

# **Emotional Analysis of E-Commerce Online Comment Data**

Lili Jin

Management Department South China College Guangdong University of Foreign Studies Guangzhou, China

Abstract—Introducing text emotion analysis into e-commerce online comment data can judge the emotional tendency of e-commerce online comments. At the same time, the application of Python Jieba participle and LDA topic model enables more detailed understanding of the emotional changes in online comments of e-commerce. Taking product reviews of Midea water heater on JD.com as an example, the empirical analysis shows that the online product review, by introducing text sentiment analysis, determines the consumer emotion tendency, which can help with the electric business improvement project for merchants to provide accurate marketing, set up new businesses evaluation way, dynamically monitor customer emotional tendency, and timely grasp the emotional trend of E-commerce industry.

Keywords—LDA Topic Model; Python Programming; Online Review of E-Commerce; Emotional Analysis

#### I. INTRODUCTION

With the development of mobile Internet, people's demand on online shopping has become higher and higher, it brings great development opportunities for e-commerce platforms such as Jingdong and Taobao. At the same time, the electronic business platform improves the quality of goods, lower commodity prices. Otherwise, it's very necessary that understanding consumer's emotional tendency, and there is also a very important way for company to review data and mine internal data, and helping businesses improve the quality of service, attract more consumers. Therefore, these papers selected the online comment data of Midea water heater on Jingdong, improved the traditional model of text emotion analysis, directly reflected the operation and management effect of e-commerce merchants by using Jieba segmentation and LDA theme model in Python, and finally give valuable Suggestions for improvement.

#### II. THE RESEARCH METHODS

The article builds the model based on the data of Midea water heater from Jingdong. We analyze these data through preprocessing the data, segmenting Chinese word, filtering the stop words. Next, we mine the model using the methods of coding depth study, semantic network and the theme of LDA. We judge the tendency and mine the information, then provide feasible business countermeasures and Suggestions. In addition, this paper is implemented by Python programming, including comment data extraction, primary data filtering, text emotion analysis, etc.

### III. THE EMPIRICAL ANALYSIS

### A. The Source of Data

To analyze the water heater review data from the electric business platform, we need to collect review data. Therefore, this paper chooses to crawl data with an octopus collector. After processing the collected comment data, some comment data of the original text can be obtained as shown in Table I.

TABLE I. ORIGINAL COMMENT TEXT

| ID | Collected | Sent  | Electric<br>Business<br>Platform | Brand | Comments                                     | Time      | PageUrl           |
|----|-----------|-------|----------------------------------|-------|--|-----------|-------------------|
| 1  | TRUE      | FALSE | Jingdong                         | Midea | Jingdong mall trust, buy at ease, with       | 2019/3/21 | http://s.club.jd. |
| 2  | TRUE      | FALSE | Jingdong                         | Midea | Bought for company dormitory, door-to        | 2019/3/21 | http://s.club.jd. |
| 3  | TRUE      | FALSE | Jingdong                         | Midea | The beauty is trustworthy and the qualit     | 2019/3/18 | http://s.club.jd. |
| 4  | TRUE      | FALSE | Jingdong                         | Midea | Good good , the first time to buy on jd.c    | 2019/3/18 | http://s.club.jd. |
| 5  | TRUE      | FALSE | Jingdong                         | Midea | I'm very satisfied. The water is still hot a | 2019/3/15 | http://s.club.jd. |
| 6  | TRUE      | FALSE | Jingdong                         | Midea | Oneself start work installation, bought n    | 2019/3/15 | http://s.club.jd. |
| 7  | TRUE      | FALSE | Jingdong                         | Midea | I've been using it for several rental units  | 2019/3/15 | http://s.club.jd. |
| 8  | TRUE      | FALSE | Jingdong                         | Midea | Not bad, is a little slow delivery, no call  | 2019/3/15 | http://s.club.jd. |
| 9  | TRUE      | FALSE | Jingdong                         | Midea | Things are very good, have a discount re-    | 2019/3/1  | http://s.club.jd. |
| 10 | TRUE      | FALSE | Jingdong                         | Midea | Cost-effective, next time will patronize!    | 2019/2/26 | http://s.club.jd. |
| 11 | TRUE      | FALSE | Jingdong                         | Midea | The order was placed the night before a      | 2019/2/24 | http://s.club.jd. |
| 12 | TRUE      | FALSE | Jingdong                         | Midea | Not so bad.                                  | 2019/2/24 | http://s.club.jd. |
| 13 | TRUE      | FALSE | Jingdong                         | Midea | A brand that should be trusted               | 2019/2/23 | http://s.club.jd. |
| 14 | TRUE      | FALSE | Jingdong                         | Midea | Cheap, shopping convenient and fast          | 2019/2/20 | http://s.club.jd. |
| 15 | TRUE      | FALSE | Jingdong                         | Midea | Good, good, good, good, good, good, go       | 2019/1/30 | http://s.club.jd. |
| 16 | TRUE      | FALSE | Jingdong                         | Midea | (Midea f40-15a1 40 liter) electric water     | 2019/1/30 | http://s.club.jd. |
| 17 | TRUE      | FALSE | Jingdong                         | Midea | Help colleagues to buy he said good, bu      | 2019/1/25 | http://s.club.jd. |
| 18 | TRUE      | FALSE | Jingdong                         | Midea | It's been a while. It works. No questions    | 2019/1/25 | http://s.club.jd. |
| 19 | TRUE      | FALSE | Jingdong                         | Midea | How come? I bought it the day before y       | 2019/1/20 | http://s.club.jd. |
| 20 | TRUE      | FALSE | Jingdong                         | Midea | Very easy to use, very convenient! The       | 2019/1/18 | http://s.club.jd. |
| 21 | TRUE      | FALSE | Jingdong                         | Midea | For the company, just convenient             | 2019/1/17 | http://s.club.jd. |
| 22 | TRUE      | FALSE | Jingdong                         | Midea | The water that 2 people bathe is ok, say     | 2019/1/13 | http://s.club.jd. |

Then we extracted the comment column of "brand", and part of the Python code for comment extraction is shown in listing 1.

| import pandas as pd                                    |
|--|
| inputfile = 'D:/data/huizong.csv'                      |
| outputfile = 'D:/data/meidi_jd.txt'                    |
| data = pd.read_csv(inputfile, encoding = 'utf-8')      |
| data = data[[u'Comments']][data[u'Brand'] == u'Midea'] |
| data.to_csv(outputfile, index = False, header = False) |
|  |

#### Listing 1 comment extraction code

B. Preprocessing the Comment Data

There are three parts of preprocessing the text comment data: text deduplication, mechanical corpus compression and short sentence deletion.

#### 1) Text deduplication

Text deduplication is removed duplicates from the text comment data. There are three reasons: (1) if the customer



does not make a comment out of time, the system will automatically make a comment for the customer, of course, most of the results of comments will be good evaluation. (2) There are many same or similar comments because many customers want to save time or avoid trouble, they may even directly copy the comments of others. (3) there will be some meaningless comments, such as "well, well, well", "XX brand water heater XX liter" and so on, this kind of comment is obviously only meaningful in the first comment. Although it's not completely repetitive, they also belong to meaningless comments.

Therefore, this paper adopts the method of comparative deletion, that is, pairwise comparison, the method of removing only the identical ones. The text-de-weighted portion of the Python code is shown in listing 2.

| import pandas as pd  |
|--|
| inputfile = 'D:/data/meidi_jd.txt'   |
| outputfile = 'D:/data/meidi_jd_process_1.txt'                              |
| data = pd.read_csv(inputfile, encoding = 'utf-8', header = None)           |
| 11 = len(data)   |
| data = pd.DataFrame(data[0].unique())                                      |
| 12 = len(data)   |
| data.to_csv(outputfile, index = False, header = False, encoding = 'utf-8') |
| print(u'deleted % comments.' %(11 - 12))                                   |
|  |

#### Listing 2 Original data deduplication

# 2) Mechanical Compressed Words

In fact, the corpus to be processed by mechanically compressed words is the part of corpus that has continuous and redundant repetition. From the perspective of general comment preference, the general production of meaningless continuous repetition will only be carried out at the beginning or end, such as: "why? Why? "And" really, really, really, really, really, really, really, really, really." Therefore, continuous repetition at the beginning and end needs to be mechanically compressed to remove words. In this paper, according to the judgment of placing continuous redundant repetition in the processing of word removal by mechanical compression and 7 compression rules. Compare the compressed statement and the original statement by intercepting part of the forward mechanical compression, as shown in Fig. 1.

| Can, CanCanCanCan                      |  |
|--|--|
| Can, Can                               |  |
| Easy to use Easy to use Easy to use!!! |  |
| Easy to use                            |  |
| Good, good, cheap                      |  |
| Good,cheap                             |  |

Fig.1. The compressed statement is compared to the original statement

#### 3) Deleting Short Eentences

In the comment data, comments with too few words are meaningless comments, so delete the short comment text data to remove meaningless comments, including:

(1) There are too short original comment texts, such as "very good".

(2) After the mechanical compression of the word processing short comment text, that is, the original for the existence of continuous repetition and meaningless long text, such as "good, good, good,

limit for reserved comments. Generally, 4 to 8 international characters are reasonable lower limit. This paper sets the lower limit as 7 international characters, that is, if the corpus obtained after the first two steps of preprocessing is less than or equal to 7 international characters, the corpus shall be deleted. After the above three steps pretreatment, can filter out a lot of junk information.

#### (3) Participle analysis of text comments

In this paper, Python's Chinese word segmentation package "Jieba" stammering word segmentation is adopted to conduct Chinese word segmentation on the commodity comment data in the document. "Stammering segmentation" provides segmentation, part of speech tagging, unregistered word recognition, support user dictionary and other functions. After relevant tests, the segmentation accuracy of this system is as high as 97%. In order to carry out word frequency statistics, part of speech tagging is removed in word segmentation.

# 4) Building Emotional Orientation Model and LDA Model for Thematic Analysis

## a) Construct the text emotional tendency model

Compared with artificial intelligence algorithms such as Bayes, neural network and SVM, text emotion judgment adopts natural language segmentation in NLP and relies on emotion dictionary for text emotion judgment. In this paper, using Jieba participle online reviews of the electricity suppliers to natural language shard, establish emotional dictionary for electric merchants electric fields, the introduction of NLP research field of n-gram after 2 yuan grammar phrases collocation, relying on the center words basic judgment, relying on the former influence emotional words effect direction and degree of emotion judgment, the model is shown in Fig. 2.

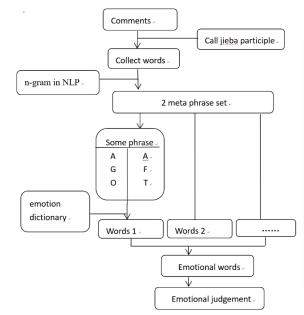


Fig.2. Model graph of judgment of textual emotional tendency



b) Implementation process of analysis based on LDA theme model

The LDA topic model, also known as the three-layer bayesian probability model, is a statistical model used to discover abstract topics in a series of documents in fields such as machine learning and natural language processing. The Bag of Words (BOW) model is used to treat each document as a word frequency vector, thus transforming textual information into numerical information that can be easily modeled. Then Gibbs sampling in markov chain monte carlo algorithm is used to estimate the parameters of the LDA model according to (1):

$$P(z_i = s \mid Z_{-i}, W) \propto (n_{s,-i} + \beta_i) / (\sum_{i=1}^{V} n_{s,-i} + \beta_i) \times (n_{s,-j} + \alpha_s)$$
(1)

In the LDA topic mining of commodity reviews in this paper, the feature words in the comments are observable variables in the model. The steps of its realization are as follows.

Step 1: In order to improve the accuracy of subject analysis in the reflection of hot concerns under different emotional tendencies, this paper, on the basis of the result of emotion classification, mined and analyzed the potential subjects under different emotional tendencies, so as to obtain the user's response to different aspects of the water heater under different emotional tendencies.

Step 2: The statistics of the comment in the corpus distribution, the theme of the positive and negative emotional tendency under the two kinds of emotional tendency, every subject the number of occurrences of from high to low rank, according to the analysis, choose the top several topics as hot focus in essays, and then according to the key words on the underlying theme of the probability distribution of comment get hot focus of the word.

Step 3: Use Gibbs sampling method to approximate the parameters of the LDA model. According to the above introduction of the model, there are 3 variables in the model that need to determine the best value, namely the prior parameters of dirichlet function and the number of topics. In this paper, the prior parameters and of dirichlet function are set as empirical values, respectively,. The number of topics is selected by using the confusion degree of evaluation standard commonly used in the statistical language model, i.e. =50.Therefore, the text is divided into "positive emotional results" and "negative emotional results", and LDA theme analysis is conducted to find out the advantages and disadvantages of goods. Use Python's Gensim library to complete the code for LDA, part of which is shown in listing 3

| neg_dict = corpora.Dictionary(neg[2])<br>neg_corpus = [neg_dict.doc2bow(i) for i in neg[2]]   |
|---|
| neg_lda = models.LdaModel(neg_corpus, num_topics = 3, id2word = neg_dict)   |
| for i in range(3):  |
| neg_lda.print_topic(i)  |
| <pre>pos_dict = corpora.Dictionary(pos[2]) pos_corpus = [pos_dict.doc2bow(i) for i in pos[2]] pos_lda = models.LdaModel(pos_corpus, num_topics = 3, id2word = pos_dict) for i in range(3):     neg_lda.print_topic(i)</pre> |



After LDA topic analysis, the comment text is grouped into 3 themes, and 10 most likely words and corresponding probabilities are generated under each theme. Table II shows the potential themes in the positive evaluation text of beauty, while table 3 shows the potential themes in the negative evaluation text of beauty.

 
 TABLE II.
 POSITIVE EVALUATION OF THE UNDERLYING THEME OF MIDEA

| Topic 1    | Topic 2   | Topic 3      |
|------------|-----------|--------------|
| very good  | good      | installation |
| delivery   | the       | the          |
| quick      | things    | master       |
| is         | not bad   | oneself      |
| good       | Jingdong  | Midea        |
| heating    | Midea     | the          |
| speed      | price     | yuan         |
| very quick | feel      | no           |
| service    | very good | after-sales  |
| very       | worth     | the door     |

TABLE III. NEGATIVE EVALUATION OF THE UNDERLYING THEME OF MIDEA

|              | 11110111     |          |
|--------------|--------------|----------|
| Topic 1      | Topic 2      | Topic 3  |
| installation | is           | the      |
| master       | not bad      | the      |
| Midea        | heating      | things   |
| yuan         | don't know   | no       |
| delivery     | however      | Jingdong |
| After-sales  | a little bit | oneself  |
| service      | ok           | or       |
| bad          | use          | but      |
| the door     | speed        | this     |
| good         | the          | ok       |

According to the extraction of feature words of three potential topics praised by Midea water heater, the high-frequency feature words in theme 1 are "good", "fast delivery", "heating", "fast", "service", "very", etc. Midea's water heater heats quickly; High frequency keywords in theme 2, that is, the hot focus is mainly price, things, worth, etc., mainly reflecting the beauty of the water heater is good, the price is appropriate and worth buying, etc. High-frequency keywords in theme 3, namely, the hot focus is mainly on after-sales service, master, door-to-door, installation, etc., mainly reflecting jd's after-sales service and master door-to-door installation, etc.

From the three potential topics of Midea's water heater poor evaluation, we can see that the high-frequency keywords in theme 1 are mainly installation, service, yuan, etc., that is, theme 1 mainly reflects the high installation fee of Midea's water heater and the poor after-sales service of the water heater. The high-frequency keywords in theme 2 are mainly emotional words such as "little", "little" and "ok", while theme 3 mainly reflects that the water heater of beauty may not meet its needs. Topic 3 in the high frequency keywords are mainly no, but, oneself, etc., theme 3 May mainly reflect the United States water heater installation and so on.

Comprehensive above to the theme and its high-frequency keywords can be seen, the advantages of the United States water heater has the following aspects: affordable, cost-effective, good-looking appearance, water heater practical,



easy to use, heating speed is fast, good service. Relative to character, the complaint point that the user is opposite beautiful water heater basically reflects the following a few respects: the charge that beautiful water heater installs is expensive reach after-sales service.

#### IV. CONCLUSIONS AND SUGGESTIONS

## A. Providing accurate marketing improvement programs

Through the above empirical analysis, it can be concluded that customers buy Midea water heater for the following reasons: Midea's big brand is trustworthy; Midea's water heater is affordable and cost-effective. Therefore, we propose the following precision marketing Suggestions for Midea brands: 1, on the basis of the advantage such as keeping water heater convenient to use, price affordable, undertake improving to water heater, the quality that elevates water heater from whole. 2. Improve the overall quality of installation personnel and customer service personnel, and improve the service quality. Only installation fee collection clear details, and open and transparent, reduce the installation process of arbitrary fees. Reduce installation cost and material cost moderately; highlight advantage in the competition of big brand in this.

#### B. Establishing new management evaluation methods

At present, there are many ways to evaluate e-commerce merchants, including from the perspective of commodities, logistics channels and customer experience. For example, Mei Hu (2015) explored the impact of logistics service quality on consumers through factor analysis and regression analysis [1]. Zhao Huifen (2015) constructed a consumer network social ecosphere of commodity evaluation system to explore the impact of commodities on consumers [2]. Victor Ho. (2017) established online evaluation indicators, including cross-class variation indicators, structural distribution indicators and internal variation indicators.[3] However, through text segmentation algorithm and LDA topic model, this paper can judge the emotional tendency of customers in online comments, and evaluate the effect of customers' recognition of merchant management services and the operation and management of e-commerce merchants.

# C. Monitoring the emotional tendency of customers

Traditional methods of monitoring customer emotional tendency are mainly obtained through consumer complaints, target market survey, etc., with high cost, long time and inaccurate data. Now using text participle, LDA theme model such as machine learning algorithms for the customer's online reviews text mining, the data can not only distinguish the customer's emotional tendency, also can undertake dynamic monitoring, such as can be with month, day, for the time dimension to grab comment on data analysis, grasp the emotional tendency of customer changes, precise management service and marketing strategy.

#### D. Grasping the emotional trend of the whole industry

In model based on text participle and LDA subject emotional analysis, application of electric business online reviews data, digging to the store for the dimensions, to accurately grasp the customer emotional tendency and degree of each electricity businesses, service satisfaction reflects the businessman, grasp the trend of the development of the service, so that the electric business platform and businesses to grasp the customer emotional level and trend of the industry. To sum up, based on the text participle and LDA theme online reviews data mining model accurately determine the customer's emotional tendency, do precise marketing provides powerful basis for businesses, is a kind of new electricity business appraisal through the camp, to dynamic state prosecutor enough measuring customer emotion change, make sure electrical business platforms and business must to grasp the whole emotional trend of the industry.

#### REFERENCE

- Mei Hu, Lin Lingxia, Ma Zicheng. Construction and Analysis of Key Indicators for Evaluation of E-Commerce Logistics Service Quality [J]. Logistics Technology, 2015(6).
- [2] Zhao Huifen. Study on Commodity Evaluation System in the Process of E-Commerce Marketing [D]. Jilin University, 2015(5).
- [3] Victorho. Achieving Service Recovery through Responding to Negative Online Reviews [J]. Discourse & Communication, 2017(11).