creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

외교학석사 학위논문

북한이탈주민 관련 기사에 대한 댓글의 분화

2021년 2월

서울대학교 대학원

정치외교학부 외교학전공

최 자 인

북한이탈주민 관련
기사에 대한 댓글의 분화

지도교수 조 동 준

이 논문을 외교학석사 학위논문으로 제출함
2021년 2월

서울대학교 대학원
정치외교학부 외교학전공
최 자 인

최자인의 외교학석사 학위논문을 인준함
2021년 2월

위 원 장 _____ 임 경 훈 _____ (인)

부위원장 _____ 박 종 희 _____ (인)

위 원 _____ 조 동 준 _____ (인)

국문초록

신문사의 이념적 분화는 독자측의 이념적 분화를 이끌어 낼 수 있는가? 독자가 어떤 신문사를 이용하는지에 따라서 동일한 대상과 사건에 대한 독자의 태도가 서로 다르게 나타나는가? 언론의 틀짓기효과에 대한 고려를 바탕으로, 본 연구는 보수 성향의 신문과 진보 성향의 신문에서 제공하는 기사의 내용에 차이가 존재하는지, 이와 같은 차이가 기사에 달린 댓글의 내용에서도 나타나는지 확인하고자 하는 데에서 출발하였다. 그리고 해당 내용을 확인하기 위해 조선일보, 중앙일보, 한겨레신문, 경향신문의 북한이탈주민에 대한 기사와 댓글이라는 특정 사례를 선택하였고, 방대한 양의 기사와 댓글 자료를 다루기 위해 토픽 분석과 연관어 분석 등을 실시하였다.

그 결과, 보수 성향의 신문과 진보 성향의 신문에서 제공하는 기사들은 몇 가지 공통점과 차이점을 보여주었고, 이와 같은 양상은 보수 성향의 신문사 댓글과 진보 성향의 신문사 댓글에서도 매우 유사하게 나타나고 있었다. 이에 따라 본 연구는 신문사가 특정한 관점이 담긴 기사의 내용을 독자측에 제시함으로써 독자측의 반응을 일정한 방향으로 유도할 수 있다는 점과 혹은 신문사별로 참여하는 독자의 특성이 서로 다르다는 것을 상관관계적인 측면에서 보여줄 수 있었다.

주요어 : 북한이탈주민, 기사, 댓글, 태도, 분화, 텍스트 마이닝
학 번 : 2017-24750

목 차

제 1 장 서론 ····· 1
제 1 절 문제 제기 ····· 1
제 2 절 논문의 구성 ····· 3

제 2 장 선행 연구 검토 ····· 5
제 1 절 기사와 댓글 간 일방향적 관계에 대한 연구 ····· 5
제 2 절 북한이탈주민에 대한 한국 사람들의 태도 연구 7
제 3 절 소셜 미디어 텍스트를 활용한 연구 ····· 9

제 3 장 연구 설계 ····· 14
제 1 절 연구 질문 ····· 14
제 2 절 연구 방법 ····· 19
1. 자료 수집 ····· 19
1) 신문 기사 ····· 19
2) 댓글 ····· 22
2. 자료 전처리 ····· 25
3. 텍스트 마이닝 방법 ····· 27
1) 토픽 분석: LDA 토픽 모델링 ····· 27
2) 연관어 분석: 워드투벡 ····· 30
3) 감성분석 ····· 31
(1) 감성사전을 이용한 감성분석 ····· 33
(2) 지정된 단어를 이용한 감성분석 ····· 35

제 4 장 분석 결과 ····· 36
제 1 절 기사 토픽 분석 결과 ····· 36
1. 신문사의 성향에 따른 기사 토픽 분석 결과 비교 ····· 41
1) 공통점 ····· 41
2) 차이점 ····· 42
제 2 절 댓글 연관어 분석 결과 ····· 44
1. 신문사의 성향에 따른 댓글 연관어 분석 결과 비교 ··· 44
1) 공통점 ····· 44
2) 차이점 ····· 45
제 3 절 댓글 감성분석 결과 ····· 57
1. 감성사전을 이용한 댓글 감성분석 결과 ····· 57
2. 지정된 단어를 이용한 댓글 감성분석 결과 ····· 58

제 5 장 결론 ····· 68

부록 ····· 72
참고문헌 ····· 122
Abstract ····· 130

표 목 차

[표 1] 네 개의 신문 기사 댓글에 사용된 상위 80개 단어 17
[표 2] 기사 수집 ····· 23
[표 3] 댓글 수집 ····· 24
[표 4] 조선일보 기사 토픽 분석 결과 해석 ····· 37
[표 5] 중앙일보 기사 토픽 분석 결과 해석 ····· 38
[표 6] 한겨레신문 기사 토픽 분석 결과 해석 ····· 39
[표 7] 경향신문 기사 토픽 분석 결과 해석 ····· 40
[표 8] 군산대학교 감성사전을 이용한 댓글 감성분석 결과 59
[표 9] 지정된 단어를 이용한 댓글 감성분석 결과 ····· 60
[표 10] ‘북괴’ 단어를 이용한 댓글 카이제곱검정 ····· 61
[표 11] ‘인권’ 단어를 이용한 댓글 카이제곱검정 ····· 62
[표 12] ‘동포’ 단어를 이용한 댓글 카이제곱검정 ····· 63
[표 13] ‘보호’ 단어를 이용한 댓글 카이제곱검정 ····· 64
[표 14] ‘간첩’ 단어를 이용한 댓글 카이제곱검정 ····· 65
[표 15] ‘위장’ 단어를 이용한 댓글 카이제곱검정 ····· 66
[표 16] ‘빨갱이’ 단어를 이용한 댓글 카이제곱검정 ····· 67

부록 [표 1] 조선일보 기사 LDA 토픽 모델링 결과 ····· 76
부록 [표 2] 중앙일보 기사 LDA 토픽 모델링 결과 ····· 81
부록 [표 3] 한겨레신문 기사 LDA 토픽 모델링 결과 ····· 84
부록 [표 4] 경향신문 기사 LDA 토픽 모델링 결과 ····· 87
부록 [표 5] 조선일보 댓글 LDA 토픽 모델링 결과 ····· 94
부록 [표 6] 중앙일보 댓글 LDA 토픽 모델링 결과 ····· 98
부록 [표 7] 한겨레신문 댓글 LDA 토픽 모델링 결과 ····· 102
부록 [표 8] 경향신문 댓글 LDA 토픽 모델링 결과 ····· 106
부록 [표 9] 신문사들 간의 토픽 현저(topic salience) 비교:
조선일보 기사 LDA 토픽 모델링 결과 ····· 110
부록 [표 10] 신문사들 간의 토픽 현저(topic salience) 비교:
중앙일보 기사 LDA 토픽 모델링 결과 ····· 113
부록 [표 11] 신문사들 간의 토픽 현저(topic salience) 비교:
한겨레신문 기사 LDA 토픽 모델링 결과 ····· 116
부록 [표 12] 신문사들 간의 토픽 현저(topic salience) 비교:
경향신문 기사 LDA 토픽 모델링 결과 ····· 119

그 림 목 차

[그림 1-1] 조선일보 댓글 연관어 분석 결과 ····· 49
[그림 1-2] 조선일보 댓글 연관어 분석 결과 부분 확대 ···· 50
[그림 2-1] 중앙일보 댓글 연관어 분석 결과 ····· 51
[그림 2-2] 중앙일보 댓글 연관어 분석 결과 부분 확대 ···· 52
[그림 3-1] 한겨레신문 댓글 연관어 분석 결과 ····· 53
[그림 3-2] 한겨레신문 댓글 연관어 분석 결과 부분 확대 54
[그림 4-1] 경향신문 댓글 연관어 분석 결과 ····· 55
[그림 4-2] 경향신문 댓글 연관어 분석 결과 부분 확대 ···· 56

부록 [그림 1] 토픽 개수에 따른 조선일보 기사의
일관성 점수 변화 ····· 72
부록 [그림 2] 토픽 개수에 따른 중앙일보 기사의
일관성 점수 변화 ····· 72
부록 [그림 3] 토픽 개수에 따른 한겨레신문 기사의
일관성 점수 변화 ····· 73
부록 [그림 4] 토픽 개수에 따른 경향신문 기사의
일관성 점수 변화 ····· 73
부록 [그림 5] 토픽 개수에 따른 조선일보 댓글의
일관성 점수 변화 ····· 74
부록 [그림 6] 토픽 개수에 따른 중앙일보 댓글의
일관성 점수 변화 ····· 74
부록 [그림 7] 토픽 개수에 따른 한겨레신문 댓글의
일관성 점수 변화 ····· 75
부록 [그림 8] 토픽 개수에 따른 경향신문 댓글의
일관성 점수 변화 ····· 75
부록 [그림 9-1] pyLDAvis를 이용한 조선일보 기사
LDA 토픽 모델링 결과 시각화 예시 (1) ····· 79
부록 [그림 9-2] pyLDAvis를 이용한 조선일보 기사
LDA 토픽 모델링 결과 시각화 예시 (2) ····· 80
부록 [그림 10] 조선일보 본문 연관어 분석 결과 ····· 90
부록 [그림 11] 중앙일보 본문 연관어 분석 결과 ····· 91
부록 [그림 12] 한겨레신문 본문 연관어 분석 결과 ····· 92
부록 [그림 13] 경향신문 본문 연관어 분석 결과 ····· 93

제 1 장 서 론

제 1 절 문제 제기

언론은 의제설정효과(agenda-setting effects)라는 일차적 효과와 일차적 효과의 전제 하에 발생하는 이차적 효과 곧, 점화효과(priming effects)와 틀짓기효과(framing effects)를 지닌다.[1] 의제설정효과란 "언론은 특정 이슈를 강조함으로써 공중의 논의 주제를 결정하는 효과"를, 점화효과란 "공중은 언론이 강조한 이슈와 관련된 개념이나 용어를 지배적으로 사용하게 됨으로써, 그들의 정치적 판단이나 선택이 영향을 받게 된다"는 것을, 틀짓기효과란 "뉴스가 현실을 특정한 방식으로 선택, 강조, 요약해서 나타냄으로써 수용자의 이해와 해석의 범위에 일정한 한계를 설정하는 효과"를 의미한다.[2]

이와 같은 언론과 공중의 관계에 대한 이론들은 실험 연구 등 다양한 경험적 연구를 통해 뒷받침 되어왔는데, 최근에는 온라인 공간의 발전에 따라 댓글이라는 새로운 형태의 공중의 의견을 파악하여 언론과 공중 간의 관계를 살펴보고자 하는 연구들이 진행되어 오고 있다.

이러한 맥락 속에서 본 연구는 신문사의 이념적 분화가 독자측의 이

---

1) 이준웅, "의제설정 이론,"『신문과 방송』, 2001.10, p. 45.

2) 이준웅, 2001.10, p. 40. 재인용: M. E. McCombs and D. L. Shaw, "The agenda-setting function of mass media," Public opinion quarterly, Vol. 36, No. 2 (1972), pp. 176-187; 이준웅, "점화효과 이론",『신문과 방송』, 2001.11, p. 63. 재인용: S. Iyengar and D. R. Kinder, *News that matters: Agenda-setting and priming in a television age* (Chicago: University of Chicago Press, 2010); 이준웅, "틀짓기효과,"『신문과 방송』, 2001.12, p. 50. 재인용: 이준웅, "프레임, 해석 그리고 커뮤니케이션 효과,"『언론과 사회』, Vol. 29, (2000), pp. 85-153. 특히, 틀짓기효과는 이슈에 대한 언론의 묘사 방식이 이슈에 대한 공중의 '이해' 방식에 영향을 줄 수 있다는 점에서 단발적인 반응을 설명한 접근성 기반의 모델 곧, 의제설정효과 및 점화효과와는 구분된다(D. A. Scheufele and D. Tewksbury, "Framing, agenda setting, and priming: The evolution of three media effects models," *Journal of communication*, Vol. 57, No. 1 (2007), p. 11.)

념적 분화를 이끌어 낼 수 있는지, 신문사별로 참여하는 독자의 특성이 서로 다르다고 말할 수 있는지 알아보기 위해 북한이탈주민[3]에 대한 보수 및 진보 성향 신문사들의 기사 및 댓글 분석을 시도하고자 한다.[4] 다시 말해, 틀짓기효과를 고려한다면 보수 성향의 신문사에서 제공하는 북한이탈주민 관련 기사와 진보 성향의 신문사에서 제공하는 북한이탈주민 관련 기사에서는 서로 내용적 차이가 발생할 수 있고, 이러한 기사의 내용적 차이가 댓글의 내용적 차이를 유발할 수 있을 것이다.

이에 따라 본 연구는 보수 및 진보 성향의 신문사들에서 제공하는 기사의 내용적 차이가 첫째, 같은 이슈에 해당되지만 서로 다르게 부각되는 내용, 둘째, 강조되는 서로 다른 이슈를 통해 드러나는지 확인하고, 이로 인해 북한이탈주민에 대한 댓글 작성자의 태도[5]가 달라질 수 있는

---

3) 북한으로부터의 이탈 과정을 거쳐 한국에 정착한 사람들을 지칭하는 용어는 여러 차례 변화를 겪었다. 냉전 시기에는 진영 논리가 반영된 것으로 보이는 '귀순자', '귀순용사', '귀순북한동포' 등이 사용되다가 1990년대 북한의 경제난이 시작되면서 정치적인 망명보다 경제적인 망명을 택하는 사람들이 늘어남에 따라 '탈북자', '북한이탈주민'이라는 용어가 사용되기 시작했다(『한겨레신문』, 2019년 10월 29일; 『자유아시아방송』, 2016년 1월 15일). 이후 노무현 정부는 '새터민'이라는 용어를 새로이 채택해 기존 용어가 주었던 부정적인 어감을 개선하고자 했으나 이명박 정부로 넘어오면서 '새터민'이라는 용어가 지니는 어색함에 대해 문제가 제기되는 등 여러 가지 이유들로 인해 해당 용어를 더 이상 사용하지 않겠다고 발표되었고, 이후 법률상 공식적인 용어에 해당되는 '북한이탈주민'과 별칭으로 '탈북민'이 사용되기 시작했다(『연합뉴스』, 2012년 9월 6일). 본 연구에서는 위 다양한 용어들 가운데 법률상 공식적인 용어에 해당되는 '북한이탈주민'을 사용하고자 한다.

4) 본 연구는 언론과 공중의 관계를 탐구하기 위한 하나의 사례로 북한이탈주민에 대한 기사 및 댓글을 선정하는 과정에서 북한에 대한 언론사의 보도방식이 언론사의 성향에 따라 다르게 나타났다고 주장한 김경희・노기영(2011)의 연구 결과와 남북한 관계에 대한 고려가 북한이탈주민에 대한 인식으로 전이될 수 있다고 주장한 윤인진・송영호(2013)의 연구 결과를 참고하였다(김경희・노기영, "한국 신문사의 이념과 북한 보도방식에 대한 연구,"『한국언론학보』, Vol. 55, No. 1 (2011), pp. 361-387; 윤인진・송영호, "북한이주민에 대한 남한주민의 민족의식과 다문화의식,"『재외한인연구』, Vol. 30, (2013.6), p. 18.).

5) 성동규(2010)는 국내외 여러 학자들의 문헌 고찰을 통해 의견과 태도, 두 개념 간의 명확한 구분은 이루어지기 어려우나 태도란 사람들의 내적인 성향을 가리키는 것이며 이러한 태도가 반영되어 외적으로 표출되는 것이 의견이라고 보았다(성동규・양승찬 연구책임 한국언론진흥재단 조사분석팀 연구진행,『여론과 미디어』, (여론집중도조사위원회

지 댓글에 사용된 표현을 통해 살펴보고자 한다.

제 2 절 논문의 구성

본 연구는 다음과 같이 구성된다.

먼저, 2장에서는 총 세 가지 절을 통해서 선행 연구 검토를 진행할 것이다. 2장 1절에서는 기사와 댓글 간의 관계를 탐구하고자 하는 본 연구의 목적과 관련된 선행 연구를 중심으로, 2장 2절에서는 해당 목적을 확인하기 위해 선택한 사례 곧, 북한이탈주민에 대한 사람들의 태도와 관련된 선행 연구를 중심으로, 마지막 2장 3절에서는 해당 사례를 파악하기 위해 선택한 텍스트 마이닝 방법과 관련된 선행 연구를 중심으로 살펴볼 것이다. 그리고 이러한 선행 연구 검토를 통해 기존 연구들의 흐름 가운데 자리할 수 본 연구의 위치를 확인할 것이다.

다음으로, 본 연구는 3장 연구 설계 부분을 통해서 본 연구에서 살펴보고자 하는 연구 질문들을 제시하고, 제시된 연구 질문들에 대답하기 위해 설계한 연구 방법에 대해서 설명할 것이다. 특히 연구 방법은 본 연구에서 활용하고자 하는 자료가 무엇인지, 그리고 선정된 해당 자료를 어떻게 수집하고 처리할 것인지에 대해 소개하고자 한다. 더불어 다양한 텍스트 마이닝 기법들 가운데 본 연구에서 사용하고자 하는 텍스트 마이닝 기법들의 종류, 원리, 그리고 활용 방법에 대해서 간략하게 소개할

발간자료, 2010), pp. 32-37. http://www.nl.go.kr/app/nl/search/common/download.jsp?file_id=FILE-00010270982 (검색일: 2021년 1월 27일)). 또한, 크로스닉, 쥬드, 그리고 휘튼브린크(J. A. Krosnick, C. M. Judd, and B. Wittenbrink, 2018) 역시 선행 연구 고찰을 통해 "태도란 잠재적인 구성으로서 상대적으로 구체화된 자세 안에 존재하거나 아직 명료한 의견으로 통합되지 않은 것"이라고 요약했다(J. A. Krosnick, C. M. Judd, and B. Wittenbrink, "The Measurement of Attitudes," in D. Albarracin and B. T. Johnson (eds.), *The Handbook of Attitudes, Volume 1: Basic Principles* (Routledge, 2018), p. 48). 이에 따라 본 연구에서 북한이탈주민에 대한 댓글 작성자들의 '태도'를 언급할 때, 국립국어원의 한국어기초사전에 등재된 "어떤 일에 대한 입장이나 생각"으로서의 태도의 의미를 사용하는 것은 큰 무리가 되지 않을 것으로 보인다. https://www.korean.go.kr/ (검색일: 2020년 11월 5일).

것이다.

이후 4장 분석 결과에서는 연구 설계 단계에서 제시된 연구 질문들과 연구 방법에 따라 도출될 연구 결과에 대해 기술하고 해석하는 과정을 거칠 것이다. 특히, 연구 결과에 대한 기술과 해석은 본 연구에서 앞서 연구 설계를 통해 제시한 연구 질문들에 상응할 수 있도록 보수 성향의 매체에서 추출한 자료와 진보 성향의 매체에서 추출한 자료 간의 공통점과 차이점을 중심으로 이루어질 것이다.

마지막으로 본 연구는 5장 결론 부분을 통해 본 연구가 지금까지 무엇을 다루고자 하였는지, 이에 따라 도출된 분석 결과의 핵심적인 내용은 무엇이었는지에 관해 요약할 것이다. 또한, 본 연구의 의의와 한계에 관해 짧게 언급하고자 한다.

제 2 장 선행 연구 검토

제 1 절 기사와 댓글 간 일방향적 관계에 대한 연구

본 연구는 보수 및 진보 성향의 신문사들에서 제공하는 북한이탈주민에 대한 기사와 기사에 달린 댓글 간의 관계에 대한 탐구를 목적으로 하고 있기에, 한국 신문을 대상으로 기사와 댓글 간의 연관성을 다룬 기존 문헌들을 중심으로 살펴보고자 한다.

기사와 댓글 간의 연관성을 다룬 기존 연구들에는 제공자와 수용자라는 일방향적 관계에 대한 탐구를 탈피하여 기사 프레임과 댓글 프레임 간의 융합과 융합의 영향력에 대한 연구, 기사 프레임과 댓글 프레임의 쌍방향적 영향 관계에 대한 연구 등이 등장하기도 하였다.[6] 하지만 본 연구는 보수 및 진보 성향의 신문사들에서 제공하고 있는 북한이탈주민과 관련된 기사와 기사에 대한 수용적 반응이라고 볼 수 있는 댓글 간의 관계를 파악하고자 한다는 점에서 이와 같은 기사와 댓글 간의 융합이나 상호성에 대한 연구보다는 일방향적 관계에 대한 연구를 중점적으로 고찰하고자 하며, 이에는 이재신・김지은・류재미・강재혁(2010)과 좌보경・백혜진・서필교(2014), 김은미・조윤용・임영호・송보영(2015), 김가영・강경필・손민지・이청아・홍신범・김성태(2020)가 해당된다.[7]

---

6) 김혜미・이준웅, “인터넷 뉴스와 댓글의 뉴스 프레임 융합 효과 연구: 해석의 복잡성 및 태도의 극단성 분석을 중심으로,”『한국언론학보』, Vol. 55, No. 2 (2011), pp. 32-55; 이찬주・임종섭, “언론 프레임의 온라인 문화적 공명이 여론 틀 짓기에 미치는 영향 관계: 김영란법 관련 기사와 댓글에 대한 텍스트 마이닝 및 그랜저 인과 관계 분석을 중심으로,”『한국언론학보』, Vol. 62, No. 4 (2018), pp. 82-121.

7) 이재신・김지은・류재미・강재혁, “기사 프레임과 장르가 댓글 유형에 미치는 영향,”『한국언론학보』, Vol. 54, No. 2 (2010), pp. 116-137; 좌보경・백혜진・서필교, “금연정책 관련 온라인 뉴스와 댓글 유형의 내용 분석,”『홍보학연구』, Vol. 18, No. 3 (2014), pp. 13-43; 김은미・조윤용・임영호・송보영, “다문화 범죄 보도에서 기사 구성 방식과 출신국에 대한 태도가 댓글에 미치는 영향,”『한국언론학보』, Vol. 59, No. 6 (2015), pp. 107-136; 김가영・강경필・손민지・이청아・홍신범・김성태, “북한 뉴스 담론에 대한 미디어

이재신・김지은・류재미・강재혁(2010)은 정치・연예 기사 장르와 일화적・주제적 기사 프레임이 이성과 감성 혹은 욕설과 비욕설이라는 댓글의 유형 변화에 유의미한 영향을 끼친다는 연구 결과를 보여주었다.[8] 이후 좌보경・백혜진・서필교(2014)는 이재신・김지은・류재미・강재혁(2010)의 연구가 기사의 장르와 프레임만을 다뤘다는 한계점에서 출발하여 프레임, 정보원, 하위주제라는 뉴스의 속성에 따라 변화하는 댓글의 내용을 관측하고자 하였다.[9] 그 결과, 금연정책에 대한 댓글의 지지 혹은 반대 감성은 주제적 프레임이 사용되었을 때 높게 나타나는 경향이 발견되었고, 특히 전문적인 정보원이 제시되었을 때에는 지지의 감성이 높게 나타났다.[10]

그밖에도 김은미・조윤용・임영호・송보영(2015)은 외국인 이민자에 대한 기사의 댓글 수와 댓글 가운데 드러난 감성이 기사의 관점, 프레임, 내러티브 방식, 그리고 사람들이 기존에 지니고 있었던 외국인 이민자에 대한 태도에 영향을 받아 작성된다는 것을 보여주었다.[11]

한편, 김가영・강경필・손민지・이청아・홍신범・김성태(2020)는 토픽모델링 기법을 활용하여 북한 관련 뉴스의 의제와 이에 달린 댓글의 의제를 파악하고자 하였고, 그 결과 뉴스와 댓글 간의 동일한 토픽의 등장, 동일한 토픽 내 유사한 단어의 사용을 통해 북한 관련 의제가 미디어 제공자에서 미디어 수용자로 전이된다는 것을 상관관계 측면에서 보여주었다.[12]

본 연구는 김가영・강경필・손민지・이청아・홍신범・김성태(2020)와 같이 빅데이터 분석에 유용한 토픽모델링과 워드임베딩 기법을 활용하되 뉴스의 의제설정효과에 대한 확인에서 그치는 것이 아니라, 김은미・조윤용・임영호・송보영(2015)과 같이 언론의 틀짓기효과에 주목하여 북한

의제설정효과 연구: 빅데이터 토픽모델링과 워드임베딩 분석을 중심으로,"『평화연구』, Vol. 28, No. 1 (2020), pp. 287-332.

8) 이재신・김지은・류재미・강재혁, 2010, pp. 126-129.

9) 좌보경・백혜진・서필교, 2014, p. 14.

10) 좌보경・백혜진・서필교, 2014, pp. 27-32.

11) 김은미・조윤용・임영호・송보영, 2015, pp. 123-127.

12) 김가영・강경필・손민지・이청아・홍신범・김성태, 2020, pp. 309-314.

이탈주민에 대한 기사와 댓글 간의 관계를 확인하고자 한다.

제 2 절 북한이탈주민에 대한 한국 사람들의 태도 연구

북한이탈주민에 대한 한국 사람들의 태도 연구는 주로 주관식 설문조사를 통해 얻을 수 있는 정량화된 자료나 종종 심층 인터뷰[13]의 활용을 통해 이루어져 왔다. 북한이탈주민에 대한 한국 사람들의 인식을 파악할 수 있는 설문조사는 여러 기관들을 통해 지속적으로 진행되어 왔는데, 대표적으로 서울대학교 통일평화연구원[14], 고려대학교 아세아문제연구소・중앙일보・동아시아연구원[15], 국가인권위원회[16], KBS[17] 등이 있다

주관식 설문조사를 활용한 연구들 가운데 권수현(2011)은 서울대학교 통일평화연구원의 2007년 <통일의식조사> 자료를 활용하여 북한이탈주민에 대한 한국 사람들의 태도 결정 요인에 경제적인 요인과 정치적인 이념의 영향력이 제한적인 반면, 민족 정체성이 강하게 작용한다는 것을 확인하였다.[18]

하지만 손애리・이내영(2012), 윤인진・송영호(2013)는 고려대학교 아

---

13) 김희경, "북한이탈주민과 남한주민의 교류 질적 연구,"『한국심리학회지: 상담 및 심리치료』, Vol. 28, No. 2 (2016), pp. 529-562; 전주람・신윤정, "남한출신 복지관련 종사자는 탈북 동료를 어떻게 바라보는가?"『한국사회복지질적연구』, Vol. 12, No. 2 (2018), pp. 39-67.

14) 서울대학교 통일평화연구원은 2007년부터 매년 <통일의식조사>를 실시하고 있다. http://tongil.snu.ac.kr/ (검색일: 2020년 6월 29일).

15) 고려대학교 아세아문제연구소, 중앙일보, 그리고 동아시아연구원의 <한국인의 정체성 조사>는 2005년부터 5년 주기로 수행되고 있다. http://kossda.snu.ac.kr/ 또는 http://www.asiaticresearch.org/ (검색일 2020년 6월 29일).

16) 윤인진 연구책임 북한인권정보센터 연구수행,『북한이탈주민에 대한 국민인식 및 차별실태조사』, (국가인권위원회 발간자료, 2016), pp. 1-316. https://kossda.snu.ac.kr/ (검색일: 2020년 6월 30일).

17) KBS의 <국민 통일의식 조사>는 1999년에 시작되어 2005년부터 5년을 주기로 두 차례 시행되어 왔으나 2011년부터는 매년 이루어지고 있다. http://office.kbs.co.kr/tongil/ (검색일: 2020년 6월 29일).

18) 권수현, "북한이탈주민에 대한 남한국민의 태도,"『한국정치연구』, Vol. 20, No. 2 (2011), pp. 137-147.

세아문제연구소・중앙일보・동아시아연구원의 2010년 <한국인의 정체성 조사> 자료를 활용하여 북한이탈주민에 대한 한국 사람들의 태도 결정 요인에 민족 정체성이 약하게 작용한다는 것과 북한에 대한 생각, 다문화주의에 대한 인식이 중요하게 작용한다는 것을 밝혔다.[19]

하지만 앞서 소개된 권수현(2011), 손애리・이내영(2012), 그리고 윤인진・송영호(2013)의 연구는 특정 시기 곧, 1년 동안의 자료만을 다루었기 때문에 북한이탈주민에 대한 한국 사람들의 태도를 시계열적으로 다루지 못했다는 점에서 한계를 보여주었다.[20]

이와 같은 한계점이 지적된 뒤로 권수현・송영훈(2015)은 서울대학교 통일평화연구원이 실시한 <통일의식조사> 자료 가운데 2007년부터 2014년까지의 자료를 활용하여 기존 연구들이 상정했던 변수들의 영향력을 재분석하였다.[21] 황정미(2016)는 이에 더 나아가 서울대학교 통일평화연구원의 2007년부터 2015년까지의 <통일의식조사> 자료를 활용하여 현실적 위협과 인지된 개인적・사회적 위협이 북한이탈주민에 대한 한국 사람들의 태도에 영향을 줄 수 있는지 연구함으로써 새로운 틀을 마련하고자 하였다.[22] 정진원(2019) 역시도 한걸음 더 나아가 서울대학교 통일평

---

19) 손애리・이내영, "탈북자에 대한 한국인의 태도 연구: 국가정체성과 다문화수용성을 중심으로,"『아태연구』, Vol. 19, No. 3 (2012), pp. 5-33; 윤인진・송영호, 2013, pp. 7-40. 이와 같은 손애리・이내영(2012), 윤인진・송영호(2013)의 연구 결과는 오랜 기간의 남북 분단으로 인해 한국 사람들에게 '한민족' 혹은 '동포'라는 인식이 옅어졌다는 것과 북한이탈주민 역시도 외국인 이민자 가운데 하나라는 인식이 존재한다는 것을 보여준다(손애리・이내영, 2012, p. 27; 윤인진・송영호, 2013, p. 34.). 참고로 권수현(2011)과 손애리・이내영(2012), 윤인진・송영호(2013)의 연구가 서로 상이한 연구 결과를 보여준 데에는 여러 가지 원인들이 존재하겠으나 추측컨대 권수현(2011)이 활용한 자료와 손애리・이내영(2012), 윤인진・송영호(2013)가 활용한 자료가 서로 상이했다는 점과 손애리・이내영(2012), 윤인진・송영호(2013)는 다문화주의에 대한 한국 사람들의 인식과 북한에 대한 한국 사람들의 생각을 추가적인 변인으로 고려했다는 점에서 차이가 발생한 것으로 예상된다.

20) 권수현, 2011, p. 148.

21) 권수현・송영훈, "피를 나눈 형제, 이웃보다 못한 친척? 북한이탈주민에 대한 남한주민의 사회적 거리감과 결정요인, 2007-2014,"『OUGHTOPIA』, Vol. 30, No. 2 (2015), pp. 123-160.

22) 황정미, "사회적 위협 인식과 북한이탈주민에 대한 사회적 거리,"『아태연구』, Vol. 23, No. 2 (2016), pp. 311-346.

화연구원의 <통일의식조사> 자료 가운데 2012년부터 2016년까지의 자료를 활용하여 북한인권문제에 대한 한국 사람들의 인식과 태도가 북한이탈주민에 대한 한국 사람들의 태도에 영향을 줄 수 있는지 알아보고자 하였다.[23]

한편, 최윤형・김수연(2013)은 2011년 12월 북한 김정일 전 국방위원장의 사망 이후 1년 동안 네이버와 다음 포털에 게재된 조선일보, 중앙일보, 동아일보, 경향신문, 한겨레신문 총 5개 종합일간지에서 보도한 기사 가운데 북한이탈주민과 관련된 기사들에 달린 댓글만을 활용하여 북한이탈주민에 대한 댓글 작성자들의 시각을 분석하였다.[24]

본 연구는 북한이탈주민에 대한 한국 사람들의 태도를 파악하기 위해 주관식 설문조사를 이용한 연구 검토를 통하여 오랜 시간, 여러 연구들에 걸쳐 한 개인의 정치적인 입장이 북한이탈주민에 대한 태도를 결정지을 수 있는 중요한 요인으로 고려되어 왔다는 것을 알 수 있었다. 또한, 최윤형・김수연(2013)이 공백으로 남아 있는 향후 연구에 대해 제언했던 바과 같이 보수 성향의 신문사 댓글과 진보 성향의 신문사 댓글 간의 차이에 대한 연구의 부재가 존재한다는 점을 고려했을 때, 본 연구는 정치적으로 서로 다른 입장을 지닌 보수 성향의 언론과 진보 성향의 언론 간의 기사와 기사에 달린 댓글 비교가 의미 있는 작업이 될 것이라고 생각하는 바이다.[25]

제 3 절 소셜 미디어 텍스트를 활용한 연구

비교적 최근에는 정형화된 자료가 지닌 여러 가지 한계에 대한 인식이 고조되면서 비정형 자료 곧, 텍스트 형태의 자료를 이용해 다수의 사

---

23) 정진원, “북한이탈주민에 대한 한국인의 태도 결정요인,”『사회과학연구』, Vol. 30, No. 1 (2019), pp. 155-182.
24) 최윤형・김수연, ““대한민국은 우릴 받아줬지만, 한국인들은 탈북자를 받아준 적이 없어요.” 댓글에 나타난 남한사람들의 탈북자에 대한 인식과 공공 PR의 과제,”『한국관광홍보학보』, Vol. 15, No. 3 (2013), pp. 187-218.
25) 최윤형・김수연, 2013, p. 214.

람들의 생각을 파악하고자 하는 텍스트 마이닝 기법을 활용한 분석이 주목을 받고 있다.[26] 텍스트 마이닝(text mining)이란 구조화되지 않은 텍스트 집합으로부터 이전에는 알려지지 않았지만 새롭게 조명할 가치가 있는 규칙과 지식을 발견하고 추출해내는 과정을 일컫는다.[27] 텍스트 마이닝은 20세기 후반부터 점차 많은 사람들에게 인터넷 접근이 용이해짐에 따라 누구든지 자신의 생각을 웹상에 자유롭게 남길 수 있게 된 환경적 요인과 더불어, 대중의 반응과 사회적 유행을 민감하게 좇아가야 하는 다양한 산업분야에서 웹상에 쌓인 다량의 텍스트가 중요한 자료로 기능하게 됨에 따라 컴퓨터공학, 산업공학, 신문방송학, 문헌정보학, 정치외교학, 경영학 등 다양한 영역에서 연구가 진행되어져 왔다.[28]

텍스트 마이닝과 관련된 연구들은 다양한 기준으로 분류될 수 있겠으나 본 연구는 기사와 기사에 달린 댓글 분석을 시도하고자 하기 때문에, 소셜 미디어에서 얻을 수 있는 한국어 텍스트를 이용해 텍스트 마이닝 기술을 활용한 연구들을 중점적으로 살펴보고자 한다.

먼저, 트위터 텍스트를 이용한 연구들이 있다.[29] 실시간으로 쏟아지는

---

26) 논외로 여론조사기관에서 실시한 여론조사 결과와 실제 선거 결과가 맞지 않는 경우가 증가함에 따라 기존의 여론조사 방식에 대한 회의가 증가하고 있는 것 역시도 텍스트 마이닝에 대한 사회적 관심과 요구를 증대시킨 것으로 볼 수 있을 것이다(『조선일보』, 2020년 4월 14일). 더불어, 텍스트 마이닝을 활용했을 경우 기존의 설문조사 방식에 드는 비용보다 저렴하게 다량의 자료를 수집하고 분석 할 수 있다는 점 역시 텍스트 마이닝 기술이 지닌 강점이다.

27) M. A. Hearst, “Untangling text data mining,” *Proceedings of the 37th Annual meeting of the Association for Computational Linguistics*, (1999), pp. 3-10; A. H. Tan, “Text mining: The state of the art and the challenges,” *Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases*, Vol. 8, (1999), p. 65; A. Stavrianou, P. Andritsos, and N. Nicoloyannis, “Overview and semantic issues of text mining,” *ACM Sigmod Record*, Vol. 36, No. 3 (2007), p. 23.

28) 남길임・조은경, 『한국어 텍스트 감성분석』, (서울: 커뮤니케이션북스, 2017), pp. vii-viii; 서상현・김준태, “딥러닝 기반 감성분석 연구동향,” 『한국멀티미디어학회지』, Vol. 20, No. 3 (2016), p. 8.

29) 트위터 텍스트를 이용한 연구에 대한 본 연구자의 두 가지의 거시적인 분류는 계량정보학적 방법론을 통해 트위터 텍스트를 활용한 500편 이상의 국내 연구 동향을 조사한 연구의 결과와도 일치한다. 해당 내용은 다음의 논문을 참조. 강범일・이재윤, “트위터 관련 연구에 대한 계량정보학적 분석,” 『정보관리학회지』, Vol. 31, No. 3 (2014), pp.

방대한 양의 트위터를 통해서 의제가 설정되거나 여론이 조성되는 등 트위터의 영향력을 감안하여 2012년 대선 당시 트위터 내부의 선거 관련 트렌드를 분석하거나 대선의 결과를 예측한 연구들, 그리고 '헬조선', '미투운동', '구제역' 등과 같이 특정한 사회적 이슈에 대해 키워드, 주제, 감성분석 등을 시도한 연구들이 있다.[30] 또한, 접근 가능성이 높고 간결한 동시에 신속하다는 트위터 텍스트의 장점을 이용해 특정 상품에 대한 평가를 파악하여 기업의 마케팅 측면에서의 니즈(needs)를 충족시킬 수 있다는 결과를 발표한 연구들이 있다.[31]

다음으로, 뉴스 텍스트를 이용한 연구들이 있다. 뉴스 텍스트를 이용한 연구들은 먼저 언론사들의 정파성을 파악하고자 하였던 연구들로서, 각 언론사에서 생산해내는 방대한 양의 기사들을 이용해 상시에 드러나는 언론사의 보도 색채를 정량적으로 탐색하고자 한 연구와 이에 더 나아가 언론사의 다양한 색채가 선거 혹은 '세월호 참사'와 같이 특정한 사건이 발생했을 경우에 어떻게 드러나는지를 다룬 연구 등이 존재한다.[32] 또한, 이처럼 큰 규모로 발생한 특정 사건이 아니더라도 '수저계

293-311.

30) 배정환・손지은・송민, "텍스트 마이닝을 이용한 2012년 한국대선 관련 트위터 분석,"『지능정보연구』, Vol. 19, No. 3 (2013), pp. 141-156; 이서영・권상집, "트위터 메시지 분석을 통한 선거 결과 예측 고찰: 18대 대선을 중심으로,"『한국콘텐츠학회논문지』, Vol. 19, No. 4 (2019), pp. 174-186; 차준범・성정웅・김정근・박상현, "트위터 데이터에 기반한 헬-조선 키워드 분석,"『한국멀티미디어학회』, Vol. 19, No. 2 (2016), pp. 195-198; 곽수정・김현희, "텍스트 마이닝과 토픽 모델링을 기반으로 한 트위터에 나타난 사회적 이슈와 키워드 및 주제 분석,"『정보처리학회논문지 소프트웨어 및 데이터공학』, Vol. 8, No. 1 (2019), pp. 13-18; 채희찬・이종욱・최윤아・박대희・정용화, "트윗 텍스트 마이닝 기법을 이용한 구제역의 감성분석,"『정보처리학회논문지 소프트웨어 및 데이터공학』, Vol. 7, No. 11 (2018), pp. 419-426.

31) 진설아・허고은・정유경・송민, "트위터 데이터를 이용한 네트워크 기반 토픽 변화 추적 연구,"『정보관리학회지』, Vol. 30, No. 1 (2013), pp. 285-302; 이혜진・최자영, "텍스트 마이닝 기법을 활용한 편의점 이용 고객의 트위터 리뷰에 대한 감성분석,"『글로벌경영학회지』, Vol. 16, No. 4 (2019), pp. 143-164; 서대호・한지은, "텍스트 마이닝 기법 기반 국내 홈 퍼니싱 브랜드의 소비자 인식 및 동향 비교・분석: IKEA와 한샘을 중심으로,"『디자인융복합연구(구.인포디자인이슈)』, Vol. 17, No. 6 (2018), pp. 229-243.

32) 감미아・송민, "텍스트 마이닝을 활용한 신문사에 따른 내용 및 논조 차이점 분석,"『지능정보연구』, Vol. 18, No. 3 (2012), pp. 53-77; 강범일・송민・조화순, "토픽 모델링

급', '기후환경', '저출산' 등과 같이 세간의 관심이 쏠리는 이슈에 대한 사회적 인식이나 파급효과, 그리고 언론 보도의 동향이 어떠한지 뉴스 텍스트를 통해 확인하고자 한 연구들이 있다.[33]

다음으로, 다양한 종류의 사이트에서 추출할 수 있는 댓글 텍스트를 활용한 연구들이 있다. 댓글 텍스트를 이용한 연구들 가운데 일부는 뉴스 등의 보도 자료와 함께 연구되었는데, 주로 사회적으로 이슈가 되는 '난민', '낙태' 등의 특정한 주제나 주제와 관련된 사건에 대한 이해를 돕고 이에 대한 사람들의 인식을 조사하는 방식으로 작성되었다.[34]

마지막으로 트위터, 뉴스, 댓글 텍스트 이외에도 다양한 출처의 텍스트들 역시 연구의 분석대상으로 활용되었다. 메갈리아와 같은 온라인 커뮤니티의 게시글과 해당 게시글 작성자의 익명성 유무 간의 관계를 분석한 연구, 청와대 국민청원 자료를 활용해 정치참여와 감정의 관계를 탐색한 연구, 뉴스 텍스트와 더불어 커뮤니티 및 카페의 게시글, 뉴스 댓글, 그리고 인터넷 방송 실시간 채팅 텍스트를 통해 온라인 성차별 현상을 이해하고자 한 연구, 기존의 논문들을 활용해 연구 동향을 살피거나 새롭게 연구할 수 있는 영역을 고찰한 연구 등이 있다.[35]

을 이용한 신문 자료의 오피니언 마이닝에 대한 연구,"『한국문헌정보학회지』, Vol. 47, No. 4 (2013), pp. 315-334; 박종희, "세월호 참사 1년 동안의 언론보도를 통해 드러난 언론매체의 정치적 경도,"『한국정치학회』, Vol. 50, No. 1 (2016), pp. 239-269.

33) 김용회・한창근, "'수저계급' 관련 웹 뉴스 기사에 대한 의미연결망 분석,"『한국사회복지학』, Vol. 71, No. 3 (2019), pp. 55-81; 강성원・김도연・최희선・한국진・진대용, "기후환경 이슈 분석을 위한 텍스트 마이닝 활용방안 연구,"『기후환경정책연구』, (2018), pp. 29-64; 노법래・양경은, "한국 사회 저출산 논의 구조와 그 변환에 관한 텍스트 마이닝 분석 - 2000년 이후 언론 기사문에 대한 토픽 모델링과 토픽 구조의 시계열적 변동을 중심으로,"『한국사회복지학』, Vol. 71, No. 4 (2019), pp. 153-175.

34) 강진구・이기성, "텍스트마이닝(Text Mining)을 통해 본 제주 예멘 난민: 네이버 뉴스 댓글을 중심으로,"『다문화콘텐츠연구』, Vol. 30 (2019), pp. 103-135; 김찬우, "온라인 기반 낙태 이슈 변화 분석: 온라인 뉴스와 댓글을 중심으로,"『예술인문사회 융합 멀티미디어 논문지』, Vol. 8, No. 10 (2018), pp. 559-568.

35) 송준모・강정한, "메갈리아의 두 딸들: 익명성 수준에 따른 온라인 커뮤니티의 정체성 분화,"『한국사회학』, Vol. 52, No. 4 (2018), pp. 161-206; 송준모・박영득, "청와대 국민청원에서는 무엇이 일어나는가?: 자연어 처리를 활용한 청와대 국민청원 분석,"『한국정치학회보』, Vol. 53, No. 5 (2019), pp. 53-78; 이수연・김현정・정수연, "텍스트마이닝 기반 토픽 분석을 통한 온라인 성차별성의 이해,"『사이버커뮤니케이션학보』, Vol. 33, N

지금까지 텍스트의 원천에 대한 분류를 기준으로 텍스트 마이닝 기법을 활용한 연구들을 살펴보았다. 이를 통해 최근에는 트위터, 뉴스, 댓글, 온라인 커뮤니티/카페, 청와대 국민청원, 인터넷 방송 실시간 채팅 등 다양한 출처들로부터 얻을 수 있는 다량의 텍스트들이 연구에 적극적으로 활용되고 있음을 알 수 있었다. 본 연구 역시 텍스트에 해당되는 기사 자료와 댓글 자료를 활용한 텍스트 마이닝 기법을 통해 앞서 2장 1절과 2절에서 언급했던 연구의 공백을 메우고자 한다.

---

o. 3 (2016), pp. 159-199; 김성근・조혁준・강주영, “학술연구에서의 텍스트 마이닝 활용 현황 및 주요분석기법,”『정보화연구(구 정보기술아키텍처연구)』, Vol. 13, No. 2 (2016), pp. 317-329.

제 3 장 연구 설계

제 1 절 연구 질문

본 연구는 보수 및 진보 성향의 신문사들의 기사 자료와 댓글 자료를 활용하여 기사와 댓글 간의 관계를 상관관계 측면에서 파악하고자 한다.[36] 이에 따라 먼저 신문사의 성향에 따라 나타나는 기사의 토픽 차이를 확인하기 위해 기사 자료를 활용한 토픽모델링을 실시할 것이다. 다음으로 댓글의 경우 짧은 내용으로 구성되어 있어 토픽 확인이 어렵다는 특성을 고려하여 연관어 분석을 실시하고 이를 통해 어떤 단어들이 서로 함께 등장하여 어떠한 맥락을 이루고 있는지 알아보고자 한다. 이와 같은 두 가지 분석은 기사에서 보수 성향 신문 대(對) 진보 성향 신문이라는 구도의 차이가 반영되어 나타나는지, 이와 같은 구도의 차이가 댓글에서도 나타나는지, 특히 북한이탈주민에 대한 댓글 작성자들의 태도가 어떤 단어들로 표현되는지 확인해 볼 수 있도록 도와줄 것이다. 이에 따라 본 연구는 두 가지 연구 질문을 제기한다.

연구 질문 1. 보수 성향과 진보 성향의 신문사 홈페이지에 게재된 북한이탈주민 관련 기사들의 토픽은 무엇인가? 해당 토픽은 신문사의 성향에 따라 차이를 보이는가?

연구 질문 2. 보수 성향과 진보 성향의 신문사 홈페이지에서 추출된 북한이탈주민에 대한 댓글은 어떤 단어로 구성되어 있는가? 단어의 구성은 신문사의 성향에 따라 차이를 보이는가?

---

36) 공중이 이용하는 언론의 성향 혹은 종류에 따라 발생하는 공중의 반응을 인과적인 측면에서 탐구하는 연구 분야는 다음의 연구를 참고. 이준웅, “언론 매체 이용 및 해석적 틀이 통일 및 대북 정책에 대한 의견에 미치는 효과,”『한국언론학보』, Vol. 48, No. 1 (2004), pp. 28-56; 손영준, “미디어 이용이 보수 진보적 의견에 미치는 영향,”『한국언론학보』, Vol. 48, No. 2 (2004), pp. 240-266.

더불어, 본 연구는 위 두 가지 연구 질문과는 별도로 북한이탈주민에 대한 댓글 작성자들의 태도를 파악하기 위해 북한이탈주민에 대한 댓글을 대상으로 감성사전을 이용한 감성분석과 본 연구에서 지정한 일부 단어들을 이용한 감성분석, 총 두 가지의 감성분석을 추가적으로 진행하고자 한다.

특히, 지정된 단어를 이용한 감성분석의 경우 북한이탈주민에 대한 한국 사람들의 태도가 북한과 북한 인권에 대한 한국 사람들의 인식과 관련이 있다는 윤인진・송영호(2013), 정진원(2019)의 연구 결과로부터 착안하였다.[37]

하지만 본 연구는 북한에 대한 한국 사람들의 긍정적이거나 부정적인 태도가 북한이탈주민에 대한 한국 사람들의 긍정적이거나 부정적인 태도로 동일하게 전이된다고 주장한 윤인진・송영호(2013), 정진원(2019)의 연구 결과와 달리, 북한에 대한 적대적인 태도가 북한이탈주민에 대한 우호적인 태도와 함께 나타날 수 있다는 점을 확인하고자 한다.

이는 먼저, 김경희・노기영(2011)이 보수 성향을 띠는 조선일보, 동아일보가 진보 성향을 띠는 한겨레신문, 경향신문에 비해 북한에 대해 부정적인 논조의 기사를 게재해왔고, 특히 보수 성향의 신문사는 신문사가 지향하는 이념을 드러내는 데 있어 북한이탈주민을 취재원으로서 이용한 것으로 보인다고 주장했던 내용을 통해 뒷받침 될 수 있다.[38] 또한, 곽정래・이준웅(2009), 권호천(2017) 역시도 조선일보, 중앙일보가 북한이탈주민에 대해서 보도할 때 한겨레신문, 경향신문에 비해 인권을 강조하는 방식으로 기사를 다뤄왔다는 연구 결과를 보여주었다.[39]

이에 따라 본 연구는 보수 성향의 신문사에서 수집한 댓글에서는 북한에 대한 적대적인 정서를 나타내는 '북괴'라는 단어가 북한이탈주민에 대한 우호적인 정서를 나타내는 '인권' 혹은 '동포' 혹은 '보

---

37) 윤인진・송영호, 2013, pp. 18, 26-27; 정진원, 2019, pp. 163, 172-173.
38) 김경희・노기영, 2011, p. 361.
39) 곽정래・이준웅, "김대중・노무현 정부시기 탈북자 문제에 관한 언론의 프레임 유형 연구,"『한국언론학보』, Vol. 53, No. 6 (2009), pp. 196-217; 권호천, "박근혜 정부시기 탈북민 관련 신문보도의 의미구조 차이 비교 분석: 조선일보와 한겨레신문 기사의 텍스트 네트워크 분석을 중심으로,"『사회과학연구』, Vol. 29, No. 2 (2017), pp. 103-141.

호'라는 단어와 함께 사용되는 비율이 진보 성향의 신문사에서 수집한 댓글에 비해 높을 것으로 예상한다. 또한, 본 연구는 보수 성향의 신문사에서 수집한 댓글에서는 북한에 대한 적대적인 정서를 나타내는 '북괴'라는 단어가 북한이탈주민에 대한 적대적인 정서를 나타내는 '간첩' 혹은 '위장' 혹은 '빨갱이'[40]라는 단어와 함께 사용되는 비율이 진보 성향의 신문사에서 수집한 댓글에 비해 낮을 것으로 예상한다. 이와 같은 일부 단어 지정은 위 윤인진・송영호(2013), 정진원(2019), 김경희・노기영(2011), 곽정래・이준웅(2009), 권호천(2017)의 연구 결과뿐만 아니라 아래 [표 1][41]을 통해 제시된 바와 같이 네 개의 신문 기사 댓글에서 사용된 상위 80개의 단어들을 참고하였다.

마지막으로 본 연구는 감성사전을 이용한 감성분석을 시도할 경우, 보수 성향의 신문사에서 수집한 댓글 가운데 나타난 북한이탈주민에 대한 긍정적 감정 수치가 진보 성향의 신문사에서 수집한 댓글 가운데 나타난 북한이탈주민에 대한 긍정적 감정 수치보다 높을 것으로 예상한다.

---

40) '빨갱이'라는 단어는 '북괴', '인권', '동포', '보호', '간첩', '위장' 단어들과는 달리 아래 [표 1] 가운데 등재되어 높은 빈도수를 보여주는 단어는 아니다. 하지만 본 연구는 '빨갱이'라는 단어가 북한이탈주민에 대한 한국 사람들의 적대적인 태도를 보여줄 수 있는 단어라고 판단하였기에 추가적으로 함께 고려하였다.

41) [표 1]에 등재된 상위 80개 단어들은 별도의 전처리 과정 없이 'konlpy' 패키지 내 꼬꼬마(Kkma) 형태소 분석기의 명사 추출 기능을 통해 추출되었다. https://konlpy-ko.readthedocs.io/ko/v0.4.3/ (검색일: 2020년 11월 26일).

[표 1] 네 개의 신문 기사 댓글에 사용된 상위 80개 단어

순위	단어	빈도수	순위	단어	빈도수
1	탈북자	10,054	21	간첩	837
2	북한	3,814	22	통일	816
3	사람	1,944	23	재인	804
4	북민	1,937	24	국가	734
5	우리	1,733	25	문재인	693
6	대한	1,706	26	사회	683
7	민국	1,675	27	단체	681
8	대한민국	1,657	28	김정은	652
9	정부	1,576	29	때문	652
10	국민	1,500	30	주민	649
11	정권	1,467	31	인간	639
12	인권	1,416	32	미국	633
13	한국	1,374	33	지금	598
14	나라	1,222	34	대통령	597
15	생각	1,220	35	가족	588
16	남한	1,182	36	좌파	580
17	중국	1,180	37	지원	558
18	자유	1,121	38	하나	554
19	문제	1,100	39	북송	533

20	정은	908	40	민주	483
41	무엇	474	61	조사	353
42	사실	473	62	자신	351
43	이것	472	63	대북	344
44	국정원	461	64	강제	340
45	주의	436	65	임수경	338
46	동포	436	66	이용	336
47	필요	429	67	남북	335
48	기사	427	68	이해	334
49	김정일	417	69	정보	334
50	보호	415	70	민변	334
51	목숨	412	71	집단	331
52	하다	411	72	소리	326
53	조선	400	73	출신	323
54	종북	395	74	위장	323
55	수경	389	75	정신	323
56	이상	387	76	세력	320
57	이유	370	77	의원	320
58	자기	368	78	언론	317
59	북괴	361	79	탈출	317
60	그것	359	80	정도	314

제 2 절 연구 방법

1. 자료 수집

본 연구는 자료 수집 대상을 선정하는 과정에서 사회적 통념에 따라 나뉘어져 있는 보수 성향의 신문사들과 진보 성향의 신문사들의 색채를 실증적으로 증명한 연구들을 참고하였다. 그 가운데 특히 북한과 북한이탈주민에 대한 언론의 보도 방식에서도 신문사의 정치・경제적 이념에 따라 차이가 나타난다고 주장한 기존 연구들의 결과를 참고하였다.[42] 이에 따라 본 연구는 자료 수집 대상으로 조선일보, 중앙일보, 한겨레신문, 경향신문 총 네 곳을 채택하였다.[43]

채택된 네 곳의 신문사들에 대해서는 파이썬의 'BeautifulSoup' 패키지를 활용하여 각각의 웹크롤러를 제작한 뒤 북한이탈주민, 탈북자, 탈북민, 그리고 새터민으로 키워드를 검색하였을 경우 확인할 수 있는 뉴스 기사들과 기사들에 달린 댓글들을 수집하였다. 본 연구는 북한이탈주민 집단에 대해 여러 가지 용어들이 사용되고 있는 현실을 고려하여 북한이탈주민 외에도 탈북자, 탈북민, 그리고 새터민을 키워드 검색에 포함하여 더 많은 기사와 댓글 자료들을 확보하고자 하였다.

1) 신문 기사

키워드 검색을 통해 확인할 수 있는 기사 자료에는 각 신문사에서 발

---

42) 곽정래・이준웅, 2009, pp. 196-217; 김경희・노기영, 2011, pp. 361-387; 권호천, 2017, pp. 103-141.

43) 연구 초기에 조선일보와 중앙일보 외에도 대표적인 보수 성향의 신문으로 고려되는 동아일보 역시 본 연구의 웹크롤링 대상으로 포함하였다. 하지만 동아일보 웹에서 댓글을 수집할 경우 동일한 페이지 안에서도 랜덤키가 설정되어 있어 이를 제어할 수 있는 기술을 찾지 못해 연구를 진행하는 도중에 동아일보 자료는 제외되었다. 참고로 앞서 소개한 최윤형・김수연(2013)의 연구는 분석 대상으로 동아일보 기사 댓글이 포함하고 있는데, 이는 포털 사이트라는 특정 플랫폼을 이용하였기 때문에 가능하였을 것으로 사료된다.

행하는 다양한 종류의 콘텐츠 자료가 포함된다. 예를 들면 조선일보의 경우 뉴스, 조선비즈, 헬스조선, 주간조선, 이코노미조선, 어린이조선 등 총 18개의 콘텐츠가 존재한다. 하지만 동일한 신문사의 뉴스 기사 자료일지라도 콘텐츠가 변경될 경우 웹크롤러의 작동이 중지되는 어려움이 있어 본 연구는 각 신문사의 '뉴스'를 중심으로 자료를 수집하였다. 이에 따라 조선일보의 '뉴스', 중앙일보의 '뉴스', 한겨레신문의 '뉴스', 그리고 경향신문의 '경향신문'이 본 연구의 자료 수집 대상으로 선정되었다.[44]

조선일보의 경우 2020년 6월 20일에 파이썬 웹크롤러를 실행하여 총 12,822개 기사들을 모을 수 있었다. 이 가운데 탈북민으로 검색했을 때 1,337개, 탈북자로 검색했을 때 10,051개, 새터민으로 검색했을 때 663개, 북한이탈주민으로 검색했을 때 822개의 자료들을 확보할 수 있었다. 하지만 조선일보 웹페이지에서는 10개씩 기사 목록을 보여주는 웹페이지의 숫자가 1,000이 넘어가는 경우 그 이상의 기사들을 출력해주지 않기 때문에, 사실상 탈북자로 검색했을 때 10,000개의 기사들을 수집할 수 있었다. 이후 기사와 댓글 간의 조응 관계에 대한 고려를 바탕으로 조선일보의 댓글이 시작된 것으로 추정되는 2006년 1월부터 2020년 3월 1일까지의 9,141개 기사를 추출하였고, 기사의 본문 내용을 기준으로 중복과 결측치를 제거하는 과정을 통해 총 8,326개의 기사를 최종 분석 대상으로 확보하였다.[45]

---

44) 이때, 경향신문의 '향이네신문' 기사 본문과 댓글은 자료 수집에서 제외되었다. '향이네신문'의 경우 수집하는 과정에서 홈페이지 접속 이후에 추가적으로 어떤 좌표든 클릭과 스크롤다운 기술이 필요한데, 클릭할 좌표를 설정할 때 홈페이지마다 고정적으로 클릭해도 무방한 좌표를 지정하는 데 어려움이 있었다. 그러나 '향이네신문'으로 수집되는 기사의 개수가 17개에 불과했다는 점에서 '향이네신문'의 수집 여부가 전체 분석 결과에 큰 영향을 미치지 않을 것으로 예상된다.

45) 조선일보에서 추출한 뉴스 기사들 가운데에는 종종 "요청하신 페이지의 주소가 잘못 입력되었거나 주소의 변경 혹은 삭제로 인하여 사용할 수 없는 상태입니다."라는 오류가 존재하였다. 이로 인해 조선일보 기사 자료에 대해 중복 및 결측치 제거 작업을 실시할 경우 자료의 크기가 대폭 감소하는 경향이 나타났다. 해당 오류가 혹여 본 연구에서 사용한 웹크롤러에 수정이 필요한 코드가 존재하거나 인터넷 연결 오류로부터 파생되었을 수 있는 가능성을 고려해 2020년 3월 22일 1차 수집 이후 2020년 6월 20일 2차 수집을

중앙일보의 경우 2020년 3월 24일에 파이썬 웹크롤러를 실행하여 탈북민으로 검색 시 520개, 북한이탈주민으로 검색 시 792개, 새터민으로 검색 시 394개, 탈북자로 검색 시 4,506개의 기사들로 총 6,212개의 기사들을 수집하였다. 한겨레신문의 경우 2020년 3월 24일에 파이썬 웹크롤러를 실행하여 탈북민으로 검색했을 때 207개, 새터민으로 검색했을 때 510개, 북한이탈주민으로 검색했을 때 683개, 탈북자로 검색했을 때 1,000개로 총 2,400개의 기사들을 수집하였다.[46] 경향신문의 경우 2020년 3월 25일 파이썬 웹크롤러를 실행하여 탈북민으로 검색했을 경우 210개, 새터민, 북한이탈주민, 탈북민으로 검색했을 경우 3,280개로 총 3,490개의 기사들을 수집하였다.[47]

이렇게 수집한 중앙일보, 한겨레신문, 경향신문 기사 자료는 앞서 설명된 조선일보 기사 본문과 동일한 과정을 거쳤고, 그 가운데 중앙일보 기사 총 3,287개, 한겨레신문 기사 총 1,367개, 경향신문 기사 총 1,659개가 최종 분석 대상으로 선정되었다. 자세한 결과는 아래 [표 2]를 통해 확인할 수 있다.

시도하였지만 동일한 페이지에서 동일한 오류가 발생함을 확인하고 난 뒤 뉴스 기사 홈페이지의 주소가 변경된 것으로 이해하였다.

46) 2020년 3월 24일 한겨레신문의 웹페이지에서 '탈북자' 키워드로 검색을 했을 때, 99번째 페이지의 목록이 100번째 페이지에서 237번째 페이지까지 동일하게 반복되는 오류가 존재하여 0번째에서 99번째 페이지까지 출력되는 기사 목록만을 크롤링하였다.

47) 구글에서 경향신문을 검색하여 접속할 경우, 접속기기가 PC이더라도 경향신문의 모바일 버전으로 접속되는 현상이 존재한다. 경향신문은 PC버전 경향신문 사이트(http://www.khan.co.kr)와 모바일 경향신문 사이트(http://www.m.khan.co.kr)에서 수집할 수 있는 기사의 양적인 면에서 차이를 보여준다. 본 연구에서 확인한 바로는 2020년 3월 25일 PC버전 경향신문 사이트에서 '탈북민' 키워드 검색을 제외하고 '새터민', '북한이탈주민', '탈북자' 키워드 검색 시 출력되는 기사의 목록이 모두 3,280개로 동일하였다. 반면, 모바일 경향신문 사이트의 경우 '새터민' 키워드 검색 시 359개, '북한이탈주민' 키워드 검색 시 527개, '탈북자' 키워드 검색 시 2,613개로 총 3,499개의 기사 목록이 출력되었다. 이에 따라 본 연구는 PC버전 경향신문 사이트에서 '새터민', '북한이탈주민', '탈북자'를 모두 같은 단어로 취급하고 있을 것으로 예상하고 있다. 또한 모바일 경향신문 사이트의 결과물 3,499개에서 일반 경향신문 사이트의 결과물 3,280개로 총량이 감소한 것은 서로 다른 키워드로 검색하였을 때 출력될 수 있는 중복 기사가 제외되었기 때문으로 추측하고 있다. 본 연구는 모바일 경향신문 사이트가 아닌 PC버전 경향신문 사이트를 통해 자료를 수집하였다.

2) 댓글

조선일보 기사들로부터 수집할 수 있었던 댓글의 규모는 조선일보 댓글 기능이 도입된 것으로 추정되는 2006년 1월부터 2020년 3월 1일까지 총 174,154개에 해당된다. 그렇지만 본 연구는 수집한 댓글을 그대로 사용하지 않고 최근 댓글 조작에 대한 논란을 반영하여 완전히 동일하지만 여러 차례 등록된 댓글과 몇 글자만 바꾸어서 등록된 거의 비슷한 댓글을 최대한 배제하기 위해서 두 차례의 중복 제거 작업을 거쳤다. 먼저, 조선일보의 경우 댓글 수집 이후 완전히 동일한 댓글을 제외하는 1차 중복 제거 작업을 통해 148,027개의 댓글이 남겨지게 되었다.

하지만 수집된 댓글에는 북한이탈주민과 관련되지 않은 댓글도 다량 포함되어 있기 때문에 별도의 추출 과정이 필요하다고 판단하였다. 이에 따라 본 연구는 148,027개의 댓글에 'hanja' 패키지[48]를 이용하여 댓글 내용에 존재하는 한자를 한글로 바꾼 뒤, 북한이탈주민과 관련된 태도가 드러나는 댓글을 최대한 다량으로 수집하기 위해 '북한이탈주민', '탈북자', '탈북민', '새터민', '귀순자', '귀순용사', '귀순동포', '귀순북한동포', '월남민'이라는 글자가 포함된 9,939개의 댓글들을 추출하였다.

이후 파이썬의 're' 패키지를 활용하여 댓글의 핵심적인 내용들을 담고 있다고 판단되는 글자 곧, 자음과 모음이 합쳐진 한글, 숫자, 영어 그리고 한 글자의 공백을 제외하고 특수문자나 두 글자 이상의 공백을 모두 제거한 뒤 2차 중복 제거 과정을 진행하였다. 이로써 최종 분석 대상으로 총 9,937개의 댓글을 확보할 수 있었다.

중앙일보, 한겨레신문, 경향신문에서 수집한 댓글 자료 역시 위 조선일보 댓글 자료와 동일한 과정을 거쳤고, 본 연구는 이로부터 각각 1,252개, 735개, 542개의 댓글 자료를 확보할 수 있었다. 자세한 결과는 아래 [표 3]을 통해 확인할 수 있다.

---

48) https://github.com/suminb/hanja (검색일: 2020년 9월 15일).

[표 2] 기사 수집

	조선일보	중앙일보	한겨레신문	경향신문
댓글이 시작된 시기부터 2020년 3월 1일까지 추출된 기사 개수	9,141	3,707	1,585	1,700
댓글 시작 시기	(2006. 1. 4.~)	(2009.12.31.~)	(2013. 5. 2.~)	(2012. 5. 11.~)
기사 본문을 기준으로 제거된 중복 기사 개수	814	420	218	41
기사 본문을 기준으로 제거된 결측치 기사 개수	1	0	0	0
본 연구에 사용되는 총 기사 개수	**8,326**	**3,287**	**1,367**	**1,659**

註: 단위(개), 出處: 조선일보 https://www.chosun.com (검색일: 2020년 6월 20일), 중앙일보 https://joongang.joins.com (검색일: 2020년 3월 24일), 한겨레신문 http://www.hani.co.kr (검색일: 2020년 3월 24일), 경향신문 http://www.khan.co.kr (검색일: 2020년 3월 25일)

[표 3] 댓글 수집

	조선일보	중앙일보	한겨레신문	경향신문
댓글이 시작된 시기부터 2020년 3월 1일까지 추출된 댓글 개수	174,154	23,758	15,718	10,298
댓글 내용을 기준으로 제거된 1차 중복 댓글 개수	26,127	4,327	2,559	1,114
댓글 내용을 기준으로 제거된 결측치 댓글 개수	0	0	1	1
댓글 내용에 북한이탈주민을 지칭하는 용어가 불포함 되어 제거된 댓글 개수	138,088	18,179	12,422	8,639
댓글 내용을 기준으로 제거된 2차 중복 댓글 개수	2	0	1	2
본 연구에 사용되는 총 댓글 개수	**9,937**	**1,252**	**735**	**542**

註: 단위(개), 出處: 조선일보 https://www.chosun.com (검색일: 2020년 6월 20일), 중앙일보 https://joongang.joins.com (검색일: 2020년 3월 24일), 한겨레신문 http://www.hani.co.kr (검색일: 2020년 3월 24일), 경향신문 http://www.khan.co.kr (검색일: 2020년 3월 25일)

2. 자료 전처리

본 연구는 추후 사용하게 될 형태소 분석기의 성능을 높이기 위해 'soynlp' 패키지[49] 내 정규화(normalizer) 작업 기능을 이용하여 "ㅋㅋㅋㅋㅋㅋㅋㅋ우와와와와와와와와"와 같이 댓글에서 의미 없이 과도하게 연속적으로 등장하는 글자들을 "ㅋㅋ우와와"처럼 두 글자의 형태로 정리하였다.[50]

---

49) https://github.com/lovit/soynlp (검색일: 2020년 8월 21일). 참고로 'soynlp' 패키지는 다양한 기능을 제공하는데 특히, 명사에서 자주 발생하는 미등록단어(Out of Vocabulary)를 추출한 뒤 추출된 미등록단어를 형태소 분석기가 인식할 수 있도록 도움을 줄 수 있다는 데 강점을 지니고 있는 패키지이다. 본 연구 역시 연구를 진행하는 과정 가운데 'soynlp' 패키지 내 명사추출(LRNounExtractor_v2) 기능과 'ckonlpy' 패키지 내 트위터(前 Twitter, 現 Okt) 형태소 분석기를 결합하여 텍스트 마이닝을 시도해보았다. 하지만 전반적으로 꼬꼬마 형태소 분석기를 사용했을 때의 결과에 비해 해석하기 용이한 결과가 나타나지 않아 결국 'soynlp' 패키지의 명사추출 기능은 사용하지 않았다. 'ckonlpy' 패키지는 다음을 참조. https://github.com/lovit/customized_konlpy (검색일: 2020년 8월 21일). 트위터 형태소 분석기는 다음을 참조. https://github.com/twitter/twitter-korean-text (검색일: 2020년 8월 21일). https://github.com/open-korean-text/open-korean-text (검색일: 2020년 8월 21일).

50) 전처리 과정에서 한글 텍스트의 띄어쓰기는 형태소 분석기가 인식할 수 있는 텍스트의 범위에 영향을 줄 수 있기 때문에 매우 중요한 부분에 해당된다. 이에 따라 'soyspacing', 'py-hanspell', 'PyKoSpacing' 등을 비롯해 한글의 띄어쓰기 문제를 교정해 줄 수 있는 여러 패키지들이 개발되었다. 본 연구는 이와 같은 문제의식에 동의하여 댓글에서 잘 지켜지지 않는 띄어쓰기 문제를 교정하고자 노력하였으나 아래의 여러 가지 어려움으로 인해 띄어쓰기 전처리를 시행하지 못하였다. 먼저, 'soyspacing'의 경우 띄어쓰기를 해결하려는 댓글 자료 내에서 띄어쓰기가 잘 지켜진 댓글 자료를 일부 추출한 다음 모델을 훈련시켜 사용할 것을 권하고 있다. 그러나 본 연구에서 수집한 약 20만 개의 댓글 텍스트 가운데 모델의 한글 띄어쓰기 훈련을 위해 최소 10%를 추출한다고 하더라도 2만 개의 자료를 추출하는 데 드는 물리적 비용을 감당하기에는 다소 어려움이 있었다. 다음으로 'py-hanspell'은 네이버 맞춤법 검사기 API와 연동되어 작동하는 패키지이다. 하지만 최근 네이버 맞춤법 검사기 주소 변경에 따라 관리자 업데이트가 시행되지 않아 패키지 설치에 어려움이 있었다. 마지막으로 'PyKoSpacing'은 100만개 이상의 뉴스 자료로 훈련된 띄어쓰기 교정 패키지라는 점에서 강점을 지니고 있지만 한 번에 198자 이상을 지원하지 않는다는 점과 댓글 자료로 훈련되어 있지 않다는 점에서 본 연구에서 사용하기에 어려움이 있을 것이라 판단하였다. 자세한 내용은 이하를 참조. https://github.com/lovit/soyspacing/blob/master/tutorials/presentation.pdf (검색일: 2020년 8월 21일). https://github.co

한편, 댓글을 이용한 감성분석의 경우 위 정규화 과정까지만 전처리가 진행된 반면, 신문 기사와 댓글을 이용한 토픽 분석과 연관어 분석의 경우 이하의 전처리 과정을 추가적으로 진행하였다.

본 연구는 신문 기사와 댓글을 토큰화(tokenization) 하는 과정에서 'konlpy' 패키지 내 꼬꼬마(Kkma) 형태소 분석기를 사용하였다.[51] 그리고 해당 과정에서는 어떤 의미를 지니고 있는지 파악하기 어려운 한 글자 길이의 단어와 유의미한 단어라고 보기 어려운 불용어(stopwords)를 제거하였다. 불용어는 길호현(2018)이 제시하고 있는 한국어 불용어 목록을 사용하였다.[52]

이후 토큰화된 기사 본문에 대해서는 보통명사와 고유명사만을 추출하였다. 이는 기사에서 어떠한 토픽들이 존재하였는지 파악하는 데에는 명사 추출을 통한 유추만으로 충분하다고 판단하였기 때문이다.

---

m/ssut/py-hanspell (검색일: 2020년 8월 21일). https://github.com/haven-jeon/PyKoSpacing (검색일: 2020년 8월 21일).

51) 본 연구는 꼬꼬마, 한나눔, 트위터 등의 다양한 형태소 분석기 가운데 꼬꼬마 형태소 분석기를 사용하는 것으로 결정하기까지 여러 차례의 시도를 거쳤다. 본 연구에서 기사와 댓글의 분석을 위해 꼬꼬마 형태소 분석기를 사용하는 것으로 결정한 데에는 다음의 이유가 존재한다. 먼저 꼬꼬마 형태소 분석기는 총 56가지의 품사를 지원하고 있기 때문에 꼬꼬마 형태소 분석기를 사용할 경우 9가지 혹은 22가지의 품사를 지원하는 한나눔, 19가지의 품사를 지원하는 트위터 형태소 분석기보다 더욱 구체적인 형태소 분석이 가능하다. 또한, 이처럼 꼬꼬마 형태소 분석기가 상대적으로 더욱 구체적인 형태소 분석을 가능하게 하기 때문에 꼬꼬마 형태소 분석기를 사용했을 경우, 한나눔, 트위터 형태소 분석기를 사용했을 경우에 도출되는 결과에 비해 훨씬 더 연구자가 해석하기에 용이하고 정확한 결과를 도출된다. 형태소 분석기 간의 품사 분석은 다음을 참조. https://docs.google.com/spreadsheets/d/1OGAjUvalBuX-oZvZ_-9tEfYD2gQe7hTGsgUpiiBSXI8/edit#gid=0 (검색일: 2020년 10월 15일).

52) 길호현, "텍스트마이닝을 위한 한국어 불용어 목록 연구," 『우리말글』, Vol. 78, (2018), pp. 1-25. 참고로 길호현(2018)의 경우 세종 품사 태그를 기준으로 불용어의 품사를 규정하고 있다. 하지만 본 연구에서 사용하는 꼬꼬마 형태소 분석기의 경우 꼬꼬마 형태소 분석기에서 자체적으로 품사를 규정하고 있는데, 꼬꼬마 형태소 분석기에서 규정하고 있는 품사와 세종 품사 태그 간에는 다소 차이가 존재하여, 본 연구에서는 길호현(2018)이 제시하는 불용어를 사용하되 불용어의 품사는 고려하지 않고 사용하였다. 이처럼 언어학적인 측면에서의 지식이 결여됨으로 인해 길호현(2018)의 의도에 맞게 사용되지 않은 부분 곧, 연구의 한계가 존재함을 밝힌다.

하지만 토큰화된 댓글에 대해서는 사람들의 의견과 감정을 최대한 읽어낼 수 있도록 보통명사, 고유명사뿐만 아니라 동사와 형용사도 함께 추출하였다. 이후 추출된 자료에 대해서는 등장 횟수가 1회에 그친 단어들을 제거하였는데 이는 등장 횟수가 1회인 단어들의 중요도가 상대적으로 낮다고 판단하였기 때문이다. 위 설명된 자료 전처리 과정은 앞서 제시한 [표 2], [표 3] 자료들에 대해 모두 실시하였다.

3. 텍스트 마이닝 방법

이번에는 본 연구에서 사용하고자 하는 텍스트 마이닝 방법들에 대해 간략히 소개하고 이를 활용한 방식에 대해 설명하겠다. 본 연구에서는 잠재디리클레할당(Latent Dirichlet Allocation, LDA) 토픽 모델링(이하 LDA 토픽 모델링)을 이용한 토픽 분석, 워드투벡(Word2Vec)을 활용한 연관어 분석, 그리고 감성사전 및 지정된 단어를 활용한 감성분석으로 크게 세 가지의 텍스트 마이닝 방법을 사용하고자 한다. 또한, 앞서 설명하였듯이 기사 자료에 대해서는 토픽 분석을 시도하고 댓글 자료에 대해서는 연관어 분석과 감성분석을 진행하되, 추가적으로 기사 자료를 활용한 연관어 분석과 댓글 자료를 활용한 토픽 분석을 함께 실시하여 해당 결과에 대한 해석의 난해함도 함께 보여줄 것이다.

1) 토픽 분석: LDA 토픽 모델링

토픽 모델링(topic modeling)은 직역을 통해서도 알 수 있듯이 주제를 모형화 하는 것에 목적을 둔다. 텍스트를 수집하되 대량으로 수집할 경우 단시간에 모든 자료를 수작업을 통해 주제별로 분류해내기는 어렵다. 하지만 비지도 기계학습(unsupervised machine learning)에 속하는 토픽 모델링을 활용할 경우 대량의 텍스트를 효율적으로 분류할 수 있다. 이로 인해 사람은 토픽 모델링을 통해 출력되는 결과를 활용하여 다량의 텍스트에 대한 대략적인 내용을 쉽게 파악할 수 있다. 본 연구는 다양한

토픽 모델링 기법 가운데 높은 활용도를 보이고 있는 LDA 토픽 모델링을 사용하고자 한다.

줄리아 실기(Julia Silge)와 데이비드 로빈슨(David Robinson)은 LDA 토픽 모델링에 대해서 수학적으로 이해하지 않아도 "모든 문서는 토픽들의 혼합체이다.", "모든 토픽은 단어들의 혼합체이다."라는 두 가지 원리를 통해 LDA 토픽 모델링을 이해할 수 있다고 설명하였다.[53] 다시 말해, LDA 토픽 모델링은 "문서에서 관측되는 단어가 잠재된 토픽을 생성하는 과정에서 확률적으로 관측되었다고 가정하는" '생성적 확률 모형(generative probabilistic model)'이자, 단어-토픽-문서 간의 배속적(embedded) 또는 위계적(hierarchical) 관계, 그리고 사전확률과 관측된 데이터에 의해 구성되는 사후확률이 반영된 '3층 위계적 베이지안 모형(three-level hierarchical Bayesian model)'인 것이다.[54] 이와 같은 LDA 토픽 모델링의 원리에 대해서 윤태일・이수안(2018)은 "글쓰기 과정을 거꾸로 되밟아 가는 방법"으로 비유했는데, 이는 우리가 어떤 주제로 글을 쓸지 고민하는 과정에서 먼저 주제를 구상하고 그 다음 주제를 이루는 데 핵심이 되는 단어들을 구상 및 배치하는 전체적인 과정을 LDA 토픽 모델링은 반대로 뒤집어 토픽을 추론하고자 하기 때문이다.[55]

본 연구는 LDA 토픽 모델링의 결과를 도출하기 위해서 파이썬의 'gensim' 패키지를 사용할 것이다.[56] 일반적으로 토픽 모델링을 통해

---

53) 줄리아 실기・데이비드 로빈슨 지음, 박진수 옮김, 『R로 배우는 텍스트마이닝』, (서울: 제이펍, 2019), pp. 106-107. LDA 토픽 모델링 개발에 대한 자세한 내용은 이하를 참조. D. M. Blei, A. Y. Ng, M. I. Jordan, "Latent Dirichelet Allocation," *Journal of Machine Learning*, Vol. 3, (January 2003), pp. 993-1022.

54) 백영민, 『R을 이용한 텍스트 마이닝』, (서울: 한울아카데미, 2017), pp. 198-202. 재인용: D. M. Blei, A. Y. Ng, M. I. Jordan, 2003, pp. 996-997.

55) 윤태일・이수안, 『파이썬으로 텍스트 분석하기』, (고양: 늘봄, 2018), p. 217.

56) https://radimrehurek.com/gensim/models/ldamodel.html (검색일: 2020년 8월 13일). 'gensim' 패키지를 이용해 도출된 토픽 모델링의 결과는 'pyLDAvis' 패키지를 이용해 시각화할 수 있고 결과에 대한 이해를 돕는다. 자세한 내용은 다음을 참고. C. Sievert and K. Shirley. "LDAvis: A method for visualizing and interpreting topics," *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, (2014) pp. 63-70. 'pyLDAivs'를 이용한 LDA 토픽 모델링 결과 시각화 예시는 본 연구의 부록 [그림 9-1], [그림 9-2]를 참고.

도출되는 결과는 텍스트 전처리 과정과 토픽 모델링 과정에서의 토픽 개수 설정에 큰 영향을 받는다.[57] 따라서 본 연구는 최적의 토픽 개수를 결정하기 위해 혼란도(perplexity score)와 토픽의 일관성 점수(coherence score)를 참고하되 주로 토픽의 일관성 점수를 중점적으로 참고할 것임을 밝힌다.[58]

---

57) 본 연구는 토픽의 개수 설정 파라미터 이외에도 두 개의 파라미터를 추가하여 매번 LDA 토픽 모델링에 있어서 총 세 개의 파라미터를 사용하였다. 청크사이즈(chunksize)는 매 훈련에 사용되는 문서의 양을 의미하며, 본 연구는 사용하는 자료마다 전체 문서 크기를 확인한 뒤 전체 문서 크기의 1/5로 설정하였다. 패시즈(passes)는 코퍼스를 통한 훈련 횟수를 의미하며, 본 연구에서는 10회로 설정하였다.

58) 혼란도가 감소한다는 것은 확률 모델이 결과값을 적절하게 예측하고 있다는 것을 의미하기 때문에, 혼란도가 낮아지고 토픽의 일관성 점수가 높아지는 토픽 개수의 지점이 연구자가 토픽 모델링의 결과를 해석하는 데 있어 최적에 해당되는 토픽 개수라고 볼 수 있다(서대호, 『잡아라! 텍스트 마이닝 with 파이썬』, (서울: BJpublic, 2019), pp. 97-98). 하지만 혼란도에 로그(log)를 취할 경우에는 로그(log)를 취한 혼란도가 높은 값을 보일수록 확률 모델이 관측된 값을 적절하게 예측하고 있다는 것을 의미한다(구글의 'gensim' 패키지 대화를 참고. https://groups.google.com/g/gensim/c/uQxQiR2oC98 (검색일: 2020년 10월 23일)). 본 연구는 혼란도가 사람이 토픽 모델링의 결과를 해석하는 데 있어서 적절한 도움을 제공하지 않을 수도 있다는 연구 결과를 기반으로, 혼란도보다 토픽의 일관성 점수를 더욱 중점적으로 참고하여 토픽의 개수를 정하였다(J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, D. M. Blei, "Reading Tea Leaves: How Humans Interpret Topic Models," *Advances in neural information processing systems*, (2009), pp. 288-296). 또한, 최적의 토픽 개수를 설정할 때 고려되는 혼란도는 더 이상 변동하는 추세를 보여주지 않고 적절히 수렴되는 지점인데(서대호, 2019, p. 97), 본 연구에서는 토픽 개수가 40이 되어도 로그(log)를 취한 혼란도가 수렴되는 지점을 발견할 수 없었고 토픽 개수가 0일 때부터 시작해 지속적으로 감소하는 추세만을 확인할 수 있었다. 또한 로그(log)를 취한 혼란도는 앞서 설명하였듯이 높은 값을 보여줄수록 적절한 토픽 모델링을 시도할 수 있는데 본 연구에서는 주로 토픽 개수가 10을 넘기 이전에 가장 높은 값을 보여주었다. 하지만 본 연구는 토픽 모델링을 시도할 때 단순 빈도 분석과 차이를 두기 위해 10 이상의 토픽 개수를 지향하고 있다는 점에서 로그(log)를 취한 혼란도를 기준으로 최적의 토픽 개수를 찾기 어렵다고 판단하였다. 참고로 토픽의 일관성 점수를 계산하는 데에는 'gensim' 패키지에서 'c_v', 'u_mass', 'c_uci', 'c_npmi' 총 네 가지의 계산방식을 제공하고 있으며, 본 연구는 그 가운데 사람이 매기는 것과 가장 비슷한 결과를 낸다고 밝혀진 'c_v' 계산방식을 사용하였다(M. Röder, A. Both, and A. Hinneburg, "Exploring the Space of Topic Coherence Measures," *Proceedings of the eighth ACM international conference on Web search and data mining*, (2015), pp. 399-408).

더불어 본 연구는 토픽 모델링의 경우 최적의 토픽 개수를 찾을 때 단순 빈도 분석과의 차별성을 두기 위해 10 이하의 토픽 개수를 배제할 것이다. 또한 토픽의 개수가 20을 초과하게 될 경우 동일한 토픽에 대한 과도한 세분화가 이루어져 연구자의 입장에서 결과에 대한 해석이 난해해질 수 있고, 또 지면상 할애의 어려움이 존재하여 토픽의 개수가 10에서 20 사이일 때 가장 높은 일관성 점수를 보여주는 토픽의 개수를 최적의 토픽 개수로 판단하고자 한다.

그 결과, 조선일보 기사의 경우 10, 중앙일보 기사의 경우 18, 한겨레신문과 경향신문의 기사의 경우 20으로 평균적으로 17개의 최적의 토픽 개수를 보여주었다. 또한 조선일보와 중앙일보의 댓글의 경우 17, 한겨레신문 댓글의 경우 20, 그리고 경향신문 댓글의 경우는 19로 평균적으로 18.25개의 최적의 토픽 개수를 보여주었다. 이로써 본 연구는 기사 자료를 활용한 토픽 모델링에는 최적의 토픽 개수로 17을 설정하고 댓글 자료를 활용한 토픽 모델링에는 최적의 토픽 개수로 18을 설정한다. 해당 내용은 부록의 [그림 1]부터 [그림 8]까지의 자료들을 통해 확인할 수 있다.

2) 연관어 분석: 워드투벡

LDA 토픽 모델링은 구조화되지 않은 대량의 텍스트를 구조화한다는 점에서 장점을 지니고 있지만, 단어들의 순서가 아닌 단어들의 존재 유무만 가정하는 단어 교환성 법칙에 의거하고 있기 때문에, 단어 수준에서의 문맥을 고려하지 못한다는 데 단점이 안고 있다.[59] 이에 본 연구는 LDA 토픽 모델링의 문장 수준 임베딩 결과에서 한 걸음 더 나아가, 단어 수준 임베딩을 시도하는 워드투벡을 활용하고자 한다.[60] 이를 통해 동일한 단어일지라도 각기 신문사의 기사와 댓글에서 어떠한 문맥적 차

59) 서대호, 2019, p. 94.
60) 한국어 임베딩과 관련된 자세한 내용은 다음의 책을 참고. 이기창, 『한국어 임베딩: 자연어 처리 모델의 성능을 높이는 핵심 비결 Word2Vec에서 ELMo, BERT까지』, (서울: 에이콘출판사, 2019).

이가 나타나는지 알아볼 수 있을 것이다.

본 연구는 파이썬의 'gensim' 패키지를 활용하여 워드투벡 모델 학습을 진행할 것이다.[61] 이후 워드투벡의 결과를 파이썬의 'scikit-learn' 패키지 내 't-SNE(t-Stochastic Nearest Neighbor)'를 이용하여 2차원으로 축소시킨 뒤 용이한 해석을 위해 200개의 단어로 시각화를 시도할 것이다.[62]

3) 감성분석

"감성분석(sentiment analysis) 혹은 오피니언 마이닝(opinion mining)"이란 텍스트에 표현된 "사람들의 의견(opinions), 감성

---

61) https://radimrehurek.com/gensim/models/word2vec.html (검색일: 2020년 10월 21일). 워드투벡 모델 학습에서는 여러 가지 파라미터 설정이 요구된다. 본 연구에서는 단어 벡터의 차원을 의미하는 사이즈(size)는 300으로, 하나의 문장 안에서 예측하려는 단어와 타 단어 간의 최대 거리를 뜻하는 윈도우(window)는 10으로 설정하였다. 또한, 단어의 최소 등장 빈도수를 정하여 모델 학습에 해당 단어를 포함할 것인지를 결정하는 최소카운트(min_count)의 경우, 활용하려는 자료의 종류가 기사일 경우 50으로 설정하였고, 활용하려는 자료의 종류가 댓글일 경우 5로 설정하였다. 모델 학습에 몇 개의 프로세스를 사용할지 결정하는 워커즈(workers)는 4로 설정하였고, 'sg=1', 'hs=0'으로 설정하여 스킵그램(skip-gram)과 네거티브 샘플링(negative sampling)을 사용하였다. 워드투벡의 'CBOW(Continuous Bag of Words)'와 스킵그램(skip-gram) 알고리즘, 위계적 소프트맥스(hierarchical softmax)와 네거티브샘플링(negative sampling) 학습 방법에 대한 내용은 다음을 참고. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *In Advances in neural information processing systems*, (2013), pp. 3111-3119; T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," (2013), https://arxiv.org/abs/1301.3781 (검색일: 2020년 10월 25일).

62) https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html (검색일: 2020년 10월 25일) 't-SNE(T-Distributed Stochastic Neighboring Embedding)'는 고차원의 데이터를 저차원으로 축소시킬 수 있는 수학적 기법 가운데 하나이다. 이와 같은 차원 축소 기법은 데이터의 용량 증대가 오히려 모델의 학습을 저해할 수 있다는 차원의 저주(curse of dimensionality)를 해결하기 위해 개발되었다. 't-SNE'의 개발과 기존의 차원 축소 기법에 비해 상대적으로 뛰어난 't-SNE' 성능과 관련된 내용은 다음의 연구를 참고. L. van der Maaten, G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, Vol. 9, (2008), pp. 2579-2605.

(sentiments), 평가(evaluations), 태도(attitudes), 판단(appraisal), 영향(affects), 관점(views), 감정(emotions), 주관성(subjectivity)” 등을 컴퓨터를 활용해 파악하는 방법이다.[63] 이러한 감성분석은 크게 두 가지의 자동화된 방법들로 수행되어 왔는데, 하나는 텍스트에 포함된 단어 혹은 구의 의미로부터 텍스트의 경향성을 계산하는 어휘기반 접근 방법(lexicon-based approach)이고, 다른 하나는 텍스트 내 분류가 되어있는(labeled) 예시들을 통해 분류기(classifier)를 제작하여 주어진 텍스트를 분류하는 지도 기계학습(supervised machine learning)과 심화학습(deep learning)을 이용한 방법이다.[64]

다시 말해, 어휘기반 접근 방법은 먼저 특정 감성을 나타내는 단어들을 찾고 분류하여 감성사전을 제작한 뒤, 쪼개어진(tokenized) 텍스트와 감성사전에 포함되어 있는 단어들의 감성 점수 대응을 통해 감성을 판단하는 방법이라면,[65] 지도 기계학습과 심화학습을 활용한 접근 방법은 텍스트 가운데 긍정 혹은 부정과 같이 정답이 매겨진(labeled) 훈련 모델(train set)을 이용하여 학습 모델을 생성하고 이를 통해 나머지 텍스트(test set)의 정답을 예측하고자 하는 방법인 것이다.[66]

이처럼 지도 기계학습과 심화학습을 이용한 접근 방법의 경우 어휘기반 접근 방법에 비해 어휘의 표면적인 의미 너머에 숨어있는 문맥을 반영할 수 있다는 데 강점을 지닌다. 특히, 어휘기반 접근 방법과 지도 기계학습을 이용한 접근 방법은 어떠한 성격의 텍스트를 기반으로 감성사전과 분류기를 제작하느냐에 따라 주어진 텍스트에 대한 분석이 달라지는 영역 적응(domain adaption)의 문제를 안고 있는 반면, 심화학습을 이용한 접근 방법은 영역 의존성 문제를 비교적 해소해 줄 수 있어 최근 큰 주목받고 있다.[67]

---

63) B. Liu, “Sentiment analysis and opinion mining,” *Synthesis lectures on human language technologies*, Vol. 5, No. 1 (2012), p. 7.
64) M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, “Lexicon-Based Methods for Sentiment Analysis,” *Computational linguistics*, Vol. 37, No. 2 (2011), p. 268.
65) 고승연, “SR3. 송민 연세대 교수 인터뷰‘감정 분석 뜬다는데 해볼까’는 위험 어떤 문제에 왜 필요한지 정의 먼저,” 『동아비즈니스리뷰』, 11월 Issue 2, 2018.
66) 서대호, 2019, p. 108.

하지만 본 연구의 경우 지도 기계학습과 심화 학습을 이용한 감성분석을 시도하기 위한 분류 작업은 1인 연구자가 구현하기에는 물리적 비용과 객관성 및 정확도 유지의 측면에서 어려움이 있다고 판단하였기에, 감성사전을 이용한 감성분석과 더불어 지정된 단어를 이용한 감성분석을 시도하고자 한다.

(1) 감성사전을 이용한 감성분석

감성사전(sentiment or opinion lexicon)이란 사람들의 주관적인 느낌, 감성 혹은 의견을 드러내기 위해 사용된 어휘들의 집합체이다. 감성사전을 이용한 감성분석은 감성사전에 어떤 어휘들이 등재되어 있는지에 따라서 분석 결과가 다르게 도출될 수 있다. 즉, 어떤 영역(domain)의 텍스트를 기반으로 감성사전을 구축했는지에 따라서 긍정, 부정, 중립을 나타내는 어휘들이 달라질 수 있는 특수성의 문제가 존재한다.[68] 이로 인해 주어진 텍스트에 따라서 텍스트의 특성을 최대한 반영할 수 있는 특수한 감성사전을 구축하는 것이 가장 정확한 결과를 도출할 수 있는 방법이겠지만 시간과 비용이 과다하게 드는 데 비해 재사용의 어려움이 존재해, 한국어 범용 사전을 구축하려는 시도들이 있었다.[69]

대표적인 한국어 감성사전으로는 서울대학교 언어학과 컴퓨터언어학 연구실에서 개발된 'KOSAC(Korean Sentiment Analysis Corpus, 이하 서울대학교 감성사전)'과 군산대학교 소프트웨어공학과 데이터인텔리전스랩(Data Intelligence Lab)에서 개발된 'KnuSentiLex(이하 군산대학교 감성사전)'가 있다.[70] 두 사전은 모두 특정한 도메인에 맞춰져서 개발되

---

67) 서상현・김준태, 2016, p. 9.
68) 박상민・나철원・최민성・이다희・온병원, "Bi-LSTM 기반의 한국어 감성사전 구축 방안,"『지능정보연구』, Vol. 24, No. 4 (2018), p. 220.
69) 박상민・나철원・최민성・이다희・온병원, 2018, p. 220.
70) 서울대학교 감성사전, 군산대학교 감성사전 외에도 '오픈한글'이 있다. 하지만 '오픈한글'은 현재 잠정적으로 서비스를 종료한 상태이다. http://openhangul.com/ (접속일: 2020년 10월 23일). 오픈한글 개발과 관련된 논문은 다음을 참조. 안정국・김희웅, "집단지성을 이용한 한글 감성어 사전 구축,"『지능정보연구』, Vol. 21, No. 2 (2015), pp. 49-67.

지 않고 보편적인 특성을 지닌 사전으로 개발되었고, 최종적으로는 특정한 도메인에 맞춰진 감성사전의 기초 자료로 활용되는 것에 목적을 두고 있다는 점에서 공통점을 지닌다.

한편, 서울대학교 감성사전의 경우 꼬꼬마 형태소 분석기와의 연계를 통해 감성을 드러내는 단어와 해당 단어의 형태소가 감성 사전 내에 함께 구성되어 있는 반면, 군산대학교 감성사전의 경우 형태소에 대한 분석 없이 단어만으로 구성되어 있다는 점에서 차이점이 존재한다.

본 연구는 두 가지의 한국어 감성사전 가운데 군산대학교 감성사전을 사용하여 댓글을 이용한 감성분석을 진행할 것이다.[71] 이는 서울대학교 감성사전을 공식 홈페이지에서 내려 받아 사용할 경우, 유니그램(uni-gram), 바이그램(bi-gram), 트라이그램(tri-gram) 모두 고려되어야 하는데 바이그램과 트라이그램에 대한 고려는 본 연구에서 사용하고자 하는 꼬꼬마 형태소 분석기에 대한 재구성 작업을 요구한다고 판단하였기 때문이다.[72]

---

71) 군산대학교 감성사전은 유니그램(uni-gram), 바이그램(bi-gram), 엔그램(n-gram, n은 최대 8), 축약어, 이모티콘을 포함하여 총 14,843개의 감성단어를 제공한다. 14,843개의 단어는 매우 긍정 2,597개, 긍정 2,266개, 중립 154개, 부정 5,029개, 매우 부정 4,797개로 구성되어 있다. 하지만 깃허브를 통해 군산대학교 감성사전을 다운로드 받을 경우, 14,843개가 아닌 총 14,855개의 감성단어로 구성되어 있음을 확인할 수 있다. 본 연구는 14,855개의 단어 가운데 중복된 세 단어('버릇없이', '적극적이다', '울컥하다')를 제외한 14,852개의 단어를 분석에 사용하고, -1이 아닌 +1로 매겨진 '적극적이다'라는 용어와 -1이 아닌 -2로 매겨진 '울컥하다'라는 용어를 사용하였다. 깃허브는 다음을 참조. https://github.com/park1200656/KnuSentiLex (검색일: 2020년 10월 20일).

72) 서울대학교 감성사전은 3,476개의 유니그램(uni-gram), 6,579개의 바이그램(bi-gram), 6,307개의 트라이그램(tri-gram)으로 총 16,362개의 감성단어를 제시하고 있다. 자세한 내용은 다음을 참조. 김문형・장하연・조유미・신효필, "KOSAC(Korean Sentiment Analysis Corpus): 한국어 감정 및 의견 분석 코퍼스,"『한국정보과학회 학술발표논문집』, (2013.6), pp. 650-652. 현재 서울대학교 감성사전은 초창기의 파이썬 라이브러리 형태로는 제공되고 있지 않으며, 서울대학교 감성사전이 등재된 공식 홈페이지에서만 내려 받을 수 있다.

(2) 지정된 단어를 이용한 감성분석

본 연구는 또한 감성사전을 이용한 감성분석 외에도 북한이탈주민에 대한 댓글 작성자들의 태도를 파악할 수 있다고 판단되는 단어들을 지정하여 추가적인 감성분석을 시도하고자 한다. 이는 앞서 살펴본 서울대학교 감성사전이나 군산대학교 감성사전의 경우 자료의 특수성을 초월하는 보편성을 지향하여 제작된 사전이기 때문에, 자료의 특수성에 맞게 특화된 감성분석을 진행하는 데 있어서 어려움이 존재하기 때문이다.[73] 감성분석에 이용될 지정 단어는 연구 설계에서 언급한 단어들 곧, '북괴', '인권', '동포', '보호', '간첩', '위장', '빨갱이'를 사용하고자 한다.

73) 박종희・박은정・조동준(2015)의 경우 연구자들의 전문성에 의거하여 핵, 남조선, 미제에 대한 북한의 태도를 측정하는 데 핵심이 된다고 판단되는 단어들을 추출한 뒤 시간에 따라 변화해왔던 핵, 남조선, 미제에 대한 북한의 태도를 살펴보았다(박종희・박은정・조동준, "북한 신년사(1946-2015)에 대한 자동화된 텍스트 분석,"『한국정치학회보』, Vol. 49, No. 2 (2015), pp. 42-50).

제 4 장 분석 결과

제 1 절 기사 토픽 분석 결과[74)]

토픽 모델링의 결과는 각각의 토픽을 구성하는 단어가 무엇인지, 그리고 그 토픽이 분석하려는 전체 텍스트 가운데 어느 정도의 비중을 차지하고 있는지에 대해서만 알려줄 뿐, 해당 토픽이 무엇에 관한 것인지에 대해서는 알려주지 않는다. 이에 따라 본 연구는 기사 자료를 활용해 도출된 LDA 토픽 모델링의 결과 곧, 부록의 [표 1]부터 [표 4]까지의 결과들을 기반으로 해당 토픽이 어떠한 내용으로 요약 될 수 있는지 이하 [표 4], [표 5], [표 6], [표 7]로 정리하고자 한다. 또한 토픽 내용에 대한 요약을 통해 보수 성향 신문사들의 기사와 진보 성향 신문사들의 기사 간 공통점과 차이점, 그리고 동일한 보수 성향 신문사들의 기사 내에서 혹은 동일한 진보 성향 신문사들의 기사 내에서 발견되는 토픽의 공통점과 차이점에 주목함으로써 각 신문사들이 어떤 것에 중점을 두고 기사를 게재하고자 하였는지 알아볼 것이다. 그 과정에서 특히, 토픽의 비중이 5% 미만에 해당되는 토픽보다는 5% 이상의 결과를 차지한 토픽을 중심으로 살펴볼 것이다.

---

74) 본 연구는 기사의 내용을 파악하기 위해 기사 자료를 활용한 토픽 분석을 시도하는 것 외에도 연관어 분석을 추가적으로 실시하였다. 하지만 연관어 분석 결과가 토픽 분석 결과만큼 연구자가 해석하기에 용이한 결과를 제시하지 못한다고 판단하였기 때문에 부록으로 첨부하였다. 기사 자료를 활용한 연관어 분석 결과는 부록의 [그림 10], [그림 11], [그림 12], [그림 13]을 참고.

[표 4] 조선일보 기사 토픽 분석 결과 해석

	비중	토픽의 내용
1	11%	북한 내부 정치 및 경제 상황에 대한 소식
2	10.7%	동북아시아 국가들 간의 정상 회담 관련 논의
3	10.3%	북한이탈주민의 탈북 경로
4	9.9%	북한의 인권 문제 해결 촉구 논의
5	9%	한국 내 북한이탈주민의 대입 전형
6	8%	북한이탈주민의 이야기를 소재로 삼은 작품들
7	7.9%	보수 성향을 지닌 북한이탈주민들의 대북 전단 살포 관련 논란
8	5.7%	서울시 간첩 사건 수사 진행 과정
9	5.4%	북한이탈주민들의 한국 정치 참여
10	5.1%	고향과 가족에 대한 그리움을 담은 북한이탈주민들의 이야기
11	4.9%	(1) 해상 탈북, (2) 북한의 핵실험 관련 정보 등
12	4.9%	북한이탈주민을 대상으로 한 사회복지 서비스 및 지원 사업
13	3.3%	북한의 지도자에 대한 묘사 및 정보
14	1.7%	모란봉클럽 프로그램을 통해 방송된 북한의 지도자의 친인척과 관련된 소식
15	1.1%	북한이탈주민의 스포츠 경기 출전
16	0.7%	북한 내 기독교 전파
17	0.4%	(1) 북한이탈주민 지원 사무소, (2) 북한이탈주민 출연과 관련된 방송 제작 등

[표 5] 중앙일보 기사 토픽 분석 결과 해석

	비중	토픽의 내용
1	20.8%	동북아시아 국가들 간의 정상 회담 관련 논의
2	13%	북한 인권 문제에 대한 국제사회의 주목
3	12.9%	북한매체를 통해 드러난 북한 정치 상황
4	10.8%	북한이탈주민들의 이야기를 소재로 한 영화 제작 관련
5	6.4%	북한이탈주민 정착 지원 및 보호 정책
6	5.7%	북한이탈주민 학생들의 교육 과정 관련
7	5.6%	북한이탈주민 대상 지원 사업
8	4%	관악구 북한이탈주민 모자 아사 사건
9	3.9%	북한의 군사적 도발 관련
10	3.3%	국정원의 간첩 수사
11	3.1%	북한의 지도부 관련 내부 소식
12	3%	북한이탈주민들의 한국 정치 참여
13	2.6%	북한의 새로운 사업 도전
14	2%	자유를 외치는 북한이탈주민들의 집회 관련
15	1.1%	(1) 한국 내 북한이탈주민에 대한 거주지 지원, (2) 한국 내 북한이탈주민의 대입 전형
16	0.9%	남북 간 포로 해결을 위한 소통 노력
17	0.9%	국제스포츠 출전에서의 남북 연합

[표 6] 한겨레신문 기사 토픽 분석 결과 해석

	비중	토픽의 내용
1	15.9%	북한이탈주민의 한국 정착 과정
2	15.8%	한반도를 둘러싼 북한, 한국, 중국, 미국 간 협력
3	14.7%	북한이탈주민을 대상으로 한 한국 내 다양한 분야의 지원
4	13.6%	북한이탈주민에 대한 간첩 의혹 조사 관련
5	7.3%	북한이탈주민과의 북한 이탈 관련 인터뷰
6	5.4%	보수 성향을 지닌 북한이탈주민들의 대북 전단 살포 관련 논란
7	4.9%	서울시 간첩 조작 사건 당사자 가족의 진술
8	4.1%	북한의 도발에 따른 개성 공단 폐쇄
9	3.8%	한국 내 북한이탈주민의 대입 전형
10	3.7%	평창남북평화영화제 상영작
11	3.5%	류경식당 종업원 집단탈북에 대한 박근혜 정부의 기획 논란
12	3.2%	유엔 인권이사회의 북한인권결의안 연속 채택
13	1.8%	세월호 반대 집회를 위해 북한이탈주민을 동원한 어버이연합
14	1%	탈북을 소재로 한 문학
15	0.6%	한국을 방문한 프란치스코 교황에게 북한 인권과 관련해 기도를 요청한 북한이탈주민들
16	0.4%	(1) 영화 <국가대표2>, (2) <연극> 목란언니, (3) 북한이탈주민 가장 사고사, (4) 성소수자 북한이탈주민 등
17	0.3%	(1) 북한이탈청소년 은주, (2) 호주에 체포된 북한 공작원 등

[표 7] 경향신문 기사 토픽 분석 결과 해석

	비중	토픽의 내용
1	12.7%	북한의 납치와 관련된 인권 문제
2	11.5%	서울시 간첩 조작 사건 수사 진행 과정
3	10.4%	북한 인권 문제에 대한 국제사회의 주목
4	9.3%	북한이탈주민 대상 국가 및 지역적 차원의 지원 사업
5	9.1%	북한이탈주민을 소재로 한 예술작품
6	8.4%	라오스 북한이탈주민 북송사태
7	6.5%	박근혜 전 대통령에 대한 북한이탈주민의 지지
8	5.8%	뉴라이트전국연합과 북한이탈주민
9	5.7%	북한이탈주민 대상 지원 사업 내 체계
10	5.1%	한국 내 북한이탈주민의 대입 전형
11	5.1%	북한이탈주민이 한국 정착 과정에서 겪게 되는 어려움
12	3.2%	북한이탈주민을 위한 공적·사적 영역의 지원
13	2.2%	장경욱 변호사와의 국가 폭력 관련 인터뷰
14	1.9%	(1) 한국 사회 내 북한이탈주민의 범죄, (2) 기무사 사찰과 이석기 내란음모 사건, (3) 2004년 북한이탈주민 집단입국 등
15	1.5%	(1) 북한이탈주민의 해외 이민, (2) 영화 <댄스 타운> 등
16	1.2%	인신매매의 위험에 처해있는 북한이탈주민
17	0.5%	(1) 한국 내 북한이탈주민에 대한 재산 상속 인정, (2) 예술로서의 분단 등

1. 신문사 성향에 따른 기사 토픽 분석 결과 비교

1) 공통점

보수 성향을 띠는 신문사들과 진보 성향을 띠는 신문사들 간에는 몇 가지 공통점과 차이점이 존재한다. 첫 번째 공통점은 조선일보, 중앙일보, 경향신문 모두 북한의 인권 탄압 문제에 주목하고 있다는 점이다. [표 4] 조선일보 기사 토픽 분석 결과에서는 북한의 인권 문제 해결 촉구 논의가 9.9%, [표 5] 중앙일보 기사 토픽 분석 결과에서는 북한 인권 문제에 대한 국제사회의 주목이 13%, [표 7] 경향신문 기사 토픽 분석 결과에서는 북한의 납치와 관련된 인권 문제가 12.7%, 라오스 북한이탈주민 북송사태가 8.4%, 그리고 북한 인권 문제에 대한 국제사회의 주목이 10.4%로 나타났다.

두 번째 공통점은 보수 성향의 신문사들과 진보 성향의 신문사들 모두 북한이탈주민의 한국 정착 문제에 대해 주목하고 정착을 위한 여러 가지 지원 사업에 대해 언급하고 있다는 점이다. [표 4] 조선일보 기사 토픽 분석 결과에서는 한국 내 북한이탈주민의 대입 전형이 9%, 그리고 북한이탈주민을 대상으로 한 사회복지 서비스 및 지원 사업이 4.9%, [표 5] 중앙일보 기사 토픽 분석 결과에서는 북한이탈주민 정착 지원 및 보호 정책이 6.4%, 북한이탈주민 학생들의 교육 과정 관련이 5.7%, 그리고 북한이탈주민 대상 지원 사업이 5.6%, [표 6] 한겨레신문 기사 토픽 분석 결과에서는 북한이탈주민을 대상으로 한 한국 내 다양한 분야의 지원이 14.7%, [표 7] 경향신문 기사 토픽 분석 결과에서는 북한이탈주민 대상 국가 및 지역적 차원의 지원 사업이 9.3%, 한국 내 북한이탈주민의 대입 전형이 5.1%로 나타났다.

세 번째 공통점은 보수 성향의 신문사들과 진보 성향의 신문사들 모두 북한이탈주민들이 북한 이탈을 선택하기 이전에 겪었던 북한에서의 생활과 북한에서 이탈하는 과정에 대해 기사로 다루고 있다는 점이다. [표 4] 조선일보 기사 토픽 분석 결과에서는 북한이탈주민의 탈북 경로

가 10.3%, 북한이탈주민의 이야기를 소재로 삼은 작품들이 8%, 그리고 고향과 가족에 대한 그리움을 담은 북한이탈주민들의 이야기가 5.1%, [표 5] 중앙일보 기사 토픽 분석 결과에서는 북한이탈주민들의 이야기를 소재로 한 영화 제작 관련이 10.8%, [표 6] 한겨레신문 기사 토픽 분석 결과에서는 북한이탈주민과의 북한 이탈 관련 인터뷰가 7.3%, [표 7] 경향신문 기사 토픽 분석 결과에서는 북한이탈주민을 소재로 한 예술작품이 9.1%로 나타났다.

네 번째 공통점은 조선일보, 중앙일보, 한겨레신문 모두 북한이탈주민을 둘러싼 주변 국가들의 반응과 관련된 기사들을 게재해 왔다는 점이다. [표 4] 조선일보 기사 토픽 분석 결과에서는 동북아시아 국가들 간의 정상 회담 관련 논의가 10.7%, [표 5] 중앙일보 기사 토픽 분석 결과에서는 동북아시아 국가들 간의 정상 회담 관련 논의가 20.8%, [표 6] 한겨레신문 기사 토픽 분석 결과에서는 한반도를 둘러싼 북한, 한국, 중국, 미국 간 협력이 15.8%로 나타났다.

다섯 번째 공통점은 조선일보, 한겨레신문, 경향신문 모두 한국 사회 내에서 오랜 시간 동안 발생해온 북한이탈주민 위장 간첩 문제에 대해 언급해왔다는 점이다. [표 4] 조선일보 기사 토픽 분석 결과에서는 서울시 간첩 사건 수사 진행 과정이 5.7%, [표 6] 한겨레신문 기사 토픽 분석 결과에서는 북한이탈주민에 대한 간첩 의혹 조사 관련이 13.6%, [표 7] 경향신문 기사 토픽 분석 결과에서는 서울시 간첩 조작 사건 수사 진행 과정이 11.5%로 나타났다.

여섯 번째 공통점은 조선일보와 한겨레신문에 일부 북한이탈주민들로 형성된 특정 단체에서 대북 전단을 살포한 것과 관련된 기사들이 게재되어 왔다는 점이다. [표 4] 조선일보 기사 토픽 분석 결과에서는 보수 성향을 지닌 북한이탈주민들의 대북 전단 살포 관련 논란이 7.9%, [표 6] 한겨레신문 기사 토픽 분석 결과에서는 보수 성향을 지닌 북한이탈주민들의 대북 전단 살포 관련 논란이 5.4%로 나타났다.

2) 차이점

보수 성향의 신문사들과 진보 성향의 신문사들의 토픽 분석 결과는 위 언급된 공통점들뿐만 아니라 차이점들 역시 보여 준다. 먼저, 보수 성향의 신문사들은 북한이탈주민들의 증언을 통해 얻을 수 있는 북한 내부 정보와 관련된 기사들을 게재해왔다. [표 4] 조선일보 기사 토픽 분석 결과에서는 북한 내부 정치 및 경제 상황에 대한 소식이 11%, [표 5] 중앙일보 기사 토픽 분석 결과에서는 북한매체를 통해 드러난 북한 정치 상황이 12.9%로 나타났다.

또한, 조선일보는 한국에 정착한 북한이탈주민들 가운데 한국의 제도권 정치에 참여한 북한이탈주민과 관련된 기사들을 게재해왔다. [표 4] 조선일보 기사 토픽 분석 결과에서는 북한이탈주민들의 한국 정치 참여가 5.4%로 나타났다.

한편, 진보 성향의 신문사들은 한국 사회 내에서 정착하는 과정 가운데 이방인으로 취급되며 특히, 북한과 한국 간의 특수한 정치·외교적 관계로 인해 이들이 사람으로서 대우받지 못했던 차별 현상에 주목하고 있다는 점에서 보수 성향의 신문사들과 큰 차이를 보여준다. [표 6] 한겨레신문 기사 토픽 분석 결과에서는 북한이탈주민의 한국 정착 과정이 15.9%, [표 7] 경향신문 기사 토픽 분석 결과에서는 북한이탈주민이 한국 정착 과정에서 겪게 되는 어려움이 5.1%로 나타났다.

또한, 경향신문은 한국에 정착한 북한이탈주민들의 보수적인 성향에 주목하여 이와 관련된 기사들을 게재하였다. [표 7] 경향신문 기사 토픽 분석 결과에서 박근혜 전 대통령에 대한 북한이탈주민의 지지가 6.5%, 뉴라이트전국연합과 북한이탈주민이 5.8%로 나타났다.[75]

---

75) 위 사용된 방법과 같이 보수 성향의 신문사들과 진보 성향의 신문사들 간의 차이점은 각각의 신문사들이 게재해온 기사들의 내용을 바탕으로 유사한 것끼리 묶어서 비교할 수 있을 뿐만 아니라, 각각의 신문사마다 가장 높은 비율을 보여준 토픽을 이용해서도 비교해 볼 수 있다. 하지만 본 연구의 경우 각각의 신문사마다 댓글이 시작된 시기를 바탕으로 기사 자료를 추출함으로써 각 신문사들의 기사 자료 수집 기간에 차이점이 발생하였기 때문에 완전한 토픽 순위 비교는 어려움을 알린다. 신문사들 간의 온전한 토픽 순위 비교가 필요할 경우, 부록의 [표 9], [표 10], [표 11], [표 12]를 참고. 부록의 [표 9], [표 10], [표 11], [표 12]는 부록의 [표 1], [표 2], [표 3], [표 4]과 달리, 네 개의 신문사

제 2 절 댓글 연관어 분석 결과[76)]

1. 신문사 성향에 따른 연관어 분석 결과 비교

1) 공통점

보수 성향의 신문사 댓글 연관어 분석 결과와 진보 성향의 신문사 댓글 연관어 분석 결과에서 가장 두드러지게 나타나는 공통점은 댓글 작성자들에게 북한이탈주민은 도움과 지원이 필요한, 어려운 상황에 놓여 있는 사람들이라는 생각이 나타난다는 점이다.

먼저, 조선일보 댓글을 이용한 연관어 분석 결과에서는 [그림 1]의 '탈북자' 단어 주변에 '도와주', '사람'이라는 단어가 가깝게 위치한 것을 확인할 수 있다. 이는 댓글 작성자들의 입장에서 북한이탈주민은 다른 사람의 도움이 필요한 존재로 비춰지고 있다는 점을 시사한다. 북한이탈주민에 대한 댓글 작성자들의 긍휼한 마음과 관심이 드러나는 대목이다.

또한, 한겨레신문 댓글을 이용한 연관어 분석 결과에서도 [그림 3]의 '극우' 위편에 존재하는 '지원', '국내', '이해', '힘들'이라는 단어들을 통해 댓글 작성자들이 어려움을 겪는 북한이탈주민을 국내적 차원에서의 지원과 이해가 필요한 대상으로 바라보고 있다는 점을 보여준다.

더불어, 경향신문 댓글 연관어 분석 결과에서도 [그림 4]의 '탈북자' 단어 좌측에 '정착', '난민', '이용', '사회', '노력'이라는 단어들이 나열되어 있다는 점을 통해 댓글 작성자들은 북한이탈주민들이 사회적으로 난

들의 공통된 기간 곧, 2005년 1월 9일부터 2020년 3월 1일까지의 기간을 대상으로 도출된 기사 본문 분석 결과이다. 더불어, 이와 같이 동일한 기간을 대상으로 보수 성향의 신문사와 진보 성향의 신문사 간의 북한 및 북한이탈주민에 대한 보도방식 비교 연구는 다음을 참고. 곽정래・이준웅, 2009, pp.196-217. 권호천, 2017, pp. 103-141.

76) 본 연구는 댓글의 내용을 통해 드러난 북한이탈주민에 대한 댓글 작성자들의 태도를 파악하기 위해 댓글 자료를 활용한 연관어 분석을 시도하는 것 외에도 토픽 분석을 추가적으로 실시하였다. 하지만 토픽 분석 결과가 연관어 분석 결과만큼 연구자가 해석하기에 용이한 결과를 제시하지 못한다고 판단하였기 때문에 부록으로 첨부하였다. 댓글 자료를 활용한 토픽 분석 결과는 부록의 [표 5], [표 6], [표 7], [표 8]을 참고.

민의 위치에 있다고 인식하고 있으며, 북한이탈주민이 정착할 수 있도록 사회적 차원에서의 노력을 기울이고, 그들이 이용할 수 있는 자원들을 마련해주어야 한다는 생각을 하고 있음을 알 수 있다.

2) 차이점

한편, 보수 성향의 신문사 댓글 연관어 분석 결과와 진보 성향의 신문사 댓글 연관어 분석 결과에서는 위 언급된 공통점들뿐만 아니라 신문사의 성향에 따른 차이점들 역시 나타났다. 그리고 그 차이점들은 같은 성향의 신문사라고 할지라도 신문사의 종류에 따라 다르게 나타나기도 하였다.

먼저, 조선일보 댓글을 이용한 연관어 분석 결과에서는 [그림 1]의 '간첩', '검증', '사상', '혐의'라는 단어가 '탈북자'라는 단어의 왼쪽 하단에 위치된 것을 확인할 수 있다. 이는 댓글 작성자이 북한이탈주민을 생각할 때 간첩 혐의를 함께 떠올린다는 것을 의미하며 북한이탈주민을 사상 검증이 필요한 존재로서 바라보고 있음을 의미한다.[77]

하지만 중앙일보 댓글을 이용한 연관어 분석에서는 [그림 2]의 '탈북자' 단어가 '김정일', '멀어지', '사상' 단어들과는 가깝게 위치하고 있는 반면, '의심' 단어와는 다소 멀리 위치하고 있다는 것에 주목해 볼 수 있다. 이와 같은 위치는 댓글 작성자들이 북한이탈주민을 떠올릴 때 북한이탈주민 스스로 북한의 사상을 거부하였기 때문에 북한으로부터의 이탈을 선택했을 것이라 생각하고 있음을 의미한다. 또한, 댓글 작성자들이 북한이탈주민에 대한 의심을 전혀 갖지 않고 있는 것은 아니지만 '의심'이라는 단어가 '김정일', '멀어지', '사상'이라는 단어보다 '탈북자' 단어

77) 결국 조선일보 댓글을 이용한 연관어 분석에서 확인할 수 있는 것은 댓글 작성자들이 북한이탈주민에 대한 긍휼한 마음과 의심 곧, 긍정적인 감정과 부정적인 감정을 동시에 양가감정의 형태로 지니고 있다는 점이다. 그러나 여기서 주목해야할 부분은 긍휼한 마음의 해석을 가능하게 하는 단어들의 위치가 사상 검증의 필요성과 관련된 단어들의 위치보다 '탈북자' 단어에 가깝게 위치해있다는 점이다. 따라서 북한이탈주민에 대한 댓글 작성자들의 긍휼한 마음과 의심이 혼합된 양가감정은 긍휼한 마음에 좀 더 기울어진 양가감정이라고 볼 수 있다.

로부터 멀리 위치하고 있다는 점에서 중앙일보 댓글에서는 북한이탈주민에 대한 댓글 작성자들의 의심이 조선일보 댓글에 비해 비교적 낮게 나타나고 있는 것으로 볼 수 있다.

이와 같은 해석은 앞서 조선일보 댓글에 대한 연관어 분석에서 댓글 작성자들이 북한이탈주민을 떠올릴 때, 북한이탈주민이 지녔을 사상을 함께 생각한다는 점에서 서로 유사해 보인다. 하지만 차이점이 있다면 조선일보 댓글에서는 북한이탈주민과 함께 '사상' 을 떠올릴 때 '검증'을 함께 떠올린다는 점에서 북한이탈주민이 갖고 있는 사상을 의심하는 측면이 강조된 반면, 중앙일보 댓글에서는 북한이탈주민이 지닌 사상에 대해 의심하기보다 그들이 북한의 사상과 멀어졌을 것이라는 생각이 나타난다는 점에 주목해야 한다. 이는 또한 중앙일보 댓글의 연관어 분석 결과에서 '탈북자' 단어가 좌측의 '위장', '간첩'이라는 단어와 상당한 거리를 보여주고 있다는 점으로도 뒷받침 될 수 있다.

한편, 중앙일보 댓글을 이용한 연관어 분석에서는 [그림 2]의 '탈북자' 단어의 오른편 하단에 '국정원', '조작', '강제', '사실'에 대한 언급이 존재하지만 다소 멀리 떨어져 있다는 점에 주목해 볼 수 있다. 이는 이후 한겨레신문 댓글의 연관어 분석 결과와 경향신문 댓글의 연관어 분석 결과에서 주되게 언급될 내용이지만, 중앙일보 댓글에서는 북한이탈주민이 종종 간첩으로 조작되어 온 상황에 대한 관심이 나타나고 있기는 하지만 한겨레신문의 댓글, 경향신문의 댓글에 비해 그 관심이 비교적 낮게 나타나고 있는 것으로 보인다. 결국 중앙일보 댓글을 이용한 연관어 분석 결과는 조선일보 댓글을 이용한 연관어 분석과 한겨레신문 및 경향신문 댓글을 이용한 연관어 분석을 양 끝 지점으로 상상해보았을 때 가운데 지점에 있을 것으로 생각해 볼 수 있다.

다음으로 한겨레신문 댓글 연관어 분석 결과에서는 [그림 3] 내 '탈북자'라는 단어가 왼편의 '극우'라는 단어와 가깝게 위치하고 있는 것에 주목해 볼 수 있다. 이는 댓글 작성자들이 북한이탈주민을 떠올릴 때 북한이탈주민은 한국 사회 내에서 정치적으로 극우적인 성향을 띠고 있다는 생각을 지니고 있음을 의미한다. 실제로 북한이탈주민은 북한에서 이탈

하였기 때문에 정치적으로 공산주의에 대한 부정적인 정서를 강하게 나타내는 경향이 있다. 이로 인해 한국 내 적응하는 과정에서 공산주의에 강하게 대응하는 한국의 보수 진영 집단과의 가치관의 공유가 이루어 질 수 있는 교집합이 존재한다. 그리고 이는 정치적으로 진보 진영에 서있는 사람들에게 있어서 극우로 보여 질 수 있는 부분이다.

또한 [그림 3]의 '지원', '국내', '이해', '힘들' 단어 좌측의 '어렵', '차별' 단어들의 등장은 댓글 작성자들이 다양한 곳에서 차별받는 북한이탈주민의 상황에 대해 관심을 갖고 있음을 보여준다. 이와 같은 [그림 3]의 '지원', '국내', '이해', '힘들' 단어들의 등장 곧, 북한이탈주민에 대한 댓글 작성자들의 관심은 앞서 보수 성향의 신문사 댓글과 진보 성향의 신문사 댓글 간의 공통점으로 언급되었다. 하지만 보수 성향의 신문사 댓글 연관어 분석 결과와 진보 성향의 신문사 댓글 연관어 분석 결과 간의 공통점 내 차이점은 '차별'이라는 단어의 존재유무에 있다. 그리고 이는 결국 국내정치적 입장에 따라 북한이탈주민에 대한 입장이 변화될 수 있는 가능성을 보여준다. 정치적으로 진보 진영에 가까울수록 차별에 대해 더욱 민감한 태도를 보여줄 가능성이 높게 나타나기 때문이다.

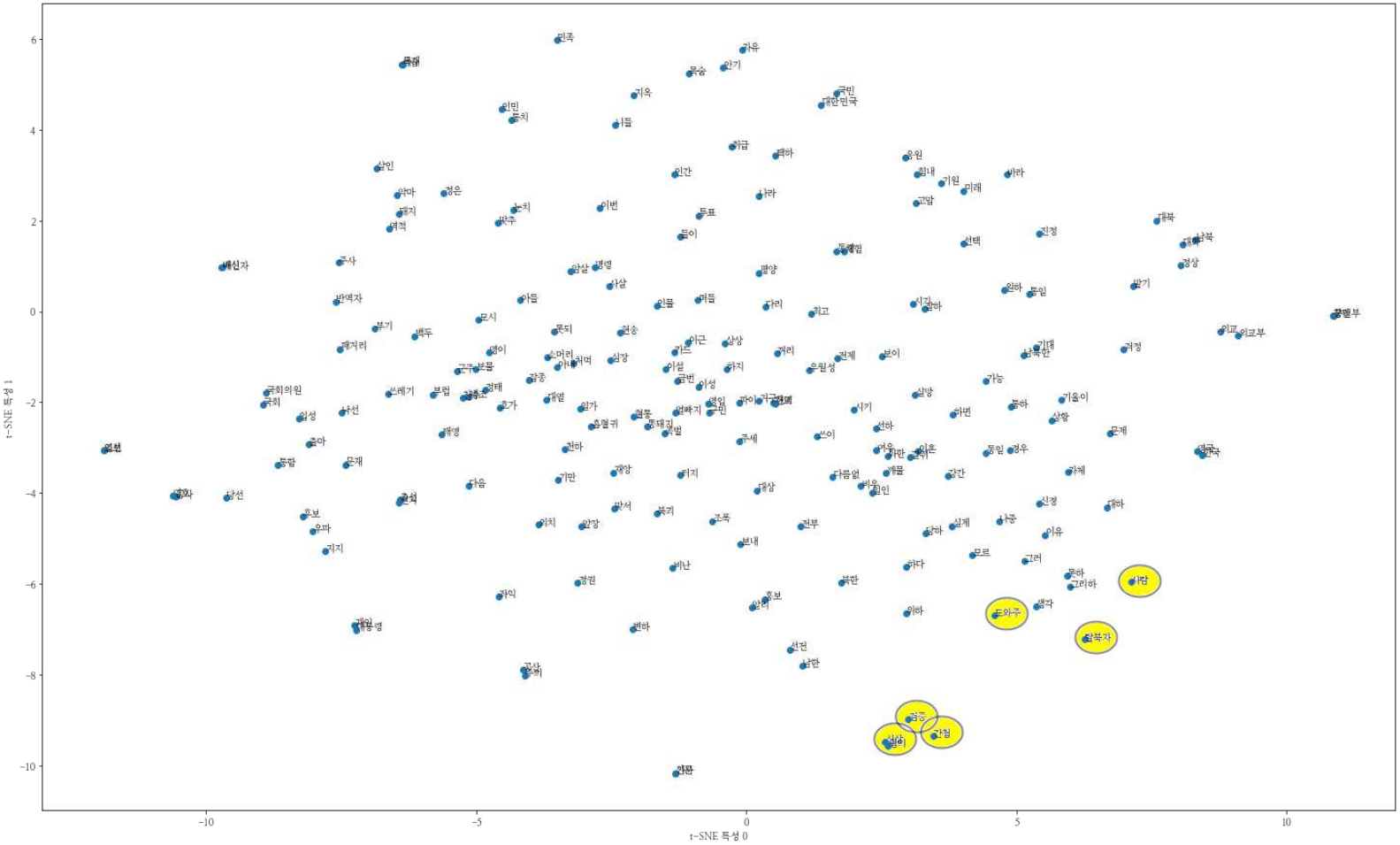
한편, 한겨레신문 댓글 연관어 분석 결과에서는 '간첩'이라는 단어가 등장하지 않지만 '탈북자'를 기준으로 좌측 하단에 '빨갱이'라는 단어가 '조작'과 함께 등장하고 있는 것을 볼 수 있다. 이는 댓글 작성자들이 한국 사회 내에서 북한이탈주민이 남파된 빨갱이로 의심받는 것과 관련해 조작의 가능성을 어느 정도 염두에 두고 있다는 점을 시사한다.

마지막으로 경향신문 댓글 연관어 분석 결과에서도 '정착', '난민', '이용', '사회', '노력'이라는 단어들을 통해 앞서 보수 성향의 신문사 댓글과 진보 성향의 신문사 댓글 간의 공통점 곧, 북한이탈주민에 대한 댓글 작성자들의 긍정적인 관심이 존재함을 확인할 수 있다. 하지만 경향신문 댓글 연관어 분석 결과는 댓글 작성자들이 난민이라는 특수한 사회적 지위로 인해 다양한 사회적 차별을 겪었을 북한이탈주민에 대해 고려하고 있고, 북한이탈주민은 제도적인 보완이 필요한 위

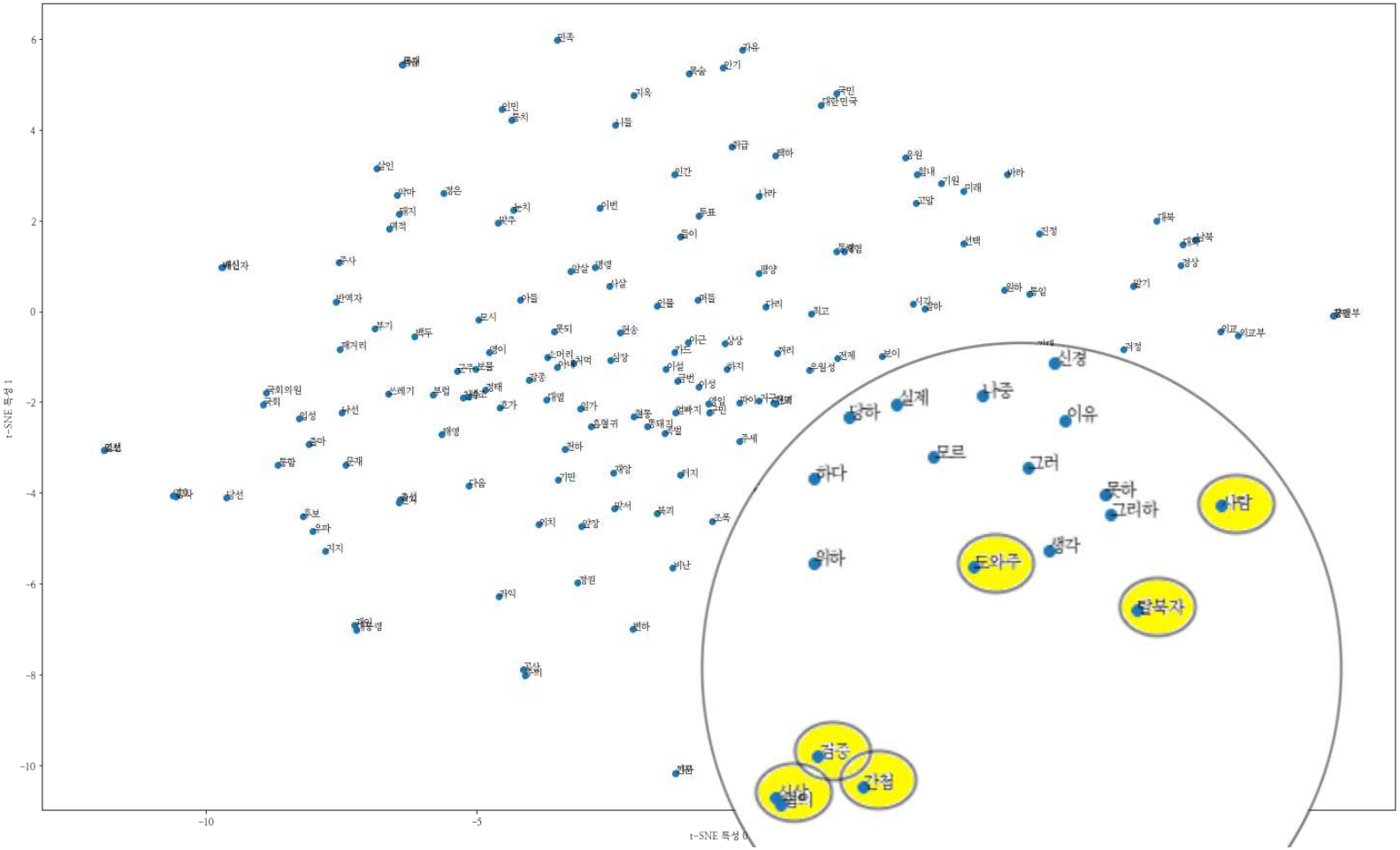
치에 있다는 생각을 했을 것으로 추정된다는 점에서, 조선일보 댓글의 연관어 분석 결과와는 다소 거리가 존재하고 한겨레신문 댓글의 연관어 분석 결과와 매우 가깝다고 볼 수 있겠다.

다음으로 [그림 4]의 '탈북자' 단어 하단에 '조사'가 위치해있고 위편에는 '진실', 우측에는 '정권', '간첩', '사람'이라는 단어가 위치해있다는 것에도 주목할 필요가 있다. 이는 북한이탈주민이 대한민국 정권의 변화에 따라서 간첩일지도 모른다는 의심과 이에 따른 부당한 조사를 지속적으로 받아왔다는 내용을 시사하며, 댓글 작성자들에게는 북한이탈주민 역시도 사람이기 때문에 이와 관련된 진실 규명이 필요하다는 생각이 존재함을 보여준다. 이와 같은 해석은 한겨레신문 댓글을 이용한 연관어 분석 결과를 해석하는 과정에서 '빨갱이'라는 단어가 '조작'이라는 단어와 함께 등장하였던 것에 초점을 맞추어 해석했던 결과와 유사하다.

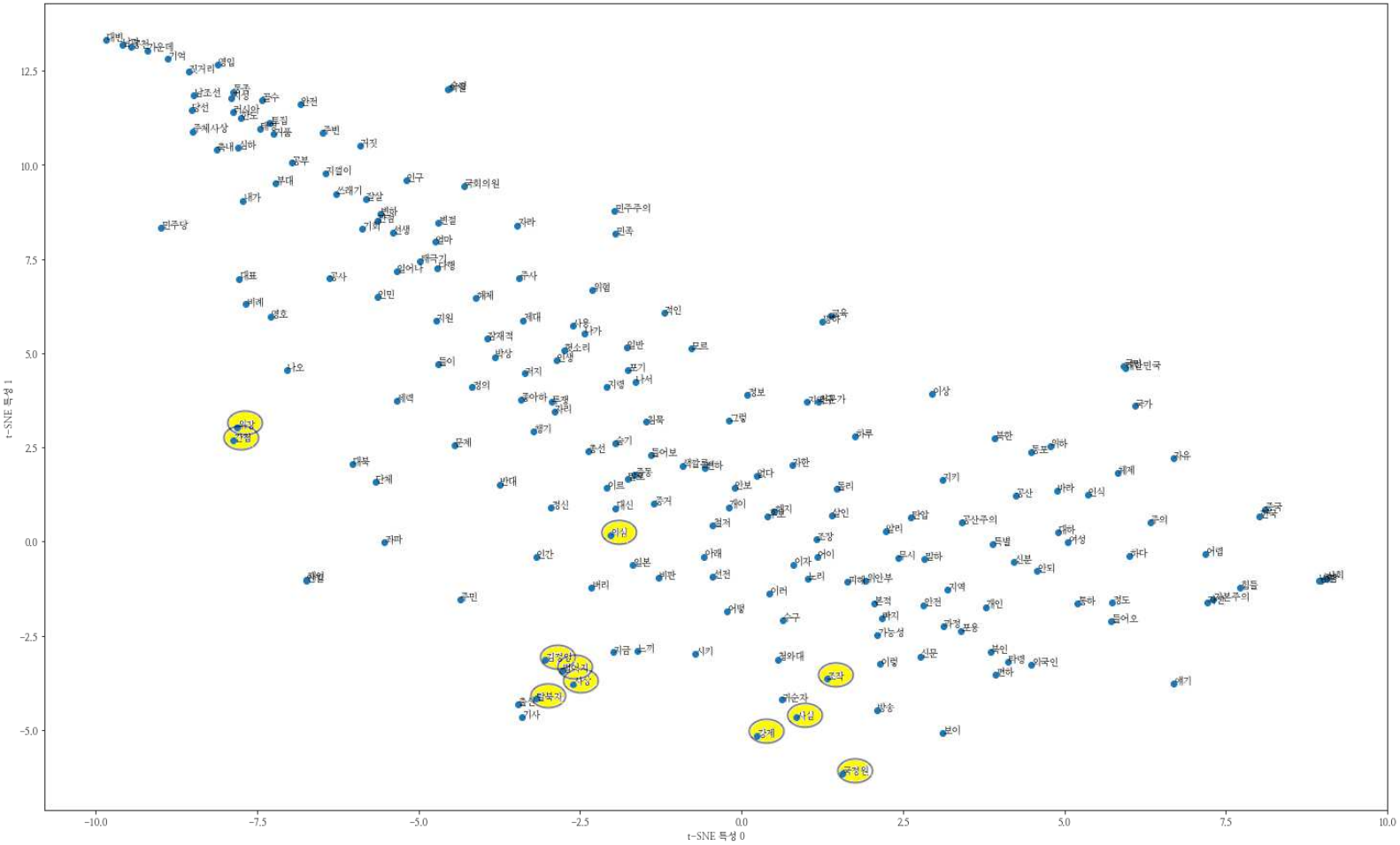
[그림 1-1] 조선일보 댓글 연관어 분석 결과



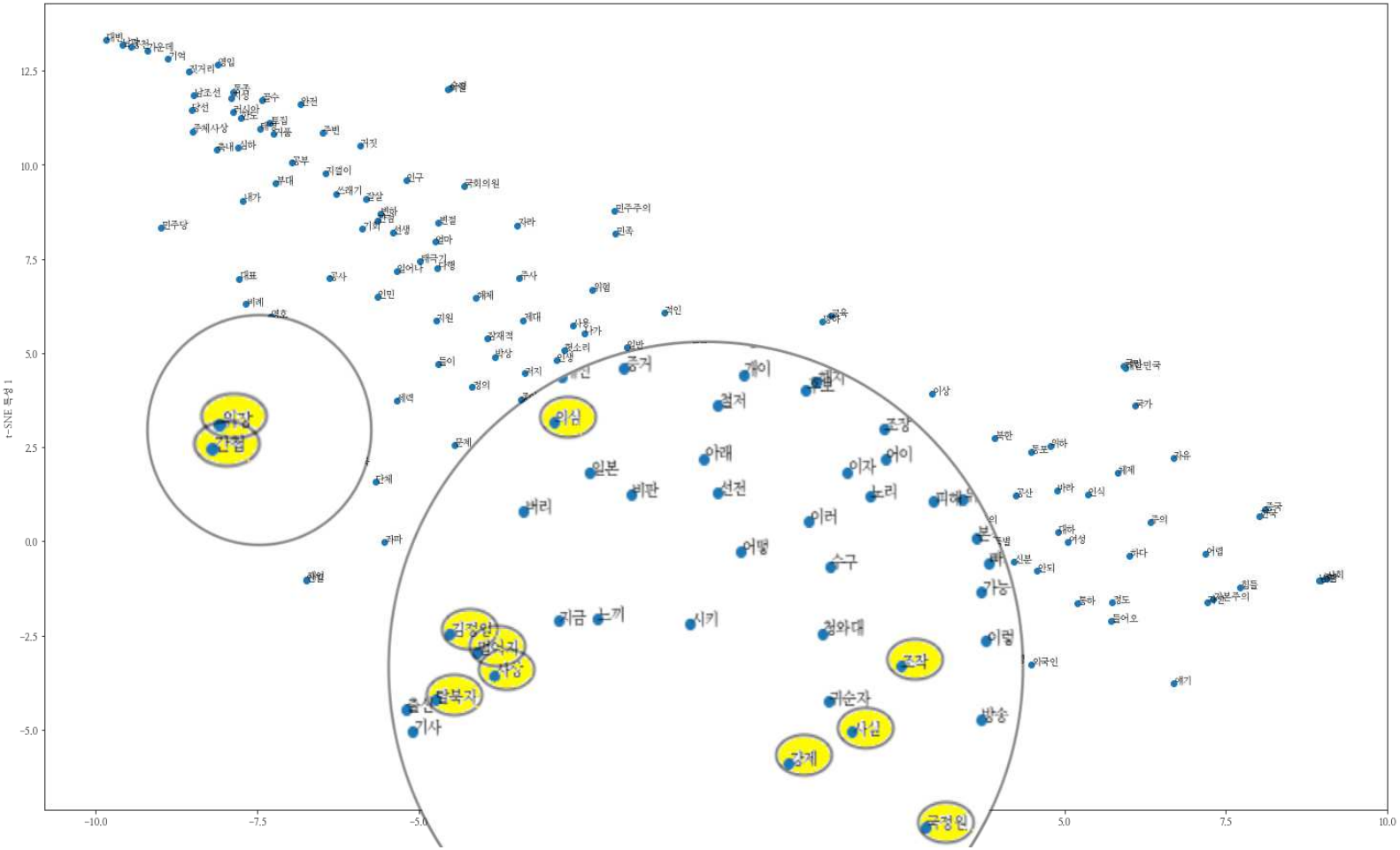
[그림 1-2] 조선일보 댓글 연관어 분석 결과 부분 확대



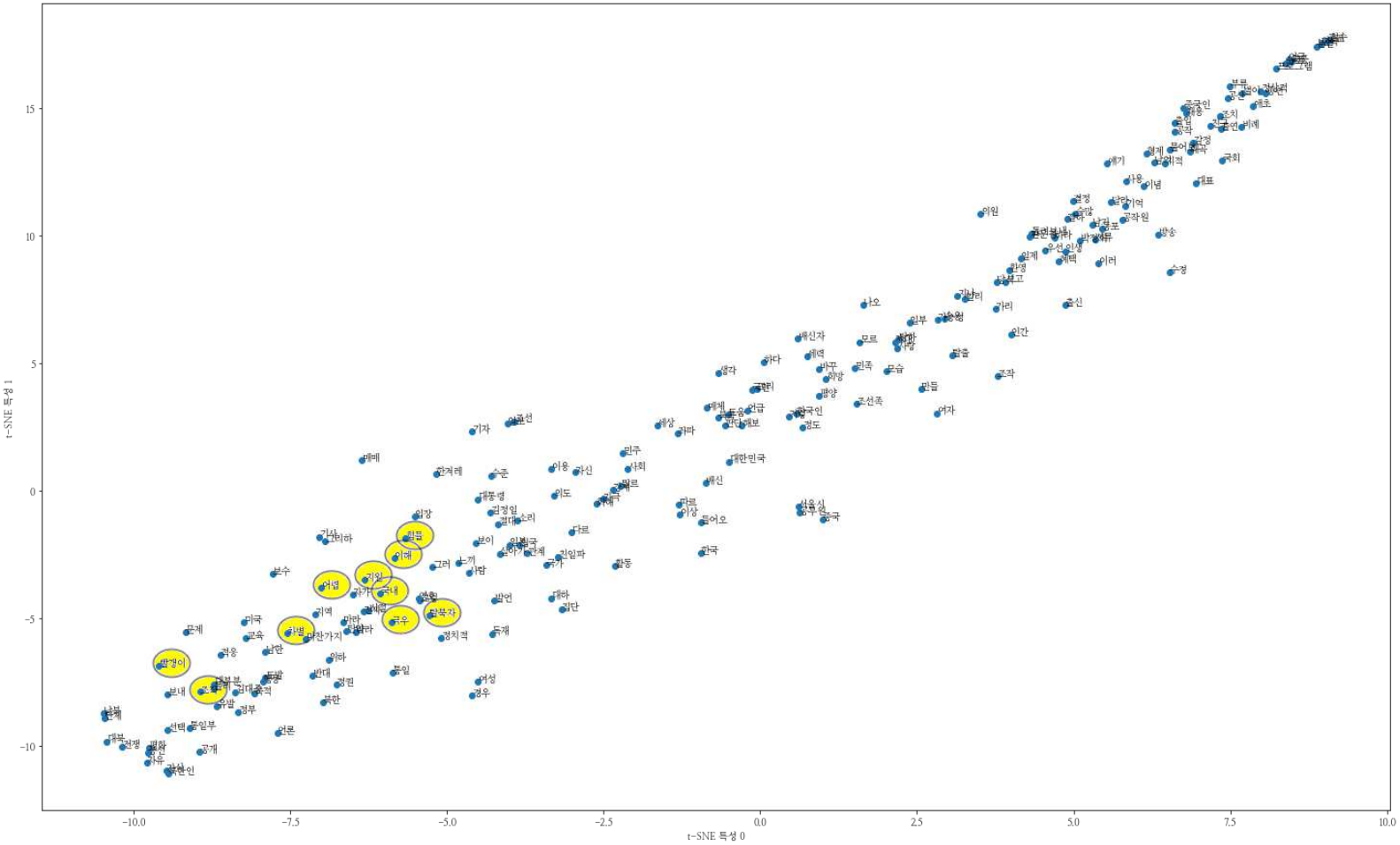
[그림 2-1] 중앙일보 댓글 연관어 분석 결과



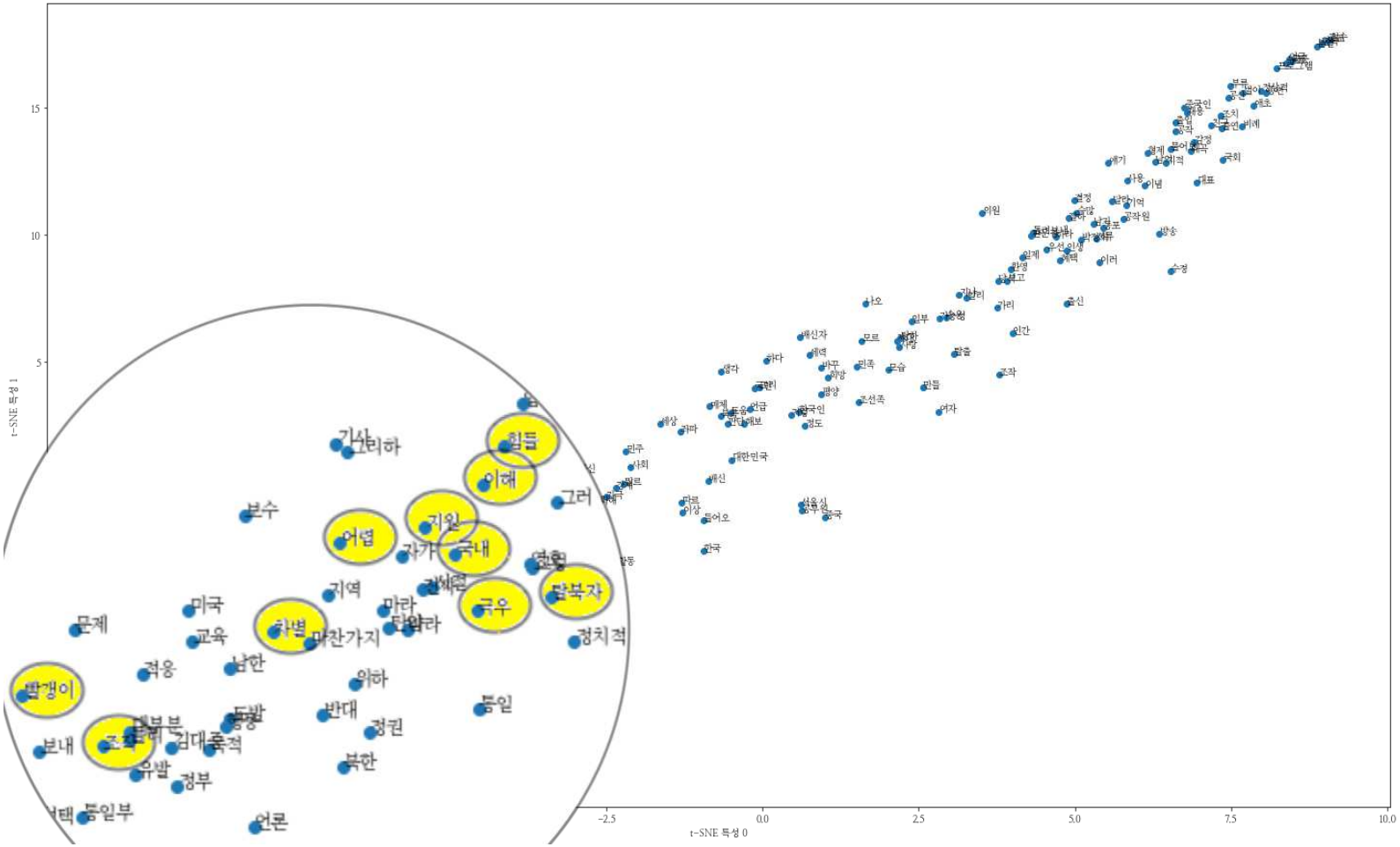
[그림 2-2] 중앙일보 댓글 연관어 분석 결과 부분 확대



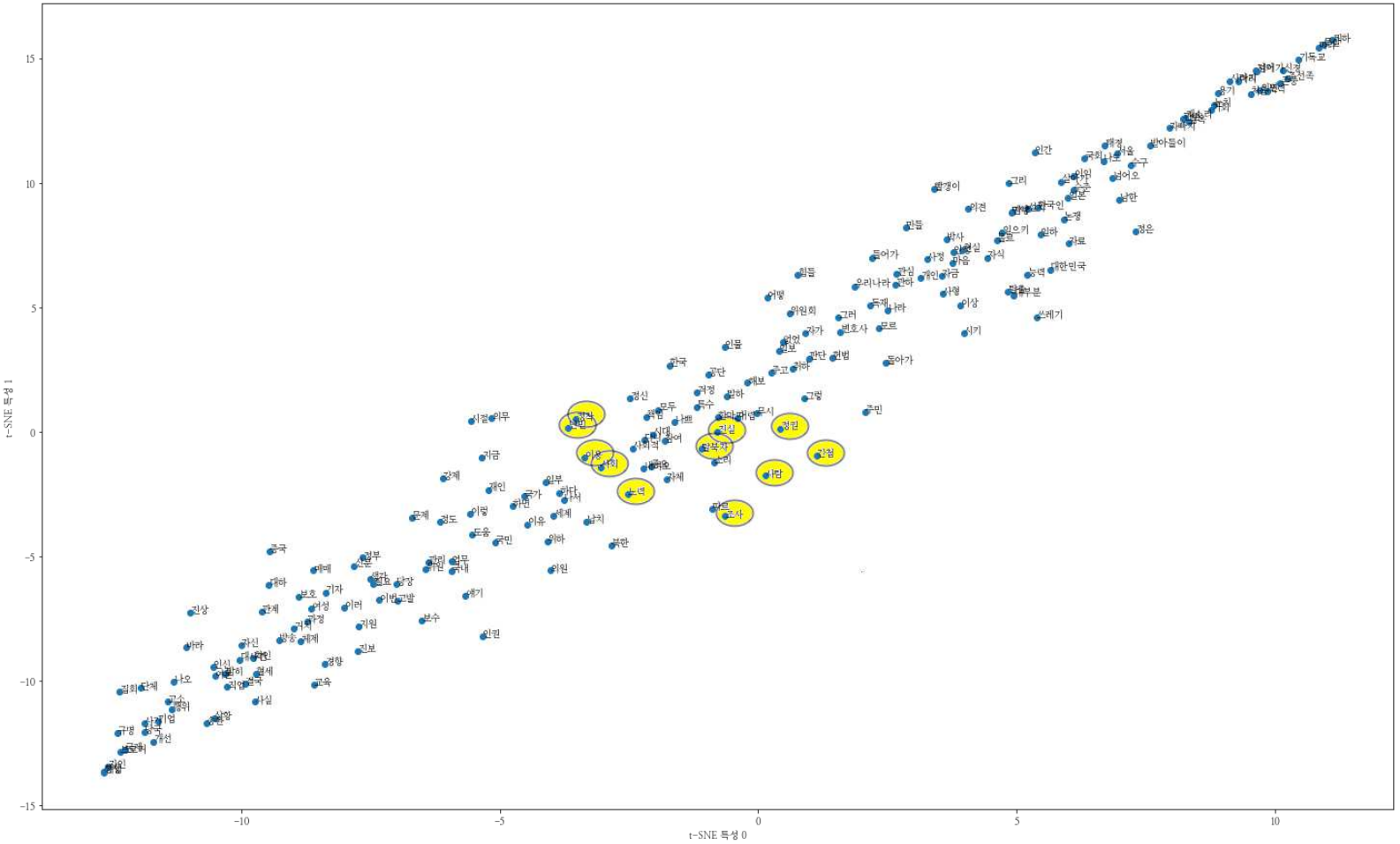
[그림 3-1] 한겨레신문 댓글 연관어 분석 결과



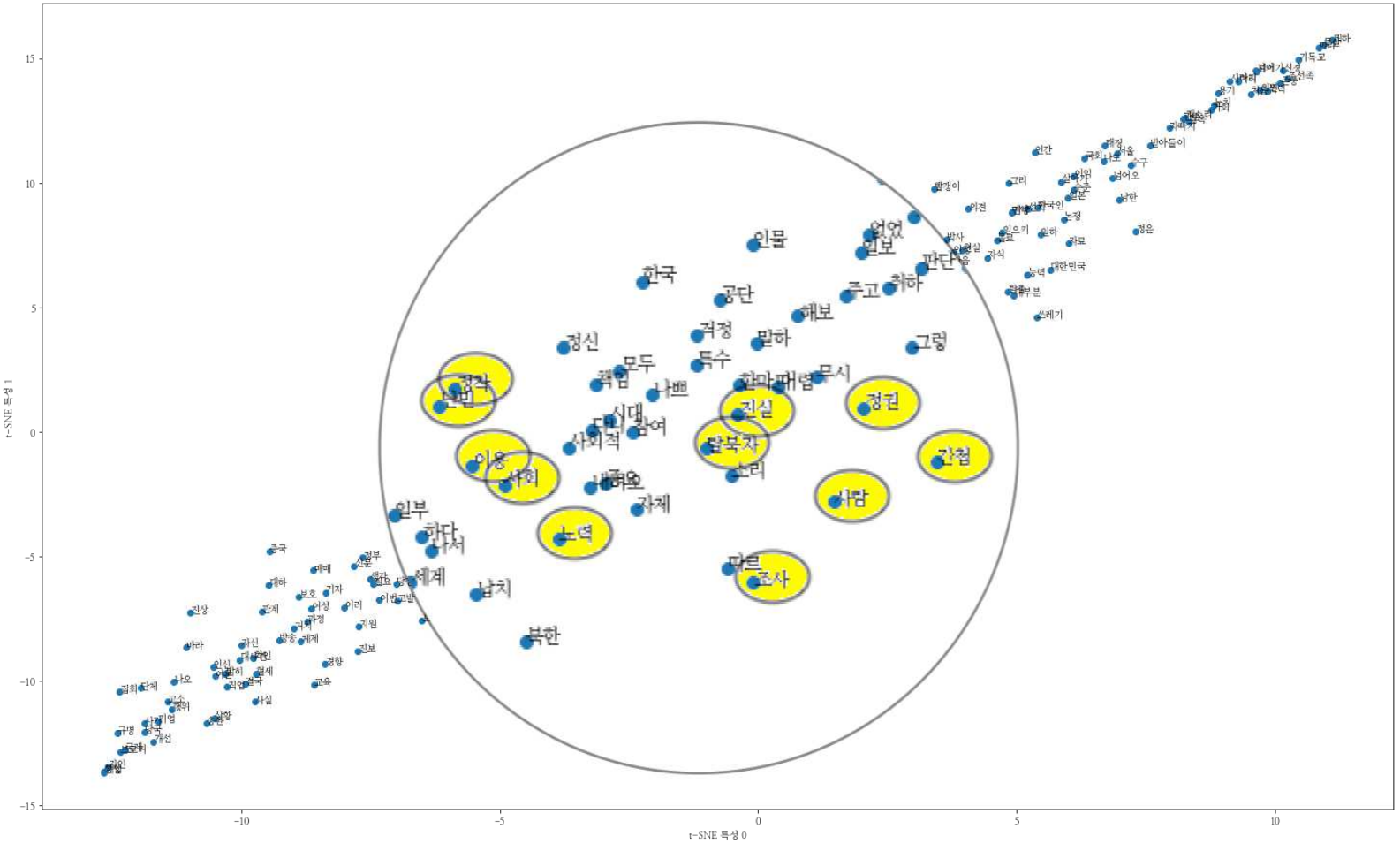
[그림 3-2] 한겨레신문 댓글 연관어 분석 결과 부분 확대



[그림 4-1] 경향신문 댓글 연관어 분석 결과



[그림 4-2] 경향신문 댓글 연관어 분석 결과 부분 확대



제 3 절 댓글 감성분석 결과

1. 감성사전을 이용한 댓글 감성분석 결과

먼저, [표 8]의 군산대학교 감성사전을 이용한 댓글 감성분석 결과를 살펴보면 본 연구의 연구 설계 부분을 통해 예상했던 바와는 상이한 결과가 나타난 것을 확인할 수 있다. 본 연구는 앞서 감성사전을 이용한 감성분석을 시도할 때, 진보 성향을 띠는 신문사 댓글보다는 보수 성향을 띠는 신문사 댓글에서 긍정적인 댓글의 비율이 좀 더 높게 나타날 것으로 예상하였다. 하지만 실제 결과에서는 네 개의 신문사들의 긍정적인 댓글의 비율이 대략적으로 대동소이하고, 자세히 살펴보아도 오히려 진보 성향을 띠는 한겨레신문 댓글에서 긍정적인 댓글의 비율이 가장 높게 나타난 것을 확인할 수 있다.

더 나아가 부정적인 댓글의 비율에서도 역시 네 개의 신문사들은 서로 큰 차이를 보여주지 않았고, 자세히 보더라도 진보 성향의 신문사 댓글보다는 보수 성향의 신문사 댓글에서 부정적인 댓글의 비율이 좀 더 높게 나온 것을 확인 할 수 있다. 이는 본 연구의 연구 설계 부분을 통해 제시했던 것과는 완전히 다른 결과가 나타났음을 보여준다.

본 연구는 이와 같은 결과가 도출된 데에 다음과 같은 이유가 존재할 것으로 생각한다. 먼저는 군산대학교 감성사전의 단어들과 본 연구에서 분석 대상으로 삼고 있는 댓글 텍스트 내 단어들 간의 대응(matching)이 잘 이루어지지 않았다는 점이다. 군산대학교 감성사전은 대략적으로 보아도 약 1만 4천 8백 50개의 단어들로 구성되어 있는데, 그 가운데 부정적인 단어들이 약 1만개, 긍정적인 단어들이 약 4,700개, 그리고 중립적인 단어들이 약 150개로 구성되어 있다. 하지만 본 연구의 결과에서 중립적인 것으로 분류된 댓글은 네 개의 신문사 모두에서 평균적으로 60% 초반 대의 비율을 보여주고 있다는 점에서, 군산대학교 감성사전에 등재된 긍정적 단어들과 부정적 단어들이 북한이탈주민에 대한 댓글 작성자들의 태도를 판단하는 데에는 활용도가 낮은 자료라고 볼 수 있다. 이에

따라 감성사전을 이용한 감성분석에서 감성사전을 연구하려는 자료에 맞게 재구성(customizing)하지 않을 경우 해당 감성사전에 상당한 한계가 발생함을 알 수 있다. 그리고 이는 보편성을 띠는 감성사전을 제작한 사람들의 본 의도와도 연결된다.

혹은 감성사전을 이용한 댓글 분석 결과에서 본 연구가 연구 설계를 통해 제시했던 예상과는 전혀 다른 방향성을 보여주는 결과가 도출되었기 때문에, 본 연구의 추측이 현실을 제대로 반영하지 못한 것이었다고도 볼 수 있다.

2. 지정된 단어를 이용한 댓글 감성분석 결과

본 연구에서는 북한에 대한 적대적인 감정과 북한이탈주민에 대한 긍정적 혹은 부정적인 감정을 나타낼 수 있는 단어를 지정하는 연구 설계 과정에서 '북괴', '인권', '동포', '보호', '간첩', '위장', '빨갱이'를 선택하였다. 하지만 [표 9]를 통해 알 수 있듯이, 제시된 단어들은 네 개의 모든 신문사 댓글에서 골고루 사용되기보다 조선일보와 중앙일보의 댓글에서만 비교적 활발하게 사용되고 있었다. 이에 따라 제시된 단어들이 빈번하게 사용되었을 것이라 가정하고 사용 비율을 따져보고자 하였던 본 연구 설계 내용에 대한 확인이 어려워지게 되었다. 이와 같은 결과는 북한이탈주민에 대한 댓글 작성자들의 생각이 다양한 단어로 드러나고 있기 때문에 지정된 단어를 이용한 댓글 감성분석이 북한이탈주민에 대한 댓글 작성자들의 태도를 측정하는 데 있어서 실증적인 이해를 높일 수 있는 방법이 아니라고도 생각해볼 수 있는 부분이다.

한편, 위 지정된 단어를 이용한 댓글 감성분석 결과를 바탕으로 카이제곱검정을 실시할 경우 신문사 댓글 간 특정 단어 사용에 있어 유의미한 차이가 존재하는지 확인할 수 있다. 본 연구는 한겨레신문 댓글과 경향신문 댓글이 각 1,000개를 넘지 않아 조선일보 댓글과 중앙일보 댓글에 대해서만 카이제곱검정을 실시하였다. 해당 결과는 아래 [표 10], [표 11], [표 12], [표 13], [표 14], [표 15], [표 16]을 통해 확인할 수 있다.

[표 8] 군산대학교 감성사전을 이용한 댓글 감성분석 결과

	조선일보	중앙일보	한겨레신문	경향신문
본 연구에서 사용되는 총 댓글 수	9,937	1,252	735	542
긍정적인 댓글 수	1,436	176	111	66
중립적인 댓글 수	6,154	765	454	352
부정적인 댓글 수	2,347	311	170	124
긍정적인 댓글의 비율	**14.5%**	**14.1%**	**15.1%**	**12.2%**
중립적인 댓글의 비율	**61.9%**	**61.1%**	**61.8%**	**64.9%**
부정적인 댓글의 비율	**23.6%**	**24.8%**	**23.1%**	**22.9%**

註: 단위(개), 군산대학교 감성사전을 이용하여 도출되는 감성지수 결과 값이 0보다 클 경우 긍정적인 것으로, 0인 경우 중립적인 것으로, 0보다 작을 경우 부정적인 것으로 판단하였다. 표 하단의 비율의 경우 소수점 두 번째 자리에서 반올림하였다.

[표 9] 지정된 단어를 이용한 댓글 감성분석 결과

	조선일보	중앙일보	한겨레신문	경향신문
본 연구에서 사용되는 총 댓글 수	9,937	1,252	735	542
(1) ‘북괴’를 포함한 댓글 수	329	35	12	5
(2) ‘북괴’를 포함한 댓글 가운데 ‘인권’ 혹은 ‘동포’ 혹은 ‘보호’를 포함한 댓글 수	72	13	4	0
(3) ‘북괴’를 포함한 댓글 가운데 ‘간첩’ 혹은 ‘위장’ 혹은 ‘빨갱이’를 포함한 댓글 수	43	9	3	1
(2) / (1)의 비율	**21.88%**	**37.14%**	**33.33%**	**0%**
(3) / (1)의 비율	**13.07%**	**25.71%**	**25%**	**20%**

註: 단위(개), 표 하단의 (2) / (1)의 비율과 (3) / (1) 비율의 경우 소수점 세 번째 자리에서 반올림하였다.

[표 10] ‘북괴’ 단어를 이용한 댓글 카이제곱검정

	‘북괴’ 포함 댓글	‘북괴’ 불포함 댓글
조선일보	**329**	**9,608**
기대값	323.269997	9613.73
카이제곱	0.10156504	0.00341521
중앙일보	**35**	**1,217**
기대값	40.7300027	1211.27
카이제곱	0.80611168	0.0271062
	카이제곱 총합	**0.93819813**

註: df = 1, α = 0.1 일 때, X^2 = 2.71
X^2 ≥ 2.71 경우, *로 표기.
df = 1, α = 0.05 일 때, X^2 = 3.84
X^2 ≥ 3.84 경우, **로 표기.
df = 1, α = 0.01 일 때, X^2 = 6.63
X^2 ≥ 6.63 경우, ***로 표기.

위 [표 10]의 카이제곱검정 결과, 조선일보 댓글과 중앙일보 댓글 내 ‘북괴’ 단어 이용과 관련해 두 집단 간의 유의미한 차이를 발견하기 어렵다.

[표 11] ‘인권’ 단어를 이용한 댓글 카이제곱검정

	‘인권’ 포함 댓글	‘인권’ 불포함 댓글
조선일보	**1,235**	**8,702**
기대값	1226.47216	8710.52784
카이제곱	0.05929531	0.00834898
중앙일보	**146**	**1,106**
기대값	154.52784	1097.47216
카이제곱	0.47062104	0.06626505
	카이제곱 총합	**0.60453039**

註: df = 1, α = 0.1 일 때, X^2 = 2.71
$X^2 \geq 2.71$ 경우, *로 표기.
df = 1, α = 0.05 일 때, X^2 = 3.84
$X^2 \geq 3.84$ 경우, **로 표기.
df = 1, α = 0.01 일 때, X^2 = 6.63
$X^2 \geq 6.63$ 경우, ***로 표기.

위 [표 11]의 카이제곱검정 결과, 조선일보 댓글과 중앙일보 댓글 내 ‘인권’ 단어 이용과 관련해 두 집단 간의 유의미한 차이를 발견하기 어렵다.

[표 12] ‘동포’ 단어를 이용한 댓글 카이제곱검정

	‘동포’ 포함 댓글	‘동포’ 불포함 댓글
조선일보	**385**	**9,552**
기대값	390.765931	9546.23407
카이제곱	0.08507896	0.00348263
중앙일보	**55**	**1,197**
기대값	49.2340692	1202.76593
카이제곱	0.67526326	0.02764125
	카이제곱 총합	**0.7914661**

註: df = 1, α = 0.1 일 때, X^2 = 2.71
$X^2 \geq 2.71$ 경우, *로 표기.
df = 1, α = 0.05 일 때, X^2 = 3.84
$X^2 \geq 3.84$ 경우, **로 표기.
df = 1, α = 0.01 일 때, X^2 = 6.63
$X^2 \geq 6.63$ 경우, ***로 표기.

위 [표 12]의 카이제곱검정 결과, 조선일보 댓글과 중앙일보 댓글 내 ‘동포’ 단어 이용과 관련해 두 집단 간의 유의미한 차이를 발견하기 어렵다.

[표 13] ‘보호’ 단어를 이용한 댓글 카이제곱검정

	‘보호’ 포함 댓글	‘보호’ 불포함 댓글
조선일보	**346**	**9,591**
기대값	362.34659	9574.65341
카이제곱	0.73744593	0.02790817
중앙일보	**62**	**1,190**
기대값	45.6534096	1206.34659
카이제곱	5.8530353	0.22150435
	카이제곱 총합	**6.83989374 *****

註: df = 1, α = 0.1 일 때, X^2 = 2.71
$X^2 \geq 2.71$ 경우, *로 표기.
df = 1, α = 0.05 일 때, X^2 = 3.84
$X^2 \geq 3.84$ 경우, **로 표기.
df = 1, α = 0.01 일 때, X^2 = 6.63
$X^2 \geq 6.63$ 경우, ***로 표기.

위 [표 13]의 카이제곱검정 결과, 조선일보 댓글과 중앙일보 댓글 내 ‘보호’ 단어 이용과 관련해 두 집단 간의 유의미한 차이가 존재한다.

[표 14] ‘간첩’ 단어를 이용한 댓글 카이제곱검정

	‘간첩’ 포함 댓글	‘간첩’ 불포함 댓글
조선일보	**648**	**9,289**
기대값	667.8545	9269.1455
카이제곱	0.59025007	0.04252832
중앙일보	**104**	**1,148**
기대값	84.1455	1167.8545
카이제곱	4.68475638	0.33754305
	카이제곱 총합	**5.65507783 ****

註: df = 1, α = 0.1 일 때, X^2 = 2.71
$X^2 \geq 2.71$ 경우, *로 표기.
df = 1, α = 0.05 일 때, X^2 = 3.84
$X^2 \geq 3.84$ 경우, **로 표기.
df = 1, α = 0.01 일 때, X^2 = 6.63
$X^2 \geq 6.63$ 경우, ***로 표기.

위 [표 14]의 카이제곱검정 결과, 조선일보 댓글과 중앙일보 댓글 내 ‘간첩’ 단어 이용과 관련해 두 집단 간의 유의미한 차이가 존재한다.

[표 15] ‘위장’ 단어를 이용한 댓글 카이제곱검정

	‘위장’ 포함 댓글	‘위장’ 불포함 댓글
조선일보	**254**	**9,683**
기대값	256.662168	9680.33783
카이제곱	0.02761272	0.00073212
중앙일보	**35**	**1,217**
기대값	32.3378318	1219.66217
카이제곱	0.21915939	0.00581074
	카이제곱 총합	**0.25331496**

註: df = 1, α = 0.1 일 때, X^2 = 2.71
$X^2 \geq 2.71$ 경우, *로 표기.
df = 1, α = 0.05 일 때, X^2 = 3.84
$X^2 \geq 3.84$ 경우, **로 표기.
df = 1, α = 0.01 일 때, X^2 = 6.63
$X^2 \geq 6.63$ 경우, ***로 표기.

위 [표 15]의 카이제곱검정 결과, 조선일보 댓글과 중앙일보 댓글 내 ‘위장’ 단어 이용과 관련해 두 집단 간의 유의미한 차이를 발견하기 어렵다.

[표 16] ‘빨갱이’ 단어를 이용한 댓글 카이제곱검정

	‘빨갱이’ 포함 댓글	‘빨갱이’ 불포함 댓글
조선일보	**0**	**9,937**
기대값	16.8739834	9920.12602
카이제곱	16.8739834	0.02870239
중앙일보	**19**	**1233**
기대값	2.12601662	1249.87398
카이제곱	133.927135	0.22780802
	카이제곱 총합	**151.057629 *****

註: df = 1, α = 0.1 일 때, X^2 = 2.71
X^2 ≥ 2.71 경우, *로 표기.
df = 1, α = 0.05 일 때, X^2 = 3.84
X^2 ≥ 3.84 경우, **로 표기.
df = 1, α = 0.01 일 때, X^2 = 6.63
X^2 ≥ 6.63 경우, ***로 표기.

위 [표 16]의 카이제곱검정 결과, 조선일보 댓글과 중앙일보 댓글 내 ‘빨갱이’ 단어 이용과 관련해 두 집단 간의 유의미한 차이가 존재한다.

위 [표 10], [표 11], [표 12], [표 13], [표 14], [표 15], [표 16]에 나타난 카이제곱검정 결과들을 통해 ‘북괴’, ‘인권’, ‘동포’, ‘도움’, ‘간첩’, ‘위장’, ‘빨갱이’ 단어 가운데 조선일보 댓글과 중앙일보 댓글 간의 유의미한 차이를 보여준 단어는 ‘보호’, ‘간첩’, ‘빨갱이’인 것임을 알 수 있다. 이와 같은 결과는 앞서 동일한 보수 성향의 신문사 댓글일지라도 조선일보 댓글 연관어 분석 결과와 중앙일보 댓글 연관어 분석 결과 간에 차이점이 드러난다고 언급했던 부분을 상기시킨다.

제 5 장 결 론

본 연구는 보수 성향의 신문사들과 진보 선향의 신문사들이 서로 다른 관점의 기사를 제시할 때, 서로 다른 독자의 반응을 이끌어낼 수 있는지 알아보기 위하여 각 신문사 홈페이지에서 북한이탈주민과 관련된 기사들과 기사에 달린 댓글들을 수집하여 토픽 분석, 연관어 분석을 중점적으로 실시하되 추가적으로 감성분석과 카이제곱검정을 실시하였다.

본 연구는 먼저 제시된 연구 질문 1에 대답하기 위해 기사를 활용한 토픽 분석을 실시하였다. 그 결과, 보수 성향의 신문사들의 기사와 진보 성향의 신문사들의 기사 간의 공통점과 차이점이 몇 가지 발견되었다. 보수 성향의 신문사들과 진보 성향의 신문사들은 모두 북한 인권 문제, 북한이탈주민들의 한국 내 정착 지원 제도 문제, 북한이탈주민들의 북한에서의 생활 혹은 탈북 과정 이야기, 북한이탈주민을 둘러싼 주변국들의 반응, 한국 사회 내 북한이탈주민을 가장한 간첩 문제, 일부 보수적인 성향을 지닌 북한이탈주민 단체의 대북전단 문제에 대해서 기사를 게재해왔다.

하지만 보수 성향을 지닌 신문사들의 경우 북한이탈주민에 대한 주변국들의 반응과 북한이탈주민을 통해 들을 수 있는 북한 내부 소식을 강조하는 방향으로 기사를 게재해온 반면, 진보 성향을 지닌 신문사들의 경우 국정원의 간첩 수사에 대한 진상 규명, 한국 사회 내 차별로 인해 정착에 어려움을 겪는 북한이탈주민들의 모습, 그리고 보수 성향을 띠는 북한이탈주민들과 관련해 기사를 게재해왔다는 점에서 어느 정도 차이를 보여주었다. 더불어 조선일보는 북한이탈주민의 한국 내 제도권 정치 참여에 대해 기사를 게재해온 반면, 경향신문은 박근혜 전 대통령과 뉴라이트전국연합을 언급하며 일부 보수적인 성향을 지닌 북한이탈주민에 대해 기사를 게재해왔다.

다음으로 본 연구는 연구 질문 2에 대답하기 위해 댓글을 활용한 연관어 분석을 실시하였다. 그 결과, 보수 성향을 지닌 신문사들의 댓글과 진보 성향을 지닌 신문사들의 댓글 모두에서 북한이탈주민은 사회적으로

도움이 필요하고 그 도움은 한국 사회 내에서 제공되어야 한다는 생각이 공통적으로 나타남을 알 수 있었다.

또한, 각 신문사 댓글 간의 차이점 역시 두드러지게 나타났다. 먼저, 조선일보 댓글 작성자들은 다른 신문사의 댓글 작성자들과 달리 북한이탈주민을 가장한 간첩의 가능성에 대해서 가장 많이 염두에 두고 있었다.

하지만 중앙일보 댓글의 경우 중앙일보가 조선일보와 같은 보수 성향의 신문사로 분류됨에도 불구하고 다소 차이를 보여주었는데, 중앙일보 댓글 작성자들은 북한이탈주민이 지니고 있는 사상에 대해 의심을 갖고 있기는 하지만 북한이탈주민들이 북한에서 강조되는 사상과는 멀어졌을 것이라 생각하고 있는 것으로 나타났다. 또한, 중앙일보 댓글 작성자들은 한국 사회 내에서 위장 간첩으로 의심 받아온 북한이탈주민에 대해 언급하고 있었다는 점에서 한겨레신문 댓글 작성자들과 다소 유사한 모습을 보여주기도 하였다.

한편, 한겨레신문 댓글 작성자들은 북한이탈주민의 극우적인 성향과 북한이탈주민들이 한국 사회 내에서 정착하는 과정 가운데 겪었던 차별에 대해 관심을 가지고 있는 것으로 나타났다. 또한, 한겨레신문 댓글 작성자들은 보수 성향의 신문사 댓글들과는 달리, 북한이탈주민의 위장 간첩 문제와 관련해 조작 가능성을 염두에 두고 있는 것으로 나타났다.

경향신문 댓글 작성자들 역시도 한겨레신문 댓글 작성자들과 같이 한국 내에서 정착하는 과정 가운데 겪었던 사회적 차별로 인해 고통 받아온 북한이탈주민들에 대해서 관심을 표명하고 있는 것으로 나타났다. 또한, 경향신문 댓글 작성자들은 앞서 설명된 한겨레신문 댓글 작성자들에서 보다 더 나아가 북한이탈주민의 위장 간첩 문제는 정권의 부당한 조사에 대한 진실 규명이 필요한 문제라고 생각하고 있음을 알 수 있었다.

결국 댓글의 연관어 분석 결과는 중앙일보를 제외하고, 신문사의 보수적 혹은 진보적인 성향이 댓글의 내용과 조응을 이루고 있는 것으로 나타났다. 이와 같은 결과는 신문사에서 신문을 이용하는 사람들에게 어떤 내용의 기사를 어떠한 관점에서 제공해주느냐에 따라, 기사 댓글의 내용

과 관점이 특정 범위 안에서 제약될 수 있는 가능성을 보여준다. 특히 북한이나 북한이탈주민에 대한 내용의 경우 분단의 현실로 인해 댓글 작성자들의 입장에서는 언론에서 제공해주는 정보에 의존적일 수밖에 없는 구조가 형성되어 있어, 언론에서 제공해주는 기사의 내용이 댓글 가운데 짙게 나타날 가능성이 존재한다.

다음으로 본 연구에서 시도했던 감성분석과 관련된 한계점들에 대해 언급하고자 한다.[78] 첫째, 감성분석의 경우 결과로 도출되는 값의 타당성을 검증하기 위해 예측값과 인간의 판단을 비교하는 준거타당도(criterion-related validity) 검증이 필요한데, 본 연구는 이를 검증하지 못했다는 점에서 한계가 지니고 있다.[79] 둘째, 감성사전을 이용한 감성분석의 경우 단순히 긍정과 부정을 나타내는 단어를 사용하였다는 점만을 근거로 문서의 긍정과 부정을 판단하지만, 실제로 감성을 나타내는 단어 주변에 어떠한 단어들이 존재하는지 혹은 문서 전체에 어떠한 맥락이 존재하는지 전혀 고려하고 있지 않다는 점에서 감성사전을 이용한 감성분석의 원초적인 한계점이 존재한다.[80] 더불어 보편성을 지향하는 감성사전을 분석 대상 자료에 맞게 재구성하지 않아 정확한 결과를 도출해내기 어려웠고, 또한 댓글이라는 자료가 신조어나 오탈자를 내포하고 있을 수 있는 가능성에 대해서 고려하지 못했다.

그럼에도 불구하고 본 연구는 북한이탈주민과 관련된 신문 기사들에 대한 토픽 분석을 시도하고 댓글 자료를 활용하여 연관어 분석과 감성분석을 시도함으로써 기사와 댓글 간의 조응 관계를 확인할 수 있었다. 특히, 댓글 작성자들은 기본적으로 북한이탈주민은 도움이 필요한 존재이며 그 도움은 한국사회 내에서 제공되어야 한다고 생각하지만, 자신이 구독하는 언론이 보수인지 혹은 진보인지 그 여부에 따라 북한이탈주민

---

78) 한국어 감성분석의 근본적인 어려움들에 대한 자세한 내용은 다음 자료를 참고. 남길임・조은경, 2017.

79) 백영민, 2017, p. 237.

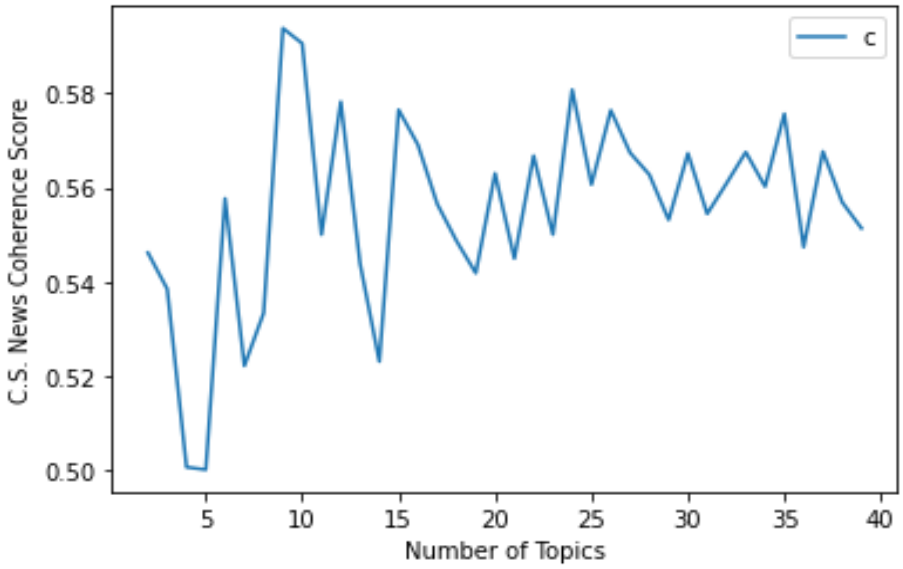
80) 다시 말해, 부정을 나타내는 단어 주변에 감성사전에 등재되지 않은 부정어휘가 사용되는 경우나(백영민, 2017, p. 253) 표면적으로 긍정적일지라도 반어법이나 야유를 드러내는 경우를 의미한다(윤태일・이수안, 2018, p. 257. 재인용: 임동훈, 『R을 이용한 빅데이터 분석』, (파주: 자유아카데미, 2015))

과 관련된 간첩 조사를 다르게 받아들이고 있음을 알 수 있었다.

끝으로 본 연구는 댓글 자료를 통하여 북한이탈주민에 대한 일반적인 한국 사람들의 태도를 확인하는 데 그 목적을 두고 있지는 않았지만, 본 연구의 결과를 통해서 북한이탈주민에 대한 한국 사람들의 태도 결정 요인에는 한 개인이 어떤 언론을 즐겨 찾는지가 중요한 요인으로 작용할 수 있다는 가능성을 제시할 수 있을 것이다.

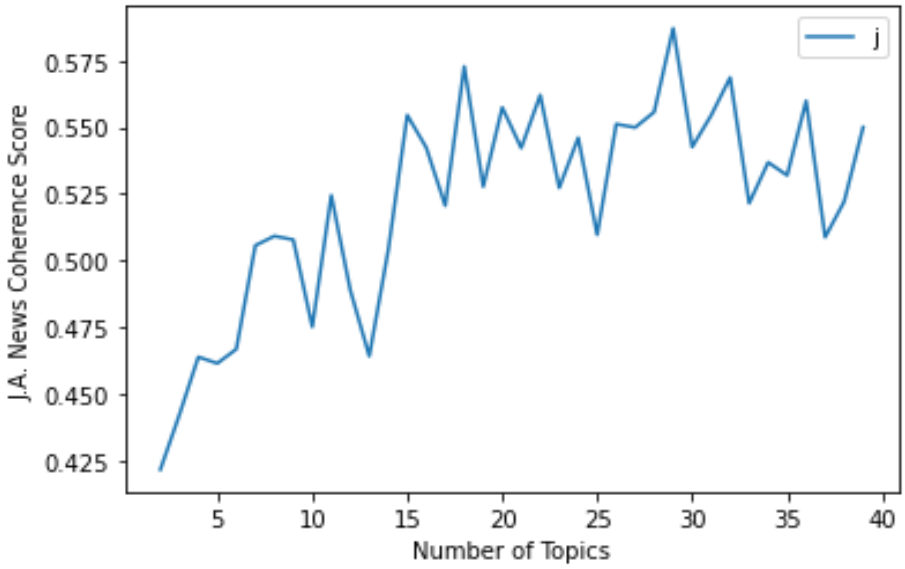
부 록

[그림 1] 토픽 개수에 따른 조선일보 기사의 일관성 점수 변화



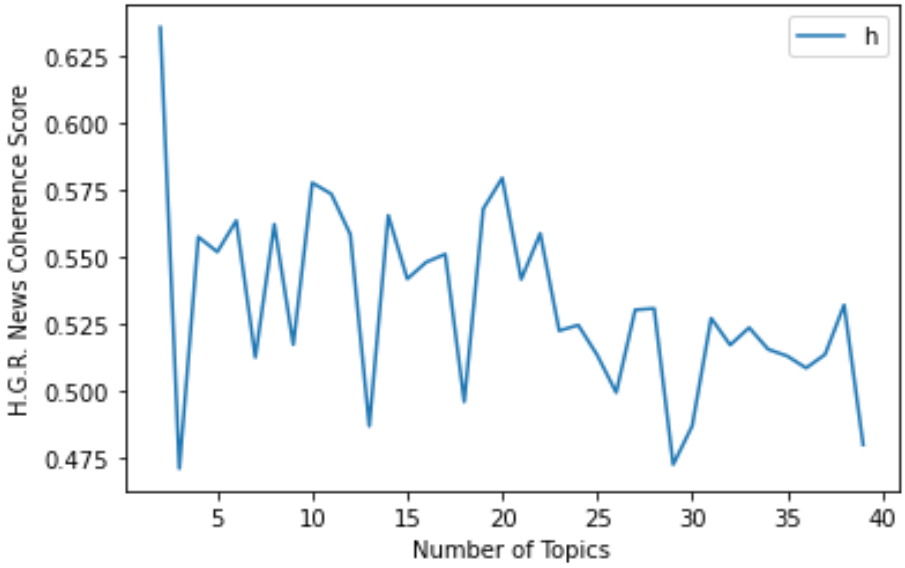
註: 10부터 20까지의 토픽 개수에서 가장 높은 일관성 점수를 보여줄 때의 토픽 개수는 10이다.

[그림 2] 토픽 개수에 따른 중앙일보 기사의 일관성 점수 변화



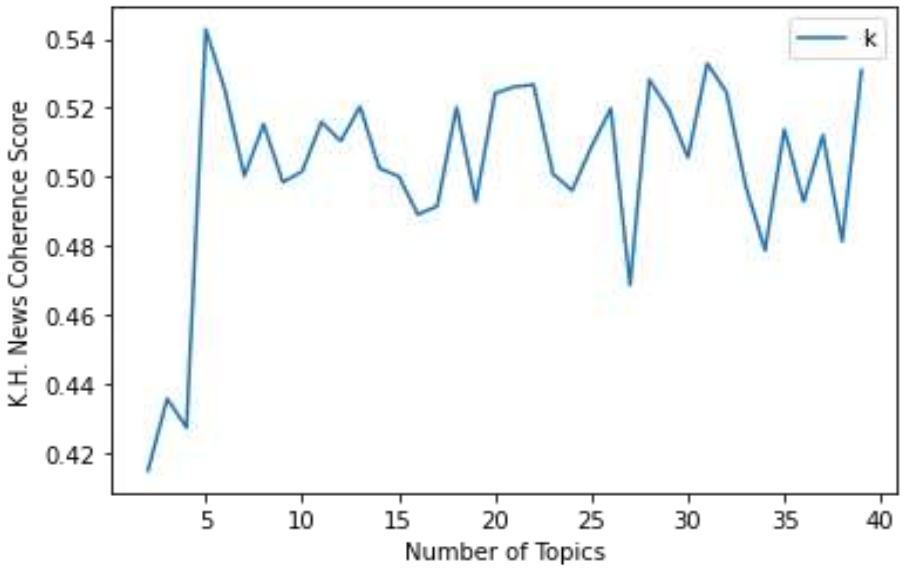
註: 10부터 20까지의 토픽 개수에서 가장 높은 일관성 점수를 보여줄 때의 토픽 개수는 18이다.

[그림 3] 토픽 개수에 따른 한겨레신문 기사의 일관성 점수 변화



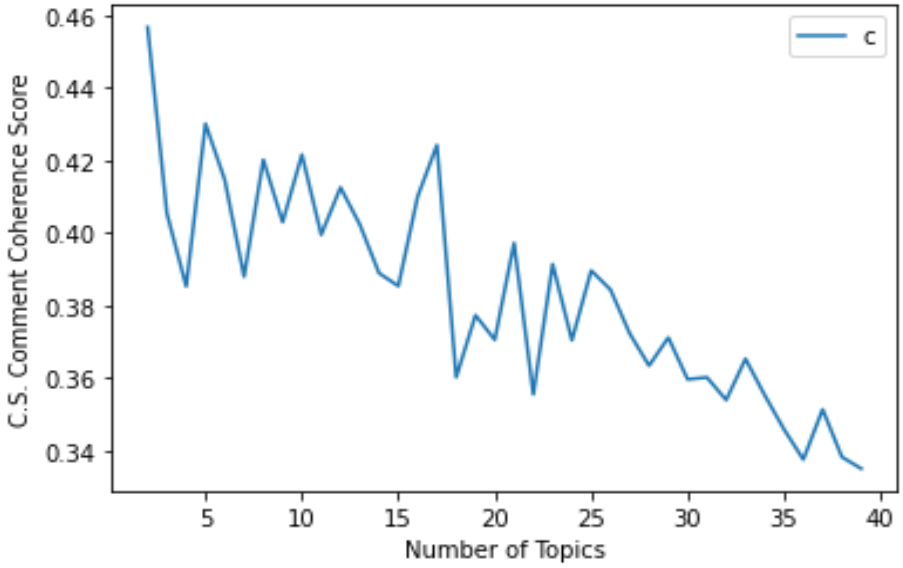
註: 10부터 20까지의 토픽 개수에서 가장 높은 일관성 점수를 보여줄 때의 토픽 개수는 20이다.

[그림 4] 토픽 개수에 따른 경향신문 기사의 일관성 점수 변화



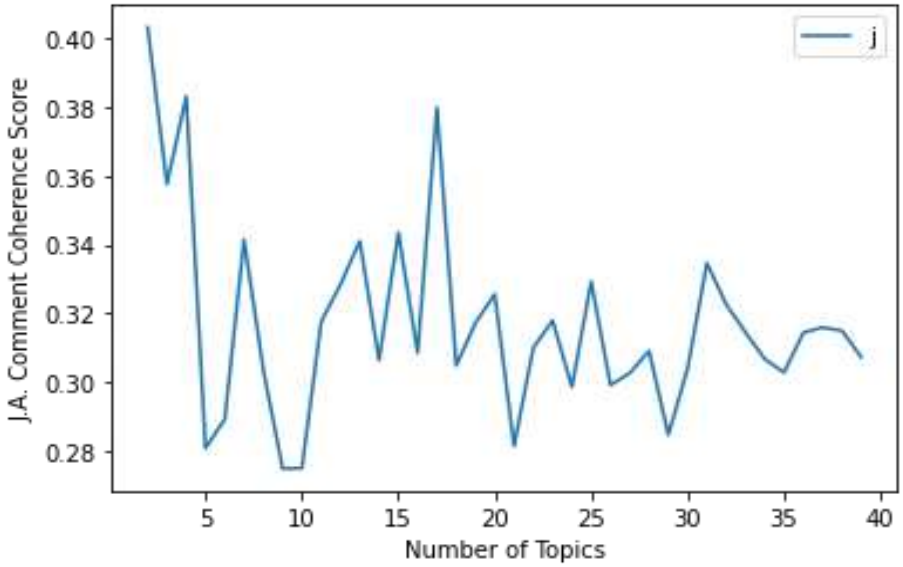
註: 10부터 20까지의 토픽 개수에서 가장 높은 일관성 점수를 보여줄 때의 토픽 개수는 20이다.

[그림 5] 토픽 개수에 따른 조선일보 댓글의 일관성 점수 변화



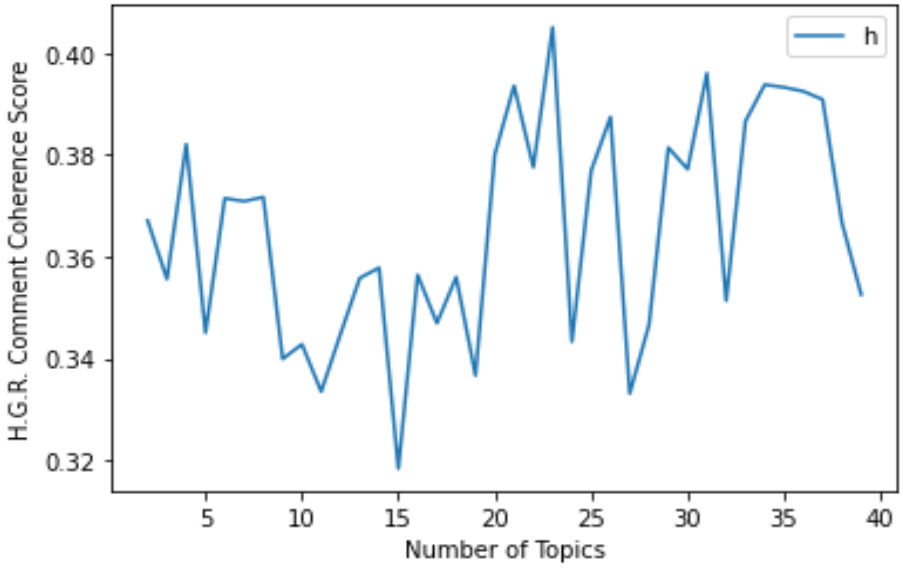
註: 10부터 20까지의 토픽 개수에서 가장 높은 일관성 점수를 보여줄 때의 토픽 개수는 17이다.

[그림 6] 토픽 개수에 따른 중앙일보 댓글의 일관성 점수 변화



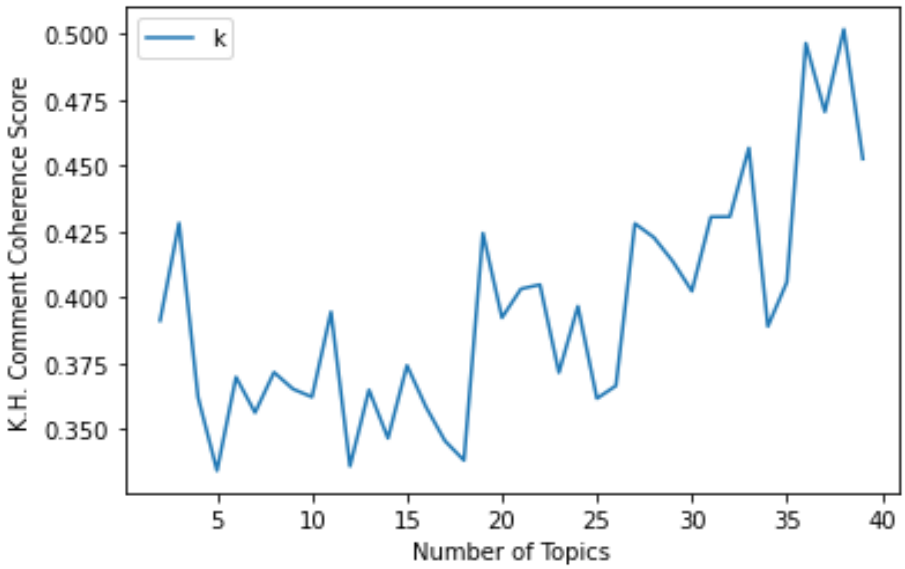
註: 10부터 20까지의 토픽 개수에서 가장 높은 일관성 점수를 보여줄 때의 토픽 개수는 17이다.

[그림 7] 토픽 개수에 따른 한겨레신문 댓글의 일관성 점수 변화



註: 10부터 20까지의 토픽 개수에서 가장 높은 일관성 점수를 보여줄 때의 토픽 개수는 20이다.

[그림 8] 토픽 개수에 따른 경향신문 댓글의 일관성 점수 변화



註: 10부터 20까지의 토픽 개수에서 가장 높은 일관성 점수를 보여줄 때의 토픽 개수는 19이다.

[표 1] 조선일보 기사 LDA 토픽 모델링 결과

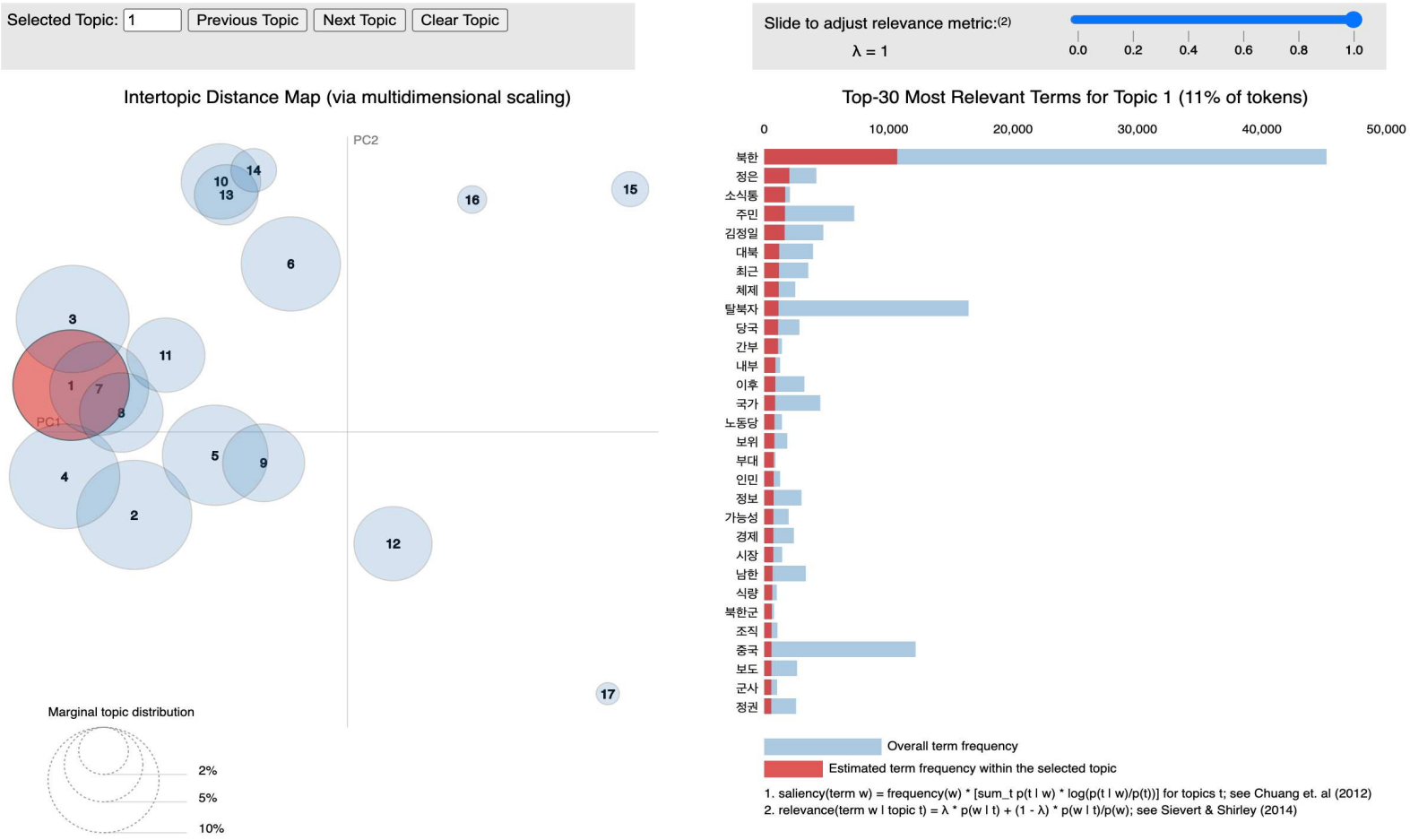
	비중	토픽에 해당되는 상위 30개 단어
1	11%	북한, 정은, 소식통, 주민, 김정은, 대북, 최근, 체제, 탈북자, 당국, 간부, 내부, 이후, 국가, 노동당, 보위, 부대, 인민, 정보, 가능성, 경제, 시장, 남한, 식량, 북한군, 조직, 중국, 보도, 군사, 정권
2	10.7%	북한, 대통령, 문제, 미국, 정부, 남북, 통일, 회담, 한국, 정책, 중국, 관계, 한반도, 경제, 정상, 평화, 협력, 국가, 대북, 상황, 안보, 외교, 대화, 생각, 장관, 국민, 일본, 국제, 필요, 해결
3	10.3%	중국, 탈북자, 북한, 한국, 국경, 사람, 체포, 가족, 일본, 일본, 공안, 대사관, 탈출, 입국, 보도, 정부, 여성, 당국, 미국, 지역, 마약, 조선, 최근, 태국, 라오스, 보위, 기자, 전화, 취재, 확인, 강제
4	9.9%	북한, 인권, 정부, 탈북자, 문제, 미국, 유엔, 국제, 수용소, 강제, 난민, 주민, 한국, 단체, 지원, 조사, 정치범, 자유, 관련, 사회, 송환, 위원회, 중국, 보고서, 국가, 상황, 고문, 이날, 이번, 북한인
5	9%	교육, 학교, 학생, 대학, 탈북자, 한국, 사회, 지원, 북한, 주민, 학과, 통일, 전형, 교수, 생활, 정착, 영어, 과정, 청소년, 이탈, 서울, 이상, 경우, 남한, 입학, 취업, 선발, 공부, 모집, 교사

6	8%	영화, 방송, 감독, 사람, 이야기, 배우, 기자, 드라마, 공연, 작품, 출연, 생각, 모습, 연기, 프로그램, 사랑, 영화제, 한국, 노래, 자신, 조선, 작가, 제작, 촬영, 다큐멘터리, 관객, 오후, 사진, 무대, 관심
7	7.9%	북한, 자유, 단체, 대북, 전단, 탈북자, 운동, 사람, 대표, 정권, 살포, 대한민국, 기자, 삐라, 정부, 조선, 주민, 비판, 연합, 주장, 민족, 민주화, 국민, 내용, 체제, 사회, 세력, 비난, 생각, 방송
8	5.7%	혐의, 사건, 간첩, 국정원, 수사, 검찰, 조사, 경찰, 국가, 정보, 탈북자, 북한, 사실, 서울, 구속, 부장, 검사, 재판, 주장, 법원, 변호사, 공작원, 진술, 기소, 증거, 국내, 선고, 관련, 보위, 위장
9	5.4%	의원, 통일, 대표, 서울, 국민, 회장, 국회, 이날, 누리, 집회, 민주당, 단체, 후보, 행사, 오후, 위원회, 한국, 참석, 시민, 위원, 인사, 위원장, 선거, 통합, 탈북자, 민주, 장관, 한나라, 정치, 참여
10	5.1%	사람, 가족, 여성, 아버지, 북한, 생각, 남한, 아들, 어머니, 결혼, 엄마, 포로, 남편, 국군, 마음, 생활, 아내, 납북, 고향, 이야기, 눈물, 자신, 처음, 지금, 당시, 출신, 부모, 남자, 병원, 모습
11	4.9%	북한, 선상, 주민, 전화, 방송, 정보, 평양, 휴대, 지역, 사용, 귀순, 시설, 아파트, 당국, 정도, 경우, 구역, 최근, 실험, 확인, 발견, 사고, 이용, 건설, 조사, 어선, 사람, 보도, 공급, 공장

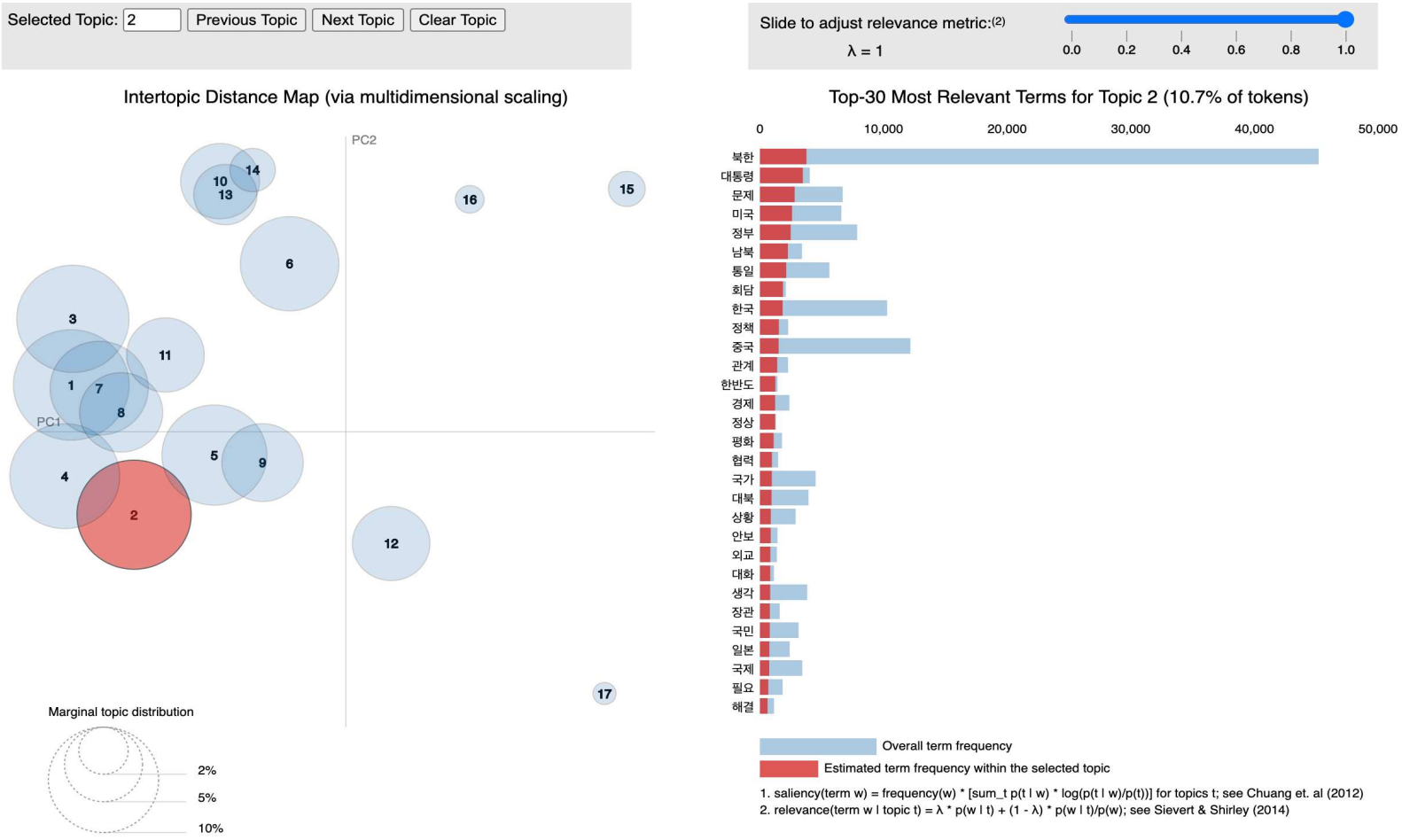
12	4.9%	기획부, 지원, 사업, 사회, 기업, 지역, 과장, 문화, 재단, 복지, 활동, 운영, 나눔, 기부, 센터, 봉사, 교육, 사회적, 프로그램, 사이버, 제공, 관리, 장애인, 금융, 계층, 다문화, 서비스, 창업, 계획, 올해
13	3.3%	김정일, 김일성, 평양, 북한, 위원장, 정은, 사진, 신문, 조선, 사망, 행사, 노동당, 비서, 중앙, 모습, 방문, 보도, 주민, 국방, 최고, 탈북자, 당시, 동상, 모자, 생일, 기사, 노동, 인민, 부자, 방북
14	1.7%	정은, 정남, 성택, 경희, 김정일, 조선, 모란봉, 북한, 암살, 공개, 출신, 망명, 부장, 부인, 후계자, 권력, 설주, 정철, 인물, 측근, 활동, 여정, 정보, 클럽, 처형, 천리마, 고영희, 공연, 당시, 민방위
15	1.1%	아나운서, 상준, 시하, 선수, 경기, 축구, 대회, 야구, 스포츠, 올림픽, 영란, 프로, 현미, 한주, 챔피언, 고양, 야구단, 골프, 시즌, 훈련, 체육, 감독, 여자, 리그, 세계, 대표팀, 경기장, 월드컵, 기자, 농구
16	0.7%	목사, 교회, 기독교, 종교, 스페인, 선교, 기도, 교인, 활동, 백씨, 한국, 예배, 신자, 선교사, 하나님, 습격, 신앙, 담임, 성경, 기도회, 지하, 자유, 복음, 예수, 건물, 기독교인, 목회자, 조선, 서울, 화해
17	0.4%	팀장, 방송, 영상, 제작, 기술, 편성, 서기관, 기획, 본부, 국장, 회담, 실무, 불교, 보도국, 총무국, 승진, 경영, 재정, 울산, 총무, 설립, 기획실, 정책, 총괄, 단장, 사무소, 남북, 정착, 교육, 이탈

註: 토픽 개수가 17일 때, 로그(log)를 취한 혼란도는 -8.2996, 일관성 점수는 0.5566이다.

[그림 9-1] pyLDAvis를 이용한 조선일보 기사 LDA 토픽 모델링 결과 시각화 예시 (1)



[그림 9-2] pyLDAvis를 이용한 조선일보 기사 LDA 토픽 모델링 결과 시각화 예시 (2)



[표 2] 중앙일보 기사 LDA 토픽 모델링 결과

	비중	토픽에 해당되는 상위 30개 단어
1	20.8%	대통령, 북한, 문제, 미국, 중국, 프라이, 정부, 회담, 한국, 관계, 정상, 남북, 평화, 국가, 정책, 안보, 국민, 외교, 위원장, 경제, 한반도, 회의, 사회, 상황, 어진, 트럼프, 대화, 장관, 생각, 일본
2	13%	북한, 탈북자, 중국, 인권, 한국, 정부, 하달, 문제, 유엔, 미국, 정치범, 관련, 강제, 당국, 대사관, 독식, 국제, 기자, 대사, 국가, 현지, 수용소, 난민, 망명, 체포, 보고서, 신변, 사건, 주장, 송환
3	12.9%	북한, 지시문, 평양, 김일성, 사진, 주민, 중앙, 조선, 각급, 연초, 방송, 위원장, 중국, 모습, 신문, 탈북자, 체제, 대북, 최근, 등장, 제재, 정은, 인민, 공개, 위원, 노동, 이후, 경우, 매체, 당국
4	10.8%	한국, 사람, 영화, 생각, 사진, 기자, 탈북자, 가족, 서울, 여성, 교수, 사회, 이야기, 마음, 대표, 자신, 감독, 시작, 친구, 사랑, 모습, 처음, 남한, 문화, 관심, 세계, 지금, 행사, 출신, 활동
5	6.4%	북한, 주민, 통일부, 정부, 이탈, 센터, 정보, 조사, 지원, 정착, 통일, 한국, 연구, 남한, 기자, 귀순, 연구원, 탈북자, 사회, 여성, 경우, 남북, 종업원, 기관, 입국, 정책, 관련, 식당, 국내, 보호

6	5.7%	교육, 대학, 학생, 학교, 학과, 활동, 사이버, 지원, 사회, 한국, 과정, 입학, 장학금, 학부, 졸업, 청소년, 과학, 학년, 전형, 분야, 공부, 수업, 평가, 시험, 서울, 대학생, 다양, 교수, 전문, 이상
7	5.6%	지원, 그랜드, 대국민, 일괄, 일회성, 태영, 기업, 사업, 사회, 지역, 신변, 금융, 복지, 사회적, 활동, 회장, 은행, 계층, 대상, 재단, 운영, 제공, 그룹, 업체, 나눔, 올해, 봉사, 서비스, 공헌, 계획
8	4%	탈북자, 경찰, 서울, 기자, 전화, 병원, 여성, 발견, 확인, 모자, 관계자, 조사, 아들, 오후, 사기, 아파트, 사건, 번호, 사망, 신청, 중국, 신고, 대구, 가족, 사람, 경찰서, 보험, 발굴, 브로커, 장세정
9	3.9%	북한, 대북, 미사일, 전단, 발사, 살포, 군사, 공격, 정보, 훈련, 실험, 북한군, 도발, 무대, 무기, 개발, 미국, 풍선, 연평도, 전쟁, 위협, 탄도, 천안, 대응, 시설, 국방부, 관계자, 당국, 지역, 서해
10	3.3%	국정원, 혐의, 검찰, 간첩, 사건, 수사, 조사, 국가, 부장, 정보, 재판, 증거, 변호사, 기소, 형사, 서울, 조작, 사실, 진술, 국내, 선고, 보위, 검사, 범죄, 기록, 법원, 마약, 구속, 중앙, 위장
11	3.1%	정은, 북한, 생일, 김정일, 확인, 단체, 후계자, 비서, 정보, 국가, 조직, 당국자, 탈북자, 내부, 진행, 정부, 기념, 충성, 기념일, 노동당, 위원장, 국방, 간부, 행사, 지도자, 관련, 가능성, 지시, 보도, 영종

12	3%	의원, 후보, 대표, 국회, 민주당, 누리, 정치, 선거, 통합, 민주, 대선, 기자, 보수, 국민, 자유, 총선, 출신, 원내, 한국, 국회의원, 야당, 위원장, 위원, 비례, 정당, 진보, 공천, 철수, 수경, 여야
13	2.6%	경제, 영상, 추천, 과장, 코로나, 시장, 선언, 보기, 맥주, 정책, 이전, 성장, 다음, 산업, 마스크, 국가, 기획, 기업, 박진, 관리, 전보, 통합, 부장, 강남, 개혁, 열풍, 도전, 튜브, 회담, 종로
14	2%	단체, 운동, 연합, 목사, 대표, 집회, 자유, 기자, 회원, 민주화, 시민, 서울, 당시, 광주, 보수, 주장, 이날, 시위, 뉴스, 경찰, 교회, 진보당, 탈북자, 반대, 행사, 임진각, 비판, 주사, 경기, 개입
15	1.1%	주택, 공급, 특별, 선발, 전형, 모집, 반영, 아파트, 일반, 임대, 계열, 가구, 경우, 수능, 영역, 성적, 기준, 인문, 등급, 면접, 입주, 단계, 수시, 생물, 소득, 학생부, 인원, 단지, 건축, 서류
16	0.9%	남북, 해결, 추진, 방안, 대화, 지원, 국군, 포로, 납북, 북한, 회담, 문제, 인도적, 송환, 방식, 개성, 대북, 통일부, 협력, 계획, 제시, 관련, 정부, 업무, 필요, 당국, 과제, 본격, 창의, 타결
17	0.9%	통일, 동독, 녹화, 주안점, 선수, 독일, 남북, 올림픽, 서독, 경이, 아이스하키, 준비, 축구, 금리, 대회, 비용, 아시안, 기금, 남북한, 체육, 게임, 단일팀, 분단, 국민, 민간, 부의장, 농구, 저축, 베를린, 인터넷

註: 토픽 개수가 17일 때, 로그(log)를 취한 혼란도는 -8.6575, 일관성 점수는 0.5207이다.

[표 3] 한겨레신문 기사 LDA 토픽 모델링 결과

	비중	토픽에 해당되는 상위 30개 단어
1	15.9%	사회, 사람, 생각, 한국, 사회적, 문제, 시장, 시민, 도덕, 경제, 얘기, 평양, 시작, 모습, 지금, 정치, 국민, 시대, 기업, 자신, 중요, 교육, 청년, 정도, 주의, 국가, 지역, 차별, 경험, 이야기
2	15.8%	북한, 미국, 통일, 중국, 문제, 한국, 체제, 관계, 정부, 평화, 국가, 남북, 전쟁, 한반도, 정책, 남한, 경제, 상황, 이후, 진입, 제재, 대통령, 처사, 전가, 안보, 대북, 지역, 협력, 정은, 독일
3	14.7%	지원, 서울, 교육, 한국, 사업, 센터, 기획, 제창, 정책, 재단, 연구, 지역, 사회, 복지, 협력, 운영, 관리, 문화, 오후, 주민, 본부, 활동, 기술, 금융, 누리, 국제, 과학, 정보, 이탈, 청소년
4	13.6%	사건, 검찰, 수사, 국가, 대통령, 주장, 정보, 당시, 보도, 관련, 서울, 경찰, 기자, 법안, 내용, 경찰청, 법원, 조사, 의원, 판결, 국민, 인론, 변호사, 비호, 재판, 개입, 활동, 자료, 의혹, 정부
5	7.3%	탈북자, 북한, 사람, 중국, 한국, 남한, 주민, 자신, 특권, 교감, 공기업, 가족, 센터, 솜씨, 난민, 자백, 생각, 아버지, 사건, 어머니, 인터뷰, 이야기, 여성, 전화, 내가, 처음, 신문, 조사, 이탈, 조선족
6	5.4%	전단, 살포, 단체, 대북, 탈북자, 정부, 자유, 운동, 연합, 사태, 북한, 대통령, 봉쇄, 경기도, 파주, 기자, 대표, 회원, 주장, 담화, 주민, 경찰, 방송, 진정, 비난, 군사, 상황, 여명, 오전, 인기

7	4.9%	국정원, 간첩, 진술, 여동생, 증언, 탈북자, 사진, 북한, 행위, 사실, 공개, 가족, 화교, 증거, 아버지, 공소장, 입북, 보위, 내용, 노래방, 한겨레, 출신, 생각, 혐의, 사건, 변호사, 폭행, 중국, 두만강, 대부분
8	4.1%	북한, 개성, 공단, 남북, 남쪽, 회담, 당국, 대화, 북쪽, 연습, 대통령, 군사, 방침, 문제, 정부, 정상, 훈련, 위원장, 통일부, 폐쇄, 대북, 제의, 한겨레, 기념일, 김포시, 준수, 청와대, 진정, 실무, 합동
9	3.8%	학교, 대학, 사원, 신입, 학생, 전형, 학과, 입학, 드라마, 스쿨, 방송, 졸업, 공부, 교과서, 이상, 경우, 성적, 기준, 선발, 교육, 기자, 부모, 전공, 사이버, 자동차, 제도, 일반, 자녀, 학년, 소득
10	3.7%	영화, 문화, 한국, 이야기, 감독, 일본, 작품, 교수, 영화제, 작가, 배우, 주제, 공연, 기자, 촬영, 국제, 동포, 위안부, 사진, 기지, 콘서트, 통일, 분단, 연기, 강원도, 평화, 장면, 승진, 제공, 세계
11	3.5%	정부, 북한, 주민, 명성, 기자, 보도, 종업원, 지원, 통일부, 보호, 탈북자, 식당, 관련, 집단, 확인, 공개, 입국, 사실, 관계자, 이탈, 외교부, 센터, 정보, 언론, 재단, 공사, 요청, 당국자, 총선, 국내
12	3.2%	인권, 북한, 조사, 유엔, 형사, 침해, 국제, 사회, 문제, 위원회, 보고서, 국가, 권고, 수용소, 결의, 범죄, 고문, 결의안, 위원장, 개선, 경위, 주장, 상황, 권법, 보호, 재판소, 수사, 보장, 구금, 지난해

13	1.8%	집회, 연합, 어버이, 단체, 보수, 목사, 전경련, 교회, 세월, 서울, 기독교, 기자, 청와대, 계좌, 자금, 극우, 탈북자, 반대, 동원, 대표, 시위, 대한민국, 한겨레, 재심, 의혹, 회원, 광장, 사무실, 세력, 일당
14	1%	의원, 작가, 소설, 문학, 비례, 냉면, 음식, 요리, 인터뷰, 치료, 이주, 공연, 사랑, 명상, 이주민, 마음, 누리, 기자, 작품, 다문화, 밥상, 여성, 유치원, 국회, 문학상, 처음, 시인, 광장, 국회의원, 참가자
15	0.6%	교황, 처장, 부장, 지청, 공항, 도시, 인천, 주교, 지검, 부부장, 건설, 철도, 세월, 미사, 이날, 교통, 인사, 신자, 천주교, 스님, 봉제, 공장, 참석, 환영, 건축, 지점장, 출입문, 이주, 사제, 차장
16	0.4%	상무, 원상, 완성, 광주, 선수, 동성애, 유족, 송도, 경희, 중앙일보, 노래, 기념식, 보훈처, 목란, 포스, 착수, 상원, 출전, 답방, 이정, 낙원, 스키, 소비자, 대기자, 박찬, 청진시, 중원, 장례식, 전임, 씨알
17	0.3%	달력, 은주, 한국행, 오스트, 법상, 결혼식, 억류자, 행방불명, 레일리, 국적, 밀입국, 박태, 하루, 일력, 할머니, 믿음, 무국적, 판정, 인쇄, 숫자, 노숙인, 선언문, 장시간, 제공, 도움, 병원, 감사, 유전자, 단련, 조립

註: 토픽 개수가 17일 때, 로그(log)를 취한 혼란도는 -9.103, 일관성 점수는 0.5511이다.

[표 4] 경향신문 기사 LDA 토픽 모델링 결과

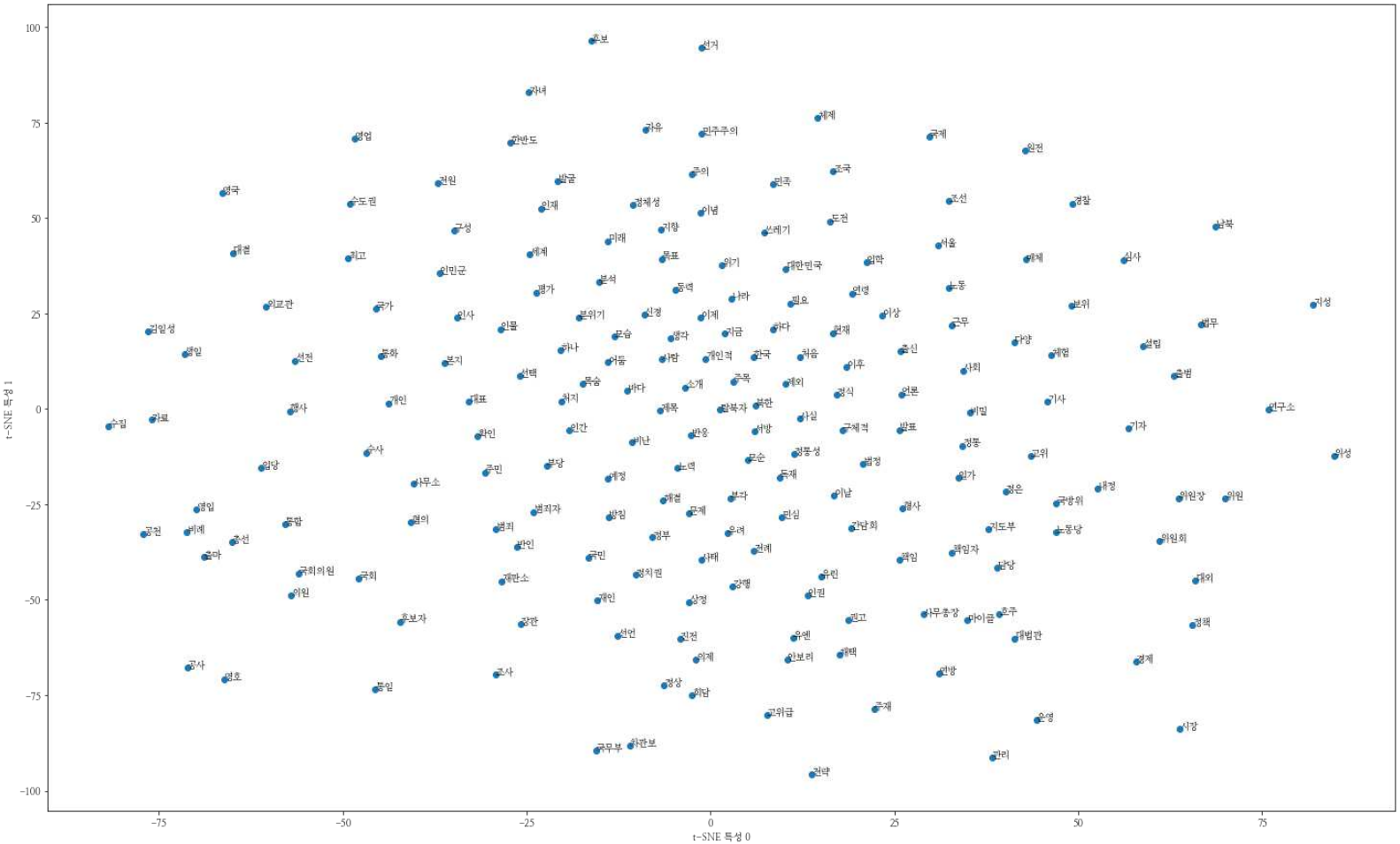
	비중	토픽에 해당되는 상위 30개 단어
1	12.7%	북한, 중국, 활동, 탈북자, 미국, 남조선, 납치, 확인, 당국, 인권, 지원, 대북, 보도, 요청, 주체사상, 핵심, 운동권, 일본, 가능성, 유학, 통신, 국가, 정부, 설명, 체포, 안전, 석방, 당시, 대변인, 전단
2	11.5%	검찰, 혐의, 탈북자, 국정원, 사건, 선양, 북한, 조사, 수사, 국가, 중국, 간첩, 증가, 사실, 기소, 조작, 재판, 서울, 확인, 정보, 진술, 제출, 위조, 과정, 법원, 공무원, 검사, 선고, 내용, 주장
3	10.4%	인권, 북한, 문제, 남자, 연구원, 사람, 탈북자, 해외, 주장, 유엔, 반박, 통일, 최근, 제기, 국제, 탈출, 규정, 조국, 지적, 당의, 정부, 언론, 비판, 체제, 지옥, 대응, 평화, 내용, 출신, 위원회
4	9.3%	북한, 주민, 이탈, 지원, 의료, 사회, 건강, 복지, 병원, 경제, 필요, 수익, 아산, 국가적, 사업, 교육, 지역, 재정적, 제도, 조사, 대상, 정보, 충남, 서비스, 기관, 제공, 센터, 공공, 생활, 국민
5	9.1%	사람, 생각, 한국, 통일, 영화, 이야기, 전쟁, 지금, 세계, 평화, 사회, 무마, 자신, 남북, 모습, 소설, 현실, 취재, 마음, 교수, 감독, 청년, 이후, 작가, 분단, 얘기, 정도, 미국, 시작, 나라

6	8.4%	의원, 정부, 국회, 통합, 남북, 탈북자, 대표, 라오스, 한국, 공단, 누리, 진보당, 민주, 장관, 위원, 개성, 통일부, 위원장, 관련, 민주당, 회의, 총영사, 대책위, 혁명당, 거취, 대책, 북한, 외교, 입장, 이번
7	6.5%	대통령, 국민, 정치, 국가, 박근, 대한민국, 발언, 보수, 세력, 정당, 시민, 세월, 자유, 청와대, 후보, 인사, 의원, 선거, 대선, 안보, 이념, 고수, 경제, 개혁, 누리, 자신, 나라, 표현, 노무현, 한국
8	5.8%	단체, 연합, 대표, 운동, 기자, 수거, 서울, 집회, 이날, 사무총장, 수색, 판로, 행사, 비드, 재활용, 회원, 오전, 시민, 회견, 의혹, 라이트, 보수, 탈북자, 목사, 피해자, 광주, 교회, 자유, 오후, 당시
9	5.7%	과장, 부장, 기획, 지원, 정책, 소각, 팀장, 본부, 센터, 협력, 관리, 본부장, 원장, 실장, 안전, 통일, 사업, 교육, 감사, 위원회, 남북, 산업, 국장, 지방, 서울, 연구, 기술, 공무원, 전보, 승진
10	5.1%	한화, 교육, 다문화, 사업, 학생, 워크숍, 대학, 전형, 구미, 사이버, 문화, 해양, 학과, 동반, 프로그램, 사회, 지원, 활동, 모집, 입학, 학교, 참여, 선발, 올해, 국제, 제공, 봉사, 다양, 학습, 운영
11	5.1%	학교, 난민, 사람, 사회, 한국, 탈북자, 생활, 생각, 남한, 자신, 청소년, 정착, 차별, 북한, 주민, 어머니, 경우, 학생, 아들, 이유, 임대, 가족, 자녀, 보호, 남편, 여명, 부모, 아파트, 가명, 경제적

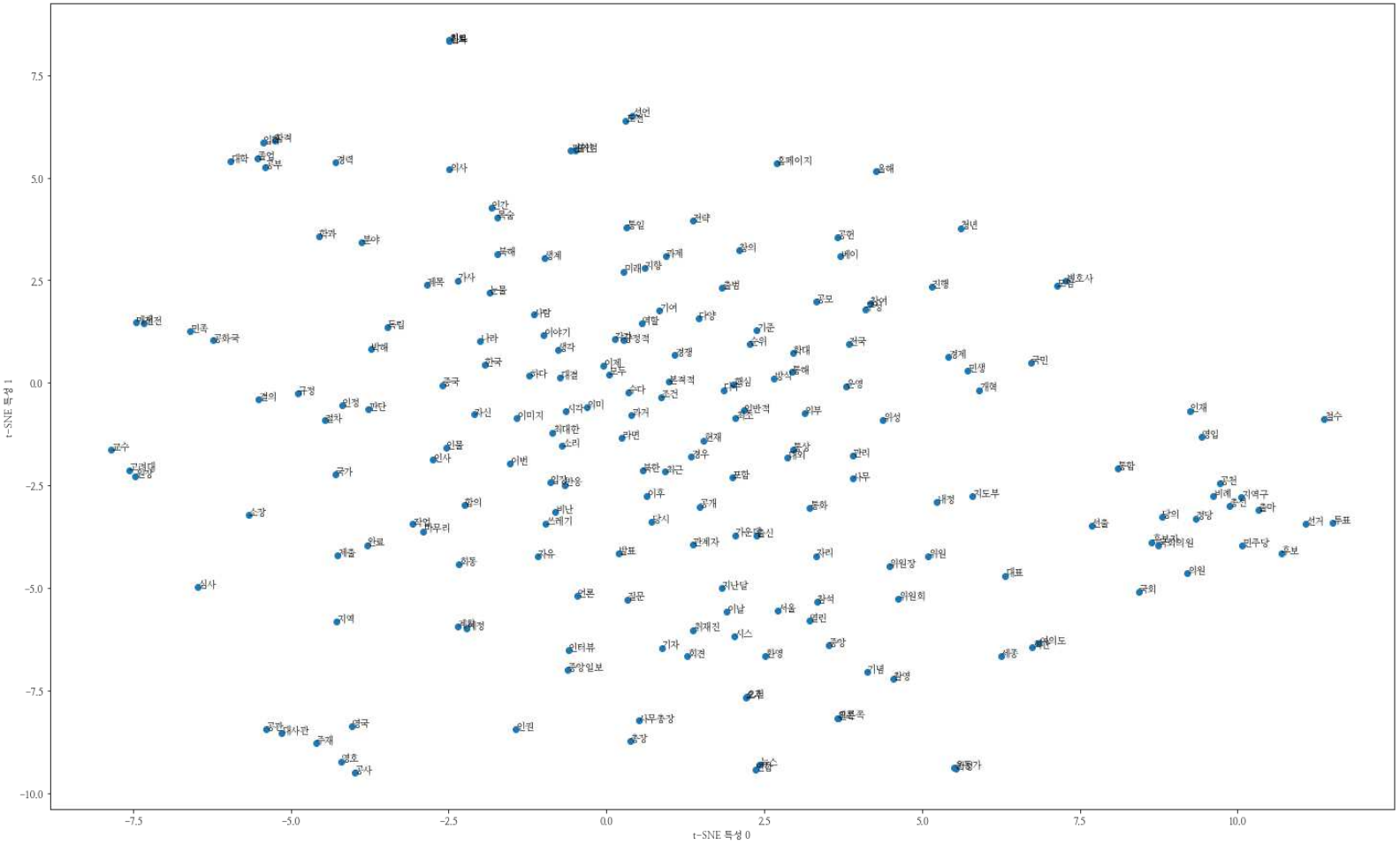
12	3.2%	기업, 지원, 사회적, 환경, 그룹, 지역, 고용, 문제, 창출, 선정, 경영, 위치, 제품, 마케팅, 백두, 주민, 이탈, 나눔, 북한, 계층, 금융, 소외, 성장, 회장, 기부, 재단, 전달, 대상, 기금, 분야
13	2.2%	변호사, 사건, 접견, 법대, 간첩, 필명, 선처, 국정원, 센터, 조작, 자백, 민주, 국가, 고문, 조사, 사회, 민변, 강철, 신문, 보호, 북한, 동안, 보안법, 탈북자, 장경, 허위, 재판, 변호인, 합동, 종업원
14	1.9%	경찰, 국방부, 경찰서, 마약, 여수, 개혁, 인천, 텍스, 부대, 경기, 기무사, 콘서트, 필로폰, 인력, 관계자, 기무, 내란, 경찰청, 경찰관, 계획, 지역, 살인, 청와대, 국군, 투약, 구속, 병원, 사찰, 음모, 선물
15	1.5%	한국, 평양, 이주, 출신, 일본, 가족, 영국, 결혼, 노동자, 공연, 공사, 연극, 카페, 이주민, 할머니, 사람, 베트남, 결혼식, 마이, 독일, 한인, 문화, 극단, 영철, 초청, 식당, 센터, 패션, 서울, 다문화
16	1.2%	중국, 여성, 한국, 국경, 남성, 강제, 한국인, 중국인, 매매, 노동자, 조선족, 아동, 부부, 인신매매, 외국인, 장애, 공안, 장애인, 결핵, 노동, 입양, 현지, 인신, 남편, 탈북자, 탈출, 여자, 자국, 러시아, 가정
17	0.5%	의류, 상속, 작품, 형제, 드라마, 작가, 재산, 대행, 영상, 중계, 압록강, 생존자, 대사, 현실, 소멸, 설치, 지정, 비디오, 가이드, 시청자, 도시, 신자, 정과, 타이틀, 전시, 후원자, 보좌관, 비엔날레, 원형, 무진

註: 토픽 개수가 17일 때, 로그(log)를 취한 혼란도는 -8.8609, 일관성 점수는 0.4915이다.

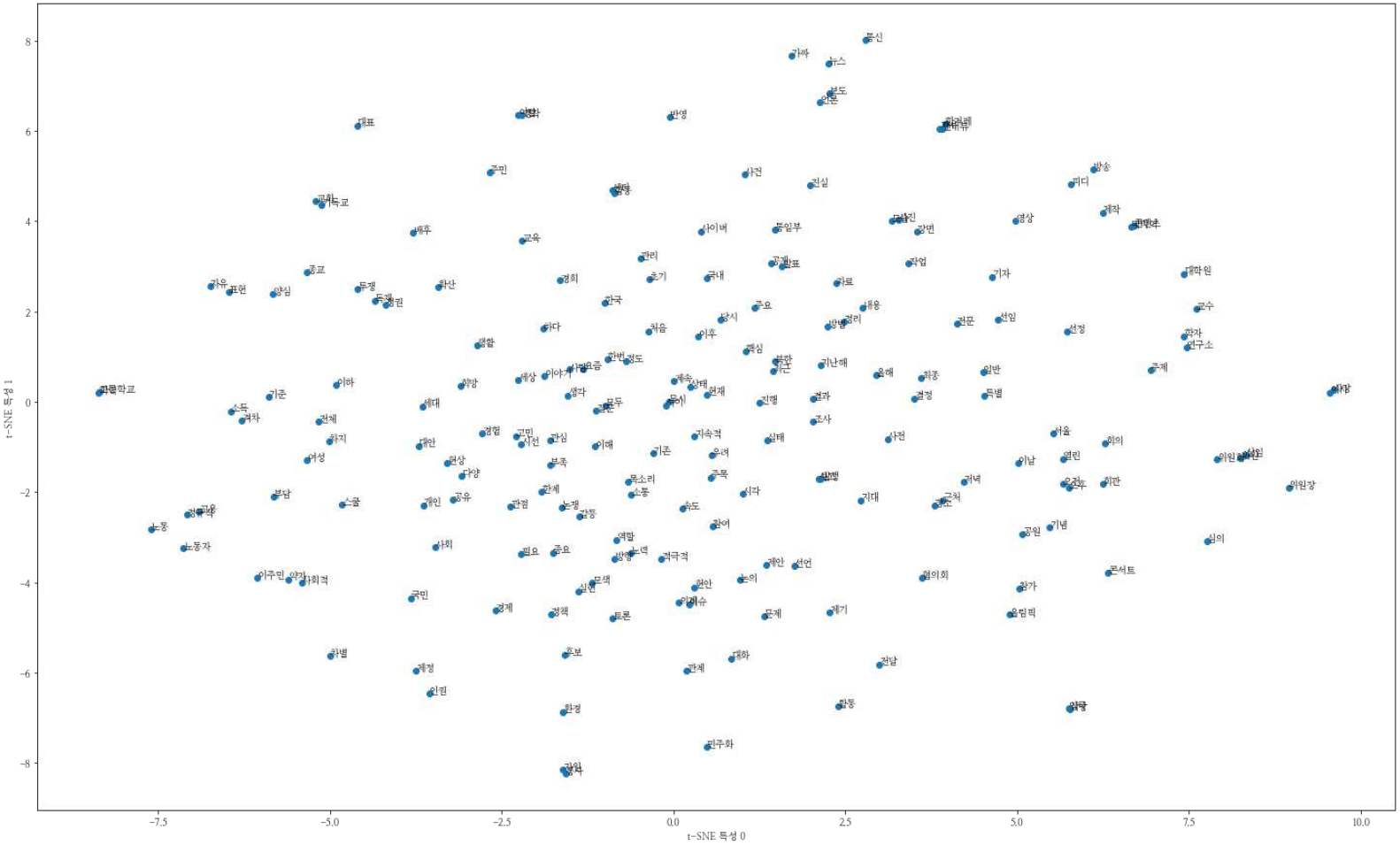
[그림 10] 조선일보 본문 연관어 분석 결과



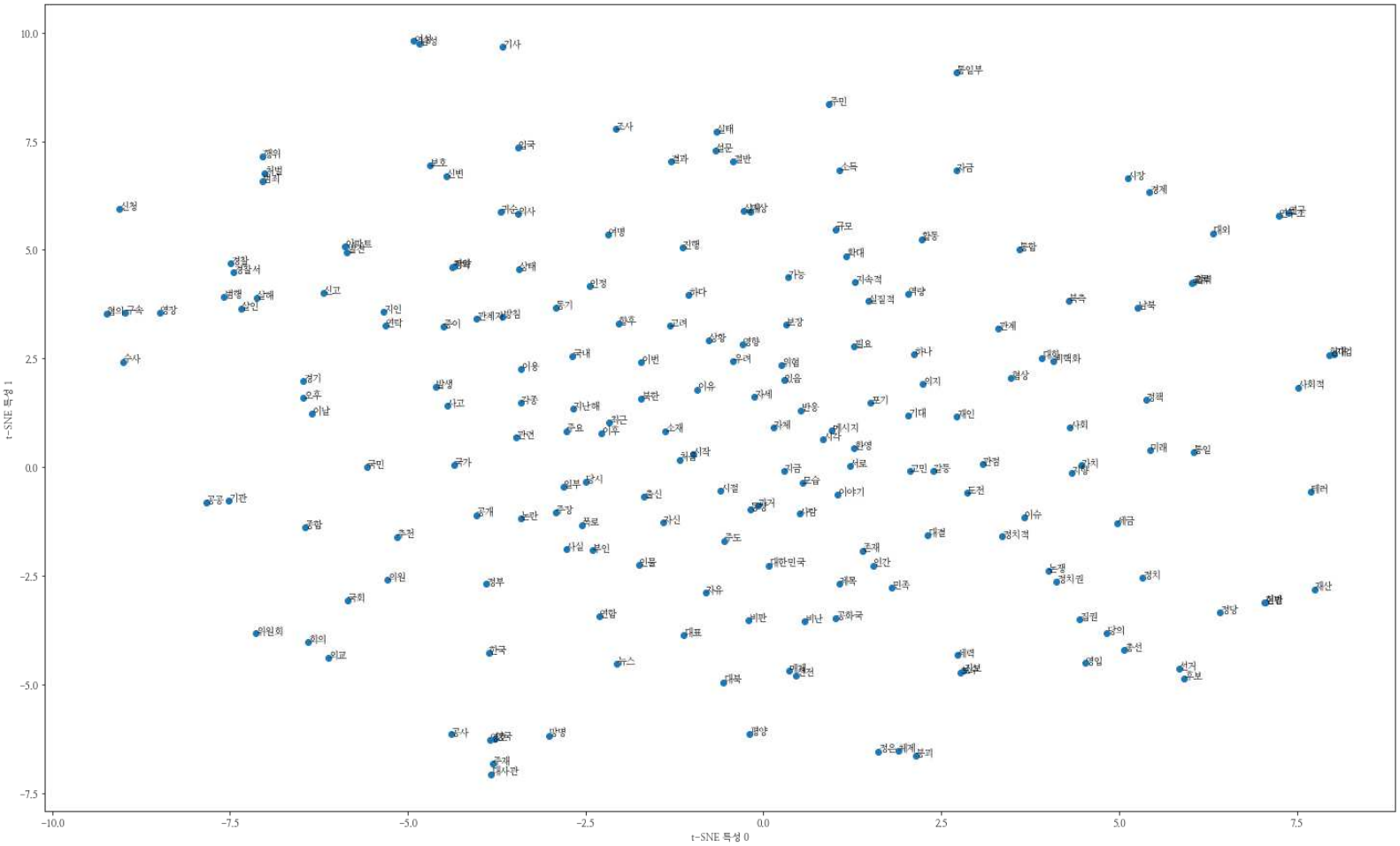
[그림 11] 중앙일보 본문 연관어 분석 결과



[그림 12] 한겨레신문 본문 연관어 분석 결과



[그림 13] 경향신문 본문 연관어 분석 결과



[표 5] 조선일보 댓글 LDA 토픽 모델링 결과

	비중	토픽에 해당되는 상위 30개 단어
1	11.7	국민, 탈북자, 대한민국, 정부, 북한, 자유, 문제, 정권, 국가, 대하, 보호, 나라, 중국, 위하, 만들, 좌파, 이렇, 보내, 지금, 외교, 정은, 눈치, 대통령, 북괴, 외교부, 목사, 헌법, 그러, 사람, 안보
2	10.5	탈북자, 한국, 중국, 북한, 남한, 사회, 사람, 미국, 적응, 생각, 난민, 국가, 나라, 대사관, 조선족, 그러, 한국인, 어떻, 그리, 만들, 힘들, 들어오, 중국인, 어렵, 민족, 정부, 필요, 경제, 국제, 일하
3	9.2%	북한, 자유, 탈북자, 주민, 통일, 위하, 대한민국, 체제, 남한, 대하, 독재, 정권, 동포, 정부, 정은, 목숨, 사람, 세습, 민주주의, 공산, 붕괴, 탈출, 실상, 세력, 통하, 지금, 평화, 세계, 누리, 국가
4	9%	인권, 북한, 탈북자, 문제, 김정일, 대하, 중국, 위하, 정권, 주민, 사람, 인간, 외면, 동포, 이야기, 유엔, 관심, 생각, 노무현, 납북, 침묵, 유린, 나라, 모르, 해결, 중요, 대통령, 운동, 그렇, 세계

5	8.4%	탈북자, 민변, 북한, 정은, 수경, 인권, 대한민국, 재인, 변호사, 좌파, 변절자, 배신자, 배신, 민주당, 보내, 조국, 생각, 의원, 국회의원, 자라, 사람, 집단, 말하, 나라, 만들, 단체, 인민, 대하, 출신, 위하
6	6.4%	사람, 탈북자, 생각, 대한민국, 남한, 정부, 정착, 국민, 생활, 이해, 힘들, 어렵, 모르, 고향, 노력, 돌아가, 그렇, 도와주, 책임, 정도, 나라, 성공, 정착금, 바라, 필요, 지원, 세상, 데려오, 그러, 위하
7	5.6%	탈북자, 간첩, 정보, 보내, 북한, 위장, 국정원, 당국, 사실, 처벌, 남한, 수사, 사람, 광주, 방송, 관리, 명단, 돌리, 사건, 하다, 사태, 나오, 입북, 공작, 색출, 이번, 의심, 넘어가, 모르, 돌려보내
8	5.1%	조선, 일보, 탈북자, 영호, 의원, 공사, 국회, 국회의원, 기사, 바라, 당선, 대표, 국민, 쓰레기, 생각, 선영, 인간, 내용, 모르, 기자, 사과, 대한민국, 김정일, 돼지, 북한, 만들, 정일, 보이, 의심, 비난
9	4.9%	탈북자, 가족, 언론, 기자, 나오, 공개, 기사, 보도, 정은, 수용소, 통일부, 북한, 처형, 방송, 인간, 모양, 니들, 생각, 사람, 죽이, 모르, 그러, 송금, 취재, 이러, 보이, 회담, 조금, 그렇, 사실

10	4.9%	정권, 재인, 탈북자, 노무현, 김대중, 정은, 정책, 북한, 좌파, 탄압, 대북, 생각, 통일부, 대통령, 그러, 방송, 찬탈, 황장, 보내, 장관, 시절, 위원장, 보수, 심기, 심정, 사기꾼, 정부, 무너지, 대하, 공산당
11	4.7%	탈북자, 정부, 단체, 지원, 대북, 북한, 대하, 활동, 살포, 시민, 위하, 전단, 방법, 적극적, 날리, 정책, 차원, 보호, 중단, 안하, 대책, 민주화, 바라, 환영, 선거, 비용, 조사, 받아들이, 변화, 적극
12	3.9%	탈북자, 정권, 송환, 강제, 주사, 살인, 좌익, 촛불, 저지르, 범죄, 인간, 인권, 재인, 범죄자, 시위, 죽이, 요구, 만행, 지지, 애국자, 국제, 독재자, 수구, 보내, 사건, 이번, 행위, 대하, 정은, 찬성
13	3.2%	미국, 나라, 탈북자, 대통령, 망명, 한국, 아프, 가슴, 감사, 공무원, 트럼프, 마음, 고생, 민족, 서울, 사람, 신청, 외국, 친구, 동남아, 경찰, 걱정, 대사관, 일어나, 관계자, 불만, 부끄럽, 담당, 아들, 눈물
14	3.1%	조사, 의사, 귀순, 탈북자, 귀순자, 중국, 강제, 추방, 모르, 주장, 밝히, 잘못, 요구, 그렇, 내용, 학교, 북괴, 우리나라, 이번, 처리, 마약, 살인, 매매, 넘기, 어떻, 호소, 사건, 돌려보내, 방문, 도망

15	2.8%	정신, 탈북자, 차리, 계속, 나가, 인간, 지도자, 조심, 방법, 무섭, 문가, 빠지, 돼지, 힘내, 세력, 전략, 매국노, 폭정, 마음, 모르, 신문, 무료, 수도, 민족, 조명, 의심, 만나, 무너뜨리, 대한민국, 이제
16	2.7%	통일, 교육, 탈북자, 이북, 기사, 북쪽, 수구, 지역, 이용, 남북, 의도, 학생, 필요, 한반도, 세대, 남한, 일부, 선전, 직업, 나오, 바라, 사회, 남조선, 적화, 대부분, 문제, 이념, 사기, 평생, 벌어지
17	2.4%	탈북자, 남한, 국군, 포로, 출신, 여성, 국민, 개성, 전체, 뉴스, 공단, 지원금, 갈등, 인터뷰, 나오, 따지, 존경, 근무, 실향민, 물어보, 한명, 외교관, 직원, 시키, 이상, 그리하, 배우, 공부, 소리, 취업
18	1.4%	삐라, 좌파, 진보, 지상, 여론, 낙원, 영토, 뿌리, 부기, 사설, 관광, 전쟁, 미군, 살아남, 세력, 증오, 동상, 싫어하, 입장, 추종, 종부기, 도발, 정리, 오해, 오르, 타령, 운동가, 김일성, 조성, 보복

註: 토픽 개수가 18일 때, 로그(log)를 취한 혼란도는 -8.9435, 일관성 점수는 0.3602이다.

[표 6] 중앙일보 댓글 LDA 토픽 모델링 결과

	비중	토픽에 해당되는 상위 30개 단어
1	23.9%	탈북자, 북한, 한국, 사람, 중국, 간첩, 보내, 국가, 문제, 미국, 정부, 가족, 남한, 북괴, 그렇, 국정원, 정권, 지금, 대한민국, 대하, 행위, 김정일, 생각, 그러, 위장, 대북, 저러, 국민, 이용, 송환
2	12.6%	탈북자, 북한, 통일, 남한, 사람, 생각, 세력, 대하, 주민, 자유, 자신, 위하, 동포, 이해, 그렇, 단체, 이야기, 정권, 누리, 사회, 대한민국, 변절자, 어렵, 그리하, 날리, 공개, 모르, 망명, 주의, 의원
3	9.3%	인권, 탈북자, 북한, 자유, 정권, 증언, 좌파, 대한민국, 문제, 주장, 대통령, 말하, 위하, 대하, 주민, 언론, 방법, 사람, 정부, 보호, 국민, 노력, 인민, 일으키, 통일, 시작, 기자, 한국, 재인, 상황
4	5.9%	탈북자, 중앙, 일보, 수경, 출신, 기사, 노무현, 김대중, 서울, 수사, 공무원, 떨어지, 인간, 뭐하, 들이, 선동, 악마, 세금, 좌익, 국회의원, 과거, 혜택, 배신자, 나라, 시청, 공격, 상대, 이용, 미치

5	5.4%	탈북자, 사회, 지원, 정책, 정부, 얘기, 남한, 적응, 정도, 체제, 대하, 당국, 나쁘, 행동, 조국, 망하, 통일, 정착, 배우, 인도적, 국내, 인식, 연합, 인민, 교육, 비난, 이후, 경쟁, 사람, 줄이
6	5.3%	정은, 정권, 대한민국, 탈북자, 만들, 나라, 국민, 재인, 정신, 김일성, 박근, 북한, 개성, 정부, 이명, 차리, 국회의원, 공단, 당시, 생각, 판단, 출신, 대통령, 박수, 역사, 통제, 노릇, 당하, 소리, 어리
7	5.2%	탈북자, 인권, 민주, 방송, 들어가, 사건, 단체, 필요, 보이, 나오, 어떻, 변호사, 정치, 부모, 북한, 수구, 거짓, 조작, 정은, 라도, 조사, 사람, 수용소, 고통, 자식, 한마디, 하다, 위하, 지원, 욕하
8	5.2%	탈북자, 중국, 단체, 한국, 주민, 일부, 이상, 보내, 지금, 남쪽, 돌아가, 수준, 대사관, 국경, 정부, 사실, 지원, 떠나, 지역, 인구, 거주, 기관, 말하, 마음, 배신, 더럽, 정치, 내려오, 남북, 수가
9	4.8%	인간, 인권, 평화, 북한, 탈북자, 남북, 그러, 경제, 권력, 대부분, 재인, 반대, 대선, 들이, 사고, 체제, 전라도, 현실, 절대, 떠들, 국제, 극우, 촛불, 대통령, 국민, 무리, 유린, 남한, 과거, 단식

10	3.9%	나라, 국민, 대한민국, 사람, 탈북자, 우리나라, 세계, 존재, 일반, 마라, 보안법, 때리, 배려, 자유, 신경, 독재자, 준비, 자리, 감사, 차이, 선택, 근거, 모르, 이상, 이북, 전체, 대국, 기지, 무식, 그리
11	3.8%	탈북자, 문제, 집단, 나오, 민변, 국회, 식량, 의원, 대표, 모르, 니들, 천안, 현재, 처리, 개발, 재판, 이러, 의지, 당하, 갈등, 입장, 비례, 범죄자, 주민, 대량, 여행, 유전선, 후보, 조선, 이명
12	3.8%	목숨, 넘어오, 탈북자, 대한민국, 소리, 돌려보내, 주고, 챙기, 한번, 위하, 죽이, 통일부, 포격, 국민, 쓰레기, 북한, 이번, 도발, 알리, 나라, 지령, 경비, 난민, 장관, 부르, 통일, 몰르, 이제, 태도, 모르
13	3.8%	활동, 다르, 의사, 사실, 밝히, 귀순자, 자유, 진실, 국정원, 처단, 자격, 이전, 이해, 국정, 관련, 영국, 총선, 지급, 보수, 이적, 북한군, 탈북자, 증거, 거부, 재인, 검찰, 조치, 안보, 귀순, 정의
14	1.9%	도와주, 돼지, 조금, 모습, 민주화, 시위, 판문점, 목적, 일어, 추천, 넘어가, 세상, 수상, 고향, 무섭, 계속, 이산가족, 특수, 얼굴, 가족, 자의, 한민족, 이정, 방문, 똑같, 보라, 가하, 만나, 위험, 사퇴

15	1.8%	민주당, 이제, 지지, 해지, 안보, 재앙, 나라, 기회, 의원, 태어나, 자라, 짓거리, 변절, 뻔하, 적국, 트집, 정신, 식당, 이러, 행태, 점령, 자한, 역겹, 방향, 인기, 간부, 목숨걸, 몰이, 신경, 언사
16	1.5%	시민, 평양, 간첩, 정치인, 처형, 가짜, 망하, 적화, 피하, 일제, 달리, 한국행, 특별, 경찰, 공작원, 불르, 수고, 탈북자, 남파, 보고, 시도, 노출, 학습, 협박, 의하, 얼마, 기자, 생산성, 숭배, 중앙일보
17	1.1%	여성, 세상, 여자, 입국, 남자, 해외, 운운, 주기, 결혼, 최선, 갖추, 마라, 이주, 제외, 호소, 외모, 어린애, 엉망, 기자, 자신감, 후진국, 들먹이, 먹거리, 변명, 우리나라, 이상, 생각, 말하, 굴욕적, 자존
18	0.8%	관심, 붕괴, 해결, 과장, 미군, 가면, 싫어하, 어쩌, 동무, 경험, 흥분, 실정, 오만, 형국, 참상, 영향력, 미래, 자극적, 갖추, 목표, 설사, 재판, 극단적, 광란, 위원장, 조합, 흉악, 증오, 청중, 변호사

註: 토픽 개수가 18일 때, 로그(log)를 취한 혼란도는 -10.7711, 일관성 점수는 0.3049이다.

[표 7] 한겨레신문 댓글 LDA 토픽 모델링 결과

	비중	토픽에 해당되는 상위 30개 단어
1	17.4%	북한, 탈북자, 인권, 남한, 국민, 정부, 주민, 신문, 위하, 국가, 지금, 조사, 기사, 한겨레, 입장, 사회, 나오, 다르, 체제, 나라, 한국, 시키, 정권, 배신, 말하, 이상, 대하, 하다, 의미, 안되
2	15.6%	탈북자, 간첩, 국정원, 북한, 정권, 만들, 걸레, 이용, 자유, 언론, 주장, 사람, 위하, 조작, 이유, 그렇, 문제, 정보, 선거, 어떻, 사건, 세력, 가치, 중국, 집단, 민주주의, 탈출, 대한민국, 국민, 모르
3	8.9%	탈북자, 사람, 대한민국, 변절자, 사실, 북한, 그러, 자유, 수경, 대하, 정신, 누리, 좌파, 주의, 정은, 어렵, 인권, 한겨레, 김일성, 돌아가, 마음, 진보, 힘들, 안되, 거지, 이념, 취급, 설치, 생각, 도와주
4	6.5%	기사, 탈북자, 대통령, 노무현, 김정일, 이해, 공무원, 느끼, 대상, 서울시, 사회, 박근, 불법, 수준, 교육, 그러, 내가, 인사, 당하, 진정, 부정, 출연, 기본, 머리, 당시, 김대중, 인생, 보고, 감정, 터지

5	6.4%	탈북자, 사람, 나오, 죽이, 생각, 우리나라, 이야기, 보내, 들이, 정도, 무섭, 대한민국, 얘기, 도움, 경우, 모두, 문제, 차별, 대하, 국민, 돌리, 그리, 대부분, 종편, 만나, 이상, 지적, 자리, 여성, 받아들이
6	5.8%	탈북자, 위장, 국적, 간첩, 출신, 보내, 그리하, 북한, 의도, 평화, 가리, 남쪽, 모르, 마라, 데이, 불안, 국민, 정원, 이름, 반대, 미국노, 수사, 선택, 나쁘, 투쟁, 조선, 마시, 의심, 제기, 일보
7	5.2%	탈북자, 니들, 그렇, 나라, 증언, 범죄, 인정, 보이, 정은, 평양, 관계, 자유, 세습, 북괴, 도발, 노력, 북한, 거짓, 이제, 요즘, 탄압, 죽이, 이익, 사퇴, 따르, 색깔론, 데리, 대한민국, 양심, 자신
8	4.8%	가족, 탈북자, 민변, 그리, 정치적, 생활, 북한, 걱정, 위험, 존재, 잘못, 도망, 고향, 미국, 인신, 웃기, 문제, 부모, 하라, 범죄자, 지나, 구제, 망하, 대체, 종업원, 방문, 자신, 대하, 만들, 자가
9	4.4%	통일, 인간, 대하, 탈북자, 신분, 해보, 들이, 민주당, 바라, 자신, 생각, 북쪽, 수용소, 한마디, 공개, 남북, 부르, 의문, 말르, 사고, 위장, 나오, 무시, 왜곡, 얼마, 이해, 남한, 행복, 뉴스, 우익

10	4.1%	탈북자, 활동, 소리, 총살, 세금, 민주, 보이, 전체, 일으키, 넘어오, 말하, 저러, 정착금, 남한, 좋아하, 정치, 북한인, 목숨, 살리, 수많, 알바, 그러, 먹이, 처형, 채용, 아깝, 농담, 다하, 수령, 돌려보내
11	3.8%	수경, 의원, 빨갱이, 국회의원, 발언, 문제, 탈북자, 변절, 국회, 배신자, 사과, 거짓말, 김일성, 대한민국, 자라, 북한, 뭐하, 과정, 그렇, 나대, 당의, 잘못, 인식, 건지, 지가, 말대, 공안, 상태, 그럴듯하, 조국
12	3.8%	단체, 대북, 탈북자, 민주화, 운동, 정부, 전단, 행동, 뿌리, 대하, 보수, 지원, 시민, 북한, 청년, 연합, 집회, 목숨, 라도, 이해, 자신, 구하, 민주, 매국, 강화, 고치, 통합, 의무, 비판적, 회원
13	3.7%	중국, 한국, 이렇, 북한, 여자, 일하, 어쩌, 외국, 안보, 미치, 중국인, 마찬가지, 매매, 조선, 생각, 정부, 들어오, 나서, 조치, 한국인, 기대, 생기, 여성, 기사, 존중, 굶주리, 들오, 염려, 불만, 의견
14	3.6%	전쟁, 탈북자, 풍선, 남한, 추방, 짓거리, 조선족, 안하, 힘들, 꼴통, 입국, 대한민국, 인간, 들어오, 잘살, 사회, 입맛, 외국인, 배우, 생각, 환영, 유발, 학생, 거지, 맞추, 유지, 탈출, 대대, 이민, 잘하

15	2.2%	방송, 보호, 이번, 자극, 고통, 확인, 통하, 가족, 더럽, 신청, 매체, 주제, 보이, 하루, 통제, 못살, 한쪽, 세끼, 희생자, 일성, 요청, 그러, 아래, 차원, 필요, 배우, 구로, 원인, 말씀, 사건
16	2%	세상, 바꾸, 어떻, 교수, 지지, 라이트, 지역, 진정, 언급, 물어보, 국내, 좌익, 근거, 인물, 가르치, 개념, 잘못되, 시야, 중심, 재산, 노동자, 파렴치, 안하, 처지, 힘없, 드라마, 남도, 출신, 기자, 농민
17	1.5%	근본, 민족, 대가리, 선동, 동포, 분열, 그라, 의심, 돌려보내, 화해, 사상, 부칙, 당장, 황장, 하제, 가시, 놀리, 네트워크, 조오, 드러나, 진실, 붙어먹, 탈북자, 앞장서, 충격적, 독교, 조국, 투쟁, 연민, 동지회
18	0.4%	물불, 몸뚱이, 도처, 짓밟, 응분, 낮잠, 자우, 같애, 질서, 식용, 조타, 지소, 라인, 흩어지, 부익부, 비극적, 인건, 결단, 굴비, 바가지, 제보, 외교관, 판무관, 선생, 미로, 판단력, 악조건, 퍼먹, 잡아두, 대개

註: 토픽 개수가 18일 때, 로그(log)를 취한 혼란도는 -11.323, 일관성 점수는 0.356이다.

[표 8] 경향신문 댓글 LDA 토픽 모델링 결과

	비중	토픽에 해당되는 상위 30개 단어
1	17.4%	탈북자, 북한, 정부, 국정원, 위하, 그리하, 미국, 나라, 조작, 가족, 국민, 남한, 상황, 인권, 기사, 매매, 그리, 집단, 단체, 어렵, 웃기, 의사, 생각, 주장, 생활, 직원, 나오, 북풍, 나쁘, 요구
2	14.7%	탈북자, 북한, 인권, 주민, 생각, 빨갱이, 국정원, 이름, 대통령, 그러, 대한민국, 갈등, 사람, 말하, 신문, 배신, 문제, 이탈, 그리, 지금, 부대, 모르, 대하, 삐라, 남남, 자라, 악마, 존재, 정권, 한마디
3	9.9%	탈북자, 북한, 대한민국, 사람, 한국, 국민, 정권, 행위, 탈출, 이해, 그러, 비판, 돌아가, 기자, 대부분, 정도, 개선, 적응, 쓰레기, 자체, 남쪽, 힘들, 빨갱이, 대하, 기업, 시절, 사회, 넘어오, 독재, 북쪽
4	9.8%	사람, 나오, 모르, 탈북자, 보이, 태경, 집단, 일부, 사실, 정은, 사회, 그렇, 국회, 박근, 들어가, 협박, 거지, 진실, 생기, 김일성, 이유, 방송, 독재자, 남한, 다르, 자료, 이용, 자신, 심하, 결국

5	6.5%	탈북자, 시키, 남한, 자유, 추방, 국가, 국민, 생각, 전단지, 모르, 주의, 저러, 바라, 남북, 대북, 권력, 사회, 하나, 보내, 사실, 민족, 입장, 단어, 조장, 나라, 국정원, 허위, 잘살, 체제, 지금
6	6.4%	탈북자, 국민, 의원, 정치, 연합, 어버이, 언론, 소리, 정신, 세금, 지나, 미국, 경향, 민주주의, 가짜, 차리, 보수, 사상, 단체, 동원, 일보, 모습, 이상, 기사, 책임, 개소리, 부모, 깡패, 시키, 경찰
7	5.8%	탈북자, 나라, 조사, 안되, 인간, 어떻, 자유, 가족, 보내, 정권, 만들, 대한민국, 공개, 미국, 부정, 북한, 걱정, 정부, 방해, 시대, 다하, 집회, 기독교, 확인, 표현, 쓰레기, 성향, 집단, 논리, 목숨
8	5.5%	문제, 대하, 탈북자, 필요, 얘기, 하다, 하면, 활동, 대한민국, 해결, 사회, 보호, 민주, 사람, 하지, 대사관, 어떻, 특수, 모두, 나도, 국가, 국민, 몰르, 목숨, 한반도, 결국, 수준, 나이, 미래, 생각
9	5%	탈북자, 북한, 방송, 민주당, 세력, 그렇, 말하, 한국, 사상, 내가, 반대, 사건, 남북, 정책, 바라, 위하, 대통령, 도망, 못살, 마라, 경제, 국민, 이상, 책임, 희망, 대화, 범죄, 박상, 조선, 돌려보내

10	3.5%	단체, 보수, 탈북자, 진보, 그러, 서울, 평화, 시민, 받아들이, 자신, 머리, 조선, 예전, 기사, 지원, 아가, 전원, 통일, 주의자, 따지, 지방, 국민, 북한, 참석, 세뇌, 힘들, 공화국, 들러리, 버스, 자유당
11	3%	행동, 기관, 통하, 오해, 수입, 경우, 자인, 어리, 벗어나, 중국, 자가, 밝히, 전화, 무료, 진료, 정부, 단체, 관계, 대학, 가능, 병원, 방법, 개성, 상황, 어이, 내려오, 기다리, 서류, 공단, 장애인
12	2.7%	중국, 들이, 세상, 다니, 만들, 동포, 사고, 일하, 탈북자, 인식, 어쩌, 정착, 지원, 문제, 이용, 차별, 그렇, 체제, 난민, 북한, 인정, 돈벌이, 핍박, 넘어오, 덜하, 대하, 빠지, 게으르, 약화, 여론
13	2.7%	간첩, 탈북자, 국정원, 국가, 설치, 문제, 교육, 일으키, 중요, 사람, 시달리, 고민, 상태, 제정신, 정보기관, 이주, 섞이, 경기, 합심, 김진태, 정의, 예의, 경상도, 대중, 제주도, 교육대, 망상증, 기도, 신세, 전라도
14	2.4%	알바, 생명, 전쟁, 정도, 짓거리, 종자, 일삼, 탈북자, 이렇, 수령, 다음, 어버이, 문제, 자본주의, 드러내, 도망치, 정착금, 사례, 독일, 기반, 시행, 분단국가, 범위, 교육, 살아가, 관계, 더러운, 모란봉, 클럽, 전세계

15	1.9%	직업, 강제, 신분, 훈련, 마음, 당하, 도움, 생존, 업무, 부적응, 진심, 관련, 거부, 현실, 투사, 나가, 아부, 드럽, 힘들, 교포, 감수, 우기, 세계, 처하, 모르, 출국, 어려움, 심각, 드리, 일하
16	1.2%	이용, 관계, 포기, 반성, 관심, 중동, 뭐하, 놀리, 아시아, 정상적, 실현, 주지, 자국민, 미안, 병신, 대학교, 월북자, 가깝, 어제, 총선, 통일부, 사라지, 어린아이, 연락, 가난, 부가, 한의대, 보내, 모두, 수경
17	1.1%	우리나라, 죽이, 들어오, 잘하, 포격, 아프, 펼치, 속이, 동포, 목숨, 안타깝, 김정일, 조선족, 영화, 사람, 부끄럽, 살인, 달라, 꼬이, 간자, 범죄, 용도, 시작, 자위, 최선, 회유, 떨구, 변절자, 미친놈, 새끼
18	0.6%	위장, 떠나, 느낌, 이야기, 상식적, 마세, 재판, 미쿡, 귀순자, 보라, 새끼, 처형, 혼란, 풍파, 따먹, 시민권, 얼굴, 노역, 국적, 잡히, 뚫리, 정치범, 서울시, 공무원, 수용소, 술수, 남조선, 순천, 일본국, 침투

註: 토픽 개수가 18일 때, 혼란도는 -11.4147, 일관성 점수는 0.3383이다.

[표 9] 신문사들 간의 토픽 현저(topic salience) 비교: 조선일보 기사 LDA 토픽 모델링 결과

	비중	토픽에 해당되는 상위 30개 단어
1	12.8%	중국, 탈북자, 북한, 한국, 미국, 정부, 국경, 수용소, 강제, 인권, 난민, 사람, 일본, 체포, 여성, 대사관, 문제, 공안, 탈출, 보도, 당국, 입국, 가족, 최근, 단체, 정치범, 자유, 지역, 납치, 목사
2	11.7%	북한, 인권, 대북, 정부, 문제, 단체, 미국, 유엔, 국제, 자유, 주민, 지원, 전단, 남북, 탈북자, 정권, 상황, 사회, 살포, 대표, 개선, 제재, 보고서, 삐라, 위원회, 체제, 한국, 관련, 국가, 연합
3	9.5%	대통령, 의원, 문제, 정부, 미국, 회담, 장관, 한국, 대표, 국회, 정상, 위원장, 외교, 국민, 남북, 관계, 이날, 정책, 북한, 관련, 회의, 일본, 안보, 탈북자, 부시, 민주당, 국가, 누리, 후보, 기자
4	9.4%	통일, 지원, 사회, 사업, 지역, 재단, 문화, 과장, 기업, 센터, 복지, 서울, 교육, 주민, 회장, 활동, 나눔, 협력, 기획, 기부, 한국, 대표, 연구, 사회적, 정책, 경제, 운영, 팀장, 단체, 관리
5	8.9%	사람, 생각, 대한민국, 통일, 북한, 한국, 사회, 자유, 국민, 나라, 운동, 정권, 자신, 교회, 탈북자, 세계, 정치, 민족, 체제, 문제, 지금, 집회, 국가, 목사, 교수, 서울, 주의, 역사, 세력, 반대

6	8.3%	북한, 김정일, 정은, 정보, 노동당, 탈북자, 조선, 성택, 위원장, 부대, 정남, 보위, 비서, 사건, 출신, 국가, 가능성, 조직, 활동, 군사, 대남, 북한군, 김일성, 이후, 인민, 훈련, 암살, 소식통, 관련, 당시
7	8%	영화, 방송, 감독, 이야기, 공연, 기자, 배우, 드리마, 출연, 작품, 프로그램, 사람, 생각, 모습, 엄마, 조선, 한국, 사랑, 연기, 노래, 영화제, 오후, 제작, 탈북자, 자신, 가족, 사진, 다큐멘터리, 출신, 마음
8	7.4%	북한, 주민, 김정일, 평양, 정은, 소식통, 최근, 김일성, 당국, 경제, 간부, 시장, 남한, 방송, 지역, 식량, 보도, 이후, 탈북자, 내부, 매체, 체제, 개혁, 사람, 정도, 인민, 단속, 대북, 위원장, 통제
9	5.7%	학교, 교육, 학생, 대학, 학과, 탈북자, 전형, 지원, 청소년, 한국, 영어, 사회, 과정, 이탈, 입학, 모집, 선발, 사이버, 주민, 생활, 취업, 교사, 공부, 북한, 학년, 자녀, 경우, 서울, 프로그램, 출신
10	5.2%	혐의, 간첩, 탈북자, 수사, 경찰, 검찰, 사건, 조사, 북한, 서울, 국정원, 국가, 변호사, 마약, 구속, 재판, 검사, 법원, 사실, 기소, 선고, 중국, 국내, 증거, 정보, 진술, 형사, 부장, 위반, 범죄

11	4.7%	사람, 여성, 선수, 북한, 여자, 경기, 모습, 올림픽, 얼굴, 축구, 사진, 한국, 남자, 대표, 대회, 남성, 평양, 스포츠, 생각, 머리, 친구, 출신, 기사, 신이, 현미, 병원, 영란, 세계, 수술, 정도
12	3.4%	북한, 통일부, 전화, 정부, 주민, 조사, 사건, 귀순, 이날, 종업원, 오후, 확인, 휴대, 의사, 관계자, 국정원, 기자, 무진, 경찰, 국민, 사실, 보호, 선원, 관련, 오전, 발견, 주장, 모자, 추방, 공개
13	2.8%	가족, 북한, 남한, 아버지, 남북, 포로, 아들, 어머니, 국군, 사람, 고향, 생활, 전쟁, 남편, 결혼, 부모, 생각, 당시, 탈출, 아내, 할머니, 출신, 동독, 독일, 소식, 동생, 서독, 한국, 마음, 눈물
14	2.1%	회사, 아파트, 노동자, 공단, 개성, 식당, 사람, 야구, 음식, 공장, 무역, 북한, 기업, 은행, 근로자, 판매, 업체, 정도, 월급, 가구, 거래, 외화, 공급, 지원, 운영, 자금, 탈북자, 수입, 건물, 제품

註: 토픽 개수가 14일 때, 혼란도는 -8.1717, 일관성 점수는 0.5825이다.

[표 10] 신문사들 간의 토픽 현저(topic salience) 비교: 중앙일보 기사 LDA 토픽 모델링 결과

	비중	토픽에 해당되는 상위 30개 단어
1	17.4%	북한, 대북, 김정일, 정보, 정은, 미국, 미사일, 체제, 당국, 탈북자, 공개, 관련, 방송, 전단, 보도, 정부, 국가, 실험, 가능성, 단체, 중국, 위원장, 이후, 노동당, 기자, 경우, 발사, 한국, 군사, 국방
2	14.7%	북한, 중국, 인권, 문제, 탈북자, 미국, 한국, 정부, 관계, 유엔, 국제, 외교, 관련, 일본, 대통령, 강제, 대사, 난민, 국가, 회담, 상황, 대사관, 기자, 입장, 장관, 송환, 대북, 협력, 이번, 수용소
3	10.3%	사람, 한국, 영화, 생각, 탈북자, 사진, 기자, 가족, 여성, 감독, 이야기, 자신, 모습, 중국, 친구, 마음, 처음, 남한, 미국, 출신, 세계, 북한, 일본, 생활, 시작, 정도, 서울, 얘기, 아버지, 지금
4	10.1%	이원, 국민, 대통령, 대표, 후보, 정치, 문제, 국회, 정책, 경제, 정부, 생각, 자유, 국가, 민주당, 한국, 보수, 선거, 사람, 통합, 민주, 정권, 대한민국, 사회, 누리, 주장, 기자, 박근, 주의, 안보

5	9.5%	위원장, 대통령, 북한, 회담, 남북, 정은, 정상, 평양, 기자, 사진, 트럼프, 영상, 올림픽, 김정일, 선언, 중앙, 방문, 김일성, 판문점, 대화, 뉴스, 위원, 노동당, 북측, 추천, 청와대, 재인, 행사, 회의, 국무
6	6.3%	북한, 주민, 탈북자, 지원, 사회, 정부, 남한, 통일부, 남북, 정착, 이탈, 조사, 통일, 사람, 경우, 식량, 입국, 생활, 문제, 여성, 필요, 한국, 기자, 정책, 국내, 가족, 지난해, 이후, 결과, 이상
7	5.4%	탈북자, 경찰, 사건, 혐의, 국정원, 조사, 간첩, 중국, 검찰, 수사, 서울, 기자, 사실, 확인, 국가, 목사, 체포, 국내, 공안, 관계자, 증거, 재판, 입국, 구속, 위장, 활동, 당시, 보위, 관련, 여성
8	5.1%	활동, 사회, 봉사, 지원, 청소년, 지역, 서울, 회장, 단체, 행사, 다문화, 사진, 사업, 기자, 문화, 재단, 프로그램, 복지, 참여, 센터, 사랑, 대표, 나눔, 어린이, 올해, 아동, 가족, 기부, 장애인, 한국
9	5%	북한, 경제, 평양, 제재, 중국, 시장, 생산, 공장, 개성, 공단, 개발, 기술, 건설, 제품, 국가, 무역, 중앙, 조선, 거래, 최근, 맥주, 근로자, 수입, 대북, 사진, 공급, 정도, 판매, 수출, 지역

10	4.6%	교육, 대학, 학생, 학교, 학과, 전형, 선발, 사이버, 입학, 과정, 모집, 학부, 지원, 영어, 장학금, 졸업, 학년, 일반, 시험, 교사, 성적, 수업, 공부, 평가, 서울, 한국, 사회, 외국어, 과목, 이상
11	4.6%	기업, 주택, 지원, 금융, 은행, 사업, 예산, 복지, 사회적, 정부, 임대, 아파트, 가구, 공급, 소득, 대상, 특별, 이상, 분양, 업체, 청약, 계획, 서비스, 기초, 경우, 국민, 일자리, 보험, 고용, 추진
12	4.1%	통일, 교수, 한국, 북한, 연구, 평화, 박사, 독일, 남북, 한반도, 서울, 문화, 주제, 대학, 준비, 서울대, 과정, 학생, 원장, 의사, 공동, 교육, 오후, 연구원, 국제, 사회, 기자, 대학생, 연구소, 행사
13	2%	과장, 부장, 기획, 지원, 교육, 팀장, 한국, 관리, 정책, 사업, 원장, 실장, 협력, 과학, 행정, 승진, 전보, 문화재, 소장, 연구, 석기, 산업, 기술, 형사, 공무원, 위원회, 센터, 국장, 본부, 파견
14	1%	변호사, 변호, 사이버, 소송, 정보, 해킹, 금괴, 헌법, 동화사, 등록, 법률, 주민, 재산, 개인, 자료, 공격, 해커, 인터넷, 판결, 코드, 주소, 경우, 대구, 사법, 악성, 수집, 사례, 메일, 피해, 보안

註: 토픽 개수가 14일 때, 혼란도는 -8.3773, 일관성 점수는 0.5477이다.

[표 11] 신문사들 간의 토픽 현저(topic salience) 비교: 한겨레신문 기사 LDA 토픽 모델링 결과

	비중	토픽에 해당되는 상위 30개 단어
1	17%	문제, 국가, 사회, 미국, 북한, 정치, 대통령, 경제, 한국, 국민, 후보, 주의, 시민, 정권, 생각, 교수, 세력, 체제, 사람, 시장, 전쟁, 정책, 보수, 정부, 선거, 나라, 이후, 세계, 시대, 일본
2	13.2%	서울, 한국, 지원, 기획, 교육, 센터, 문화, 사업, 연구, 관리, 기술, 협력, 정책, 운영, 재단, 본부, 오후, 부장, 원장, 교수, 회장, 국제, 과장, 주제, 팀장, 지역, 과학, 경영, 정보, 산업
3	12.3%	사람, 생각, 한국, 이야기, 자신, 탈북자, 아버지, 사회, 남한, 북한, 마음, 활동, 얘기, 기자, 처음, 시작, 인터뷰, 어머니, 사진, 생활, 친구, 노동자, 서울, 정도, 내가, 여성, 세상, 가족, 지금, 모습
4	10.2%	북한, 남북, 대통령, 정부, 대북, 관계, 중국, 미국, 평화, 통일, 문제, 회담, 위원장, 대화, 정상, 한반도, 협력, 외교, 장관, 정책, 안보, 북쪽, 제재, 관련, 군사, 합의, 청와대, 상황, 이번, 공동

5	10.2%	북한, 조사관, 탈북자, 인권, 교화소, 중국, 전단, 군가, 정부, 살포, 주민, 국가, 국정원, 기자, 조사, 한국, 보호, 주장, 증언, 사실, 당국, 유엔, 입국, 관련, 공개, 위조, 문제, 자백, 대북, 지난해
6	7.9%	사건, 간첩, 검찰, 수사, 변호사, 조작, 증거, 혐의, 진술, 국가, 내용, 서울, 주장, 유우, 기록, 재판, 법원, 판결, 공무원, 기소, 제출, 검사, 형사, 서울시, 당시, 관련, 의혹, 무죄, 재판부, 선고
7	7.8%	북한, 물거품, 검색어, 손맛, 통일, 남한, 주민, 내용, 사람, 한겨레, 언론, 보도, 삐라, 탈북자, 방송, 기자, 남쪽, 정부, 김일성, 독일, 주장, 북한군, 북쪽, 사실, 당시, 작전, 이후, 교과서, 미군, 비판
8	6%	지원, 주민, 기업, 사회, 사회적, 사업, 복지, 이탈, 지역, 정부, 정착, 예산, 활동, 교육, 계층, 금융, 아동, 대상, 제도, 은행, 기관, 취업, 운영, 다문화, 지난해, 생활, 올해, 재단, 가정, 계획
9	5.4%	학교, 교육, 대학, 학생, 사이버, 전형, 학과, 입학, 모집, 선발, 이상, 교사, 지원, 학년, 자녀, 과정, 졸업, 학부, 대학교, 일반, 고교, 성적, 경우, 학년도, 전공, 교육청, 학력, 한국, 수업, 한양

10	2.9%	의원, 한국, 대표, 난민, 목사, 교회, 광주, 누리, 기자, 운동, 국회, 미국, 외국인, 기독교, 영국, 일본, 민주당, 원내, 출신, 당시, 동포, 민주화, 국회의원, 이주, 소설, 공사, 차별, 위원회, 가짜, 소수자
11	2.3%	국정원, 조사, 가족, 기사, 조선, 남한, 북한, 정보, 보도, 신문, 보위, 사실, 센터, 합동, 자신, 기자, 유출, 한겨레, 확인, 과정, 통화, 사과, 국방부, 생각, 아들, 피해, 동안, 국가, 상황, 시작
12	2.1%	단체, 경찰, 연합, 집회, 어버이, 기자, 회원, 보수, 서울, 운동, 탈북자, 대표, 세월, 청와대, 전경련, 시민, 오전, 이날, 경찰서, 보안, 박근, 사퇴, 행사, 의혹, 경기도, 자유, 계좌, 목사, 광장, 한겨레
13	1.7%	영화, 방송, 문화, 감독, 드라마, 공연, 작품, 이야기, 종편, 작가, 프로그램, 배우, 영화제, 연기, 출연, 시사, 촬영, 티브이, 채널, 노래, 제공, 연극, 달력, 콘서트, 관객, 장면, 피디, 여성, 사랑, 연출
14	1%	아파트, 임대, 주택, 가구, 단지, 공급, 공진, 공사, 기초, 병원, 영구, 주민, 서울시, 알바, 의료, 입주, 도시, 분양, 장애인, 수급자, 대출, 특별, 최저, 국민, 생활, 소득, 유공자, 부모, 이하, 수급

註: 토픽 개수가 14일 때, 혼란도는 -8.7255, 일관성 점수는 0.5946이다.

[표 12] 신문사들 간의 토픽 현저(topic salience) 비교: 경향신문 기사 LDA 토픽 모델링 결과

	비중	토픽에 해당되는 상위 30개 단어
1	19.3%	북한, 인권, 문제, 중국, 탈북자, 정부, 미국, 관계, 한국, 남북, 회담, 대북, 대통령, 국제, 정책, 유엔, 상황, 관련, 지원, 외교, 국가, 난민, 입장, 이번, 위원회, 장관, 해결, 대화, 강제, 개선
2	11.4%	문제, 영화, 사람, 보수, 생각, 사회, 국민, 국가, 진보, 정부, 운동, 정치, 경제, 한국, 시민, 대통령, 교수, 정권, 세력, 주의, 정책, 시장, 시대, 나라, 감독, 이념, 지금, 진영, 기자, 변화
3	10.5%	한국, 여성, 사람, 북한, 탈북자, 사회, 남한, 가족, 노동자, 생각, 결혼, 생활, 이주, 자신, 주민, 출신, 외국인, 차별, 정착, 하다, 중국, 친구, 어머니, 정도, 엄마, 일본, 과정, 문제, 조사, 한국인
4	8.1%	북한, 탈북자, 중국, 미국, 한국, 보도, 당국, 조선, 방송, 기자, 김정일, 정은, 국경, 일본, 사실, 라오스, 공안, 주장, 체제, 신문, 확인, 체포, 입국, 대사관, 인민, 정보, 활동, 정부, 통신, 단체

5	7.7%	의원, 대표, 국회, 대통령, 누리, 후보, 위원장, 민주당, 발언, 한나라, 통합, 출신, 인사, 위원, 국민, 기자, 정치, 위원회, 민주, 박근, 대선, 원내, 국회의원, 장관, 관련, 논란, 회의, 자유, 정당, 이날
6	7.6%	지원, 사회, 사업, 복지, 지역, 활동, 교육, 의료, 다문화, 기업, 계층, 건강, 제공, 문화, 봉사, 재단, 센터, 사회적, 주민, 프로그램, 병원, 참여, 대상, 장애인, 소외, 보건, 노인, 운영, 서울, 나눔
7	7.1%	통일, 남북, 평화, 학교, 북한, 공연, 소설, 문화, 작가, 작품, 한국, 이야기, 세계, 서울, 문학, 전쟁, 올해, 평양, 행사, 모습, 시작, 도시, 탈북자, 분단, 예산, 사진, 선수, 환경, 예술, 노래
8	5.5%	국정원, 사건, 정보, 변호사, 조사, 북한, 탈북자, 국가, 간첩, 사실, 센터, 공개, 법원, 주장, 입국, 국내, 가족, 신문, 확인, 소송, 보호, 합동, 공작원, 종업원, 관련, 조작, 판결, 정부, 정보원, 보위
9	4.7%	검찰, 혐의, 수사, 탈북자, 사건, 중국, 경찰, 간첩, 증거, 기소, 조사, 국정원, 구속, 국가, 재판, 서울, 선고, 위반, 조작, 북한, 위조, 검사, 진술, 총영사관, 제출, 공무원, 재판부, 기록, 영사, 공안

10	4.6%	주민, 이탈, 지원, 북한, 주택, 정착, 이상, 생활, 임대, 국민, 금융, 경우, 기초, 가구, 대출, 조사, 은행, 아파트, 상품, 정부, 소득, 기준, 대상, 경찰, 공사, 가입, 기간, 보호, 제도, 서민
11	4.4%	과장, 부장, 기획, 지원, 팀장, 정책, 본부, 관리, 실장, 본부장, 교육, 협력, 사업, 센터, 남북, 전보, 승진, 연구, 원장, 안전, 회담, 운영, 국장, 정보, 기술, 감사, 통일, 소장, 과학, 교류
12	4.1%	단체, 연합, 전단, 살포, 대북, 어버이, 삐라, 북한, 대표, 운동, 탈북자, 자유, 목사, 집회, 회원, 행사, 이날, 교회, 기자, 주민, 임진각, 서울, 오전, 보수, 지역, 풍선, 오후, 경찰, 국방부, 주장
13	3.4%	학교, 교육, 대학, 학생, 전형, 학과, 사이버, 모집, 입학, 선발, 지원, 과정, 공단, 학부, 국제, 학습, 시험, 일반, 성적, 졸업, 대학교, 이상, 수업, 학년, 자녀, 장학금, 서울, 고교, 반영, 대학원
14	1.7%	아동, 남북, 대표, 가족, 교황, 탈북자, 피해자, 위원회, 청소년, 무역, 종교, 한국, 한화, 입양, 통일부, 포함, 서울, 정부, 학대, 투자, 어선, 전화, 하나, 중국, 보호, 피해, 이용, 범죄, 법률, 한성

註: 토픽 개수가 14일 때, 혼란도는 -8.3512, 일관성 점수는 0.5935이다.

참 고 문 헌

1. 단행본

남길임・조은경. 2017.『한국어 텍스트 감성분석』.

백영민. 2017.『R을 이용한 텍스트 마이닝』.

서대호. 2019.『잡아라! 텍스트 마이닝 with 파이썬』.

성동규・양승찬 연구책임 한국언론진흥재단 조사분석팀 연구진행. 2010.『여론과 미디어』.

윤인진 연구책임 북한인권정보센터 연구수행. 2016.『북한이탈주민에 대한 국민인식 및 차별실태조사』.

윤태일・이수안. 2018.『파이썬으로 텍스트 분석하기』.

이기창. 2019.『한국어 임베딩: 자연어 처리 모델의 성능을 높이는 핵심 비결 Word2Vec에서 ELMo, BERT까지』.

줄리아 실기・데이비드 로빈슨 지음. 박진수 옮김. 2019.『R로 배우는 텍스트마이닝』.

Iyengar, S. and D. R. Kinder. 2010. *News that matters: Agenda-setting and priming in a television age*, Chicago: University of Chicago Press.

Krosnick, J. A., C. M. Judd, and B. Wittenbrink. 2018. "The Measurement of Attitudes," in Albarracin D. and B. T. Johnson (eds.), *The Handbook of Attitudes, Volume 1: Basic Principles*. Routledge.

2. 학술지

감미아・송민. 2012. "텍스트 마이닝을 활용한 신문사에 따른 내용 및 논조 차이점 분석."『지능정보연구』, Vol. 18, No. 3.

강범일・송민・조화순. 2013. "토픽 모델링을 이용한 신문 자료의 오피니언 마이닝에 대한 연구."『한국문헌정보학회지』, Vol. 47, No. 4.

강범일・이재윤. 2014. “트위터 관련 연구에 대한 계량정보학적 분석.” 『정보관리학회지』, Vol. 31, No. 3.

강성원・김도연・최희선・한국진・진대용. 2018. “기후환경 이슈 분석을 위한 텍스트 마이닝 활용방안 연구.”『기후환경정책연구』.

강진구・이기성. 2019. “텍스트마이닝(Text Mining)을 통해 본 제주 예멘 난민: 네이버 뉴스 댓글을 중심으로.”『다문화콘텐츠연구』, Vol. 30.

곽수정・김현희. 2019. “텍스트 마이닝과 토픽 모델링을 기반으로 한 트위터에 나타난 사회적 이슈와 키워드 및 주제 분석.”『정보처리학회논문지 소프트웨어 및 데이터 공학』, Vol. 8, No. 1.

곽정래・이준웅. 2009. “김대중・노무현 정부시기 탈북자 문제에 관한 언론의 프레임 유형 연구.”『한국언론학보』, Vol. 53, No. 6.

권수현. 2011. “북한이탈주민에 대한 남한국민의 태도.”『한국정치연구』, Vol. 20, No. 2.

권수현・송영훈. 2015. “피를 나눈 형제, 이웃보다 못한 친척? 북한이탈주민에 대한 남한주민의 사회적 거리감과 결정요인, 2007-2014.” 『OUGHTOPIA』, Vol. 30, No. 2.

권호천. 2017. “박근혜 정부시기 탈북민 관련 신문보도의 의미구조 차이 비교 분석: 조선일보와 한겨레신문 기사의 텍스트 네트워크 분석을 중심으로.”『사회과학연구』, Vol. 29, No. 2.

길호현. 2018. “텍스트마이닝을 위한 한국어 불용어 목록 연구.”『우리말글』, Vol. 78.

김가영・강경필・손민지・이청아・홍신범・김성태. 2020. “북한 뉴스 담론에 대한 미디어 의제설정효과 연구: 빅데이터 토픽모델링과 워드임베딩 분석을 중심으로.”『평화연구』, Vol. 28, No. 1.

김경희・노기영. 2011. “한국 신문사의 이념과 북한 보도방식에 대한 연구.”『한국언론학보, Vol. 55, No. 1.

김문형・장하연・조유미・신효필. 2013. “KOSAC(Korean Sentiment Analysis Corpus): 한국어 감정 및 의견 분석 코퍼스.”『한국정보

과학회 학술발표논문집』.
김성근・조혁준・강주영. 2016. “학술연구에서의 텍스트 마이닝 활용 현황 및 주요분석기법.” 『정보화연구(구 정보기술아키텍처연구)』, Vol. 13, No. 2.
김용회・한창근. 2019. “‘수저계급’ 관련 웹 뉴스 기사에 대한 의미연결망 분석.” 『한국사회복지학』, Vol. 71, No. 3.
김은미・조윤용・임영호・송보영. 2015. “다문화 범죄 보도에서 기사 구성 방식과 출신국에 대한 태도가 댓글에 미치는 영향.” 『한국언론학보』, Vol. 59, No. 6.
김찬우. 2018. “온라인 기반 낙태 이슈 변화 분석: 온라인 뉴스와 댓글을 중심으로.” 『예술인문사회 융합 멀티미디어 논문지』, Vol. 8, No. 10.
김혜미・이준웅. 2011. “인터넷 뉴스와 댓글의 뉴스 프레임 융합 효과 연구: 해석의 복잡성 및 태도의 극단성 분석을 중심으로.” 『한국언론학보』, Vol. 55, No. 2.
김희경. 2016. “북한이탈주민과 남한주민의 교류 질적 연구.” 『한국심리학회지: 상담 및 심리치료』, Vol. 28, No. 2.
노법래・양경은. 2019. “한국 사회 저출산 논의 구조와 그 변환에 관한 텍스트 마이닝 분석 - 2000년 이후 언론 기사문에 대한 토픽 모델링과 토픽 구조의 시계열적 변동을 중심으로.” 『한국사회복지학』, Vol. 71, No. 4.
박상민・나철원・최민성・이다희・온병원. 2018. “Bi-LSTM 기반의 한국어 감성사전 구축 방안.” 『지능정보연구』, Vol. 24, No. 4.
박종희. 2016. “세월호 참사 1년 동안의 언론보도를 통해 드러난 언론매체의 정치적 경도.” 『한국정치학회』, Vol. 50, No. 1.
박종희・박은정・조동준. 2015. “북한 신년사(1946-2015)에 대한 자동화된 텍스트 분석.” 『한국정치학회보』, Vol. 49, No. 2.
배정환・손지은・송민. 2013. “텍스트 마이닝을 이용한 2012년 한국대선 관련 트위터 분석.” 『지능정보연구』, Vol. 19, No. 3.

서대호・한지은. 2018. “텍스트 마이닝 기법 기반 국내 홈 퍼니싱 브랜드의 소비자 인식 및 동향 비교・분석: IKEA와 한샘을 중심으로.”『디자인융복합연구(구.인포디자인이슈)』, Vol. 17, No. 6.

서상현・김준태. 2016. “딥러닝 기반 감성분석 연구동향.”『한국멀티미디어학회지』, Vol. 20, No. 3.

손애리・이내영. 2012. “탈북자에 대한 한국인의 태도 연구: 국가정체성과 다문화수용성을 중심으로.”『아태연구』, Vol. 19, No. 3.

손영준. 2004. “미디어 이용이 보수 진보적 의견에 미치는 영향.”『한국언론학보』, Vol. 48, No. 2.

송준모・강정한. 2018. “메갈리아의 두 딸들: 익명성 수준에 따른 온라인 커뮤니티의 정체성 분화.”『한국사회학』, Vol. 52, No. 4.

송준모・박영득. 2018. “청와대 국민청원에서는 무엇이 일어나는가?: 자연어 처리를 활용한 청와대 국민청원 분석.”『한국정치학회보』, Vol. 53, No. 5.

안정국・김희웅. 2015. “집단지성을 이용한 한글 감성어 사전 구축.”『지능정보연구』, Vol. 21, No. 2.

윤인진・송영호. 2013. “북한이주민에 대한 남한주민의 민족의식과 다문화의식.”『재외한인연구』, Vol. 30, (2013.6).

이서영・권상집. 2019. “트위터 메시지 분석을 통한 선거 결과 예측 고찰: 18대 대선을 중심으로.”『한국콘텐츠학회논문지』, Vol. 19, No. 4.

이수연・김현정・정수연. 2016. “텍스트마이닝 기반 토픽 분석을 통한 온라인 성차별성의 이해.”『사이버커뮤니케이션학보』, Vol. 33, No. 3.

이준웅. 2000. “프레임, 해석 그리고 커뮤니케이션 효과.”『언론과 사회』, Vol. 29.

_____. 2004. “언론 매체 이용 및 해석적 틀이 통일 및 대북 정책에 대한 의견에 미치는 효과.”『한국언론학보』, Vol. 48, No. 1.

이재신・김지은・류재미・강재혁. 2010. “기사 프레임과 장르가 댓글 유

형에 미치는 영향.”『한국언론학보』, Vol. 54, No. 2.
이찬주・임종섭. 2018. “언론 프레임의 온라인 문화적 공명이 여론 틀 짓기에 미치는 영향 관계: 김영란법 관련 기사와 댓글에 대한 텍스트 마이닝 및 그랜저 인과 관계 분석을 중심으로.”『한국언론학보』, Vol. 62, No. 4.
이혜진・최자영. 2019. “텍스트 마이닝 기법을 활용한 편의점 이용 고객의 트위터 리뷰에 대한 감성분석.”『글로벌경영학회지』, Vol. 16, No. 4.
전주람・신윤정. 2018. “남한출신 복지관련 종사자는 탈북 동료를 어떻게 바라보는가?”『한국사회복지질적연구』, Vol. 12, No. 2.
정진원. 2019. “북한이탈주민에 대한 한국인의 태도 결정요인.”『사회과학연구』, Vol. 30, No. 1.
좌보경・백혜진・서필교. 2014. “금연정책관련 온라인 뉴스와 댓글 유형의 내용 분석.”『홍보학연구』, Vol. 18, No. 3.
진설아・허고은・정유경・송민. 2013. “트위터 데이터를 이용한 네트워크 기반 토픽 변화 추적 연구.”『정보관리학회지』, Vol. 30, No. 1.
차준범・성정웅・김정근・박상현. 2016. “트위터 데이터에 기반한 헬-조선 키워드 분석.”『한국멀티미디어학회』, Vol. 19, No. 2.
채희찬・이종욱・최윤아・박대희・정용화. 2018. “트윗 텍스트 마이닝 기법을 이용한 구제역의 감성분석.”『정보처리학회논문지 소프트웨어 및 데이터 공학』, Vol. 7, No. 11.
최윤형・김수연. 2013. ““대한민국은 우릴 받아줬지만, 한국인들은 탈북자를 받아준 적이 없어요.” 댓글에 나타난 남한사람들의 탈북자에 대한 인식과 공공 PR의 과제.”『한국광고홍보학보』, Vol. 15, No. 3.
황정미. 2016. “사회적 위협 인식과 북한이탈주민에 대한 사회적 거리.”『아태연구』, Vol. 23, No. 2.
Blei, D. M., A. Y. Ng and M. I. Jordan. 2003. “Latent Dirichelet

Allocation." *Journal of Machine Learning*, Vol. 3.

Chang, J., J. Boyd-Graber, S. Gerrish, C. Wang and D. M. Blei. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." *Advances in neural information processing systems*.

Hearst, M. A. 1999. "Untangling text data mining." *Proceedings of the 37th Annual meeting of the Association for Computational Linguistics*.

Liu, B. 2012. "Sentiment analysis and opinion mining." *Synthesis lectures on human language technologies*, Vol. 5, No. 1.

Maaten, L. van der and G. Hinton. 2008. "Visualizing Data using t-SNE." *Journal of Machine Learning Research*, Vol. 9.

McCombs, M. E. and D. L. Shaw. 1972. "The agenda-setting function of mass media." *Public opinion quarterly*, Vol. 36, No. 2.

Mikolov, T., K. Chen, G. Corrado and J. Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." https://arxiv.org/abs/1301.3781.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado and J. Dean. 2013. "Distributed representations of words and phrases and their compositionality." *In Advances in neural information processing systems*.

Röder, M., A. Both and A. Hinneburg. 2015. "Exploring the Space of Topic Coherence Measures." *Proceedings of the eighth ACM international conference on Web search and data mining*.

Scheufele, D. A. and D. Tewksbury. 2007. "Framing, agenda setting, and priming: The evolution of three media effects models." *Journal of communications*, Vol. 57, No. 1.

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

Stavrianou, A., P. Andritsos and N. Nicoloyannis. 2007. “Overview and semantic issues of text mining.” *ACM Sigmod Record*, Vol. 36, No. 3.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll and M. Stede. 2011. “Lexicon-Based Methods for Sentiment Analysis.” *Computational linguistics*, Vol. 37, No. 2.

Tan, A. H. 1999. “Text mining: The state of the art and the challenges.” *Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases*, Vol. 8.

3. 잡지 및 신문

고승연. 2018. “SR3. 송민 연세대 교수 인터뷰 ‘감정 분석 뜬다는데 해볼까’는 위험 어떤 문제에 왜 필요한지 정의 먼저.” 『동아비즈니스리뷰』, (11월 Issue 2).

이준웅. 2001.10. “의제설정 이론.” 『신문과 방송』.

_____. 2001.11. “점화효과 이론.” 『신문과 방송』.

_____. 2001.12. “틀짓기효과.” 『신문과 방송』.

『연합뉴스』, 2012년 9월 6일

『자유아시아방송』, 2016년 1월 15일

『조선일보』, 2020년 4월 14일

『한겨레신문』, 2019년 10월 29일

4. 웹페이지

https://www.korean.go.kr/

http://tongil.snu.ac.kr/

http://kossda.snu.ac.kr

http://www.asiaticresearch.org/

http://office.kbs.co.kr/tongil/

https://www.chosun.com

https://joongang.joins.com
http://www.hani.co.kr
http://www.khan.co.kr
http://www.m.khan.co.kr
https://github.com/suminb/hanja
https://github.com/lovit/soynlp
https://github.com/lovit/soyspacing/blob/master/tutorials/presentation.pdf
https://github.com/lovit/customized_konlpy
https://github.com/twitter/twitter-korean-text
https://github.com/open-korean-text/open-korean-text
https://github.com/ssut/py-hanspell
https://github.com/haven-jeon/PyKoSpacing
https://konlpy-ko.readthedocs.io/ko/v0.4.3/
https://docs.google.com/spreadsheets/d/1OGAjUvalBuX-oZvZ_-9tEfYD2gQe7hTGsgUpiiBSXI8/edit#gid=0
https://radimrehurek.com/gensim/models/ldamodel.html
https://groups.google.com/g/gensim/c/uQxQiR2oC98
https://radimrehurek.com/gensim/models/word2vec.html
https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html
http://openhangul.com/
https://github.com/park1200656/KnuSentiLex

Abstract

Differentiation of Comments on Articles related to North Korean Defectors

Jain Choi
Dept. of Political Science & International Relations
Master's Degree
The Graduate School
Seoul National University

Can a newspaper's ideological differentiation lead to the reader's ideological differentiation? Does the reader's attitude towards the same object and event appear differently depending on which newspaper they use? Based on the consideration of the framing effect of the press, this study aims to determine whether there is a difference in the content of articles provided by the conservative and progressive newspapers, and whether such differences also appear in the content of comments on the articles. It started from that. In order to confirm the contents, specific cases such as articles and comments on North Korean defectors from the Chosun Ilbo, JoongAng Ilbo, Hankyoreh Shinmun, and Kyunghyang Shinmun were selected. And to deal with a

vast amount of article and comment data, topic analysis and related word analysis were conducted.

As a result, the articles provided by the conservative-oriented and the progressive-oriented newspapers showed some commonalities and differences, and this pattern was very similar in the comments of the conservative-oriented newspaper and the progressive-oriented newspaper. Accordingly, in terms of correlation, this study shows that newspaper companies can induce the reader's reaction in a certain direction by presenting the content of the article containing a specific point of view to the reader, or that the characteristics of the readers participating in each newspaper company are different.

keywords : North Korean Defectors, Articles, Comments, Attitude, Differentiation, Text mining
Student Number : 2017-24750