Journal of the Korea Convergence Society
Vol. 11. No. 5, pp. 199-207, 2020

ISSN 2233-4890 / eISSN 2713-6353
https://doi.org/10.15207/JKCS.2020.11.5.199

텍스트 마이닝을 활용한 2017년 한국 대선 분석

안은희[1], 안정국[2*]
[1]연세대학교 경영학과 박사과정, [2]선문대학교 경영학과 교수

An Analysis of the 2017 Korean Presidential Election Using Text Mining

Eunhee An[1], Jungkook An[2*]
[1]Ph. D. Course, School of Business, Yonsei University
[2]Professor, Department of Business Administration, Sun Moon University

요 약 최근 빅데이터 분석은 대량의 데이터로부터 미래를 예측하여 가치를 창출할 수 있어 다양한 분야에서 주목받고 있으며, 정치 캠페인 운영이나 결과 예측에도 활용되고 있다. 하지만 기존의 연구는 특정 SNS 데이터만을 분석하여 후보자들에 대한 정보를 취합하는데 한계가 있었다. 이에 본 연구는 2017년 한국 대선 후보별 뉴스와 댓글을 수집하여 뉴스 생성 추이, 토픽 추출, 감성 분석, 키워드 분석, 키워드 감성 분석을 하였다. 분석 결과, 대선 후보 간 다양한 토픽들이 생성되는 것을 확인하였으며, 후보별 이슈가 되는 중점 키워드와 이에 대한 유권자들의 호응도가 추출되었다. 본 연구는 포털 뉴스에서 생성되는 대선 캠페인에 대한 동향을 마이닝 할 수 있게 했다는 점과 감성 분석을 통해 대권주자들에 대한 유권자들의 관심과 의견들을 정량화하여 수치화한 것에 의의가 있다. 본 연구가 여론 수렴의 도구적 방법을 제시함으로써 이를 바탕으로 전략적인 행동 방안을 도출할 수 있을 것을 기대한다.

주제어 : 대선 분석, 텍스트 마이닝, 토픽 추출, 감성 분석, 댓글 분석

Abstract Recently, big data analysis has drawn attention in various fields as it can generate value from large amounts of data and is also used to run political campaigns or predict results. However, existing research had limitations in compiling information about candidates at a high-level by analyzing only specific SNS data. Therefore, this study analyses news trends, topics extraction, sentiment analysis, keyword analysis, comment analysis for the 2017 presidential election of South Korea. The results show that various topics had been generated, and online opinions are extracted for trending keywords of respective candidates. This study also shows that portal news and comments can serve as useful tools for predicting the public's opinion on social issues. This study will This paper advances a building strategic course of action by providing a method of analyzing public opinion across various fields.

Key Words : Presidential election, Text mining, Topic extraction, Sentiment analysis, Comment analysis

1. 서론

정보 기술의 발전으로 인해 인터넷은 사람들의 일상적인 커뮤니케이션의 채널로 자리를 잡았고, 이는 사람들의 뉴스 생산 및 소비 방식에도 큰 변화를 일으켰다. 기성 언론사 외에도 플랫폼에 뉴스를 생성할 수 있는 주체들이 증가하면서 포털에 유통되고 있는 뉴스량이 과거보다 훨씬 방대해졌음은 물론, 주로 검색 및 뉴스 플랫폼을 통

*Corresponding Author : Jungkook An(jungkook@sunmoon.ac.kr)
Received February 26, 2020 Revised April 20, 2020
Accepted May 20, 2020 Published May 28, 2020

해 주로 뉴스를 소비한다. 대다수의 뉴스의 생산과 소비가 포털을 통해 이루어지는 만큼, 포털 뉴스는 사람들의 생활 및 의사결정에 중요한 영향력을 행사한다고 볼 수 있다. 또, 과거에는 뉴스를 소비하는 방식이 수동적이었다면, 현재는 온라인상에서 지식과 의견들을 실시간 댓글 등으로 적극적으로 참여를 하여 하나의 정보 생산 주체의 역할을 하여 기존의 커뮤니케이션 방식을 새롭게 변화시켰다. 이러한 현상에서 빅데이터는 사회 현상의 이해하고 예측하는데 새로운 기회로서 관심을 받게 되었으며, 빅데이터 분석기법을 통한 데이터 마이닝, 오피니언 마이닝 등 다양한 연구 방식이 등장했다. 더불어 시시각각 새롭게 등장하는 빅데이터를 활용하여 사회 다양한 분야에 적용하려는 노력 역시 활발하게 이루어지고 있다.

그러나 지금까지 빅데이터를 토대로한 대선 분석 연구는 특정 SNS의 데이터를 기반으로 진행되었는데, 이는 SNS 사용자 외의 유권자들의 의견을 대표하지 못한다는 점과 후보자에 대한 정보를 취합하는데 한계가 있다. 또한, 연구 자료에 의하면 나이, 성별, 인종, 사회경제적 지위, 온라인 경험 등 다양한 요인들이 유권자들이 사용하는 SNS 플랫폼에 영향을 미친다는 것으로 나타났다[1]. 특정 SNS에서 도출된 데이터로부터 도달 할 수 있는 결론 또한 편향성이 과도하다고 볼 수밖에 없다.

따라서 본 연구는 포털 뉴스 데이터를 수집하여 대선 후보자들과 관련된 기사 및 이슈를 파악하고, 기존의 단순한 텍스트 마이닝의 방식에서 한층 더 나아가 텍스트에 대한 감성 분석을 통해 사람들의 의견, 성향과 같은 주관적인 데이터의 특성을 규명하여 후보별 화두 및 여론에 대한 긍정적 인식 및 부정적 인식을 알아보고자 하였다. 이와 더불어 기사의 댓글 및 연관검색어 분석을 통해 유권자들의 관심사 및 선호도를 추출하여 여론을 분석하고자 하였다.

선거철의 후보자들에 대한 보도는 선거에 관한 전반적인 추이를 가늠하고 판단할 수 있는 중요한 기준이 된다. 이에 본 연구는 후보들의 특정 행동 및 발언에 대한 유권자들의 반응을 신속하게 파악하고 더 나아가 선거 결과를 예측할 수 있는 유권자 선호도를 추론하는데 효과적으로 사용될 수 있을 것이다. 기존 대선 예측 및 분석과는 다르게 정확성을 높이고 유권자들의 반응을 신속하게 파악하는 선거 결과 예측 효과를 기대할 수 있다. 또한, 유권자 관심사 및 선호도 데이터 기반으로 효과적인 선거 및 유세 전략을 세우며 추후 변경 사항에 유용한 참고 자료가 될 수 있을 것이다.

2. 선행 연구

토픽모델링은 문서들의 집합에서 키워드들을 비지도 학습의 텍스트 마이닝 기법으로 추상적인 토픽과 키워드들을 군집화하는 기법으로 주제에 대한 토픽들을 탐색하는데 많이 쓰인다[2, 3]. 가장 대표적으로 쓰이고 있는 토픽모델링 알고리즘은 LDA(Latent Dirichlet Allocation)이며, 최근에는 뉴스와 소셜미디어에서의 방대한 텍스트에서 토픽들을 추출하기 위해 토픽모델링 기법이 많이 활용되었다[4].

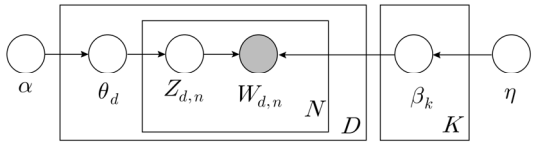

Fig. 1. Topic modeling[2, 3]

토픽모델링과 관련된 최근 연구들을 보면 광고 PR 분야 논문들의 핵심 주제어들을 분석하여 연구 동향을 분석한 연구[5]가 있었고, 텍스트 마이닝과 토픽모델링을 기반으로 트위터 데이터를 분석하여 사회적 이슈와 관련된 키워드들이 시간적 흐름에 따라 어떠한 변화가 있었는지[6], 각 산업 분야별 모바일 증강현실 앱 사용자들의 리뷰를 통해 만족과 불만족에 관련된 요소를 추출한 연구[7]가 있었다.

토픽모델링을 활용해 주제와 관련된 토픽을 추출하는 것은 사람들이 무엇(what)에 관심이 있는지를 알 수 있다. 더 나아가, 사람들이 토픽에 대해 어떠한(how) 의견을 가졌는지를 알기 위해서는 감성 분석이 쓰일 수 있으며, 온라인 쇼핑에서 상품에 대한 소비자들의 의견을 분석하는데 효율적으로 쓰일 수 있다. 한글 감성 분석의 경우 집단지성을 활용한 오픈한글[8]은 단어들의 정량적인 감성적 수치를 연구와 실무에서 다양한 분석에 활용이 되도록 API를 제공했다. 감성 분석은 소셜미디어, 뉴스, 웹에서의 텍스트에서 사람들의 감성을 정량화하여 분석하는 텍스트 마이닝 기법의 하나로 고급 분석기법이다. 또한, 특정 단어들에 대한 긍정, 부정 등의 의견을 추출하여 트렌드를 분석할 수 있는 기술이다[9]. 텍스트에 나타난 사람들의 의견, 태도, 성향과 같은 주관적인 데이터를 분석하고, 제품과 서비스에 대한 사용자들의 의견들에 대한 분석, 특히 기업들에 대한 부정적인 의견이나 특정 이슈들에 대해 실시간으로 모니터링을 하거나, 시장 현황,

경쟁 업체들에 대한 모니터링을 통해 기업과 관련된 다양한 활동이나 이슈들에 대한 고객이나 미디어의 반응 등을 정량적으로 측정하기 위해 사용되고 있다[15]. 한글 감성 분석의 경우, 집단지성을 활용한 감성 사전인 오픈 한글[8]이 있었고, 이를 활용한 감성 분석과 관련된 연구가 있다[4]. 감성분석과 관련된 최근 연구들을 보면, Yaqub et al 2017은 미국 대선과 관련된 트윗들의 감성 분석을 하였다[10]. 또한, CNN-LSTM 조합모델을 이용하여 영화 리뷰 감성 분석을 하여, 기존 방법보다 성능 개선을 시켰다는 연구가 있었고[11], 인터넷에서 자주 쓰이는 용어들의 감성분석을 통해 동영상 컨텐츠를 이용하는 사용자들의 만족도를 분석한 연구[12], 에어비앤비와 같은 공유숙박 서비스에 대한 사용자들의 감성을 분석하여 고객의 의견을 다양한 관점으로 이해를 하는 연구[13], 편의점 이용 고객들의 트위터 데이터를 통해 편의점들에 대한 감성을 비교한 연구가 있었다[14].

3. 연구 방법

3.1 데이터

본 연구의 연구 과정은 Fig. 2와 같다. 2017년 5월 1일부터 2017년 5월 7일 동안의 대선 후보 이름이 언급된 총 24,817건의 국내 포털 뉴스를 수집했다. 또한, 후보자들에 대한 여론을 살피기 위하여 같은 기간의 국내 포털 뉴스에 대한 댓글 총 485,811건을 수집하였다. 뉴스를 수집하여 데이터베이스에 저장을 하였고, 저장된 데이터베이스에서 뉴스의 컨텐츠와 댓글을 불러와서 전처리하는 과정을 진행하였다. 데이터를 검색하여 수집하는 과정에 있어서는 불필요한 데이터 또는 필터링이 되지 않아 잘못 수집이 된 데이터들이 많으므로, 키워드 선정이나 전처리 과정이 중요하다. 데이터 필터링을 제대로 하지 않을 경우 뉴스 버즈를 분석하는 과정에서 잘못된 결과를 낼 수 있어, 필터링할 키워드를 선정하는데 있어서 각 후보별로 일관성있는 키워드를 적용시키는 것이 중요하다. 그러므로 제외할 불용어가 특정 후보에게 많이 적용을 시키는 것을 방지하였다.

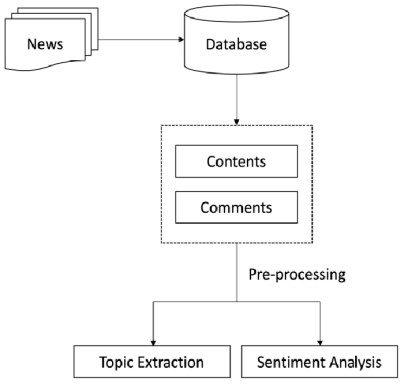

Fig. 2. Research Process

본 연구에서는 대선 후보자들에 관한 토픽들을 추출하기 위해 LDA 토픽모델링을 사용하였다[2, 3]. 토픽모델링은 문서의 집합에서 추상적인 토픽을 찾는 방법으로 최근 다양한 분석에서 많이 활용되었으며, 기계학습을 기반으로 한 텍스트 마이닝의 대표적인 비지도 학습이다[2, 3]. 그 중에서도 LDA(Latent Dirichlet Allocation) 알고리즘은 확률분포를 기반으로 하는 모델로을 본 연구에 적용을 하였다. 감성 분석은 국내 유일의 집단 지성 기반의 한글 감성어 사전인 오픈한글을 활용하여 감성의 깊이에 대한 분석의 정확도를 높였다. 또한 토픽을 추출하거나 감성 분석을 하는 과정에 있어서 특정 뉴스에 한 명 이상의 후보가 언급이 되는 경우에는 해당 뉴스는 분석에서 제외를 하는 방식을 택하였다.

4. 연구 결과

네이버 포털 뉴스 분석 결과, Fig. 3과 같이 뉴스에서 언급되는 비율은 문재인 후보가 32.1%, 홍준표 후보가 23.0%, 안철수 후보가 22.7%, 유승민 후보가 14.1%, 심상정 후보가 8%로 나타나 문재인 후보에 관한 기사가 가장 많음을 확인할 수 있었다.

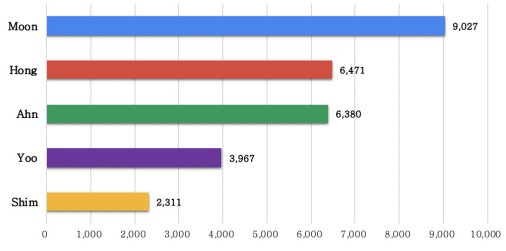

Fig. 3. News buzz by candidate

Fig. 4에서 뉴스 생성 추이 변화에 대한 조사 결과를 살펴보면 5월 2일을 기점으로 후보별 뉴스 생성 추이에 큰 변화가 생겼다. 5월 2일의 뉴스 생성 추이 변화 결과

가 조사 종료일까지의 순위와 거의 변동이 없다는 것을 확인할 수 있다. 문재인 후보의 경우 5월 2일을 기점으로 뉴스 생성량 3위에서 1위로 올랐으며, 홍준표 후보와 유승민 후보, 심상정 후보는 해당 일을 기점으로 뉴스 생성량이 급락한 것을 확인할 수 있다. 반면 안철수 후보는 뉴스 생성량에 큰 변화 없었다.

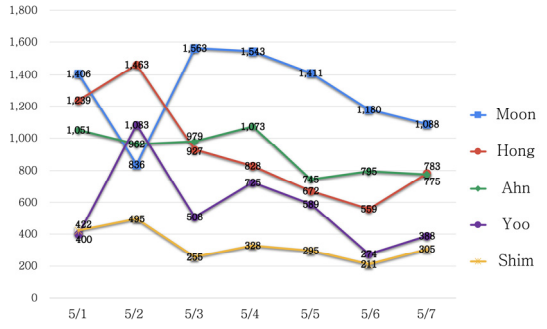

Fig. 4. News buzz trend by candidate

후보자들과 관련된 뉴스의 주요 분포를 살펴보면 주로 이슈, 지역, 정책에 집중되어 있음을 확인할 수 있다. Fig. 5에서 문재인 후보의 경우 이슈에 관한 뉴스가 가장 많고 인물과 지역 관련이 두 번째였다. 반면 홍준표 후보는 지역 관련 이슈가 가장 많음을 확인할 수 있는데, 이는 보수 중심의 TK 지역에서 상승세가 시작되었기 때문으로 추측된다. 안철수 후보는 다른 후보들과 다르게 정책에 관한 뉴스 집중도가 높게 나타났으며, 유승민 후보는 '이슈'에 관한 뉴스가 가장 많은 분포를 차지하는데 바른정당이 선거 막바지 탈당 사태를 겪은 데 따른 것으로 보인다. 심상정 후보는 지역과 정책에 대한 뉴스 집중도가 높게 나타났다.

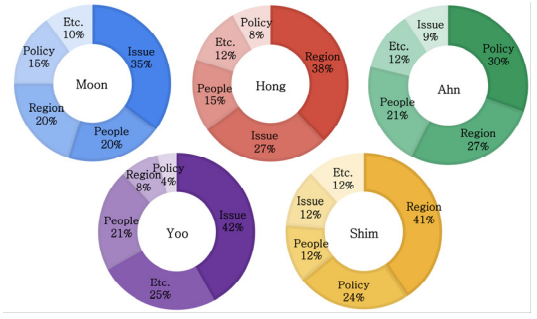

Fig. 5. News distribution by candidate

뉴스 데이터에서 각 후보에 대한 뉴스 생성량이 가장 많았던 날의 주요 키워드를 조사해보았다. 우선 문재인 후보의 경우 '세월호 인양', 'SBS', '사과', '의혹', '해수부' 등 정치적 이슈에 관한 키워드가, 안철수 후보의 경우 '경북', '국민 속으로', '도보', '버스', '뚜벅' 등의 자신의 정책적 방향 및 유세 상황에 대한 키워드가 추출되었다. 홍준표 후보의 경우 '집단 탈당', '긴급 회동', '유승민', '여론조사'의 키워드가 추출되어 당시 유승민 후보 및 바른정당 탈당 사태와 관련해 주목을 받았음을 확인할 수 있다. 유승민 후보의 경우 '집단 탈당', '홍준표'. '토론', '완주', '대학가', '성희롱' 등의 키워드가, 심상정 후보의 경우 '토론', '분권', '청년', '개헌'의 키워드가 추출되었다.

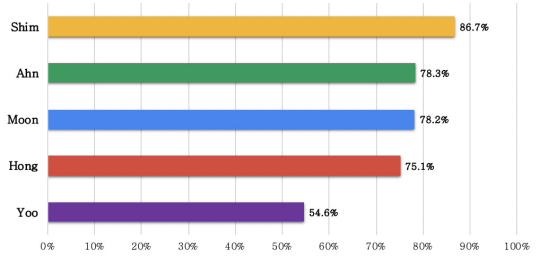

Fig. 6. Positive news by candidate

각 후보자가 언급된 뉴스 기사에 대한 감성 분석을 통해 후보별 긍정·부정 점수를 살펴보았다. 후보자가 단독으로 언급된 총 34,061건의 기사를 분석에 사용하였다. 긍정점수의 경우 Fig. 6과 같이, 심상정 후보가 86.7%로 가장 높았고, 그 후 안철수 후보가 78.3%, 문재인 후보가 78.2%, 홍준표 후보 75.1%, 유승민 후보 54.6% 순으로 나타났다. 부정 점수의 경우 Fig. 7과 같이, 유승민 후보가 45.4%로 가장 높았고, 그 후 홍준표 후보가 24.9%, 문재인 후보가 21.8%, 안철수 후보 21.7%, 심상정 후보 13.3% 순으로 나타났다. 뉴스 내용에 대한 감성 분석 결과와 대선 결과를 비교해보면 당선 가능성과 뉴스 내용에 대한 긍정·부정 점수 사이에는 큰 상관관계가 없는 것을 알 수 있다.

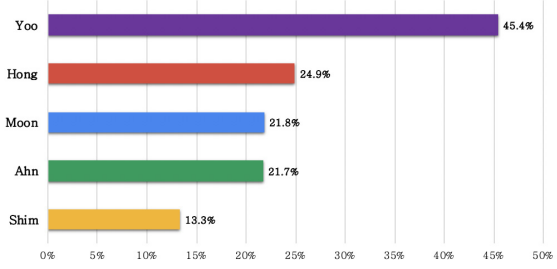

Fig. 7. Negative news by candidate

Table 1. Sentiment keywords from news

	Positive Keyword	Negative Keyword
Moon	Declaration of support, Focus, Appeal, Unity, Gwang-Ju, Overwhelming, Federation of Korean Trade Unions, Pledge	Sewol-hom Accusation, SBS, Son, Suspision, Recruitment privilege
Hong	Support, Conservative union, Park Geun-hye, Desertion from the party, Park Geun-ryeong	Rough talk, Buggin, Suspected involvement, Resignation, Park Jie-won, Cho Won-jin, Blame
Ahn	Honam, Future, People, Youth, Joint government, Communication, Reform, Mentor	Jeon Tae-il, Labor Organization, Unity, Campaign cancellation, Woo Sang-ho, Vote hearts, Poll result
Yoo	Conservatism, Reform and conservatism, Vote hearts, Full run, Reform, Encourage, Yoo Dam	Incident, Yoo Seung-min daughter, HR request, Ahn Jong-bum, Unity, Full run, Group Desertion from the party
Shim	Labor contributors, Confident, KTCU, Make, Discussion, Street, Disabled	Resignation, Argue, Criticism, Progress

감성 분석을 통한 키워드 분석 결과의 경우 Table 1에서 확인할 수 있다. 문재인 후보와 관련하여 긍정적인 키워드는 '지지 선언', '집중' 등 지역, 유세와 관련한 키워드가 추출되었고, 홍준표 후보의 경우 '지지', '보수 결집', '박근혜' 등 보수 세력과 전 정권에 관한 키워드가 등장했다. 안철수 후보의 경우 '호남', '미래', '국민' 등 지역 및 정책에 관련된 키워드가, 유승민 후보는 '보수', '개혁 보수', '표심', '완주' 등의 키워드가 추출되었다. 심상정 후보의 경우 '노동 헌장', '당당하다' 등 노동에 관한 키워드가 주로 등장한 것을 확인할 수 있다. 후보별 부정 키워드를 추출한 결과는 다음과 같다. 문재인 후보와 관련해서는 '세월호', '고발', 'SBS' 등 아들 및 세월호에 관련된 언급이 많았으며, 홍준표 후보는 '막말', '도청' '개입 의혹' 등의 키워드가 언급되었다. 안철수 후보는 '전태일', '노동단체', '단일화' 등의 키워드가, 유승민 후보는 '사건', '유승민 딸', '인사청탁' 등의 키워드가 추출되었다. 심상정 후보의 경우 부정 키워드가 상대적으로 적어 키워드가 부정확하지만 주로 '사표론', '반박' 등의 키워드가 추출된 것을 확인할 수 있다.

후보자들에 대한 여론을 살피기 위하여 후보자들이 단독으로 언급된 댓글에 대한 긍정어 비율을 산출한 결과 Fig. 8과 같이 심상정 후보 67.7%, 문재인 후보 66.3%, 안철수 후보 62.2%, 홍준표 후보 55.9%, 유승민 후보 50.0% 순으로 나타났다. 심상정 후보의 경우에는 상대 후보 지지자들의 견제를 덜 받았던 것으로 판단할 수 있다.

뉴스 댓글에 대한 호응의 경우 대댓글 수, 추천수, 비추천수 세 가지 요인으로 세부 분석이 가능하다. 후보별 뉴스 댓글에 대한 호응 분석 결과는 Table 2와 같다. 문재인 후보의 긍정 댓글은 평균적으로 0.4개의 댓글, 32.6개의 추천, 4.4개의 비추천이 생성되었고 부정 댓글은 평균적으로 0.3개의 댓글, 18개의 추천, 3.5개의 비추천이 생성되었다. 문재인 후보를 제외한 모든 후보자의 평균 긍정·부정 대댓글 수는 일치했고 문재인 후보만 긍정 대댓글 수가 부정 대댓글 수보다 0.1 높았다. 댓글 추천 수 경우에는 문재인 후보와 유승민 후보만 긍정 댓글 추천 수가 높았고 나머지 후보자들은 부정 댓글 추천 수가 더 높았다. 뉴스 댓글에 긍정어 비율이 높다는 것은 해당 후보의 콘크리트 지지자들의 성향이 나타났기 때문이란 해석과 상대 후보 지지자들의 견제를 덜 받는다는 해석이 가능하다.

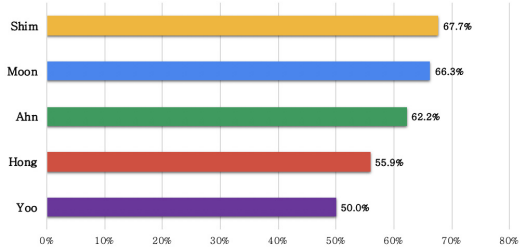

Fig. 8. Positive comments by candidate

Table 2. Sentiments from the comments of the comments

Cand.	S(1)	C(2)	Like	Dislike
Moon	Pos.	0.4	32.6	4.4
	Neg.	0.3	18.0	3.5
Hong	Pos.	0.2	12.0	2.1
	Neg.	0.2	22.4	2.0
Ahn	Pos.	0.2	9.0	3.3
	Neg.	0.2	24.5	3.5
Yoo	Pos.	0.1	22.6	1.6
	Neg.	0.1	16.0	1.6
Shim	Pos.	0.1	13.6	2.6
	Neg.	0.1	17.3	2.7

1) Sentiment
2) Comments of the Comments

후보별 댓글의 연관 키워드를 추출하여 분석한 결과 Table 2와 같이 문재인 후보는 '적폐', '당선' 같은 키워드가, 안철수 후보는 '미래', '박지원', 홍준표 후보는 '보수', '박근혜', 유승민 후보는 '보수', '바르다', 심상정 후보는 '진보', '사표'의 키워드가 각각 추출된 것을 확인할 수 있다.

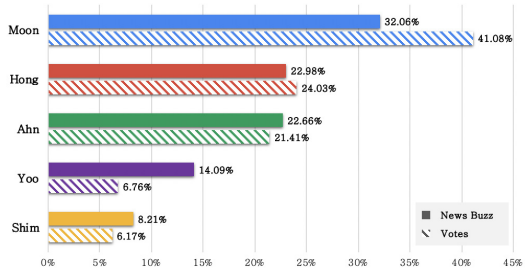

Fig. 9. News buzz and voting result

Table 3. Related keywords by candidate

	Moon	Ahn	Hong	Yoo	Shim
1	Jeokpye	Future	Conservative	Conservatism	Progressive
2	Elected	Park Jie-won	Park Geun-hye	Be right	Resignation
3	Park Geun-hye	Conservative	Pig aphrodisiac	Cheering	Discuss
4	Sewol-ho	Discuss	Rough talk	Desertion from the party	Jeokpye
5	Overwhelming	Honam	Discuss	Full run	Worker
6	Clearing	Policy	Forces	Traitor	Discuss
7	Forces	Jeokpye	Party	Park Geun-hye	candlelight
8	SBS	New politics	Lie	Reform	Clearing
9	Media	Economy	Jeokpye	Fair	Pledge
10	Son	Pledge	Left wing	Kim Moo-sung	Reality

5. 연구 결과 토의 및 시사점

5.1 연구결과 토의

기존의 텍스트 마이닝을 사용하여 선거 결과를 예측에 관한 연구들의 경우, 트위터를 많이 분석해왔지만 트위터는 한국에서 쓰는 비율이 높지 않아 한국 유권자들에 대한 대표성 확보가 어려워 대선 결과 예측 연구에 있어 보완의 필요성이 있다고 판단되었다. 이에 본 연구는 대선 후보들에 대한 네이버 포털 뉴스 및 해당 기사에 달린 댓글을 토대로 텍스트 마이닝 및 감성 분석을 진행하였다. 여론 공표 금지 기간을 포함한 2017년 5월 1일부터 2017년 5월 7일까지의 네이버 포털 뉴스 데이터를 분석한 결과는 다음과 같다.

첫째, Fig. 9에서 뉴스 생성량을 살펴보면 후보자에 따른 생성된 뉴스량의 순위와 대선 후보 득표 순위가 같았다. 문재인 후보의 뉴스 생성 비율이 실제 득표율에 비교하여 적은 수치로 측정이 되었다. 반면 유승민 후보의 경우에는 실제 득표율보다 뉴스 생성 비율이 비교적 높게 측정이 되었다. 이에 텍스트 마이닝을 바탕으로 한 뉴스 생성 결과에 대한 분석은 현재 진행되고 있는 여론조사의 한계점을 보완하며 대선 결과를 예측할 수 있는 주요 지표가 될 수 있음을 시사한다. 실제로 현재 진행되고 있는 여론조사의 결과에 대해 대중들이 많은 의구심이 생기고 있다. 2016년 총선 당시에는 기존 여론조사 기관들이 한 번도 예측하지 못한 여소야대 상황이 벌어졌고 [16], 2017년 미국 대선에서도 부정적인 이미지를 가지고 있는 대권 주자인 트럼프를 지지했던 유권자들이, 주위의 시선을 의식하여 공개적인 지지 의사를 비치지 않아, 여론조사와 투표 결과가 상이하게 다르게 나왔던 적이 있다. 포털 뉴스 기사 생성 추이를 분석한 결과 뉴스에서 후보자에 대한 언급 빈도수가 실제 대선 결과 순위와 일치하며, 득표율 역시 소수의 후보를 제외한 다른 후보들의 경우, 실제 득표율과 굉장히 유사한 결괏값을 지닌다는 점에서 이는 기존의 여론조사를 보완하는 새로운 수단임을 확인할 수 있다.

둘째, 뉴스 생성 추이에 대한 분석을 토대로 대부분의 후보자 뉴스 생성 추이가 5월 2일을 기점으로 급변하는 것을 확인할 수 있다. 5월 2일은 '사회'를 주제로 하는 선거 전 마지막 대선 후보 토론회가 진행되었다는 점을 고려해보았을 때, 해당 토론으로 인해 후보자 뉴스 생성 및 연관 감성어 트랜드 패턴이 변화한 것이란 해석이 가능하다. 토론 직후 생성된 뉴스 버즈량이 최종 선거 결과 순위와 비슷하게 나타났다는 점이 의미가 있으며, 이를 통해 선거를 앞둔 마지막 토론이 뉴스 생성 추이에 직접적인 영향력을 행사하고, 더 나아가 유권자들에게 간접적인 영향을 미친다고 해석할 수 있다. 실제 연구 결과에 의하면 투표 후보 결정 시 참고한 매체로는 'TV토론'이 59%로 가장 많았고 '신문/방송 보도' 23%, '인터넷 뉴스' 17%, '가족/주위사람' 14%, '페이스북, 카카오톡 등 SNS' 12%, '선거 유세' 11%, '선거공보/벽보' 7%, '신문/방송/인터넷 광고' 6% '본인 생각/판단' 2%, '기타' 1% 순이었다[17]. 32.1%

셋째, 감성 분석을 통해 추출한 뉴스 기사의 긍정도는 일정 수준 이상이 되면 대선 결과와 크게 유관하지 않다는 것을 확인하였다. 심상정 후보의 경우 뉴스 긍정점수가 86%가량을 달성했고 뉴스 댓글 긍정점수 또한 후보 중 제일 높은 67%를 달성했는데도 불구하고, 실제 대선 결과 득표율 5위에 그치고 말았다. 이는 긍정점수가 높다는 것은 상대적으로 당선 가능성이 작아 다른 후보들의 견제를 덜 받아 나타나는 현상으로도 해석할 수 있다. 혹은 진보 정치적인 성향을 띄고 있는 유권자들은 심상정 후보 또한 선호했지만, 확실히 지지율이 더 높은 문재인 후보를 뽑아 민주당의 승리를 기원했을 것으로도 해석할 수 있다. 반면 문재인 후보의 뉴스 댓글 점수는 1위 심상정 후보와 1% 차이밖에 나지 않은 높은 점수였고 대선 결과 역시 유사했다. 이는 문재인의 콘크리트 지지자들의 성향이 뉴스 댓글에서도 나타나기 때문이라고 볼 수 있다. 이와 연계하여 여론을 수렴하고자 포털 뉴스의 댓글의 감성 분석을 진행한 결과, 기사 긍정·부정도의 순위와 댓글 긍정·부정도의 순위는 동일하게 나타났다. 이에 기사의 내용이 댓글에 영향을 미친다는 것을 추론해볼 수 있다. 다만, 뉴스 기사보다 댓글에 전반적으로 부정도가 높게 나타나는 것에서, 유권자들이 기사를 비판적인 시각으로 보는 경향, 또는 자신이 지지하지 않는 후보자에 대해 악의적인 댓글을 다는 경향이 있다는 것을 유추해낼 수 있다.

5.2 연구의 한계 및 향후 연구 방향

본 연구는 네이버 포털 뉴스 및 해당 뉴스의 댓글들을 수집하여 빈도수 분석, 토픽 추출, 감성 분석과 같은 다양한 텍스트 마이닝 기법을 활용하여 19대 대선 후보자들을 분석하였다. 수집한 데이터는 아래와 같은 한계점을 지닌다.

첫째, 뉴스 기사가 모든 사람들의 의견을 대변한다고 볼 수 없다. 특히 정치와 기업과의 연계로 인해 치우친 정치적 편향을 가진 언론들의 경우, 정치적 편향성으로 인해 이슈에 대해 방어적이거나 공격적인 주관적 보도를 진행하는 성향이 있다. 따라서 정파적 언론의 경우 특정 이슈에 대한 보도 양의 차이에 대한 편차가 클 확률이 높다[18]. 즉 포털 뉴스를 기반으로 한 대선 관련 연구는 사회적 논의 트렌드를 대표하기 위해서는 언론사별로 기사를 제한하여 데이터 표본을 추출할 필요성이 있다.

본 연구는 뉴스 기사에 한정된 댓글을 토대로 여론을 조사한 것이다. 하지만 댓글은 전체성보다는 편향성을 가질 수 있으므로, 후보자에 대한 의견을 적극적으로 개진할 수 있는 공간이 아니라는 한계가 있다. 이에 추후 조사에서는 한국 사람들이 자주 사용하는 SNS 채널들을 추가적으로 분석하여 표본을 모집단에 가깝게 추출해야 할 필요가 있다.

둘째, 본 연구의 대선 분석 기간은 5월 1일에서 5월 7일, 5월 9일 본 선거일 8일 전부터 시작했기에 장기 선거 전략에 줄 수 있는 큰 파급력은 기대하기 어렵다. 선거 운동 기간은 23일밖에 되지 않아 그 짧은 기간 동안 효과적인 유세를 펼치기 위해 예비 후보자들은 몇 년 전부터 선거 전략을 준비하기 시작한다. 이에 비해 두 주일 전부터 시작되는 데이터 마이닝 분석은 대선 후보들의 유세가 시작되고 이슈 뉴스가 생성되는 짧은 시기에만 가능하다. 많은 변화가 일어날 수 있는 본 선거일 전 일주일 동안의 자료를 수집하여 대선 결과를 예측할 수는 있지만, 그 결과에 큰 영향은 미치지 못할 것이다.

따라서 차기 후보자들은 장기적으로 본 연구의 사회 트렌드 마이닝 기술을 사용하기 위해 뉴스와 데이터가 상대적으로 적은 선거 운동 기간 전에는 과거 결과 및 분석을 활용하여 어느 이슈가 후보자의 호응도에 제일 큰 영향을 미쳤는지 확인하고 전략적으로 사용할 필요가 있다.

5.3 시사점

본 연구는 다음과 같은 학술적 시사점을 갖는다. 우선, 본 연구는 네이버 포털 뉴스 생성 추이 분석을 진행하였고, 해당 결과는 2017년 대선 후보 득표 순위와 일치함을 확인하였다. 기존의 대선 분석에서 사용되는 전통적인 여론조사의 경우 대선 후보들에 대한 유권자들의 선택을 예측하기가 쉽지 않다는 점에서, 본 연구는 뉴스의 생성 추이 분석을 통해 이를 보완할 수 있다는 가능성을 보여주었다.

둘째, 본 연구는 감성 분석을 적용하여 대선 후보자들에 대한 긍정 및 부정적 키워드를 추출하고, 이를 토대로 긍정 점수 및 부정 점수에 관한 결과를 도출하였다. 단순히 뉴스의 생성량 추이를 넘어 후보자들에 대한 긍정 및 부정적인 반응을 모아 시각화함으로써 후보자들에 대한 자질 및 이슈들을 한눈에 이해할 수 있게 하였다. 유권자들은 후보자에 대한 객관적인 정보를 바탕으로 합리적인 판단을 해야 하는데, 난립하는 개개의 뉴스만을 통해서는 후보자들에 대한 이슈를 한 번에 파악하기 쉽지 않다. 이에 후보자들의 최근 이슈 및 관련 키워드를 긍정과 부정으로 나누어 정리함으로써, 현재 뉴스화된 후보자들의 정보를 시각화하여 유권자들에게 차별화된 가치를 제공한다.

더불어 네이버 뉴스의 댓글에 대해 감성 분석을 진행

하여 온라인 유권자들의 대선 후보에 대한 기대치와 반응을 시각화하고, 댓글에서 언급된 후보자들과의 연관검색어를 추적하였다. 이는 뉴스 기사를 작성하는 언론사와 댓글을 작성하는 온라인 유권자들 사이에서 대선 후보자들에 대한 관점 및 시선의 온도 차를 비교 분석할 수 있는 지표가 된다는 데 의의가 있다.

본 연구는 학술적 시사점 외에도 다음과 같은 실무적 시사점을 지닌다. 최근 텍스트 마이닝을 사용하여 선거 결과를 예측하는 연구나 보고서가 점차 증가하고 있기는 하지만[19], 기존의 연구들은 SNS 데이터나 뉴스 댓글들에만 초점을 맞춘 경우가 대부분이었다. 하지만 본 연구는 포털의 뉴스를 대상으로 연구를 진행하여 후보자에 대한 개인의 주관이 배제된 자료들을 바탕으로 객관적이고 실증적인 데이터를 수집했다는 데 의의가 있다. 민주사회에서 공정한 투표는 유권자들이 후보자를 판단할 때 객관적이고 정확한 정보를 기반으로 판단했다는 전제가 충족되어야 성립한다. 본 연구는 정보 전달에 있어 최대한의 객관성을 담보하는 뉴스 기사를 토대로 연구를 진행하고 이를 시각화함으로써 대선 후보자들의 행동과 사회적 반응에 대한 객관적이고 직관적인 이해를 가능케 했다. 이러한 연구 결과는 유권자들이 후보들의 행동 및 발언에 대해 직관적으로 파악하고, 더 나아가 선거에 영향을 미치는 유권자 선호를 형성하는데 이바지할 것이다.

둘째, 본 연구는 정치 캠페인 전략 수립을 위한 자료로 사용될 수 있다. 데이터 분석 기반의 선거 전략은 2012년 미국 오바마 대통령 선거부터 시작되었다. 오바마 캠프는 선거에서 지지자들의 자료를 분석해 효과적으로 활용했다[20]. 선거 운동을 시작하기도 전에 유권자들에 대한 정보를 수집할 수 있었기 때문에 맞춤형 선거 전략을 짜고 적용할 수 있었다. 2012년 미국 대통령 선거는 데이터 기반 선거 캠페인의 파급력을 분명히 보여주었다. 오바마 캠프는 4년 만에 데이터 분석팀의 규모를 5배로 늘리고 데이터 분석팀의 예산을 3배로 늘리면서 전략을 수립하는데 투자를 아끼지 않았다[21, 22]. 따라서 한국 선거 또한 데이터 분석 기반의 선거 자료를 활용하여 후보자가 내놓은 정책 및 발언에 대해 유권자들이 어떻게 반응하는지 파악할 수 있으며, 후보자 행위에 대한 유권자의 선호도를 추론하는 데 효과적으로 사용될 수 있다. 사람들의 반응을 실시간으로 파악함으로써 후보자가 최선의 선거 전략을 토대로 효과적인 유세 전략을 펼치는 데 도움을 줄 수 있을 것이라 기대한다.

더 나아가 본 연구에서 시도한 포털 뉴스에 대한 분석은 비단 선거뿐만 아닌 뉴스에서 다루는 사회 다방면의 이슈들에 대한 객관적 분석과 여론을 파악하는데 적절하게 활용될 수 있을 것이다. 사안에 대한 텍스트 마이닝의 분석은 논의되고 있는 모든 안건에 대한 정보들을 객관적으로 수집하고 이를 수치화함으로써 개인이 가지고 있는 편협한 관점에서 벗어나 객관적이고 포괄적으로 사안에 관한 결과를 제공함으로써, 우리 사회가 합리적이고 효과적인 판단을 내리는데 이바지할 것으로 판단한다.

REFERENCES

[1] E. Hargittai. (2015). Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites. *The ANNALS of the American Academy of Political and Social Science, 659(1),* 63-76.
DOI : 10.1177/0002716215570866

[2] D. M. Blei, A. Y. Ng & M. I. Jordan. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research, 3,* 993-1022.

[3] D. M. Blei, (2012). Probabilistic Topic Models, *Communications of the ACM, 55(4),* 77-84.
DOI : 10.1145/2133806.2133826

[4] J. K. An, S. H. Lee, E. H. An & H. W. Kim. (2016). Fintech Trends and Mobile Payment Service Analysis in Korea: Application of Text Mining Techniques, *Informatization Policy, 23(3),* 26-42.
DOI : 10.22693/NIAIP.2016.23.3.026

[5] H. M. Lee, J. W. Park & J. K. Lee. (2020). How Social Media Have Been Studied in the Academic Field of Advertising and Public Relations?: A Semantic Network and Topic Modeling Analysis, *Journal of Practical Research in Advertising and Public Relations, 13(1),* 130-158.

[6] S. J. Kwak & H. H. Kim. (2019). Keywords and Topic Analysis of Social Issues on Twitter Based on Text Mining and Topic Modeling. Korea Information *Processing Society, 8(1),* 13-18.
DOI : doi.org/10.3745/KTSDE.2019.8.1.13

[7] J. Hong, M. R. Yu & B. R. Choi. (2019). An Analysis of Mobile Augmented Reality App Reviews Using Topic Modeling, *Journal of Digital Contents Society, 20(7),* 1417-1427
DOI : 10.9728/dcs.2019.20.7.1417

[8] J. An & H. W. Kim. (2015). Building a Korean Sentiment Lexicon Using Collective Intelligence, *Journal of Intelligence and Information Systems, 21(2),* 49-67.
DOI : 10.13088/jiis.2015.21.2.49

[9] Y. N. Lee, E. J. Choi & M. J. Kim. (2018). Analysis of the Influence of Presidential Candidate's SNS Reputation on Election Result: focusing on 19th Presidential Election. *Journal of Digital Convergence, 16(2),* 195-201.

DOI : 10.14400/JDC.2018.16.2.195

[10] U. Yaqub, S. A. Chun, V. Atluri & J. Vaidya. (2017). Analysis of political discourse on twitter in the context of the 2016 US presidential elections. Government *Information Quarterly, 34(4),* 613-626. DOI : doi.org/10.1016/j.giq.2017.11.001

[11] H. Park & K. Kim. (2019). Sentiment Analysis of Movie Review Using Integrated CNN-LSTM Model, *Journal of Intelligence and Information Systems, 25(4),* 141-154 DOI : doi.org/10.13088/jiis.2019.25.4.141

[12] S. Kim, J. E. Kim., W. W. Seung & Y. Kim. (2019). Design of Video Advertisement Analysis via Analysis of Internet Term Sensitivity, *Journal of KIISE, 46(9),* 919-925. DOI : 10.5626/JOK.2019.46.9.919

[13] H. S. Park & J. Kim. (2019). Image change through emotional analysis of Shared accommodation service company, *Culinary Science & Hospitality Research, 25(5),* 67-74. DOI : 10.20878/cshr.2019.25.5.007

[14] H. Lee & J. Choi. (2019). Sentiment Analysis of Twitter Reviews toward Convenience Stores Customer in Korea, *Global Business Administration Review, 16(4),* 143-164.

[15] D. Hong, H. Jeong, S. Park E. Han, H. Kim & I. Yoon. (2017). Study on the Methodology for Extracting Information from SNS Using a Sentiment Analysis. *Intelligent Transportation Systems, 16(6),* 141-155. DOI : 10.12815/kits.2017.16.6.141

[16] J. Kim. (2016. 4. 14.). *Another wrong poll ... 'should be improved survey methods'*, The Chosun Ilbo. https://news.chosun.com/site/data/html_dir/2016/04/14/2016041401093.html

[17] Gallup Korea. (2017). Post-election survery of the 19th presidential election, Gallup Report. https://www.gallup.co.kr/gallupdb/reportContent.asp?seqNo=831

[18] S. W. Lee, H. J. Lee & B. K. Lee. (2018). The Varieties of Newsgathering and Processing Activities of Korean Press: Focusing on Choi Soon-Sil Scandal. *Journal of Research Methodology, 3(1),* 1-24. DOI : 10.21487/jrm.2018.5.3.1.1

[19] J. W. Bae, J. E. Son & M. Song. (2013). Analysis of Twitter for 2012 South Korea Presidential Election by Text Mining Techniques, *Journal of Intelligence and Information Systems, 19(3),* 141-156. DOI : 10.13088/jiis.2013.19.3.141

[20] K. S. Noh. (2013). A Exploratory Study on Big-data based Election Campaign Strategy Model in South Korea, *Journal of Digital Convergence, 11(12),* 113-120. DOI : 10.14400/JDPM.2013.11.12.113

[21] J. Kim. (2014. 2. 25). *Obama was elected president with big data*, Future Korea, http://www.futurekorea.co.kr/news/articleView.html?idxno=26124

[22] N. Lee. (2014. 12. 7). *[World Change Maker] Obama's Secret Weapon... Two presidential victories programming*, JoongAng Sunday, https://news.joins.com/article/16626818

안 은 희(Eunhee An) [학생회원]

- 2016년 2월 : 연세대학교 정보대학원 (석사)
- 2019년 3월 ~ 현재 : 연세대학교 경영학과(박사과정)
- 2020년 3월 ~ 현재 : 서울여자대학교 정보보호학과 겸임교수
- 관심분야 : 사이버보안, 인공지능, 데이터사이언스
- E-Mail : eunhee@yonsei.ac.kr

안 정 국(Jungkook An) [정회원]

- 2018년 8월 : 연세대학교 정보대학원 (정보시스템박사)
- 2019년 3월 ~ 2018년 8월 : 연세대학교 정보대학원 겸임교수
- 2019년 9월 ~ 현재 : 선문대학교 경영학과 조교수
- 관심분야 : 사이버보안, 인공지능, 데이터사이언스
- E-Mail : jungkook@sunmoon.ac.kr