International Journal of Advanced Culture Technology Vol.11 No.4 378-384 (2023)
DOI https://doi.org/10.17703/IJACT.2023.11.4.378

IJACT 23-12-45

An Exploratory Study on Issues Related to chatGPT and Generative AI through News Big Data Analysis

[1]Jee Young Lee

[1]*Assistant Professor, Department of Software, SeoKyeong University, Korea*
J.Ann.LEE@skuniv.ac.kr

Abstract

In this study, we explore social awareness, interest, and acceptance of generative AI, including chatGPT, which has revolutionized web search, 30 years after web search was released. For this purpose, we performed a machine learning-based topic modeling analysis based on Korean news big data collected from November 30, 2022, when chatGPT was released, to August 31, 2023. As a result of our research, we have identified seven topics related to chatGPT and generative AI; (1)growth of the high-performance hardware market, (2)service contents using generative AI, (3)technology development competition, (4)human resource development, (5)instructions for use, (6)revitalizing the domestic ecosystem, (7)expectations and concerns. We also explored monthly frequency changes in topics to explore social interest related to chatGPT and Generative AI. Based on our exploration results, we discussed the high social interest and issues regarding generative AI. We expect that the results of this study can be used as a precursor to research that analyzes and predicts the diffusion of innovation in generative AI.

Keywords: *ChatGPT, Generative AI, News Big Data Analysis, Topic Modeling, LDA*

1. Introduction

It has been 30 years since the W3Catalog, the first primitive search engine, was released on September 2, 1993. The world changed by web search engines is being turned upside down by a new game changer called Open AI. ChatGPT had 1 million users within 5 days of its launch, and showed rapid growth with Monthly Active Users (MAU) exceeding 100 million within 2 months [1]. Existing search engine companies are competitively releasing search engines that apply generative AI to respond to chatGPT. Many companies are releasing generative AI based services.

ChatGPT, which was released by OpenAI on November 30, 2022, quickly became a protagonist by attracting the attention of the general public. Unlike previous AI models, chatGPT received rapid attention because it was released as a chat interface that the general public can easily use. ChatGPT is a Large Language Model (LLM) that learned large amounts of data in various fields. Therefore, customized answers are provided not only for specific areas of expertise but also for most questions that may arise in everyday life. This allowed users to experience improvements in productivity and efficiency in their daily lives. The personalized AI assistant service that users have been wanting for so long has been realized. ChatGPT changed the Internet

Manuscript Received: September 6, 2023 / Revised: October 10, 2023 / Accepted: October 20, 2023
Corresponding Author: J.Ann.Lee@skuniv.ac.kr
Tel: +82-2-940-7757
Assistant Professor, Dept. of Software, SeoKyeong Univ., Korea

from ‘an ocean of information for everyone’ to ‘an ocean of knowledge for me.’

This study sought to explore social awareness, interest, and acceptance of chatGPT, which brought about an innovative evolution in web search, 30 years after the initial work on web search was released. To this end, this study analyzed Korean news big data related to chatGPT and generative AI and explored significant words and topics.

2. Related Works

2.1 Generative AI and Information Retrieval

Generative AI is an artificial intelligence technology that actively generates results in response to requests based on user input or data sets [2]. While existing deep learning-based AI technology was limited to making predictions or classifications based on data, generative AI finds, analyzes, and understands data on its own in response to questions asked by users, summarizes the optimal answers, and presents them as results. It is an evolved AI technology [2]. A representative service of generative AI is chatGPT, applied to chatbots and search engine services. Existing web information retrieval services such as Google and Naver provided a list of search results containing the keyword when the user entered a search keyword. Since the advent of chatGPT, generative AI-based information retrieval analyzes and understands questions through natural language processing when an interactive question is entered, summarizes the searched content through an AI learning algorithm, and presents a complete answer.

ChatGPT is a generative AI model developed by OpenAI and uses Generative Pretrained Transformer (GPT) as an LLM. OpenAI announced chatGPT on November 30, 2022, and in March 2023, chatGPT-4, which uses a learning model about 500 times larger than chatGPT-3.5. In response, Google released ‘Bard’, a chatbot service using Pathways Language Model (PaLM) as an LLM, and Meta released an LLM called ‘Large Language Model Meta AI (LLaMA)’. In Korea, Naver developed ‘OCEAN’, a huge language model specialized for Korean, and launched ‘HyperCLOVER X’, an OCEAN-based chatbot service.

2.2 News Big Data Analysis

News text is unstructured data whose content is important because it contains rich facts and opinions about issues across society, such as politics, economy, society, and culture. News big data analysis is a methodology that extracts structured data by extracting entity names, quotations, and sentences such as people, places, and events through natural language processing from unstructured news articles [3, 4]. This is used as a method to analyze social issues or social trends. News big data analysis methodology is a series of methods to analyze news big data by applying text mining techniques. Among text mining techniques, Latent Dirichlet Allocation (LDA)-based topic modeling is widely used to derive keywords and topics latent in news text [3, 5-8].

3. Research Methods

3.1 Data Collection

Since the purpose of this study is to identify social issues related to generative AI, represented by chatGPT, news big data was the subject of analysis. For news collection, BIGKinds, a news big data service of the Korea Press Foundation, was used. The news data collection period was set from November 30, 2022, when chatGPT was released, to August 31, 2023, at the time of the current study. The search expression was set to (‘chatGPT’ OR ‘Generative AI’). News search targets are 54 media outlets provided by BIGKinds. As a result of collecting news, 16,471 news items were collected. Excluding 557 exceptional articles (obituaries, sympathy, greetings, etc.), 15,914 news articles were selected for analysis.

3.2 Data Analysis Methods

Since the analysis data is news big data, this study used LDA-based topic modeling among the text mining techniques based on machine learning algorithms to extract latent topics in the news [9]. In this study, the Python Gensim library was used in the Anaconda 3 (Python 3.11.3) environment for analysis work.

As for the analysis procedures performed in this study, first, preprocessing to remove meaningless characters was performed on the collected news texts using Python. Second, the words included in the news were separated using the Python Gensim library. And based on the frequency of occurrence, a corpus and dictionary to be input into the LDA model were created. Third, an LDA model was constructed and machine learning was performed using an unsupervised learning method.

As the final task of analysis, we analyzed the topic probability of words extracted from the LDA topic model, the topic probability of the document, and topic-specific words extracted from the LDA visualization task, and with the results, we labeled each topic.

Then, the coherence value of the model was calculated to determine and set the optimal topic coefficient [10]. Fourth, to explore the results of topic modeling, we explored changes in the monthly appearance frequency of topics and performed LDA visualization [8, 11]. This study used LDAvis, a web-based interactive visualization library, for LDA visualization work [11]. LDAvis uses Principal Component Analysis (PCA), a dimensionality reduction method, and keyword extraction method to visualize relationships between topics and keywords. As shown in Figure 1, the left panel visualizes the Inter-topic Distance Map showing the distribution of topics in the four quadrants, and the right panel shows the main terms for the topic selected in the left panel. Additionally, the closer you set the relevance metric λ on the right panel to a value of 0, the more you can find terms that appear specifically for the current topic [11]. This study set λ to 0.6 and used visualization of the frequency estimate in the selected topic compared to the overall frequency of the term for analysis.

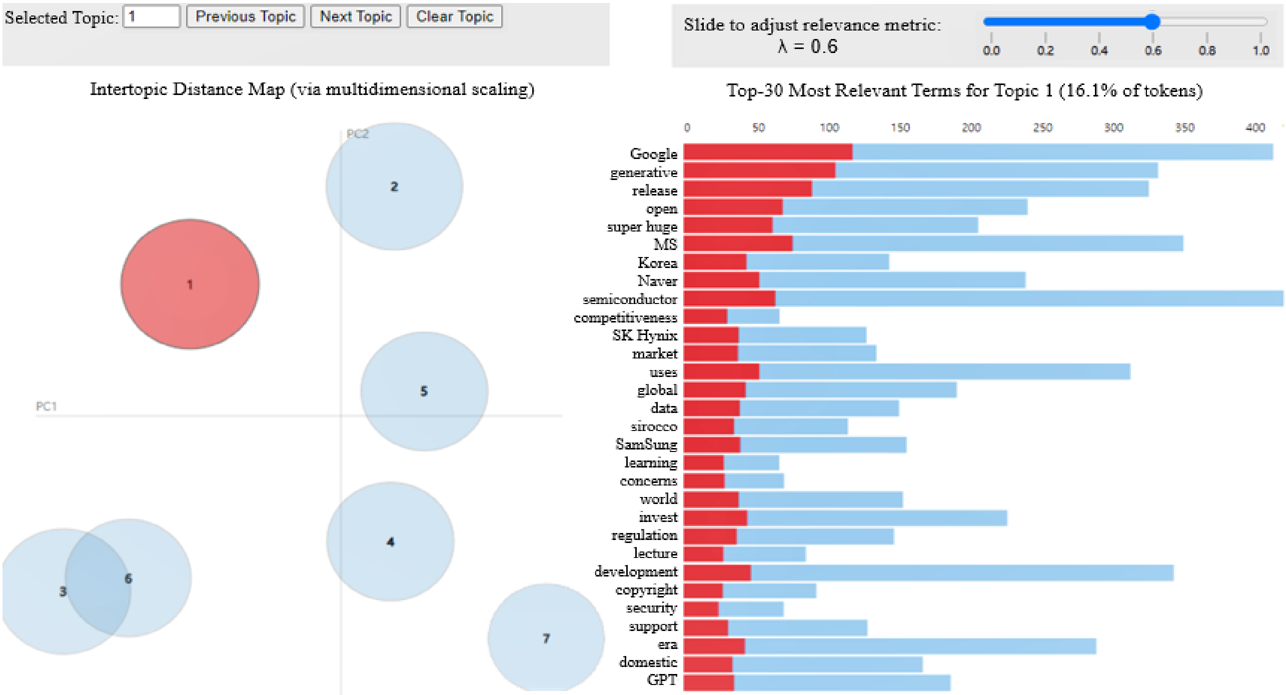

Figure 1. Visualization of LDA model results using LDAvis

4. Results and Discussion

As a result of analyzing 15,914 news big data, 7 topics were derived as shown in Table 1. The second column of Table 1 shows the number of news by topic, and the third column shows the proportion. The fourth column indicates the document number and potential probability values of the two documents (news) with the highest topic potential probability values. The last column is the label of the topic determined by analyzing words classified in the topic, topic-specific words extracted by setting λ to 0.6 in the LDA visualization task, and documents with high topic potential probability.

Table 1. Summary of topics extracted using LDA topic modeling

Topic ID	Number of news	Proportion by topic	Top 2 representative news	Topic label
Topic-1	2613	16.4%	d6449(0.934), d11870(0.922)	Growth of the high-performance hardware market
Topic-2	2476	15.6%	d9012(0.943), d894(0.928)	Technology development competition
Topic-3	2404	15.1%	d9663(0.928), d10795(0.928)	Service contents using generative AI
Topic-4	2138	13.4%	d8563(0.933), d6615(0.928)	Human resource development
Topic-5	2108	13.3%	d2688(0.921), d2108(0.914)	Instructions for use
Topic-6	2096	13.2%	d12548(0.913), d2743(0.913)	Revitalizing the domestic ecosystem
Topic-7	2076	13.0%	d5361(0.938), d4216(0.922)	Expectations and concerns

Figure 2 shows the news volume from November 2022 to August 31, 2023, to explore changes in news volume on the topic. The graph at the top of Figure 2 shows the monthly news volume over the data collection period. The graph at the bottom of Figure 2 shows the monthly topic frequency ratio for the same period. In December 2022, there was high interest in Topic-4(Human resource development). Social interest shifted to Topic-1(Growth of the high-performance hardware market) in January 2023, and continues to receive high attention to this day. As of August 2023, it appears that the news ratios for the seven topics have become similar. Since chatGPT and generative AI have passed the early adoption stage and now in the early majority stage, it can be interpreted that they are receiving high interest in various topics rather than any specific topic.

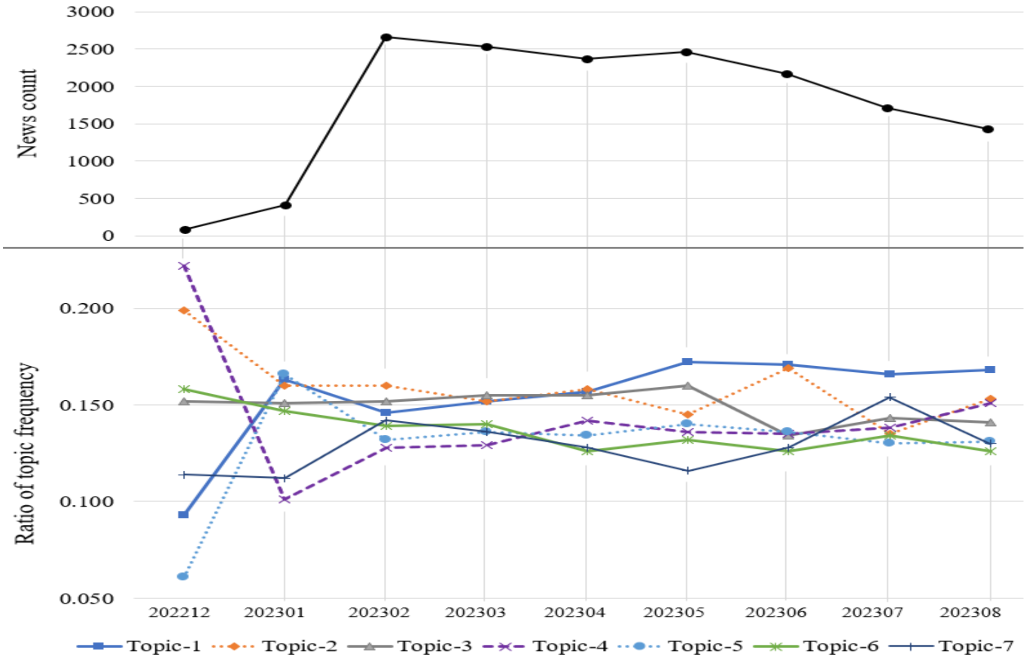

Figure 2. Ratio of news volume and news topic frequency monthly

The seven topics derived are summarized as follows.

Topic-1: Growth of the high-performance hardware market. The first representative document of Topic-1, d6449 was published by Seoul Economic News on May 18, 2023, and was an article on the price rise of high-performance next-generation DRAM (DDR5), which was essential for generative AI and ultra-large data center and the growth drivers of the memory industry. The second representative document of Topic-1, d11870, was published by Money Today on March 13, 2023, was an article about high-performance computing (HPC)-based data centers required for generative AI. Looking at related news comprehensively, it was discussed that the demand for high-performance hardware will increase for the use and deployment of chatGPT and generative AI, which will lead to growth in the semiconductor and HPC hardware market.

Topic-2: Technology development competition. The representative document of Topic-2, d9012 was published by Munhwa Ilbo on April 17, 2023, stated that Microsoft, stimulated by chatGPT, combined generative AI with the Bing search engine, and Google built a completely new AI technology search engine to improve performance. In addition, d894, published on August 16, 2023 in the Hankook Ilbo, the representative document of Topic-2, stated that even the Middle East had joined the race to develop AI technology led by the United States and China. Topic 2 was about the fierce technology competition and global technology investment of big tech companies such as Microsoft, Google, Naver, and Kakao in the generative AI field, including ChatGPT.

Topic-3: Service contents using generative AI. News classified in Topic-3, including d9663, published on April 7, 2023 in Maeil Business Newspaper, and d10795, published on March 26, 2023 in Seoul Economic Daily, which are representative documents of Topic-3, are service contents using generative AI. It became an issue that it is becoming a profit model and innovation engine in all industries.

Topic-4: Human resource development. News, including representative document of Topic-4, d8563, published on April 20, 2023 in Seoul Economy, and d6615, published on May 16, 2023 in Maeil Newspaper, discuss the topic that the government, corporations, and society should work together to foster talent needed in the generative AI era.

Topic-5: Instructions for use. The representative documents for Topic-5, such as d2688, published on July 10, 2023 in Seoul Economy, and d2108, published on July 19, 2023 in Chosun Ilbo, discussed the need for ethical standards, certification, and impact assessment in using generative AI effectively.

Topic-6: Revitalizing the domestic ecosystem. News classified in Topic-6, including representative documents d12548, published on March 3, 2023 in Money Today, and d2743, published on July 10, 2023, in Electronic Newspaper, emphasized the need for mutual cooperation and strategic alliance for the development of the domestic generative AI ecosystem. They also discussed the importance of institutional and cultural establishment, as well as the necessity of nurturing policies.

Topic-7: Expectations and concerns. News articles clustered under Topic-7, including representative documents d5361, published on June 1, 2023 in Electronic Newspaper) and d4216, published on June 16, 2023 in National Daily), expressed expectations that ChatGPT would assist in enhancing tasks such as HR and finance and help the shortage of human resources. They also showed interest in ChatGPT's responses to various questions related to societal issues. In contrast, there were critical articles regarding the use of ChatGPT to write condolences for a shooting incident at a U.S. university. Additionally, there were reports on errors in ChatGPT's responses, such as the incident involving King Sejong throwing a MacBook Pro. Furthermore, there were warnings about the potential harm to human emotions if the capabilities of generative AI were misused.

5. Conclusion

This study was conducted to explore social interests and major topics using Korean news big data related to chatGPT and generative AI. Through this study, we were able to explore issues related to chatGPT, which has spread beyond the initial acceptance stage. ChatGPT and generative AI are developing competitively at a rapid pace as they are released for free as interactive services that general users can easily use in various fields. As a result, the market for related industries, such as the high-performance hardware industry, is growing. In addition, there are generative AI-utilizing service contents and talent training, and an ecosystem is rapidly being formed. Everything is moving rapidly. Reminiscent of the metaverse hype of just a year or two ago, concerns over too many expectations and fast pace is also being discussed. The UN Security Council held its first meeting on July 18, 2023 to discuss and respond to future risks brought about by AI.

As a result of this study, we have found that chatGPT and generative AI have progressed beyond the early adoption stage at an unprecedented pace and are now in the mid to late stages of the early majority stage. We identified both positive reactions, such as performance, expectations, competition, and investment, as well as negative responses, including apprehension and acceptance resistance. The era of AI-based intelligent information, referred to as the 4th Industrial Revolution, has now matured to a stage where it can be perceptibly felt. In the midst of these changes, we will need systems and capabilities to support them to be used well in problem solving, based on a correct understanding of the possibilities and limitations of generative AI.

This study has limitations, notably that it exclusively focused on Korean news and conducted a comprehensive analysis without distinguishing between news categories. Additionally, the data collection period was relatively short, spanning about ten months. Nevertheless, this study is significant in that it provides an initial search for user technology adoption and technology prediction research of generative AI. We expect that this study can be used as a precursor for future research exploring and analyzing the long-term perspective of innovation diffusion of generative AI.

Reference

[1] The Guardian, ChatGPT Reaches 100 Million Users Two Months after Launch [Internet]. *https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app.*

[2] N. F. Liu, T. Zhang, and P. Liang, "Evaluating Verifiability in Ggenerative Search Engines," *arXiv preprint arXiv:2304.09848*, 2023. DOI: https://doi.org/10.48550/arXiv.2304.09848

[3] J. Y. Lee, "A Study on Metaverse Hype for Sustainable Growth," *International Journal of Advanced Smart Convergence (IJASC),* Vol. 10, No. 3, pp. 72-80, 2021. DOI: https://doi.org/10.7236/IJASC.2021.10.3.72

[4] J. Y. Lee, *Exploring Hype Dynamics of Virtual Reality and Augmented Reality for Predicting Success Factors of Metaverse*, Ph. D. Thesis. Graduate School, Yonsei University, Seoul, Korea, 2022.

[5] J. Y. Lee, "Deep Learning Research Trend Analysis using Text Mining," *International Journal of Advanced Culture Technology (IJACT),* Vol. 7, No. 4, pp. 295-301, December 2019. DOI: https://doi.org/10.17703/IJACT.2019.7.4.295

[6] J. Y. Lee, "A Study on Research Trend Analysis and Topic Class Prediction of Digital Transformation using Text Mining," *International Journal of Advanced Smart Convergence (IJASC),* Vol. 8, No. 2, pp. 183-190, 2019. DOI: https://doi.org/10.7236/IJASC.2019.8.2.183

[7] A. Amado, P. Cortez, P. Rita, and S. Moro, "Research Trends on Big Data in Marketing: A Text Mining and Topic Modeling based Literature Analysis," *European Research on Management and Business Economics*, Vol. 24, No. 1, pp. 1-7, 2018. DOI: https://doi.org/10.1016/j.iedeen.2017.06.002

[8] K. Bastani, H. Namavari, and J. Shaffer, "Latent Dirichlet allocation (LDA) for Topic Modeling of the CFPB Consumer Complaints," *Expert Systems with Applications*, Vol. 127, pp. 256-271, 2019. DOI: https://doi.org/10.1016/j.eswa.2019.03.001

[9] R. Alghamdi and K. Alfalqi, "A Survey of Topic Modeling in Text Mining," *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 6, No. 1, pp. 147-153, 2015.

[10] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, "Evaluation Methods for Topic Models," in *Proc. 26th Annual International Conference on Machine Learning (ICML)*, pp. 1105-1112, June 2009. DOI: https://doi.org/10.1145/1553374.1553515

[11] C. Sievert and K. Shirley, "LDAvis: A Method for Visualizing and Interpreting Topics," in *Pro c. the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pp. 63-70, Jun e 27, 2014.