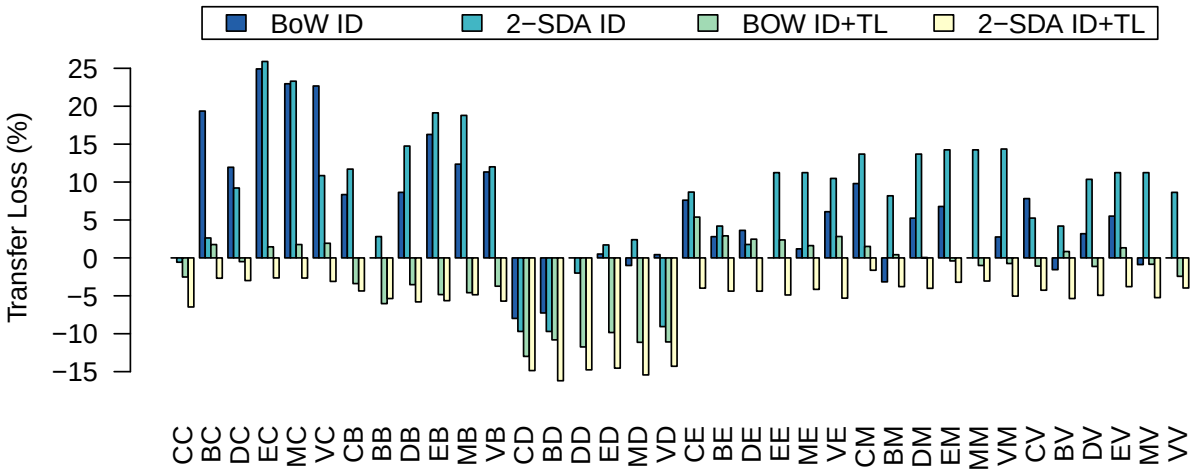

Figure 1: Transfer losses (%) for DA. Training-test pairs grouped by target domain and abbreviated by first letter (e.g., CD: training on CINEMA, evaluating on DVDS). In-domain results shown for comparison to Table 2.

setups with BoW and 2-SDA features. We measure the transfer losses we suffer from training on one domain and evaluating on another (Figure 1). The overall picture is the same as above: ID+TL 2-SDA models perform best. In the baseline BoW ID setup, domain shifts have a strong influence on the results. The combination of out-of-domain and out-of-task data in ID+TL keeps losses uniformly low. 2-SDA features lower almost all losses further. On average, 2-SDA ID+TL reduces transfer loss by 12.7 points compared to the baseline (Table 3). As expected, pairings of thematically strongly related domains (e.g., BOOKS and DVDS) have lower losses in all setups.

The biggest challenge is the strong domain shift between the CINEMA and PRODUCTS domains (concerning mainly the retail aspects). With BoW ID, losses on CINEMA reach up to 25 points, and using CINEMA for training causes high losses for PRODUCTS in most cases. Our key result is that the ID+TL 2-SDA setup successfully compensates for these problems, reducing the losses below 0.

Losses across the PRODUCTS domains are less pronounced. The DVDS baseline classifier has the lowest F_1 (cf. Table 2) and shows the highest improvements in domain adaptation: BoW models of other domains perform better than the in-domain classifier. Analyzing the DVDS model shows overfitting to specific movie terms which occur frequently across each review in the training data. SNR content in movies is mostly concerned with named entity types which cannot easily be learned from BoW representations. Out-of-domain models are less specialized and perform better than in-domain models. TL and SDA increase the coverage of movie terms and provide better generalization, which improves performance further.

	BoW	2-SDA
ID	6.7	8.9
ID+TL	-1.8	-6.0

Table 3: Mean transfer losses (%) for the different training data and feature representation setups. In-domain results not included.

BOOKS is the most challenging domain in all setups. It is particularly heterogeneous, containing both fiction and non-fiction reviews which feature different SNR aspects. Both results illustrate that domain effects depend on how diverse SNR content is within the domain.

Overall, the results show that ID+TL leads to a successful compensation of cross-domain effects. SDA features improve the results significantly† for ID+TL. In particular, we find that the SDA successfully compensates for the strong domain shift between CINEMA and PRODUCTS.

5 Conclusion

We presented experiments on multi- and cross-domain sentiment relevance classification. We showed that transfer learning (TL) using stacked denoising autoencoder (SDA) representations significantly increases performance by 6.8% F_1 for in-domain classification. Moreover, the average transfer loss in domain adaptation is reduced by 12.7 percentage points where the SDA features compensate for strong domain shifts.

References

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (ACM TIST)*, 2(3):1–27.

Minmin Chen, Zhixiang Xu, Kilian Weinberger, and Fei Sha. 2012. Marginalized denoising autoencoders for domain adaptation. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 513–520.

Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *Proceedings of the International Conference on Web Search and Web Data Mining (WSDM)*, pages 219–230.

Eric W. Noreen. 1989. *Computer Intensive Methods for Hypothesis Testing: An Introduction*. Wiley.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL)*, pages 271–278.

Christian Scheible and Hinrich Schütze. 2013. Sentiment relevance. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 954–963.

Maite Taboada, Julian Brooke, and Manfred Stede. 2009. Genre-based paragraph classification for sentiment analysis. In *Proceedings of the SIGDIAL 2009 Conference*, pages 62–70.

Oscar Täckström and Ryan McDonald. 2011. Semi-supervised latent variable models for sentence-level sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 569–574.

Sebastian Thrun. 1995. Is learning the n-th thing any easier than learning the first? In *Proceedings of Advances in Neural Information Processing Systems 8 (NIPS)*, pages 640–646.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research (JMLR)*, 11:3371–3408.

Sida Wang and Christopher Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 90–94.