Frontiers in Business, Economics and Management
ISSN: 2766-824X | Vol. 4, No. 3, 2022

Sentiment Analysis and Personalized Recommendations Based on JD.com Reviews

Yutong Zhang[1, a]

[1]School of Computer Science and Information Engineering, Hubei University, Wuhan 430000 China
[a]Corresponding Author's Email: 201931119020041@stu.hubu.edu.cn

Abstract: The general big data personalized recommendation is based on the number of times or the length of time users click on a related content, but in many cases, these cannot be the most direct basis for accurate recommendation, and there may be cases such as wrong clicks by users. These factors and a large number of related products or articles recommended to users may cause users' disgust. This article conducts sentiment analysis on JD.com reviews as an example, obtains the user's likes and dislikes, and then makes accurate personalized recommendations, so that a greater understanding of preferences can improve the effect of recommendations, and more accurate personalized recommendations.

Keywords: Natural language processing, Sentiment analysis, Personalized recommendation, JD.com review.

1. Introduction

Nowadays, the science and technology are developed rapidly, people are in an environment where information is almost transparent. Big data can track any user using the internet and obtain all their information exposed on the Internet. Many merchants saw business opportunities from it, obtained a large amount of information through users' use of their software, and analyzed the data through a series of methods to obtain each user's preferences, and made personalized recommendations, which greatly increased traffic or profits. profit target. Personalized recommendation exists in almost every software, but the data of most software personalized recommendation basically comes from the length of time users stay in a certain aspect of content or the number of clicks, but in fact not all of these data are is the user's real preferences.

In China, JD.com which is one of the biggest e-commerce companies collects a large mass of data on user reviews as many users place orders for shopping every day. It is not difficult to find that there are many emotional factors in these user comments. For example, users who dislike will publish some negative words, while users who like them will have positive words. According to these emotional comments, the software background can intuitively see the user's satisfaction with the products and analyze whether the products of the same category or derived supporting products can attract users, so as to infer future recommendations for users. Content, the platform adds this method to greatly improve the effect of the recommendation. Therefore, sentiment analysis based on comments must be considered in personalized analysis.

In the field of the computer science and artificial intelligence, the natural language processing is a very important orientation. It refers to the technology which human could use the natural language to communicate with the machines. Various theories and methods of communication. Nasukawa et al. [1] were the first group of people who originally proposed the natural language processing technology of the sentiment analysis, which refers to extracting people's emotions or opinions from written words. Sentiment analysis is currently a popular research direction at home and abroad. Compared with sentiment analysis in Chinese, sentiment analysis in English will be more accurate and simple, because English has no major semantic and grammatical changes, but Chinese sentiment analysis will be more accurate and simple. Since it is ever-changing, it is relatively difficult to process, and the accuracy of the processing results will also decrease. Sentiment analysis and topic model are two hot research directions. According to different application fields, text sentiment analysis technology can be divided: (1) Text sentiment analysis of product reviews is generally used to assist consumers in decision-making and business public opinion monitoring [2-3]; (2) Text sentiment analysis for news comments is used to process text comments published by news events, and is generally used to guide public opinion based on the correct public opinion [4-5].

Two main research methods in the field of the sentiment analysis include the unsupervised one and the supervised one. In the early period, the supervised learning is about shallow models such as Support Vector Machine (SVM), Maximum Entropy, Naïve Bayes, etc. The unsupervised learning is based on semantic analysis, dictionary and the other methods. Because of the appearance of the deep learning, regression tasks and many classifications have achieved best results. In recent years, it has become a hotspot of the research in the field of the sentiment analysis which is applicated with the deep learning.

In terms of model establishment, the LDA model is the mainstream model in the current topic model, it's about the PSLA which is added with the Bayesian framework, meanwhile it's also a typical bag model. It can give the subject of each document in the document set in the form of a probability distribution, so as to extract their subject by analyzing the documents, and perform clustering or text classification on the subject.

2. Technical Implementation Method

2.1. Technical Assumptions

According to the articles or some paragraphs published by

the user, the sentiment analysis is carried out to obtain the content that the user likes and dislikes, and the obtained results are used to make personalized recommendations, so as to obtain the user's preferences more accurately. To solve the key problem, we need to collect the data of these segments, and use the relevant libraries in python to complete word segmentation, model establishment, etc. to obtain the results.

2.2. Word Vector

Computers cannot directly read human vocabulary, so they cannot directly process text. It needs to be converted into data that can be recognized by computers. Therefore, first of all, the acquired text data must be converted into word vectors so that computers can recognize them. This process can also be called the process of word embedding, and there are mainly two ways based on statistics and based on language models. Statistics-based methods include bag-of-words model, One-Hot encoding and Term Frequency-Inverse Document Frequency (TF-IDF) [6], etc.; language model-based methods include well-known word2vec and BERT. Word2vec[7-8] is an open source word embedding tool based on deep learning proposed by Google, which mainly includes two models, namely Skip-Gram and Continuous Bag Of Words (CBOW). TF-IDF, Word2vec and a weighted word2vec make three methods of extracting word vectors [9]

2.3. Stop Words

The concept of the stop words is that in order to save the storage space and promote the efficiency of searching the information in the information retrieval, there will automatically filtered out some word or words before or after the processing of the natural language data (or text). These word or words are named stop words. Those stop words are input by human and generated without automation. Then, the generated stop words would become a stop word list. Nevertheless, it's no certain list of the stop words which could work for all tools.

Stop words can be roughly divided into two categories:

(1) It is widely used and can be seen everywhere on the Internet. For example, the word "Web" appears on almost every website. Search engines couldn't make sure that those words could work out the true search results. Also, it's hard to help narrow the search, meanwhile it also will reduce the search's efficiency;

(2) Those modal particles, adverbs, prepositions, conjunctions and so on usually don't have clear meanings by their own. They just have certain useful effects when they are added to complete sentences.

2.4. LDA Model

The LDA model is a potential Dirichlet allocation, and it's a very common topic model which often used in experiment. This kind of model belongs to unsupervised learning model which could be applied in identifying the potential topic information in the large-scale document sets or corpora. The LDA model also known as a three-layer Bayesian probability model, includes three-layer structures: words, topics and documents.

For each document in the corpus, the generation process is as follows:

(1) Extracting a topic from the topic distribution in each document;

(2) Extracting a word from the word distribution corresponding to the step one's topic;

(3) Repeating the step two's process until that all the words in the document are traversed

The overall process of LDA is as follows (define the meaning of some letters: document set D, topic set T):

(1) Each document d in D is regarded as a word sequence <w1,w2,…,wn>, wi represents the i-th word, and d has n words. (It is called wordbag in LDA, in fact, the appearance position of each word has no effect on the LDA algorithm)

(2) All the different words involved in D form a large set VOCABULARY (VOC for short), LDA takes the document set D as input, and hopes to train two result vectors (set to be clustered into k topics, and VOC contains m words in total));

(3) For each document d in D, the probability θd<pt1,…,ptk> corresponding to different topics, where pti represents the probability that d corresponds to the ith topic in T. The calculation method is intuitive, pti=nti/n, where nti represents the number of words in d corresponding to the ith topic, and n is the total number of all words in d.

(4) For each topic in T, the probability of generating different words φt<pw1,…,pwm>, where pwi represents the probability that t generates the i-th word in VOC. The calculation method is also very intuitive, pwi=Nwi/N, where Nwi represents the number of the i-th word in the VOC corresponding to topict, and N represents the total number of words corresponding to topict.

The core formula of LDA:

$$p(w|d)=p(w|t)*p(t|d) \quad (1)$$

3. Jingdong Comments Are Expected to Be Obtained

3.1. Obtaining Method

In order to obtain the latest comment data, it is implemented through crawler technology. The crawler technology can automatically crawl the information on the webpage by writing a program, and write the code according to the URL and user-agent obtained in the webpage to crawl the information of the specified page. The crawler technology is often limited by various websites, and sometimes it is frequently accessed. It will be monitored by the platform in the background to limit the number of visits, which is not applicable when obtaining a large amount of data. The amount of data obtained here is not large, so the crawler is directly used to obtain the data.

The data crawled from JD.com comments in this paper are divided into the two categories, one is about negative comments and another is about the positive comments. According to the corresponding parameters in the transformation code, the two types of data are saved in different documents for backup (the data content is the comments of a certain brand of computer).

comment
Not good, there is no matching computer bag, you should know not to buy it
The right mouse button can't be used, and the delivery is slow
The touchpad button is loose, and it will shake if you press it lightly
Always won't turn on after shutting down
Playing an exciting battlefield still keeps the screen flickering
Not bad, still blue screen, hit CF but no response
You do not fill in the content, the default is good
It took me a week to get a black screen
There is a problem with the quality, I exchanged it twice
It's really a bit stuck. It gets stuck from time to time.

Figure 1. Extraction of bad reviews

comment
The all-in-one machine has a beautiful atmosphere. It can be turned on and off in seconds. It is completely stress-free when running some basic programs and small games. It is completely silent. T
It's pretty, and it's simple. The screen is matte and looks comfortable. The keyboard and mouse are fine. Boot up is fast. It is very suitable for people like me who only use the computer to view the v
The children at home want to take online classes, and the computers provided by the company are all DELL. The quality of Dell is still unremarkable, and the after-sales service is also in place. Just
This computer is very nice to use in office, and the appearance is also very atmospheric. The original configuration of win10 is also very smooth, and the cost performance is particularly high.
Appearance: beautiful and clean
Performance configuration: especially suitable for office
Running speed: very fast
Screen effect: very high refresh rate, very clear
Product packaging: intact
I am very satisfied with the appearance, performance configuration, running speed, and the effect of the screen. I am very satisfied with the product packaging. The product packaging is also very
Dell computer all-in-one machine, the value of the **. Jingdong logistics is very fast, it is for office use, and the configuration is sufficient. It runs quickly and installs easily. Follow-up use and come
Good quality and low price, the appearance is high-end and high-grade, the boot speed is fast, practical, the operation is very smooth and the furniture style is very suitable, I like it very much and
The computer is very smooth to use, boots in seconds, comes with office software, the screen is super large and the border is negligible, and the white body and keyboard are beautiful
The office computers have been replaced by a new batch. They have high appearance and are very fast. They are very good for office use. satisfy! ! !
Great shopping experience, the network speed is very smooth, and the memory is large enough. Finally, I can improve my work efficiency. After changing to a new computer, my performance will

Figure 2. Extraction of positive comments

4. Data Preprocessing

4.1. The Overall Preprocessing Process

First, perform word segmentation on the obtained Jingdong comment text data, then perform part-of-speech tagging, stop word deletion, etc. The next step is to perform syntactic analysis, and finally generate word vectors as needed. It should be noted that when analyzing Chinese text, a special word segmentation algorithm is required to assist the computer in identifying words, phrases and fixed collocations in Chinese text.

There are roughly four kinds of common stop word lists in Chinese:

(1) The Harbin Institute of Technology Chinese stop word list;

(2) The Chinese stop word list;

(3) Baidu stop word list;

(4) Sichuan University machine Smart Lab stop-word library.

In this experiment, the using stop word list in Chinese is the forth one above.

5. Model Establishment

5.1. LDA Model Establishment

First use Gensim for model building: the corpora. Dictionary method builds a specific object to store corpus data, and converts the cleaned information into a Gensim-approved corpus form. The second step uses models. LdaModel to train the model on the processed expectations.

6. Result

After writing and running the code, the topic extraction situation after LDA modeling is obtained (the experimental results of the positive comments and the negative comments are illustrated in the Figure 3 and Figure 4), which are divided into two categories: (1) positive comment topics extracted from positive comments; (2) negative comment topics extracted from negative comments. Through the theme results extracted from the experiment, the emotional trend of each user for this product is analyzed, so as to make a more accurate recommendation for the next step according to the emotional trend.

Table 1. Experimental results of positive

-	0	1	2	3	4	5	6	7	8	9
Theme 1	very	one machine	run	computer	totally	Dell	very	complete	equipment	buy
Theme 2	very	elegant	smooth	exterior	start	computer	quick	worth	one machine	like
Theme 3	computer	run	special	very	very	speed	screen	very	high	equipment

Table 2. Experimental results of negative

-	0	1	2	3	4	5	6	7	8	9
Theme 1	twice	quality	change	goods	problem	a little	carton	shaking	tap	bottom
Theme 2	reaction	blue screen	expensive	heavy	old	play	exciting	splash screen	battlefield	none
Theme 3	nothing	after that	don't buy	all	not good	knowledge	computer game	shutdown	high	useless

For users who are on the positive side, it can be found from the keywords that the satisfaction is very high, and the products of the same category can no longer be pushed, and can continue to recommend some derivative products about the product, such as peripherals, accessories, etc. For users on the negative side, it can be found from the subject words that they are not satisfied with the product. According to the main dissatisfaction problems raised by the negative comments, other products of the same category and different brands that meet these expectations are recommended, such as providing higher quality and more beautiful. Product push with better performance.

7. Conclusion

By using the sentiment analysis, the efficiency and accuracy of the recommendation can be greatly improved, and a large number of repeated products can be avoided, thereby increasing the benefits.

References

[1] Nasukawa T & Yi J. (2003) Sentiment analysis: Capturing favorability using natural language processing. Proc of the 2nd International Conference on Knowledge Capture, 70-77.

[2] Pang B & Lee L. (2008) Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2): 1-135.

[3] Hu M & Liu B. (2004) Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 168-177.

[4] Ren F & Wu Y. (2013) Predicting user-topic opinions in twitter with social and topical context. IEEE Transactions on Affective Computing, 4(4): 412-424.

[5] Hirschberg J & Manning C D. (2015) Advances in natural language processing. Science, 349(6245): 261-266.

[6] Aizawa A. (2003) An information-theoretic perspective of T F-IDF Measures. Information Processing & Management, 39(1): 45-65.

[7] Le Q & Mikolov T. (2014) Distributed representations of sentences and documents. Proc of the 31st International Conference on Machine Learning, 1188-1196.

[8] Mikolov T, Chen K, Corrado G & et al. (2014) Efficient estimation of word representations in vector space. arXiv: 1301,3781.

[9] Li Rui, Zhang Qian & Liu Jia-yong. (2017) Microblog sentiment analysis based on weighted word2vec. Communications Technology, 50(3): 502-506.