ATLANTIS PRESS
Advances in Engineering Research, volume 118
2nd International Conference on Automation, Mechanical Control and Computational Engineering (AMCCE 2017)

A Hybrid HDP-ME-LDA Model for Sentiment Analysis

Bo Yuan[1, a], Gang Wu[2, b]

[1] School of Software, Shanghai Jiao Tong University, Shanghai, 200240, China

[2] School of Software, Shanghai Jiao Tong University, Shanghai, 200240, China

[a]yblelouch@gmail.com, [b] dr.wugang@sjtu.edu.cn

Keywords: Aspect Detection, Sentiment Analysis, LDA, Maximum Entropy, Hierarchical Dirichlet Process

Abstract. Sentiment Analysis is an important research area in data mining. Recently, various topic models for aspect detection and sentiment analysis have been proposed, many of which are based on Latent Dirichlet Allocation(LDA) unsupervised machine learning approach. LDA requires the number of topics in advance which are often based on unreliable experience for different areas. On the other hand, it is important to identify aspect and opinion words of topics, especially aspect-specific opinion words and analyze sentiment polarity. But few research did them all. To solve these problems, this paper proposes a hybrid Hierarchical Dirichlet Process and Maximum Entropy-Latent Dirichlet Allocation(HDP-ME-LDA) model. It uses HDP to automatically determine the number of topics and utilizes maximum entropy classifier to separate aspect and opinion words, including aspect-specific opinion words. We evaluate our model on data sets of reviews of restaurant and electronic devices qualitatively and quantitatively. The result shows that we perform better than other topic models, like JST, ASUM, MaxEnt-LDA, HDP-LDA.

Introduction

There is more and more information on the Internet with the development of the network. Many of them are user reviews. For a customer, it is impossible to read all the complicated reviews. He or she wants to learn key information from the reviews. Sentiment analysis has great significance for providing technical support for this purpose. Aspect detection and sentiment analysis are the main research of opinion mining in recent years. The former is extracting the described objects of different granularities from the text, and the latter is finding the writer's opinions. Many of today's models focus on both.

Even though sentiment analysis and aspect detection has been studied for a long time[1], there are still some problems. First, most of the current models are based on LDA, which requires the number of topics in advance. It's an important parameter for LDA. But we can't know the exact number for different areas and data sets with our experience. Second, many models mix aspect and opinion words together. And few models have considered finding the aspect-specific opinion. For example, "exciting" provides more specific information compared with "good" for "movie" aspect. On the other hand, the same word may have different meanings, just like "high" for "price" and "quality". Third, many studies use the sentiment dictionary. But dictionaries are usually laggard and not professional. Fourth, LDA model doesn't have context information.

This paper combines HDP-LDA[2] with MaxEnt-LDA[3] and improves MaxEnt-LDA, finally proposes HDP-ME-LDA model to solve these problems. Firstly, we divide sentences or reviews into "clause". Then we use MaxEnt classifier, which has two variable u and v for indicating the word is local or general and the word is aspect or opinion. And we improve MaxEnt by adding a sentiment layer, which makes it can get sentiment polarity. The main step is that we use HDP to generate the topic distribution, and then use adapted MaxEnt-LDA which is based on the HDP's result to get the sentiment distribution.

Related Work

Researchers usually use syntax-based[4], supervised and unsupervised machine learning to do aspect detection and sentiment analysis. And these methods can be combined to perform better.

Joint Sentiment/Topic Model (JST) [5] adds a sentiment layer in LDA for getting sentiment polarity. Aspect and Sentiment Unification Model (ASUM) [6] considers about the context information by assuming that a sentence has the same aspect and the same sentiment. It increases efficiency, but it is not practical. JST and ASUM doesn't separate aspect and opinion words. [3] enriches LDA by adding a Maximum Entropy component. MaxEnt-LDA can separate aspect and opinion words, even separate general and local opinion words. But it has the same assume for sentence with ASUM. The other disadvantage is that it does not classify sentiment polarity. To automatically generate the number of topics, HDP-LDA is purposed in [2]. HDP is a nonparametric Bayesian model which replaces Dirichlet allocation with Dirichlet processes. It can automatically generate the number of topics. We interpret DP as Chinese Restaurant Process (CRP). In CRP, one document is deemed as a restaurant, and each word in the document is deemed as a customer, who chooses a table to sit. A table represents an aspect. A dish represents a topic. CRP assumes that one table has the same dish. Each customer first chooses a table. The probability of choosing an existing table is $c/(n-1+\alpha)$, and the probability of choosing a new one is $\alpha/(n-1+\alpha)$, α is the para meter to control the generation of new tables, c is the number of customers who have already been sitting at the table. If choosing a new table, this table will choose a dish. The probability is like choosing a table. These two DPs are Hierarchical DP.

The Hybrid HDP-ME-LDA Model

In this section, we introduce the hybrid HDP-ME-LDA model. Before the generation process, we need to introduce two basic methods in our model. One is construction of clauses, the other is adapted MaxEnt-LDA.

Different from other LDA-based probability models, we use a "clause" as the basic unit of the model, rather than a single word or a natural sentence. Using a single word will lose context information. But using a sentence need to assume that a sentence has the same aspect and the same sentiment. It is unrealistic. For example, the sentence "The soup tastes not good, but the waiters are friendly" has two aspects and opposite sentiment. So we split a sentence into some clauses, and assume that a clause has the same aspect and the same sentiment. We do POS tagging at first, and then split according to these regulars: (1) meeting punctuation marks (2) meeting conjunction (3) if there is no verb in the clause, then the original sentence uses the same verb-object structure, e.g. "the food and drinks are awful" will be "the food are awful" and "the drinks are awful".

We use MaxEnt classifier to separate aspect and opinion words (we call it word subjectivity). We know that completely unsupervised approaches can't get good classification of word subjectivity from [7] and [5]. So we use MaxEnt, a good supervised machine learning approach to be our classifier. We use two features to represent the feature vector $\mathbf{x}$, which are the lexical features and POS tag features of the previous, the current and the next words($\{w_{i-1}, w_i, w_{i+1}\}$ and $\{POS_{i-1}, POS_i, POS_{i+1}\}$). We training λ as the weight of $\mathbf{x}$ to be the input of model. Eq. 1 shows the probability of word subjectivity v of i^{th} word in review d, clause m. R is 2, which means word subjectivity type is 2. And the variable u indicates the word is local or general which is generated by binomial distribution ρ.

$$P(\mathrm{V}_{d,m,i} = b) = \frac{\exp(\lambda_b \cdot \mathbf{X}_{W_{d,m,i}})}{\sum_r^R \exp(\lambda_r \cdot \mathbf{X}_{W_{d,m,i}})} \quad (1)$$

And then we improve MaxEnt-LDA to make it have the ability to analyze sentiment polarity. The way is to add a sentiment layer between topic layer and word layer and to make it a 4-layer model from 3-layer model. The adapted MaxEnt-LDA can get sentiment polarity at the time of getting sentiment distribution.

We use CRP explanation to introduce HDP[8]. Assume that there are D reviews, and we can find M clauses from them. The i^{th} clause in the j^{th} review $C_{j,i}$ has an aspect word $A_{j,i}$ according to MaxEnt classifier. $A_{j,I}$ is generated by a table distribution G_j in CRP. Whether $A_{j,i}$ is a new table or an existing table is up to G_j[2]. G_j draws from a global distribution G_0, a CRP dish distribution. The dish distribution is the topic distribution. The number of dish distributions is the number of topics. We add context information to HDP by adding a coefficient ε to balance the number of words in each table or dish and context information. It is shown in Eq. 2. Q_t is the number of words assigned to the table t, $N_{j,i,t}$ is the number of times the word in $C_{j,i}$ is assigned to the table t.

$$P(a_{j,i} = t \mid a_{-(j,i),w,\alpha,\delta}) \propto \begin{cases} (1-\delta) * Q_t + \delta * N_{j,i,t} & (\text{if t existed}) \\ \alpha & (\text{if t is new}) \end{cases} \tag{2}$$

Now we can combine HDP and adapted MaxEnt-LDA to get HDP-ME-LDA. HDP-ME-LDA uses clauses as inputs, and its generation process is as following. Firstly, the model uses HDP to generate topic distribution and detect global and local aspect word with the MaxEnt classifier. This step is described above. We can get topic distribution θ, general aspect word distribution Φ^A, local aspect word distribution Φ^a, and the number of topics T. Secondly, the model use MaxEnt-LDA to generate sentiment distribution. For every topic, it generates sentiment distribution. Then every clause gets its sentiment by multinomial distribution which is generated by prior parameter β. According to u and v of classifier, the word will get his type. So the model has the sentiment distribution π, general opinion word distribution Φ^B, local opinion word (aspect-specific word) distribution Φ^b. At last, we calculate the sentiment polarity in the order of word, clause, reviews. Fig. 1 shows the complete process. In addition to the notations that have already been introduced, α and γ are the parameter to control the generation of table or dish in DP, δ and ε are the parameter to balance the context information in DP, ζ and ω are the prior Dirichlet parameter, η are the prior Beta parameter.

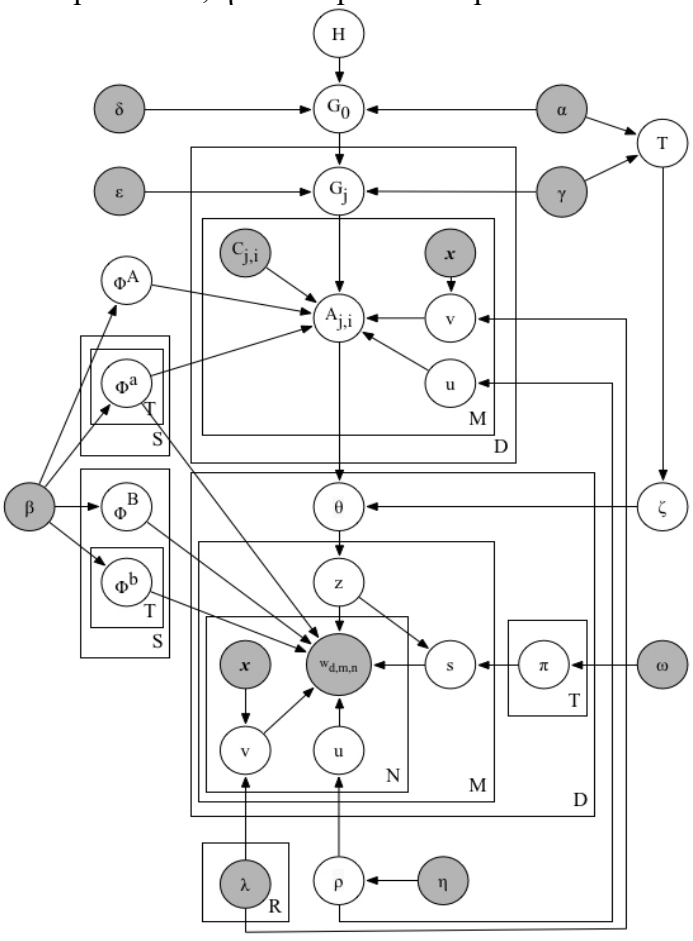

Fig. 1 Graphical Representation of HDP-ME-LDA

Experiment and Evaluation

This section introduces the data sets and qualitative and quantitative experimental results. It shows our model performs better than other baseline models.

Considering that there should be strong distinctions between the selected datasets, this paper decided to select the Amazon electronic product data set and the Citysearch New York restaurant data set.

Table 1 shows the qualitative results of HDP-ME-LDA model on restaurant data set. Result shows these 3 points. First, the model can separate aspect and opinion words exactly. Second, the consistency of each topic is high, that is, the correlation degree between the words within each topic is high. Third, the model can separate general and local words. Local opinion words are aspect-specific words. They provide specific information about the topic.

Table 1 Sample Result of Citysearch New York Data Set for HDP-ME-LDA

Topic 0	**Positive**	**Aspect**	dinner	beef	coffee	wine	salad
		Opinion	delicious	fresh	cheap	amazing	tasty
	Negative	**Aspect**	cheese	rice	meat	bread	coffee
		Opinion	oily	cooked	fat	hot	mediocre
Topic 1	**Positive**	**Aspect**	dinner	waiter	manager	bowl	staff
		Opinion	funny	friendly	helpful	pro	impressive
	Negative	**Aspect**	bartender	pork	staff	waitress	place
		Opinion	rude	evil	slow	rushing	simple
Topic 2	**Positive**	**Aspect**	bowl	space	music	bar	area
		Opinion	big	rich	romantic	clean	neat
	Negative	**Aspect**	area	light	wine	cheese	rest
		Opinion	small	tiny	noise	dark	expensive
General Aspect			food	service	restaurant	menu	America
			New York	NYC	Chinese	France	experience
General Opinion	**Positive**		good	best	fine	right	nice
			enjoy	well	excellent	love	ok
	Negative		bad	suck	wtf	worst	problem
			wrong	never	hate	horrible	complain

We do quantitative experiments by calculating the accuracy of sentiment classification. Because HDP-ME-LDA and HDP generate the number of topics as 12 on Amazon data set, we set the number of topics as 7, 12, 17, 22 and 27 to evaluate the impact of topic numbers.

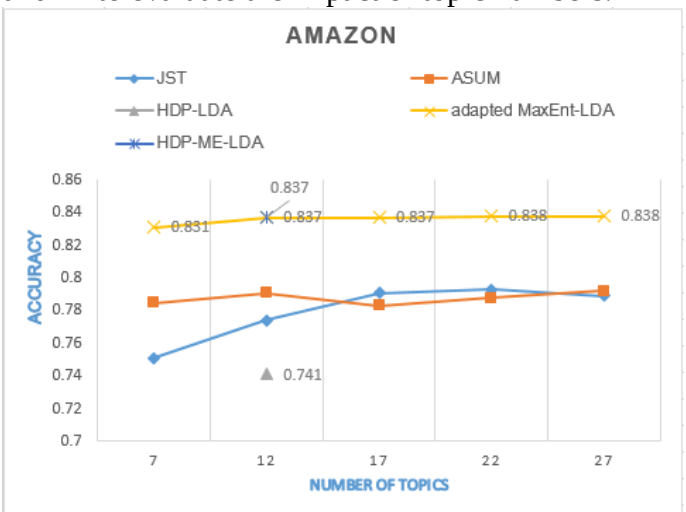

Fig. 2 Experiment Result of Accuracy of Sentiment Classification on Amazon Data Set

Fig. 2 shows the result. We can find 3 points. First, basically, with the increase in the number of topics, the accuracy of sentiment classification has increased. But the promotion is not large, the overall trend is stable. Second, we use the general sentiment dictionary for HDP, the accuracy of which is the worst. It shows the importance of professional and complete dictionary. So HDP-ME-LDA doesn't use dictionary but use clustering method to get sentiment polarity. Third, HDP-ME-LDA has the best performance in all models and the number of topics is appropriate.

Conclusion

This paper proposed an aspect detection and sentiment analysis model, HDP-ME-LDA. The contributions are:

(1) This model can automatically determine the number of topics.

(2) It can separate aspect and opinion words and find not only local and general aspect word but also local and general opinion word.

(3) It doesn't use sentiment dictionary to do sentiment analysis.

(4) It adds context information to HDP and MaxEnt classifier, and uses clause to help context reflect better.

The future work is that we may merge the two steps into one step, and let the sentiment and aspect distribution are generated together.

References

[1] Schouten K, Frasincar F. Survey on aspect-level sentiment analysis[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(3): 813-830.

[2] Ding W, Song X, Guo L, et al. A novel hybrid HDP-LDA model for sentiment analysis[C]//Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 01. IEEE Computer Society, 2013: 329-336.

[3] Zhao W X, Jiang J, Yan H, et al. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010: 56-65.

[4] Zhuang L, Jing F, Zhu X Y. Movie review mining and summarization[C]//Proceedings of the 15th ACM international conference on Information and knowledge management. ACM, 2006: 43-50.

[5] Lin C, He Y. Joint sentiment/topic model for sentiment analysis[C]//Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009: 375-384.

[6] Jo Y, Oh A H. Aspect and sentiment unification model for online review analysis[C]//Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011: 815-824.

[7] Mei Q, Ling X, Wondra M, et al. Topic sentiment mixture: modeling facets and opinions in weblogs[C]//Proceedings of the 16th international conference on World Wide Web. ACM, 2007: 171-180.

[8] Blei D M, Griffiths T L, Jordan M I. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies[J]. Journal of the ACM (JACM), 2010, 57(2): 7.