Sentiment Analysis Using Modified LDA

Jingyi Ye[1,2(✉)], Xiaojun Jing[1,2], and Jia Li[3]

[1] School of Information and Communication Engineering,
Beijing University of Posts and Telecommunications, Beijing, China
yejingyi@naver.com
[2] Key Laboratory of Trustworthy Distributed Computing and Service (BUPT),
Ministry of Education, Beijing University of Posts and Telecommunications, Beijing, China
[3] School of Engineering and Computer Science, Oakland University, Rochester, USA

Abstract. The technology of the Internet develops rapidly recent years, the public tends to share their reviews, opinions and ideas on the Internet. The forms of these subjective texts are free and concise, and they contain a wealth of sentiment information. In this paper, a modified latent Dirichlet allocation (LDA) model and support vector machine (SVM) are used for sentiment analysis of subjective texts. Analysis of sentiment could help producer to enhance the products and guide user make better choices as well. We apply a modified LDA model using term frequency-inverse document frequency (TF-IDF) algorithm to mine potential topics, find the most relevant words of the topic and represent the document. Then we use SVM to categorize the texts into two classes: positive and negative. Experiment results show that the performance of the modified LDA approach is better than the traditional LDA model.

Keywords: Opinion mining · Sentiment analysis · Latent Dirichlet allocation
TF-IDF algorithm

1 Introduction

The technology of the Internet and the integration of dynamic web developed rapidly in the past decades. People begin to enjoy the fun of network applications and express their mood, thoughts and ideas on the network. Thus, a large number of potentially valuable sentiment texts appear on the Internet.

Sentiment analysis has been widely used in many areas, and can be beneficial to many aspects. From the perspective of individual users, sentiment analysis could affect the individual's understanding and attitude towards the specific object. Form an enterprise point of view, sentiment analysis of user views and service experience can enable enterprises to know their products for the merits of an accurate grasp and develop

Project 61471066 supported by NSFC.

S. Sun et al. (eds.), *Signal and Information Processing, Networking and Computers*, Lecture Notes in Electrical Engineering 473,
https://doi.org/10.1007/978-981-10-7521-6_25

effective strategies in business competition and product development. From the government point of view, sentiment analysis can help government promptly understand people's views and attitudes [1].

Recently, researchers show a great interest in exploring new text representation models for improving the accuracy and efficiency of text processing. The theoretical idea of the topic model is that document is a mixture of several topics, each of which contains multiple terms of word distribution. Topic model obtains the semantic related topic collection hidden in the document through the common information of words in the document. Topic model transforms the document from word space to topic space, and expresses the document in a lower dimension space. The origination of topic model was from latent semantic indexing (LSI), and then topic model evolved to a variety of forms, especially latent Dirichlet allocation (LDA) and LDA based models. The topic model has a sentiment feature that can extract the implicit sentiment topic from the data set. Due to the good mathematical foundation and flexible extensible of topic model, it has been widely used in text mining and information processing tasks [2].

It is time-consuming to extract words manually in a document, so we need to extract words automatically. People use algorithms to divide the document into words, give the corresponding weight of each word according to some methods. Many approaches can be utilized for the weight calculation, such as frequency function, entropy function and term frequency-inverse document frequency (TFIDF). TF-IDF is a statistical method with a high accuracy and recall, and it can be applied to a document in a corpus for evaluating the importance of a word. The principle of this method is that if a word appears in the document at a high frequency and is rarely present in other documents, it is considered that the word has a good class distinction and is suitable for the classification [3].

In this paper, we analyze the sentiment of subjective documents with a modified LDA model and SVM. We use TF-IDF algorithm to modify the traditional LDA model. According to the potential topic mining method, the documents are represented by the modified LDA model. Then the topics are divided into two classes via support vector machine (SVM), namely positive and negative. Experiment results prove the effectiveness and practicability of the modified model.

The outline of this work is shown as follows. In Sect. 2, a brief introduction of sentiment analysis, LDA model and TF-IDF algorithm is presented. A specific description of the proposed approach is followed in next section. After that, the experiment results are shown in Sect. 4, and conclusions are made in the final section.

2 Related Work

2.1 TF-IDF Algorithm

Jones first proposed the concept of Opposed to Document Frequency (IDF) in [4] in 1972. He pointed out that the weight of words can be given according to the frequency of words appeared in the corpus. If a word has a higher number of occurrences in the corpus, it has a lower information entropy and the corresponding weight, and vice versa.

Back to 1973, the TF-IDF algorithm was first proposed by Salton, and it was effectively applied to the field of information retrieval [5]. In 1988, a detailed description of the use of multiple words weighting method was proposed [6]. TF-IDF mainly reflects the following idea: the higher frequency of a word, the stronger its ability to distinguish the content of the document (TF) and the wider the scope of a word in the corpus, the lower attributes of the document content (IDF) [6].

2.2 Latent Dirichlet Allocation

LDA is a generation probability model, which is the method of modeling the topic information of text data [2], with good mathematical basis and flexible expansibility. LDA could mine words with topic, through the three-tier Bayesian model. Titov [6] proposed a multi-granularity LDA model and applied it to sentiment summary generation with a multi-topic sentiment model. Zhao [7] proposed the ME-LDA model, which combines the maximum entropy with the topic model.

2.3 Sentiment Analysis

With the arrival of large data age, the analysis of massive data could access valuable products or services. However, most of the information is unstructured and difficult to manually analyze. Sentiment analysis is a new research direction that rises to deal with the analysis of implied emotional information [8]. Sentiment analysis can automatically analyze the subjective text, effectively identify and excavate the emotional information. The main tasks of sentiment analysis include the extraction, classification and retrieval. Sentiment extraction refers to the extraction of relevant sentiment words and evaluation of objects. Sentiment classification is to distinguish subjective and objective texts and determine the polarity of subjective text. Sentiment retrieval is used to retrieve documents containing relevant sentiment information to meet users' query needs. In this paper, we mainly concerns on the extraction and classification tasks.

3 The Proposed Approach

Basic steps of the proposed approach are as follows: we organize and preprocess the documents in the corpus at first, then we use the modified model to learn text representation of training set and test set, finally train the SVM classifier to complete the sentiment classification of test set.

3.1 Pretreatment

The documents in the corpus are chapters composed of sentences. In order to fit the structured text needed by the classification algorithm, it is necessary to preprocess the corpus.

The main pretreatment in this paper is to remove the stop words. The purpose of this process is to reduce the spatial complexity and time complexity of the method and to

improve the accuracy of the follow-up feature selection. Stop words usually do not make sense or have contribution to the classification results, including punctuation, modal particles, prepositions and conjunctions, such as “are”, “above”, “and”, etc. The implementation of this process requires a vocabulary of stop words, and these words can be filtered out by applying an approach of string matching. For instance, the target is “a little girl held an apple from the basket and cried”, “a”, “an”, “from”, “the”, “and” are stop words, and the remaining words are “little girl held apple basket cried”.

3.2 Model Description

After the preprocessing, the modified LDA model is used to deal with the corpus. We use TF-IDF algorithm to modify LDA model.

TF-IDF Algorithm
TF-IDF is a widely used algorithm in information retrieval field. In recent years, researchers have used TF-IDF algorithm to calculate the weight of features and achieved good results [5].

$$W = TF \times IDF = TF \times \frac{1}{DF}. \tag{1}$$

The frequency of word T in document D is TF, which is used to calculate the capability of the word to describe the document. IDF represents the inverse of the frequency of the document D containing the word T in the corpus, which is used to calculate the capability of the word to distinguish the document. If the frequency of a word is high in its own document but low in other documents, this word has a strong ability to distinguish it from other documents and is assigned to a high weight.

Latent Dirichlet Allocation
The generation probability model LDA imitates the process of human writing. If we want to write a document, it is often necessary to determine what topics to write. After determining the topics, we will use some words highly related to the topics to describe them. Thus, a document usually consists of multiple topics, each topic is described with the high frequency words associated with the topic.

Blei proposed an unsupervised full probability generation model, LDA, in 2003 [2]. LDA has a clear internal structure and mathematical basis, which could be calculated by efficient probability estimation algorithm. LDA model has been widely used in text classification, text modeling, image processing and information retrieval and other fields.

LDA model is a method of modeling documents, topics and words, which transforms the traditional word vector expression into the topic vector expression [9]. The benefit of doing this is obvious that the expression based on topics reduces the dimension of feature space. LDA model has a clear logical structure, containing the document layer, the topic layer and the word layer. Each layer is adjusted by variables and parameters. Figure 1 shows the graphical representation of LDA model.

The box indicates that the content is repeated, the number of repetitions is in the lower right corner, the shadow represents the observed value, the empty node represents the implied random variable or parameter, and the arrow indicates the dependency.

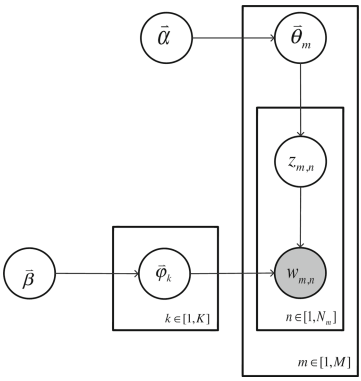

Fig. 1. Graphical model representation of LDA

The symbols in LDA model are explained as follows:

(1) K: the number of topics,
(2) M: the number of documents,
(3) N_m: the total number of words in document m,
(4) α: Dirichlet prior parameters of multinomial distribution θ_m under each document,
(5) β: Dirichlet prior parameters of multinomial distribution φ_k under each topic,
(6) $z_{m,n}$: the topic of the word n of the document m,
(7) $w_{m,n}$: the word n of the document m,
(8) θ_m: the topic distribution under the document m,
(9) φ_k: the word distribution under the topic k.

Modified LDA

TF-IDF algorithm is used to improve the performance of LDA. The main idea of the modified LDA model is to replace the untreated words in the traditional LDA model using the words treated by TF-IDF algorithm. The words treated by TF-IDF algorithm have a weight and have the ability to express the importance of the word to the current document and to distinguish it from other documents. The structure of modified LDA is shown in Fig. 2, and t_k is used to describe the word distribution which is treated by TF-IDF algorithm in the slash part.

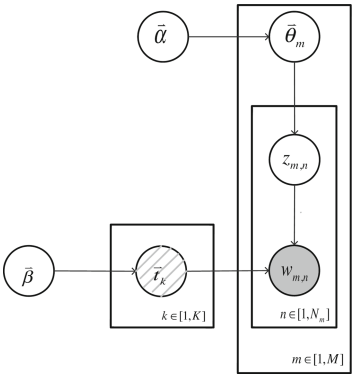

Fig. 2. Graphical model representation of modified LDA

The corpus D contains all documents, which served as the input part of the modified LDA model. K is the number of topics. The output of modified LDA is the most relevant words of each topic. The generative process of the modified model is as follows:

For each document, we generate the multinomial distribution of topic θ_m, which is derived from the Dirichlet distribution of the parameter α.

For each topic, we generate the multinomial distribution of word t_k, which is derived from the Dirichlet distribution of the parameter β. For each topic, the sum of $z_{m,n}$ is calculated by using Gibbs sampling and an approximate posterior on θ_m and t_k is obtained. The formula is as follows:

$$p(\Theta, \Phi | D^{train}, \alpha, \beta) = \sum_{z} p(\Theta, \Phi | z, D^{train}, \alpha, \beta) \times p(z | D^{train}, \alpha, \beta). \tag{2}$$

We repeat the above procedure to get the output of modified LDA. Through the above process, we get a list of topics for each document, and for each topic we get some of the most relevant words with the topic.

4 Experiment Results

We demonstrate the performance of modified LDA model by experiment in this section. The dataset we used is the sentiment dataset from Cornell University, which contains 2000 movie reviews with both positive and negative data of 1000. In the experiment, we use 1500 reviews as training data and 500 reviews as test data. The classifier in the experiment we used is SVM. Also we use five-fold cross validation method and measure the performance with precision, recall and F1-value.

The experiment compares the influence of different topic numbers on LDA and modified LDA and the result of F1-value is shown in Fig. 3 and the result of precision and recall in shown in Table 1. The results show that the modified LDA has a better performance than traditional LDA model. When topic number is 40, we get the best performance with precision at 0.9, recall at 0.87 and F1-value at 0.87.

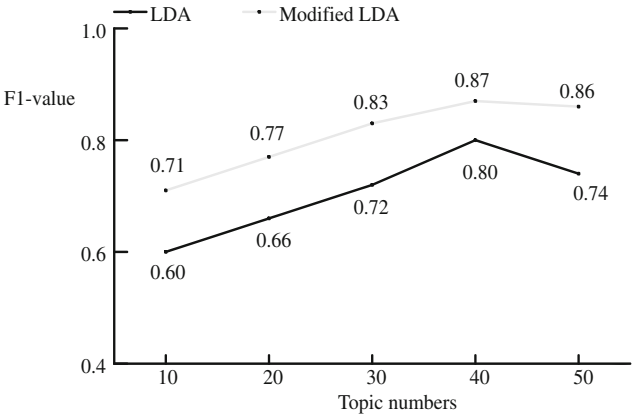

Fig. 3. The classification performance of F1-value using LDA and modified LDA to represent the document, with different topic numbers.

Table 1. The classification performance of precision and recall using LDA and modified LDA to represent the document, with different topic numbers.

Topic numbers	LDA		Modified LDA	
	Precision	Recall	Precision	Recall
10	0.63	0.61	0.82	0.73
20	0.67	0.67	0.85	0.78
30	0.72	0.72	0.88	0.84
40	0.8	0.8	0.9	0.87
50	0.74	0.73	0.86	0.86

This section mainly discusses the corpus selection, evaluation criteria, LDA model and modified LDA model experimental results. The text sentiment classification of modified LDA model has improved the classification quality than traditional LDA model. The comparative experiment using movie reviews data show the effectiveness of the modified LDA model.

5 Conclusions

Sentiment analysis is an important research area in natural language processing field. It is of great significance for the research of Internet public opinion supervision, commodity sales and information screening. We have studied the text representation model in sentiment analysis problem, and modified the LDA model with TF-IDF algorithm. Experiments show the effectiveness of modified LDA. In this paper, we use modified LDA model in binary classification problem, in the future work we can use the modified model in multi-classification problem.

References

1. Cambria, E.: Affective computing and sentiment analysis. IEEE Intell. Syst. **31**(2), 102–107 (2016)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
3. Guo, A., Yang, T.: Research and improvement of feature words weight based on TFIDF algorithm. In: 2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference, pp. 415–419 (2016)
4. Jones, K.S.: A statistical interpretation of term specificity and its application in retrieval. J. Documentation **28**(1), 11–21 (1972)
5. Salton, G., Yu, C.T.: On the construction of effective vocabularies for information retrieval. In: Proceedings of the 1973 Meeting on Programming Languages and Information Retrieval, pp. 11–21 (1972)
6. Titov, I., McDonald, R.: Modeling online reviews with multi-grain topic models. In: The 17th International World Wide Web Conference, pp. 111–120 (2008)
7. Zhao, W.X., Jiang, J., Yan, H., Li, X.: Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 56–65 (2010)
8. Poria, S., Gelbukh, A., Hussain, A., Howard, N., Das, D., Bandyopadhyay, S.: Enhanced SenticNet with affective labels for concept-based opinion mining. IEEE Intell. Syst. **28**(2), 31–38 (2013)
9. Li, Y., Zhou, X., Sun, Y., Zhang, H.: Design and implementation of Weibo sentiment analysis based on LDA and dependency parsing. China Commun. **13**(11), 91–105 (2016)