Semantic Analysis of Online Initial and Additional Textual Reviews of Pharmaceutical Products

Yumei Luo[1], Jin Zhang[1], Changlin Cao[2], YaWen Zhong[2] and Qiongwei Ye[3]
[1]*Department of Management Science, Yunnan University, Kunming, China*
[2]*Guangzhou Institute for Modern Industrial Development in Greater Bay Area, Guangzhou, China*
[3]*Business School, Yunnan University of Finance and Economics, Kunming, China*

Keywords: Pharmaceutical Products, Initial Reviews, Additional Reviews, Semantic Analysis.

Abstract: In this paper, we selected the textual reviews of drugs on Ali Health Pharmacy as sample data. Firstly, the text was pre-processed, including data cleaning, word separation, lexical annotation and removal of deactivated words, in order to lay the data foundation for the subsequent analysis process. Secondly, we conducted semantic analysis on the processed review texts, including word frequency analysis, LDA topic clustering analysis and word frequency co-occurrence analysis. The results showed that the overall distribution of semantic features in the text content of initial and additional comments was similar, but there were differences in the attention of consumers to different topics in initial and additional comments.

1 INTRODUCTION

More and more consumers making their purchase decisions based on online reviews. Additional reviews as a new form of online reviews, contained the product information that is somewhat different from consumers' initial reviews in both richness and objective veracity (Wang et al. 2015), it is necessary to conduct a comparative analysis of initial and additional reviews. Text mining of the review content of this kind of online reviews with no fixed form can help to better understand the large amount of information contained in the review text. With the development of the economy, the popularization of medical knowledge and the impact of the new pneumonia epidemic, consumers are increasingly inclined to buy common and safe drugs on the Internet. The pharmaceutical products involved in the research sample of this paper are all over-the-counter drugs.

The mining and research on online reviews of pharmaceutical products, as a special product, need yet to be strengthened (Terra and Clarke,2003). The safety of pharmaceutical products is directly related to the safety of consumers' lives, and consumers will be more cautious in making purchase decisions, and online reviews of pharmaceutical products will play a more important role in consumers' purchase decisions. Therefore, in this paper, we use text mining analysis to study the pharmaceutical products' semantic aspects of online initial and additional reviews. On the basis of word frequency analysis, this paper further conducts cluster analysis to analyse the topics that consumers pay attention to medical products. In order to better grasp the semantics of short texts, in addition to using LDA theme analysis, this paper also conducts word frequency co-occurrence analysis to better highlight the relevance between words.

2 LITERATURE REVIEW

2.1 Comparative Study of Initial and Additional Reviews

Many researchers pay attention to text reviews, Zhou and Li (2017) and Wang and Zhou (2016) used questionnaires to study the impact of consistency and ambiguity between initial and additional comments on consumer information adoption. Sun and Li (2017) and Li and Chen (2016) explored the effect of different additional forms of reviews and word-of-mouth emotional direction on purchase intention through a questionnaire study of college students; Hu and Ning (2017) investigated the usefulness of online initial and follow-up reviews based on time interval and product type by subjects completing

Luo, Y., Zhang, J., Cao, C., Zhong, Y. and Ye, Q.
Semantic Analysis of Online Initial and Additional Textual Reviews of Pharmaceutical Products.
DOI: 10.5220/0012044000003612
In *Proceedings of the 3rd International Symposium on Automation, Information and Computing (ISAIC 2022)*, pages 715-722
ISBN: 978-989-758-622-4; ISSN: 2975-9463

questionnaire scales. Zhang et al. (2020) crawled the cell phone reviews on JingDong Mall as a research sample and studied the association characteristics of the time series of online users' review chasing behaviour. Shi et al. (2016) selected review data on Mall for a comparative study of online initial reviews and online additional reviews in terms of product type, product price, number of reviews, review length, review interval, and review sentiment intensity.

Therefore, more and more scholars have begun to pay attention to the comparative analysis of the initial reviews and additional reviews, but most of them have used questionnaires to investigate the impact of questionnaires on consumer satisfaction, purchase intention, and the usefulness of reviews, etc., and to a large extent, a large amount of online data has been ignored.

2.2 Text Analytics Research

For the mining analysis of text content, Blei and Jordan (2003) proposed the thematic mining model LDA (latent Dirichlet allocation) model in 2003. Jiang et al. (2013) proposed an associative LDA model for opinion mining problem domains and applied it to user online reviews. Zhang and Qiao (2020) established a text mining model based on Bayesian correlation principles, completed the mining of user comments, and calculated the corresponding sentiment values. Liu and Jian (2018) completed the data mining of Pecan fruit user reviews on the Tmall website by establishing the user's emotional index. Wen and Ding (2010) completed the judgment of the user's product review sentiment and calculated the related emotional intensity. Chatterjee et al. (2020) use text mining techniques, machine learning, and econometric techniques to discover those sentiments that are more important in reflecting and predicting customer satisfaction. Zhang et al. (2020) used analytic methods such as word frequency co-occurrence analysis for text mining and analysis of online reviews. Yang et al. (2020) combined a sentiment dictionary in the field of cigarettes to analyze the word frequency of their high-frequency sentiment words in order to compare two different brands of cigarettes. Qiu and Sun (2021) used word frequency analysis methods to find popular majors in the three sample colleges, using ERNIE Pre-trained models make further sentiment classifications of comment data.

At present, domestic and foreign online review research involves many business fields such as movies, finance, hotels, social networks, etc., and the research content involved in network review mining is also very extensive, scholars mainly have more research on the number, length and polarity of online reviews, while the comparative analysis of the initial evaluation and the lack of online user reviews in the field of medical e-commerce drugs are relatively lacking. Therefore, this paper takes the comment content of related drugs on Alibaba Health Pharmacy as a sample to conduct semantic analysis of the initial post-evaluation, which not only enriches the theoretical research of pharmaceutical e-commerce and the comparative analysis of the initial and additional comments, but also enriches the application scope of text mining and analysis technology.

3 RESEARCH MODELS AND DATA PROCESSING

3.1 Research Model

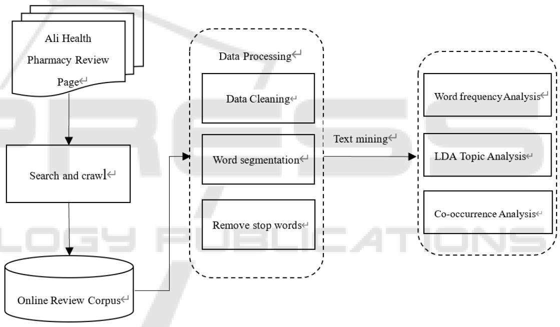


Figure 1: Research model.

According to the research model in Figure1, it can be seen that first, how to pre-process the text, including data cleaning, word segmentation, lexical annotation, removal of stop words, etc., to lay a data foundation for the relevant analysis process of the following text; Secondly, the semantic analysis of the processed comment content includes: word frequency analysis, LDA topic cluster analysis, and word frequency co-occurrence analysis.

3.2 Data Sources and Pre-Processing

According to the data, more than 40% of Chinese consumers will choose to buy drugs on Alibaba Health, and the top 5 usage rates of China's pharmaceutical e-commerce platforms in the first half of 2021 are shown in Figure2.

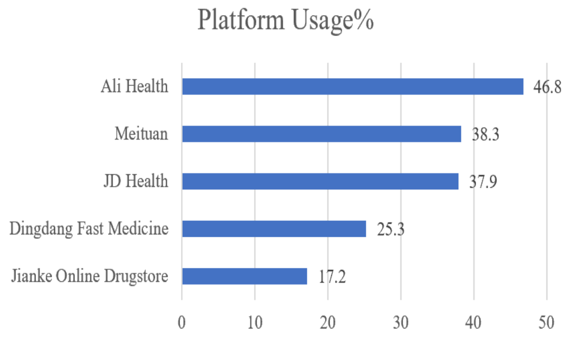


Figure 2: Top 5 in the first half of 2021 for the use of China's pharmaceutical e-commerce platforms.

Therefore, this article chooses Ali Health as a research platform, this article mainly through Python crawler technology to crawl the online user review information data of 320 drugs in the pharmaceutical e-commerce plat-form Ali Health Pharmacy, and there are 242 kinds of products left after removing duplicate products. After screening and eliminating meaningless comments such as automatic replies, a total of 1249 initial reviews and 3667 additional reviews were left, of which the existence of posthumous reviews was used as the filter, and a large number of comment data that were only posthumously commented on were deleted.

The data is then pre-processed, including data cleaning, word segmentation, part-of-speech annotation, and removal of manual annotations of stop words, in order to lay the data foundation for subsequent analysis.

3.2.1 Participle and Lexical Annotation

This article uses the jieba package in the Python language to segment each comment data, and the word segmentation supports three modes, namely precision mode, full mode, and search engine mode. Among them, the precision mode tries to make the most accurate segmentation of the sentence, which is suitable for text analysis. Because this article requires text analysis of the review data of pharmaceutical products, the precise mode in the stutter participle is selected to classify the comment text with participles and part-of-speech labels.

3.2.2 Remove Stop-Words

In natural language processing, in order to improve the efficiency of storage utilization and the efficiency of computational analysis, certain words or words are filtered out as needed before or after text analysis, and these filtered out are stop words. This article collates the commonly used Chinese stop words and adds some stop words according to the corpus of comment data to be analysed, a total of 786, some of which are stopped words as shown in Table 1 below: After removing the stop words and word segmentation from the comment data, provide corpus for the following analysis.

Table 1: Stop Words (partial).

1	①	a few	up	say	think
2	②	some	down	buy	bottle
3	③	all	under	eat	one time
4	”	One	still	drink	few days
5	、	an	still	afterwards	any

4 SEMANTIC ANALYSIS

4.1 Word Frequency Analysis

In this paper, after using Python's Jieba package to segment the comment data, the frequency of the vocabulary is counted, the high-frequency knowledge meta words in the initial comment and the additional comment are extracted, the synonyms are merged, and the Wordcloud2 package is used to map the vocabulary cloud.

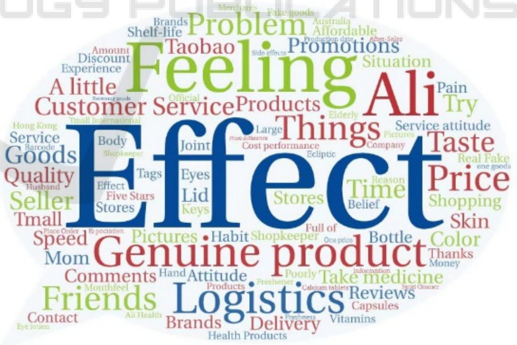


Figure 3: Word cloud map for initial reviews.

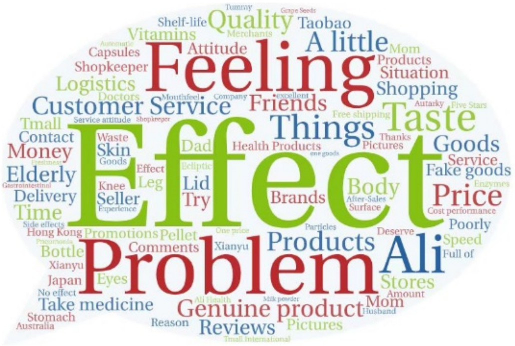


Figure 4: Word cloud map for additional reviews.

From Figure 3 and Figure 4 above, it can be found that the online initial reviews and additional reviews of drugs on the Alibaba Health platform, the overall concern of consumers includes product features such as "effect", "price" and "quality" of drugs, and the attention to service features such as "customer service" and "logistics" of drugs is also high. At the same time, it can also be seen that consumers will compare and analyse similar products on "Ali Health" and "Tmall International". In this paper, the top 10 high-frequency knowledge meta-vocabularies in the initial reviews and the additional review are selected to make word frequency statistics table, as shown in Table 2 and Table 3 below:

Table 2: Word frequency statistics for initial comments.

Ranking	Entry	Frequency	Ranking	Entry	Frequency
1	Effect	247	6	Problem	68
2	Genuine product	84	7	Price	45
3	Products	83	8	Promotions	44
4	Logistics	82	9	Quality	43
5	Customer Service	70	10	Taste	43

Table 3: Word frequency statistics for additional comments.

Ranking	Entry	Frequency	Ranking	Entry	Frequency
1	Effect	200	6	Customer Service	39
2	Products	71	7	Quality	38
3	Problem	59	8	Logistics	31
4	Genuine product	48	9	Take Medicine	26
5	Taste	39	10	Reviews	25

Further divide "effect", "appearance", "price" and "date" into product characteristics of pharmaceutical products, divide "customer service" and "logistics" into service characteristics of pharmaceutical products, analyze the similarity and difference of semantic features of online users' initial reviews and additional comments, and select the top 100 high-frequency knowledge meta-vocabularies respectively, and the statistical results of the two are shown in Table 4 below:

Table 4: Comparison of high-frequency vocabulary in the initial and additional reviews.

Comment type	Classify					
	Products				Logistics	Customer Service
	Effect	Appearance	Price	Date		
Initial comments	23.73 %	16.95 %	16.95 %	5.08%	13.56 %	23.73%
Additional comments	37.10 %	14.52 %	12.90 %	3.23%	11.29 %	20.97%

4.2 LDA Topic Analysis

For unstructured text content like comments, computers cannot directly understand, we need to introduce text mining to help us extract high-value information hidden in text descriptions. In natural language processing, LDA is an important generative model that can be used to identify hidden topic information in large scale documents or corpora (Suominen and Toivanen 2016; Shan and Li 2010). The LDA uses the bag-of-words approach and a joint distribution to calculate the conditional distribution of hidden variables under a given observable variable. Observable variables are a set of words, and latent variables are subjects (Lyu et al. 2022). The graphical representation of LDA model is shown in Figure5.

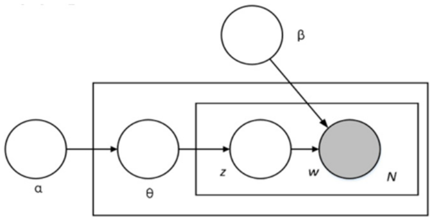


Figure 5: Graphical model representation of LDA.

α and β each control a Dirichlet distribution to obtain the distribution of words under different topics, so it is important to determine the number of topics in the LDA model (Lamba and

Madhusudhan,2019; Kushkowski et al. 2020). This article selects the optimal number of topics by calling the sklearn package in Python, performing unsupervised topic clustering analysis for initial comments and additional comments, respectively, and through topic confusion, GridSearchCV grid search and cross-validation, and visualization.

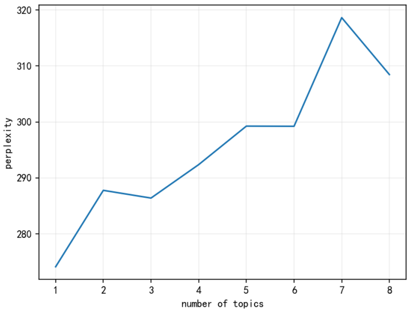


Figure 6: Perplexity trend of LDA for initial comments.

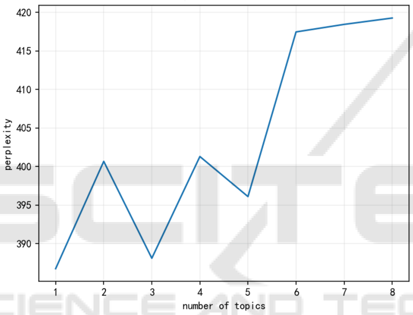
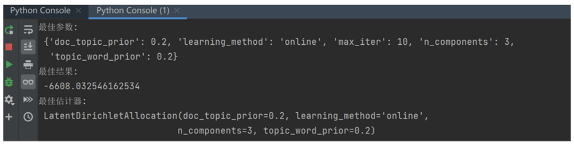


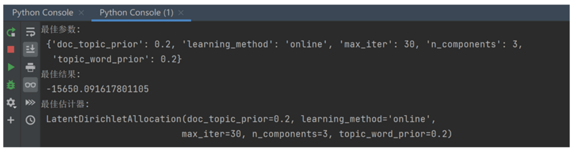
Figure 7: Perplexity trend of LDA for additional comments.

As can be seen from the above Figure6 and Figure7, in the case of not considering the number of topics of 1, whether for the initial comment or additional comment, the LDA model has the lowest degree of topic confusion when the number of topics is 3, that is, the optimal number of topics at this time is 3. In order to further verify the optimal number of topics, this paper also uses grid search and cross-validation to tune the model. The number of topics is set to 3-9 when tuning this model, and the optimal number of topics is searched in turn, and the data is cross-validated in five folds, and the parameters of the optimal model obtained are shown below:



```
最佳参数:
 {'doc_topic_prior': 0.2, 'learning_method': 'online', 'max_iter': 10, 'n_components': 3,
  'topic_word_prior': 0.2}
最佳结果:
 -6608.032546162534
最佳估计器:
 LatentDirichletAllocation(doc_topic_prior=0.2, learning_method='online',
                           n_components=3, topic_word_prior=0.2)
```

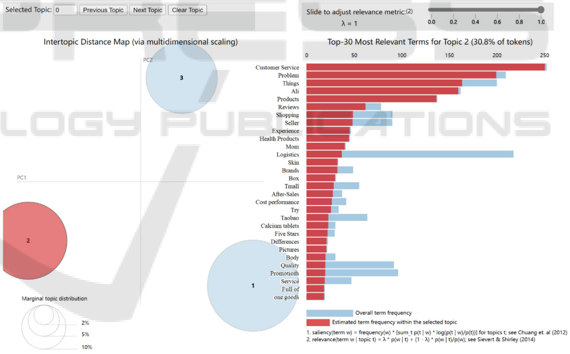
Figure 8: The best parameters of the LDA model for initial comments.



```
最佳参数:
 {'doc_topic_prior': 0.2, 'learning_method': 'online', 'max_iter': 30, 'n_components': 3,
  'topic_word_prior': 0.2}
最佳结果:
 -15650.091617801105
最佳估计器:
 LatentDirichletAllocation(doc_topic_prior=0.2, learning_method='online',
                           max_iter=30, n_components=3, topic_word_prior=0.2)
```

Figure 9: The best parameters of the LDA model for additional comments.

It can also be clearly seen from the above Figure8 and Figure9 that the parameters in the optimal model n components = 3, which is a further proof that for online reviews of pharmaceutical products, whether it is a first review or an additional review, the focus of consumer attention is on three major themes. We evaluated models with different numbers of topics (1 to 9 topics) on preliminary and post-evaluation data, respectively, with the help of pyLDAvis in Python Tool (Sievert and Shirley 2014) to visualize each model. When the number of topics is 3, there are relatively few connections be-tween the topics, which means that the models are relatively independent and have little to do with each other, indicating that the choice of the number of topics in this article is appropriate. However, the words of the LDA topic model are discrete, and taking the preliminary evaluation data as an example, as shown in Figure10.


Selected Topic: 0

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 10: Example diagram of LDA model visualization.

It is relatively inaccurate to determine what each type of topic represents and which ones are included under each topic Characteristic words, therefore, it is necessary to do further word frequency co-occurrence analysis on the comment text data, and word frequency co-occurrence analysis pays more attention to the connection between words than LDA model analysis, which can more clearly show the connection between subject words.

4.3 Word Frequency Co-Occurrence Analysis

In this article, the initial and additional comments are sorted by the number of word frequencies, and the top 1,000 are selected to fit the number of word frequencies, and the results are shown in the following figure:

As can be seen from Figure11 and Figure12, the number of word frequencies for initial comments and additional comments conforms to a stricter power index distribution, and the goodness of fit of the two is 0.92 and 0.97, respectively. In this paper, the words in the top 100 of the word frequency quantity are selected for word frequency co-occurrence analysis, and the high-frequency knowledge meta-vocabulary is selected It is to remove words that have only appeared once or twice, to eliminate some of the noise in the corpus, and to make high-frequency words more representative and more effective in reflecting the information contained in the comments.

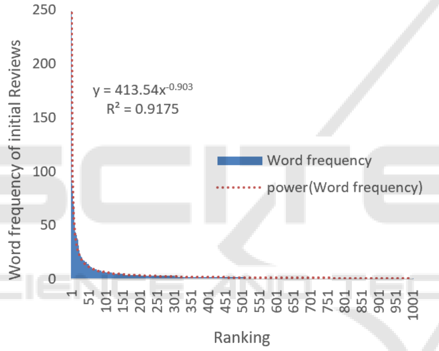


Figure 11: Word frequency count fitting for initial reviews.

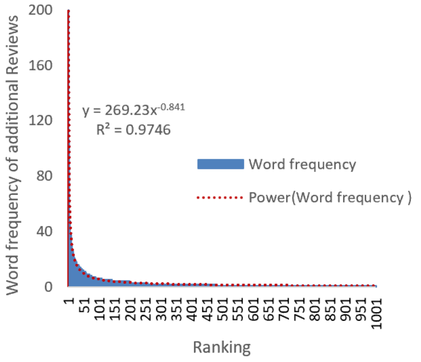


Figure 12: Word frequency count fitting for additional reviews.

Word frequency co-occurrence analysis is an unsupervised cluster analysis. It is generally believed that the more words appear in the same document, the more closely related the two topics are. According to the literature, there are many ways to calculate the similarity between words, such as Pointwise Mutual Information, Likelihood ratio, Average Mutual Information and so on (Terra and Clarke 2003). This paper calculates the semantic similarity between high-frequency words in word frequency analysis based on the Pointwise Mutual Information, that is, the PMI value, so as to achieve co-occurrence between words. This article uses Gephi to visualize the word frequency co-occurrence matrix, as shown in the following Figure13 and Figure14.

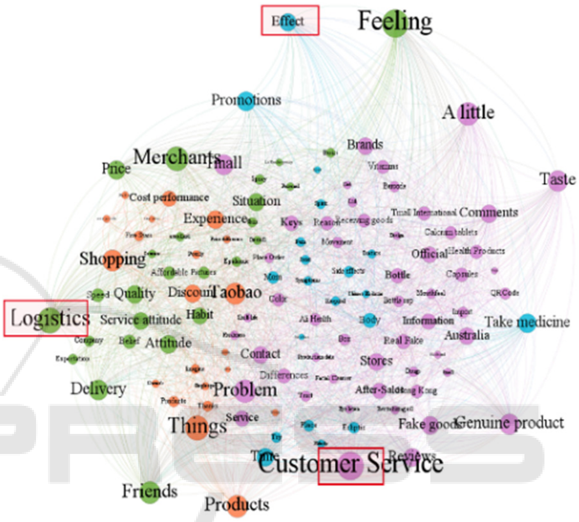


Figure 13: Initial comments word frequency co-occurrence.

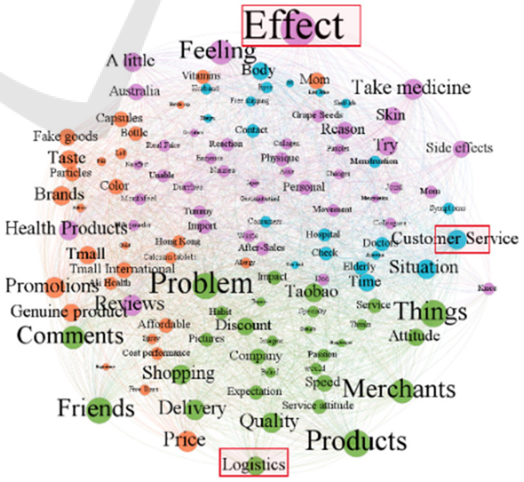


Figure 14: Additional comments word frequency co-occurrence.

As can be clearly seen from Figures 13 and 14 above, for unsupervised clustering analysis, the model automatically divides high-frequency words into 3 major themes according to the co-occurrence relationship between words. Each color represents a theme, and the co-occurrence relationship between words of the same color is stronger, and they belong to the same theme. The blue and orange themes in the preliminary review are also the themes under the theme of product characteristics, and the distribution of topics in the post-evaluation is similar to the above, so the description below this article focuses on the analysis of the three themes of "effect", "customer service" and "logistics". The larger the node, the higher the centrality, the more frequently it appears in the comment, and the more vocabulary it is connected to. For the theme of "effect", the size of the node in the post-evaluation is significantly larger than that of the preliminary evaluation, and the words associated with it are also greater than the preliminary evaluation in the post-evaluation, and in the preliminary evaluation, for the special commodity of the drug, consumers are more likely to hold a try-and-see attitude, and will only be in the post-evaluation after a period of use Further evaluation of the efficacy of the drug, so that the attention to the topic of "effect" in the follow-up evaluation has increased significantly; For the two major themes of "logistics" and "customer service", the node size in the post-evaluation is significantly smaller than that of the preliminary evaluation, and the "logistics" theme word is closely related to the words "delivery" and "attitude", and the keywords such as "after-sales service", "store" and "problem" are closely related to the theme word "customer service". This shows that consumers' focus on pharmaceutical products is reflected in both online initial and additional reviews, and there are both similarities and differences between the two.

5 CONCLUSIONS

Combined with the above figure and the above table, coupled with the above series of analysis, it can be concluded that the first comment and the additional comment have the similarity of the overall trend in the distribution of semantic features of the text content, that is, the description of the product accounts for the highest proportion, the description of logistics accounts for the lowest, the description of customer service is centered. Consumers' attention to pharmaceutical products is concentrated in the three themes of "effect", "customer service" and "logistics". From the perspective of difference, the additional content focuses more on the effect experience of drugs. Compared with the initial comment, the proportion of additional comments in other aspects is lower than the initial comment, and the attention to other aspects has decreased to a certain extent. This paper reveals the as-sociations and differences in semantic features of the initial and additional reviews, and performs a visual analysis of word frequency co-occurrence of high-frequency vocabulary, and points out feature words that are closely related to different topics.

It should be pointed out that there are certain limitations in this paper, first of all, the medical product review data selected in this paper is relatively small, and future studies can verify whether there is a consistent conclusion in the context of big data. Secondly, this article only conducts textual mining on the semantic aspects of the initial review of online reviews, and subsequent research can pay more attention to the emotions contained in the comment text. Finally, we hope that this research will help pharmaceutical e-commerce platforms to discover consumers' concerns about pharmaceutical products and make corresponding adjustments. Make the most of your website and improve the quality of its services while helping them better manage customer reviews.

ACKNOWLEDGEMENTS

This research was also supported by Kunming E-commerce and Internet Finance R&D Center (KEIRDC[2020]), the Prominent Educator Program (Yunnan [2018]11), Yunnan Province Young Academic and Technical Leader candidate Program (2018HB027) & Yunnan Provincial E-Business Entrepreneur Innovation Interactive Space (2017DS012), National Natural Science Foundation of China [71402159], Yunnan University Eastland Yong and Middle-aged Backbone Teachers Training Program (2018), 2021 Yunnan Provincial Academic Degrees Committee Program [SJZYXWALK202113], 2021 Yunnan University Program [XJKCSZ202117].

REFERENCES

Blei, D. M., & Jordan, M. I. (2003). Modeling annotated data ACM SIGIR FORUM,

Chatterjee, S., Goyal, D., Prakash, A., & Sharma, J. (2020). Exploring healthcare/health-product ecommerce satisfaction: A text mining and machine learning application. Journal of Business research, 131(7), 815-825.

Ghose, A., & Ipeirotis, P. G. (2007). Designing Novel Review Ranking Systems: Predicting Usefulness and Impact of Re-views. International Conference on Electronic Commerce,

Hu, C., & Ning, C. (2017). When is online post-evaluation more useful than preliminary evaluation? -Analysis of regulatory effects based on time interval and product type. Prediction, 36(04), 36-42.

Hu, N., Koh, N., & Reddy, S. K. (2014). Ratings lead you to the product, reviews help you clinch it? The mediating role of online review sentiments on product sales[J]. , 2014. Decision Support Systems (57), 42-53.

Jiang, W., Zhang, L., DaiI, Y., Jiang, J., & Wang, G. (2013). Analysis on the Usefulness of Online Reviews For User Needs. Chinese Journal of Computing Science, 36(01), 119-131.

Kushkowski, J. D., Shrader, C. B., Anderson, M. H., & White, R. E. (2020). Information flows and topic modeling in corporate governance. Journal of Documentation, 76(6), 1313–1339.

Lamba, M., & Madhusudhan, M. (2019). Mapping of topics in DESIDOC Journal of Library and Information Technology, India: A study. Scientometrics, 120(2), 477–505.

Li, X., & Chen, Y. (2016). The Influence of Word-of-Mouth Supplement Form on Purchase Intention: The Moderating Effect of Word-of-Mouth Direction. Psychological Journal, 48(06), 722-732.

Lyu, Y., Yin, M., Xi, F., & Hu, X. (2022). Progress and Knowledge Transfer from Science to Technology in the Research Frontier of CRISPR Based on the LDA Model. Journal of Data and Information Science, 7(1), 1-19.

Qiu, J., & Sun, Y. (2021). Online Review Research on University Environment Based on Text Mining. Library Theory and Practice(5), 16-22.

Shan, B., & Li, F. (2010). A survey of topic evolution based on LDA (in Chinese). Journal of Chinese Information Processing, 24(06), 43-49+68.

Shi, W., Lu, W., Na, S., & Cai, J. (2018). A Comparative Study into the Impact of Initial and Follow-on Online Comments on Sales. Management Review, 30(01), 144-153.

Shi, W., Xue, G., Zhang, Q., & Wang, L. (2016). A Comparative Study of Online Initial Review and Online Additional Review. Management Science, 29(04), 45-58.

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In Proceedings of the workshop on interactive language learning, visualization, and interfaces,

Sun, R., & Li, X. (2017). Research on the Influence of Contradictory Post-evaluation on Consumers' Purchase Intention. Geomatics and Philosophy and Social Sciences of Wuhan University, 70(01), 75-86.

Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. Journal of the Association for Information Science and Technology, 67(10-), 2464–2476.

Terra, E. L., & Clarke, C. L. A. (2003). Frequency estimates for statistical word similarity measures[C]. Proceedings of the 2003 human language technology conference of the North American Chapter of the Association for Computational Linguistics,

Wang, C., He, S., & Wang, K. (2015). Research on How Additional Review Affects Perceived Usefulness of Review. Journal of Management Sciences, 28(3), 102-114.

Wang, J., & Zhou, S. (2016). Research on the Influence of Consistent and Contradictory Online Reviews on Consumer Information Adoption: The Mediating Role of Perception Usefulness and the Moderating Role of Self-Efficacy Based on Perception. Library and Information Services (22), 94-101.

Wen, N., & Ding, S. (2010). A Review of Tendency Analysis of Subjective Product Review Information. Journal of Intelligence, 29(12), 70-74+48.

Yang, C., Zhang, H., Huang, J., & Wan, J. (2020). Text sentiment analysis of cigarette online reviews. Chinese Journal of Tobacco, 26(02), 92-100.

Zhang, N., & Qiao, D. (2020). Research on Sentiment Analysis of Online Learning Review Based on Deep Learning. Journal of Henan University of Urban Construction, 29(04), 63-71+92.

Zhang, Y., Wang, Y., Peng, L., & Liu, Y. (2020). Research on The Content Intelligence of Online User Additional Comments Based on Text Mining: A Case Study of Jingdong Mall Mobile Phone Review Data. Modern Intelligence, 40(09), 96-105.

Zhou, H., & Li, S. (2017). The Impact of Online Additional Comments on Consumers' Information Adoption. Sociology Mind, 07(02), 60-71.