Topic Sentiment Joint Model with Word Embeddings

Xianghua Fu, Haiying Wu, Laizhong Cui

College of Computer Science and Software Engineering, Shenzhen University, Shenzhen Guangdong 518060, China

fuxh@szu.edu.cn, whywuhaiying@gmail.com, cuilz@szu.edu.cn

Abstract. Topic sentiment joint model is an extended model which aims to deal with the problem of detecting sentiments and topics simultaneously from online reviews. Most of existing topic sentiment joint modeling algorithms infer resulting distributions from the co-occurrence of words. But when the training corpus is short and small, the resulting distributions might be not very satisfying. In this paper, we propose a novel topic sentiment joint model with word embeddings (TSWE), which introduces word embeddings trained on external large corpus. Furthermore, we implement TSWE with Gibbs sampling algorithms. The experiment results on Chinese and English data sets show that TSWE achieves significant performance in the task of detecting sentiments and topics simultaneously.

1 Introduction

With the rapid development of e-commerce and social media, it is extremely urgent and valuable to automatically analyze the reviews to detect sentiments and topics simultaneously. Great effort on new methodologies for detecting topics and sentiments simultaneously has flourished in the recent years [1-5].

Several works extending probabilistic topic models[6,7] have been designed to tackle the problem of the joint extraction of sentiments and latent topics from documents in the recent years [2, 3, 8]. The joint sentiment topic model (JST) [2] extends LDA to a four-layer model by adding an additional sentiment layer between the document and the topic layers. Topic sentiment mixture (TSM) [8] jointly models topics and sentiments in the corpus built on the basis of PLSI. These approaches infer sentiment and topic distributions from the co-occurrence of words within documents. However, when the training corpus is small or when the documents are short, the sentiment and topic distributions might be not very satisfactory. Additionally, most of recent works [2, 3, 9] try to incorporate some polarity lexicons into their models as the prior knowledge. However, these approaches still have their limitations, for example if the polarity lexicons are not rich, the improvement of the prior is very limited. As a result, we have to seek for other approaches.

Most recently, word embeddings are gaining more and more attention, since they show very good performance in a broad range of natural language processing (NLP) tasks [10-12]. For example, [10] incorporates latent feature vector representations of

In: P. Cellier, T. Charnois, A. Hotho, S. Matwin, M.-F. Moens, Y. Toussaint (Eds.): Proceedings of DMNLP, Workshop at ECML/PKDD, Riva del Garda, Italy, 2016.

words to LDA model, and [11] employs latent topic models to assign topics for each word in the text corpus, and learns topical word embeddings (TWE). But these models only complete the task of mining topics. Little attention has been devoted to topic sentiment model with word embeddings so far. In this paper, we propose a new topic sentiment model which incorporates word embeddings. To the best of our knowledge, it is the first work to formulate topic sentiment model with word embeddings.

In contrast with other topic sentiment modeling frameworks, our model is distinguished from them as follows: (1) we incorporate word embeddings trained on very large corpora. It significantly improves the sentiment-topic-word mapping and extends semantic and syntactic information of words. (2) experiments are performed on four real online review data sets for two kinds of language (English and Chinese), which show that our model is used more extensive. (3) we also compare the performance on incorporating the sentiment polarity and without introducing sentiment polarity respectively to demonstrate that our new model is fully unsupervised. We find that our unsupervised model is highly portable to other domains for the sentiment classification task and achieves significant performance in the task of sentiment analysis, and extracting sentiment-specific topics.

2 Topic and Sentiment Model with Word Embeddings

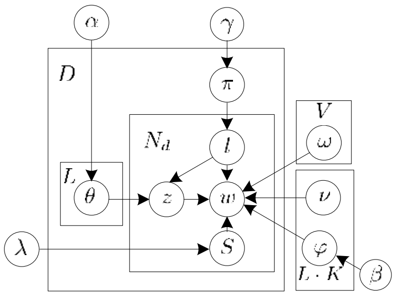


Fig. 1. Graphical representation of TSWE model

2.1 Topic and Sentiment Model with Word Embeddings

In this section, we propose a novel topic sentiment model with word embeddings called TSWE, as shown in Fig. 1. TSWE is formed by taking the original topic sentiment model JST [2, 3] and replacing their Dirichlet multinomial component with a two components mixture of a sentiment-topic-to-word Dirichlet multinomial component and a word embeddings component. Our model defines the probability that it generates a word from embeddings component as the multinomial distribution $MulT$ with:

$$MulT(w_i|\nu_k\omega^T) = \frac{\exp(\nu_k \cdot \omega_{w_i})}{\sum_{w_i' \in W} \exp(\nu_k \cdot \omega_{w_i'})} \quad (1)$$

The negative log likelihood L according to our model factorizes topic-wise into factors L_k for each topic associated with sentiment. we derive:

$$L_k = \mu \parallel \nu_k \parallel_2^2 - \sum_{w_i \in W} N^{k,w_i} \left(\nu_k \omega_{w_i} - \log \left(\sum_{w_i' \in W} \exp \left(\nu_k \omega_{w_i'} \right) \right) \right) \tag{2}$$

Then we apply L-BFGS implementation [13] from the Mallet toolkit [14] to derive the topic vector ν_k that minimizes L_k.

2.2 Generative process for the TSWE model

The formal definition of the generative process of TSWE model is as follows:
For each of sentiment-topic pair (l,z)
generate the word distribution of the sentiment-topic pair $\varphi_{l,k} \sim Dir(\beta)$
For each document d
draw a multinomial distribution $\pi_d \sim Dir(\gamma)$
For each sentiment label l under document d
draw a multinomial distribution $\theta_{d,l} \sim Dir(\alpha)$
For each word w_i in document d
-draw a sentiment label $l_i \sim Mul(\pi_d)$
-draw a topic $z_i \sim Mul(\theta_{d.l})$
-draw a binary indicator variable $s_i \sim Ber(\lambda)$
-draw a word $w_i \sim (1 - s_i)Mul(\varphi_{z_i}) + s_i MulT(\nu_{z_i}\omega^T)$

2.3 Gibbs sampling for TSWE model

In this section, we introduce the Gibbs sampling algorithm [15] for the TSWE model.The detailed derivation process on Gibbs Sampling for topic models can refer the literature [16].

The Posterior probability can be obtained from the joint probability as follows:

$$P(z_i = k, l_i = l | w, z^{\neg i}, l^{\neg i}, \alpha, \beta, \gamma, \lambda, \nu, \omega) \propto \left((1 - \lambda \cdot \frac{N_{l,k,w_i}^{\neg i} + \beta}{N_{l,k}^{\neg i} + V\beta} + \lambda \cdot MulT(w_i | \nu_k \omega^T) \right) \cdot \frac{N_{d,l,k}^{\neg i} + \alpha}{N_{d,l}^{\neg i} + T\alpha} \cdot \frac{N_{d,l}^{\neg i} + \gamma}{N_d^{\neg i} + L\gamma} \tag{3}$$

Samples derived from the Markov chain are then used to estimate π, θ and φ as depicted in equation (4), (5), (6).

$$\pi_{d,l} = \frac{N_{d,l} + \gamma}{N_d + L\gamma} \tag{4}$$

$$\theta_{d,l,k} = \frac{N_{d,l,k} + \alpha}{N_{d,l} + T\alpha} \tag{5}$$

$$\varphi_{l,k,i} = (1\text{-}\lambda) \cdot \frac{N_{l,k,i} + \beta}{N_{l,k} + V\beta} + \lambda M\mathrm{ulT}(\mathrm{w}_i | \nu_k \omega^T) \tag{6}$$

3 Experiment

In this section, we explore the performance of TSWE model on document-level sentiment classification and topic extraction evaluations on different kinds of datasets for English and Chinese.

3.1 Experimental setup

3.1.1 Training word embeddings

We train 300 dimensional word embeddings on two corpus by using the Google word2vec toolkit [17]: Chinese Wikipedia[1] and English Wikipedia[2].

3.1.2 Experimental datasets

We perform experiments on two kinds of sentiment mining datasets, Chinese and English. Chinese datasets consists of three categories of product reviews datasets[3] including book, hotel, and computer, with 1000 positive and 1000 negative examples for each domain. English corpora is the polarity dataset version 2.0[4] which is introduced by Pang and Lee in 2004, consisting of 1000 positive and 1000 negative movie reviews, which we call MR04 dataset.

Preprocessing: We remove the repetitive comments and stop words, the words that word frequencies are less than 2 or larger than 15 and the words that are not found in Google embeddings representations trained from Chinese Wikipedia corpus and English Wikipedia corpus. In addition, we perform word segment for Chinese datasets

3.2 Parameter Setting

We set the symmetric prior hyper-parameter β=0.01 in our TSWE model. The symmetric hyper-parameter γ is set to $\frac{0.05 \cdot A}{L}$, where A is the average document length and L is total number of sentiment labels, as noted by [3]. The α is set to the standard setting $\frac{50}{K}$.

3.3 Experimental Results and Analysis

In this section, we present and discuss the experimental results of both document-level sentiment classification and topic extraction.

3.3.1 Sentiment classification evaluation

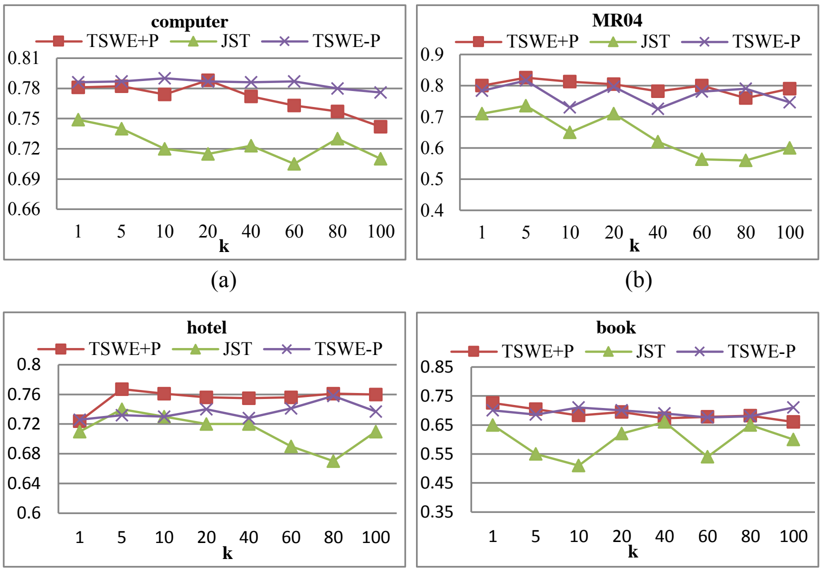
We use the common metrics to evaluate classification performance: Accuracy. Table 1 presents classification accuracy results obtained by TSWE on the computer data set with the number of topics K set to either 1 or 20. By varying λ, as shown in Table 1, the TSWE model obtains its best result at λ=0.1, where the λ is set 0.1 to 0.5 is better

[1] http://download.wikipedia.com/zhwiki/latest/zhwiki-latest-pages-articles.xml.bz2
[2] http://nlp.stanford.edu/data/WestburyLab.wikicorp.201004.txt.bz2
[3] http://www.datatang.com/data/11937
[4] www.cs.cornell.edu/people/pabo/movie-review-data/

than λ=0.0 on computer data sets. That shows the word embeddings is effective in capturing positive and negative sentiments. So we fix λ at 0.1, and report experimental results based on this value for the rest of this section.

Table 1. Accuracy on the computer and MR04 .

data	accuracy											
	λ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
computer	K=1	0.765	**0.791**	**0.786**	**0.791**	**0.781**	**0.776**	0.726	0.689	0.653	0.653	0.561
	K=20	0.781	**0.797**	**0.788**	**0.782**	**0.791**	**0.786**	**0.791**	0.772	0.745	0.602	0.552



(a) (b)

(c) (d)

Fig. 2. Accuracy with different topic number settings on the four datasets.

With lexicons vs no lexicons:

In the experiments, we compare the classification results of introducing lexicon and no lexicon, as shown in Fig.2, TSWE+P represents the accuracy of incorporating sentiment prior, TSWE-P denotes sentiment prior is not introduced. The lexicon includes two subjectivity lexicons, the English lexicon is the MPQA[5] and the Chinese lexicon is Hownet emotional word set[6]. On most tests, the classification results of incorporating lexicon are almost similar to the the classification results of no lexicon on the same topic number. That shows the word embeddings have already captured positive and negative sentiments.

TSWE vs JST with different number of topics:

[5] http://www.cs.pitt.edu/mpqa/

[6] http://www.datatang.com/datares/go.aspx?dataid=603399

Fig. 2 shows classification results produced by TSWE and the JST models on the four datasets with different numbers of topics. TSWE significantly outperforms JST in all of the datasets, particularly on the MR04 dataset where we get 20.0% improvement on accuracy at K = 80. The above results show that the word embeddings can help to extend the semantic information of words, and also can capture the sentiment information of words.

3.3.2 Topic extraction evaluation.
The other goal of evaluation task is to extract topics and evaluate the effectiveness of sentiment topic. First we need to evaluate the topic clustering performance under the corresponding sentiment polarity. We use two common metrics to evaluate the performance: perplexity and normalized mutual information (NMI) [18]. More formally, for a test set of D documents, the perplexity is:

$$perplexity = \exp\left(-\frac{\sum_{d=1}^{D} \log p(w_d)}{\sum_{d=1}^{D} N_d}\right)$$

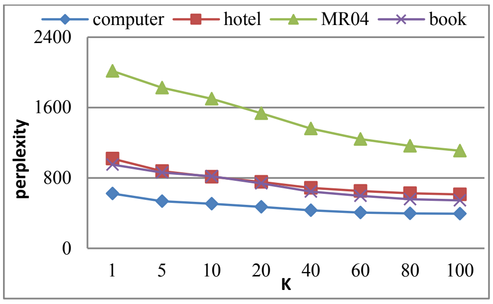


Fig. 3. Perplexity in TSWE model with different topic number settings on the four data sets

Fig. 3 shows that the perplexity on the MR04 dataset is higher than the other datasets. The reason is that the word in the corpus is more than others.

From Table 2 we can learn that the TSWE model has better NMI than JST, the NMI for TSWE model is around 0.268~0.600, and the JST obtains only around 0.10~0.420, which shows the effectiveness of the topic cluster under the sentiment with TSWE.

Table 2. NMI results in TSWE and JST on data sets book and MR04.

Data	Model	NMI							
		K=1	K=5	K=10	K=20	K=40	K=60	K=80	K=100
book	TSWE	0.392	0.353	0.309	0.338	0.293	0.302	0.315	0.268
	JST	0.260	0.083	0.070	0.195	0.270	0.062	0.24	0.168
MR04	TSWE	0.542	0.600	0.572	0.554	0.507	0.550	0.472	0.540
	JST	0.358	0.420	0.248	0.370	0.195	0.101	0.100	0.164

A topic is a multinomial distribution over words conditioned on both topics and sentiments. The most probable words for each sentiment-topic distribution could approximately reflect the meaning of the topic. Table 3 shows the selected examples of global topics extracted from computer data set with JST and TSWE. Each row shows the top 15 words for corresponding topics. We can see that some words of TSWE such as "cooling, fan, radiator, voice, temperature, workmanship, operation" are about the computer Heat-dissipation problem, and some words such as "good, quietness, perfect, like, nice, suitable" are the emotional tendencies of the computer Heat-dissipation problem. It shows that TSWE can extract topic and sentiment simultaneously. Overall, the above analysis illustrates the effectiveness of TSWE in extracting opinionated topics under sentiment from a corpus.

Table 3. extracted topic under different sentiment labels by JST and TSWE

JST	**Pos**	漂亮/nice ; 散热/cooling; 外观/appearance; 喜欢/like; 设计/design; 配置/configuration; 比较 /very; 时尚/fashion ; 硬盘/hard disk; 噪音/noise; 内存/memory; 本本 /machine; 完美 /perfect; 钢琴/piano; 键盘/keyboard
	Neg	声音/voice;风扇/fan;温度/temperature;发热量/calorific value;散热/cooling; 硬盘/hard disk; 接受/accept; 开机/starting up; 噪音/ noise; 发热/heat; 感觉/feeling; 确实/indeed; 运行 /operation; 控制/contro; 触摸/touch
TSWE	**Pos**	散热/cooling; 风扇/fan; 不错/good; 声音/voice; 安静/quietness; 温度/temperature;完美/perfect;散热器/radiator;喜欢/like;做工/workmanship; 漂亮/nice; 运行/operation; 游戏/game; 合适/suitable; 效果/effect
	Neg	散热/cooling; 风扇/fan; 声音/ voice; 温度/ temperature; 一般/general; 不好/bad;噪音/noise;散热器/ radiator;发热量/calorific value;机器/machine; 运行/operation; 发热/heat; 游戏/game; 硬盘/hard disk; 效果/effect

4 Conclusions and Future Work

In this paper, we propose a novel unsupervised generative model (TSWE) for jointly mining sentiments, sentiment-specific topics from online reviews. To the best of our knowledge, this is the first work to model topic sentiment joint model with word embeddings. Most importantly, the experiments on real review data sets for English and Chinese show that TSWE is effective in discovering sentiments and topics simultaneously. In the future work, we will explore how to properly introduce the lexicon with HowNet lexicon to improve the performance of detecting sentiments and sentiment-specific topics.

Acknowledgements.

This research is supported by the National Nature Science Foundation of China under Grants 61472258, 61402294, National Key Technology Research and Development Program of the Ministry of Science and Technology of China (2014BAH28F05), Science and Technology Foundation of Shenzhen City under Grants JCYJ20140509172609162 and JCYJ20130329102032059.

5 References

1. Dermouche, M., Kouas, L., Velcin, J., Loudcher, S.: A Joint Model for Topic-Senti ment Modeling from Text. In: ACM/SIGAPP Symposium On Applied Computing (SAC). (2015)
2. Lin, C., He, Y.: Joint sentiment/topic model for sentiment analysis. In: Proceedings of the 18th ACM conference on Information and knowledge management, pp. 375-384. ACM, (2009)
3. Lin, C., He, Y., Everson, R., Rüger, S.: Weakly supervised joint sentiment-topic detec tion from text. Knowledge and Data Engineering, IEEE Transactions on 24, 1134-1145 (2012)
4. Pavitra, R., Kalaivaani, P.: Weakly supervised sentiment analysis using joint sentiment topic detection with bigrams. In: Electronics and Communication Systems (ICECS), 2015 2nd International Conference on, pp. 889-893. IEEE, (2015)
5. Brody, S., Elhadad, N.: 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, pp. 804-812. Association for Computational Linguistics, (2010)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. the Journal of machine Learning research 3, 993-1022 (2003)
7. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd an nual international ACM SIGIR conference on Research and development in information retrieval, pp. 50-57. ACM, (1999)
8. Mei, Q., Ling, X., Wondra, M., Su, H., Zhai, C.: Topic sentiment mixture: modeling facets and opinions in weblogs. In: Proceedings of the 16th international conference on World Wide Web, pp. 171-180. ACM, (2007)
9. Chen, Z., Li, C., Sun, J.-T., Zhang, J., Li, C., Zhang, J., Sun, J.-T., Chen, Z.: Sentiment Topic Model with Decomposed Prior. In: SDM, pp. 767-775. (2013)
10. Nguyen, D.Q., Billingsley, R., Du, L., Johnson, M.: Improving Topic Models with La tent Feature Word Representations. Transactions of the Association for Computational Linguistics 3, 299-313 (2015)
11. Liu, Y., Liu, Z., Chua, T.-S., Sun, M.: Topical Word Embeddings. In: AAAI, pp. 2418-2424. (2015)
12. Das, R., Zaheer, M., Dyer, C.: Gaussian LDA for topic models with word embeddings. In: Proceedings of the 53nd Annual Meeting of the Association for Computational Linguistics. (2015)
13. Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large scale optimiza tion. Mathematical programming 45, 503-528 (1989)
14. McCallum, A.K.: {MALLET: A Machine Learning for Language Toolkit}. (2002)
15. Walsh, B.: Markov chain monte carlo and gibbs sampling. (2004)
16. Heinrich, G.: Parameter estimation for text analysis. Technical report (2005)
17. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representa tions of words and phrases and their compositionality. In: Advances in neural information processing systems, pp. 3111-3119. (2013)
18. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval. Cambridge university press Cambridge (2008)