A Non-negative Matrix Tri-factorization Approach to Sentiment Classification with Lexical Prior Knowledge

Tao Li   Yi Zhang
School of Computer Science
Florida International University
{taoli,yzhan004}@cs.fiu.edu

Vikas Sindhwani
Mathematical Sciences
IBM T.J. Watson Research Center
vsindhw@us.ibm.com

Abstract

Sentiment classification refers to the task of automatically identifying whether a given piece of text expresses positive or negative opinion towards a subject at hand. The proliferation of user-generated web content such as blogs, discussion forums and online review sites has made it possible to perform large-scale mining of public opinion. Sentiment modeling is thus becoming a critical component of market intelligence and social media technologies that aim to tap into the collective wisdom of crowds. In this paper, we consider the problem of learning high-quality sentiment models with minimal manual supervision. We propose a novel approach to learn from lexical prior knowledge in the form of domain-independent sentiment-laden terms, in conjunction with domain-dependent unlabeled data and a few labeled documents. Our model is based on a constrained non-negative tri-factorization of the term-document matrix which can be implemented using simple update rules. Extensive experimental studies demonstrate the effectiveness of our approach on a variety of real-world sentiment prediction tasks.

1 Introduction

Web 2.0 platforms such as blogs, discussion forums and other such social media have now given a public voice to every consumer. Recent surveys have estimated that a massive number of internet users turn to such forums to collect recommendations for products and services, guiding their own choices and decisions by the opinions that other consumers have publically expressed. Gleaning insights by monitoring and analyzing large amounts of such user-generated data is thus becoming a key competitive differentiator for many companies. While tracking brand perceptions in traditional media is hardly a new challenge, handling the unprecedented scale of unstructured user-generated web content requires new methodologies. These methodologies are likely to be rooted in natural language processing and machine learning techniques.

Automatically classifying the sentiment expressed in a blog around selected topics of interest is a canonical machine learning task in this discussion. A standard approach would be to manually label documents with their sentiment orientation and then apply off-the-shelf text classification techniques. However, sentiment is often conveyed with subtle linguistic mechanisms such as the use of sarcasm and highly domain-specific contextual cues. This makes manual annotation of sentiment time consuming and error-prone, presenting a bottleneck in learning high quality models. Moreover, products and services of current focus, and associated community of bloggers with their idiosyncratic expressions, may rapidly evolve over time causing models to potentially lose performance and become stale. This motivates the problem of learning robust sentiment models from minimal supervision.

In their seminal work, (Pang et al., 2002) demonstrated that supervised learning significantly outperformed a competing body of work where hand-crafted dictionaries are used to assign sentiment labels based on relative frequencies of positive and negative terms. As observed by (Ng et al., 2006), most semi-automated dictionary-based approaches yield unsatisfactory lexicons, with either high coverage and low precision or vice versa. However, the treatment of such dictionaries as forms of prior knowledge that can be incorporated in machine learning models is a relatively less explored topic; even lesser so in conjunction with semi-supervised models that attempt to utilize un-

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 244–252,
Suntec, Singapore, 2-7 August 2009. ©2009 ACL and AFNLP

labeled data. This is the focus of the current paper.

Our models are based on a constrained non-negative tri-factorization of the term-document matrix, which can be implemented using simple update rules. Treated as a set of *labeled features*, the sentiment lexicon is incorporated as one set of constraints that enforce domain-independent prior knowledge. A second set of constraints introduce domain-specific supervision via a few document labels. Together these constraints enable learning from partial supervision along both dimensions of the term-document matrix, in what may be viewed more broadly as a framework for incorporating dual-supervision in matrix factorization models. We provide empirical comparisons with several competing methodologies on four, very different domains – blogs discussing enterprise software products, political blogs discussing US presidential candidates, amazon.com product reviews and IMDB movie reviews. Results demonstrate the effectiveness and generality of our approach.

The rest of the paper is organized as follows. We begin by discussing related work in Section 2. Section 3 gives a quick background on Non-negative Matrix Tri-factorization models. In Section 4, we present a constrained model and computational algorithm for incorporating lexical knowledge in sentiment analysis. In Section 5, we enhance this model by introducing document labels as additional constraints. Section 6 presents an empirical study on four datasets. Finally, Section 7 concludes this paper.

2 Related Work

We point the reader to a recent book (Pang and Lee, 2008) for an in-depth survey of literature on sentiment analysis. In this section, we briskly cover related work to position our contributions appropriately in the sentiment analysis and machine learning literature.

Methods focussing on the use and generation of dictionaries capturing the sentiment of words have ranged from manual approaches of developing domain-dependent lexicons (Das and Chen, 2001) to semi-automated approaches (Hu and Liu, 2004; Zhuang et al., 2006; Kim and Hovy, 2004), and even an almost fully automated approach (Turney, 2002). Most semi-automated approaches have met with limited success (Ng et al., 2006) and supervised learning models have tended to outperform dictionary-based classification schemes (Pang et al., 2002). A two-tier scheme (Pang and Lee, 2004) where sentences are first classified as *subjective* versus *objective*, and then applying the sentiment classifier on only the *subjective* sentences further improves performance. Results 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 driving the polarity prediction stage with minimal manual effort.

In this regard, our model brings two inter-related but distinct themes from machine learning to bear on this problem: *semi-supervised learning* and *learning from labeled features*. The goal of the former theme is to learn from few labeled examples by making use of unlabeled data, while the goal of the latter theme is to utilize weak prior knowledge about term-class affinities (e.g., the term "awful" indicates negative sentiment and therefore may be considered as a negatively labeled feature). Empirical results in this paper demonstrate that simultaneously attempting both these goals in a single model leads to improvements over models that focus on a single goal. (Goldberg and Zhu, 2006) adapt semi-supervised graph-based methods for sentiment analysis but do not incorporate lexical prior knowledge in the form of labeled features. Most work in machine learning literature on utilizing labeled features has focused on using them to generate weakly labeled examples that are then used for standard supervised learning: (Schapire et al., 2002) propose one such framework for boosting logistic regression; (Wu and Srihari, 2004) build a modified SVM and (Liu et al., 2004) use a combination of clustering and EM based methods to instantiate similar frameworks. By contrast, we incorporate lexical knowledge directly as constraints on our matrix factorization model. In recent work, Druck et al. (Druck et al., 2008) constrain the predictions of a multinomial logistic regression model on unlabeled instances in a Generalized Expectation formulation for learning from labeled features. Unlike their approach which uses only unlabeled instances, our method uses both labeled and unlabeled documents in conjunction with labeled and unlabeled words.

The matrix tri-factorization models explored in this paper are closely related to the models proposed recently in (Li et al., 2008; Sindhwani et al., 2008). Though, their techniques for proving algorithm convergence and correctness can be readily adapted for our models, (Li et al., 2008) do not incorporate dual supervision as we do. On the other hand, while (Sindhwani et al., 2008) do incorporate dual supervision in a non-linear kernel-based setting, they do not enforce non-negativity or orthogonality – aspects of matrix factorization models that have shown benefits in prior empirical studies, see e.g., (Ding et al., 2006).

We also note the very recent work of (Sindhwani and Melville, 2008) which proposes a dual-supervision model for semi-supervised sentiment analysis. In this model, bipartite graph regularization is used to diffuse label information along both sides of the term-document matrix. Conceptually, their model implements a co-clustering assumption closely related to Singular Value Decomposition (see also (Dhillon, 2001; Zha et al., 2001) for more on this perspective) while our model is based on Non-negative Matrix Factorization. In another recent paper (Sandler et al., 2008), standard regularization models are constrained using graphs of word co-occurences. These are very recently proposed competing methodologies, and we have not been able to address empirical comparisons with them in this paper.

Finally, recent efforts have also looked at transfer learning mechanisms for sentiment analysis, e.g., see (Blitzer et al., 2007). While our focus is on single-domain learning in this paper, we note that cross-domain variants of our model can also be orthogonally developed.

3 Background

3.1 Basic Matrix Factorization Model

Our proposed models are based on non-negative matrix Tri-factorization (Ding et al., 2006). 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-classes:

$$X \approx FSG^T. \tag{1}$$

where F is an $m \times k$ 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 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 classes, and S is an $k \times k$ nonnegative matrix providing a condensed view of X.

The matrix factorization model is similar to the probabilistic latent semantic indexing (PLSI) model (Hofmann, 1999). 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

$$\bar{X} \approx WSD^T, \sum_k W_{ik} = 1, \sum_k D_{jk} = 1, \sum_k S_{kk} = 1. \tag{2}$$

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

$$F^T F = I, \; G^T G = I. \tag{3}$$

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., (Ding et al., 2006). Our goal now is to bias these models with constraints incorporating (a) labels of features (coming from a domain-independent sentiment lexicon), and (b) labels of documents for the purposes of domain-specific adaptation. These enhancements are addressed in Sections 4 and 5 respectively.

4 Incorporating Lexical Knowledge

We used a sentiment lexicon generated by the IBM India Research Labs that was developed for other text mining applications (Ramakrishnan et al., 2003). It contains 2,968 words that have been human-labeled as expressing positive or negative sentiment. In total, there are 1,267 positive (e.g. “great”) and 1,701 negative (e.g., “bad”) unique terms after stemming. We eliminated terms that were ambiguous and dependent on context, such as “dear” and “fine”. It should be noted, that this list was constructed without a specific domain in mind; which is further motivation for using training examples and unlabeled data to learn domain specific connotations.

Lexical knowledge in the form of the polarity of terms in this lexicon can be introduced in the matrix factorization model. By partially specifying term polarities via F, the lexicon influences the sentiment predictions G over documents.

4.1 Representing Knowledge in Word Space

Let F_0 represent prior knowledge about sentiment-laden words in the lexicon, i.e., if word i is a positive word $(F_0)_{i1} = 1$ while if it is negative $(F_0)_{i2} = 1$. Note that one may also use soft sentiment polarities though our experiments are conducted with hard assignments. This information is incorporated in the tri-factorization model via a squared loss term,

$$\min_{F,G,S} \|X - FSG^T\|^2 + \alpha \mathrm{Tr}\left[(F - F_0)^T C_1 (F - F_0)\right] \tag{4}$$

where the notation $\mathrm{Tr}(A)$ means trace of the matrix A. Here, $\alpha > 0$ is a parameter which determines the extent to which we enforce $F \approx F_0$, C_1 is a $m \times m$ diagonal matrix whose entry $(C_1)_{ii} = 1$ if the category of the i-th word is known (i.e., specified by the i-th row of F_0) and $(C_1)_{ii} = 0$ otherwise. The squared loss terms ensure that the solution for F in the otherwise unsupervised learning problem be close to the prior knowledge F_0. Note that if $C_1 = I$, then we know the class orientation of all the words and thus have a full specification of F_0, Eq.(4) is then reduced to

$$\min_{F,G,S} \|X - FSG^T\|^2 + \alpha \|F - F_0\|^2 \tag{5}$$

The above model is generic and it allows certain flexibility. For example, in some cases, our prior knowledge on F_0 is not very accurate and we use smaller α so that the final results are not dependent on F_0 very much, i.e., the results are mostly unsupervised learning results. In addition, the introduction of C_1 allows us to incorporate partial knowledge on word polarity information.

4.2 Computational Algorithm

The optimization problem in Eq.(4) can be solved using the following update rules

$$G_{jk} \leftarrow G_{jk} \frac{(X^T FS)_{jk}}{(GG^T X^T FS)_{jk}}, \tag{6}$$

$$S_{ik} \leftarrow S_{ik} \frac{(F^T XG)_{ik}}{(F^T FSG^T G)_{ik}}. \tag{7}$$

$$F_{ik} \leftarrow F_{ik} \frac{(XGS^T + \alpha C_1 F_0)_{ik}}{(FF^T XGS^T + \alpha C_1 F)_{ik}}. \tag{8}$$

The algorithm consists of an iterative procedure using the above three rules until convergence. We call this approach Matrix Factorization with Lexical Knowledge (MFLK) and outline the precise steps in the table below.

Algorithm 1 Matrix Factorization with Lexical Knowledge (MFLK)

```
begin
1. Initialization:
  Initialize $F = F_0$
  $G$ to K-means clustering results,
  $S = (F^T F)^{-1} F^T XG(G^T G)^{-1}$.
2. Iteration:
  Update G: fixing $F, S$, updating $G$
  Update F: fixing $S, G$, updating $F$
  Update S: fixing $F, G$, updating $S$
end
```

4.3 Algorithm Correctness and Convergence

Updating F, G, S using the rules above leads to an asymptotic convergence to a local minima. This can be proved using arguments similar to (Ding et al., 2006). We outline the proof of correctness for updating F since the squared loss term that involves F is a new component in our models.

Theorem 1 *The above iterative algorithm converges.*

Theorem 2 *At convergence, the solution satisfies the Karuch, Kuhn, Tucker optimality condition, i.e., the algorithm converges correctly to a local optima.*

Theorem 1 can be proved using the standard auxiliary function approach used in (Lee and Seung, 2001).

Proof of Theorem 2. Following the theory of constrained optimization (Nocedal and Wright, 1999), we minimize the following function

$$L(F) = ||X - FSG^T||^2 + \alpha \mathrm{Tr}\left[(F-F_0)^T C_1 (F - F0)\right]$$

Note that the gradient of L is,

$$\frac{\partial L}{\partial F} = -2XGS^T + 2FSG^T GS^T + 2\alpha C_1(F - F_0). \tag{9}$$

The KKT complementarity condition for the non-negativity of F_{ik} gives

$$[-2XGS^T + FSG^T GS^T + 2\alpha C_1(F - F_0)]_{ik} F_{ik} = 0. \tag{10}$$

This is the fixed point relation that local minima for F must satisfy. Given an initial guess of F, the successive update of F using Eq.(8) will converge to a local minima. At convergence, we have

$$F_{ik} = F_{ik} \frac{(XGS^T + \alpha C_1 F_0)_{ik}}{(FF^T XGS^T + \alpha C_1 F)_{ik}}.$$

which is equivalent to the KKT condition of Eq.(10). The correctness of updating rules for G in Eq.(6) and S in Eq.(7) have been proved in (Ding et al., 2006). □

Note that we do not enforce exact orthogonality in our updating rules since this often implies softer class assignments.

5 Semi-Supervised Learning With Lexical Knowledge

So far our models have made no demands on human effort, other than unsupervised collection of the term-document matrix and a one-time effort in compiling a domain-independent sentiment lexicon. We now 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 G_0 where $(G_0)_{i1} = 1$ if the document expresses positive sentiment, and $(G_0)_{i2} = 1$ for negative sentiment. As with F_0, one can also use soft sentiment labeling for documents, though our experiments are conducted with hard assignments.

Therefore, the semi-supervised learning with lexical knowledge can be described as

$$\min_{F,G,S} ||X - FSG^T||^2 + \alpha \mathrm{Tr}\left[(F-F_0)^T C_1 (F - F_0)\right] + \beta \mathrm{Tr}\left[(G-G_0)^T C_2 (G - G_0)\right]$$

Where $\alpha > 0, \beta > 0$ are parameters which determine the extent to which we enforce $F \approx F_0$ and $G \approx G_0$ respectively, C_1 and C_2 are diagonal matrices indicating the entries of F_0 and G_0 that correspond to labeled entities. The squared loss terms ensure that the solution for F, G, in the otherwise unsupervised learning problem, be close to the prior knowledge F_0 and G_0.

5.1 Computational Algorithm

The optimization problem in Eq.(4) can be solved using the following update rules

$$G_{jk} \leftarrow G_{jk} \frac{(X^T FS + \beta C_2 G_0)_{jk}}{(GG^T X^T FS + \beta GG^T C_2 G_0)_{jk}} \tag{11}$$

$$S_{ik} \leftarrow S_{ik} \frac{(F^T XG)_{ik}}{(F^T FSG^T G)_{ik}}. \tag{12}$$

$$F_{ik} \leftarrow F_{ik} \frac{(XGS^T + \alpha C_1 F_0)_{ik}}{(FF^T XGS^T + \alpha C_1 F)_{ik}}. \tag{13}$$

Thus the algorithm for semi-supervised learning with lexical knowledge based on our matrix factorization framework, referred as SSMFLK, consists of an iterative procedure using the above three rules until convergence. The correctness and convergence of the algorithm can also be proved using similar arguments as what we outlined earlier for MFLK in Section 4.3.

A quick word about computational complexity. The term-document matrix is typically very sparse with $z \ll nm$ non-zero entries while k is typically also much smaller than n, m. By using sparse matrix multiplications and avoiding dense intermediate matrices, the updates can be very efficiently and easily implemented. In particular, updating F, S, G each takes $O(k^2(m+n) + kz)$ time per iteration which scales linearly with the dimensions and density of the data matrix. Empirically, the number of iterations before practical convergence is usually very small (less than 100). Thus, computationally our approach scales to large datasets even though our experiments are run on relatively small-sized datasets.

6 Experiments

6.1 Datasets Description

Four different datasets are used in our experiments.

Movies Reviews: This is a popular dataset in sentiment analysis literature (Pang et al., 2002). 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 (Sindhwani and Melville, 2008). 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: Kitchen, Books, DVDs, and Electronics (Blitzer et al., 2007). 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. Genuinely unlabeled posts for Political and Lotus were used for semi-supervised learning experiments in section 6.3; they were not used in section 6.2 on the effect of lexical prior knowledge. In the experiments, we set α, the parameter determining the extent to which to enforce the feature labels, to be 1/2, and β, the corresponding parameter for enforcing document labels, to be 1.

6.2 Sentiment Analysis with Lexical Knowledge

Of course, one can remove all burden on human effort by simply using unsupervised techniques. Our interest in the first set of experiments is to explore the benefits of incorporating a sentiment lexicon over unsupervised approaches. Does a one-time effort in compiling a domain-independent dictionary and using it for different sentiment tasks pay off in comparison to simply using unsupervised methods? In our case, matrix tri-factorization and other co-clustering methods form the obvious unsupervised baseline for comparison and so we start by comparing our method (MFLK) with the following methods:

- Four document clustering methods: K-means, Tri-Factor Nonnegative Matrix Factorization (TNMF) (Ding et al., 2006), Information-Theoretic Co-clustering (ITCC) (Dhillon et al., 2003), and Euclidean Co-clustering algorithm (ECC) (Cho et al., 2004). These methods do not make use of the sentiment lexicon.

- Feature Centroid (FC): This is a simple dictionary-based baseline method. Recall that each word can be expressed as a "bag-of-documents" vector. In this approach, we compute the centroids of these vectors, one corresponding to positive words and another corresponding to negative words. This yields a two-dimensional representation for documents, on which we then perform K-means clustering.

Performance Comparison Figure 1 shows the experimental results on four datasets using accuracy as the performance measure. The results are obtained by averaging 20 runs. It can be observed that our MFLK method can effectively utilize the lexical knowledge to improve the quality of sentiment prediction.

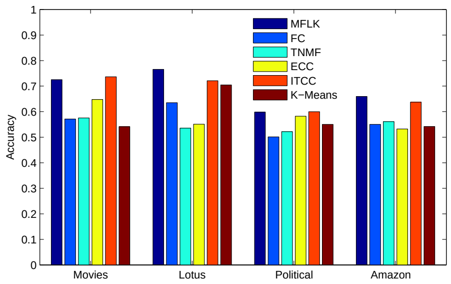

Figure 1: Accuracy results on four datasets

Size of Sentiment Lexicon We also investigate the effects of the size of the sentiment lexicon on the performance of our model. Figure 2 shows results with random subsets of the lexicon of increasing size. We observe that generally the performance increases as more and more lexical supervision is provided.

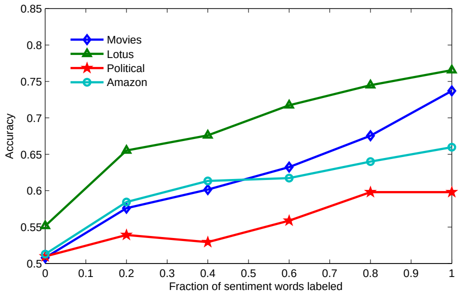

Figure 2: MFLK accuracy as size of sentiment lexicon (i.e., number of words in the lexicon) increases on the four datasets

Robustness to Vocabulary Size High dimensionality and noise can have profound impact on the comparative performance of clustering and semi-supervised learning algorithms. We simulate scenarios with different vocabulary sizes by selecting words based on information gain. It should, however, be kept in mind that in a truely unsupervised setting document labels are unavailable and therefore information gain cannot be practically computed. Figure 3 and Figure 4 show results for Lotus and Amazon datasets respectively and are representative of performance on other datasets. MLFK tends to retain its position as the best performing method even at different vocabulary sizes. ITCC performance is also noteworthy given that it is a completely unsupervised method.

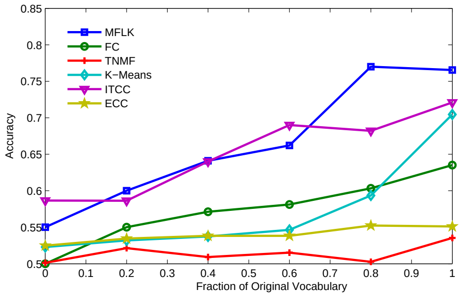

Figure 3: Accuracy results on Lotus dataset with increasing vocabulary size

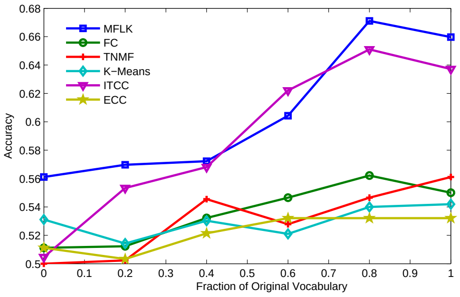

Figure 4: Accuracy results on Amazon dataset with increasing vocabulary size

6.3 Sentiment Analysis with Dual Supervision

We now assume that together with labeled features from the sentiment lexicon, we also have access to a few labeled documents. The natural question is whether the presence of lexical constraints leads to better semi-supervised models. In this section, we compare our method (SSMFLK) with the following three semi-supervised approaches: (1) The algorithm proposed in (Zhou et al., 2003) 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) (Zhu et al., 2003); and (3) Green's function learning algorithm (Green's Function) proposed in (Ding et al., 2007).

We also compare the results of SSMFLK 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 (Pang et al., 2002) initially sparked interest in using machine learning methods for sentiment classification. Note that none of these competing methods utilizes lexical knowledge.

The results are presented in Figure 5, Figure 6, Figure 7, and Figure 8. We note that our SSMFLK 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.

Learning Domain-Specific Connotations In our first set of experiments, we incorporated the sentiment lexicon in our models and learnt the sentiment orientation of words and documents via F, G factors respectively. In the second set of

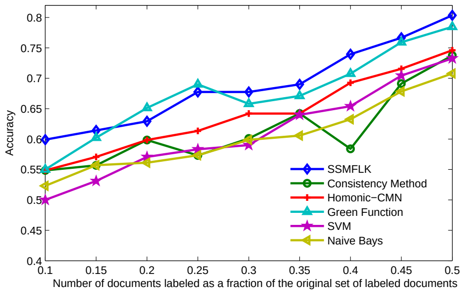

Figure 5: Accuracy results with increasing number of labeled documents on Movies dataset

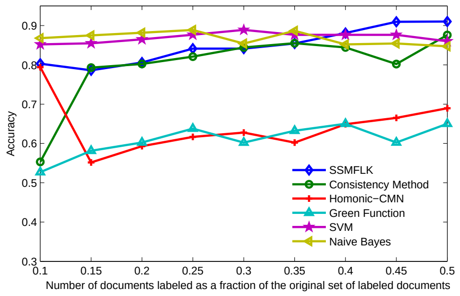

Figure 6: Accuracy results with increasing number of labeled documents on Lotus dataset

experiments, we additionally introduced labeled documents for domain-specific adjustments. Between these experiments, we can now look for words that switch sentiment polarity. These words are interesting because their domain-specific connotation differs from their lexical orientation. For amazon reviews, the following words switched polarity from positive to negative: *fan, important, learning, cons, fast, feature, happy, memory, portable, simple, small, work* while the following words switched polarity from negative to positive: *address, finish, lack, mean, budget, rent, throw*. Note that words like *fan, memory* probably refer to product or product components (i.e., computer fan and memory) in the amazon review context but have a very different connotation say in the context of movie reviews where they probably refer to movie fanfare and memorable performances. We were surprised to see *happy* switch polarity! Two examples of its negative-sentiment usage are: *I ended up buying a Samsung and I couldn't be more happy* and *BORING, not one single exciting thing about this book. I was happy when my lunch break ended so I could go back to work and stop reading*.

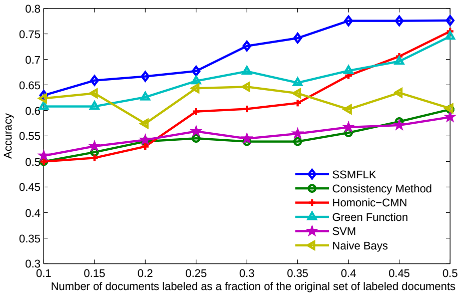

Figure 7: Accuracy results with increasing number of labeled documents on Political dataset

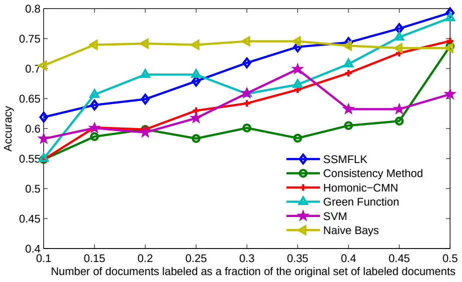

Figure 8: Accuracy results with increasing number of labeled documents on Amazon dataset

7 Conclusion

The primary contribution of this paper is to propose and benchmark new methodologies for 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 build effective sentiment models by appropriately constraining the factors using lexical prior knowledge and document annotations. To more effectively utilize unlabeled data and induce domain-specific adaptation of our models, several extensions are possible: facilitating learning from related domains, incorporating hyperlinks between documents, incorporating synonyms or co-occurences between words etc. As a topic of vigorous current activity, there are several very recently proposed competing methodologies for sentiment analysis that we would like to benchmark against. These are topics for future work.

Acknowledgement: The work of T. Li is partially supported by NSF grants DMS-0844513 and CCF-0830659. We would also like to thank Prem Melville and Richard Lawrence for their support.

References

J. Blitzer, M. Dredze, and F. Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of ACL*, pages 440–447.

H. Cho, I. Dhillon, Y. Guan, and S. Sra. 2004. Minimum sum squared residue co-clustering of gene expression data. In *Proceedings of The 4th SIAM Data Mining Conference*, pages 22–24, April.

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

I. S. Dhillon, S. Mallela, and D. S. Modha. 2003. Information-theoretical co-clustering. In *Proceedings of ACM SIGKDD*, pages 89–98.

I. S. Dhillon. 2001. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of ACM SIGKDD*.

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

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

G. Druck, G. Mann, and A. McCallum. 2008. Learning from labeled features using generalized expectation criteria. In *SIGIR*.

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

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

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

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

D.D. Lee and H.S. Seung. 2001. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 13*.

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

B. Liu, X. Li, W.S. Lee, and P. Yu. 2004. Text classification by labeling words. In *AAAI*.

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

J. Nocedal and S.J. Wright. 1999. *Numerical Optimization*. Springer-Verlag.

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

B. Pang and L. Lee. 2008. *Opinion mining and sentiment analysis*. Foundations and Trends in Information Retrieval: Vol. 2: No 12, pp 1-135 http://www.cs.cornell.edu/home/llee/opinion-mining-sentiment-analysis-survey.html.

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

G. Ramakrishnan, A. Jadhav, A. Joshi, S. Chakrabarti, and P. Bhattacharyya. 2003. Question answering via bayesian inference on lexical relations. In *ACL*, pages 1–10.

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

R.E. Schapire, M. Rochery, M.G. Rahim, and N. Gupta. 2002. Incorporating prior knowledge into boosting. In *ICML*.

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

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

P. Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424.

X. Wu and R. Srihari. 2004. Incorporating prior knowledge with weighted margin support vector machines. In *KDD*.

H. Zha, X. He, C. Ding, M. Gu, and H.D. Simon. 2001. Bipartite graph partitioning and data clustering. *Proceedings of ACM CIKM*.

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

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

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