KIPS Trans. Softw. and Data Eng.
Vol.8, No.1 pp.13~18 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2019.8.1.13

Keywords and Topic Analysis of Social Issues on Twitter Based on Text Mining and Topic Modeling

Soo Jeong Kwak† · Hyon Hee Kim††

ABSTRACT

In this study, we investigate important keywords and their relationships among the keywords for social issues, and analyze topics to find subjects of the social issues. In particular, we collected twitter data with the keyword 'metoo' which has attracted much attention in these days, and perform keyword analysis and topic modeling. First, we preprocess the twitter data, identified important keywords, and analyzed the relatedness of the keywords. After then, topic modeling is performed to find subjects related to 'metoo'. Our experimental results showed that relatedness of keywords and subjects on social issues in twitter are well identified based on keyword analysis and topic modeling.

Keywords : Topic Modeling, Keyword Analysis, Text Mining, Twitter, Metoo

텍스트 마이닝과 토픽 모델링을 기반으로 한 트위터에 나타난 사회적 이슈의 키워드 및 주제 분석

곽 수 정† · 김 현 희††

요 약

본 연구는 커뮤니케이션이 활발한 SNS 속에서 사회적 이슈가 어떤 주제별로 나뉘어져 있고, 어떤 키워드들이 유기적으로 연결되었는지 그 연결 관계를 알아보고자 하였다. '미투'라는 새로운 단어가 생겨남과 동시에 큰 운동으로 번지고 있는 '미투운동'을 사회적 이슈로 간주하였고, 여러 SNS 중 특히 실시간 소통이 가장 활발한 트위터를 중심으로 분석을 실시하였다. 우선 키워드를 '미투'로 하여 관련된 키워드를 각 날짜별로 추출하였고, 주요 키워드를 파악한 후 토픽 모델링을 수행하였다. 이를 통해 사회적 이슈를 둘러싼 키워드들이 시간의 흐름에 따라 어떻게 변화하였는지 파악하고, 각 토픽 내의 키워드를 종합하여 토픽별 사회적 이슈의 다양한 관점을 해석하였다.

키워드 : 토픽 모델링, 키워드 분석, 텍스트 마이닝, 트위터, 미투

1. 서 론

과거 Social Network Service (SNS)가 없던 시절을 다시는 상상하지 못할 만큼 오늘날의 SNS는 우리의 일상 속 깊숙이 침투했다. 특히 불특정 다수와 메시지를 주고받고 실시간으로 의사소통을 한다는 특징은 사회적 이슈에 대해 즉각적인 반응을 확인할 수 있다. 또한 SNS의 전파속도는 방대한 양의 데이터가 빠르게 확산되어 특정 사건이 많은 사람들에게 쉽게 노출되기 때문에 그 사건을 바라보는 다양한 관점을 파악할 수 있다.

최근 사회적으로 큰 관심을 받은 이슈가 연구 논문과 뉴스와 같은 데이터를 이용한 주제 분석 연구도 진행되었다[1]. 본 연구에서는 SNS의 이러한 특징을 이용하여 사회적 이슈를 바라보는 다양한 관점을 파악하고 이슈의 중심에 있는 주요 키워드와 연결 관계를 분석하기 위해 실시간으로 관찰이 가장 용이한 트위터(Twitter)를 선택하였다.

'#MeToo'는 미국 할리우드의 유명 영화 제작자 하비 웨인스타인의 성추문 사건 이후 영화배우 알리사 밀라노가 지난해 10월 15일 처음 제안하면서 시작되었다. '미투운동'이란 이름으로 우리나라까지 그 파장이 확산되었고, 권력관계 때문에 불가피하게 성희롱, 성폭행을 당하는 여성들이 SNS에 '#

※ 이 논문은 동덕여자대학교 교내연구비에 의하여 연구되었음.
※ 이 논문은 2018년도 한국정보처리학회 춘계학술발표대회에서 '트위터에 나타난 미투운동의 키워드 연관성 및 키워드 네트워크 분석'의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 동덕여자대학교 정보통계학과 학사
†† 정 회 원 : 동덕여자대학교 정보통계학과 부교수
Manuscript Received : July 6, 2018
First Revision : August 17, 2018
Accepted : August 26, 2018
* Corresponding Author : Hyon Hee Kim(heekim@dongduk.ac.kr)

미투'로 해시태그를 걸고 성희롱, 성폭행 등 피해사실을 고백하여 세상에 알리는 운동으로 자리 잡았다. 현재까지도 각계 분야에서 활발히 진행되고 있는 '미투운동'은 고발과 폭로에서 그치지 않고 이를 해결하기 위한 방안을 모색하는 방향으로 나아가고 있다.

본 연구는 사회 전반적으로 영향을 끼치는 '미투운동'을 사회적 이슈로 간주하고 이에 대해 중심 키워드와 시간의 흐름에 따른 변화를 살펴보고자 한다. 먼저 텍스트 마이닝을 활용하여 주요 단어를 분석해 키워드 변화를 알아보고, 이후 토픽 모델링을 이용하여 '미투운동'에 대해 각 중요 키워드들이 나타내는 주제를 파악하였다[2].

본 연구의 공헌은 다음과 같다. 먼저, '미투운동'이라는 사회적 이슈에 대해 소셜 미디어를 과학적으로 분석하였고, 시간의 흐름에 따른 키워드 변화를 통해 대중들의 시선이 어떻게 이동하였는지 찾아내었다. 다음으로, 토픽 모델링을 통하여 소셜 미디어 사용자들이 사회적 이슈에 대해 갖는 여러 관점을 해석하였다.

본 논문의 구성은 다음과 같다. 제 2장에서 관련 연구에 관해 알아보고, 제 3장에서는 본 연구의 설계 과정을 자세히 설명한다. 이어 제 4장에서 분석 결과를 서술하고, 마지막으로 제 5장에서 결론을 맺는다.

2. 관련 연구

트위터를 이용하여 사회현상을 분석한 연구들이 최근 많은 관심을 얻고 있다. 특정 상품명에 관한 키워드를 기준으로 동시출현 단어 네트워크를 분석하고 토픽의 변화 시점 및 패턴을 파악하는 연구[2, 3]와 2014년 등장한 신조어 '헬 조선' 키워드를 분석하여 사회적 이슈에 대한 키워드와 그 맥락을 파악하는 연구[4] 등 트윗 데이터를 기반으로 다양한 분석 기법을 활용하는 연구들이 증가하고 있다.

[2]에서는 신문 기사로부터 토픽 모델링을 실시하여 주제를 추출하고 각 주제 오피니언 마이닝을 실행하여 주제의 구조와 내용을 분석하였다. 분석 결과 진보매체와 보수 매체 모두 이데올로기에 따라 보도하는 경향을 찾아내었다. [3]의 연구는 트위터 데이터로부터 토픽의 시간적 변화를 분석하였다. 마지막으로 [4]의 연구는 본 논문과 가장 유사한 연구로서, 키워드 '헬 조선'을 기반으로 트위터 데이터를 수집한 뒤 키워드 분석을 통하여 헬 조선의 사회적 맥락을 파악하였다.

이러한 연구들을 통해서 소셜 미디어 데이터를 분석하여 사회적 이슈에 대한 사용자들의 관점 및 인식을 파악하고 중요 키워드들을 통해서 주제를 해석하는 것이 유용한 접근 방법임을 알 수 있다.

3. 연구 설계

3.1 연구개요

본 연구의 연구 개요는 Fig. 1과 같다. 우선 국내 트위터에서 해시태그 '#미투'를 사용한 트윗 데이터를 수집하여 전처리 과정을 거쳤다. 수집한 데이터는 Term Frequency - Inverse Document Frequency (TF-IDF) 가중치를 적용하여 상위 키워드만 추출해 워드 클라우드를 생성하여 처음 미투운동이 시작한 지 얼마 되지 않은 초기와 시간이 흐르고 이슈의 진행방향이 꽤 뚜렷해진 현재의 키워드 양상을 비교하였다. 이후 토픽 모델링을 수행하여 시간의 흐름에 따라 각 토픽에서 나타나는 주제별 관점이 어떻게 달라졌는지 해석하였다. 본 연구에서의 토픽 모델링 및 시각화는 모두 통계 분석 언어인 R [5]을 사용하였다.

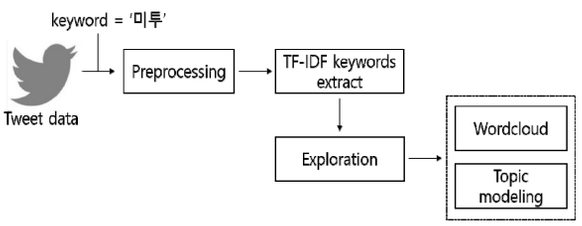

Fig. 1. Research Summary

3.2 분석 방법

먼저 국내 트위터에서 2018년 3월 20일을 기준으로 3일간 '미투' 단어가 포함된 20,000개의 트윗 데이터와 8월 20일을 기준으로 3일간에 걸쳐 13,997개의 트윗 데이터를 각각 수집하였다. 이 때, 추출 대상 범위를 국내로 한정시켰기 때문에 '미투'의 영문 표기(metoo)는 검색 단어로 쓰지 않았다.

수집한 데이터는 리트윗한 트윗을 제외한 맨션(mention)에 해당하는 부분만 사용하였다. 미투운동의 특성상 실명이 자주 등장하고 민감한 부분인 만큼 수집한 데이터에서 이름명사는 제외시켰고, 해시태그(#)와 불용어 및 인터넷 용어, 특수문자, 한자, 숫자를 없애는 전처리 과정[6]을 거쳐 모두 합친 총 4,403개의 트윗 데이터를 데이터베이스화 하였다. 단, '2차 피해'라는 단어가 자주 등장하여 임의로 해당 단어는 숫자를 없애지 않고 사전에 새로운 단어로 추가하였다.

본 연구에서 사용된 토픽 모델링 기법은 MALLET의 LDA (Latent Dirichlet Allocation)이다. LDA 모델은 단어의 사전(prior) 분포와 문서의 사전 분포가 있음을 가정하는 베이지안(Bayesian) 기법을 통해, 단어와 문서의 분포를 추정하여 해당 문서의 주요 단어와 분류를 추측하는 방식이다[7]. 해당 모델링 기법을 선택한 이유는 다음과 같다.

첫째, 분류에서 가장 중요도가 높은 단어들을 추출하여 해당 분류의 주제(topic)가 무엇인지 추측해볼 수 있다.

둘째, Dirichlet 분포를 사용하여 모수를 가정하지 않으며, 계산이 간편하고 많은 문서에 대해서도 처리시간이 빠르다.

그리하여 본 연구에서는 주제가 비교적 한정되어 있는 사회적 이슈에 대해 사용되어 한 단어가 갖는 주제별 의미를 명확히 파악하고자 하였다.

먼저 각 일자별 토픽의 수를 10개와 20개로 나누어 해석의 여지가 타당한 모델을 분석대상으로 하였다. 이를 바탕으로 '미투'와 관련하여 각 토픽의 주제를 파악하고 트위터 내에서 '미투'라는 사회적 이슈에 대해 어떤 관점이 존재하는지 알아

보았다. 이후 토픽 모델링 결과로 나온 토픽 단어들이 과거와 현재에 어떤 차이가 있는지 비교분석하였다.

4. 분석 결과

4.1 데이터 탐색

전처리 과정을 거쳐 만든 데이터베이스에 TF-IDF를 적용하여 '미투' 단어와 연관된 키워드를 알아보는 워드 클라우드를 시행하였다. TF-IDF는 TF(Term Frequency)는 특정 문서 하나에서 특정 단어가 나온 횟수를 나타내고, IDF(Inverse Document Frequency)는 특정 단어의 전체 문서내의 빈도를 역수로 취한 값이다[8]. 즉, TF-IDF는 단순 빈도에 가중치를 부여하여 문서 내에 얼마나 많은 비중을 차지하는지 나타내기 때문에 보다 정확한 중요도를 파악할 수 있다.

Fig. 2는 3월 20일을 기준으로 한 TDM을 매트릭스 형식으로 변환한 후, 행의 합을 계산하고 내림차순으로 정렬하여 최소 빈도수를 350으로 했을 때 생성한 워드 클라우드이다. TF-IDF를 적용했기 때문에 가중치가 높은 순으로 키워드 빈도수가 정렬되어 주요 단어들을 한 눈에 볼 수 있다. 중심 키워드로 가중치가 높은 단어는 '미투'이고, 그 뒤로 '미투운동', '공무원', '이유', '여자', '법정', '요건', '비서관', '짤린', '2차피해' 등이 따른다.

Fig. 2. 〈20180320〉 TF-IDF Word Cloud

Fig. 3은 8월 20일을 기준으로 최소 빈도수를 100으로 했을 때 생성한 워드 클라우드이다. 과거에 비해 현재 SNS에 언급되는 횟수가 현저히 줄어 수집할 수 있는 데이터의 양이 이전과 동일하지 않아 최소 빈도수를 낮게 설정하였다.

Fig. 2와 Fig. 3에서 볼 수 있는 바와 같이, '미투'와 '미투운동'을 제외한 다른 키워드들이 시간이 지남에 따라 달라진 것을 알 수 있다. 또한, 8월 20일 기준의 워드 클라우드에서는 '미투' 다음으로 '여성'이 먼저 등장했고 그 다음이 '미투운동'이었다. 이어서 '피해자', '운동', '사람', '사건', '무죄', '심판', '성폭력' 등이 따른다.

Fig. 3. 〈20180820〉 TF-IDF Word Cloud

3월 20일 기준 워드 클라우드에서는 우후죽순으로 터지는 제보와 사건에 대중들의 분노가 그대로 드러난 직설적인 욕과 함께 '팩트', '열풍', '적극', '안터진' 등의 단어가 등장하였고, 대부분 사태의 심각성과 사건 자체에 대한 내용이 분포되어 있다. 이에 반해 5개월이 지난 8월 20일 기준 워드 클라우드에서는 여전히 미투운동을 지지하는 단어들이 등장하긴 하지만 '불륜', '꽃뱀', '가짜', '허위', '무고' 라는 미투운동에 대해 부정적인 단어들도 함께 등장한다. 또한, '시위', '집회'는 여성인권 운동과 관련하여 시위나 집회가 많이 열렸기 때문에 등장한 것으로 보인다.

4.2 토픽 모델링

앞서 살펴본 워드 클라우드에서 사회적 이슈인 미투운동에 대한 키워드 양상이 매우 달라졌음을 알 수 있었다. 하지만, 키워드들이 어느 토픽에 속함으로써 그 의미가 어떻게 반영되는 지는 파악하기 어렵다. 따라서 키워드들을 개별적으로 보아서는 찾아내지 못하는 주제를 찾아내기 위해 TF-IDF를 적용한 DocumentTermMatrix(문서-단어 행렬, DTM)에 LDA기법을 이용하여 토픽 모델링을 수행하였다. 그리고 토픽들을 적절하게 분류하기 위해서 사분면 위에 시각화 하였고, 토픽 수를 날짜 별로 10개와 20개로 나누어 실행하였다. 또한, 사회적 이슈의 초기 발달 단계에서의 토픽 모델링 결과와 진화 단계에서의 토픽 모델링 결과의 차이를 살펴보고 각 토픽 내의 키워드가 어떻게 변화하였는지 살펴보았다. 토픽 모델링 시각화는 R의 LDAvis 패키지를 이용하였다[5].

토픽 모델링을 실행하면 사분면 위에 k개의 토픽 수만큼 원이 생성되며 각 원은 하나의 토픽을 나타낸다. 토픽 원들은 2차원 좌표 위에 축약하여 표현되고 토픽 원끼리 접점이 많을수록 겹치는 단어가 많다는 것을 의미한다. 각 토픽 원의 넓이는 코퍼스(corpus) 내에서 N개의 전체 토큰들에 대한 비율이다.

Fig. 4와 Fig. 5는 3월 20일자를 기준으로 수집한 데이터를 각각 토픽 수 10개와 20개로 나누어 만든 토픽 분포이다. Fig. 4의 토픽이 10개인 경우, 토픽 분포가 1번 토픽과 2번 토픽을 제외하고 두 토픽씩 겹쳐 있는 형태이다. Fig. 5의 토픽

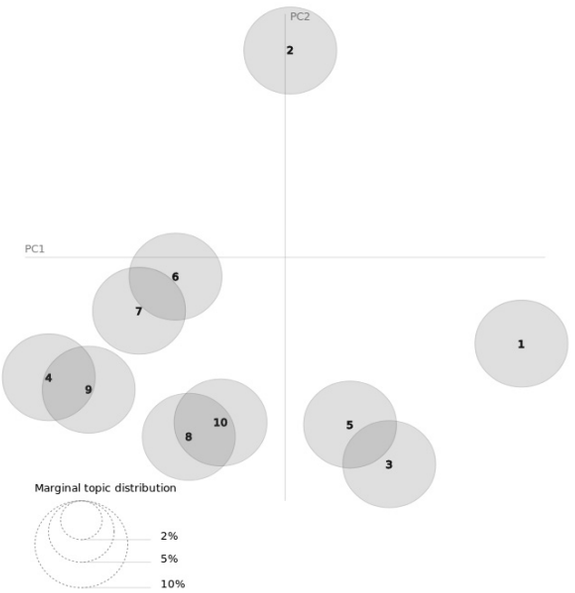

Fig. 4. 〈20180320〉 Topic Modeling – Topic 10

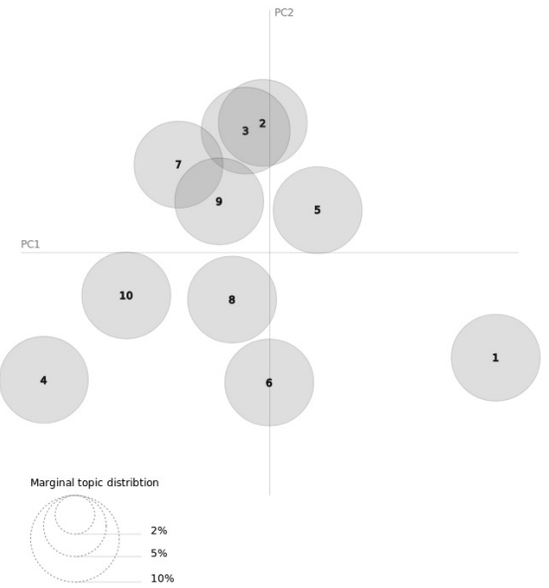

Fig. 6. 〈20180820〉 Topic Modeling – Topic 10

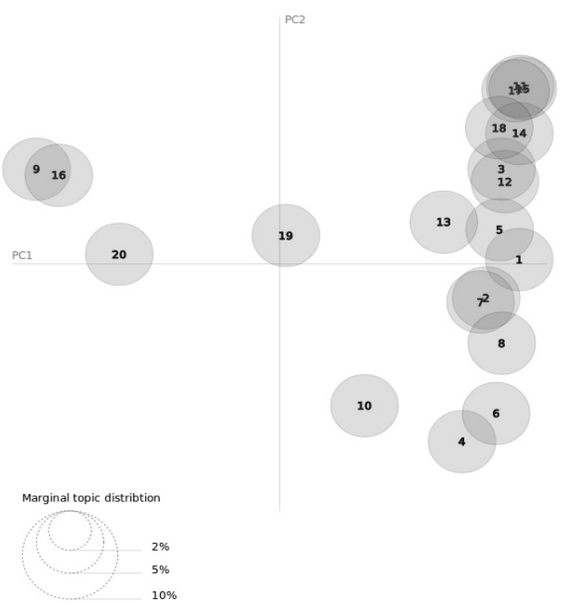

Fig. 5. 〈20180320〉 Topic Modeling – Topic 20

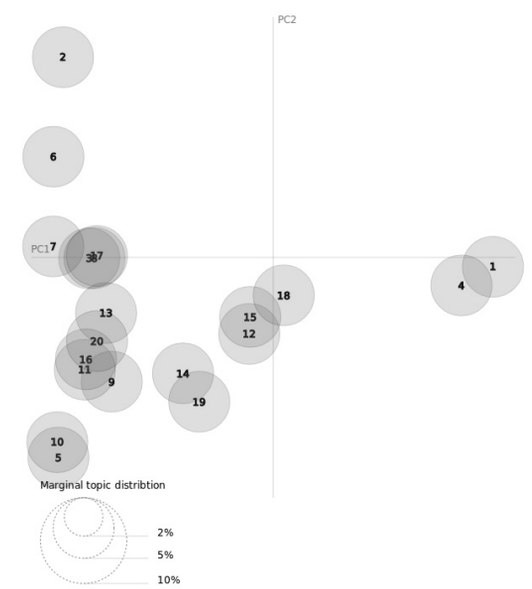

Fig. 7. 〈20180820〉 Topic Modeling – Topic 20

이 20개인 경우에는 9번, 10번, 16번, 19번, 20번 총 5개의 토픽을 제외하고 대부분의 토픽들이 오른쪽으로 치우쳐져 뭉쳐있는 것을 확인할 수 있다.

Fig. 6과 Fig. 7은 8월 20일자를 기준으로 수집한 데이터를 이전과 마찬가지로 토픽 수 10개와 20개로 나누어 만든 토픽 분포이다. Fig. 6의 토픽이 10개인 경우, 너무 퍼져 있지도 않고 심하게 뭉쳐있지 않은 반면 Fig. 7의 토픽이 20개인 경우는 몇 개의 토픽을 제외하고 대부분 뭉쳐있는 Fig. 5와 비슷한 분포를 보인다.

이에 본 연구는 수집한 데이터가 특정한 사회적 이슈를 대상으로 분석했기 때문에 다양한 토픽이 나올 수 있는 주제가 아니라는 점에서 토픽의 수가 많을수록 사회적 이슈를 설명하기에 일관성이 떨어질 것이라 판단하였다. 그리하여 토픽 수를 10개로 한 토픽 모델링 결과로 토픽별 주제를 해석하였다.

Fig. 4와 Fig. 6을 바탕으로 나온 결과를 좌표 상의 토픽 간 거리를 기준으로 다시 나누어 토픽 모델링을 만든 것이 Table 1과 Table 2이다. 이 때, 토픽 모델링을 생성하는 과정에서 모든 토픽에서 '미투'와 '미투운동'이 등장했다. 본 연구는 사회적 이슈의 큰 틀 자체가 미투, 미투운동이라는 점에서 토픽 모델링을 생성할 때 해당 단어를 제외하고 생성하였다. 그리하여 각 토픽에 대해 조금 더 객관적인 키워드를 파악하고자 하였다.

Table 1. 〈20180320〉 Topic Modeling - Topic 10

Topic1	Topic2	Topic3,5	Topic4,9	Topic6,7	Topic8,10
성추행	이유	면직	생각	여자	관련
공무원	면직	2차피해	사건	미성년자	여자
피해자	펜스	성욕	공무원	2차피해	공작
2차피해	여자	이해	우파	피해자	음모
요건	현직	처음	짤린	가해자	카톡
짤린	인정	조직	법정	속보	라인
댓글운동	사건	좌빨목사	운동	이슈	심정
조직	새끼	확실	자한당	안터진	경찰
공개	좌파	꼬리자르	승산	강간	팩트
존나	음모	기	시사	씨발	펜스
		입장			

Table 2. 〈20180820〉 Topic Modeling - Topic 10

Topic1	Topic2,3,7,9	Topic4	Topic5	Topic6,8	Topic10
누구	피해자	무죄	집회	여자	여성
꽃뱀	가해자	가짜	사건	성폭력	심판
불륜	판결	인권	사람	무죄	성폭력
사람	성추행	생각	여성	증거	남성
위안부	인권	운동	무죄	폭로	집회
성폭행	사법부	사법부	사회	분노	사회
관련	강간문화	남성	분노	발언	문제
악용	증거	행동	판단	페미니즘	행동
침묵	무죄	인정	참가	사실	시위
위력	한남	잘못	민주	iwpg	고발

먼저 Table 1은 Topic1, Topic2, Topic3,5 Topic4,9 Topic6,7 Topic8,10임의의 6개로 나눈 토픽 모델링이다. 토픽을 나눌 땐 토픽들 사이의 거리를 기준으로 하였고, Topic1과 Topic2를 제외한 합쳐진 토픽들은 키워드가 서로 고르게 나타나는 비율에 따라 단어를 선정하였다.

Topic1에서는 '성추행', '피해자', '2차피해'가 등장했다. 이는 미투운동의 특성 상 특정 인물을 지목해서 고발하는 것이기 때문에 피해자와 가해자 모두 실명으로 공개될 우려 속에 피해자에게 2차피해가 가해진다는 것을 알 수 있다. 또한, '댓글운동', '조직'은 2차피해를 가하는 요인이라고 해석될 여지가 있다. Topic2에서는 '펜스', '여자', '좌파', '음모'가 등장했다. 특히 '펜스'는 일부 사람들이 '미투운동'에 대응하는 방법이라며 생겨난 '펜스룰'[1] 때문에 등장한 것으로 보인다. 또한, '음모'는 미투 음모론이 제기되면서 '좌파' 단어와 함께 등장한 것으로 보인다.

이어 Topic3, Topic5에서는 '좌빨목사', '꼬리자르기'가 등장하였는데 여기서 '좌빨목사'는 좌파를 지칭하는 은어이고, '꼬리자르기'는 미투 고발된 인물을 제명 조치하는 모습을 비유한 것이다. 이는 미투운동이 시작되고 음모론이 언급되면서 좌파에 대한 부정적인 이미지가 일부 생겨났다고 볼 수 있다.

Topic4와 Topic9는 반대로 '우파'와 '자한당'이 등장했다. 그리고 '승산' 단어가 등장했는데 이는 선거운동을 앞두고 미투가 고발된 당의 후보와 경쟁하면 보다 유리하게 작용된다는 의미로 해석할 수 있다.

Topic6과 Topic7은 '미성년자', '여자', '피해자'가 등장했는데, 이는 미투운동의 피해자가 직장 내에서 뿐만 아니라 학교 내에서도 발생했음을 알려준다. 그리고 '안터진', '이슈', '속보'를 보아 사회적 위치 때문에 아직까지 밝혀지지 못한 피해가 더 있을 수 있다는 점을 알 수 있다.

마지막으로 Topic8, Topic10은 '여자', '카톡', '팩트'가 등장했다. 주로 미투를 고발할 때 카톡 내용을 증거로 공개하기 때문에 해당 카톡 내용의 팩트 여부가 중요할 것으로 해석된다. '공작'은 '음모'와 미투운동에 대해 같은 의미로 쓰였는데 미투 음모론을 미투 공작설이라고 지칭하기도 하여 함께 등장한 것으로 보인다. 토픽 중간에 욕설과 비하단어들이 등장한 것으로 보아 악용 사례가 나타나면서 '미투운동'에 대해 부정적인 입장이 생겨난 것으로 보인다.

Table 2도 Table 1과 마찬가지로 '미투'와 '미투운동' 단어를 제외한 후 토픽 모델링을 생성하였다. Topic2, Topic3, Topic7, Topic9를 하나의 토픽으로 묶고, Topic6과 Topic8을 하나의 토픽으로 묶었다. 그리고 나머지는 개별 토픽으로 선정하였다. 사분면의 위치를 기준으로 하였지만, Topic4와 Topic10이 가지는 키워드가 매우 상이하여 분리하였다. 또한 Topic6은 사분면의 세로축에 정 가운데로 놓여있지만 Topic8과 의미가 상통하는 단어가 겹쳐 등장해서 하나의 토픽으로 분류하였다.

Topic1에서는 '꽃뱀', '불륜', '악용'이 등장했다. 이는 누군가 미투운동을 의도적으로 악용하는 사례로 인해 피해자를 피해자로 보지 않고 분명 나쁜 의도가 숨어있을 것이라는 생각하는 관점에서 등장한 것으로 보인다. 또한 '위안부', '성폭행', '침묵'은 미투운동이 이슈로 떠오름과 동시에 아직 해결되지 않은 위안부 문제도 짚어야 할 과제로 등장했음을 알 수 있다.

Topic5와 Topic6, Topic8 그리고 Topic10은 비슷한 성격의 주제로 해석이 가능하며 이는 Topic1의 부정적인 성격과 반대된다. 우선 모두 공통적으로 '집회', '시위', '페미니즘', 'iwpg', '참가'와 같은 여성운동과 관련한 단어들이 '무죄', '분노', '폭로', '문제', '고발', '심판' 단어들과 함께 등장한다. 여기서 'iwpg'는 International Women's Peace Group의 약자로 세계여성평화그룹이다. 이는 미투운동에 분노한 여성들이 페미니즘 운동으로 확산시켜 나가는 현시점에 대한 토픽으로 보인다. 즉, 끊이지 않는 미투 고발에 대해 무죄 판결이 내려지는 사회에 대해 분노한 여성들이 집회에 참가하고, 시위를 함으로써 그 목소리를 내고 있는 것으로 해석할 수 있다.

Topic2, Topic3, Topic7, Topic9는 '피해자', '가해자', '판결', '사법부' 단어를 보아 미투운동의 가해자와 피해자가 법정에서 어떤 판결을 받는지에 대한 토픽으로 보인다. 그리고 이어지는 '강간문화', '무죄'는 '한남'이라는 신조어가 생겨나게 된 배경이 될 수 있는 해석을 불러온다. 마지막으로 Topic4는 '무죄', '가짜', '잘못', '인정'이 등장했다. 여기서 '가짜'는 미투를 거짓으로 고발하는 가짜 미투를 지칭한다. 이는 '무죄' 판결이 곧 '가짜' 미투로 이어질 수 있음을 알 수 있다.

1) 2002년 마이크 펜스 미국 부통령이 인터뷰에서 "아내 외의 여자와는 절대로 단둘이 식사하지 않는다."라고 말한 발언에서 유래된 용어이다.

5. 결 론

본 연구는 사회적으로 이슈화되고 있는 문제에 대해 트윗 데이터를 중심으로 중심 키워드를 파악한 후 토픽 모델링을 수행하여 시간의 흐름에 따른 각 토픽별 주제를 해석하였다. 연구 결과를 요약하면 다음과 같다.

먼저 날짜별 중심 키워드를 살펴보았을 때, 3월 20일 기준 이슈의 키워드와 8월 20일 기준 이슈의 키워드는 비슷한 것 같으면서도 '미투'와 '미투운동'을 제외하면 해석이 상이한 키워드가 많음을 알 수 있었다. 단 5개월의 경과임에도 불구하고 미투운동이라는 사회적 이슈에 대한 시선이 눈에 띄게 변화했고, 대중적이었던 관점이 현재에 와서는 확연히 상반된 입장으로 첨예하게 나뉘었다. 이는 사람들이 점차 객관적으로 사건을 바라보는 충분한 시간이 주어졌기 때문으로 생각된다.

다음으로 토픽 모델링을 통해 토픽별 키워드 변화를 눈에 띄게 파악할 수 있었다. 특히, 미투운동이 시작되고 한창 관심을 끌던 초기에는 피해자를 옹호하는 입장이 두드러질 것이라는 예상과 달리 변질 가능성을 고려해 검증의 필요성을 강조하는 입장이 뚜렷하게 나타나는 걸로 그쳤다. 하지만 미투운동이 진행된 지 꽤 시간이 흐른 현재에는 변질된 미투에 대해 부정적인 입장이 이전과 달리 자극적인 단어의 등장과 함께 두드러졌고, 미투운동을 옹호하는 입장은 집회와 시위를 하고 활발한 여성운동을 진행하는 등 보다 적극적으로 변화하였다. 또한 페미니즘 운동으로 번지는 '미투운동'에 대해 대처하는 '펜스룰'과 이에 대응하는 '한남' 이라는 신조어가 새롭게 등장했다.

한편 이슈 초기 시점에는 직장 내 성폭력 고발이라는 미투운동의 특성상 여자, 남자의 성 대립 구조가 아닌 갑과 을의 대립이었는데 미투의 악용 사례와 납득할 수 없는 무죄 판결로 인해 현재 시점에는 여성과 남성의 대립으로 점차 번져가고 있다.

본 연구는 사회적으로 이슈가 되는 문제에 대해 초기 단계와 그 이후 단계를 중심 키워드와 함께 토픽 모델링으로 변화 형태를 파악하여 시간의 흐름에 따라 어떻게 해석이 달라지는지에 의의를 두고 있다. 그러나 사회적 이슈는 처음은 뚜렷해도 끝은 구체적으로 존재하지 않아 변화 과정을 단계별로 해석할 때 마지막 시점에 대한 분석이 어렵다는 데에 한계점이 있다. 따라서 향후 사회적 이슈에 대해 분석하고자 할 때 본 연구에서 이용한 기법에 단계별로 장기간에 걸친 시계열 분석을 추가한다면 특정 시기뿐만 아니라 사회적 이슈의 구체적인 발전 과정을 전반적으로 파악하는데 활용될 수 있을 것이라 기대한다.

References

[1] J. Y. An, K. B. Ahn, and M. Song, "Text Mining Driven Content Analysis of Ebola on News Media and Scientific Publications," *Journal of the Korean Society for Library and Information Science*, Vol.50, No.2, pp.289-307, 2016.

[2] B. I. Kang, M. Song, and W. S. Jho, "A Study on Opinion Mining of Newspaper Texts based on Topic Modeling," *Journal of the Korean Society for Library and Information Science*, Vol.47, No.4, pp.315-334, 2013.

[3] S. A. Jin, C. E. Heo, Y. K. Jeong, and M. Song, "Topic-Network based Topic Shift Detection on Twitter," *Journal of the Korean Society for Information Management*, Vol.30, No.1, pp.285-302, 2013.

[4] J. B. Cha, J. W. Sung, J. G. Kim, and S. H. Park, "Hell-Chosun Keyword Analysis based on Twitter," *Journal of the Korean Multimedia Society*, Vol.19, No.2, pp.195-198, 2016.

[5] The R Project for Statistical Computing [Internet], https://www.r-project.org/

[6] S. Y. Kim, Y. M. Chung, "An Experimental Study on Selecting Association Terms Using Text Mining Techniques," *Journal of the Korean Society for Information Management*, Vol.23, No.3, pp.147-165, 2006.

[7] J. H. Park and M. Song, "A Study on the Research Trends in Library & Information Science in Korea using Topic Modeling," *Journal of the Korean Society for Information Management*, Vol.30, No.1, pp.7-32, 2013.

[8] Y. Y. Na, J. G. Park, and I. C. Moon, "Analysis of approval ratings of presidential candidates using multidimensional Gaussian process and time series text data," *Proceedings of the Korean Operations Research And Management Society*, Yeosu, 2017, pp.1151-1156.

[9] C. W. Kwak, "Subject Association Analysis of Big Data Studies : Using Co-citation Networks," *Journal of the Korean Society for Information Management*, Vol.35, No.1, pp.13-32, 2018.

곽 수 정

https://orcid.org/0000-0003-0483-2632
e-mail : sujeng21@gmail.com
2013년~2018년 동덕여자대학교 정보통계학과(학사)
관심분야 : Deep Learning & Unstructured Data Analysis

김 현 희

https://orcid.org/0000-0002-7507-8342
e-mail : heekim@dongduk.ac.kr
1996년 이화여자대학교 컴퓨터학과(학사)
1998년 이화여자대학교 컴퓨터학과(석사)
2005년 이화여자대학교 컴퓨터공학과 (공학박사)
2005년~2006년 LG전자 디지털미디어 연구소 선임연구원
2006년~현 재 동덕여자대학교 정보통계학과 부교수
관심분야 : Machine Learning, Ontology, Big Data Analysis