Utilizing Big Data Analysis in Problem Finding of Design Thinking: Deriving Meaning from Climate Change in South Korea Using LDA Topic Modeling

Jiyoung Min[1], Hyesun Lee[2*]

[1]Intelligence Technology Design Major, Division of Design, PhD Candidate, Ewha Womans University, Seoul, Korea

[2]Division of Design, Professor, Ewha Womans University, Seoul, Korea

Abstract

Background The utility of big data analysis has been increasing along with various kinds of big data accumulated by popularized Internet access. When we try to understand users in the human-centered design thinking process, it is needed to narrow the digital divide between users and designers using big data analysis. Hence, this research aims to apply big data analysis and identify its effects from the desk research of problem finding in the design thinking process.

Methods Under the theme of climate change, with a view to understanding the meaning climate change has to people, desk research was conducted using big data analysis. Comments on climate change were collected on YouTube news videos, and preprocessing and Latent Dirichlet Allocation(LDA) topic modeling, which is a big data analytic technique, were executed. Based on the analytic findings, the study integrated and visualized topics from a perspective of design thinking, deriving findings which may be linked to user insights.

Results The research also derived the results of the general naming-visualization with a distribution ratio and a coherence score of each topic as axes. They were addressed together with the subjective analysis of the best comments on each topic, the average length of texts with a ratio of 0.5 or more per topic, the ratio of proper words, and the number of words, generating 10 findings.

Conclusions The study reveals that big data analysis helps to find the direction of the next stage of the design thinking process. Unexpected significance is also derived from a different perspective. In addition, findings from a quantitative basis such as numerical comparison increased the reliability of desk research. This study has significance in that it derives the meanings of user-related topics in desk research using big data analysis from a perspective of design thinking.

Keywords Design Thinking, Desk Research, Problem Finding, LDA Topic Modeling, Climate Change

This work was supported by the Ewha Womans University Research Grant of 2023.

*Corresponding author: Hyesun Lee (lhs@ewha.ac.kr)

Citation: Min, J., & Lee, H. (2023). Utilizing Big Data Analysis in Problem Finding of Design Thinking: Deriving Meaning from Climate Change in South Korea Using LDA Topic Modeling. *Archives of Design Research, 36*(4), 211-221.

http://dx.doi.org/10.15187/adr.2023.11.36.4.211

Received : Jul. 17. 2023 ; **Reviewed :** Aug. 14. 2023 ; **Accepted :** Sep. 04. 2023
pISSN 1226-8046 **eISSN** 2288-2987

1. 서론

1. 1. 연구 배경 및 목적

인터넷과 스마트 기기의 대중화로 기존 정형 데이터에 비해 새로운 형태를 띠는 비정형 데이터가 대량으로 생성되고 있다(Noh, K., Kim, J., Nam, S., Park, S., Choi, C., Kim, M. & Kim, Y., 2020). 비정형 데이터를 포함한 빅데이터는 분석 시 인과관계뿐만 아니라 상관관계 규명이 가능해 다양한 활용 가능성을 지닌다(Kim, 2020).

무티(Mootee, 2019)에 따르면 급진적인 변화가 있는 기술 환경에서 디자인 씽킹은 인간 중심 접근법이자 전략적 혁신 프레임워크로서 가치를 창출할 수 있다. 모르타티, 마지스트레티, 카우텔라, 델 에라(Mortati, Magistretti, Cautela & Dell'Era, 2023)는 디자인 과정에서의 데이터 활용은 불확실성 대처와 실제적 이해에 도움을 줄 수 있고, 빅데이터를 활용해 혁신을 구현하는 관행이 부족하다고 언급하였다. 린(Lin, 2018)은 데이터의 규모가 커지고 형태가 복잡해짐에 따라 디자이너와 사용자 간 정보의 격차가 생성된다고 언급하였다. 디자인 씽킹 과정에서 사용자를 이해하기 위해 데이터의 활용이 필요한 시점이다. 이에 따라 디자인 씽킹의 문제 발견 과정에서 보다 심층적이고 본질적인 문제를 파악하기 위하여 빅데이터 분석 활용 연구가 필요하다.

본 연구에서는 디자인 씽킹 과정 중 문제 발견 과정에서 사용자를 이해하는 방법으로 빅데이터 분석을 활용하고자 한다. 익명성으로 인해 표현이 정제되지 않은 온라인상의 비정형 데이터 분석을 통해서 사용자를 이해하고자 한다. 비정형 데이터 수집 및 LDA 토픽 모델링을 활용하여 사용자의 의견을 분석하고 이를 토대로 디자인 인사이트를 정리해 봄으로써 디자인 씽킹을 수행하는 데에 빅데이터 활용이 발생시키는 효과를 살펴보고자 하는 것이 목적이다.

1. 2. 연구 범위 및 방법

디자인 씽킹의 문제 발견 과정에서 빅데이터 분석을 활용하기 위하여 텍스트 데이터 분석을 통해 기후변화에 관한 대중의 생각을 파악하였다. 연구 과정은 다음과 같다. 첫째, 기후변화에 관한 데스크 리서치를 위한 빅데이터 분석 계획을 설정한다. 둘째, 유튜브(Youtube) 영상의 댓글에 해당하는 비정형 데이터를 수집하고 가공하여 LDA 토픽 모델링을 실행한다. 셋째, 빅데이터 분석 결과를 바탕으로 디자인 씽킹 과정에 대한 분석을 진행하여 발견점을 도출한다. 넷째, 디자인 씽킹의 문제 발견 과정에서 빅데이터 분석 활용의 효과를 파악한다.

2. 이론적 배경

2. 1. 디자인 씽킹의 문제 발견 과정과 데스크 리서치

브라운(Brown, 2008)에 따르면 디자인 씽킹은 사람들의 니즈를 기술 실현과 시장 기회 전환이 가능한 비즈니스 해결책에 연결시킨다.

영국 디자인 문화원(Design Council, n.d.)에서는 디자인 씽킹 과정을 시각적으로 표현한 더블 다이아몬드(Double Diamond) 모델을 제시하였다. 문제 발견 과정은 조사와 공감을 통해 디자인 문제와 관련된 사용자를 이해하는 과정이다. 디자인 씽킹 과정에서 문제 질문부터 사용자의 요구 파악을 위한 조사가 실행되는 발견하기(Discover)까지의 과정이 문제 발견 과정에 해당되며 데스크 리서치(Desk research)와 직접조사인 필드 리서치(Field research)가 실행된다(Min & Lee, 2022). 비아나, 비아나, 알더, 루세나, 루소(Vianna, Vianna, Adler, Lucena & Russo, 2012)는 디자인 씽킹 과정의 데스크 리서치는 웹 사이트, 문헌과 같은 정보원에서 주제 관련 정보를 찾는 것으로, 디자인 초기에 주제에 관한 관점과 범위에 대한 이해를 얻으며 주목할 항목의 발굴 및 상호 관계에 대한 인식이 중요하다고 하였다. 민지영, 이혜선(Min & Lee, 2022)의 연구에서 실제 디자이너들이 디자인 씽킹의 문제 발견 과정에서 실행하는 데스크 리서치의 내용을 도출하였고, 인간 중심 정보로는 사람들의 인식과 변화, 사람들에게 주제가 주는 의미, 관련된 사람들에 대한 정보, 사람들의 주제 관련 행동과 이유, 페인포인트의 추정이 있다.

2. 2. 빅데이터 분석과 토픽 모델링

빅데이터 분석은 비정형 데이터를 분석하는 텍스트 마이닝을 포괄하는 개념이며 새로운 통찰을 위해 대량의 데이터에서 패턴과 정보를 파악하는 수학적, 과학적 방법이다(Joo, Kim & Kim, 2021). 본 연구에서 실행하는 빅데이터 분석 기법은 토픽 모델링이다. 텍스트에서 많이 등장하는 단어의 빈도 분석을 통해 주요점이 압축적으로 제시되는 것을 핵심어라 한다(Yoon & Lee, 2018). 토픽 모델링은 텍스트 마이닝 기법 중 많이 활용되는 기법에 해당하며 다양한 문서 집합에 내재된 주제를 파악하기 위해 사용된다(Park & Kang, 2023). 핵심어를 바탕으로 문서에서 주제인 토픽을 추출하는 확률 모델 알고리즘이며 단어를 파악함으로써 잠재된 토픽과 단어의 드러난 정도에 대한 계량적 추론이 가능하다(Yoon & Lee, 2018).

3. 연구 과정

3. 1. 설계

디자인 주제는 '기후변화'로 설정하였다. 민지영, 이혜선(Min & Lee, 2022)의 논문의 데스크 리서치 내용에서 같은 주제에 대한 의미 및 반응이 상이한 그룹이 있음을 알 수 있었고, 이를 바탕으로 선행 연구에서 활용하지 않았던 새로운 빅데이터 분석법을 활용하여 비정형 데이터가 내포하는 가능성을 탐색하고자 하였다. 즉, 비정형 데이터의 분석을 통해 '기후변화는 사람들에게 어떤 의미를 갖는가'라는 질문에 관련된 발견점을 제시하고자 한다. 본 연구에서 발견점은 디자인 씽킹의 문제 정의 과정에서 인사이트로 연계될 가능성이 있으며 문제 발견 과정에서의 사용자, 주제와 관련된 중간 해석이다. 소셜 네트워킹 사이트 유튜브(Youtube)는 수억 명 이상의 사용자를 보유한 동영상 공유 사이트이다(Lee, Osop, Goh, & Kelni, 2017). 유튜브 영상의 댓글은 콘텐츠와 다른 형태의 소통이며, 사람들은 댓글을 통해 영상에 대한 직접적인 반응과 더불어 자신의 감정, 떠오른 추억, 조언, 정신적인 지지를 표현한다(Madden, Ruthven & McMenemy, 2013). 이에 따라 본 연구에서는 유튜브 영상의 댓글을 데이터로 하여 수집과 분석을 통해 디자인 씽킹 과정에 활용하기로 한다.

3. 2. 빅데이터 분석 실행

3. 2. 1. 데이터 수집

본 연구의 비정형 데이터는 한국의 유튜브 영상 댓글에 해당하는 텍스트 데이터로 설정하였다. 데이터 수집은 2023년 4월 3일에 실행했으며 2020년부터 2023년까지의 다양한 기간에 한국어로 게재된 영상을 선택해 수집일 기준으로 2년 전부터 하루 전까지 한국어로 작성된 댓글 총 12,690개를 수집하였다. 유튜브 영상은 '기후 위기 뉴스', '기후변화' 등의 검색어로 검색하여, 기후변화로 고통을 받고 있는 사람들의 상황과 환경에 대한 심각성을 알려주는 내용을 모두 포함하는 총 17개의 한국어 뉴스 영상을 선정하였다. 유튜브 영상의 댓글은 파이썬(Python)의 셀레니움(Selenium) 모듈을 사용해 파이썬 크롤러를 코딩하여 수집하였다. 댓글의 댓글은 콘텐츠와 관련이 적은 반복적 논쟁이 포함되므로 수집 대상에서 제외하였다.

3. 2. 2. 텍스트 데이터 전처리

수집한 텍스트 데이터는 한 번에 모아 정렬했고 불필요한 댓글을 삭제하여 총 12,370개의 댓글을 최종 분석 데이터로 삼았다. 제외된 댓글에는 의미 없는 언어의 반복적 댓글, 앵커 혹은 정치인 등이 언급되며 콘텐츠 혹은 기후변화와 관련이 없는 것 등이 포함되었다. 한국어 형태의 텍스트 전처리는 영어 텍스트와 다른 형태소 분석 과정이 필요하며 〈Table 1〉과 같은 과정을 거쳤다. 한국어 형태소 사전으로는 소셜 미디어상의 짧은 텍스트에 적합하고 비교적 짧은 분석 시간이 요구되는 okt 사전을 사용하였다.

Table 1 Text data preprocessing process

전처리 과정	과정 설명	예시
단어 통일	같은 의미의 단어를 일치시켜줌	'사람'='인간', '기후 변화'〉'기후변화'
맞춤법 교정	띄어쓰기 및 맞춤법 교정	'우리기'〉'우리가'
불용어 선정	분석에 쓰지 않을 단어를 제외	'합니다' 등 제외
사전 신조어 추가	형태소 사전에 기존에 존재하지 않은 단어를 입력	'탄소중립', '라이프스타일' 등 추가
형태소 분석	형태소 사전으로 문장을 분리해 단어의 품사를 부여	'기후변화':'Noun'

3. 2. 3. 데이터 분석

파이썬을 활용해 LDA 토픽 모델링을 실시하였다. LDA 토픽 모델링은 토픽 모델링 중 널리 사용되는 기본적인 알고리즘이며, 문서는 몇 개의 토픽으로 이루어지고 각 토픽은 단어 집합으로 이루어짐을 가정하여 내재된 토픽을 유추하는 방법이다(Park & Kang, 2023). LDA 토픽 모델링 과정에서 전처리 과정을 거친 토큰화된 데이터를 bag-of-words 표현으로 변환하고 빈도-역문서 빈도인 TF-IDF(Term Frequency-Inverse Document Frequency) 가중치값을 부여하였다. 이러한 과정으로 벡터화된 코퍼스는 LDA 모델에서 입력값으로 사용하였다. LDA 토픽 모델링, 시각화에 파이썬 모듈 중에서 각각 gensim, pyLDAvis를 활용하였다.

4. 빅데이터 분석 결과

4. 1. 토픽 개수와 주제별 토픽어

토픽 개수에 따른 coherence score와 perplexity score를 고려하여 토픽 개수를 6개로 선정하였다. LDA 토픽 모델링에서 응집성 지수인 coherence score값은 높을수록 의미론적 일관성이 높다는 것을 의미한다(Newman, Lau, Grieser & Baldwin, 2010; Kang & Lim, 2020). 혼잡도인 perplexity score는 엔트로피를 모델링하여, 낮을수록 안정적으로 구성된 토픽으로 무질서도가 낮음을 의미한다(Lee & Yi, 2021). 〈Figure 1〉에 따르면 토픽 개수가 증가함에 따라 perplexity score는 계속 감소하지만, coherence score는 토픽 개수가 6개일 경우 급격히 증가한다. 더불어 토픽 개수가 10개 이상일 경우에는 시각화 결과의 그룹이 겹치는 것이 많기에 최종 토픽 개수를 6개로 선정하였다.

토픽 개수를 6개로 하여 분석한 시각화 결과는 〈Figure 2〉와 같다. 각 토픽별 coherence score와 가중치값이 큰 상위 20개의 토픽어 도출 결과는 〈Table 2〉와 같으며 그 중 상위 10개는 굵은 글씨체로 표현하였다. 〈Table 2〉의 시각화 결과에서 원의 상대적 크기값은 전체에서 각 토픽이 차지하는 비율을 나타낸다. 각 문서에서 토픽 분포를 합산해 문서의 총수로 나눈 것을 소수점 다섯째 자리에서 반올림한 값이다.

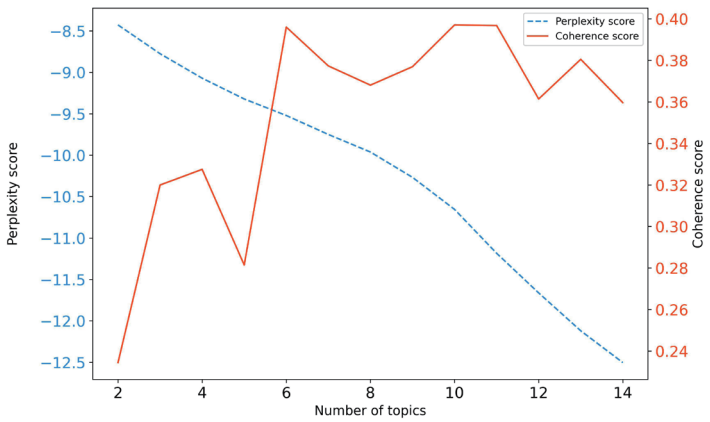

Figure 1 Perplexity score & Coherence score by number of topics

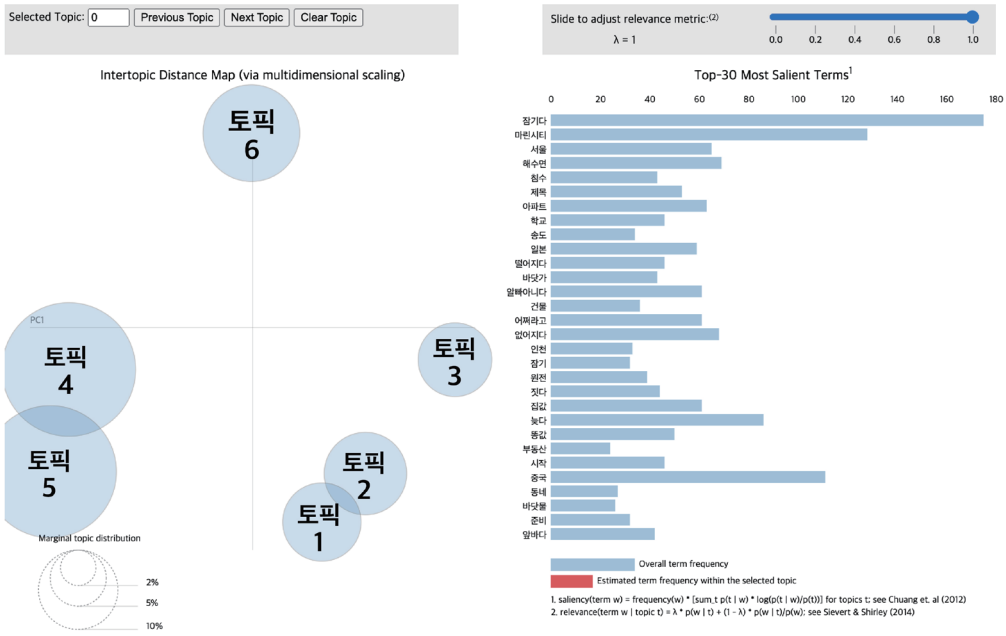

Figure 2 LDA topic modeling visualization

Table 2 LDA topic modeling results

토픽 번호 (Coherence score)	상위 토픽어	시각화 결과의 원의 상대적 크기값
Topic 1 (0.6050)	**시작, 원전, 부동산, 바닷물, 긋다, 망하다, 준비, 워터월드, 빨리, 오다,** 탄소중립, 윤석열, 떠나다, 버티다, 집다, 바다, 지구, 바람, 허다, 질문	0.0999
Topic 2 (0.5804)	**바닷가, 건물, 짓다, 지역, 방파제, 세금, 지구, 침몰, 초등학교, 원래,** 집값, 이제, 아파트, 헐다, 도망가다, 보도, 결혼, 누가, 보다, 푸틴	0.1104
Topic 3 (0.4679)	**침수, 잠기다, 일본, 송도, 학교, 떨어지다, 잠기, 집값, 지대, 사다,** 상관, 팔고, 바다, 이름, 강원도, 경상도, 큰일, 지은, 낯다, 가라앉다	0.0924
Topic 4 (0.4506)	**지구, 인간, 없다, 걱정, 무섭다, 알 바 아니다, 보다, 살다, 가다, 똥값,** 지구온난화, 인류, 좋다, 아니다, 앞바다, 지금, 대책, 도시, 이사, 어차피	0.2774
Topic 5 (0.4118)	**마린시티, 중국, 없다, 늦다, 아니다, 없어지다, 이미, 인간, 지구, 괜찮다,** 문제, 나라, 해결, 죽다, 인구, 사라지다, 자다, 많다, 보다, 줄이다	0.2678
Topic 6 (0.3297)	**잠기다, 서울, 해수면, 아파트, 제목, 어쩌라고, 인천, 상승, 동네, 고층,** 녹다, 빙하, 그렇다, 아니다, 강남, 땅값, 한국인, 살다, 인간, 보다	0.1522

4. 2. 토픽별 댓글 결과

토픽별로 관련성(relevance)이 높은 상위 댓글 20개를 도출하였고 가장 상위 첫 번째 댓글은 〈Table 3〉에 기재하였다. 관련성이 높은 댓글은 해당 주제와 관련이 많으며 각 토픽에 해당하는 비율(per-contribution)이 높은 상위 댓글에 해당한다. per-contribution은 각 댓글에서 특정 주제에 할당된 확률이다. 하나의 댓글은 두 개 이상의 주제에 할당되므로 per-contribution이 과반의 확률에 해당될 때, 해당 토픽에 주로 기여한다고 볼 수 있다. 이에 따라 각 토픽에서 per-contribution값이 0.5 이상인 데이터에서 텍스트의 길이, 어휘의 다양성을 나타내는 중복되지 않는 고유한 단어들의 비율, 단어 수를 측정해 평균값을 내었다. 이는 토픽별 텍스트의 길이, 복잡성, 정보성을 의미할 수 있으며 〈Table 3〉에서 소수점 넷째 자리까지 제시하였다. 평균 고유 단어 비율은 토픽 2가 가장 크고, 토픽 3은 평균 길이와 평균 단어수가 가장 작으며, 토픽 5는 평균 길이와 평균 단어수가 가장 크고 평균 고유 단어 비율이 가장 작은 값을 가졌다.

Table 3 Top comment with high “Relevance” values & characteristics

토픽	첫 번째 상위 댓글 (Per-contribution)	평균 길이	평균 고유 단어 비율	평균 단어 수
1	쓰레기 분리수거 하는거며 폐수버리는 대기업,중소기업만 제대로 엄벌부터 해야된다 나무심는거 한달에 한그루 무조건 심게 하자 그리고 자연 훼손 시 벌금 씨게 ㄱㄱㄱㄱ길에 담배꽁초 일회용 플라스틱컵 버릴시 벌금 100씩 ㄱㄱ(**0.7708**)	19.2184	0.9634	5.9184
2	진짜 세상 끝이다ㄷㄷ 성경에서만 보던게 눈앞에 펼쳐지니까 정말 소름돋는다... 좀 진지하게 죽고나서를 생각해볼 필요가 있는 듯(**0.7642**)	20.2461	0.9774	6.1763
3	이북은 산불이 없는지 그러면 현재 뿔이 난 이북 소행. 조사 일일이 비디오 장치를 해야함.(**0.7307**)	16.3598	0.9633	5.0819
4	아니 지구가 주는 마지막 경고 무시하지마세요 기계로 해결할라하지마세요 전기를 아끼세요 쓰레기를 함부로 버리지마세요 물도 막 틀지마세요 나 하나쯤이야라는 생각을 버리세요 그렇게 생각하는 나 하나쯤이야라는 생각으로 이세계 생물들에게 고통을 주고 그 고통이 우리에게 돌아옵니다. 그렇게 두려워하던 코로나도 인간에 의해 생긴것입니다. 제발 잘못을 한 후에 "지구가 멸망하는건가..? 너무 두려워.."라 생각하지마시고 실천을 한 후에 그런 걱정을해주세요.부탁드립니다.(**0.8393**)	27.7074	0.9578	8.3921
5	고기 적당히 먹어야돼 우리... 고기를 한명당 한달에 얼마치만 살수있게 할당량정해놓고 안먹을 사람은 먹을 권리를 돈받고 양도할수있게 하면 좋겠는데 물론 말처럼 쉽진 않겠지? ㅠ ㅋㅋ 미국전체영토 1/3정도가 소키우는데 들어가고 그거의 반정도가 소 여물 키우는데 들어가는 농경지래. 실제로 사람이 먹을 농산물 경작지가 더 작아 미국만 그런거 아니고 많이들그래. 온실가스 배출중에 정말 큰 역할을 하는게 소방구랑 소 트림이래(**0.8476**)	28.7559	0.9563	8.7124
6	그걸 모르는 사람은 없지. 근데 어차피 다들 못 막을거 알면서도 바뀌어야 한다. 탄소 줄여야한다. 이러고만 있을 뿐임. 목마르다고 소금물 마시면 죽는거 뻔히 알면서도 계속 물 들이키고 있는꼴임.(**0.7982**)	21.7098	0.9673	6.5917

5. 디자인 관점 분석

5. 1. 종합 분석

토픽 1은 상위 토픽어에 따르면 ‘원전’, ‘바닷물’, ‘부동산’ 등 다양한 이슈와 종말의 ‘시작’과 관련된 ‘준비’에 관한 내용에 해당한다. ‘-그니까 석유 가스 쓰지말고 원전쓰라고-’, ‘진짜 무서운 건 이건 시작에 불과하다는 거지’, ‘-경각심을 가지면 뭐해 근본을 때려잡아야지.’ 등의 댓글이 이를 뒷받침한다. 관련성이 높은 상위 댓글에서도 ‘원전 늘리고 석탄, 가스발전 줄이면서 친환경 늘리는 게 현실적인 대안 아닌가요?’, ‘-대기업, 중소기업만 제대로 엄벌부터 해야 한다-’ 등을 통해 토픽 1에 해당하는 사람들은 대안을 찾으려 노력하며 사회, 정책적 문제에 대한 의견을 제시하였음을 알 수 있다. 이는 토픽 1의 coherence score값이 0.6050으로 가장 커 일관적이었던 것과 상응한다. 토픽 2는 상위 토픽어에 따르면 ‘바닷가’ 주위 ‘지역’에 ‘건물’의 ‘침몰’에 대한 내용이 해당된다. 이는 한국의 인천, 부산 등의 바닷가 지역 침몰에 관한 내용이 유튜브 영상에 대부분 포함되었기 때문이다. 관련성 상위 댓글 내용에서는 생존에 위협을 느끼며 근시안적인 당장의 이슈를 논하기보다 종말에 이르렀다는 생각을 제시함을 알 수 있었다. ‘-세상 끝이다.-’. ‘지구온난화 심각해지면 고원지대로 피신해야합니다.-’, ‘-이러다간- 몰살당한다-’ 등이 이를 뒷받침한다. 평균 고유 단어 비율이 0.9774로 가장 크므로 가장 다양한 어휘를 사용했고 시각화 결과에서 토픽 1과 거리가 가까워 유사 내용을 포함하고 있음을 알 수 있다. 토픽 1보다 비교적 감정이 격한 표현을 사용하지만 기후변화에 대한 무서움에 그치는 것이 아니라 토픽 1과 같이 이에 대한 대안을 주로 함께 언급한다. 토픽 3은 상위 토픽어에 따르면 ‘침수’와 관련된 ‘집값’, ‘지대’에 관한 내용이 해당한다. 관련성 상위 댓글 내용에서는 기후변화에 관련된 이유나 구체적 행동이 아닌 다른 이야기에 대해 주로 언급하였다. 이는 평균 길이값이 가장 작았던 것과 상응한다. 토픽 4는 ‘인간’으로 인해 시작된 ‘지구’에 관한 내용이 주로 언급된다. ‘대멸종의 시작은 인간이다.’, ‘-지구 자정 작용하는 날 인류가 사라질 것이다.-’ 등의 댓글이 이를 뒷받침한다. 관련성 상위 댓글 내용에서는 인간에 대한 회의감과 체념이 나타난다. ‘잔인한 소리겠지만 인구 반정도가 되어야 지구는 원 상태로 되돌아갈

수 있을 듯 싶네요. 인간은 너무 이기적임.-', '지구가 주는 마지막 경고 무시하지 마세요.-' 등으로 인해 알 수 있다. 시각화 결과에서 상대적 원의 크기가 0.2774로 제일 크므로 가장 많은 분포 비율의 내용에 해당한다. 토픽 5은 주요 토픽어에 따르면 '이미' 한계라는 것을 직시하며 '중국'을 포함한 내용이 언급된다. '이미-노력한다고 해도 별 의미 없을 수도 있다-', '-중국처럼 거대 대륙이 발전을 시작했다- 제동을 걸 수조차 없다-' 등의 댓글이 이를 뒷받침한다. 관련성 상위 댓글 내용에 따르면 '-미국전체영토-', '-세계적으로 가축 사료작물 재배-', '중국 인도의 생활수준이 올라가면 갈수록-' 등의 댓글을 통해 세계적이고 구체적인 내용에 대해 언급함을 알 수 있다. 평균 문장의 길이와 단어 수가 가장 크지만 고유 단어 비율이 작아 댓글 안에서 구체적으로 길게 서술한다고 볼 수 있다. 토픽 6은 토픽어에 따르면 '해수면'의 '상승'으로 인한 '서울'의 '아파트', '동네'의 내용을 포함하지만 coherence score값이 가장 작으므로 단어보다는 상위 댓글 내용에서 토픽의 주요 내용을 파악할 수 있다. '-에어컨 조금만 틀고 일회용품 대신 다회용품을 애용해주세요.-', '제발 지금부터라도 친환경 제품으로 바꿔보고-.' 등의 댓글에서 주변 생활과 관련된 내용을 언급함을 알 수 있다. 각 토픽에 대한 종합 명명은 〈Table 4〉와 같다.

Table 4 Synthesis and analysis of results

토픽	주요 토픽어 기반	관련성 상위 댓글 내용 기반	토픽 주요 특성
1	원전, 탄소중립, 부동산 등 다양한 대안적 이슈와 관련 내용	대안을 찾으려 노력하며 사회적, 정책적 문제에 대한 의견을 제시한다.	- 토픽의 내용이 일관적이다.
	종합 명명		
	환경 중요성을 인지하고 사회적, 정책적 이슈에 관해 이야기한다.		
2	바닷가 주위 지역의 건물 침몰 관련 내용	생존에 위협을 느끼며 근시안적인 당장의 이슈보다 종말에 이른 생각을 제시한다.	-시각화 결과에서 토픽 1과 거리가 가장 가깝다.
	종합 명명		
	기후변화에 대해 공포를 느끼며 이에 관한 대안을 제시한다.		
3	침수와 연관된 집값, 지대 관련 내용	기후에 관련된 행동보다는 오히려 다른 이야기를 한다.	- 평균 길이값이 가장 짧다.
	종합 명명		
	무서움으로 인해 거리감이 있는 이야기를 가볍게 제시한다.		
4	인간으로 인해 시작된 지구 관련 내용	인간에 대한 회의를 가지고 있으며 체념적이다.	- 시각화 결과에서 상대적 크기값이 가장 크다. - 시각화 결과에서 토픽 5와 거리가 가깝다.
	종합 명명		
	해결책과는 다른 형태의 인간, 지구에 대한 구체적 이야기를 한다.		
5	이미 한계라는 것을 직시하는 중국 관련 내용	세계적이고 구체적인 생각에 대해 제시한다.	- 평균 길이값이 가장 크다. - 평균 고유 비율값이 가장 작다 - 평균 단어 수가 가장 길다.
	종합 명명		
	국제적인 환경 관련 이슈에 대해 비교적 자세히 언급한다.		
6	해수면 잠기는 것에서 서울 집값 관련 내용	에어컨, 일회용품 등의 주변 생활에 대한 언급을 한다.	- coherence score 값이 가장 작다.
	종합 명명		
	주변 실생활과 관련된 환경 관련 행동에 대해 언급한다.		

5. 2. 발견점 도출

〈Table 2〉의 coherence score값에 해당하는 응집성 지수, 원의 상대적 크기값에 해당하는 분포 비율을 각각 x, y축으로 하여 〈Figure 3〉과 같이 토픽 종합을 시각적으로 배치하였다. 또한 〈Table 4〉의 평균 길이, 평균 고유 단어 비율, 평균 단어 수를 토픽 1에서 3개의 값이 같도록 표준화하여 막대의 길이로 그 비율을 표현해 그래프로 시각화하였다. 상위 20개의 댓글에서 토픽 1은 기후변화 대응에 대한 생각을 발산적으로, 토픽 2는 종말에 이르는 생각을 비교적 격한 감정으로, 토픽 3은 환경과는 직접적 관련이 적은 내용을 가볍게, 토픽 4는

인간의 행동에 대해 회의적으로, 토픽 5는 국제적 이슈에 대해 비판적으로, 토픽 6은 실생활에서의 환경 관련 행동 촉구를 설득적으로, 내용을 주로 서술하는 것으로 판단할 수 있었다. 이와 더불어 〈Table 4〉, 〈Figure 3〉을 통해서 〈Table 5〉와 같이 발견점 10개를 도출하였다. 토픽 2와 토픽 3의 종합 명명에 따라 발견점 1이 도출되었다. 토픽 3에서 상위 토픽어를 통한 종합은 주로 침몰에 관한 영상 관련 내용에 해당되지만, 상위 댓글은 영상과 관련이 적은 이야기로 짧게 구성되었다. 이와 같은 차이로 인해 발견점 2가 도출되었다. 토픽 4, 토픽 5의 분포 비율이 높다는 점에서 발견점 3이 파악되었고, 토픽 4의 분포 비율이 토픽 1의 분포 비율보다 높아 발견점 4가 도출되었다. 해결책보다는 환경 이슈에 대해 많이 이야기하는 토픽 4, 토픽 5의 분포 비율이 해결안에 대해 주로 이야기하는 토픽 1, 토픽 2의 분포 비율보다 높아 발견점 5가 도출되었다. 토픽 4, 토픽 5에서는 기후변화에 대한 해결에 대해 노력하기보다는 회의적이고 비판적이므로 발견점 6이 파악되었다. 〈Table 2〉에서 토픽 2의 '집값', 토픽 3의 '집값', 토픽 4의 '똥값', 토픽 6의 '땅값'이 도출된 것처럼 사람들은 기후변화를 경제와 연관지어서 생각함을 알 수 있다. 오히려 친환경 제품, 에어컨 사용 줄이기와 같이 생활 방식에 관련된 내용인 토픽 6은 응집성 지수도 낮고 분포 비율도 높지 않았다. 이에 따라 발견점 7이 도출되었다. 토픽 1, 토픽 2의 경우 사회 및 정책적, 비현실적과 같이 방향은 다르지만 해결하기 위한 생각이 댓글에서 표현되었기에 발견점 8이 도출되었다. 특히 토픽 1의 경우 일관성이 크고 댓글의 서술 형태가 발산적이므로 발견점 9가 파악되었다. 발견점 10의 사회 및 정책적, 비현실적, 현실 도피적, 범지구적, 국제적, 라이프스타일 중심적은 각각 토픽 1, 토픽 2, 토픽 3, 토픽 4, 토픽 5, 토픽 6에서의 사람들의 관점으로 인해 파생되었다.

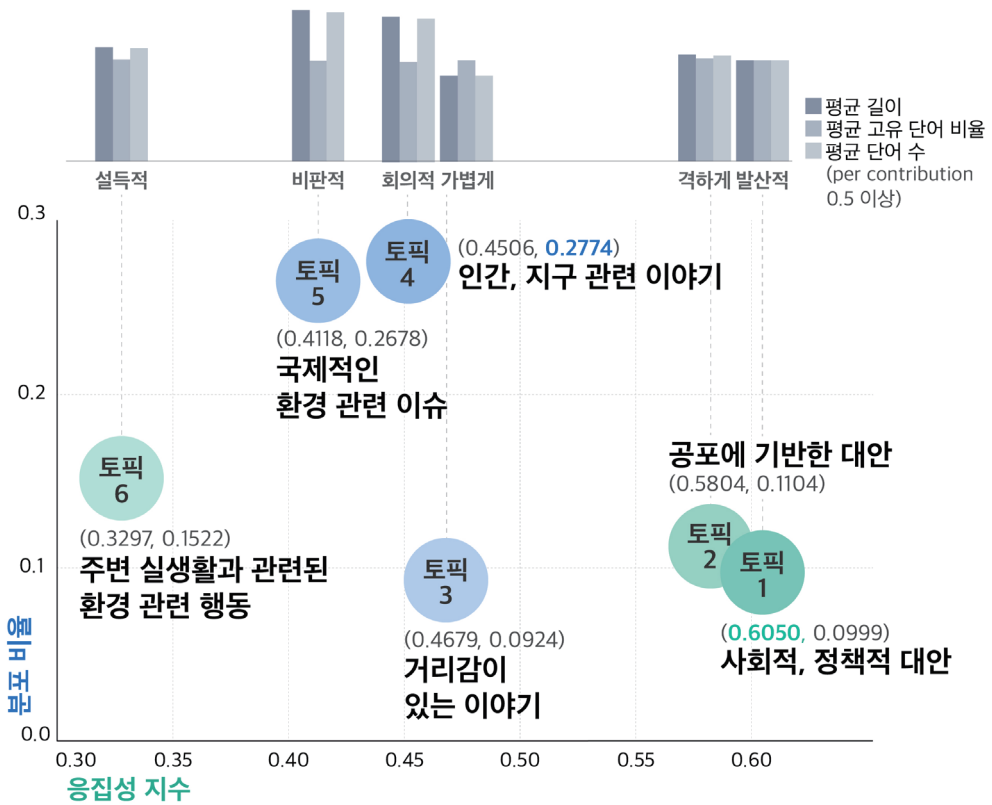

Figure 3 Topics in new matrix for searching findings

Table 5 Findings

발견점	내용	주요 관련 토픽
발견점 1	기후변화에 대해 무서움을 느끼는 사람이 많고, 해결책을 모르는 것에서 생존 관련 공포를 느낀다.	2, 3
발견점 2	사람들에게는 기후변화와 관련해 현실 도피를 하고 싶은 마음이 내재되어 있다.	3
발견점 3	사람들은 환경 이슈에 대해 비교적 많이 인지하고 있다.	4, 5
발견점 4	환경 관련 국내 사회적, 정책적 이슈보다 국제적 이슈에 관해 표현하는 사람이 많다.	1, 5
발견점 5	환경 이슈 인지도에 비해 대안을 제시하며 진지하게 노력하는 사람은 적다.	1, 2, 4, 5
발견점 6	해결책 부재에 관한 회의감과 답답함을 느끼는 사람이 많다.	4, 5
발견점 7	사람들은 환경과 직접적으로 관련된 친환경 행동 실천보다 집값과 관련된 생계에 직결된 것에 더 관심이 많다.	2, 3, 4, 6
발견점 8	사람들은 경각심 함양을 위한 예측보다 예방 방법에 대해 알기를 원한다.	1, 2
발견점 9	스스로 기후변화 예방에 관해 발산적으로 의견을 내는 사람들이 존재한다.	1
발견점 10	사람들은 기후변화에 대해 사회 및 정책, 국제적, 비현실적, 범지구적, 라이프스타일 중심적, 현실 도피적과 같이 다양한 차원에서 생각한다.	1, 2, 3, 4, 5, 6

6. 결론 및 제언

본 논문은 디자인 씽킹 과정에서 문제 발견 과정의 데스크 리서치 과정 중 대량의 텍스트 데이터를 분석하여 활용하였다. 이를 통해 기후변화의 의미와 관련된 사람들의 생각에 해당하는 발견점 10개를 도출하였다. 본 논문에서의 빅데이터 분석 활용은 디자인 씽킹과 관련되어 다음과 같은 효과를 가진다. 첫째, 디자인 씽킹의 초기 단계에서 방향성을 찾는 데에 도움을 준다. 기후변화와 관련된 지속가능한 디자인과 같이 범위가 크고 모호한 문제에서 사용자 리서치를 비롯한 디자인 씽킹 과정의 다음 실행 단계에 대한 방향과 키포인트를 제공한다. 예를 들어, 발견점 3에 집중해 현실 도피하고 싶은 사람들의 심리에 대해 더 알아보기 위한 사용자 리서치로, 현실 도피에 관련된 경험에 관한 인터뷰를 실행할 수 있다. 이를 통해 기존과 다른 새로운 시각으로 기후변화라는 문제를 해결해 나갈 수 있다. 둘째, 기존 디자인 씽킹의 데스크 리서치 과정에서는 비교적 알기 힘든, 주제가 사람에게 주는 다른 의미를 도출할 수 있다. 기존의 데스크 리서치에서는 기후변화에 대한 사람들의 생각을 알기 위해서는 반복된 데스크 리서치와 추측이 필요하고 특히, '기후변화는 사람들에게 어떤 의미를 갖는가'에 대한 양적 통계 조사가 이전에 실행되지 않았다면 파악이 쉽지 않다. 설문으로 인한 통계 데이터가 존재하더라도 과거 시점의 데이터는 디자인 프로젝트의 초기 단계의 데스크 리서치에 적합하지 않을 수 있다. 그러나 본 논문에서는 원하는 시기의 대량 빅데이터 가공 및 분석을 통해 주제가 사람들에게 주는 의미의 파악이 용이하였다. 예를 들어, 발견점 1에서 기후변화는 죽음에 대한 공포와 공통적인 부분이 있는 공포를 제공한다는 것을 파악할 수 있다. 죽음에 대한 공포를 웰리빙으로 풀어내듯이, 이후 디자인 씽킹 과정에서 기후변화에 대한 공포를 생각의 전환을 동반해 풀어낼 수 있다. 또한, 발견점 7에 따르면 기후변화가 친환경 생활보다 집값과 같이 경제적인 의미로 인지될 때 사람들의 관심이 증가한다. 기후변화와 사람들에게 영향을 주는 경제를 연결시켜 이후 디자인 씽킹 과정을 실행할 수 있다. 셋째, 수치에 기반한 양적 근거를 통해 신뢰성이 증가한다는 발견점을 도출할 수 있었다. 특히, 〈Figure 3〉과 같이 디자인 씽킹의 문제 발견 과정에서 다양한 수치 비교를 통해 발견점을 도출할 수 있으므로 디자이너의 주관에 신뢰성을 부여한다. 상위 토픽어 종합과 상위 댓글의 일치, 불일치, 응집성 지수, 분포 비율, 원의 크기, 평균 길이, 평균 고유 단어 비율, 평균 단어 수의 종합과 비교는 디자이너의 주관을 뒷받침하는 근거로 작용한다. 윤태일, 이수안(Yoon & Lee, 2018)은 SNS상의 대량의 텍스트로 이루어진 말뭉치인 코퍼스가 존재할 때 그 텍스트를 일일이 읽어 토픽을 다루고 있는지 파악하기 어렵다고 하였다. 기존 디자인 씽킹 과정에서는 유튜브 댓글을 읽으며 사용자 생각의 주요점을 파악하기 힘든 데 반해 본 연구의 빅데이터 분석 활용은 주요성과 경향성 파악을 가능하게 하였다.

본 논문은 디자인 씽킹의 문제 발견 과정에서 사용자 이해를 위해 빅데이터 분석을 활용해 방향성 제시, 의미 도출, 수치적 신뢰성 함양 측면에서 효과적인 리서치를 실행했다는 것에 의의가 있다. 디자인 씽킹의 관점으로

빅데이터 분석을 활용한 리서치를 실행하여 사용자를 이해하고 발견점을 찾은 것에 의미가 있다. 이는 기존의 빅데이터 분석 플랫폼을 통한 일률적 분석 결과 수용이 아니라 자연어 처리, 결과 분석 등 디자인 씽킹 과정에 활용하기 위한 분석을 직접 실행하였기에 가능하였다. 향후 하이퍼파라미터(hyperparameter) 조정, 단어 통일 반복 등 자연어 처리의 정교화, 분석 방법의 다양화를 통해 보다 많은 빅데이터 분석 활용이 디자인 씽킹의 문제 발견 과정에서 실행되길 기대한다. 다만, 도출된 발견점은 정량적인 데이터 분석을 기반으로 하기에 수치적인 부분에서 신뢰성이 있으나 실제 사용자의 인사이트와 연계 시 검증이 필요하다. 예를 들어 유튜브 영상에 댓글을 남기는 사람들은 그 특성이 명확하지 않다는 한계가 있다. 즉, 도출된 발견점은 이후 인터뷰, 관찰 등의 민족지학적인 사용자 리서치 방법을 실행하여 실제 사용자의 니즈와 부합하는지 검증되어야 한다.

References

1. Brown, T. (2008). Design Thinking. *Harvard Business Review, 86*(6), 84.
2. Design Council. (n.d.). Framework for Innovation [Web tutorial post]. Retrieved from https://www.designcouncil.org.uk/our-resources/framework-for-innovation/
3. Joo, H., Kim. H., & Kim, H. (2021). *빅데이터 기획 및 분석 [Big Data Planning and Analytics]*. Seoul: Crown Publishers
4. Kang, H., & Lim, H. (2020). 토픽모델링과 주성분 분석을 활용한 온라인 쇼핑 검색 질의 유형 분류 [A Study on the Types of Online Shopping Queries using Topic Modeling and Principal Components Analysis]. *Proceedings of the Korea Information Processing Society Conference (765-768), 27*(2), Seoul, KIPS.
5. Kim, H. (2020). *빅데이터 활용에 있어서의 개인정보 보호법제에 관한 연구 [A Study on the Personal Information Protection Law in the Utilization of Big Data]* (Unpublished dotoral dissertation). Donga Univesity, Busan, Korea.
6. Lee, C. S., Osop, H., Goh, D. H.-L., & Kelni, G. (2017). Making sense of comments on YouTube educational videos: a self-directed learning perspective. *Online Information Review, 41*(5), 611-625. doi:https://doi.org/10.1108/OIR-09-2016-0274
7. Lee, D., & Yi, H. (2021). LDA 토픽 모델링의 적정 토픽 수 결정 방법 탐색: 혼잡도와 조화평균법 활용을 중심으로 [Exploring methods for determining the appropriate number of topics in LDA: Focusing on perplexity and harmonic mean method]. *Journal of Educational Evaluation, 34*(1), 1-30.
8. Lin, K. -Y. (2018). A Text Mining Approach to Capture User Experience for New Product Development. *International Journal of Industrial Engineering, 25*(1), 108-121.
9. Madden, A., Ruthven, I., & McMenemy, D. (2013). A classification scheme for content analyses of YouTube video comments. *Journal of Documentation, 69*(5), 693-714. doi:https://doi.org/10.1108/JD-06-2012-0078
10. Min, J., & Lee, H. (2022). 빅데이터 분석 활용을 위한 디자인 씽킹의 데스크 리서치 과정 연구 [Study of the Desk Research Process for utilizing Big Data Analysis in Design Thinking]. *Journal of Industrial Design Studies, 16*(4), 71-83. doi:10.37254/ids.2022.12.62.07.71
11. Mootee, I. (2019). *하버드 디자인 씽킹 수업 [Harvard Design Thinking Lessons]*. Busan: UX Review.
12. Mortati, M., Magistretti, S., Cautela, C., & Dell'Era, C. (2023). Data in design: How Big Data and Thick Data Inform Design Thinking Projects. *Technovation, 122*, 1-3.
13. Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic Evaluation of Topic Coherence, *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (100-108), USA, NAACL HLT.
14. Noh, K., Kim, J., Nam, S., Park, S., Choi, C., Kim, M., & Kim, Y. (2020). *빅데이터 분석 기획 [Big Data Analytics for Business]*. Seoul: Wowpass.
15. Park, S., & Kang, J. (2023). *파이썬 텍스트 마이닝 완벽 가이드 [Complete Guide to Python Text Mining]*. Paju: Booknuri
16. Yoon, T., & Lee, S. (2018). *파이썬으로 텍스트 분석하기 [Text Analytics with Python].* Seoul: Neulbom
17. Vianna, M., Vianna, Y., Adler, I. K., Lucena, B., & Russo, B. (2012). *Design thinking : business innovation* (MJV Press ISBN 978-85-65424-01-1). Retrieved from https://www.academia.edu/9204741/business_innovaTion_Design_Thinking)

디자인 씽킹의 문제 발견에서의 빅데이터 분석 활용 연구 - LDA 토픽 모델링을 이용한 한국의 기후변화에 대한 의미 도출

민지영[1], 이혜선[2*]

[1]이화여자대학교 일반대학원 인텔리전스테크놀로지디자인 전공, 박사과정, 서울, 대한민국

[2]이화여자대학교 디자인학부, 교수, 서울, 대한민국

초록

연구배경 인터넷 사용의 대중화로 다양한 형태의 대규모 데이터가 축적됨에 따라 빅데이터 분석의 활용성이 증대되었다. 인간 중심의 디자인 씽킹 과정에서 사용자를 이해할 때 빅데이터 분석을 활용하여 사용자와 디자이너 간의 정보 격차를 줄일 필요가 있다. 이에, 디자인 씽킹 과정 중 문제 발견 과정의 데스크 리서치에서 빅데이터 분석을 활용하고 그 효과를 파악하고자 한다.

연구방법 기후변화를 주제로 사람들에게 기후변화가 갖는 의미가 무엇인지 파악하기 위하여 빅데이터 분석을 활용한 데스크 리서치를 실행하였다. 기후변화와 관련된 유튜브 뉴스 영상의 댓글을 수집하고 전처리 과정을 거쳐 빅데이터 분석 기법인 LDA 토픽 모델링을 실행하였다. 분석 결과를 바탕으로 디자인 씽킹적 관점의 분석으로 토픽을 종합하고 시각화를 거쳐 사용자 인사이트 연계 가능성이 있는 발견점을 도출하였다.

연구결과 각 토픽의 분포 비율과 응집성 지수를 축으로 종합 명명을 시각화하였다. 이를 각 토픽의 상위 댓글 주관적 분석, 각 토픽당 비율이 0.5 이상인 텍스트의 평균 길이, 고유 단어 비율, 단어 수의 수치와 종합 및 비교하여 발견점 10개를 도출하였다.

결론 빅데이터 분석 활용은 디자인 씽킹 다음 단계의 방향성을 찾는 데에 도움을 주는 것을 알 수 있었다. 또한 예상하기 어려운 다른 관점의 의미를 도출하게 하였다. 더불어 수치 비교 등의 양적 근거로 발견점을 도출하여 데스크 리서치의 신뢰성이 증가되었다. 본 연구는 디자인 씽킹의 데스크 리서치 과정에서 빅데이터 분석을 활용해 디자인 씽킹의 관점에서 사용자와 관련된 주제의 의미를 도출한 것에 의의가 있다.

주제어 디자인 씽킹, 데스크 리서치, 문제 발견, LDA 토픽 모델링, 기후변화

이 연구는 2023학년도 이화여자대학교 교내연구비 지원에 의한 연구임.

*교신저자: 이혜선 (lhs@ewha.ac.kr)