트위터 데이터에 기반한 헬-조선 키워드 분석

차준범*, 성정웅*, 김정근*, 박상현*+
*연세대학교 컴퓨터과학과
+교신저자
e-mail : sanghyun@yonsei.ac.kr

Hell-Chosun Keyword Analysis based on Twitter

Jun-Bum Cha*, Jung-Woong Sung*, Jung-Geun Kim*, Sang-Hyun Park*+
*Dept of Computer Science, Yonsei University
+Corresponding author

요 약

2014 년경 등장한 신조어인 “헬조선”은 얼마 지나지 않아 사용 빈도가 급격하게 증가하여 오프라인 매체에서도 자연스럽게 사용되는 등 완연한 신조어로 자리잡았다. 본 연구에서는 헬조선이라는 키워드를 분석하는 새로운 방법을 제안하고, 실제 트위터 데이터를 기반으로 이 키워드의 사회적 맥락에 대해 알아본다.

1. 서론

헬조선이란 지옥을 뜻하는 헬(hell)과 한국을 뜻하는 조선의 합성어로 "지옥 같은 한국" 이라는 의미를 갖는다. 처음 단어가 등장한 것은 2014 년경으로 2 년이라는 짧은 시간밖에 지나지 않았으나, 갈수록 심화되는 사회 및 경제 문제 때문에 한국에서의 삶에 지친 사람들이 이 표현에 공감하면서 빠르게 확산되었다. 2016 년에는 이미 하나의 신조어로 자리잡아 여러 오프라인 매체에서도 자연스럽게 사용되고 있다.

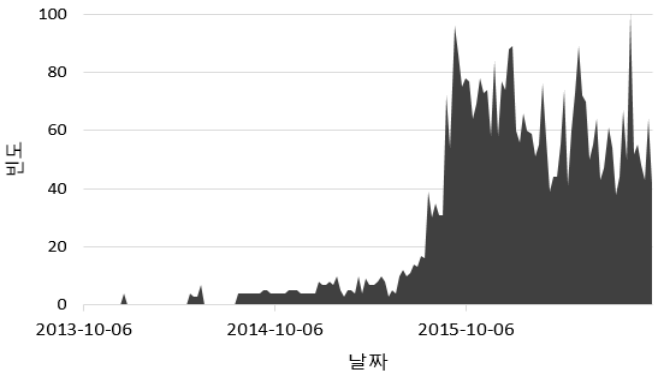

그림 1. "헬조선" 에 대한 구글 트렌드 (2013. 10. 2 ~ 2016. 10. 2)

그림 1 에 따르면 2014 년에 헬조선이라는 단어가 처음으로 등장하였고, 2015 년에 이 단어의 사용빈도가 폭발적으로 증가하여 2016 년까지 그 추세가 유지되고 있다. 본 연구에서는 이와 같은 신조어가 가장 빠르게 확산되는 트위터(twitter)의 트윗(tweet)을 기반으로 헬조선이라는 단어와 다른 단어와의 관계를 알아보고, 이를 중심으로 헬조선이라는 단어의 사회적 맥락을 짚어본다. 나아가 이러한 맥락들이 기존의 연구와 얼마나 일치하고 어떤 차이가 있는지 살펴본다.

기존에도 이러한 연구들이 제안되었는데, 사회학적으로 헬조선이라는 신조어를 분석하고 맥락을 파악하고자 하는 연구들이 있다[1,2]. 그러나 정말로 헬조선이라는 키워드를 사용하는 사람들의 실제 텍스트를 과학적으로 분석하는 연구는 없었다. 따라서 본 연구는 1) 사회적으로만 분석되던 헬조선이라는 키워드를 과학적으로 분석하였고, 2) 분석 결과와 기존의 연구 결과가 얼마나 합치되는지를 확인하여 분석 방법을 검증하였으며, 3) 기존의 연구에서는 찾아볼 수 없던 새로운 맥락을 제안하였다는 세 가지 의의를 갖는다.

2. 실험 방법

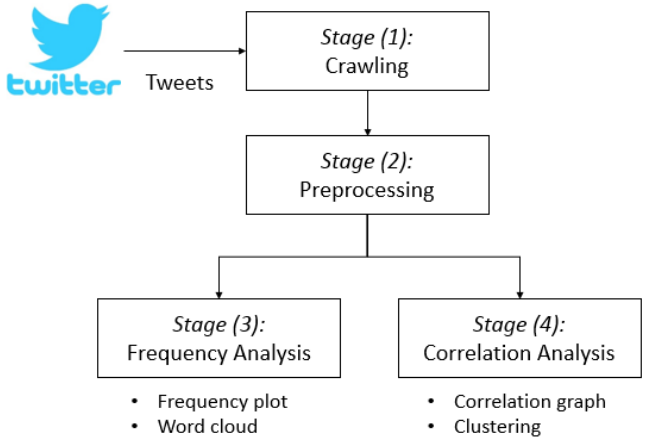

그림 2. 실험 방법 개요

그림 2 에서 볼 수 있는 바와 같이, 본 연구에서 사용한 방법은 총 4 단계로 구성된다. 트윗들을 크롤링(crawling)하여 데이터를 모으고, 이를 전처리(preprocessing)를 통해 정제한 뒤, 이렇게 정제된 데이터를 분석한다.

2.1. Crawling

트위터에서 제공하는 API 를 사용하여 정해진 "헬조선"에 대한 트윗들을 받아온다. 헬조선과 동의어로 "헬조센"이라는 단어도 사용되므로 두 단어에 대해 모두 크롤링하여 중복되는 트윗은 제거하였다. 트위터의 트윗 크롤링 API 는 최대 일주일간의 트윗만 제공하기 때문에, 본 연구에서는 많은 데이터를 모으기 위해 반복적으로 크롤링을 수행하였다.

2.2. Preprocessing

먼저, 트윗 중 광고나 리트윗(retweet) 등 의미 없는 트윗들을 걸러낸다. 이후, 각 트윗을 문장 단위로 구분하고 트위터 한국어 처리기[3]를 사용하여 문장을 분석한다. 유의미한 단어를 사용하기 위해서 다양한 형태소 중 명사만을 추출한다. 형태소 분석기는 트위터 한국어 처리기 위에도 은전한닢[4] 이나 꼬꼬마[5] 형태소 분석기 등 다양한 형태소 분석기가 있으나 트위터 데이터를 분석하는 작업이기 때문에 트위터에서 개발한 트위터 한국어 처리기를 채택하였다. 트위터 한국어 처리기는 다른 형태소 분석기에 비해 인터넷 용어가 많은 트윗 텍스트를 분석하는 데에 강점이 있으며, 한국어 정규화(normalization) 및 어근화(stemming)를 지원한다. 마지막으로 무의미한 불용어(stop word)를 제거하여 유의미한 단어만을 남긴다.

2.3. Frequency Analysis

이렇게 정제한 데이터를 분석하여 단어 빈도 플롯(word frequency plot)을 그린다. 단어 빈도 플롯은 상위 N 개의 빈발 단어를 보여주며, 각 단어의 등장 빈도를 시각적으로 확인할 수 있다. 나아가, 이 데이터를 기반으로 워드 클라우드(word cloud)를 그려 각 단어의 등장 빈도 차이를 한눈에 확인할 수 있다.

2.4. Correlation Analysis

단어 간의 연관성을 알아보기 위해, 단어 간 공기 빈도(co-occurrence frequency)를 계산하여 그래프를 생성한다. 단어 간 연관성을 구체적으로 파악하기 위해, 공기 그래프에서 Markov Clustering (MCL) 알고리즘[6]을 통해 군집화(clustering)를 수행한다. 이 때, 모든 데이터는 "헬조선" 또는 "헬조센" 이라는 키워드와 같이 등장하므로 이 두 단어는 그래프에서 제외하여 같이 등장하는 단어 간의 관계를 알아본다.

3. 실험 결과

크롤링 및 전처리, 데이터 분석에는 python 을 사용하였으며 클러스터링에는 Gephi[7]를 사용하였다. 데이터 분석을 위해 NLTK[8], koNLPy[9], NetworkX[10], matplotlib, numpy 패키지를, 워드 클라우드를 그리기 위해 pytagcloud 패키지를 사용하였다. 전반적인 데이터 분석은 python 3 기반으로 수행되었으나 pytagcloud 패키지의 경우 python 2 만을 지원하기 때문에 워드 클라우드는 python 2 에서 생성하였다.

3.1. 사용한 데이터

표 1. 데이터 정보

크롤링 기간	2016/4/30 ~ 2016/5/11
사용한 키워드	헬조선, 헬조센
트윗 수	2,124
토큰 수	24,452

표 1 은 사용한 데이터 정보를 보여준다. 광고 트윗이나 무의미한 리트윗, 중복된 내용의 트윗 등을 제거하고 유의미한 트윗만 2,124 개를 사용하여 분석하였으며 이 트윗들은 서로 다른 24,452 개의 단어로 구성되었다.

3.2. Frequency Analysis

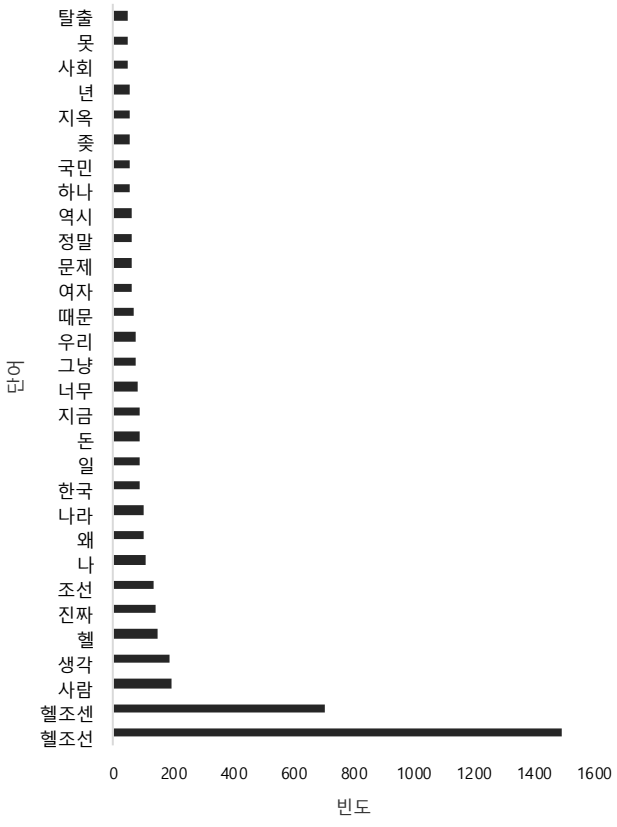

그림 3. 단어 등장 빈도 (상위 30 개)

그림 3 은 트윗에서 언급된 상위 빈발 단어 30 개를 나타낸다. 실제 실험에서는 최대 상위 100 개까지의 단어를 분석하였다.

그림 4. 워드 클라우드

단어의 빈도수를 기반으로 생성한 워드 클라우드를 그림 4 에서 볼 수 있다. 이 그림은 헬조선에 대한 트윗들의 핵심 단어들을 시각적으로 제공한다. 헬조선과 헬조센은 검색 키워드로서 다른 단어와 비교하여 굉장히 많이 등장하기 때문에 일부만 표현하였다. 오른쪽 아래의 가장 큰 글자가 헬조선의 일부다.

3.3. Correlation Analysis

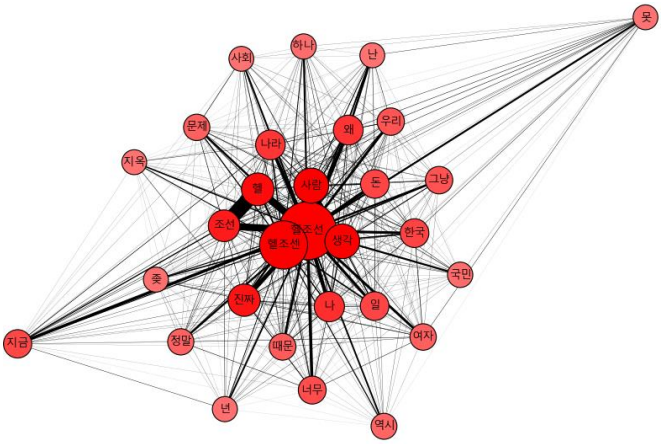

그림 5. 단어 공기 그래프

그림 5 는 단어 공기 그래프를 나타낸다. 각 정점은 단어를 나타내고, 간선은 단어 간 공기 빈도를 나타낸다. 정점의 크기 및 색의 짙음 정도는 단어 등장 빈도의 제곱근에 비례하며, 간선의 두께는 공기 빈도에 비례한다. 단어 공기 그래프는 간선을 만드는 기준치(threshold)를 설정함에 따라 다양한 그래프를 생성할 수 있다. 위 그래프는 이 기준치가 0 일 때, 즉 모든 공기 단어 간에 간선을 생성했을 때를 나타낸다. 여기서는 그림 3 과 마찬가지로 상위 30 개의 빈발 단어들만 사용하여 그래프를 구성하였으나 실제 실험에서는 최대 100 개의 단어를 사용하여 실험 및 분석하였다.

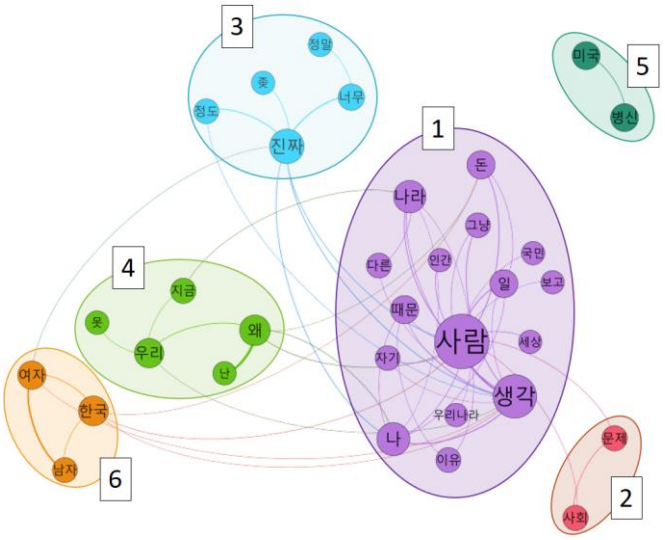

그림 6. 공기 그래프 기반 군집화

그림 6 은 공기 그래프를 기반으로 MCL 알고리즘을 이용하여 군집화 한 결과를 보여준다. Gephi 를 사용하여 군집화 및 시각화를 수행하였다. 적절한 군집화를 위해 간선 기준치를 4 로 설정하고 모든 단어와 공기가 보장되는 헬조선 및 헬조센을 데이터에서 제외하였다.

4. 분석 및 토의

그림 3 과 4 를 보면 헬조선이라는 단어를 둘러싼 부정적인 문맥을 파악할 수 있다. 등장 빈도 상위 50 개 단어를 살펴보면 새끼, 병신 등 총 7 개의 비속어와 지옥, 문제 등 15 개의 부정 혹은 문제제기 단어로 구성되어 있다. 즉, 전체 단어의 44%가 부정적인 맥락의 단어로, 이는 헬조선을 둘러싼 부정적인 맥락을 잘 나타낸다.

또한, 이 단어 중 중의적인 단어들이나 수식어구들을 제외하고 유의한 단어들만 남기면 트위터 사용자들의 헬조선에 대한 시각을 알 수 있다. 사람, 나라, 일, 돈, 여자, 놈, 사회, 현실, 게임, 탈출 등이 대표적이다. 이 중 일, 돈 등은 우리나라의 경제적 측면을 바라보는 시각을 상징하고, 나라, 사회 등은 정치사회적 측면을, 여자, 놈, 사람 등은 문화적 측면과 성별 간 대립을 그리고 탈출은 탈조선으로 대표되는 탈주의식을 나타낸다. 이와 같은 헬조선을 둘러싼 맥락은 기존의 연구에서도 찾아볼 수 있다[1,2].

헬조선과 유사한 맥락에서 최근에 많이 사용되는 신조어로 미개 와 노오-력(노오력, 노오오력 등) 등이 있어 이러한 단어들 또한 높은 빈도를 보여줄

것으로 예상하였으나, 노오-력은 등장 빈도 162 위, 미개는 223 위로 생각보다 많이 쓰이지 않는 것으로 나타났다.

그림 5 는 헬조선과 함께 등장하는 단어 간의 관계를 보여준다. 위에서 분석한 것과 같이 헬조선을 둘러싼 주요 키워드인 일, 돈, 왜, 생각, 사람, 나라 등이 헬조선과의 연관성이 깊은 것을 볼 수 있다. 또한, 수식어로서 다양한 상황에서 사용할 수 있는 못, 역시, 지금 등의 단어의 간선 수가 많은 것을 확인할 수 있다.

그림 6 은 여기서 한 걸음 더 나아가, 검색 키워드를 제외한 다른 단어 간의 관계를 나타낸다. 가장 큰 1 번 군집은 돈, 일 등의 경제적 측면을 반영하는 단어와 사람, 생각, 나라, 국민 등 정치 및 문화적 측면을 반영하는 단어가 혼합되어 있다. 이는 정치와 문화, 경제를 같은 트윗에서 한 번에 논하는 경우가 많다는 것을 의미한다. 사회와 문제로 구성된 2 번 군집은 헬조선을 사회적 문제로 바라보는 시각을 보여준다. 진짜, 너무 등으로 구성된 위쪽의 3 번 군집은 정도의 심함을 나타내는 수식어로 구성되어 있는데, 이는 헬조선이 갖는 부정적인 맥락의 정도가 심함을 의미한다. 4 번 군집은 난-왜, 우리-못 등 헬조선에서 살아가는 나와 우리에 대한 부정적인 시각을 보여준다. 5 번 군집은 흥미롭게도 미국에 대한 부정적인 시각을 보여준다. 실제로 미국에 대한 트윗들을 살펴보면 헬조선이나 미국이나 별반 다르지 않다는 글들을 많이 볼 수 있는데, 이는 미국의 대선 주자인 도널드 트럼프(Donald Trump)에 대한 부정적인 시각이 많이 반영된 것으로 보인다. 마지막으로 6 번 군집은 최근 심화되고 있는 사회문제 중 하나인 성별 간 대립을 나타낸다.

6. 결론 및 발전방향

본 연구에서는 트위터 데이터에 기반하여 최근 등장한 헬조선 또는 헬조센 이라는 신조어가 사용되는 사회적 맥락을 확인하였다. 단어 자체가 가진 뜻처럼 부정적인 맥락으로 주로 사용되었는데, 그중에서도 경제적 측면과 정치 및 사회적 측면에서 많이 사용되는 것을 보였다. 또한, 문화적 측면과 탈국가의식도 확인하였는데, 이러한 헬조선의 맥락은 기존의 연구결과들과도 합치한다[1,2].

뿐만 아니라, 본 연구에서는 미국에 대한 부정적인 인식과 남녀 간의 성별 대립이 나타나는 것 또한 확인할 수 있었는데, 이는 기존의 헬조선에 대한 연구에서 확인할 수 없었던 새로운 발견이다. 반대로, 헬조선이 유사한 맥락의 단어인 미개 또는 노오-력과 같이 사용되는 빈도는 그리 높지 않았다.

차후에는 더 많은 데이터를 기반으로 분석하여 단어의 의미 분포를 word2vec[11]을 통해 시각화 할 것이다. 또한, 본 연구는 검색 키워드로 헬조선 및 헬조센 만을 사용하였는데 유사 맥락인 미개, 노오-력 등의 단어들을 키워드로 사용한다면 더욱 다양한 결과를 얻을 수 있으리라 기대한다. 마지막으로 영어 단어의 긍정/부정의 정도를 나타내는 감성 강도 데이터셋인 SentiWordNet[12] 처럼, 한국어 단어들의 긍정/부정을 나타내는 데이터셋이 있다면 더욱 다양한 분석을 할 수 있다. 따라서 이와 같은 데이터셋을 직접 개발하여 더욱 다양한 분석을 시도할 것이다.

7. 감사의 글

이 논문은 2015 년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2015R1A2A1A05001845).

참고문헌

[1] 박권일, “‘헬조선’, 체제를 유지하는 파국론”, 황해문화, 통권, 제 90 호, pp. 73-95, 2016.

[2] 김명인, “굿모닝 헬조선; 분노와 혐오로 자욱한 신세계”, 황해문화, 통권, 제 90 호, pp. 2-10, 2016.

[3] 트위터 한국어 처리기, https://github.com/twitter/twitter-korean-text

[4] 은전한닢 형태소 분석기, http://eunjeon.blogspot.kr/

[5] 이동주, 연종흠, 황인범, 이상구, “꼬꼬마: 관계형 데이터베이스를 활용한 세종 말뭉치 활용 도구”, 정보과학회논문지: 컴퓨팅의 실제 및 레터, 16 권, 11 호, pp. 1046-1050, 2010.

[6] Van Dongen, Stijn Marinus, "Graph clustering by flow simulation,", *PhD thesis*, University of Utrecht, 2001.

[7] M. Bastian, S. Heymann, M. Jacomy, “Gephi: an open source software for exploring and manipulating networks,” *International AAAI Conference on Weblogs and Social Media*, 2009.

[8] S. Bird. “NLTK: the natural language toolkit,” *Proceedings of the COLING/ACL on Interactive presentation sessions*, Association for Computational Linguistics, 2006.

[9] 박은정, 조성준, “KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지”, 제 26 회 한글 및 한국어 정보처리 학술대회 논문집, 2014.

[10] A. Hagberg, D. Schult, P. Swart, “Exploring network structure, dynamics, and function using NetworkX,” *Proceedings of the 7th Python in Science Conferences (SciPy 2008),* 2008.

[11] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in neural information processing systems*, pp. 3111-3119, 2013.

[12] A. Esuli, F. Sebastiani, “Sentiwordnet: A publicly available lexical resource for opinion mining,” *Proceedings of LREC*, Vol. 6, 2006.