ISSN: 2288-7709 © 2020 KODISA & ICMA.
JEMM website: https://acoms.kisti.re.kr/jemm
doi: http://dx.doi.org/10.20482/jemm.2023.11.1.21

Online Shopping Research Trend Analysis Using BERTopic and LDA

Yoon-Hwang JU[1], Woo-Ryeong YANG[2], Hoe-Chang YANG[3]

Received: January 21, 2023. Revised: January 30, 2023. Accepted: February 05, 2023.

Abstract

Purpose: As one of the ongoing studies on the distribution industry, the purpose of this study is to identify the research trends on online shopping so far to propose not only the development of online shopping companies but also the possibility of coexistence between online and offline retailers and the development of the distribution industry. **Research design, data and methodology:** In this study, the English abstracts of 645 papers on online shopping registered in scienceON were obtained. For the analysis through BERTopic and LDA using Python 3.7 and identifying which topics were interesting to researchers. **Results:** As a result of word frequency analysis and co-occurrence analysis, it was found that studies related to online shopping were frequently conducted on factors such as products, services, and shopping malls. As a result of BERTopic, five topics such as 'service quality' and 'sales strategy' were derived, and as a result of LDA, three topics including 'purchase experience' were derived. It was confirmed that 'Customer Recommendation' and 'Fashion Mall' showed relatively high interest, and 'Sales Strategy' showed relatively low interest. **Conclusions:** It was suggested that more diverse studies related to the online shopping mall platform, sales content, and usage influencing factors are needed to develop the online shopping industry.

Keywords : Research Trend, Retail Industry, Online Shopping, BERTopic, LDA (Latent Dirichlet allocation).

JEL Classification Code : C10, C80, M10, M31.

1. Introduction[1]

Online shopping can be classified into various types. For example, Internet-based shopping such as the general malls, specialized malls, large portal malls, open markets, social commerce, mobile commerce, TV home shopping, and catalog shopping is also classified as online shopping. The online shopping operation type is divided into an online mall that sells goods and services to end consumers only through computer and network bases and an on/offline mall that sells goods and services to end consumers in parallel with online as well as traditional offline commerce.

Unlike offline shopping, online shopping is subject to the Act on Consumer Protection in Electronic Commerce (abbreviation: Electronic Commerce Act). According to the Electronic Commerce Act, e-commerce is defined as conducting commercial activities using electronic transactions and is classified as mail order. According to this law, mail order sales are defined as providing information and receiving subscriptions from consumers to sell goods or

* This paper was presented at IFDC2023 to be held in Yangyang in January 2023, and it is revealed that this is an extension of the paper by Ju et al. (2023).

1 First Author, Assistant Professor, Department of Online Shopping, Jangan University, South Korea. Email: marketju@jangan.ac.kr

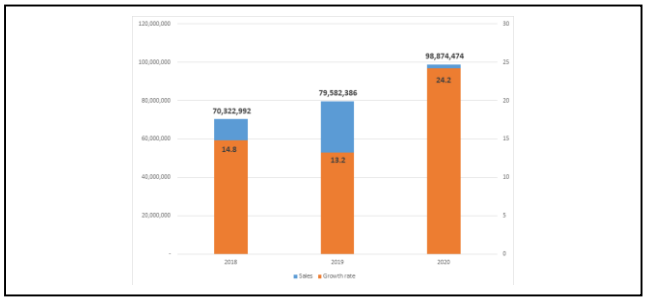
2 Second Author, Ph.D. Candidate, Integrated Course of Master & Doctoral, Dept. of Business Informatics, Hanyang University, Korea. Email: wooryeong325@gmail.com

3 Corresponding Author, Ph.D. Assistant Professor, Dept. of Distribution Management, Jangan University, South Korea, Email: pricezzang@jangan.ac.kr

services, and they are related to the sale of goods or services by using telecommunications, advertisements, advertising facilities, leaflets, broadcasting, newspapers, magazines, etc., or by using a money order, account transfer, etc. However, telephone solicitation sales under the Door-to-Door Sales Act are excluded from the scope of mail-order sales. By quoting the statutory definition, online shopping can be defined as 'A series of activities in which consumers receive information from companies and purchase goods or services using non-face-to-face media such as the Internet, catalogs, and TV'.

According to the Korea Chamber of Commerce and Industry(Korcham)'s '2021 Distribution Logistics Statistics', online shopping is classified as non-store retail stores, and non-store retail stores include internet shopping, home shopping, door-to-door, and delivery retail stores. Sales of non-store retail stores as of 2021 amounted to KRW 98.9 trillion, accounting for 26.7% of the total retail business, and the rate of change was reported at 24.2% (Korcham, 2022).

The reason this study pays attention to online shopping is that, from the consumer's point of view, despite the disadvantage that it is difficult to directly grasp the characteristics of a product, it has various advantages such as no time and space restrictions, guaranteed autonomy such as freely search for information and compare prices, and relatively low prices. In addition, this is because it is overgrowing due to the influence of changes in consumer propensity related to non-face-to-face and not contact consumption caused by COVID-19.



Note) The researcher schematizes the data of MOTIE (2022)

Figure 1: 2018-2020 Non-store retail business yearly sales trend

The purpose of this study is to identify topics that researchers are interested in by identifying research trends on online shopping so far and to explore various clues for the development of online shopping through this. In addition, this study intends to prepare the basis for proposals for developing the distribution industry and the possibility of coexistence between online and offline retailers.

The results of this study are expected to provide information necessary for online shopping companies to play a crucial role in the distribution industry and to provide various clues for coexistence with competing offline retailers.

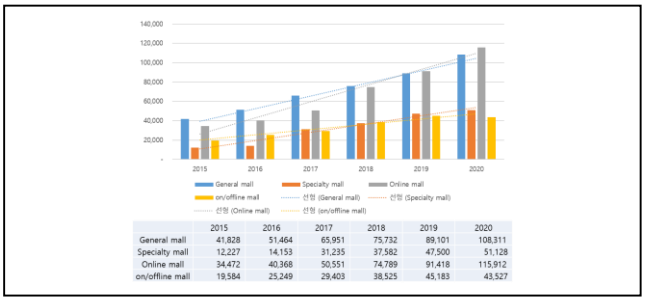
2. Literature Review

2.1. Online Shopping

The online shopping field can be classified into Internet shopping, TV home shopping, catalog shopping, etc., concerning media utilization. It can be classified into various categories such as online malls and on/offline malls, depending on the type of operation as stated above. For this reason, online shopping using internet media tends to overlap. Online shopping types classified by media utilization are presented in <Appendix 1>.

The distribution industry classified as wholesale and retail in the standard industry classification, is KRW 134.5 trillion as of 2020, accounting for 7.3% of the total GDP. It has emerged as one of the growth drivers of the national economy, accounting for 14.5% (Korcham, 2022). According to the report data of the Korea Chamber of Commerce and Industry, the total sales of wholesale and retail out of gross domestic product (GDP) by industry increased by approximately KRW 4.6323 trillion from KRW 132.6199 trillion in 2018 to KRW 137.2522 trillion in 2019. Then they increased to KRW 134.6 trillion in 2020, they decreased to KRW 621.3 trillion (Kocham, 2022).

The Korea Chamber of Commerce and Industry calculates the annual sales amount by classifying online shopping transaction amount standards into general mall/specialized mall or online mall and on/offline mall according to the type of operation. First, according to the data of the Korea Chamber of Commerce and Industry, the transaction amount of online shopping as of 2020 was reported to be 15,943.8 billion won, of which the general mall was 10,831 billion won, and the specialized mall was 5,112.7 billion won.



	2015	2016	2017	2018	2019	2020
General mall	41,828	51,464	65,951	75,732	89,101	108,311
Specialty mall	12,227	14,153	31,235	37,582	47,500	51,128
Online mall	34,472	40,368	50,551	74,789	91,418	115,912
on/offline mall	19,584	25,249	29,403	38,525	45,183	43,527

Note) The researcher schematizes the data of MOTIE (2022)

Figure 2: Transaction amount by type of online shopping from 2015 to 2020 (unit: KRW 100 million)

Meanwhile, according to data released by the National

Statistical Office in June 2022, Korea's online shopping transaction amount increased by 10.4% (1.5815 trillion won) from the same month last year to 16.7806 trillion won as of June 2022. Specifically, it increased due to the influence of travel and transportation services (102.0%), food and beverages (16.8%), cultural and leisure services (121.6%), and clothing (8.2%) (Kostat, 2022). This is not only due to the increase in external activities due to the recent lifting of social distancing and the lifting of the quarantine obligation for entrants from abroad after the spread of COVID-19 but also the increase in online shopping due to consumers learning about the convenience of transactions due to COVID-19, and this can be attributed to the effect of rapid logistics. Among them, mobile shopping increased by 15.8% (KRW 1,694.7 billion) from the same month last year to KRW 12,418.6 billion as of June 2022. The share of mobile shopping transactions among online shopping transaction amount was 74.0%, a 3.4%p increase from the same month last year. Foodservice (98.0%), e-coupon service (89.1%), pet products (85.0%), and children's/baby products (84.4%) were identified as the top items with a high percentage of mobile transaction volume by item (Kostat, 2022).

Comparing the trend lines by type in <Figure 2>, it can be seen that all types of online shopping, including general malls, are showing an upward trend. In particular, the most rapid growth of online malls can be attributed to the expansion of mobile commerce.

On the other hand, the expansion of online shopping concerns problems such as the 'Amazon effect' or 'retail apocalypse' in which traditional store-selling retailers go out of business (Mende & Noble, 2019). In a previous study, Kim et al. (2018) presented an analysis result that when the online sales growth rate increased by 1%p, the offline sales growth rate decreased by about 0.7%p. In a study by Jung and Song (2020), a 1%p increase in the national online sales growth rate reduced the regional retail sales growth rate by 0.1%p. On the other hand, Lee (2019) said that the increase in online transaction volume has a negative effect on the total number of retail businesses, but the effect is different depending on the industry. However, if this result is converted to direct overseas purchases online, it may be a factor hindering the economic development of Korea. In addition, the number of consumers who enjoy showrooming at online shopping malls after crossover shopping that combines online and offline distribution channels or comparative product search at offline stores is increasing. As a result, retailers are also facing the problem of bearing additional costs due to various types of investments, such as Omni-channel strategies. On the other hand, the sensitive attitude of suppliers is also emerging as a social issue that affect purchase decisions in online shopping such as reviews, etc. In addition, there is a structural disadvantage in that consumers cannot communicate sufficiently and make purchase decisions without fully understanding the characteristics of the product. Moreover, it is judged that there are still problems such as reliability problems in delivery, payment, and personal information protection of the distribution channels used, and problems such as returns and delays in receiving products. Also, the increase in online shopping sites has problems in that providers provide limited information about their position or are exposed to too much information, even though customers receive information-based product recommendation services.

Despite these disadvantages of online shopping, it is analyzed that the growth potential of online shopping is considerable. For example, it can overcome geographical limitations, the cost of starting and maintaining a business is relatively low compared to offline, it can save time, and it can provide buyers with various information related to characteristics and prices, as well as secure customers. Online shopping is attractive and various marketing strategies can be implemented.

Therefore, this study seeks to gain insight into ways to maximize the advantages of online shopping and compensate for the disadvantages by exploring topics that researchers have been interested in.

2.2. BERTopic & LDA (Latent Dirichlet allocation)

Natural language processing (NLP) is a challenging task in information management, semantic mining, and computer science that enables computers to derive meaning from human language processing in text documents (Jelodar et al., 2019). Among them, topic modeling is one of the most potent techniques in text mining for data mining, latent data discovery, and finding relationships between data and text documents (Yang, 2022). NMF (Non-Negative Matrix Factorization) (Févotte & Idier, 2011) or LDA (Blei et al., 2003) have various advantages in text mining. However, it has a weakness in that they require the number of topics, stopword list, and morpheme analysis to obtain optimal results and the semantic relationship between words is ignored (Yang, 2022). Therefore, text embedding technology is spreading in natural language processing (Yang, 2022).

BERT, released by Google in 2018, is an unsupervised learning model that can access input data from both sides to understand the context, and fine-tuning and pre-learning are possible (Kim & Yang, 2022). Unlike previous complex topic modeling (Bianchi et al., 2020; Devlin et al., 2018), BERTopic calls pre-trained language model data through BERT to remove elements of BoW (Bag of Word) and perform pre-learning as a technique that actively utilizes a language model (Devlin et al., 2018), it has recently been attracting attention from the academic world (Yang, 2022).

BERTopic creates context-sensitive vector representations of words and sentences. The semantic

properties of these vector representations allow similar texts to encode the meaning of the text approximately in vector space (Grootendorst, 2022). BERTopic, which utilizes clustering technology and class-based transformation of TF-IDF to create consistent topic representations, utilizes BERT-based embedding and c-TF-IDF word weights in the text embedding step, then text clustering for each domain. It is a topic modeling technique that finds latent meaningful themes in the text by doing (Ko et al., 2022).

On the other hand, LDA (Latent Dirichlet allocation) is the most used topic modeling technique in studies using topic models because it is highly applicable not only to structured data but also to unstructured data (Blei et al., 2003). LDA-based topic modeling has been applied to natural language processing, text mining, social media analysis, and information retrieval. It has characteristics that not only excel at demonstrating discrete data but also provide adequate access to find hidden structures and meanings in enormous information (Jelodar et al., 2019; Yang, 2022). LDA is a probabilistic generation model of a corpus. LDA's basic idea is that documents are included in a random mix of potential topics, and topics are organized into a distribution over words (Jelodar et al., 2019). LDA was developed by Blei et al. (2003) and was first introduced. Since LDA represents topics by word probabilities, the word with the highest probability in each topic is a characteristic that generally gives a good idea of what the word probabilities are from the LDA topic (Blei et al., 2003). Methods such as the variational method (Blei et al., 2003), expectation propagation, and Gibbs sampling (Griffiths & Steyvers, 2004) have been proposed to measure LDA parameters.

In this study, although BERTopic has various advantages, it is also noted that it is impossible to evaluate the performance of BERTopic with existing evaluation methods due to embedding characteristics (Bodrunova et al., 2020). The number of topics derived through BERTopic was much more significant than expected, so it was necessary to reduce them in the studies of Yang (2022) and Yang and Yang (2022). Also, it was pointed out that there were many outlier documents that needed to be adopted in BERTopic, and not considering them does not match the purpose of confirming actual research trends in the studies of Yang (2022) and Yang and Yang (2022). Therefore, in this study, as applied by Yang (2022) and Yang and Yang (2022) as well as Ko et al. (2022) and Kim and Yang (2022), this study tried to use BERTopic and LDA together.

3. Research Procedure

This study is a continuous study on the distribution industry, and the research procedure for analyzing online shopping research trends is presented in <Figure 3>.

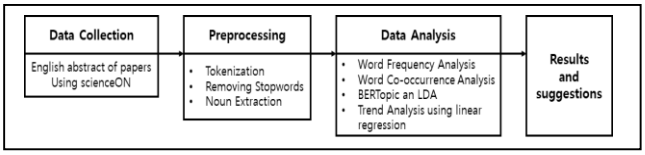


Figure 3: Research Procedure

3.1. Data Collection

As of October 10, 2022, a total of 934 papers were derived from searching for the keyword "online shopping" through scienceON (https://scienceon.kisti.re.kr). Of these, 645 English abstracts were obtained out of 901 papers, excluding redundant papers.

Table 1: Number of publications by year (n=645)

Year	1985~ 1999	2000~ 2004	2005~ 2009	2010~ 2014	2015~ 2019	2020~
Num	3	50	132	171	159	130

Note) Num means number of publications.

As a result of checking the number of publications by researchers, it was found that overall research on online shopping is increasing.

3.2. Data Preprocessing

For analysis, data were preprocessed and tokenized based on words. Along with special symbols and numbers, statistical terms such as 'aim', 'purpose', 'methodology', 'conclusion', 'regression', 'coefficient', and 'correlation', etc. that frequently appear in abstracts of academic papers, and 'online' and 'shopping' are set and removed as stop words. By doing so, this study tried to specify the topic extraction of studies related to online shopping.

4. Empirical Analysis Results

4.1. Results of Word Frequency Analysis

As a result of word frequency analysis for the real data, it was confirmed that 'product' (1,064), 'consumer' (988), 'mall' (948), 'customer' (857), 'information' (693), 'service' (547) were in the order. As a result of word frequency analysis, the top 20 words are presented in <Appendix 2>, and the word clouding results are presented in <Figure 4>.

Figure 4: Visualization results of Word Clouding

4.2. Results of Co-occurrence Frequency Analysis

As a result of analyzing the co-occurrence frequency of words based on connection centrality, mediation centrality, proximity centrality, and eigenvector centrality to confirm the relationship between words, it appeared to be words such as consumer, mall, and customer are mainly related to words such as product, service, and purchase. The top 10 words with co-occurrence frequency are presented in <Table 2>, and the results of visualizing word networks with a connection frequency of 80 or more are presented in <Figure 5>.

Table 2: Results of Word Co-occurrence Frequency Analysis

Rank	Word	Freq.	Rank	Word	Freq.
1	consumer	168	6	consumer	135
	product			information	
2	mall	159	7	information	121
	product			mall	
3	customer	152	8	product	120
	mall			purchase	
4	consumer	149	9	consumer	118
	mall			purchase	
5	information	138	10	customer	116
	product			product	

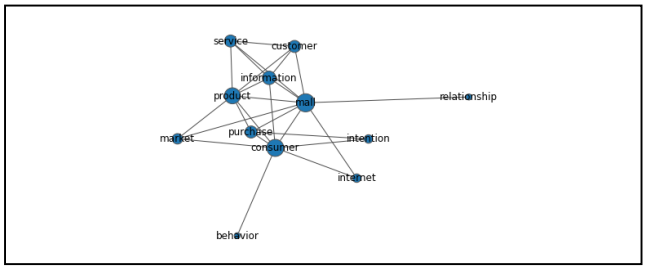


Figure 5: Network visualization results for co-occurrence frequency (connection frequency = 80 times)

As shown in <Figure 5>, studies related to online shopping were frequently conducted on factors such as information, products, services, malls, and purchases, focusing on customers or consumers.

4.3. Results of Topic Modeling

As a result of BERTopic, five topics and 286 outlier documents were derived.

<Topic 1> is composed of words such as mall, loyalty, quality, service, value, trust, etc., and it is judged to be related to the overall quality perception of online shopping malls, and it was named 'service quality'. <Topic 2> is composed of words such as business, product, sale, ecommerce, and strategy, so it was judged to be related to various strategies related to product sales in shopping malls, and it was named ‘sales strategy’. <Topic 3> consists of words such as fashion, clothing, clothes, etc., and it was judged to be related to fashion online shopping, so it was named ‘fashion mall’. <Topic 4> is composed of words such as recommendation, system, recommender, and algorithm, and it was judged to be related to the customer's post-purchase recommendation and it was named 'customer recommendation'. <Topic 5> is composed of words such as trust, risk, uncertainty, and purchase, and it is judged to be related to the safety of payment, which is the factor that consumers who use online shopping are most concerned about, and it was named ‘secure payment’.

As a result of LDA on 286 cases classified as outlier documents, the coherence score was the highest at 0.3276 when there were three topics, so 3 topics were added. <Topic 6> extracted from LDA was composed of words such as mall, purchase, experience, and user, and it was judged to be related to the experience of customers using online shopping malls, and it was named ‘purchasing experience’. <Topic 7> consists of words such as information, mall, intention, purchase, and behavior, which are judged to be related to the intention or attitude of consumers to shop online or behavior related to purchase, and it is classified as 'purchasing behavior'. Finally, <Topic 8> is composed of words such as value, image, mall, brand, satisfaction, quality, etc., and it is judged to be related to the degree of specific satisfaction with the shopping mall that customers feel, and it was named ‘shopping mall satisfaction'. The top 10 keywords by topic and each topic name determined together with experts are presented in <Table 3>.

Table 3: Results of Topic modeling using BERTopic and LDA

Topic	Topic Name	Top 10 Keywords by Topic
1	Service Quality	**mall**, customer, **loyalty, quality, service**, **satisfaction,** relationship, **value**, intention, **trust**
2	Sales Strategy	**business,** price, market, **product, sale,** customer, **ecommerce,**

		company, strategy, information
3	Fashion Mall	**fashion,** mall, image, **clothing,** service, **clothes,** consumer, product, motivation, purchase
4	Customer Recommendation	**recommendation,** product, **system, recommender,** customer, user, **algorithm,** technique, performance, network
5	Secure Payment	**trust, risk,** consumer, intention, relationship, **uncertainty, purchase,** value, experience, adoption
6	Purchasing Experience	**mall, purchase,** customer, **experience,** internet, store, site, information, **user,** system
7	Purchasing Behavior	**information, mall,** intention, **purchase,** service, internet, **behavior, attitude,** market, influence
8	Shopping Mall Satisfaction	customer, value, information, **image, mall, brand,** satisfaction, intention, **quality, service**

Note) Bold type indicates the keywords used to determine the topic name

4.4. Trend Analysis Results of Each Topic

By calculating the ratio of each topic assigned to a specific paper through topic modeling and calculating the average ratio by year of each topic using the calculated ratio and the year of publication of each paper, it is called dynamic topic modeling that the trend of the topic can be identified (Yang, 2022; Yang & Yang, 2022). Dynamic topic modeling of tracking each topic's appearance over time because the topic ratio is flexible over time (Blei & Laffety, 2006; Yang, 2022). The trend analysis results for each topic are visualized in <Figure 6>.

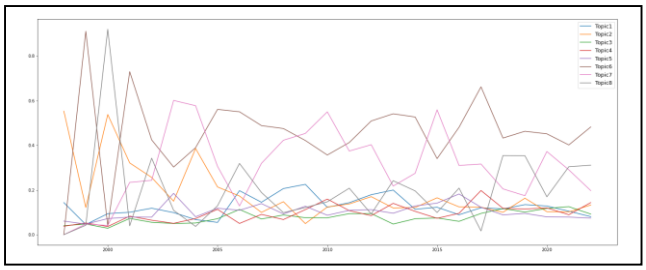


Figure 6: Visualization results of Trend Analysis

The result of the trend analysis in <Figure 6> has the disadvantage that it is difficult to predict which topics are receiving more attention. However, it is possible to confirm the degree of liquidity in which topics are attracting interest by year. Therefore, as in the studies of Yang (2022) and Yang and Yang (2022), the independent variable was set to the year of publication of the thesis, and the dependent variable was set to the average weight of the topics in that year, and then OLS regression analysis was performed (Yang, 2022; Yang & Yang, 2022).

As a result of the analysis, 'sales strategy' (coeff. = -0.0110, p<0.01) of <Topic 2> was found to be a topic with statistically significant lower interest, and 'fashion mall' of <Topic 3> ' (coeff. = 0.0025, p<0.01) and 'customer recommendation' (coeff. = 0.0035, p<0.01) in <Topic 4> were identified as topics with relatively high interest in the 95% confidence interval (See <Appendix 3>). The results of dynamic topic modeling visualization of hot topics and cold topics are presented in <Figure 7>.

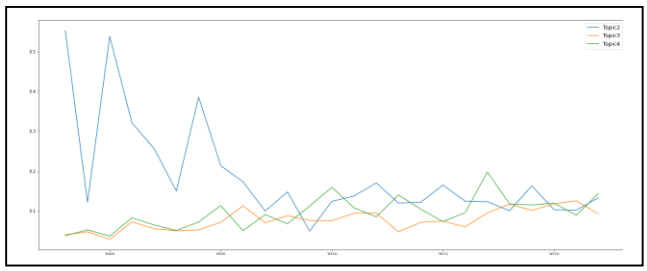


Figure 7: Visualization results of hot & cold topic

5. Conclusions

As one of the series of studies on the distribution business, this study attempted to identify topics that researchers are interested in by exploring the research trends of online shopping in response to the offline distribution business. Through this, various clues for the development of the online shopping industry and the possibility of the coexistence of distribution companies and the development of the distribution industry were intended to be laid. To this end, the following suggestions were presented through the analysis results through BERTopic and LDA using Python 3.7 for 645 English abstracts of papers registered in scienceON.

First, as a result of word frequency analysis, 'product', 'consumer', 'mall', 'customer', 'information', 'service', etc. appeared frequently, and as a result of checking the co-occurrence frequency of words such as 'consumer', 'mall', 'customer', etc. found to be mainly related. These results show that researches related to online shopping by researchers have frequently conducted related studies on factors such as products, services, and purchases targeting customers or consumers or shopping malls. Second, as a result of BERTopic, five topics such as 'Service Quality', 'Sales Strategy', 'Fashion Mall', 'Customer Recommendation', and Secure Payment' for the shopping mall were drawn. LDA was performed on 286 outlier

documents that three topics were derived such as 'Purchase Experience', 'Purchase Behavior', and 'Shopping Mall Satisfaction'. These results mean that various studies have been conducted on fashion malls among shopping malls, along with research on service quality for online shopping malls, sales strategies of shopping malls, safe payment, customer recommendation, purchasing experience, purchasing behavior, and shopping mall satisfaction. Third, as a result of trend analysis for each topic, it was found that 'Sales Strategy' was a topic with a low level of interest statistically significant, and 'Customer Recommendation' and 'Fashion Mall' were a topic with a relatively high level of interest. These results can be interpreted as the researcher's interest in the sales strategy of online shopping malls until the mid-2000s. However, since then, the importance of customer recommendations such as online word-of-mouth is increasing, and the research area of interest in online shopping is the part related to fashion.

However, the analysis results so far show limitations in the approach from various perspectives related to solving the problems of online shopping malls or developing them. Therefore, based on the analysis results so far, in this study suggests paying attention to the following research topics targeting online shopping in the future.

First, it is necessary to study the structure of online shopping malls, especially the platforms such as IT technology. This is because large companies have the capacity and capital to utilize a stable platform to some extent, but if not, it can be a very necessary factor. The shopping platform should consider not only the possibility of customization and site security but also the accessibility and convenience of consumers, such as providing information on shopping malls and search functions. In the case of one-person founders and small online shopping malls, it is difficult to use a self-developed platform, so they have no choice but to use an already-developed shopping platform. From this point of view, online shopping mall founders will likely provide various information to prospective entrepreneurs in comparative studies and effectiveness studies on the shopping platforms provided.

Second, it is necessary to conduct detailed research related to consumer use satisfaction by sales content and fashion. The research results on which part of shopping mall content is most effective will be useful information for online shopping companies. For example, in the case of online shopping mall A, it provides sufficient images/photos and detailed product descriptions. However, if online shopping mall B lacks in some areas, it is possible to revitalize the online shopping mall can increase through research related to the effectiveness of the content provided to consumers. However, it is doubtful that users, small businesses, or one-person entrepreneurs will directly explore these areas. Therefore, research results on which part of shopping mall content is most effective will be useful information for online shopping companies. From a different perspective, a study on customer satisfaction in the e-commerce industry by applying customer value, especially customer lifetime value (CLV), is also expected to provide various ideas to e-commerce companies.

Third, research on the factors affecting the use of online shopping is also needed. For example, although it is expected that research results such as ‘Service Quality’ in <Topic 1> or ‘Secure Payment’ in <Topic 5> derived from this study will include some influencing factors, it is necessary to conduct research related to delivery or sharing on social media, communication convenience related to returns/refunds, the possibility of the internet or mobile optimization of screen size, photos, fonts, etc., and consumer distrust of counterfeit brands. These results are expected to increase the effectiveness of online shopping.

Fourth, along with customer segmentation, a study on how to predict customer characteristics and develop marketing suitable for them will also help strengthen the competitiveness of e-commerce companies.

Fifth, it is necessary to study effective strategies related to the development of online shopping companies, such as the operation of experience stores, for the Omni-channel strategy in which offline retailers add online shopping. The research results related to this will solve problems such as difficulty in identifying the characteristics of products, which is pointed out as the biggest drawback of online shopping and can be clues that can contribute to the vitalization of the online shopping market.

Lastly, it is necessary to study the impact of the outflow of national wealth, such as problems related to direct overseas purchases and reverse direct purchase strategy. These results are expected to help the government establish policies to vitalize the retail industry.

Although this study confirmed the research trend of online shopping from the perspective of the distribution industry and make suggestions for development, there are some limitations, which need to be supplemented in future studies. First, the fundamental problem of topic modeling is the limitation of determining the topic name. Although in this study, the topic name was determined through collaboration with professors from J University who majored in distribution and marketing, there is a limitation that the topic name needs to reflect all the specific contents actually studied. Therefore, in future research, it will be necessary to study a method for determining a topic name that can cover words classified as topics as much as possible. Second, after deriving five topics from BERTopic, additionally conducting LDA on outlier documents was an attempt to specifically confirm the research trend, which is the purpose of this study. However, there is still a limit to

the practical-level approach academically. Therefore, in future research, it is necessary to identify a more compelling topic modeling technique by checking whether there is a difference in classified topics by performing BERTopic and LDA on the same data. Finally, in this study, 'online shopping' was limited to stopwords. However, as confirmed in the word frequency analysis, the high frequency of ‘mall’ made it difficult to determine how far to exclude it from analysis in confirming research trends in a field. Therefore, in future studies, it is necessary to consider whether to remove words that can directly affect keywords in the field.

References

Bianchi, F., Terragni, S., Hovy, D., Nozza, D., & Fersini, E. (2020). Cross-lingual contextualized topic models with zero-shot learning. arXiv preprint arXiv:2004.07737.

Blei, D. M., & Lafferty, J. D. (2006, June). *Dynamic topic models*. In Proceedings of the 23rd international conference on Machine learning (pp. 113-120).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3*(Jan), 993-1022.

Bodrunova, S. S., Orekhov, A. V., Blekanov, I. S., Lyudkevich, N. S., & Tarasov, N. A. (2020). Topic detection based on sentence embeddings and agglomerative clustering with markov moment. *Future Internet, 12*(9), 144. https://doi.org/10.3390/fi12090144

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Févotte, C., & Idier, J. (2011). Algorithms for nonnegative matrix factorization with the β-divergence. *Neural computation, 23*(9), 2421-2456.

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv preprint arXiv:2203.05794.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent dirichlet allocation(LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications, 78*(11), 15169-15211.

Ju. Y. H., Yang, W. R., & Yang, H. C. (2023). Online shopping research trend analysis using unsupervised learning. *Proceedings of the 10th International Forum on Business Convergence(ICFBC2023) of KODISA* (pp.141-144). Yangyang, Korea: KODISA.

Jung, M. S., & Shong, H. J. (2020). *Effect of online shopping growth on local retail industry: Focusing on Busan region*. Bank of Korea Regional Economic Report, Bank of Korea Busan Headquarters. p.113-127.

Kim, S. W., & Yang, K. D. (2022). Topic model augmentation and extension method using LDA and BERTopic. *Korean Society for Information Management, 39*(3), 99-132.

Kim, T. K., Park, D. J., Choi, I, H., Lee, E. W., & Jang, T. Y. (2018). Ripple effects and implications of online transaction expansion. *BOK Issue Note, 2018*(10), Bank of Korea. https://www.bok.or.kr/portal/bbs/P0002353/view.do?nttId=10048752&menuNo=200433

Ko, Y. S., Lee, S. B., Cha, M. J., Kim, S. D., Lee, J. H., Ham, J. Y., & Song, M. (2022). Topic modeling insomnia social media corpus using BERTopic and building automatic deep learning classification model. *Korean Society for Information Management, 39*(2), 111-129.

Korcham (2022). *Distribution logistics statistics for 2021*. Seoul: Korea Chamber of Commerce and Industry Distribution and Logistics Agency.

Kostat(2022). June 2022 Online Shopping Trends. (http://kostat.go.kr/wnsearch/search.jsp)

Lee, K. B. (2019). A study about the effects of online commerce on the local retail commercial area. *Economic Analysis, 25*(2), 54-95.

Mende, M., & Noble, S. M. (2019). Retail apocalypse or golden opportunity for retail frontline management? *Journal of Retailing 95*(2), 84-89.

MOTIE (2022). *Ministry of Trade, Industry and Energy. Annual '21, December '21 sales trends of major distribution companies*. Press release dated January 27, 2022

Park, C. W. (2015). *Practical distribution theory*. Seoul: Cheongnam Book Publishing House.

Yang, H. C. (2022). Analysis of distribution industry research trends using BERTopic and LDA. *Journal of Creativity and Innovation (JCI), 15*(4), 71-103.

Yang, W. R., & Yang, H. C. (2022). Topic modeling analysis of social media marketing using BERTopic and LDA. *Journal of Industrial Distribution & Business, 13*(9), 39-52.

Appendixes

Appendix 1: Types of online shopping based on media utilization

Media	Type	Definition & Characteristics	Enterprise example
Internet	General mall	- Internet shopping mall that handles various products - Mainly operated by large retailers - The main source of income is the sales commission received from the product supplier	El Lotte Hyundai H Mall Shinsegae Mall
	Specialized mall	- Internet shopping mall that handles limited category products - Operated by large companies, mid-sized companies, etc., or by individual operators using EC hosting	Kyobo Book Center Aladdin Morning glory
	Large portal mall	- Although they do not directly operate a shopping mall, they generate revenue by utilizing traffic - Limited to the case of providing only information delivery channels	Naver Daum
	Open market	- An online marketplace opens to all businesses that want to sell - Not only manufacturers, but also various types of businesses such as wholesale/retailers, class companies, and individual businesses participate. - Does not take responsibility for sales, and plays the role of providing computer infrastructure	Taobao (China) 11th Street Interpark
	Social commerce	- Started with joint purchase of initial region-based service products - A small number of products are sold at a low price, and a high proportion of sales are made through SNS on smartphones	Groupon (USA) Coupang Timon
	Mobile commerce	- Commercial transactions using mobile devices - Overall Internet-based transactions such as general malls and specialized malls are mobile-centered, so they are also included in the M-commerce area.	Baemin Yogiyo
Print	Catalog shopping	- Online shopping in which customers who receive product catalogs from home shopping companies and department stores place orders over the phone. - After peaking in 2001~2002, the decline has recently been revitalized led by home shopping companies.	Hyundai Home Shopping GS Shop
Broadcast	TV home shopping	- A type of retail business in which buyers order and pay for products they are interested in using a phone or smartphone while watching a TV home shopping channel	CJ O Shopping Hyundai Home Shopping GS Home Shopping

Note) The data of Park (2015) and the data of MOTIE (2022) are summarized and tabulated by the researcher

Appendix 2: Top 20 keywords as a result of word frequency analysis

No	Word	Freq.	No	Word	Freq.	No	Word	Freq.	No	Word	Freq.
1	product	1064	6	service	547	11	value	354	16	system	270
2	consumer	988	7	purchase	498	12	quality	341	17	site	251
3	mall	948	8	intention	419	13	market	314	18	price	237
4	customer	857	9	satisfaction	414	14	relationship	287	19	behavior	231
5	information	693	10	internet	367	15	user	281	20	experience	231

Appendix 3: Results of OLS regression analysis on topics by year

Topic No	Topic Name	Coefficient	T-value	P-value	Trend
1	Service quality	0.0006	0.493	0.627	-
2	Sales strategy	-0.0110	-3.829	0.001	Cold
3	Fashion mall	0.0025	4.876	0.000	Hot
4	Customer recommendation	0.0035	4.236	0.000	Hot
5	Secure payment	0.0008	0.867	0.395	-
6	Purchasing experience	-0.0014	-0.292	0.773	-
7	Purchasing behavior	0.0013	0.282	0.781	-

8	Shopping mall satisfaction	-0.0004	-0.080	0.937	-