Data Science and Engineering (DSE) Record, Volume 3, issue 1.

Text sentiment Analysis from Commodity Evaluation Data: Case study of Intelligent Refrigerator Reviews

Han Yang[1] and Pruet Boonma[2]

[1] Master's Degree Program in Data Science, Chiang Mai University, China, Sichuan
[2] Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand

han_yang@cmu.ac.th

Abstract. This paper takes reviews of five different refrigerators in Jindong Mall as the research object, collects text review data on the e-commerce platform through a Python web crawler, classifies user sentiment after basic pre-processing operations such as de-weighting and cleaning, and analyzes user review text data through word separation processing, word frequency statistics, and display. The LDA theme analysis model is used to analyze the review data thematically, to obtain the valuable contents of the text review data through multifaceted analysis and make suggestions for the improvement of products.

Keywords: Text data Mining, LDA topic Model, Sentiment Analysis.

1 Introduction

With the improvement of people's living standard, the popularization of the concept of Internet of Things and the climax of the development of artificial intelligence, intelligent home appliances have come into people's daily life. At present, intelligent refrigerators have become one of the most competitive products in the market of China's intelligent home appliance industry. But from the perspective of the market, after 2017, the refrigerator market scale development is relatively slow, and in recent years can be said to have entered a platform period. From 2018 to 2021, the lowest penetration of smart products in the electric category, refrigerators only from having 8.7% [1]. At present, there is no unified industry standard for smart refrigerator industry, the technical route and strategy of each company is homogeneous [2], thus leading to excessive competition between domestic manufacturers of smart refrigerators, the price war breaks out from time to time. How to quickly tap the user focus has become a major manufacturer's top priority. Therefore, this paper focuses on sentiment analysis and theme mining through the review data of the top five refrigerators of five different brands in the industry, revealing the values contained in the reviews, comparing the advantages and disadvantages of the five refrigerators, and thus proposing optimization strategies for manufacturers.

2 Methodology

2.1 Data Collection

According to the smart refrigerator sales ranking released by China State Grid 2021[3], the top five smart refrigerators are Haier, Siemens, Rong sheng, Cassady, TCL. and Jingdong Mall has a large influence in the field of China's e-commerce platform. So this paper selects the top five brand sales of Jingdong self-owned products on Jingdong Mall as the research object. Using Python to write a crawler program to collect the review data of five brands of refrigerators from Jingdong Mall website. The user ID, comment time, comment likes, comment replies, comment stars, comments 4 fields, each brand 1000 comments, a total of 5000 comments, as shown in **Figure 1.**

Content ID	Content Time	the number of content like	Content reply	Content star	Content
曲***程	2020-11-19 11:27:46	21	50	5	外形外观：外观很好看，很大气，磨砂黑的，很高级制冷效果：很棒，主要是声音很小保鲜效果：
****空	2021-11-19 22:45:55	3	3	5	冰箱墨荷灰色的外观很好看，显得非常大气；536L的超大容量能够满足大家庭的存储需求且空间分
w***x	2021-07-08 12:38:52	16	10	5	外形外观：很好看，低调奢华，深灰颜色很喜欢。制冷效果：变频风冷无霜，还可以调节温度，性
四***5	2021-02-04 20:12:43	30	14	5	容声，大品牌，值得信赖！536L超大容量，浩瀚藏鲜，海纳珍馐百味。其中，345L冷藏区+191L冷
小***号	2021-12-16 18:55:25	1	1	5	买过三个冰箱都是容声的，质量信得过，之前买过四门的，这个双门的看起来很高端，黑色的高档
S***1	2021-11-03 11:43:25	1	1	5	两口之家 买了个大冰箱 开心外形外观：颜色样子都不错的制冷效果：凉的很快 冷冻室目前不怎
u***1	2021-12-07 09:19:50	0	1	5	容升536双开门冰箱质量真好，手工很好，压缩机超静音，容量空间也大，冰箱超薄，颜色高大时
佐***7	2021-11-25 09:17:15	1	1	5	外形外观：外形很高大上，棕色的很显档次制冷效果：制冷效果很棒，冷冻室放进去不多会就冻住
H***2	2021-11-16 23:50:02	2	1	5	超出预期，质量非常满意?，材质做工非常好，实用性非常强，赞?好产品、好品质、好服务，在这
半***疯	2021-10-22 11:03:45	3	1	5	外观：金色灰色外观真的很显高级，简洁优雅内在：空间很大，隔层、抽屉应该很方便放置物品。

Fig 1. reviews.

2.2 Data pre-processing

In this paper, we select the reviews and review stars in the crawled data as the object of analysis. And set the star rating less than or equal to 3 as negative comments and greater than three as positive comments. Then four pre-processing steps of text content removal, text de-duplication, mechanical compression and word removal and phrase deletion are performed on the collected comment data.

2.3 Text splitting

In this paper, use the jieba module in Python for word separation operations. The exact mode in jieba word separation can separate the utterance most precisely, and this model is suitable for text analysis; it is able to full mode and quickly select all the words in the utterance that can be converted into words. Part of the code is shown in **Figure 2.**

```
import pandas as pd
import re
import jieba.posseg as psg
import numpy as np
strinfo = re.compile('[0-9a-zA-Z]|京东|冰箱|')
content = content.apply(lambda x: strinfo.sub('', x))
worker = lambda s: [(x.word, x.flag) for x in psg.cut(s)]
seg_word = content.apply(worker)
n_word = seg_word.apply(lambda x: len(x))
stop_path = open("stoplist.txt", 'r',encoding='UTF-8')
stop = stop_path.readlines()
stop = [x.replace('\n', '') for x in stop]
word = list(set(word) - set(stop))
result = result[result['word'].isin(word)]
n_content = [[x+1]*y for x,y in zip(list(seg_word.index), list(n_word))]
```

Fig. 2. some of word separation code

2.4 Sentiment analysis based on sentiment dictionary

The sentiment tendency of a sentence or phrase is usually determined by the sentiment words in the sentence. A good sentiment dictionary should be able to better encompass the sentiment words in the study area and accurately determine the sentiment tendency of the relevant text, therefore, building a sentiment dictionary is the basis for studying the sentiment of a text. In this paper, a preliminary set of sentiment words for intelligent refrigerator reviews is built on the basis of the How Net [4] sentiment dictionary and the "Extended Version of Synonyms" compiled by HIT. In addition, considering the influence of negation words and network words on sentiment analysis, a negation dictionary was constructed by manually collecting negation words and setting the negation weight to -1. 10 network words indicating positive sentiment and 9 indicating derogatory sentiment were added to the sentiment dictionary, as shown in the following **Table 1 and Table 2**.

Table1. Example of negative dictionary

Negative words	Weight
no（不）、（没）、no（无）、not（非）、no（莫）	-1

Table2. Example of negative dictionary

Emotional word collections names	Emotion Words	weight
Positive Network Emotion Words	bullish（牛逼），stable（稳），proper（妥妥），in place（到位）	1
Negative Network Emotion Words	Spit（吐槽）、，set（套路），pit（坑爹）	-1

The code for the sentiment analysis section is shown in the **figure 3**.

```
pos_comment = pd.read_csv("正面评价词语（中文）.txt", header=None,sep="\n",
                          encoding = 'utf-8', engine='python')
neg_comment = pd.read_csv("负面评价词语（中文）.txt", header=None,sep="\n",
                          encoding = 'utf-8', engine='python')
pos_emotion = pd.read_csv("正面情感词语（中文）.txt", header=None,sep="\n",
                          encoding = 'utf-8', engine='python')
neg_emotion = pd.read_csv("负面情感词语（中文）.txt", header=None,sep="\n",
                          encoding = 'utf-8', engine='python')
data_posneg['amend_weight'] = data_posneg['weight']
nge(0, len(data_posneg))
only_inclination = data_posneg.dropna()
only_inclination.index = np.arange(0, len(only_inclination))
index = only_inclination['id']
result = emotional_value.merge(word,
                               left_on = 'index_content',
                               right_on = 'index_content',
                               how = 'left')

result = result[['index_content','content_type', 'a_type']].drop_duplicates()
```

Fig. 3. some of the sentiment analysis code

3 Result

3.1 LDA Model Topic Analysis

In fields such as machine learning and natural language processing, a topic model is a statistical model for discovering abstract topics in a set of documents. The LDA model, as one of the topic models, is an unsupervised generative topic probability model that can be used to find patterns in the use of words in a text and to link similar words together to find useful information in an unstructured set of text.

In this paper, first build a dictionary and corpus based on sentiment data obtained from lexical sentiment analysis, and then use LDA topic models to find topic words under different topic numbers, and each model takes out several topic words (e.g., the first 100) each and merges them into a set. The frequency vector between any two topics is generated, and the cosine similarity between the two vectors is calculated, with larger values indicating more similarity; the average cosine similarity of each topic number is calculated, and the optimal topic number is found. Then use python's Genism module for LDA topic modeling. Part of the code is shown in **Fig. 4**.

```
import pandas as pd
import numpy as np
import re
import itertools
import matplotlib.pyplot as plt
from gensim import corpora, models


def cos(vector1, vector2):
    normA = 0.0;
    normB = 0.0;
    for a,b in zip(vector1, vector2):
        dot_product += a*b
        normA += a**2
        normB += b**2
    if normA == 0.0 or normB==0.0:
        return(None)
    else:
        return(dot_product / ((normA*normB)**0.5))
```

Fig. 4. some of the LDA topic analysis code

4 Analysis of LDA model theme analysis results

4.1 Analysis results

Table 3. Haier positive evaluation potential themes

1. Effect	6. Satisfied
2. Appearance	7. Like
3.Capacity	8. Shape
4.Refrigeration	9. Generous
5.Reshness	10. Size

Table 4. Haier negative evaluation potential themes

Theme 1		Theme 2		Theme 3	
1.Appe-arance	6.Refrig-era-tion	1.High	6.Haier	1.Effec-tiven-ess	6.Re-friger-ation
2.Capa-city	7.Effect	2.Refrig-era-tion	7.Very	2.Cost-ef-fectiveness	7.Really
3.Fresh-ness	8.Open the door	3.Shape	8.Model	3.Face value	8.Design
4.Gen-erous	9.Feel	4.Nice	9.Functi-on	4.Special	9.Black
5.Space	10.High	5.Size	10.Put	5.Super	10.High

From Table 3, it can be seen that the main advantages of Haier refrigerators are good effect, large capacity. From Table 4 it can be seen that the main disadvantages of Haier refrigerators are cost performance is not high, refrigeration function has problems.

Table 5. Siemens Positive Evaluation Potential Themes

Theme1		Theme 2	
1.Effect	6.Good	1.Capacity	6.Brand
2.Cooling	7.Space	2.Very large	7.Siemens
3.Freshness	8.Shape	3.Noise	8.White
4.Appearance	9.Generous	4.Master	9.Rating
5.Satisfied	10.Size	5.like	10.Installation

Table 6. Siemens Negative Evaluation Potential Themes

Theme 1		Theme 2		Theme 3	
1.Appearance	6.High	1.Capacity	6.Noise	1.High	6.ugly
2.Cost-effec-tiveness	7.Model	2.Face value	7.Size	2.Effect	7.Master
3.Very large	8.Really	3.White	8.Quality	3.Space	8.Feeling
4.Good	9.Double cycle	4.Refrigeration	9.Put	4.Siemens	9.Put in
5.Sound	10.Home	5.Shape	10.Installation	5.Freshness	10.Generous

From Table 5 it can be seen that the main advantages of Siemens refrigerators are good product quality, strong refrigeration and freshness preservation. From Table 6 it can be seen that the main disadvantages of Siemens are cost performance is not high, noisy and the staff service is problematic.

Table 7. Rong sheng Positive evaluation of potential themes

1. Effect	6. Good
2.Appearance	7. Satisfied
3. like	8. Color
4. Capacity	9. Atmosphere
5. Quality	10. Quality

Table 8. Rong sheng Negative evaluation of potential themes

Theme 1	Theme 2	Theme 3	Theme 4
Freshness	Refrigeration	Effect	High
Space	Shape	Capacity	Appearance
Buy	Good	Cost-effective-ness	Very large
Power	Put	Things	Face value
Advanced	Sound	Rong sheng	Size
Hour Gray	Special	Color	Gray
Not good	Very big	Freshness preservation	Too
Speed	Freshness	home	Delivery
Open door	Effect	Appearance	Quality
Tall	Ugly	Refrigeration	Very slow

From Table 7 it can be seen that the main advantages of Rong Sheng refrigerators are good product quality, good appearance. From Table 8 it can be seen that the main disadvantages of Rong Sheng refrigerators are poor freshness, small space and slow logistics.

Table 9. Cassady Positive evaluation of potential themes

Theme 1		Theme 2	
1.Satisfied	6.Capacity	1.like	6.Soon
2.Appearance	7.Worth	2.Effect	7.Brand
3.Good	8.Service	3.Quality	8.Good looking
4.Generous	9.Logistics	4.Master	9.Special
5.Installation	10.Delivery	5.Color	10.Shape

Table10. Cassady Negative evaluation potential themes

Theme 1		Theme 2	
1.Refrigeration	6.Shape	1.High	6.Space
2.Face Value	7.Size	2.Capacity	7.Very large
3.Freshness	8.Too	3.Effect	8.Really
4.Very large	9.1High	4.Appearance	9.Special
5.Effect	10.Nice	5.Not good	10.Generous

From Table 9 it can be seen that the main advantages of Cassady refrigerators are good appearance of the product, the buyer's satisfaction with the installation service and buyer's trust in the brand of Cassady. From Table 10 it can be seen that the main disadvantages of Cassady refrigerators are insufficient capacity and poor refrigeration.

Table 11. TCL Positive evaluation of potential themes

1. satisfied	6. Refrigeration
2. Effect	7. Good
3. Appearance	8. Quickly
4. Capacity	9. Shape
5. like	10. Very big

Table 12. TCL Negative evaluation of potential themes

Theme 1		Theme 2		Theme 3	
1.Capac-ity	6.Deliv-ery	1.High	6.Very large	1.Space	6.Cost-effec-tiveness
2.Fresh-nes-s	7.home	2.Effect	7.stuff	2.Refrig-erati-on	7.enough
3.Size	8.Cross	3.Cost-ef-fectiveness	8.Face Value	3.Ap-pearance	8.Speed
4.Shape	9.Noise	4.Nice	9.Refriger-ation	4.Special	9.Sound
5.Put	10.Door	5.Price	10.Enoug-h	5.Gener-ous	10.Not to

From Table 11 it can be seen that the main advantages of TCL refrigerators are beautiful appearance and good effect. From Table 12 it can be seen that the main disadvantages of TCL refrigerators are low-cost performance bad appearance and poor refrigeration effect.

4.2 Suggestions

As a whole, it can be seen that the relative advantages of the five refrigerators are: 1. five refrigerators look good, can attract customers; 2. good quality.
Relative disadvantages: 1. performance. From the five-refrigerator negative theme model analysis can be seen in all have freshness, poor refrigeration function, refrigerator freezing when the noise problem. 2. capacity. There is the problem of refrigerator capacity is not enough. 3. For Rong sheng, Siemens, the two major brands, there may be consumers are not satisfied with the master service, the master service attitude is poor, and the logistics speed is too slow and other problems.
This puts forward five major refrigerator brand manufacturers to improve the following suggestions.
1. In maintaining the product cost performance based on the refrigerator for refrigeration, quiet and freshness function on the improvement, improve the cooling effect, reduce the noise when running, from the overall quality of the refrigerator, improve the competitiveness in the market.
2. Establish a complete after-sales service system. Regularly carry out effective rectification for the core issues of user concerns or complaints. In turn, improve user satisfaction and increase user experience, thus making the goods more competitive.
3. Improve logistics. Jingdong platform should improve the supply chain logistics services and improve the speed of logistics.

5 Conclusion

The paper firstly crawls the review data of five brands of smart refrigerators on the website of Jingdong Mall, saves them and then performs text de-duplication, mechanical compression words, phrase filtering, word separation processing, deactivation word deletion and other processing. Then the smart refrigerator dictionary is constructed, and the sentiment analysis is carried out by using the dictionary-based sentiment analysis method, and finally the LDA topic analysis is carried out to dig out the potential topic words in the review data and the related analysis of the topic words. After analysis, it is found that the lexicon-based sentiment analysis method and the mining of review words in this paper are better. The accuracy of sentiment analysis can be improved by constructing sentiment dictionaries of related domains, and the accuracy of topic analysis can be improved by performing sentiment analysis first and then topic model analysis. This method is also applicable to the sentiment analysis of other product reviews.

References

1. China Internet Information Center, https://www.cnnic.com.cn/, last accessed 2022/1/30.
2. Cheng Hui-Hua, Huang Fa-Liang, Pan Chuan Di. (2017). User satisfaction mining based on sentiment analysis of product reviews [J]. Journal of Fujian Normal University (Natural Science Edition), 2017, 17(1): 14-21.
3. National Bureau of Statistics of China, www.stats.gov.cn/, last accessed 2022/1/30.
4. Zhi wang Emotion Dictionary, https://c,dian.cnki.net/, last accessed 2022/2/1.