Sentiment Analysis of Home Appliance Comment Based on Generative Probabilistic Model

Cao Shi
School of Information Science and Technology
Qingdao University of Science and Technology
Qingdao, China
caoshi@yeah.net

Ye Tao
School of Information Science and Technology
Qingdao University of Science and Technology
Qingdao, China
ye.tao@qust.edu.cn

Canhui Xu
School of Information Science and Technology
Qingdao University of Science and Technology
Qingdao, China
ccxu09@yeah.net

Xiaodong Wang
Department of Computer Science
Ocean University of China
Qingdao, China
wangxiaodongdong@ouc.edu.cn

Yanzhe Wang
School of Information Science and Technology
Qingdao University of Science and Technology
Qingdao, China
932811287@qq.com

Zihao Wang
School of Information Science and Technology
Qingdao University of Science and Technology
Qingdao, China
648487765@qq.com

***Abstract* — Sentiment analysis plays important roles in the field of e-commerce. Not only the customers and end-users but also the suppliers and manufacturers take advantages of sentiment analysis results for product improvement. In this paper, based on the Latent Dirichlet Allocation (LDA) algorithm, a generative probabilistic model is proposed and applied to predict sentiment opinions of customer's online comments. Experiments show that the proposed model can be exploited well to make sentiment analysis of home appliance comment.**

Keywords* — *sentiment analysis, home appliances, generative probabilistic model

I. Introduction

Sentiment analysis is one of vital challenges in the field of natural language processing (NLP). Especially, when e-commerce in China has been developed incredibly, the interest in online comment analysis stems from two points. The first is enthusiasm from researchers in the field of sentiment analysis. The second is from manufacturer. Through sentiment analysis of online comment, manufacturer can know well about customer, give more proper advise to help customer chose the right type of appliance, and most importantly improve product [1].

Before the boom of e-commerce, home appliance manufacturer had to deliver home appliance to customer through complex distribution channel. Whereas, when manufacturer expected feedback from customer to improve the product, results often become unpredictable. The reason is, on one hand, interaction between manufacturer and customer is very time consuming, and on the other hand, the information is spread with large deviation. However, Haier [1] has created a new information platform that fulfills real-time and accurate information exchange between manufacturer and customer. Manufacturer can effectively and efficiently gain customer's requirement from online comment, use obtained comment to improve production strategy, and then finally satisfies customer's need. Hence, in the whole process, sentiment analysis play an essential role.

E-commerce platform generally encourages customer to give comment and rating (*e.g.* from 1 to 5 stars). Generally, customer's comment on product contains information about quality, logistics, and service. Research [2] shows that sentiment analysis is supposed to assign different weights for comment and rating. However, in some cases, comment does not agree with rating, and then sentiment analysis is beneficial to identify this kind of situation. Finally, it will help e-commerce platform and manufacturer make right decision.

In this paper, sentiment analysis on customer's online comment is studied. Exploring Latent Dirichlet Allocation (LDA), a generative probabilistic model, to obtain effective information from comment. And Python based web crawler technology is exploited to gain comment and rating of Haier home appliance from [3].

II. Related Works

In the research field of sentiment analysis, pattern classification always be an urgent mission, such as classifying ratings according to comment. Classification plays a key role in opinion mining such as product reviews [4], document summaries [2]. Classification can be realized using lexicon-based strategy [5] or machine learning methodology [6]. One side, lexicon-based classification method needs to measure polarities from the dictionary, which is usually composed of words and phrases in original document. On the other side, utilizing machine learning classification, document is first projected into the vector space, and then classifier is used to classify vectors.

During the past three decades, topic model is frequently used to reduce dimensionality in NLP [7]. Topic can be

978-1-7281-4528-0/19/$31.00 ©2019 IEEE
DOI 10.1109/DSC.2019.00055

represented by probability distribution of words. Topic model refines semantic information to form a set of related topics. It maps words space into topics space. The dimension of topics space is smaller than dimension of words space, so topic model gets achievement of effective dimension reduction. Every element in the gained vector gains semantic feature, because each element is corresponding to a topic. Because of this advantage, research on topic model gains its bloom.

III. Generative Probabilistic Model

Over past three decades, probabilistic model contributes significantly to model corpus. In most cases, document or word is mapped into vector space, implementing convert from text to number. Meanwhile, dimension of corpus is also reduced. An intuitive and effective representation of word and document is TF-IDF model [8], which influence research field of natural language processing (NLP) until now. TF-IDF model is a classical probabilistic model, and it is defined as product of two statistics. One is term frequency which represents how many times a word occurs in a document. The other is inverse document frequency, which describe the inverse of how many documents in a corpus contain the word. However, the disadvantage of TF-IDF model is that dimension reduction is not remarkable. To overcome this disadvantage, kinds of dimension reduction algorithms are proposed. One of them is latent semantic indexing (LSI) [9]. LSI utilizes singular value decomposition to realize dimension reduction by decomposing matrix to the product of a column vector, a far smaller dimension matrix and a row vector. In order to make well use of probability theory, a study [10] analyzes probabilistic principle behind LSI. Based on LSI, an important improvement is achieved [11]. It is probabilistic latent semantic indexing (pLSI), in which topic model emerges more clearly. According to word exchangeability assumption [12], probability based models almost employ Bag of Word (BOW) assumption that order of words is supposed to be ignored. Mixture distribution of random variables is applied to the assumption [13], and then random variables representing words are exchangeable. Moreover, to consider the probabilistic distribution of documents, a probabilistic model named latent Dirichlet allocation (LDA) [7] is proposed. In this paper, LDA is applied well to sentiment analysis of home appliance comment.

As a generative probabilistic model, LDA is utilized to analyze home appliance comment. Fig.1 illustrates the core idea in this paper. Suppose that there are totally M comments on home appliance, and each comment has N words:

$$\text{comment} = [\mathbf{w}_1, \mathbf{w}_1, \cdots, \mathbf{w}_N]^T = [\mathbf{w}_i]^T_{1\times N} \qquad (1)$$

Here, $\mathbf{w}_i$ is a row vector, in which only one element is 1 and others are 0:

$$\mathbf{w}_i = [0, 0, 1, \cdots, 0]_{1\times V} = [\omega_j]_{1\times V} \qquad (2)$$

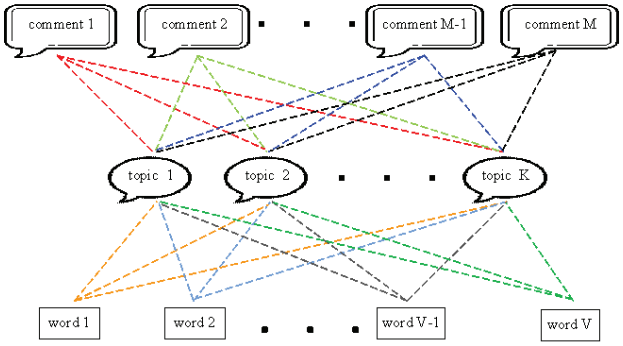


Fig. 1. “Comment-Topic-Word” mapping. Three layers of two mappings, and each mapping represents a probability distribution. The top two layers “Comment-Topic” means comment can be modeled by a probability distribution of topic. Meanwhile, the bottom two layers “Topic-Word” indicates topic can be described by a distribution of word.

As Fig.1 shows, V words are selected from M comments. Each element of $\mathbf{w}_i$ corresponds to a selected word. When the document includes a word, the corresponding element ω_j in $\mathbf{w}_i$ is 1. Otherwise, the element is 0. Generally speaking, V is far less than N.

It is obviously, each comment could has many topics. Let z_i denote a topic:

$$\sum_{i=1}^{K} p(\mathrm{z}_i) = 1 \qquad (3)$$

$$\boldsymbol{\theta} = [p(\mathrm{z}_1), p(\mathrm{z}_2), \cdots, p(\mathrm{z}_K)] = [\theta_1, \theta_2, \cdots, \theta_K] \quad (4)$$

i.e. the comment gets the probability $p(\mathrm{z}_i)$ to contain the topic z_i. And z_i obeys Multinomial distribution:

$$z_i \sim \text{Multinomial distribution}(\boldsymbol{\theta}), \qquad (5)$$

and $\boldsymbol{\theta}$ obeys Dirichlet distribution:

$$\boldsymbol{\theta} \sim \mathit{Dirichlet\ distribution}(\boldsymbol{\alpha}). \qquad (6)$$

This explains the term “latent Dirichlet allocation”: comment can be modeled by latent topics, which obeys Dirichlet distribution. In (5), $\boldsymbol{\alpha}$ is parameter of Dirichlet distribution:

$$p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \qquad (7)$$

which satisfies the constraint:

$$\theta_i \geq 0 \ \text{ and } \ \sum_{i=1}^{K} \theta_i = 1. \qquad (8)$$

In (7) and (8), K has the same meaning as shown in Fig. 1. That is the number of topics. $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ are both vectors, of which the dimensions are same: $1 \times K$. Γ is gamma function.

On the other side, in Fig.1, each topic is related to V words. Similarly, each topic has probability $p(\mathbf{w}_i)$ to be connected with word $\mathbf{w}_i$:

$$\sum_{i=1}^{V} p(\mathbf{w}_i) = 1\,. \qquad (9)$$

As for a topic, the probability distribution is:

$$\boldsymbol{\beta} = [p(\mathbf{w}_1), p(\mathbf{w}_2), \cdots, p(\mathbf{w}_V)] = [\beta_1, \beta_2, \cdots, \beta_V] \qquad (10)$$

$$\boldsymbol{\beta} \sim Dirichlet\ \ distribution(\eta) \qquad (11)$$

Different from $\boldsymbol{\theta}$, although $\boldsymbol{\beta}$ also obeys Dirichlet distribution, there is only one parameter:

$$p(\boldsymbol{\beta} \mid \eta) = \frac{\Gamma(V\eta)}{\Gamma(\eta)^V} \prod_{i=1}^{V} \beta_i^{\eta-1} \qquad (12)$$

From (1) to (12), a comment can be seemed as a result of words selection according to probability distributions: Firstly, choose topics using $\boldsymbol{\theta}$ for the comment in (5). $\boldsymbol{\theta}$ is determined by $\boldsymbol{\alpha}$ in (6) ~ (8). Secondly, based on chosen topics, and considering words distributions from (10) ~ (11), words are selected for the comment.

In the words selection progress, $\boldsymbol{\alpha}$, $\boldsymbol{\theta}$, η and $\boldsymbol{\beta}$ are parameters of probability distributions. $\boldsymbol{\alpha}$ and η determine $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ respectively. z_i is a latent topic. The number of latent topics K can be manually configured. The number of selected words V can be chosen according to word frequency, or manually configured. Based on theory of probabilistic graphical models [14], an extension of the Expectation Maximization (EM) algorithm like Variational Bayesian Inference [15, 16] or Markov Chain Monte Carlo (MCMC) methods such as Gibbs Sampling [17] are made well use of to estimate parameters of probability distributions. Among these parameters, $\boldsymbol{\theta}$ can be used to represent comments as an algebraic feature.

IV. Experiments and Results

Ten types of refrigerators manufactured by Haier [1] are used to analyze sentiment from online customer comments. Comments and rating (from 1 star to 5 stars) are extracted from [3] using Python. Table 1 lists ten types of Haier refrigerators from BC-93TMPF to BCD-325WDSD. Bad comments are cleansed. Five levels (1 star ~ 5 stars) represent 5 classes. And 1000 comments are collected for each star level of a type. In this way, the corpus consist 50,000 comments.

The generative probabilistic model in last section is employed to extract topic distribution $\boldsymbol{\theta}$ in (4) for each comment, SVM is exploited as classifier, and "Star/Stars for ratings" in Table 1 is used as "Comment ID". The corpus in Table 1 is divided to training set and validation set with a ratio 1:1.

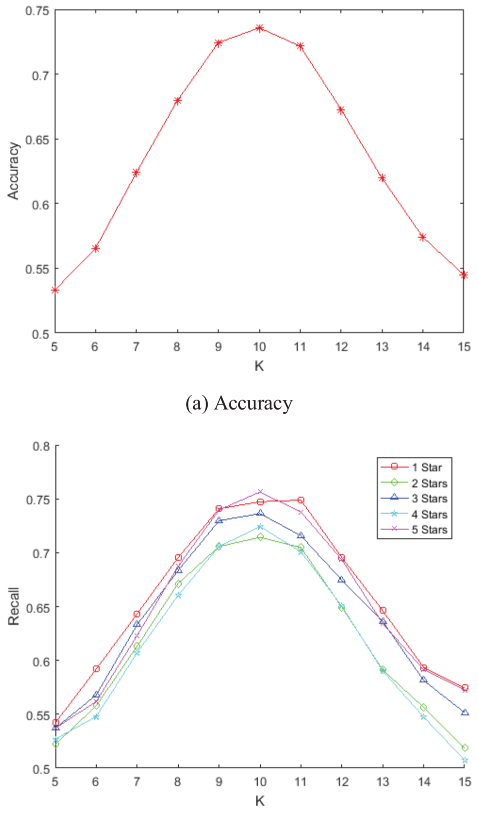
TABLE I. Comment dataset of Haier refrigerator

No.	Type	Star/Stars for ratings	Number of comments
1	BC-93TMPF	1、2、3、4、5	1000 × 5 = 5000
2	BCD-160TMPQ	1、2、3、4、5	1000 × 5 = 5000
• • •	• • • • • • •	• • • • • • •	• • • • • • •
10	BCD-325WDSD	1、2、3、4、5	1000 × 5 = 5000

In this paper, accuracy is defined as:

$$accuracy = \frac{\sum_1^5 t(i)}{\sum_1^5 t(i) + \sum_1^5 f(i)} \qquad (13)$$

Here, $t(i)$ denotes the "true i star/stars". Naturally, $f(i)$ means "false i star/stars". **Fig. 2(a)** illustrates the accuracy of the whole comment dataset in Table 1. K is the number of latent topics in (3) and (4). It is easy to find the accuracy in (13) gets biggest value when the number of latent topics is 10. And it decreases when K is over 10.



(a) Accuracy

(b) Recall

Fig. 2. Accuracy and Recall. In (a) accuracy increases with the latent topics number *K* in [5,10], and then decrease when K > 10.

Accuracy describes the global performance of the generative probabilistic model for a balanced data set in Table 1. On the other hand, to look into details of individual rating, recall is formulated as follows:

$$recall_i = \frac{t(i)}{num(i)}. \quad (14)$$

In this equation, i indicates 5 ratings and $num(i)$ is the number of comments of i star/stars. Evidently, $num(i)$ is 5000.

In **Fig. 2(b)**, 4 curves (2~5 stars) obtain biggest value at $K = 10$, and the curve of 1 star gets biggest value at the nearest position $K = 11$. Therefore, 5 recall curves confirm conclusion of accuracy curve that when the number of latent topics equals 10 the generative probabilistic model gains the best performance.

Experiments conclude that the generative probabilistic model in this paper can be exploited well to make sentiment analysis of home appliance comment.

V. Conclusion

In this paper, a study is made on sentiment analysis of home appliance comment based on the generative probabilistic model. SVM is exploited as classifier. Experiments show the approach in this paper should be useful to verify effectivity of home appliance comment. The further research will be focused on integration of sentiment analysis of home appliance comment with other key information of manufacturing industry, in order to enrich value chain of home appliance manufacture.

Acknowledgment

This research is supported by National Key Research and Development Plan (No. 2017YFB1400903).

References

[1] Haier. Available: https://www.haier.com/cn/

[2] E M Alshari, A Azman, N Mustapha, S C Doraisamy, M Alksher. Prediction of Rating from Comments Based on Information Retrieval and Sentiment Analysis. in Information Retrieval and Knowledge Management (CAMP), 2016 Third International Conference on, 2016, pp. 32-36.

[3] Jd.Com, Available: https://www.jd.com/ ; https://corporate.jd.com/home

[4] Z Singla, S Randhawa, S Jain. Statistical and Sentiment Analysis of Consumer Product Reviews. in 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2017, pp. 1-6.

[5] M Taboada, J Brooke, M Tofiloski, K Voll, M J C l Stede. Lexicon-Based Methods for Sentiment Analysis. 2011, 37 (2): 267-307.

[6] P V Rajeev, V S Rekha. Recommending Products to Customers Using Opinion Mining of Online Product Reviews and Features. in Circuit, Power and Computing Technologies (ICCPCT), 2015 International Conference on, 2015, pp. 1-5.

[7] D M Blei, A Y Ng, M I Jordan, J Lafferty. Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003, 3): 993-1022.

[8] G Salton. Introduction to Modern Information Retrieval. New York: McGraw-Hill, 1983.

[9] S Deerwester, S T Dumais, G W Furnas, T K Landauer, R A Harshman. Indexing by Latent Semantic Analysis. Journal of the American Society of Information Science, 1990, 41 (6): 391-407.

[10] C H Papadimitriou, H Tamaki, P Raghavan, S Vempala. Latent Semantic Indexing: A Probabilistic Analysis. symposium on principles of database systems, 1998, 159-168.

[11] T Hofmann. Probabilistic Latent Semantic Indexing. international acm sigir conference on research and development in information retrieval, 1999, 50-57.

[12] D J Aldous. Exchangeability and Related Topics. In: Hennequin P.L. (eds) École d'Été de Probabilités de Saint-Flour XIII — 1983. Lecture Notes in Mathematics, Berlin, Heidelberg, 1983, 1–198.

[13] B d Finetti. Theory of Probability. Vol. 1-2: John Wiley & Sons Ltd., 1990. Reprint of the 1975 translation.

[14] D Koller, N Friedman. Probabilistic Graphical Models: Principles and Techniques: Massachusetts: MIT Press, 2009.

[15] C Fox, S Roberts. A Tutorial on Variational Bayesian Inference. Artificial Intelligence Review, 2011, 38 (2): 85-95.

[16] M J Beal. Variational Algorithms for Approximate Bayesian Inference. PhD, Gatsby Computational Neuroscience Unit, University College London, 2003.

[17] C Bishop. Pattern Recognition and Machine Learning: Springer, 2007-10-1.