

SUIT: A Supervised User-Item Based Topic Model for Sentiment Analysis

Fangtao Li¹, Sheng Wang², Shenghua Liu³ and Ming Zhang⁴

¹Google Inc. 1600 Amphitheatre Pkwy, Mountain View, CA

²Department of Computer Science, University of Illinois Urbana Champaign, Urbana, IL

³Institute of Computing Technology, Chinese Academy of Sciences, Beijing

⁴School of Electronics Engineering and Computer Science, Peking University, Beijing
lifangtao@google.com, swang141@illinois.edu, liu.shengh@gmail.com, mzhang@net.pku.edu.cn

Abstract

Probabilistic topic models have been widely used for sentiment analysis. However, most of existing topic methods only model the sentiment text, but do not consider the user, who expresses the sentiment, and the item, which the sentiment is expressed on. Since different users may use different sentiment expressions for different items, we argue that it is better to incorporate the user and item information into the topic model for sentiment analysis. In this paper, we propose a new Supervised User-Item based Topic model, called SUIT model, for sentiment analysis. It can simultaneously utilize the textual topic and latent user-item factors. Our proposed method uses the tensor outer product of text topic proportion vector, user latent factor and item latent factor to model the sentiment label generalization. Extensive experiments are conducted on two datasets: review dataset and microblog dataset. The results demonstrate the advantages of our model. It shows significant improvement compared with supervised topic models and collaborative filtering methods.

Introduction

Currently, there are many ways for people to express their sentiments or opinions on the Web. The reviews at commercial web sites, the comments on News sites and many other opinions on microblog, forum etc. are expressed everyday. These opinions are very useful for the general users, who often read product reviews before making the final decision. Companies also hope to track the customers' opinion to improve the quality of the products or services. Therefore, sentiment analysis, which aims to analyze the expressed sentiment or opinion towards different items, has become a popular topic in recent years (Pang and Lee, 2008; Liu 2010).

Since the sentiment expressions are dependent on different topics or themes, probabilistic topic model has

been widely used for sentiment analysis. Probabilistic topic models, such as PLSA (Hofmann 1999) and LDA (Blei et al., 2003), aim to analyze the words of the text content to discover themes. Various topic models (see the survey Liu, 2012) have been proposed and successfully used for sentiment analysis. Most of these methods model the sentiment in an unsupervised way, which mainly focus on the textual content without any sentiment labels (Mei et al. 2007; Titov and McDonald, 2008; Lin and He, 2009;). The recent studies show that the supervised topic models (Lu et al. 2008; Zhao et al. 2010; Wang et al. 2011) can model the sentiment analysis better if we observe some sentiment labels. The sentiment labels can guide the model to learn sentiment related topics.

However, most of these supervised topic models only consider the textual content, but seldom utilize the authors of the text and the items expressed in the text. We argue that different users may use different sentiment expressions for different items. First, sentiment is directly related to users. Different users may have their own preference for sentiment expressions. For example, some users may choose "good" to describe a just-so-so item, but other users may express excellent item with "good". Second, sentiment is also highly related to items. The exact same word may express different sentiment polarities for different items. For example, if the item is a movie, word "complex" may express the positive opinion towards the item. However, if the word "complex" is expressed on the item, which is a camera, it will belong to the negative sentiment polarity. Therefore, it is necessary to incorporate the user and item information into textual topic models for sentiment analysis.

Based on the above observations, we propose a novel Supervised User-Item based Topic model, SUIT model, which can simultaneously utilize the textual topic and user-item factors for sentiment analysis. In this model, the user-item information is represented in the form of user latent factor and item latent factor. Each dimension corresponds

to a user or an item. The content information of the sentiment expression is modeled in the form of topic proportion. The sentiment label in the proposed graphical model depends on the item, user and topic proportion of the sentiment document. We finally predict user’s sentiment label to a given expression based on the tensor outer product of user latent factor, item latent factor and the text topic proportion. We conduct experiment on both review dataset and microblog dataset. The experimental results prove that the proposed model outperformed the state-of-the-art methods.

The rest of the paper is organized as follows. We first briefly introduce the notations and basic supervised topic model in Section 2 and 3. In Section 4, we propose our new supervised topic model. The experimental results are shown on Section 5. In Section 6, we discuss the related work. Finally, conclusions and future work are discussed in Section 7.

Notations

For the sentiment analysis setting, we follow the notations from Li et al. (2011): A set of N users $\mathcal{A} = \{u_1, u_2, \dots, u_n\}$ write sentiment document on a set of M items $\mathcal{P} = \{i_1, i_2, \dots, i_m\}$. Here, we use item as the target which the sentiment is expressed on. Normally, a user would have only posted sentiment on a subset of the M items. Let $S \subseteq \mathcal{A} \times \mathcal{P}$ denote the set of user-item pairs for which the user has expressed sentiment on that item. In this paper, both user and item information can be represented by a latent factor as $U_i \in \mathbb{R}^K$ for user i and $V_j \in \mathbb{R}^K$ for item j , where K is the number of latent factors. The corresponding matrixes are denoted as $\mathbb{U} = (U_i)_{i=1}^n$ and $\mathbb{V} = (V_j)_{j=1}^m$.

Let $d \in D$ be a sentiment document, posted by user $u \in \mathcal{A}$, expressed on item $i \in \mathcal{P}$. In this paper, we proposed our SUI model to predict user’s sentiment preference to the item. In the proposed model, a document can be represented by a topic proportion vector, as $\theta_d \in \mathbb{R}^K$. We predict user’s sentiment label based on the tensor outer product of topic proportion, user latent factor and item latent factor. Therefore, the topic proportion vector has the same dimension as user latent factor and item latent factor.

For each document d , there is an associated label r , which indicates user’s sentiment polarity or strength towards an item. Our goal is to design a novel topic model to model the sentiment label generalization, with consideration of the authored user i and the target item j .

Basic Supervised Topic Model

We first introduce the basic supervised topic model for sentiment analysis. The supervised topic model is effective

for short-text analysis. It models the sentiment label in the topic level. One of the most widely used supervised topic models is supervised Latent Dirichlet Allocation (sLDA) (Blei and McAuliffe 2007).

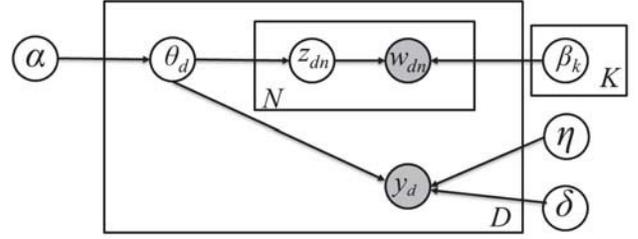


Figure 1: The graphic model for the sLDA model

Figure 1 is the graphical model representation of sLDA. Different from LDA, sLDA adds a response variable to each document. In sentiment analysis, the response variable y is user’s sentiment to the given item. α is the Dirichlet parameter, $\beta_{1:K}$ denotes the vector of term probabilities for each topic. η and δ are the response parameters. Under the sLDA model, each document and response arise from the following generative process:

1. Draw topic proportions $\theta | \alpha \sim \text{Dir}(\alpha)$
2. For each word
 - a) Draw topic assignment $z_n | \theta \sim \text{Mult}(\theta)$
 - b) Draw word $w_n | z_n, \beta_{1:K} \sim \text{Mult}(\beta_{z_n})$
3. Draw response variable $y | z_{1:N}, \eta, \delta^2 \sim \text{GLM}(\eta^T \bar{z}, \delta^2)$

where sLDA define

$$\bar{z} = \frac{1}{N} \sum_{n=1}^N z_n$$

The distribution of the response is a generalized linear model (McCullagh and Nelder, 1989),

$$p(y | z_{1:N}, \eta, \delta) = h(y, \eta) \exp \left\{ \frac{(\eta^T \bar{z})y - A(\eta^T \bar{z})}{\delta} \right\} \quad (1)$$

Supervised topic model only considers the textual content of sentiment document, but ignores the factor of users and items.

Proposed SUI Model

In this section, we will introduce our proposed Supervised User-Item based Topic model, called SUI model, for sentiment analysis. The motivation of SUI model is to incorporate user and item information into supervised topic model for sentiment analysis.

In this section, we propose to incorporate the user and item effects when predicting sentiment scales by designing a predictor function $f : \mathbb{R}^K \times \mathcal{A} \times \mathcal{P} \rightarrow \mathbb{R}$ that utilizes not only the review content θ_d but also the user identity i and item identity j . Due to the data sparsity and scale, we

cannot directly include each user and item into the supervised topic model. We also need to do the user-item clustering or dimension reduction. Here, we employ the Matrix Factorization idea from Recommender System (Koren et al. 2009). Matrix Factorization aims to learn the user latent factor and the item latent factor based on the user-item matrix. Similar users or items will have similar latent factors. Based on the observations from recommender system, we also believe that the user-item latent factor is important for sentiment analysis. We need to mine similar users based on their sentiment for the similar item. It is therefore necessary to refine the content-only topic model described in the previous section to accommodate user and item specific effects.

Based on the above discussion, we propose our supervised user-item based topic model. The graph representation of our model is shown in Figure 2. We can see that it can be considered as a natural integration of sLDA and Probabilistic Matrix Factorization (PMF) (Mnih and Ruslan, 2007). The upper part of the graph is a supervised topic model to model the textual information. The lower part of the graph is a probabilistic matrix factorization to model the user-item information. The sentiment label is dependent on both text information and user-item information.

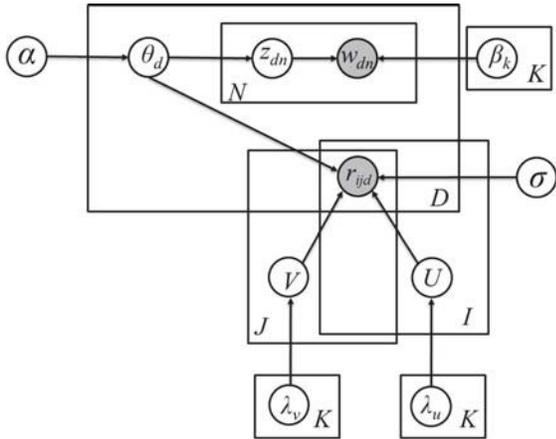


Figure 2: The graphic model for the SUIT model

The generative process of our model is as follows:

1. For each user i , draw user latent factor $U_i \sim \mathcal{N}(0, \lambda_u^{-1} I_K)$.
2. For each item j , draw item latent factor $V_j \sim \mathcal{N}(0, \lambda_v^{-1} I_K)$.
3. For each document d , expressed by user i for item j
 - a) Draw topic proportions $\theta_d | \alpha \sim \text{Dir}(\alpha)$
 - b) For each word $z_{dn} | \theta_d \sim \text{Mult}(\theta_d)$
 - i. Draw topic assignment
 - ii. Draw word $w_{dn} | z_n, \beta_{1:K} \sim \text{Mult}(\beta_{z_{dn}})$
4. For each user-item-document triple (i, j, d)

- a) Draw the sentiment label: $r_{ij} \sim \mathcal{N}(\langle U_i, V_j, \theta_d \rangle, \sigma^2)$

In our model, we use the tensor outer products (Lang, 2002) among textual topic distribution, user latent factor and item latent factor to compute the expectation of sentiment label.

$$\langle U_i, V_j, \theta_d \rangle = \sum_{f=1}^K U_{if} \cdot V_{jf} \cdot \theta_{df} \quad (2)$$

where K is the dimension of latent factor. The dimension of latent factor is equal to the number of topics in our model.

Learning

Since computing the full posterior of U_i, V_j and θ_d is intractable, we develop an EM-style algorithm similar to (Wang and Blei 2011) for learning the parameter. Maximization of the posterior is equivalent to maximize the complete log-likelihood. The log likelihood function of U_i, V_j and θ_d given β, λ_u and λ_v is

$$\begin{aligned} & -\frac{\lambda_u}{2} \sum_i U_i^T U_i - \frac{\lambda_v}{2} \sum_j V_j^T V_j \\ & + \sum_d \sum_n \log \sum_k \theta_{dk} \beta_{k, w_{dn}} \\ & - \frac{1}{2} \sum_{i,j} (r_{ij} - \langle u_i, v_j, \theta_d \rangle)^2 \end{aligned} \quad (3)$$

We set the Dirichlet parameter $\alpha = 1$ and omitted a constant term. We apply the coordinate ascent which iteratively optimizing the latent factors U_i, V_j and the topic proportion θ_d to optimize this likelihood function.

To optimize U_i and V_j , maximization follows similar to matrix factorization (Matsumoto, Takamura and Okumura 2005). Given the current value of θ_d , taking the gradient of \mathcal{L} with respect to U_i and V_j and setting it to zero leads to:

$$U_i \leftarrow \left((\mathbb{V} \otimes \theta_{u_i}) (\mathbb{V} \otimes \theta_{u_i})^T + \lambda_u I_K \right)^{-1} (\mathbb{V} \otimes \theta_{u_i}) R_i \quad (4)$$

$$V_j \leftarrow \left((\mathbb{U} \otimes \theta_{v_j}) (\mathbb{U} \otimes \theta_{v_j})^T + \lambda_v I_K \right)^{-1} (\mathbb{U} \otimes \theta_{v_j}) R_j \quad (5)$$

where $R_i = (r_{ij})_{j=1}^J$ for user i . \otimes denotes the element-wise matrix multiplication operation. For two matrixes $A \otimes B = C$, we have $C_{ij} = A_{ij} \cdot B_{ij}$. The θ_{u_i} consists of all the document topic proportion user U_i have written on

$$\theta_{u_i} = (\theta_{ij})_{j=1}^M$$

where θ_{ij} is the topic proportion of the document that user U_i has posted about item j . If user U_i has never posted about item j , we will set a random number to initialize θ_j .

Therefore, θ_{u_i} and \mathbb{V} have the same dimension. So we

can apply pairwise matrix multiplication on them. θ_{v_j} and R_j is defined in the same way.

We can see from Eq.(4) and Eq.(5) that topic proportion affects both user latent factor and item latent factor. λ_u and λ_v control the effect of topic proportion. Given U and V , we show how to learn the topic proportions.

We first define $q(z_{dn} = k) = \phi_{dnk}$ same as (Saif, He and Alani 2012). Then we separate the items that contain θ_d and apply the Jensen inequality to the likelihood function and get the tight lower bound of L .

$$\begin{aligned} & \mathcal{L}(\theta_d) \\ & \geq -\frac{1}{2} \sum_{i,j} (r_{ij} - \langle u_i, v_j, \theta_d \rangle)^2 \\ & + \sum_n \sum_k \phi_{dnk} (\log \theta_{dk} \beta_{k,wdn} - \log \phi_{dnk}) \\ & = \mathcal{L}(\theta_d, \phi_d) \end{aligned} \quad (6)$$

Let $\phi_d = (\phi_{dnk})_{n=1, k=1}^{N \times K}$. The optimal ϕ_{dnk} satisfies $\phi_{dnk} \propto \theta_{dk} \beta_{k,wdn}$.

Note that we cannot optimize θ_d analytically, so we use gradient projection to optimize $U, V, \theta_{1:j}$ and $\phi_{1:j}$. After we estimate U, V and β , we can optimize β

$$\beta_{kw} \propto \sum_d \sum_n \phi_{dnk} 1[w_{jn} = w] \quad (7)$$

The M-step of EM algorithm is the same in LDA and CTM (Blei, Ng and Jordan 2003; Saif, He and Alani 2012).

Prediction

After all the optimal parameters are learned, we can use the proposed model for prediction. Let \mathbb{D} be the observed data, in general each prediction is estimated as

$$\mathbb{E}[r_{ij} | \mathbb{D}] \approx \langle \mathbb{E}[U_i | \mathbb{D}], \mathbb{E}[V_j | \mathbb{D}], \mathbb{E}[\theta_d | \mathbb{D}] \rangle \quad (8)$$

We use the point estimate of u_i, θ_d, v_j to approximate the sentiment label expectation,

$$r_{ij} = \langle U_i, V_j, \theta_d \rangle.$$

We can see that the tensor outer product operation determines the value r_{ij} based on the pairwise interactions between the latent factors of all three entities: user, item and topic, which can naturally reflect how a user uses a topic when commenting on an item. Also we can see that the same user factor U_i are shared when computing r_{ij} values for different item and topic combinations, which effectively captures the possible correlations between r_{ij} values for the same user. Similarly, the sharing of item and term factors when determining r_{ij} for different user-item, user-topic combinations are achieved in same way.

Experiments

Dataset

We conduct our experiments on two datasets. The first is movie review data set. The second is a dataset crawled from a microblog site.

Review dataset

The first dataset is a collection of movie reviews. A subset of this collection has been used in (Pang and Lee 2005). We follow the instructions to parse the collection. We acquire 15507 reviews with explicit stars. Following Pang and Lee's setting (Pang and Lee 2005), all the review stars are mapped into a 1~4 sentiment scales. There are 458 users, 4543 products and 15507 reviews in total.

Microblog dataset

The second dataset is crawled from a Microblog site. In our experiment, we selected 14 movies, which are released in the last year. We first filter out the spam tweets, and then manually annotate the remaining tweets into three categories (0 ~ 2): negative (0), neutral (1), and positive (2). Negative, neutral and positive labels refer to the sentiment level that user has rated to the item. It is also necessary to filter out the spam users who only post advertisement. After filtering the spam tweets and the spam users, we finally got 387 users and 1299 tweets which include 403 negative tweets, 431 neutral tweets and 445 positive tweets.

Evaluation Metrics

In our experiments, we measure the performance of our method and baselines by comparing the result of our predicted labels. We use the Accuracy measure to evaluate the accuracy of sentiment analysis. We also use Root Mean Squared Error (RMSE) and the mean absolute error (MAE) measure to evaluate how much our predicted rating deviates from the true rating.

• **Accuracy**: Accuracy reports the proportion of document sentiment label that we predict correctly in the test set, which is defined as:

$$Accuracy = \frac{1}{n} \sum_{i=1}^k c(r_i) \quad (9)$$

where $c(r_i)$ is 1 if the rating of document is predicted correctly and 0 otherwise.

• **MAE & RMSE**: The mean absolute error (MAE) and the root mean squared error (RMSE) measure how much our predicted rating deviates from the true rating. A smaller value indicates a more accurate prediction.

$$MAE = \frac{\sum_i |p_i - r_i|}{n}$$

$$RMSE = \sqrt{\frac{\sum_i (p_i - r_i)^2}{n}}$$

where r_i is the true rating for the i -th document, p_i is the predicted rating for the i -th document.

Baselines

Besides the proposed method, we also implement several approaches for comparison.

Topic Level SVM (TopicSVM)

We first apply Latent Dirichlet Allocation (LDA) (Blei et al. 2003) on the text to obtain topic proportion. We then use these topic proportion vectors as the feature and run multi-class SVM as the classifier. LibSVM is used for the classification. This baseline doesn't use the user-item information and only utilize the content information.

- **Probabilistic Matrix Factorization (PMF)**

PMF(Koren et al. 2009) is the probabilistic version of Matrix Factorization. It uses a graphic model to learn the joint latent space for users and items. The final sentiment label depends on the dot product of user and item latent factor. PMF does not consider the text information in sentiment analysis, it only utilizes the user and item identity information.

- **Supervised Latent Dirichlet Allocation (sLDA)**

sLDA(Blei and McAuliffe, 2007) trains the topic proportion and topic distribution parameter based on user's rating and text information. It then predicts document's sentiment label based on these topic parameters and text information. sLDA does not consider the user and item information in the recommender system, it only considers the text information.

- **MedLDA**

MedLDA (Zhu et al. 2009) can be considered as the state-of-the-art supervised topic model. It utilizes the max-margin principle to train supervised topic models and estimate predictive topic representations. Same as sLDA, it only models the textual topic without consideration of the user and item information.

Table 1. Results on Movie Review Dataset

	TopicSVM	sLDA	PMF	MedLDA	SUIT
Accuracy	0.336	0.405	0.377	0.418	0.441
RMSE	1.263	1.003	1.106	0.995	0.984
MAE	0.942	0.726	0.768	0.713	0.697

Experimental Results on Review Dataset

Table 1 shows the sentiment label prediction result for movie review dataset. TopicSVM first trains a unsupervised topic model, and the predict the sentiment label with the SVM classifier. In this setting, TopicSVM doesn't get good performance. One reason is that the topics are learned without any sentiment labels, which may be not suitable for the sentiment label prediction task. sLDA obtains a better performance in comparison to TopicSVM, which proves the effectiveness of the supervised topic model. PMF only uses user-item information without consideration of the text, but it also beats the TopicSVM, which demonstrate the importance of user and item information for sentiment analysis. MedLDA, as the state-of-the-art supervised topic model, gets the best

performance among the baselines. Our proposed SUIT method, which models not only the textual information, but also the user and item information, achieves significant better results than MedLDA. It improves the accuracy by 5.5% (from 0.418 to 0.441).

Experimental Results on Microblogs Dataset

Table 2 shows the results for microblog dataset. Different from the review, microblog tends to be much shorter than review. Sometimes a microblog only contains one sentence. From Table 2, we can see that it has similar results as the review dataset. TopicSVM still doesn't get good performance. sLDA, which learns the latent topic with sentiment labels can achieve much better result than the unsupervised learned topic. PMF, with consideration of user and item information, still shows improvement compared with TopicSVM. MedLDA achieves much better result than other baselines. The proposed SUIT model achieves much more improvement for microblog dataset, compared to the review dataset. It improves the accuracy by 16.6% (from 0.438 to 0.511), which is a much larger improvement compared to review dataset. This may be because that it is important to incorporate the user and item information into topic model, when the textual information is not enough (microblog is much shorter than review).

To demonstrate the advantages of topic level analysis for short-text microblog, we also use another baseline word-level SVM (WordSVM): We extract words in the text as features, and employ SVM as the classifier for sentiment label prediction. The experimental result shows that when the text is short, supervised topic model can achieve better result than word level classification method.

Table 2. Results on Microblog Dataset

	Topic SVM	Word SVM	sLDA	PMF	MedLDA	SUIT
Accuracy	0.374	0.394	0.405	0.385	0.438	0.511
RMSE	1.105	1.101	1.037	1.116	0.995	0.988
MAE	0.824	0.796	0.756	0.817	0.714	0.672

Parameter Sensitivity:

In this section, we conduct parameter sensitivity analysis. Since the experiments show similar result for each metric and two datasets, to save space, we will only use accuracy as the evaluation metric and microblog as dataset for parameter analysis. One important advantage of topic model is that it can reduce the dimension of the latent space from original word space. Figure 3 shows the accuracy of our model in comparison with all the baselines in different numbers of latent factors. From Figure 3, we can see that most of topic models can achieve best results when the number of latent factors ranges from 20 to 40. The proposed SUIT model achieves best result when the

number is equal to 40. We also can see that for each number of factors, our proposed model can always outperform other models significantly.

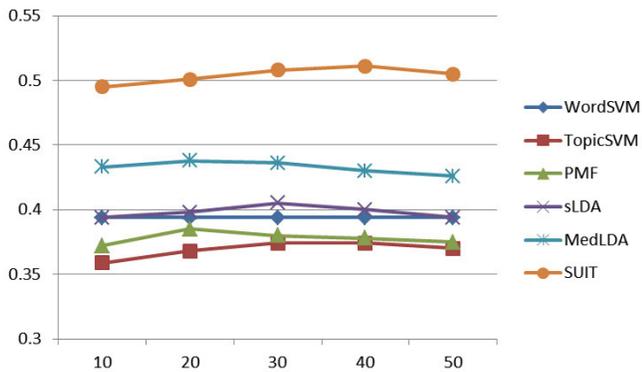


Figure 3. Impact of number of latent factors on accuracy

Related Work

Sentiment analysis is the computational study of opinions, sentiments and emotions expressed in text (Liu, 2010; Pang and Lee, 2009). Since the sentiment is related to topics or domains, probabilistic topic model (Hofmann 1999; Blei, Ng and Jordan 2003; Blei and McAuliffe 2007), which aims to analyze the words of the original texts to discover themes, has become a popular topic for sentiment analysis. Most of the methods model the sentiment in an unsupervised framework (Met et al. 2007; Brody and Elhadad, 2010; Li et al. 2010; Jo and Oh, 2011 etc). For example, Topic-Sentiment Model (TSM) (Mei et al, 2007) jointly models the mixture of topics and sentiment predictions. Multi-Grain Latent Dirichlet Allocation model (MG-LDA) (Titov and McDonald, 2008a) builds sentiment topics that are representative of ratable aspects of objects from online user reviews, by allowing terms being generated from either a global topic or a local topic. Joint Sentiment/Topic (JST) model (Lin and He, 2009), which is a four-layer topic model detects sentiment and topic simultaneously from text. The recent studies shows that if we observe some sentiment labels, the supervised topic model can model the sentiment analysis better. Several supervised variants of LDA model, including sLDA(Blei and McAuliffe 2007), DiscLDA(Simon et al. 2008), MedLDA(Zhu et al. 2009), LARA(Wang et al. 2011), achieve better results for review rating prediction. MaxEnt-LDA (Zhao et al. 2010) is a hybrid topic model to jointly discover both topic and topic-specific opinion words, which also shows that it can achieve accurate results with labeled training data. MAS (Titov and McDonald, 2008b) is a supervised extension to MG-LDA, which demonstrates the advantages of supervised topic model in multi-aspect sentiment analysis. However, all

above studies for sentiment analysis only model the textual information, but ignore the author of the text and the item expressed in the text. Our SUIT model can simultaneously model the text, user and item for sentiment analysis, which can model the sentiment label generation more accurate. Previous studies also demonstrate the importance of user and item information (Tan et al. 2010; Li et al. 2011). But they didn't model it in the framework of probabilistic topic models. In this paper, we proposed a probabilistic model with consideration of user and item.

We should emphasize that our proposed approach is different from content based collaborative filtering (CBCF) (Melville, Mooney and Nagarajan 2002; Purushotham et al. 2012; Wang and Blei 2011). The content in CBCF is the content related to only user or only item, which is the description of user or item, such as user profile, or item attribute. While in this paper, we aim to perform sentiment analysis for the content which the user expressed towards the item, which the user writes for the item. That is to say, the content information for CBCF is associated with only the user or the item whereas the content of user expression for sentiment analysis is associated with user-item pairs. For example, Wang and Blei (Wang and Blei 2011) propose a CBCF method, Collaborative Topic Model (CTM), for scientific paper recommendation, which only utilizes the content of the paper, i.e. the content of the item. However, in our task, the content refers to the sentiment document that user expressed on an item, where this content is related to a user-item pair. Therefore, the user-item based sentiment analysis is much different from the previous content-based recommendation, and cannot be handled by existing CBCF models.

Conclusion and Future Work

In this paper, we propose a novel supervised user-item based topic model, which can simultaneously model the textual topics and user-item latent factors for sentiment analysis. This model can be considered as integration between supervised LDA (sLDA) and Probabilistic Matrix Factorization (PMF). The sentiment label is generated by the tensor outer product of user latent factor, item latent factor and textual topic proportion vector. We conduct experiments on both review dataset and microblogs dataset. The experiment results show that our method achieves much better results than the state-of-the-art baselines, which also demonstrates the importance of both content information and user-item information.

In the future work, it is interesting to investigate more efficient algorithms for the SUIT model for large-scale data sets. We also plan to employ the proposed model for other applications, such as recommender system, computational ads.

Acknowledgements

This paper is partially supported by National Natural Science Foundation of China (No. 61202213, 61272343), National Grand Fundamental Research 973 Program of China (No.2014CB340401, 2012CB316303), the Doctoral Program of Higher Education of China under Grant (FSSP Grant No. 20130001110113).

References

- D. Blei, A. Ng; and A. Jordan, 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3:993–1022.
- D. Blei, and D. McAuliffe. 2007. Supervised topic models. In *NIPS-07*.
- Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10)*
- Thomas Hofmann. 1999. "Probabilistic latent semantic indexing." *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*.
- Y. Jo, and Alice H. Oh. 2011. "Aspect and sentiment unification model for online review analysis." *Proceedings of the fourth ACM international conference on Web search and data mining*.
- Koren, Yehuda, Robert Bell, and Chris Volinsky. 2009. "Matrix factorization techniques for recommender systems." *Computer* 42.8: 30-37.
- Lacoste-Julien, Simon, Fei Sha, and Michael I. Jordan. 2008. "DiscLDA: Discriminative learning for dimensionality reduction and classification." *NIPS-08*.
- Fangtao Li, Minlie Huang, and Xiaoyan Zhu. 2010. "Sentiment Analysis with Global Topics and Local Dependency." *AAAI*.
- Fangtao Li, Nathan Liu, Hongwei Jin, Kai Zhao, Qiang Yang and Xiaoyan Zhu. Incorporating Reviewer and Product Information for Review Rating Prediction. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*.
- Lang, Serge (2002), *Algebra*, Graduate Texts in Mathematics 211 (Revised third ed.), New York: Springer-Verlag
- Lin, C., & He, Y. (2009, November). Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*.
- Liu, Bing. "Sentiment analysis and subjectivity." *Handbook of natural language processing* 2 (2010): 568.
- Liu, Bing. 2012. "Sentiment analysis and opinion mining." *Synthesis Lectures on Human Language Technologies* 5.1 (2012).
- S. Matsumoto, H. Takamura and M.Okumura. 2005. Sentiment classification using word sub-sequences and dependency sub-trees. In *Proceedings of PAKDD'05, 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*.
- Yue Lu, and Chengxiang Zhai. 2008. "Opinion integration through semi-supervised topic modeling." *Proceedings of the 17th international conference on World Wide Web*. ACM.
- Q Mei, X Ling, M Wondra, H Su, CX Zhai. 2007. "Topic sentiment mixture: modeling facets and opinions in weblogs." *Proceedings of the 16th international conference on World Wide Web*. ACM.
- P. Melville, J. Mooney, and R. Nagarajan. 2002. Content-boosted collaborative filtering for improved recommendations. In *AAAI*, 187–192. Menlo Park, CA, USA.
- Mnih, Andriy, and Ruslan Salakhutdinov. 2007 "Probabilistic matrix factorization." *NIPS-07*
- B. Pang and L. Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on ACL*
- B. Pang, L. Lee and S. Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on EMNLP*
- Bo Pang, and Lillian Lee. 2008. "Opinion mining and sentiment analysis." *Foundations and trends in information retrieval* 2.1-2 (2008): 1-135.
- H. Saif, Y. He and H. Alani. 2012. *Semantic Sentiment Analysis of Twitter*. 11th International Semantic Web Conference, Boston, MA, USA, Proceedings, Part I, pages 508-524.
- Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. 2011. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD*.
- Titov, Ivan, and Ryan McDonald. 2008. "Modeling online reviews with multi-grain topic models." *Proceedings of the 17th international conference on World Wide Web*. ACM,.
- Titov, Ivan, and Ryan McDonald. 2008. "A Joint Model of Text and Aspect Ratings for Sentiment Summarization." *Proceedings of the 46th Annual Meeting of ACL*.
- C. Wang, and D. Blei. 2011. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD*, 448–456. ACM.
- Hongning Wang, Yue Lu and Chengxiang Zhai. 2010. Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach. *The 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'2010)*, p783-792,
- Wayne Xin Zhao, Jing Jiang, Hongfei Yan, Xiaoming Li. 2010. "Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid." *Proceedings of the 2010 Conference on EMNLP*.
- J. Zhu; A. Ahmed; and P. Xing. 2009. *Medlda: maximum margin supervised topic models for regression and classification*. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 1257–1264. ACM.