Bridging Domains with Words: Opinion Analysis with Matrix Tri-factorizations

Tao Li[*]   Vikas Sindhwani[†]   Chris Ding[‡]   Yi Zhang[§]

Abstract

With the explosion of user-generated web2.0 content in the form of blogs, wikis and discussion forums, the Internet has rapidly become a massive dynamic repository of public opinion on an unbounded range of topics. A key enabler of opinion extraction and summarization is sentiment classification: the task of automatically identifying whether a given piece of text expresses positive or negative opinion towards a topic of interest. Building high-quality sentiment classifiers using standard text categorization methods is challenging due to the lack of labeled data in a target domain. In this paper, we consider the problem of cross-domain sentiment analysis: can one, for instance, download *rated* movie reviews from rottentomatoes.com or IMBD discussion forums, learn linguistic expressions and sentiment-laden terms that *generally characterize* opinionated commentary and then successfully transfer this knowledge to the target domain, thereby building high-quality sentiment models without manual effort? We outline a novel sentiment transfer mechanism based on constrained non-negative matrix tri-factorizations of term-document matrices in the source and target domains. The constrained matrix factorization framework naturally incorporates document labels via a least squares penalty incurred by a certain linear model and enables direct and explicit knowledge transfer across different domains. We obtain promising empirical results with this approach.

Keywords: Sentiment analysis, Transfer learning, Non-negative matrix factorization

1 Introduction

Suppose that we download movie reviews from an online review site and compile a large term-document matrix, X_1 representing m terms and n documents. Assume further that the sentiment associated with at least some of these documents is known from the explicit ratings given by movie enthusiasts that visit these sites. From such a dataset, we can hope to identify general linguistic indicators of positive and negative opinions, e.g., sentiment-laden terms such as "great" and "awful" respectively. The end goal of this exercise would be to apply this knowledge to gauge sentiment around documents in a new *target* domain X_2, which, for example may be a collection of blogs posts talking about products and services of keen interest to a company. The process of building a sentiment classifier in this manner, applicable in the target domain, is appealing because it invokes no additional human effort. In this paper, we propose a mechanism to transform document label information in one domain to another, *via words*, using a principled matrix factorization framework.

Gleaning insights by monitoring and analyzing large amounts of user-generated online content is gaining immense importance. Recent surveys have estimated that a massive number of internet users turn to online forums to collect recommendations for products and services, guiding their own choices and decisions by the opinions that other consumers have publically expressed. The trust placed on the opinion of another consumer is often much greater than that placed on advertisements for a product. For consumers, therefore, the plethora of information and opinions from diverse sources helps them tap into the wisdom of crowds, to aid in making more informed decisions; while for producers, tracking the pulse of this ever-expanding blogosphere, enables them discern what consumers are saying about their products, which provides useful insight on how to improve or market products better. This theme is the motivation for this paper. The problem of automated sentiment classification is naturally a canonical task in this discussion.

An in-depth survey of sentiment analysis literature [18] shows emphasis on two strands of research, both mainly geared towards single-domain applications. One strand has focused on lexical rule based approaches where hand-crafted dictionaries are used to assign sentiment labels based on relative frequencies of positive and negative terms. As observed by [17], most semi-automated dictionary-based approaches yield unsatisfactory lexicons, with either high coverage and low precision or vice versa. The other strand of work, pioneered by [20] has demonstrated the value

[*]School of Computer Science, Florida International University, Miami, FL 33199. Email: taoli@cs.fiu.edu.

[†]Business Analytics and Mathematical Sciences, IBM T.J. Watson Research Center, Yorktown Heights, NY 10598. Email: vsindhw@us.ibm.com.

[‡]CSE Department, University of Texas at Arlington, Arlington, TX 76019. Email: chqding@uta.edu.

[§]School of Computer Science, Florida International University, Miami, FL 33199. Email: yzhan004@cs.fiu.edu.

of posing sentiment classification as a standard supervised learning task. In particular, a well-trained state of the art text classifier is able to outperform dictionary based approaches. On the other hand, the accuracy of such a text classifier depends on the amount of labeled documents in the domain of interest. The cost of acquiring labeled data, has started to motivate the application of semi-supervised techniques [8] which attempt to use unlabeled examples to learn high quality classifiers. Some recent papers have also attempted to combine dictionary-based approaches with supervised [16] and semi-supervised classification [23].

Transfer learning and domain adaptation (see e.g., [2]) are natural complimentary techniques to attempt to learn from limited labeled data in a target domain, provided the existence of related domains where labeled examples are more cheaply available. In this paper, we propose a novel two stage method based on constrained non-negative matrix tri-factorizations. Starting from labeled documents in the source domain, the first stage transfers document-side sentiment into word-level sentiment. Assuming that documents in source and target domains are represented over the same vocabulary, the second stage transfers word sentiment learnt in the first stage over to documents in the target domain. We show that this kind of transfer can be easily implemented using closed form update equations with convergence guarantees.

We begin this paper with a brief overview of related work. We then give a background of matrix factorization techniques. In the section that then follows, we describe a constrained matrix factorization framework where document labels are naturally incorporated via a least squares penalty incurred by a certain linear model. This framework forms the first stage of our cross-domain procedure. In the second stage, we take the learnt factor associated with terms from the first stage, and introduce it as labels on the word side. A second factorization on the target domain then produces the final document labeling. Our cross-domain procedure is then described followed by detailed empirical studies.

2 Related Work

The recent book [18] provides a detailed survey of the field of sentiment analysis from both natural language processing and machine learning perspectives. In this section, we briskly cover related work to position our contributions appropriately in the literature.

Lexical approaches attempting to generate dictionaries capturing the sentiment of words have ranged from manual approaches of developing domain-dependent lexicons [5] to semi-automated approaches [10, 30, 11], and even an almost fully automated approach [26]. Most semi-automated approaches have met with limited success [17]. Supervised learning models have tended to outperform dictionary-based classification schemes [20]. Pang and Lee [19] suggested a two-tier scheme where sentences are first classified as *subjective* versus *objective*, and then the sentiment classifier is applied on only the *subjective* sentences leading to improved performance. Empirical studies in these papers also suggest that using more sophisticated linguistic models, incorporating parts-of-speech and n-gram language models, do not improve over the simple unigram bag-of-words representation. In keeping with these findings, we also adopt a unigram text model. A subjectivity classification phase before our models are applied may further improve the results reported in this paper, but our focus is on transferring knowledge across domains to build a sentiment classifier appropriate for a target domain with minimal manual effort.

While domain adaptation and transfer learning have received great attention in machine learning recently, the application to sentiment analysis a relatively very new effort. In particular, we have empirically compared our approach with other reasonable baseline approaches and well-known semi-supervised methods, but have not been able to address comparisons with recently proposed transfer learning techniques [3, 24, 25] for this task in this paper. [1] conducted an initial empirical study on domain adaptation for sentiment analysis. Bliter et al. [3] adapted a previously proposed *structural correspondence learning* approach to this problem. The basic idea is to first choose a set of pivot words which occur frequently in both source and target domains and/or have high mutual information with the source labels. Then, their approach models the correlations between the pivot features and all other features by training linear pivot predictors to predict occurrences of each pivot in the unlabeled data from both domains. These pivot predictors are then used for define a projection for instances. If the projection defines meaningful correspondences, a classifier trained on the concatenation of original representation with this projected representation is likely to perform well on both source and target domains. [3] also suggested the use of small amount of target domain labeled data to correct misaligned projections. In a sense, the discovery of low-rank factors in the first stage of our approach influenced by labeled data in the source domain, and its subsequent use as a prior in the second stage, may be viewed as a form of pivoting. By being able to incorporate labels from the target domain, our model also provides for domain-specific context corrections.

The matrix tri-factorization models explored in this paper are closely related to the models proposed recently in [13, 22, 15, 14]. Though, their techniques for proving algorithm convergence and correctness can be readily adapted for our models, [13] did not incorporate cross-domain supervision as we do. The dual supervision models of [22] are also only applicable to single domains, and do not enforce non-negativity or orthogonality – aspects of matrix factorization models that have shown benefits in prior empirical studies, see e.g., [7]. In another recent paper [21], stan-

dard regularization models are constrained using graphs of word co-occurences. Many of these papers are very recently proposed methodologies that may well be adapted for cross-domain applications.

3 Basic Matrix Factorization Model

Our proposed models are based on non-negative matrix tri-factorization [7]. In these models, an $m \times n$ term-document matrix X is approximated by three factors that specify soft membership of terms and documents in one of k_1 and k_2 classes respectively:

$$(3.1) \qquad X \approx FSG^T.$$

where F is an $m \times k_1$ non-negative matrix representing knowledge in the word space, i.e., i-th row of F represents the posterior probability of word i belonging to the k_1 classes, G is an $n \times k_2$ non-negative matrix representing knowledge in document space, i.e., the i-th row of G represents the posterior probability of document i belonging to the k_2 classes, and S is an $k_1 \times k_2$ nonnegative matrix providing a condensed view of X.

The matrix factorization model is similar to the probabilistic latent semantic indexing (PLSI) model [9]. In PLSI, X is treated as the joint distribution between words and documents by the scaling $X \rightarrow \bar{X} = X / \sum_{ij} X_{ij}$ thus $\sum_{ij} \bar{X}_{ij} = 1$). $\bar{X}$ is factorized as

$$(3.2)$$
$$\bar{X} \approx WSD^T, \sum_k W_{ik} = 1, \sum_k D_{jk} = 1, \sum_k S_{kk} = 1.$$

where X is the $m \times n$ word-document semantic matrix, $X = WSD$, W is the word class-conditional probability, and D is the document class-conditional probability and S is the class probability distribution.

PLSI provides a simultaneous solution for the word and document class conditional distribution. Our model provides simultaneous solution for clustering the rows and the columns of X. To avoid ambiguity, the orthogonality conditions

$$(3.3) \qquad F^T F = I, \ G^T G = I.$$

can be imposed to enforce each row of F and G to possess only one nonzero entry. Approximating the term-document matrix with a tri-factorization while imposing non-negativity and orthogonality constraints gives a principled framework for simultaneously clustering the rows (words) and columns (documents) of X. In the context of co-clustering, these models return excellent empirical performance, see e.g., [7].

In the following, we take $k_1 = 2$ and $k_2 = k$ where k is a rank parameter for our algorithm. The rationale for these choices is that we intend to use sentiment labels for the source domain to estimate F whose columns reflect affinity of a word with positive or negative sentiment class; while for G we allow a higher rank k so as to associate documents with k topics. We then use G as a topical representation for documents, over which a linear predictive model is simultaneously learnt. The formulation will become clearer as we proceed.

4 Matrix Factorization Model for Sentiment Analysis

4.1 Basic Description In the section, we describe a constrained matrix factorization framework where document labels are naturally incorporated via a least squares penalty incurred by a certain linear model. In particular, this framework allows document topics are discovered and words sentiments are disambiguated simultaneously and forms the first stage of our cross-domain procedure. The matrix factorization model is essentially performing semi-supervised sentiment analysis.

We assume that a few documents are manually labeled for the purposes of capturing some domain-specific connotations leading to a more domain-adapted model. The partial labels on documents can be described using a $n \times 2$ matrix Y where $Y_{i1} = 1$ if the document expresses positive sentiment, and $Y_{i2} = 1$ for negative sentiment.

We consider the following model for topic-based semi-supervised sentiment analysis:

$$(4.4)$$
$$\underset{F,S,G,W}{\arg\min} \| X - FSG^T \|^2 + \beta \mathrm{Tr}((GW - Y)^T D(GW - Y)),$$

where $\beta > 0$ is a parameter which determines the extent to which we enforce the prior knowledge respectively, F is an $m \times 2$ non-negative matrix representing knowledge in the word space, i.e., i-th row of F represents the posterior probability of word i belonging to the sentiment classes, G is an $n \times k$ non-negative matrix representing knowledge in document space, i.e., the i-th row of G represents the posterior probability of document i belonging to the k topic classes, and S is an $2 \times k_2$ nonnegative matrix providing a condensed view of X. D is a $n \times n$ diagonal matrix whose entry $(D)_{ii} = 1$ if the sentiment category of the i-th word is known (i.e., specified by the i-th row of Y) and $(D)_{ii} = 0$ otherwise. W are the coefficients of a linear model that predicts sentiment, GW, given the topical representation G of documents. Thus, via the second term, the real valued sentiment predictions, GW, are fit to the labeled data Y.

4.2 Algorithms The optimization problem in Eq.(4.4) can be solved using the following update rules

$$(4.5) \qquad F_{ik} \leftarrow F_{ik} \frac{(XGS^T)_{ik}}{[FSG^T GS^T]_{ik}}$$

$$(4.6) \qquad G_{jk} \leftarrow G_{jk} \frac{(X^T FS + \beta D Y^T W^T)_{jk}}{[GS^T F^T FS + \beta DGWW^T]_{jk}}$$

$$(4.7) \qquad S_{ik} \leftarrow S_{ik} \frac{(F^T XG)_{ik}}{(F^T FSG^T G)_{ik}}$$

$$(4.8) \qquad W_{ik} \leftarrow W_{ik} \frac{(G^T DY)_{ik}}{(G^T DGW)_{ik}}$$

The update rules for F, S, W have been derived (See [7]). Here we derive the update rule Eq.(4.6) for G.

4.3 Update Rule for G We solve optimization Eq.(4.4) with F, S, W fixed. The objective function can be rewritten as

$$L(G) = \mathrm{Tr}\,(X - FSG^T)^T(X - FSG^T) + \alpha\|D(GW - Y)\|_{fro}^2$$

or,

$$(4.9) \qquad L(G) = \mathrm{Tr}\,(AG^T G + BG^T DG - 2G^T Q)$$

where A, B, Q are constant matrices

$$A = S^T F^T FS,\ B = \alpha WW^T,\ Q = DYW$$

In the following, we prove that a local optimal solution to $\min_G L(G)$ is given by the following update rule:

$$(4.10) \qquad G_{ik} \leftarrow G_{ik} \frac{Q_{ik}}{(GA + DGB)_{ik}}.$$

Substituting A, B, Q into Eq.(4.10), we recover the update rule Eq.(4.6) for G.

Correctness

We have the following theorem regarding to the correctness of the algorithm:

THEOREM 4.1. *If the iteration of the update rule for G converges, it converges to a local optimal solution.*

Proof.

From the constrained optimization theory, the KKT complementary slackness condition, for $G \geq 0$ is given by

$$(4.11)\ (GA^T + GA + D^T GB + DGB - 2Q)_{ik}G_{ik} = 0.$$

Because $A^T = A, B^T = B$ and D is diagonal, this is reduced to

$$(4.12) \qquad (GA + DGB - Q)_{ik}G_{ik} = 0.$$

When the update iteration Eq.(4.10)converges, the converged solution G^* satisfies

$$(4.13) \qquad G_{ik}^* i = G_{ik}^* \frac{Q_{ik}}{(G^* A + DG^* B)_{ik}},$$

which equals to $(G^* A + DG^* B - Q)_{ik}G_{ik}^* = 0$. This is identical to the KKT complementary slackness condition Eq.(4.12). $\qquad\square$

Convergence

We have the following theorem regarding to the convergence of the algorithm:

THEOREM 4.2. *The iteration of the update rule for G converges.*

The proof of Theorem 2 can be found in Appendix.

5 Cross-Domain Knowledge Transfer

In this section, we show that our matrix factorization model enables a novel knowledge transfer mechanism, which is *direct* and *explicit*.

5.1 Transfer Mechanism Let X_1 and X_2 be term-document matrices in the source and target domains respectively. We assume that the number of terms is the same: if the vocabularies differ, we simply pad zero columns and re-express the matrices under the same unified vocabulary so that the column indices in both matrices correspond to the same word. Moreover, let V represent a $m \times m$ diagonal matrix with $V_{ii} = 1$ if i is a shared word, i.e., it occurs in both domains, or in other words the associated column is non-zero for both X_1 and X_2.

In domain X_1, we have some labeled documents. The partial labels on documents can be described using Y_1 where $(Y_1)_{i1} = 1$ if the document expresses positive sentiment, and $(Y_1)_{i2} = 1$ for negative sentiment. Note that one may also use soft sentiment polarities though our experiments are conducted with hard assignments. We first transfer this document label knowledge to words by learning F in a 2-way clustering via our matrix factorization model,

(5.14)
$$\min_{F,G,S,W} \|X_1 - FSG^T\|^2 + \beta\mathrm{Tr}\left[(GW - Y_1)^T D_1(GW - Y_1)\right]$$

where the notation $\mathrm{Tr}(A)$ means trace of the matrix A. Here, $\beta > 0$ is a parameter which determines the extent to which we enforce labeled information, D_1 is a $n \times n$ diagonal matrix whose entry $(D_1)_{ii} = 1$ if the category of the i-th document is known (i.e., specified by the i-th row of Y_1) and $(D_1)_{ii} = 0$ otherwise. Note that if $D_1 = I$, then we know the class orientation of all the documents and thus have a full specification of Y_1. Let the solution of F be F_1, which contains the knowledge to be transferred to X_2. This knowledge transfer is achieved by next solving the following 2-way clustering:

$$\min_{F,G,S,W_2} \|X_2 - FSG^T\|^2 + \beta\mathrm{Tr}\left[(GW_2 - Y_2)^T D_2(GW_2 - Y_2)\right]$$

$$(5.15) \qquad\qquad +\alpha\mathrm{Tr}\left[(F - F_1)^T V(F - F_1)\right]$$

The key part here is the third term which enforces the word sentiment polarity on X_2 to be approximately close to F_1 which is learnt from X_1; the extent of this approximation is determined by parameter $\alpha > 0$. The constraint only applies to the common terms in X_1, X_2 as enforced by the diagonal matrix V. The solution for Eq.(5.15) gives (F_2, S_2, G_2, W_2). This mechanism therefore transfers the document sentiment G_1 of domain X_1 to the document sentiment $G_2 W_2$ in domain X_2, via word sentiment polarities in F_1, schematically shown in Figure 1. It can be easily seen that the transfer mechanism is direct and explicit (via word sentiment polarities).
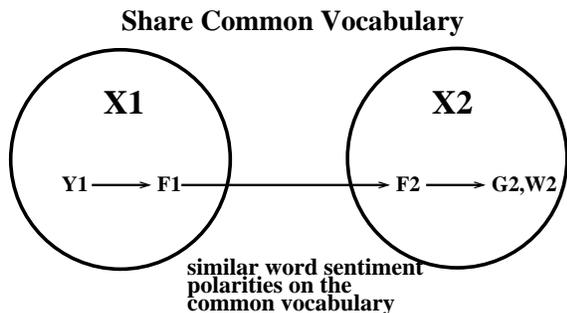
Figure 1: The Knowledge Transfer Mechanism

5.2 Algorithm Procedure The optimization in Eq.(5.14) can be solved using the update rules described in Section 4.2.

The optimization in Eq.(5.15) can be solved using the following update rule (update rules for G, S, W are the same as those in Eqs.(4.6, 4.7, 4.8)):

$$(5.16) \quad F_{ik} \leftarrow F_{ik} \frac{(XGS^T + \alpha V F_1)_{ik}}{[FF^T XGS^T + \alpha V F]_{ik}}$$

The algorithm consists of an iterative procedure using the above rules until convergence. The correctness and convergence of the updating rules can be rigorously proved using the standard auxiliary function approach [12].

6 Experiments

In this section, we perform two sets of experiments: the first set of experiments is to evaluate the performance of sentiment analysis based on matrix factorization model and the second set of experiments is used to perform cross-domain sentiment analysis.

6.1 Datasets Four different datasets are used in our experiments.

- **Movies Reviews**: This is a popular dataset in sentiment analysis literature [20]. It consists of 1000 positive and

1000 negative movie reviews drawn from the IMDB archive of the rec.arts.movies.reviews newsgroups.

- **Lotus blogs**: The data set is targeted at detecting sentiment around enterprise software, specifically pertaining to the IBM Lotus brand [23]. An unlabeled set of blog posts was created by randomly sampling 2000 posts from a universe of 14,258 blogs that discuss issues relevant to Lotus software. In addition to this unlabeled set, 145 posts were chosen for manual labeling. These posts came from 14 individual blogs, 4 of which are actively posting negative content on the brand, with the rest tending to write more positive or neutral posts. The data was collected by downloading the latest posts from each blogger's RSS feeds, or accessing the blog's archives. Manual labeling resulted in 34 positive and 111 negative examples.

- **Political Candidate blogs**: For our second blog domain, we used data gathered from 16,742 political blogs, which contain over 500,000 posts. As with the Lotus dataset, an unlabeled set was created by randomly sampling 2000 posts. 107 posts were chosen for labeling. A post was labeled as having positive or negative sentiment about a specific candidate (Barack Obama or Hillary Clinton) if it explicitly mentioned the candidate in positive or negative terms. This resulted in 49 positively and 58 negatively labeled posts.

- **Amazon Reviews**: The dataset contains product reviews taken from Amazon.com from 4 product types: Houseware-Kitchen (HK), Books, DVDs, and Electronics [3]. The dataset contains about 4000 positive reviews and 4000 negative reviews and can be obtained from `http://www.cis.upenn.edu/~mdredze/datasets/sentiment/`.

For all datasets, we picked 5000 words with highest document-frequency to generate the vocabulary. Stopwords were removed and a normalized term-frequency representation was used.

6.2 Matrix Factorization for Sentiment Analysis We use the algorithms described in Eqs.(4.5,4.6,4.7,4.8) for semi-supervised sentiment analysis. β, the corresponding parameter for enforcing document labels, are set to be 1. We compare our matrix factorization method (MF) with the following three semi-supervised approaches:

- (1) The algorithm proposed in [28] which conducts semi-supervised learning with local and global consistency (Consistency Method);

- (2) Zhu et al.'s harmonic Gaussian field method coupled with the Class Mass Normalization (Harmonic-CMN) [29];
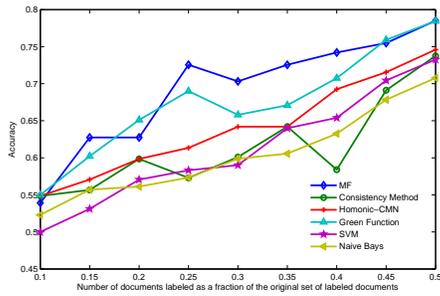
297

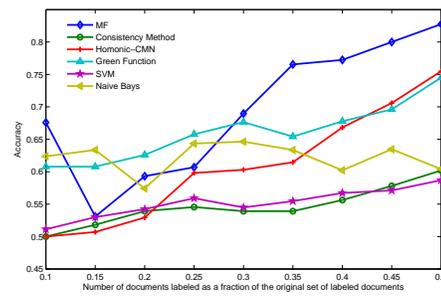Figure 2: Accuracy results with increasing number of labeled documents on Movies dataset

Figure 4: Accuracy results with increasing number of labeled documents on Political dataset
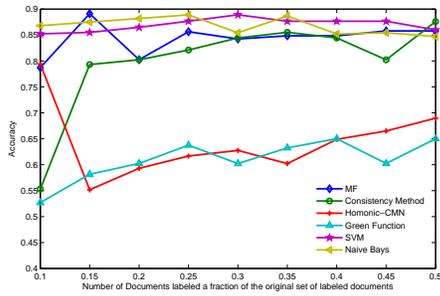

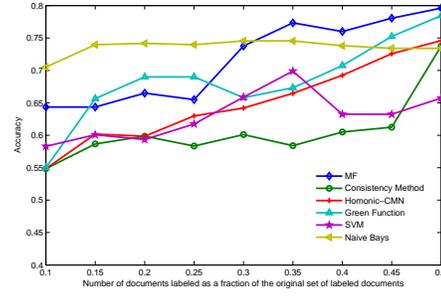Figure 3: Accuracy results with increasing number of labeled documents on Lotus dataset

Figure 5: Accuracy results with increasing number of labeled documents on Amazon dataset

- (3) Green's function learning algorithm (Green's Function) proposed in [6].

We also compare the results of our method with those of two supervised classification methods: Support Vector Machine (SVM) and Naive Bayes. Both of these methods have been widely used in sentiment analysis. In particular, the use of SVMs in [20] initially sparked interest in using machine learning methods for sentiment classification. The implementation of SVM is based on libSVM [4] and the implementation of Naive Bayes is based on the Weka software package [27].

The results are presented in Figure 2, Figure 3, Figure 4, and Figure 5. We note that our method either outperforms all other methods over the entire range of number of labeled documents (Movies, Political), or ultimately outpaces other methods (Lotus, Amazon) as a few document labels come in.
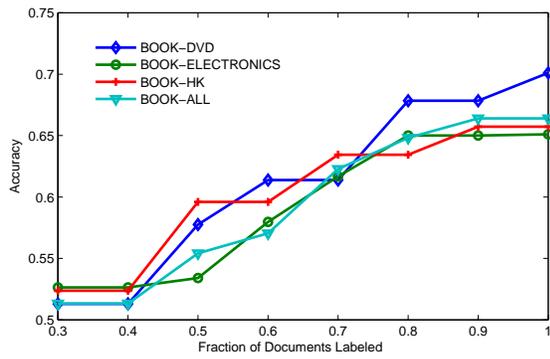
6.3 Cross-Domain Sentiment Analysis In this section, we perform cross-domain experiments on Amazon Reviews dataset which contains product reviews from different domains: Houseware-Kitchen (HK), Books, DVDs, and Electronics. Table 1 gives the characteristics of the dataset across four different domains.

In our experiments, the parameter α, which controls

the sentiment polarity approximation across two different domains, is set to be 1. Figure 6 shows the results from every source domain to the target domains. Note that in the subfigures, "All" refers all the documents in the collection excluding the training data. As can be seen, as supervision is increased in the source domain in the form of labeled data, performance gains effectively show up in the target domains also. We also note the effect of task relatedness: for example, DVD transfers more effectively than Books to Electronics but less effectively to Houseware-Kitchen.

We also take a close look at the words that are highly relevant to the sentiment analysis for each domain. For each source domain, the top 10 highest entropy words as per F_1 are as follows: (1) HK – *easy return love poor perfect excellent waste disappointed broke clean*, (2) Electronics: *return excellent price terrible waste perfect highly easy poor returned*, (3) DVD: *waste worst bad boring horrible love wonderful enjoy excellent ridiculous*, (4) BOOKS: *boring disappointing waste excellent bad wonderful poor poorly love easy*. The results demonstrate common word polarities among different domains.
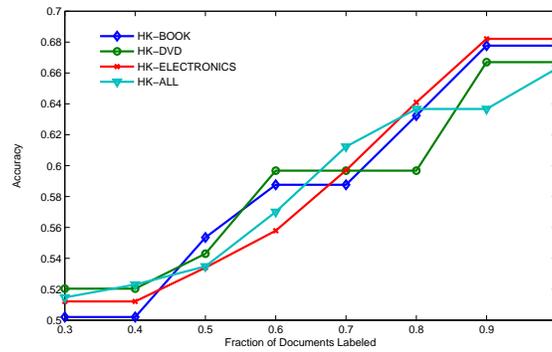
We also compare our method with SVM on cross-domain sentiment analysis. To apply SVM on cross-domain sentiment analysis, the terms appeared in all domains are used as features and the classifiers are built using the labeled
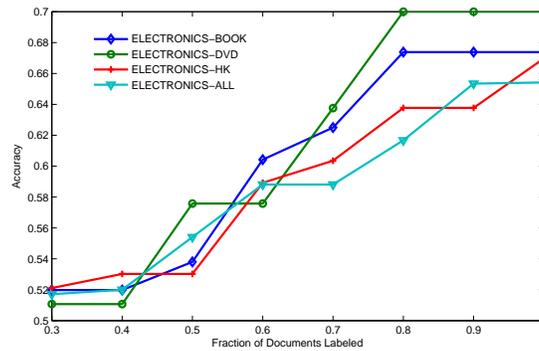
298

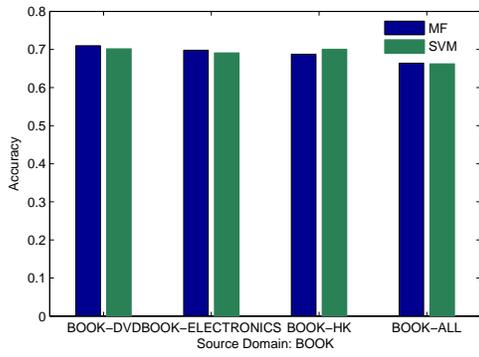(a) Source Domain: Book
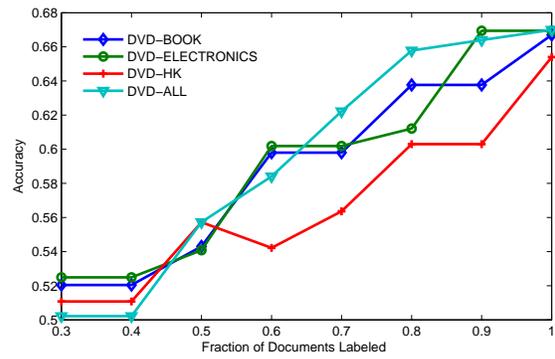
(b) Source Domain: DVD
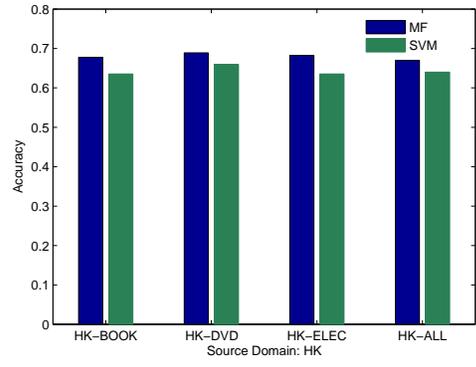
(c) Source Domain: HK
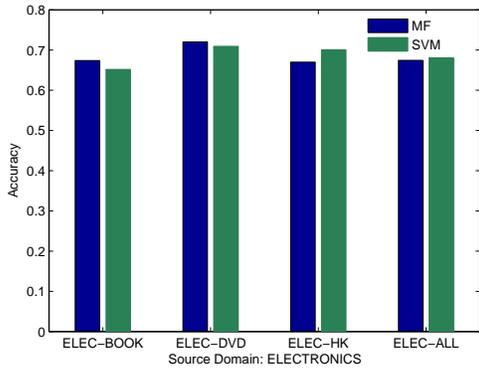
(d) Source Domain: Electronics

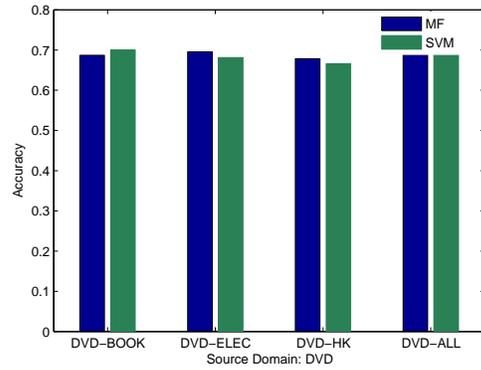Figure 6: Cross-Domain Sentiment Analysis Results

(a) Source Domain: Book
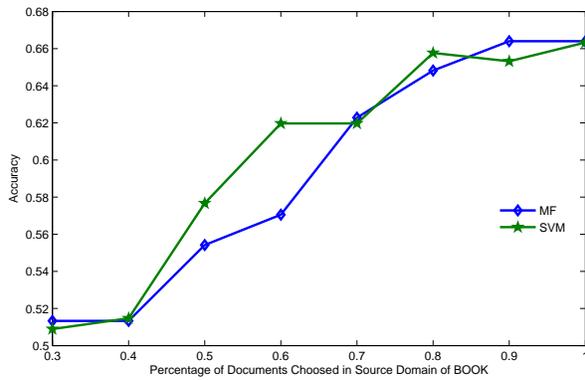

(b) Source Domain: HK

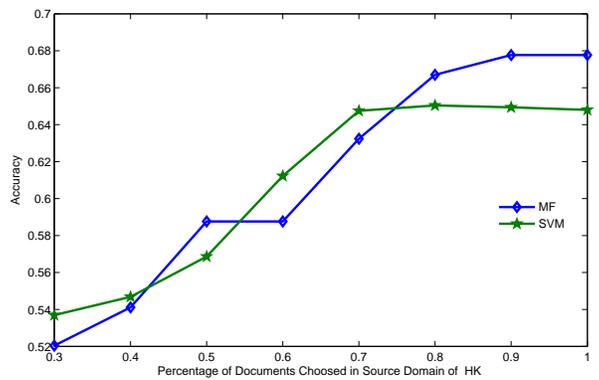(c) Source Domain: Electronics

(d) Source Domain: DVD

Figure 7: Performance comparison on different source domains

(a) Source Domain: Book; Target Domain: All

(b) Source Domain: HK; Target Domain: All

Figure 8: Performance comparison with increasing number of labeled documents

Table 1: Amazon Reviews Dataset Description

Domain	# Negative	# Positive
Book	980	1185
DVD	872	894
Electronics (Elec)	731	806
House & Kitchen (HK)	777	850

documents from the source domain only.

Figure 7 shows the performance comparison of our method and SVM for different source domains. We observe that our method for cross-domain sentiment analysis achieves competitive performance with SVM. Figure 8 shows the performance comparison of our method with SVM with increasing number of labeled documents where book and HK are used as source domains. The results on other cases are similar and thus are not included.

7 Conclusion

The primary contribution of this paper is to propose and benchmark new methodologies for cross-domain sentiment analysis. Non-negative Matrix Factorizations constitute a rich body of algorithms that have found applicability in a variety of machine learning applications: from recommender systems to document clustering. We have shown how to enable cross-domain sentiment analysis via a knowledge transfer mechanism based on constrained non-negative matrix trifactorizations of term-document matrices in the source and target domains. Document labels are naturally incorporated via a least squares penalty incurred by a certain linear model. This framework enables direct and explicit knowledge transfer via common word sentiment polarities across different domains. Several extensions are possible to extend our current model: benchmarking against several very recently proposed competing methodologies for sentiment analysis, incorporating hyperlinks between documents, and incorporating synonyms or co-occurences between words etc. These are topics for future work.

Appendix: Proof of Theorem 2

We prove the convergence of the iterative update algorithm of Eq. (4.10) by proving that $L(G)$ is monotonically decreasing (non-increasing) under the update Eq. (4.10).

We use the auxiliary function approach [12]. A function $Z(G, \tilde{G})$ is called an auxiliary function of $J(G)$ if it satisfies

$$(7.17) \qquad Z(G, \tilde{G}) \geq J(G), \quad Z(G, G) = J(G),$$

for any $G, \tilde{H}$. Define

$$(7.18) \qquad G^{(t+1)} = \arg\min_{G} Z(G, G^{(t)}),$$

where we note that we require the global minimum. By construction, we have $J(G^{(t)}) = Z(G^{(t)}, G^{(t)}) \geq$

$Z(G^{(t+1)}, G^{(t)}) \geq J(G^{(t+1)})$. Thus $J(G^{(t)})$ is monotone decreasing (non-increasing). The key is to find (1) appropriate $Z(G, \tilde{G})$ and (2) its global minimum.

The first step is to find an appropriate auxiliary function for $L(G)$. We can show that the following function

$$Z(G, G') = \sum_{ik}[-2G_{ik}Q_{ik} + \frac{(G'A + DG'B)_{ik}G'^2_{ik}}{G'_{ik}}]$$

is an auxiliary function for $L(G)$; i.e., it satisfies the requirements $L(G) \leq Z(G, G')$ and $L(G) = Z(G, G)$. This is true to the following

PROPOSITION 7.1. *For any matrices $A \in \mathbb{R}^{n \times n}_+, B \in \mathbb{R}^{k \times k}_+, G \in \mathbb{R}^{n \times k}_+, G' \in \mathbb{R}^{n \times k}_+$, with A and B symmetric, the following inequality holds:*

$$(7.19) \qquad \sum_{i=1}^{n}\sum_{p=1}^{k} \frac{(AG'B)_{ip}G^2_{ip}}{G'_{ip}} \geq Tr(G^T AGB).$$

The second step is to find the global maxima of $f(G) \equiv Z(G, G')$. To this end, we show that $Z(G, G')$ in convex in G. The gradient is

$$(7.20) \qquad \frac{\partial Z(G, G')}{\partial G_{ik}} = -2Q_{ik} + \frac{2(G'A + DG'B)_{ik}G_{ik}}{G'_{ik}}.$$

The Hessian matrix containing the second derivatives

$$\frac{\partial^2 Z(G, G')}{\partial G_{ik}\partial G_{j\ell}} = \delta_{ij}\delta_{k\ell} \frac{2(G'A + DG'B)_{ik}}{G'_{ik}}$$

is a diagonal matrix with positive entries. Thus $Z(G, G')$ is a convex function of G. and there is a unique global minimum which can obtained by setting $\partial Z(G, G')/\partial G_{ik} = 0$ in Eq. (7.20). Solving for G, we obtain

$$(7.21) \qquad G_{ik} = G'_{ik} \frac{Q_{ik}}{(G'A + DG'B)_{ik}}.$$

According to Eq. (7.18), $G^{(t+1)} \leftarrow G$ and $G^{(t)} \leftarrow G'$. we recover Eq.(4.6). □

References

[1] A. Aue and M. Gamon. Customizing sentiment classifiers to new domains: a case study. In *http://research.microsoft.com/ anthaue*, 2005.

[2] S. Ben-David, , J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *Proceedings of NIPS*, 2007.

[3] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, 2007.

[4] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[5] S. Das and M. Chen. Yahoo! for amazon: Extracting market sentiment from stock message boards. In *Proceedings of the 8th Asia Pacific Finance Association (APFA)*, 2001.

[6] C. Ding, R. Jin, T. Li, and H. D. Simon. A learning framework using green's function and kernel regularization with application to recommender system. In *Proceedings of ACM SIGKDD*, pages 260–269, 2007.

[7] C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proceedings of ACM SIGKDD*, pages 126–135, 2006.

[8] A. Goldberg and X. Zhu. Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization. In *HLT-NAACL 2006: Workshop on Textgraphs*, 2006.

[9] T. Hofmann. Probabilistic latent semantic indexing. *Proceeding of SIGIR*, pages 50–57, 1999.

[10] M. Hu and B. Liu. Mining and summarizing customer reviews. In *KDD*, pages 168–177, 2004.

[11] S.-M. Kim and E. Hovy. Determining the sentiment of opinions. In *Proceedings of International Conference on Computational Linguistics*, 2004.

[12] D. Lee and H. Seung. Algorithms for non-negative matrix factorization. In *NIPS*, 2001.

[13] T. Li, C. Ding, Y. Zhang, and B. Shao. Knowledge transformation from word space to document space. In *Proceedings of SIGIR*, pages 187–194, 2008.

[14] T. Li, V. Sindhwani, C. Ding, and Y. Zhang. Knowledge transformation for cross-domain sentiment classification. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 716–717, 2009.

[15] T. Li, Y. Zhang, and V. Sindhwani. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In *Association of Computational Lingustics (ACL)*, 2009.

[16] P. Melville, W. Gryc, and R. Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th Conference on Knowledge Discovery and Data Mining (KDD-09)*, 2009.

[17] V. Ng, S. Dasgupta, and S. M. N. Arifin. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In *COLING & ACL*, 2006.

[18] B. Pang and L. Lee. *Opinion mining and sentiment analysis*. Foundations and Trends in Information Retrieval: Vol. 2: No 1, pp 1-135.

[19] B. Pang and L. Lee. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *ACL*, 2004.

[20] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *EMNLP*, 2002.

[21] T. Sandler, J. Blitzer, P. Talukdar, and L. Ungar. Regularized learning with networks of features. In *NIPS*, 2008.

[22] V. Sindhwani, J. Hu, and A. Mojsilovic. Regularized co-clustering with dual supervision. In *Proceedings of NIPS*, 2008.

[23] V. Sindhwani and P. Melville. Document-word co-regularization for semi-supervised sentiment analysis. In *Proceedings of IEEE ICDM*, 2008.

[24] S. Tan, X. Cheng, Y. Wang, and H. Xu. Adapting naive bayes to domain adaptation for sentiment analysis. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieva*, 2009.

[25] S. Tan, G. Wu, H. Tang, and X. Cheng. A novel scheme for domain-transfer problem in the context of sentiment analysis. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieva*, 2009.

[26] P. Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of ACL*, pages 417–424, 2002.

[27] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2005.

[28] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf. Learning with local and global consistency. In *Proceedings of NIPS*, 2003.

[29] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of ICML*, 2003.

[30] L. Zhuang, F. Jing, and X.-Y. Zhu. Movie review mining and summarization. In *CIKM*, pages 43–50, 2006.