Volume: 12 Issue: 03 Pages: 4606-4612(2020) ISSN: 0975-0290

Research on Online Review Based on LDA Subject Model

Guo Tao
School of Computer and Communication, Hunan Institute of Engineering Xiangtan 411104, China
Email: 2206159759@qq.com
Wu Shi-qi
Email:1305056857@qq.com
Shi Yi-Cheng
Email:448720956@qq.com
Ma Qian-Qian
Email:459803949@qq.com
Tang Zhi-hang
Email: zhtang@hnie.edu.cn

ABSTRACT

The text topic analysis is the core element of the comprehensive review of clothing products, which can directly understand the views and consumption trends of consumer groups, taking a brand clothing store in JD.com as the research object, by using Python crawler and HANLP natural language processing technology, seven of the top-selling fashion reviews were classified and analyzed. Word frequency statistics, TF-IDF and other methods were used to quantify the text, this paper uses the visualization techniques such as word cloud graph contrast, pyLDAvis dynamic model and Sankey graph to display customers' attention points and real shopping needs from various angles. The experimental results show that the visual results of online review research based on the theme model of Lda can clearly show the advantages and disadvantages of customer-centered evaluation and clothing, and provide important reference for merchants to improve decision-making and optimize service.

Keywords**: Clothing Review; Natural Language Processing; Topic Mining; Visualization**

Date of Submission: Nov 12, 2020 Date of Acceptance: Nov 23, 2020

I. INTRODUCTION

In recent years, online shopping marketing methods have emerged one after another, from winning with quality to being the king of low prices, or providing good service to attract customers, etc. but how to capture the real demand of consumers is the key that every business is looking for. One of the main factors affecting the purchase of mall products by Internet users is the evaluation of its products by other Internet users. In particular, to purchase products with relatively high prices, netizens need information channels and platforms that can truly understand the actual situation of the products at this time. Only when consumers communicate with consumers face to face will consumer vigilance be lowered to the minimum, even direct unconscious consumption into active consumption. Therefore, strengthening the construction of information visualization is a new direction of e commerce.

In the current research on the visual analysis of online shopping reviews, Chen Huan [1] and others combined LDA with Self-Attention's short-text sentiment classification method to cluster the same topic in Galway vector space, using Self-Attention for dynamic weight assignment and classification, Liang Jiye [2] and others can achieve better short text clustering results by using the distributed representation of BTPV model to represent the more common distributed Vector quantization models

Volume: 12 Issue: 03 Pages: 4606-4612(2020) ISSN: 0975-0290

word2vec and Paragraph Vector, thus, the advantages of the model for short-text analysis are shown. Wu Fan and others can use the deep learning method to effectively improve the performance of emotion analysis and online comment quality detection based on the joint model of user and product expression Guoxian DA, Amjad Osmani, and others have introduced a new PADSLDA model for affective analysis of semantic relationships between words in text, which is based on Gaussian LDA's research on topic mining for online reviews.

This paper starts with the shopping review text, uses the word segmentation tool and LDA algorithm to model the consumer reviews into a topic model, and presents the results using visualization technology, then, we can find out the problems according to the result of the analysis, and provide important reference for the business to improve the decision-making and optimize the service.

II. WORK IN ADVANCE

2.1 Overall thinking

The whole experimental idea of this paper is shown in figure 1.

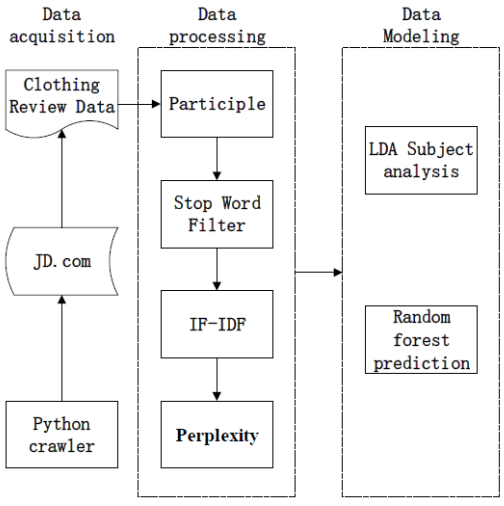

Figure 1. Flow chart of the experiment

The detailed research steps are as follows:

(1) Using python to write a crawler program to crawl the brand clothing on jd.com. Under the clothing of short-sleeved, shirt, hoodie, trousers, shorts, jacket, cotton-padded clothes, comments crawling under seven clothing comments, data including comments text, comments score, comments rate.

(2) The crawling data is cleaned and reprocessed, the HANLP segmentation module is called by Python, the stop-word list is loaded, the TF-IDF is quantified, and the confusion degree is calculated to get the best number of topics.

(3) Using Gensim module in Python to build the interface of LDA model, by debugging the optimal number of topics, constructing word bag to determine the theme and carry on the theme analysis, drawing the Sankey flow chart to visualize the experimental results.

2.2 Data Crawling

Before the experiment begins, the comment text [6] needs to be crawled. Go to the JD.com shopping mall, select the comment page and crawl the comments under short-sleeved, shirt, hoodie, trousers, shorts, jacket and cotton-padded clothes, totaling 23,000, the data that does not accord with the requirement basically is to have incomplete data, wrong data, duplicate data 3 big kinds. There are 21,263 valid data items for the spaces in the comments

Table 1. Crawling summary of comment information

Costume	Number of evaluations	Good ratings	Mean Score	Valid data
Short sleeves	4101	97%	4.8	3894
Shirt	2654	98%	4.9	2391
Hoodie	3321	97%	4.9	3056
Long Pants	2576	96%	4.8	2275
Shorts	3689	98%	4.7	3446
Jacket	1268	98%	4.6	1108
Cotton-padded clothes	5391	97%	4.9	5093
Total	23000			21263

Volume: 12 Issue: 03 Pages: 4606-4612(2020) ISSN: 0975-0290

Ⅲ. DATA PROCESSING

3.1 Stop words

The comments will be organized by the use of the segment module Hanlp segmentation module, note that the segmentation results will appear many meaningless words, so also to load the Stop Word List for the first time to remove the meaningless words, this can reduce the impact of follow-up experiments, improve the accuracy. This is a custom stop list, available on. CN. On the basis of txt, through the addition of stop words to change the nonsense words screening.

The effect of nonsense words on the whole can be seen clearly in the comparison between the non-stop words in figure 2 and the Stop Words in figure 2. After cutting out the stop words, the general words of good and good are removed, and the small but important words, such as comfort, fit, fabric, and so on, are retained.

Figure 2. Non Eliminate stop words

Figure 3. Eliminate stop words

3.2 TF-IDF Algorithm

The use of stop words can help remove most of the nonsense words, and the use of word clouds to show contrast can make a significant difference, but for a significant but infrequent keyword [9] , the word cloud is not well displayed, or even overwritten, so TF-IDF, a statistics-based Term Frequency-in-Frequency algorithm, TF is Term, and IDF is the Inverse Document Frequency. To calculate the rarity of a word, the importance of a word is not only proportional to its frequency in the document, but inversely proportional to the document containing it. The more documents that contain this term, the broader it is and the less distinctive it is. In actual use, TF is calculated by a common expression (1.1) , where nij denotes the frequency of the word I in document j, but only frequency, and a more straightforward expression is TF (word) = (number of times word appears in document)/(total number of words in document)

$$\mathrm{TF} = \frac{n_{ij}}{\sum_k n_{kj}} \tag{1.1}$$

The commonly used formula for IDF is (1.2) , | D | is the total number of documents in the document set, | Di | is the number of documents in which the word I appears in the document set. The denominator plus 1 is adopted Pierre-Simon Laplace smoothing to avoid the situation that some new words do not appear in the Corpus and result in zero denominator, which increases the robustness of the algorithm.

$$\mathrm{IDF} = \log \frac{|D|}{|\{1+D_j\}|} \tag{1.2}$$

TF-IDF Algorithm is the integrated use of TF Algorithm and IDF Algorithm, can be by two combinations, multiplication or division, after a lot of theoretical derivation and experimental research, the discovery formula (5.3) of the more effective way of calculation.

$$\begin{aligned} &\mathrm{TF} \\ &-\mathrm{IDF(t,d)} \\ &= \frac{TF(t,d)}{DF(t)} \\ &= TF(t,d) \\ &* IDF \end{aligned} \tag{1.3}$$

```
TF-IDF Weight Matrix
  (rows, line)   The importance of words
  (0, 661)   0.14587513542752567
  (0, 632)   0.2004254014853884
  (0, 154)   0.17877882966205014
  (0, 558)   0.2269003946021725
  (0, 82)    0.07799454404108228
```

Figure 4. TF-IDF weight Matrix

Volume: 12 Issue: 03 Pages: 4606-4612(2020) ISSN: 0975-0290

3.3 perplexity

In information theory, perplexity is a measure of how well a probability distribution or model predicts a sample. It can also be used to compare two probability distributions or probability models. The language model that gives higher probability values to the sentences in the test set is better. When the language model is trained, the sentences in the test set are all normal sentences. Then the trained model is the higher the probability on the test set, the better, the greater the sentence probability, the better the language MODEL, and the smaller the puzzle.

$$= \sqrt[N]{\frac{1}{P(w1w2 \cdots w_N)}} \quad (2.1)$$

From the analysis of the puzzle training model in figure 5, the model generation ability is best when the number of topics is set to three, more than three, because the number of topics is too many, the puzzle will rise to infinity and can not be analyzed.

$$\mathbf{PP(W) = P(w1w2 \cdots w_N)^{-\frac{1}{N}}}$$

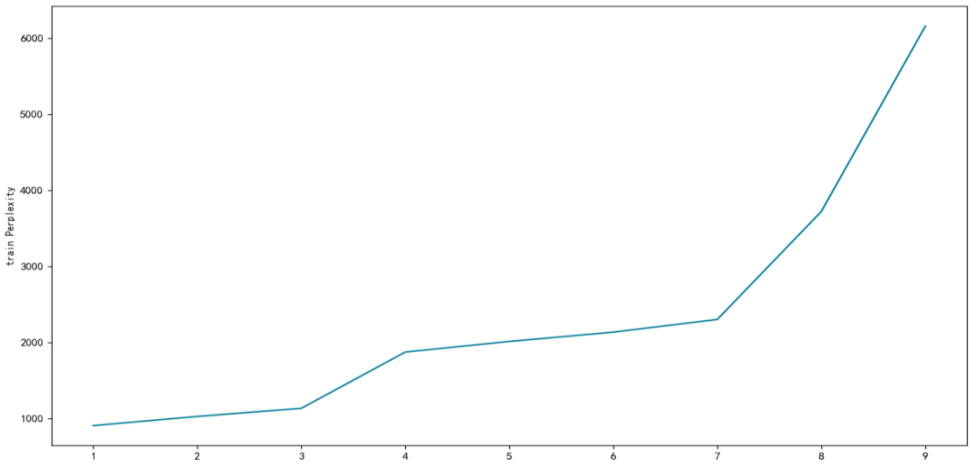

Figure 5. Perplexity calculation

After determining the number of themes, the theme can be determined by the number of Thesaurus [12] , and the visual model of LDA can be established. As you can see from the word bag in figure 6, Theme 0 can be defined as quality, theme 1 as style, and theme 2 as price.

Topic #0:
质量 物流 满意 很快 京东 衣服 非常 快递 不错 喜欢 东西 下次 好评 购买 实惠 速度 收到 还会 购物

Topic #1:
舒服 面料 大小 穿着 不错 舒适 尺码 合适 衣服 非常 起来 合身 好看 材质 版型 纯棉 是否 质感 透气

Topic #2:
物美价廉 适合 做工 信赖 款式 超级 码数 时尚 品牌 价钱 a21 衣服 觉得 完美 现在 以纯 很白 料子

Figure 6. Thesaurus bag

Ⅳ. DATA BUILDING

4.1 LDA theme model

LDA [13](Latent dirichlet allocation) is a three-layer Bayesian model proposed by David Blei et AL in 2003, which contains three layers of word, topic and document structure. The goal of unsupervised learning is to discover the hidden semantic dimension from the text by unsupervised learning.

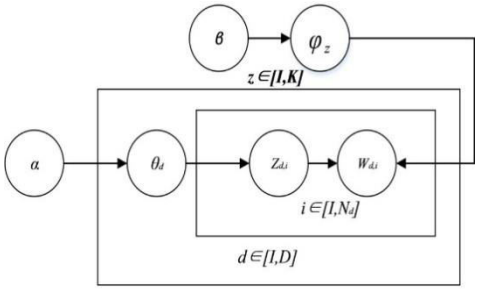

Figure 7. LDA MODEL generation process

By analyzing the contents of the word bag and defining the comment topic as quality, price and style, we can see that the proportion of comments in each document is different, and there are relatively more comments on price, quality and style, from document 1,3 is to pay

Volume: 12 Issue: 03 Pages: 4606-4612(2020) ISSN: 0975-0290

attention to clothing style; document 2,4 is to pay attention to clothing price; document 5,6 is to pay attention to clothing quality.

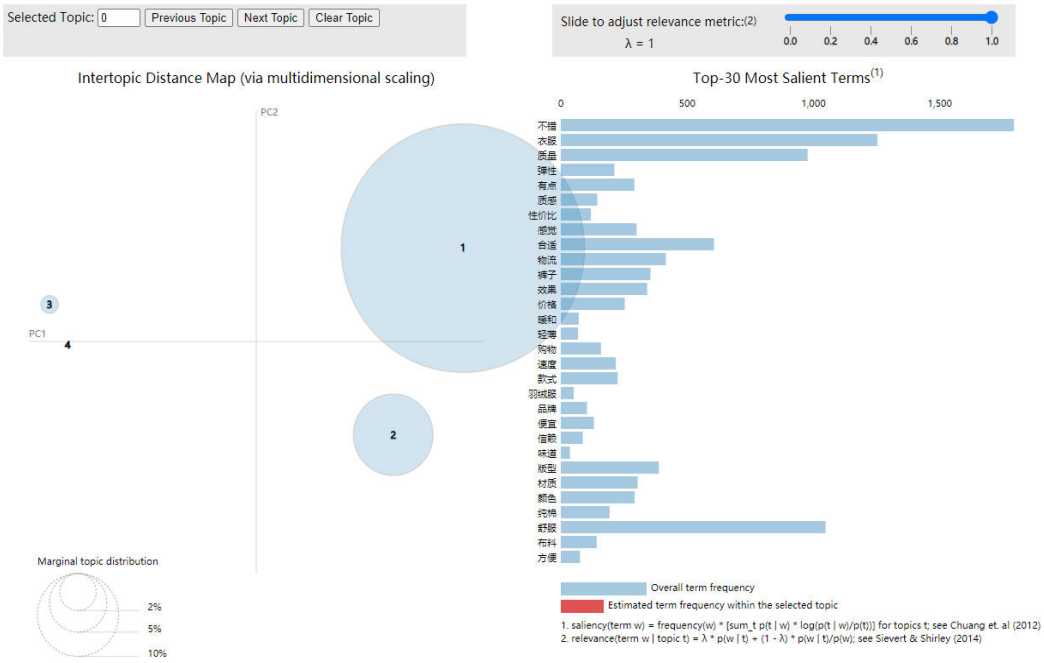

Figure 8. Distribution of clothing themes

4.2 variable parameter visualization

When all the text is gathered together and visualized using the pyLADvis model interface in Gemsim, the result in figure 9 shows that the comment text is mainly clustered into three modules. As shown in figure 9, there are four themes, but because the fourth one is too small and meaningless, we choose to ignore it, leaving three meaningful circles that do not overlap. This shows that the clustering of themes works well. Hover the mouse over topic 1. The column list on the right shows the keywords of the topic, the dark column shows the frequency of the keyword in topic 3, and the light column shows the frequency of the keyword. By adjusting the size of λ, the filtering condition and distribution difference of each subject word can be adjusted.

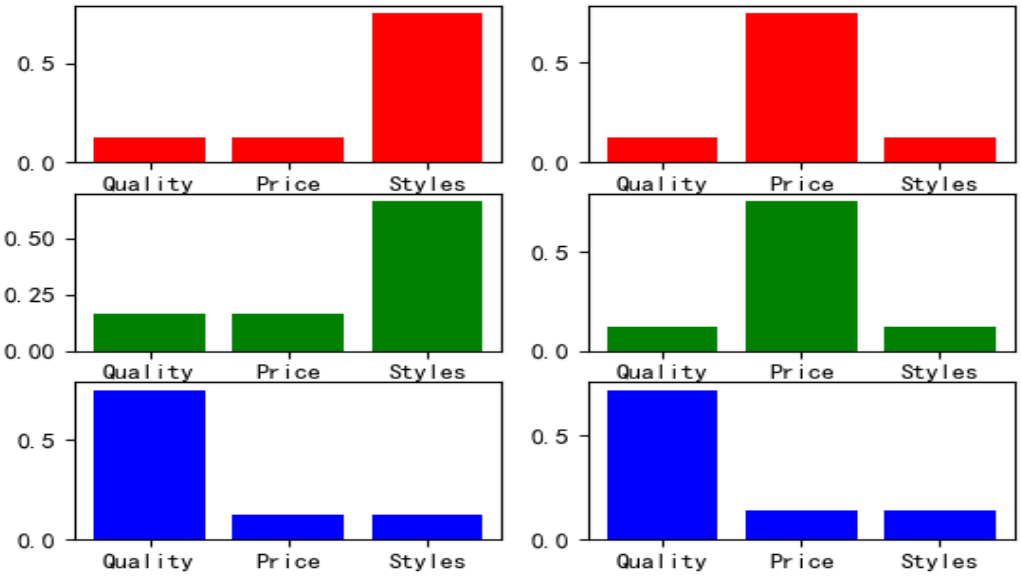

Figure 9. Clustering visualization of garment comment topics

Volume: 12 Issue: 03 Pages: 4606-4612(2020) ISSN: 0975-0290

4.3 Sankey circulations

According to figure 10, it can be seen that the width and flow direction of the branches respectively present the characteristics and proportion of different clothing attributes. As shown in the picture, cotton-padded clothes sell the most in the overall clothing category, while pants sell the least, while shorts, shirts, hoodies, short sleeves and jackets are relatively average. The style of cotton-padded clothes and shorts has a relatively large number of reviews, with cotton-padded clothes accounting for a considerable proportion, about 50% , while the price is the most evaluated jacket, general evaluation is the price and quality point of view relatively more; poor evaluation reflected in the price and style, each about 50% .

The overall rating of the brand is 50% good, 30% fair and 20% bad, which is more intuitive and true than the 99% positive rating on the website, but it still shows that the brand has high support rate among users, and cotton-padded clothes account for an important proportion in all aspects, this is the main clothing category of the brand. In addition, the overall quality ratio is less than the style and price, cost-effective slightly higher, is a light luxury brand. Merchants can mainly start from the price and style, slightly lower prices and design new clothing, in order to attract traffic.

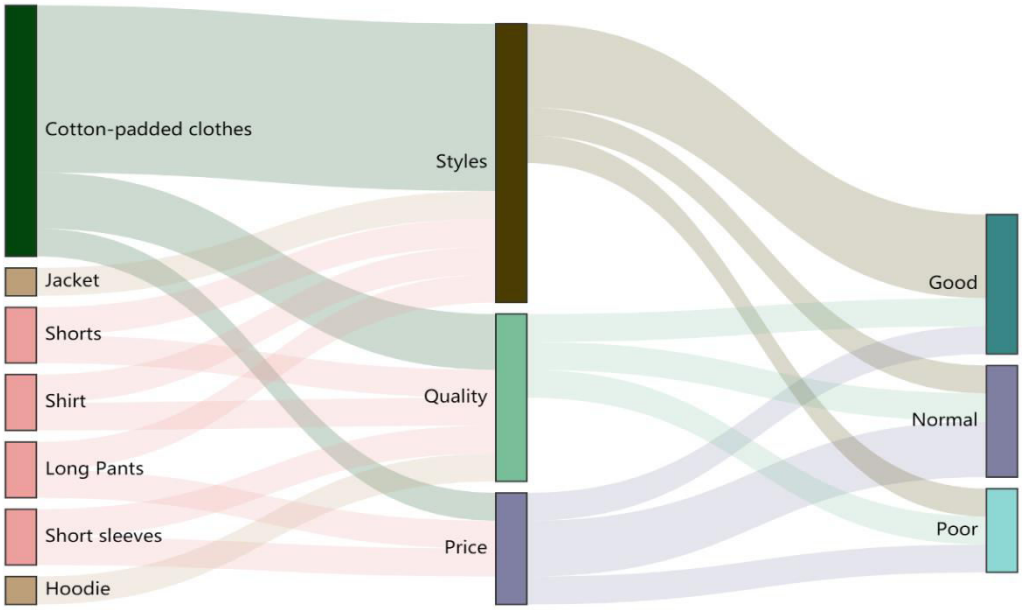

Figure 10. Relationship between clothing and rating

Ⅴ. CONCLUSION

In this paper, a natural language processing (NLP) [17] method is used to construct a garment review text and establish a subject analysis model of LDA, which objectively evaluates the quality, format and price of garment. In the experiment[18], it is necessary to save a lot of time in the process of crawling, filtering, importing python project and topic production, and the filtering of comment text is also very important, you can wash it a few more times. According to the consideration of the experimental situation, we can modify the stop-word list to clean the data many times to achieve better effect. The experimental data has the shortcoming of too little quantity, has not achieved the horizontal contrast, this is a place which needs to strengthen, the follow-up meeting increases the workload in this link.

ACKNOWLEDGEMENTS

Project supported by Provincial Natural Science Foundation of Hunan（2018JJ4047）

REFERENCES

[1]. Chen Huan, Huang Bo, Zhu Yimin, etc. . Short text sentiment classification method combining LDA and Self-Attention [j] . Computer engineering and applications, 2020,56(18) : 165-1709(in Chinese)

Volume: 12 Issue: 03 Pages: 4606-4612(2020) ISSN: 0975-0290

[2]. Liang Jiye, Qiao Jie, Cao Fuyuan, etc. . Distributed representation model for short text analysis. Computer Research and development, 2018,55(8) : 1631-1640(in Chinese)

[3]. Wu Fan, Wang Zhongqing, Zhou Xiabing, et Al. . Joint Model for sentiment analysis and review quality detection based on user and product representations [j] . Journal of Software Engineering, 2020,31(8) : 2492-2507

[4]. Amjad Osmani, Jamshid Bagherzadeh Mohasefi, Farhad Soleimanian Gharehchopogh. Enriched Latent Dirichlet Allocation for Sentiment Analysis. 2020, 37(4):n/a-n/a.

[5]. Guxian Da, Narissa, Gao Huan, etc. . Gaussian LDA based topic mining for online reviews [j] . Journal of Information Science, 2020,39(6) : 630-639(in Chinese)

[6]. Li Lin, Liu Jinxing, Meng Xiangfu, etc. . Product recommendation model based on fusion scoring Matrix and review text. Journal of Computer Science, 2018,41(7) : 1559-1573(in Chinese)

[7]. Feng Xingjie and Zeng Yunze. In-depth recommendation model based on scoring Matrix and review text. Journal of Computer Science, 2020,43(5) : 884-900(in Chinese)

[8]. Huang Jiajia, Li Peng Wei, Peng Min, etc. . Research on topic model based on deep learning. Journal of Computer Science, 2020,43(5) : 827-855. (in Chinese)

[9]. Zhang Fei, Zhang Libo, Luo Tiejian, etc. . Feature-based collaborative clustering model [j] . Computer Research and development, 2018,55(7) : 1508-1524(in Chinese)

[10]. Chen Jiaying, Yu Jiong, Yang Xingyao. A recommendation Algorithm for feature extraction based on semantic analysis. Computer Research and development, 2020,57(3) : 562-575(in Chinese)

[11]. Wang Jianxin, Prince Ya, Tian Xuan. A survey of text detection and recognition in natural scenes based on deep learning. Journal of Software Engineering, 2020,31(5) : 1465-1496(in Chinese)

[12]. Zhao Chuanjun, Wang Suge, Li Deyu. Progress in cross-domain text sentiment classification [j] . Journal of Software Engineering, 2020,31(6) : 1723-1746(in Chinese)

[13]. Veena Gangadharan, Deepa Gupta. Recognizing Named Entities in Agriculture Documents using LDA based Topic Modelling Techniques. 2020, 171:1337-1345.

[14]. Tiao-juan Han, Jian-feng Lu, Li Tao. Evaluation Research and Application of Cloud Service Provider Based on Real-Time Data. //DESTECH press, USA. 20202nd International Conference ON Advanced Control, AUTOMATION AND ARTIFICIAL INTELLIGENCE (ACAAI 2020,2ND INTERNATIONAL CONFERENCE ON ADVANCED CONTROL, AUTOMATION AND ARTIFICIAL INTELLIGENCE) proceedings. 2020:156-162.

[15]. Jiang Feng, Chu Xiaomin, Xu Sheng, et al. . Macro-text primary-secondary relation recognition method based on topic similarity [C] . //Chinese Information Society. Proceedings of the 16th National Conference on Computational Linguistics and the 5th International Symposium on Natural Language Processing based on natural labeled big data. 2017:1-10. (in Chinese)

[16]. Zhou Bei. Research on Data Mining Algorithm Based on Micro-blog of Multi-view Clustering Model. Francis. 20185th INTERNATIONAL CONFERENCE ON ELECTRICAL & Electronics Engineering and Computer Science (ICEEECS 2018) proceedings. 2018:220-226.

[17]. Sanjan S Malagi, Rachana Radhakrishnan, Monisha R, Keerthana S, Dr D V Ashoka. Content Modelling Intelligence System Based On Automatic Text Summarization. Int. J. Advanced Networking and Applications, 2020,11(6):4458-4467

[18]. Samah Osama M, Kamel, SanaaAbouElhamayed. Multi-Tenant Endorsement using Linguistic Model for Cloud Computing. Int. J. Advanced Networking and Applications,2020.11 (6): 4486-4493 (2020)