Bayseian Nonparametric Topic Model for Detecting Aspects and Sentiments from Online Reviews

Wenzhe Li
University of Washington
liwenzhe@uw.edu

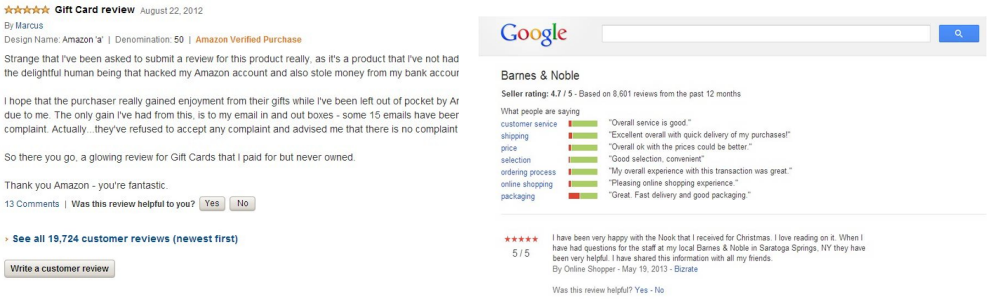
Abstract

With the increase in popularity of e-commerce, more and more product reviews become available on the internet. e.g product reviews from major sites like Amazon.com, eBay and Google Products, and discussion forums, blogs and etc. These reviews are very useful for customers to make buying decisions, and also helpful for business to improve their products based on the feedbacks. However, with the large amount of reviews exists for each product today, it's usually impossible to go through each of them, which is obviously time-consuming. In order to automate the process, we need to design a model to capture the main aspects of each review and further do sentiment classification, which in general are two-step process. In this paper, we only focus on the problem of automatically extracting aspects from product reviews. We apply nonparametric bayesian model, namely hierarchical dirichlet process(HDP), to aspect mining from reviews. As far as we know, none of the existing work consider nonparametric model. One big benefit of using nonparametric model is that it will automatically adapt as we have more and more data. We experimented HDP-LDA on yelp review data set and compared the results with classical latent dirichlet allocation model.

1 Introduction

The world wide web provides large amount of reviews of products, ranging from books, restaurants, electronic devices to many others. In these reviews, people usually give positive or negative opinions for specific aspects. For example, people who rate restaurant may provide some opinions for food, service, price, location and etc, which we call *aspects*. These online reviews are helpful for customers to make best choice, and also helpful for business to improve their products. However, the problem what is arising is that we have too many reviews to go through. The Figure 1(a) shows the customer reviews for gift card product from Amazon.com. Obviously, it's almost impossible for one to go through tens of thousand reviews, where we need a automated way for analyzing them.

In order to automatically extract useful information from the reviews, one big problem we need to solve is to extract aspects from the reviews. Different customers may talk about the different aspects for the same product. A user who is looking for computer may care about its computational power and maximum throughput, while other cares more about graphical capability. Another big problem is understanding how the sentiments for different aspects are expressed. The sentence, "The computer is great for designers" shows the positive sentiment for the graphical capability, while the sentence, "It takes forever to get the results" shows the negative sentiment for the computational power. One simple solution is to find all the positive and negative words from the sentence, and classify them based on these words. Till now, there have been lots of work done for sentiment classification using supervised learning[11,12]. But this approach typically requires a lot of labeled data, which is quite expensive in practice.



(a) (b)

Figure 1: Online reviews. (a) shows around twenty thousand reviews for Amazon gift card. (b) is categorized reviews from google. Google uses automated process for classifying the sentiment for each aspect. However, all the aspects are predefined for each type of product[].

As we see, aspect mining and sentiment analysis is a separate task. In general, we can use two-step approach to solve the problem. Firstly, we can build a model to automatically extract the aspects from the reviews. Then based on each aspect, we do sentiment classification, which has been studied a lot. Recently, there are new work came out trying to solve these two problems jointly, by using topic models. The idea is to extend the classical LDA model by adding sentiment layer.

In this paper, however, we only focus on extracting aspects from the reviews. We built bayesian nonparametric topic model for aspect mining, without explicitly defining the number of aspects. To our best knowledge, none of the previous work consider the number of aspects are unbounded. Even though it adds more complexity for the inference, it is helpful for making the model more flexible and let it automatically adapt as our data size grows. We ran the model using yelp data set, and compared the results between classical topic model and nonparametric topic model(HDP-LDA).

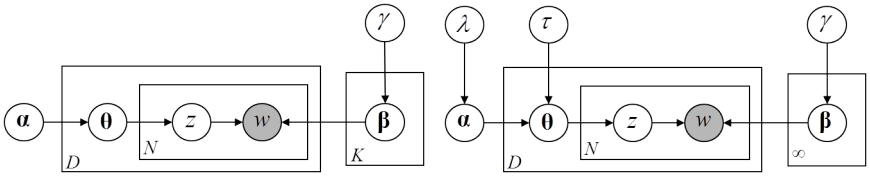
2 Previous Work

There have been quite a lot of efforts put into sentiment classification and aspects mining. For sentiment classification, Turney and Littman[13] uses an unsupervised learning algorithm to classify the semantic orientation in the word/phrase level based on mutual information, and Choi et al[14] uses conditional random fields and a variation of Autoslog to classify the sentiments. Figure 1-(b) shows the google product reviews. Where large amount of reviews are automatically classified by different aspects[15]. However, these aspects are predefined for different product types.

Comparing to the sentiment classification task, extracting aspects from the reviews seems more challenging. The earliest attempts for detecting aspects are based on frequently occurring noun phrases [9]. This approach works well when aspect are strongly tied to the single word, but less useful when aspects uses many low frequency terms. One common solution is to use clustering techniques to group the terms that associated with the same aspects. After that, they search for opinions associated with those aspects.

With the popularity of topic model(LDA)[10], there are many variations of LDA model have been applied to the product review mining. Titov et al.[7] propose a model for modeling two types of topics in reviews: global topic and local topics. The global topics correspond to a global property of review such as brand, and local topics corresponding to the product aspects. This is because aspects are fundamentally different from the global property of products. Brody and Elhadad [8] proposed a local LDA model, where they extract the aspects from sentence instead of the whole review text. Wang et al [1] uses bootstrap methods to extract the aspects first, and then use latent models to analyze the opinions. All these methods focus on extracting aspects, or dealing with the sentiment analysis as a separate step.

Some of the recent work try to analyze the aspect and opinion jointly. The work from Lin and He[6], Jo and Oh[2], Zhao et al[3] are the good examples of this type of work. They extend the LDA model



(a) LDA (b) HDP-LDA

Figure 2: (a) LDA model, with the fixed number of topic K. (b) HDP-LDA model, allowing infinite number of topics

to simultaneously capture the aspects and options. By using those models, they assume that each sentence come from one aspect.

3 Latent Dirichlet Allocation

LDA is an unsupervised topic model[10] that automatically learn the document-topic, topic-words distributions. It is a generative probabilistic model for collection of documents. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over the underlying set of topics. Each topic, in turn, is modeled as an infinite mixture over underlying set of topic probabilities. The generative process is:

1. Draw topic-word distribution β_k from Dir(γ), for $k = 1, ...K$
2. For each document d, for $d = 1, ...D$
   (a) document distribution θ_d from Dir(α)
   (b) for each word i in document d
      i. Draw z_i from multinomial(θ_d)
      ii. Draw word w_i from β_{z_i}

The model is shown in Figure 2(a). LDA has two parameters, α and γ, which prevent from overfitting. Unfortunately, exact inference in such model is intractable, and we usually turn to approximated solutions. Markov chain Monte Carlo is procedure for obtaining samples from complicated probability distributions, allowing markov chain to converge to the target distribution and then drawing samples from the markov chain [16]. Gibbs sampling, where the next state is reached by sequentially sampling all variables from their distribution when conditioned on the current values of their variables and data.

For this LDA model, we can derive the collapsed gibbs sampling, which makes the sampling process straightforward. The final gibbs sampling algorithm is:

$$P(z_i = j|z_{-i}, w) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n^{(d_i)} + \alpha}{n_{-i}^{d_i} + K\alpha} \tag{1}$$

The notations are shown as below:

- $n_{-i,j}^{(w_i)}$: number of instance of word i, assigned to the topic j, not including the current one.
- $n_{-i,j}^{(\cdot)}$: total number of words assigned to topic j, not including the current one..
- W: total number of words in vocabulary.
- β: topic-word distribution
- $n_{-i}^{d_i}$: number of words from document d_i, not including the current one.
- K: total number of topics.

For more detailed implementation of LDA, please refer to [17].

4 Hierarchical Dirichlet Process

One disadvantage of using LDA model for aspect mining is that we have to explicitly define the number of topics beforehand, which is usually unknown, but increase with the size of corpus. This is where the HDP enters the scene. Since HDP is nonparametric model, we don't need to explicitly fix the number of topics, but let the model select the correct number of topics. The powerfulness comes with computational complexity, which is much more expensive than the LDA.

The Figure 2(b) shows the HDP-LDA model, where we use infinite-dimensional multinomial α to represent the root DP, as the prior over document multinomials θ_d that represent the document DP samples, which can be used to explicitly sample topic indicator variables $z_{d,i}$. We can represent the α as stick breaking prior, where we break off a stick infinitely many times with break points distributed $Beta(1, \lambda)$.

The generative process of HDP-LDA are:

1. Draw base measure θ_k from Dir(γu).
2. Draw weights α from GEM(λ).
3. For each document d,
   (a) Draw θ_d from DP(τ, α).
   (b) For each word i in document d
      i. Draw z_i from multinomial(θ_d)
      ii. Draw word w_i from β_{z_i}

One advantage of using stick breaking process-based approach is that the infinite extension of the couple mixture keeps the structure of the finite case for Gibbs update for z_i, that is:

$$p(z_i = j|.) \propto (n_{-i,j}^{w_d} + \tau\alpha_k)\frac{n_{-i,j}^{(w_d)} + \beta}{n_{-i,j}^{(.)} + W\beta} \tag{2}$$

with α as a precision parameter. note that the sampling space has $K + 1$ dimensions, with the last point mass $\tau\alpha_{k+1}$. Whenever this mass is sampled, a new topic is created.

The prior proportions can be sampled by simulating how new components are created for $n_{d,j}$ draws from the DP under the Chinese restaurant process, which is a sequence of Bernoulli trials for each document d, and word j:

$$p(m_{d,j,r} = 1) = \frac{\tau\alpha_k}{\tau\alpha_k + r - 1} \forall r \in [1, n_{d,j}], m \in [1, D], j \in [1, K] \tag{3}$$

For more detailed implementation of HDP-LDA, please refer to [16].

5 Experiments

In this section, we will give some preliminary experimental results based on real world data sets. We experimented both for LDA and HDP-LDA model. We also give the comparison result in terms of perplexity curve.

5.1 Data Set

We chose to use yelp academic data set, which is freely available online [18]. The data set contains user reviews for the business. The data set is larger comparing to other data sets are used in previous work.

Type	Size
business	11,537
checkin set	8282
users	43873
reviews	229907

But unfortunately, we ended up using only small subset of them for our experiments (1000 of random reviews extracted from the total reviews). The primary reason is computational complexity. Since we implemented gibbs sampling for both LDA and HDP, it turns out pretty slow, and not be able to handle large amount of reviews using single machine, with limited memory size. Recently, there are new work are explicitly dealing with large data size, where they derive the stochastic variational inference [19].

5.2 Data Preprocessing

In LDA and HDP, we treat the bag of words model, without taking the orders into account. And as we see from the equation (1), the sampling process is basically calculate the number of words conditioned on different cases. It means that if the word has high frequency rate, it becomes more important for estimation process. However, our task is to extract aspects from the reviews, we are particulary interested in those words that specified different aspects. We don't need to care about other frequently appeared words like "is", "a", "the",..etc.

For data preprocessing, we simply remove the stop words and invalid words. For stop words, we mean those are appeared frequently, but does not contain any information about the aspects. And for invalid words, we defining them as containing non-alphabetical characters within the word.

5.3 Perplexity Score

In order to measure the goodness of the model, we need some measurement tool. Perplexity score is widely used in language modeling to assess the predictive power of the model. Since the documents in the corpora are treated as unlabeled, we will do density estimation, and we hope to get the high likelihood for the test data set. In particular, we use the perplexity score, which is monotonically decreasing in the likelihood of the test data, and is equivalent to the inverse of the geometric mean per-word likelihood [10]. More formally, for the test documents set,

$$\text{perplexity}(D_{test}) = exp\{-\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d}\} \tag{4}$$

In our experiment, we randomly select the 90% of the data as training samples, while the remaining 10% for testing samples.

5.4 Top word HDP-LDA

In our HDP-LDA experiments, we model automatically select the 121 topics from corpus of reviews. The table below shows the top words from the first 10 topics. It shows some meaningful aspects, but still contains some results which does not have meaningful information. For example, like "pizza." shouldn't been appeared here. This is due to the insufficient preprocessing for our data set. And another problem we can see is that the last aspect contains some sentiment-related words like "good", "happy", which does not contain any information about aspects. This actually naturally raise another problem of handling sentiment words. For simple solution, we can just remove those words from our vocabulary set. And another intuitive solution is to model our aspects and sentiment jointly, as shown in the future work section.

1	2	3	4	5	6	7	8	9	10
Chinese	pizza	Hotel	shoes	Mexican	coffee	Great	Donuts	Truth	Great
Japanese	Pizza.	Room	clothes	Salsa	movie	Place	Donut	phoenix	Food
noodles	Crust	Spa	sales	Carne	art	Good	Frozen	mountain	Good
Asian	Pizza,	Stay	shoe	Tacos	Coffee	Love	Cowboy	water	Place
korean	pizzas	pool	espresso	taco	video	well	yogurt	trail	happy

5.5 Comparison between LDA and HDP-LDA

We compared the LDA and HDP-LDA models based on perplexity curve. As we can see in Figure 4, the red line denotes the HDP-LDA model, while the blue line denotes the classical LDA model. From the blue line, as we increase the number of topics, the perplexity score drops. It drops rapidly at first, and becomes slower after that. And after at some point, the change is very small. From the figure, we can easily figure out that the optimal number of topics might be between 100-120, and in HDP-LDA model finally selects the 121 topics after running a long time.

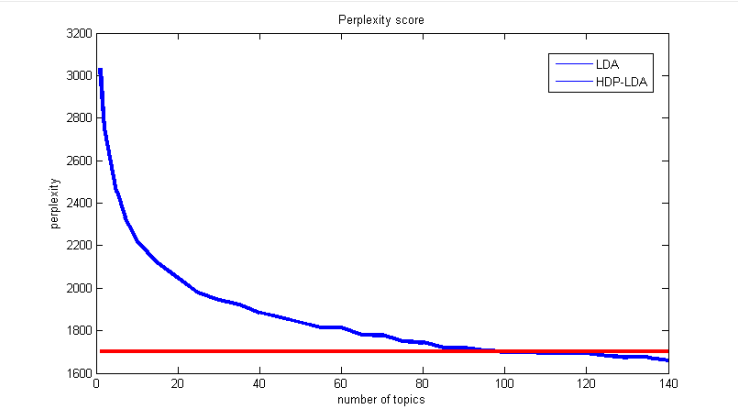


Figure 3: Comparison between LDA and HDP-LDA, in terms of perplexity score. The red line denotes the perplexity curve for the HDP-LDA model, while blue line denotes the perplexity curve for the LDA model.

6 Discussion

Our implementation and experiments have some limitations. As we can see from the result, there are some of the meaningless words are appeared as top words. It means that we need more comprehensive way for doing text preprocessing. For example, like ”pizza”,”pizzas” are equivalent, which need to be combined together. And also, as suggested in [], we didn't remove the rare words, which appears only several times in the whole corpus.

Another problem needs to figure out is that our result have relatively high perplexity score, which is between 2000 3000. We may either need to tune the parameters to get better performance score or clean the data sets to remove unnecessary information. Also, in this experiments, some reviews might from different business type, which again need filtering for the data sets.

The last problem is computational cost, as we increase the number of topics, the running time becomes really slow. In general, it takes few minutes to get out the result for LDA, and even longer for HDP-LDA. The batch processing nature of our gibbs sampling is the problem which cause such slow convergence. We need better algorithm, which can be scalable for larger data size. Because of this computational issues, we only choose a small subset for our experiments.

7 Future Work

Opinion mining and sentiment analysis are exiting research area, and recently caught many researchers' attention. But so far, the developed model is either lack of accuracy or does not scale to the read world large data set. Besides those, these techniques are open to more applications in the future. By summarize:

- Develop scalable inference algorithms for review mining/opinion mining. Most of the research done so far uses simple gibbs sampling, which is computationally infeasible for real world problems. Stochastic variational inference might be the way to try. But there are also ways to reduce the time complexity, by designing parallel algorithms.

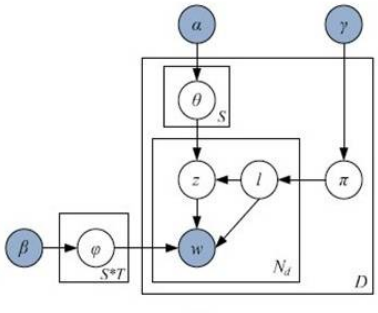


Figure 4: Comparison between LDA and HDP-LDA, in terms of perplexity score. The red line denotes the perplexity curve for the HDP-LDA model, while blue line denotes the perplexity curve for the LDA model.

- So far, all the previous work designs the topic model with fixed number of topics. This has some potential problem that if we have more and available in the future, the model is unlikely to perform well as before, since we might need more complex model to accommodate arising complexity.
- Most of current research focus on two-step process, as we've shown before. However, we can extend the topic model to solve both problems simultaneously, by incorporating sentiment information. In fact, there are already few works in this domain. One typical example is to use four layer topic model, instead of three, to incorporate the sentiment information, which is proposed by Lin[]. As shown in Figure **??**, they added another layer for sentiment. And for each sentiment type, they have topic-word distributions, in which case, we have three dimensional sentiment-topic-word distributions, instead of two in classical LDA case. However, they still fix the size of topics and number of sentiments, which can be changed based on the data set we have. Thus, one potential work may relax this condition to allow infinite number of topics and sentiment labels by using nonparametric topic model, like extension of HDP.
- Beside review mining, HDP-LDA potentially is beneficial for online research. It can be used to automatically categorize the documents returned by the search engine, which will aid the user browsing. But designing the scaling algorithms is the prerequisite for accomplishing this goal.
- Applying the topic model for asian language is another interesting direction we can approach. Like Chinese language does not have space between words, which potentially impose more difficulty than English text.

8 Conclusion

In this paper, we do preliminary experiments LDA and HDP-LDA model for aspect mining, and showed the superior of the nonparametric model, while the computational cost becomes the downside. We implemented using gibbs sampling for both of these topic models. As we see from the experimental results, the nonparametric model can automatically select the optimal number of topics based on the data set. However, in order to scale our model to huge collection of data sets, we need to design computationally inexpensive inference algorithm.

References

[1] Wang.H, Lu. Y and Zhai.C. Latent aspect rating analysis on review text data: a rating regression approach. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. pp.783–792 Washington, DC, USA 2010

[2] Jo.Y and Oh.A.H. Aspect and sentiment unification model for online review analysis. *Proceedings of the fourth ACM international conference on Web search and data mining*. pp. 815–824, Hong Kong, China 2011.

[3] Zhao.W, Jiang.J, Yan.H.F and Li.X. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. pp. 56–65. Cambridge, Massachusetts, 2010.

[4] Li.F, Huang.M and Zhu.X. Sentiment Analysis with Global Topics and Local Dependency *AAAI*. 2010.

[5] He.Y, Lin.C and Alani.H. Automatically extracting polarity-bearing topics for cross-domain sentiment classification. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1* pp.123–131. Portland, Oregon, 2011

[6] Lin.C and He.Y. Joint sentiment/topic model for sentiment analysis. *Proceedings of the 18th ACM conference on Information and knowledge management*. pp.375–384, Hong Kong, China 2009

[7] Titov.I and McDonald.R. Modeling online reviews with multi-grain topic models. *Proceedings of the 17th international conference on World Wide Web* pp. 111–120 Beijing, China 2008

[8] Brody.S and Elhadad.N. An unsupervised aspect-sentiment model for online reviews *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. pp.804–812. Los Angeles, California. 2010.

[9] Hu.M and Liu.B Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 168–177, Seattle. WA 2004

[10] David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research* pp. 993-1022 2003

[11] Pang.B, Lee.L and Vaithyanathan.S Thumbs up?: sentiment classification using machine learning techniques *Proceeding EMNLP '02 Proceedings of the ACL-02 conference on Empirical methods in natural language processing* - Volume 10 Pages 79-86

[12] Kim.S.M and Hovy.E. Determining the sentiment of opinions. *Proceeding COLING '04 Proceedings of the 20th international conference on Computational Linguistics*. Article No. 1367 .

[13] P.D. Turney and M.L.Littman. Unsupervised learning for semantic orientation from a hundred-billion-word corupus. *CoRR*, cs.LG/0212012, 2012

[14] Y.Choi, C.Cardie, E.riloff, and S.Patwardhan. Identifying sources of opinions with conditional random fields and extraction patterns. *In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing* pages 355-362, Vancouver, British Columbia, Canada, October 2005.

[15] Sasha Blair-goldensohn and Tyler Neylon and Kerry Hannan and George A. Reis and Ryan Mcdonald and Jeff Reynar Building a sentiment summarizer for local service reviews. *In NLP in the Information Explosion Era*. 2008

[16] Gregor Heinrich. Infinite LDA CImplementing the HDP with minimum code complexity. *Technical note* TN2011/1, 2011

[17] Tom Griffiths. Gibbs sampling in the generative model of Latent Dirichlet Allocation. *Note*. 2007.

[18] http://www.yelp.com/dataset_challenge/

[19] M. Hoffman, D. Blei, J. Paisley, and C. Wang. Stochastic variational inference. *Journal of Machine Learning Research* 2013