ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2018 September: 24(3): 199~219 http://dx.doi.org/10.13088/jiis.2018.24.3.199

댓글 분석을 통한 19대 한국 대선 후보 이슈 파악 및 득표율 예측

서대호
연세대학교 정보대학원
(*seodaeho91@naver.com*)

김지호
고려대학교 산업경영공학부
(*navjiho@naver.com*)

김창기
연세대학교 정보대학원
(*niceck7@daum.net*)

인터넷의 일상화와 각종 스마트 기기의 보급으로 이용자들로 하여금 실시간 의사소통이 가능하게 하여 기존의 커뮤니케이션 양식이 새롭게 변화되었다. 인터넷을 통한 정보주체의 변화로 인해 데이터는 더욱 방대해져서 빅데이터라 불리는 정보의 초대형화를 야기하였다. 이러한 빅데이터는 사회적 실제를 이해하기 위한 새로운 기회로 여겨지고 있다. 특히 텍스트 마이닝은 비정형 텍스트 데이터를 이용해 패턴을 탐구하여 의미있는 정보를 찾아낸다. 텍스트 데이터는 신문, 도서, 웹, SNS 등 다양한 곳에 존재하기 때문에 데이터의 양이 매우 다양하고 방대하여 사회적 실제를 이해하기 위한 데이터로 적합하다. 본 연구는 한국 최대 인터넷 포털사이트 뉴스의 댓글을 수집하여 2017년 19대 한국 대선을 대상으로 연구를 수행하였다. 대선 선거일 직전 여론조사 공표 금지기간이 포함된 2017년 4월 29일부터 2017년 5월 7일까지 226,447건의 댓글을 수집하여 빈도분석, 연관감성어 분석, 토픽 감성 분석, 후보자 득표율 예측을 수행하였다. 이를 통해 각 후보자들에 대한 이슈를 분석 및 해석하고 득표율을 예측하였다. 분석 결과 뉴스 댓글이 대선 후보들에 대한 이슈를 추적하고 득표율을 예측하기에 효과적인 도구임을 보여주었다. 대선 후보자들은 사회적 여론을 객관적으로 판단하여 선거유세 전략에 반영할 수 있고 유권자들은 각 후보자들에 대한 이슈를 파악하여 투표시 참조할 수 있다. 또한 후보자들이 빅데이터 분석을 참조하여 선거캠페인을 벌인다면 국민들은 자신들이 원하는 바가 후보자들에게 피력, 반영된다는 것을 인지하고 웹상에서 더욱 적극적인 활동을 할 것이다. 이는 국민의 정치 참여 행위로써 사회적 의의가 있다.

주제어 : 19대 대선, 댓글, 텍스트 마이닝, 빅데이터 분석

논문접수일 : 2018년 5월 28일 논문수정일 : 2018년 9월 6일 게재확정일 : 2018년 9월 12일
원고유형 : 일반논문 교신저자 : 김창기

1. 서론

인터넷의 일상화와 각종 스마트 기기의 보급으로 대중의 정보생산, 공유가 확대되고 있다. 기존의 전통매체는 일대다 관계로 일방적인 정보 전달이 주 목적이었지만, 현대사회에서는 다대다 관계로 쌍방향성을 가지며 활발한 정보 공유를 한다. 이에 따라 대중은 다양한 플랫폼상에서 개인의 의견을 자유로이 게재할 수 있다. 특히 이용자들이 실시간으로 국내외 사회적 이슈를 전파할 수 있어 기존 전통매체에 비해 매우 빠르게 대중의 의견을 파악할 수 있다. 이용자들은 전문적인 지식이나 기술이 필요하지 않기 때문에 수동적인 콘텐츠 소비자였던 기존의 대중들은 지식과 정보의 생산과 소비를 동시에 주도하는 생산적 소비자로 거듭나게 되었으며, 스마트폰과 같은 스마트기기의 발달은 이용자들로 하여금 콘텐츠의 실시간 생산과 공유를 가능하

게 하여 기존의 커뮤니케이션 양식을 새롭게 변화시켰다(Bae et al., 2013).

정보주체의 변화는 빅데이터라 불리는 정보의 범람을 야기하였고 이러한 빅데이터는 사회적 이슈를 이해하기 위한 새로운 기회로 각광받고 있다. 이를 효율적으로 요약, 분석하기 위한 다양한 연구가 최근 핵심 화두이며 특히 텍스트기반 데이터 분석은 텍스트 마이닝이라 불리우며 많은 연구자들의 연구주제로 관심 받고 있다. 텍스트 마이닝은 비정형 텍스트 데이터를 언어학, 통계학, 기계학습 등을 활용하여 의미있는 정보를 발견할 수 있도록 하는 기술이다(Tan, 1999). 텍스트 마이닝에 관한 초기 연구들은 문헌이나 신문기사와 같이 정형화된 형태를 띈 글을 분석하고자 하는 시도들이 주를 이루었다. 하지만 최근에는 SNS와 블로그와 같이 대중이 실제로 소통하는 공간의 글을 분석하려는 시도가 늘고 있다. 이는 여론을 즉각적으로 파악할 수 있는 유용한 방법으로 인식되어 정치, 사회, 문화적인 이슈 트래킹 연구에 이용되고 있다.

텍스트 마이닝은 선거철마다 여론조사대신 후보자들에 대한 대중의 평판을 조사하고 득표율을 예측하기 위해 많은 관심을 받고 있다. 이는 여론조사의 신뢰성에 대해 많은 사람들이 의문을 표하기 때문이다. 여론조사에서는 자신의 솔직한 응답을 거부하거나 속마음을 잘 드러내지 않는 경우가 많다. 특히 도덕적 판단과 연결된 문제일수록 비난을 모면하고자 솔직한 응답을 꺼린다(Lee et al.,, 2018). 이러한 현상은 2017년 미국 대선에서 잘 드러났는데, 부정적인 시선을 의식해서 표면적으로 트럼프 지지의사를 밝히지 않았던 많은 유권자들이 실제 트럼프에게 투표하였다. 또한 기존의 여론조사는 전화나 언론 매체를 통해서 이루어지기 때문에 전체 유권자 중 일부밖에 반영할 수 없다. 따라서 이에 대한 방안으로 웹 상에서 발생하는 대중의 의견을 수집하여 선거결과 예측에 이용하는 연구가 진행되었다. 트위터와 여론조사와의 상관관계(O's Connor et al., 2010), 실시간 트위터 감성분석을 통한 2012년 미국 대선 후보 여론 분석(Wang et al., 2012), 트위터를 이용한 18대 한국 대선 후보들에 대한 사회적 이슈 분석(Bae et al., 2013), 트위터를 이용한 19대 대선후보 득표율 예측(Lee et al., 2018)등 2010년 이후로 이와 관련한 연구가 활발하게 이루어져 왔다.

지금까지 이루어진 대부분의 연구들은 트위터에서 대중의 의견을 수집하여 분석하였다. 그러나 국내 인터넷 이용자중 트위터 이용자는 2.6%(DMC Media, 2017)에 그쳐 한국 선거 이해를 위한 다량의 데이터 수집부분에서 한계가 드러난다. 그리고 토픽모델링을 이용한 개괄적인 이슈분석 혹은 감성분석을 이용한 평판분석이 독자적으로 이루어져서 선거관련 이슈들을 단면적으로 바라볼 수 밖에 없었다.

따라서 본 연구에서는 기존 연구들의 한계점을 극복하고자 1) 국내 인터넷 플랫폼 집중도를 이해한 후 대중의 의견이 다수 분포해있는 곳에서 데이터를 수집하였다. 네이버뉴스는 국내 포털사이트 및 인터넷 언론매체 중, 인터넷뉴스 이용점유율을 55.4% 차지하고 있어(Opinion Concentration Investigation Committee, 2016) 트위터에 비해 대중의 의견을 더 정확히 반영할 수 있다고 판단하였다. 또한 2) 빈도분석, 연관감성어 분석, 토픽 감성 분석을 복합적으로 사용하여 주요 대선 후보들을 중심으로 사회적 이슈, 민심을 자세히 살펴보았다. 3) 댓글의 멘션수와 감성점수를 결합하여 후보자들의 득표율을 예측하고 실제 득표율과 비교하였다.

본 연구를 통해 유권자와 후보자 모두에게 실질적인 기여를 할 수 있을 것으로 기대한다. 유권자는 후보 별 주요 이슈를 파악하여 투표시 더욱 합리적인 결정을 내릴 수 있으며. 후보자는 자신에 대한 여론을 객관적으로 파악하여 선거 운동 전략에 반영할 수 있을 것으로 기대한다.

2. 문헌연구

2.1 텍스트 마이닝

텍스트 마이닝은 자연어 처리 기반의 비정형화된 대량의 텍스트 데이터로부터 의미 있는 정보의 추출 및 분석을 하는 기술이다(Chakraborty et al., 2013). 이전 데이터 마이닝 연구에서는 주로 숫자로 이루어진 데이터, 즉 정형데이터(Structured data)를 주된 분석 대상으로 삼았다. 그러나 최근에는 텍스트, 음성, 이미지 등과 같은 구조화되지 않은 비정형데이터(Unstructured data) 분석이 주목받고 있다(Chakrabartij, 2002). 컴퓨터의 발달을 통해 많은 양의 텍스트 문서를 저장하고 볼 수 있게 되어 많은 연구자들의 연구 수행이 활발해 지고 있으며, 정보통신 기술의 발달과 인터넷 사용자가 급격히 증가함에 따라, 텍스트 마이닝 (Text Mining) 의 중요성이 더욱이 커져가고 있다(Rebholz-Schuhmann et al., 2005; Choi and Park, 2002). 텍스트 마이닝은 문서를 분류하거나, 텍스트 내의 가치 있는 정보나 패턴들을 추출하는 기술을 제공한다. 대부분의 비정형 데이터는 텍스트로 이루어져 있어, 비정형 데이터 처리를 위해서 텍스트 마이닝은 다양하게 활용된다(Holton, 2009).

Kang et al.(2015)에 따르면 텍스트 마이닝은 다음과 같은 4단계의 과정을 거친다. 첫 번째 과정인 “데이터 수집과정”에서는 대량의 비정형화된 텍스트 데이터를 수집하는 단계이다, 예로 온라인 상의 다양한 데이터를 한 번에 수집할 수 있는 크롤링 혹은 스크래핑 기술이 있으며 이는 프로그래밍 언어로 구현 가능하다. 두 번째는 “용어 추출과정”으로 문장의 단어, 패턴 등을 연관성 분석으로 추출하여 관심 있는 후보 단어를 만들어 내는 과정이다. 다양한 통계적 방법이 적용되며, 문서 혹은 텍스트 내의 용어를 추출하는 방법의 대표적인 예로 TF-IDF 방법이 있다. 세 번째 과정은 “정보 추출과정”으로 문서 내의 사용자에게 필요한 정보를 추출해 내는 것이다. 마지막 네 번째 “정보 분석과정”은 앞서 진행된 과정에서 얻어진 키워드에 대한 분류, 군집화 등의 기법을 적용하여 유용한 정보를 분석한다.

텍스트 마이닝의 연구와 체계가 활발해 짐에 따라, 다양한 주제에 텍스트 마이닝이 활용되어지고 있다. 대표적으로는 문서의 내용을 요약하는 “문서요약”과 텍스트 혹은 문서 내에서 필요한 정보를 탐색하는 “정보검색”, 기존 설문조사나 전문가의 의견을 반영하는 연구들과 다르게 텍스트 문서 자체를 분석하여 보다 객관적인 연구를 하는 방법인 “트렌드 분석”이 있다(Kim et al, 2013; Pai et al., 2013; Cho et al., 2014). 또한, 학술지 논문 키워드 간의 연관관계 연구가 진행되었으며 텍스트 마이닝 기법을 활용하여 특허문서를 분석하거나 특허기술의 경향을 연구하기도 하였다(Cho et al., 2012; Breitzman and Mogee, 2002; Kam et al., 2013; Park et al., 2014). 이외에도 텍스트 작성자의 주관이나, 의견 혹은 감정을 분석하는 감성분석과 단어들의 관계로 네트워크를 형성하고 분석하는 네트워크 분석이 다양한 분야에서 활발하게 연구가 진행 중이다(Liu,

2012; Kim, 2013).

2.2 소셜미디어를 이용한 선거관련 연구

소셜미디어를 이용해 선거의 결과를 예측하거나 유권자의 정치적인 성향을 파악하여 선거 캠페인 도구로 활용하려는 노력이 전세계적으로 활발하게 진행되고 있다(Tumasjan et al., 2010). Williams and Gulati(2008)는 특정 선거 후보자를 지지하는 페이스북 팬의 수가 실제 득표율에 영향을 끼칠 수 있음을 밝혀내었다. Tumasjan et al.(2010)은 트위터에 언급된 후보자들 이름을 분석하여 후보자 이름 언급 빈도와 득표율간에 연관관계가 있음을 밝히고 독일 의회 연방 선거결과를 예측하였다. O'Connor et al.(2010)은 감성분석 기법을 사용하여 트위터가 대중의 의견을 반영하는 도구임을 미국 대선 사전 여론조사와 비교하여 증명하고자 하였다. Livne et al.(2011)은 미국 상, 하의원과 주지사 선거를 대상으로 후보자들이 작성한 트위터만을 분석하여 당선여부를 예측하였다. Chung and Mustafaraj(2011)는 트위터의 언급빈도와 감성분석 기법을 이용해서 미국 매사추세스 상원 선거 결과를 예측하였다. Wang et al.(2012)는 실시간으로 트워터 메시지를 받아와 감성분석을 실시하여 미국 대선후보에 대한 여론을 시시각각 변화하여 보여주었다. Bae et al.(2013)은 트위터를 이용하여 한국 대선

〈Table 1〉 researches for election using social media

research	purpose	data set	method
Williams and Gulati(2008)	Causal relationship between the number of Facebook fans and the rate of votes	Facebo-ok	regressi-on
Tumasjan et al.(2010)	Predicting Federal election results of the German parliament	Twitter	frequency analysis
O's Connor et al.(2010)	Correlation between Twitter and survey	Twitter	sentiment analysis
Livne et al.(2011)	Analyzing Twitter messages created by candidates to determine whether they will win	Twitter	logistic regressi-on
Chung and Mustafaraj (2011)	US Massachusetts Senate election results forecasted by Twitter	Twitter	sentiment analysis
Wang et al.(2012)	2012 US presidential candidate opinion analysis through real-time Twitter sentiment analysis	Twitter	sentiment analysis
Bae et al.(2013)	Analysis of social issues on 18th president candidates for the Republic of Korea using Twitter	Twitter	coocurrence, topic modeling, network analysis
Castro et el.(2017)	Predicting the region of ruling party victory in the Venezuelan election by Twitter	Twitter	Spherical K-Means, TF-IDF
Fenoll et al.(2017)	Comments in Facebook on each party during 2015 Spanish election and their relationship with each party's proponents	Facebo-ok	frequency, sentiment, network analysis
Lee et al.(2018)	Predicting the voting rate of 19th presidential candidates using Twitter	Twitter	frequency, sentiment analysis

후보들에 대한 사회적 이슈를 분석하였다. 분석 시 동시 출현 빈도분석, 다항 토픽 모델링, 네트워크 분석을 사용하여 다각적인 측면에서 이슈를 바라보았다. Castro et el.(2017)은 트위터 데이터를 Spherial K-means, TF-IDF 가중치를 이용하여 분석하였다. 트위터 데이터를 지역별로 구분하여 베네수엘라 선거기간동안 지역별 여당 또는 야당의 우세를 예측하였다. 실제 선거결과와 비교했을 때 87.5%의 정확도를 보여주었다. Fenoll et al.(2017)는 페이스북 이용자들의 2015 스페인 선거기간중 각 정당에 대한 코멘트와 각 정당 지지자들간의 관계를 분석하였다. 빈도분석, 감성분석, 네트워크 분석을 복합적으로 이용하였다. Lee et al.(2018)은 트위터를 이용하여 한국 19대 대선후보의 득표율을 예측하였다. 트위터상에 후보 이름이 언급된 빈도를 산출하여 실제 득표율과 비교하였으며 각 후보들간의 이슈를 감성분석하여 가장 긍정적인 후보가 당선될 것이라 예상하였다. 소셜미디어를 활용한 선거 관련 연구를 정리하면 <Table 1>과 같다.

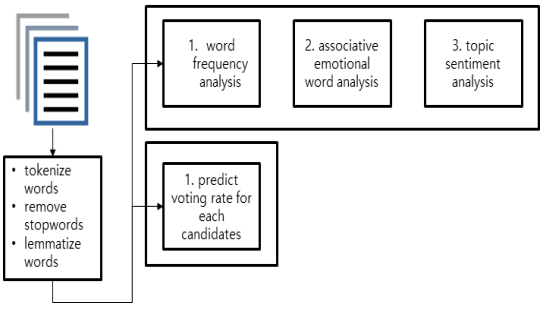
3. 연구 방법

<Figure 1>은 본 연구의 전체적인 연구 방법을 도식화했다. 본 연구는 크게 데이터 각 후보자 이슈 분석 단계와 득표율 예측 단계로 나뉜다.

우선 226,447건의 댓글을 수집하였다. 댓글은 문장별로 구분된 후 형태소분석, 불용어 삭제, 단어 정규화를 거쳐 분석하기 간편한 형태로 전처리하였다.

첫 번째로 단어 빈도분석, 연관 감성어 분석, 토픽 감성 분석을 통해 후보별 이슈를 파악하고 여론이 각 후보에게 긍정적인지 혹은 부정적인지 살펴보았다.

두 번째로 텍스트마이닝 기법 기반 각 후보의 득표율을 예상해본 후 실제 득표율과 비교해 보았다.



〈Figure 1〉 Overview for research methods

3.1 데이터 수집

데이터 수집을 위해 웹 크롤링 기법을 사용하여, 선거 직전 9일 동안(2017년 4월 29일 ~ 2017년 5월 7일)의 네이버의 뉴스 기사 중 '19대 대선'이라는 키워드를 가진 기사의 댓글들을 수집하였다. 댓글 수는 총 226,447건이다. 데이터 수집 기간인 2017년 4월 29일 ~ 2017년 5월 7일은 대선 선거일 6일 전부터 여론조사 결과 공표가 금지되는 공표금지 기간과 맞물려 있다. 하지만 선거 직전의 여론이 가장 선거결과에 근접할 것이라는 판단과 2017년 5월 2일 마지막 TV 대선 토론이 있었기 때문에 이 기간의 여론을 파악하는게 선거결과를 예상하는데 매우 중요하다고 생각했다. 따라서 본 연구에서는 2017년 4월 29일 ~ 2017년 5월 7일의 기간 동안의 댓글을 이용하였다. <Table 2>는 댓글 수집수를 날짜 기준으로 나타내었다.

〈Table 2〉 Number of comments collected by date

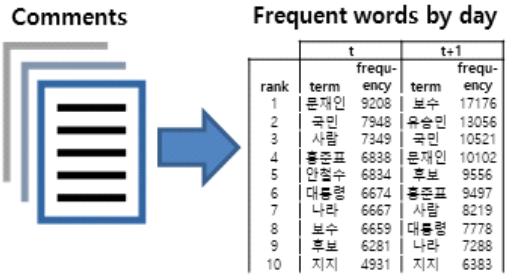
day	4/29	4/30	5/1	5/2	5/3
number of comments	9584	19662	21314	37684	11394
day	5/4	5/5	5/6	5/7	
number of comments	53260	34987	17149	21413	

3.2 데이터 전처리

수집된 데이터는 불용어를 제거한 후 한나눔 한글 형태소 분석기를 이용하여 품사 태깅을 하였다. 한국어 정보검색에서는 문서를 대표하는 색인어 또는 키워드로서 명사를 사용한다(Shin and Lee, 2009). 따라서 명사만을 추출하여 분석 기법을 적용하였다.

3.3 단어 빈도 분석

댓글은 단문으로 개인의 의견이나 생각을 쉽게 공유할 수 있다. 댓글은 남녀노소 누구나 시간과 장소의 제약을 받지 않으며 자신의 생각을 게시할 수 있고 정보를 빠르게 공유할 수 있기에 대중의 소통 공간으로 이미 널리 자리 잡았다. 특히 스마트폰의 대중화로 다수의 댓글 이용자들이 휴대폰을 이용하여 댓글을 공유하고 있다. 이에 장황하게 글을 늘여쓰기보다는 자신이 하고 싶은 말을 핵심 단어 위주로 게재하는 경향이 있다. 이에 본 연구는 텍스트 분석시 간단하면서도 강력한 전통적인 단어 빈도 분석을 이용하였다. 단어 빈도 분석은 단어 인식 분석에서 가장 주요한 분석방법으로 사용되고 있다(Balota and Chumbley, 1984, 1985; Chumbley and Balota, 1984). 많은 실험에서 대부분의 독자들이 낮은 빈도의 단어들을 읽는데 많은 시간을 할애한다고 밝혔다(Inhoff, 1984; Just and Carpenter, 1980; Rayner, 1977). 따라서 수 많은 댓글들 중 가장 많이 언급된 단어들을 집중적으로 살펴보면서 단 시간 내에 여론의 주요 관심 키워드를 알아보았다. 더욱이 댓글 수집기간별 단어 빈도 분석을 각각 수행하면 시간에 따라 변화하는 양상을 한눈에 쉽게 알 수 있다. <Figure 2>에서 본 연구에서 수행한 단어 빈도분석 절차를 나타내었다.



	t		t+1	
rank	term	frequency	term	frequency
1	문재인	9208	보수	17176
2	국민	7948	유승민	13056
3	사람	7349	국민	10521
4	홍준표	6838	문재인	10102
5	안철수	6834	후보	9556
6	대통령	6674	홍준표	9497
7	나라	6667	사람	8219
8	보수	6659	대통령	7778
9	후보	6281	나라	7288
10	지지	4931	지지	6383

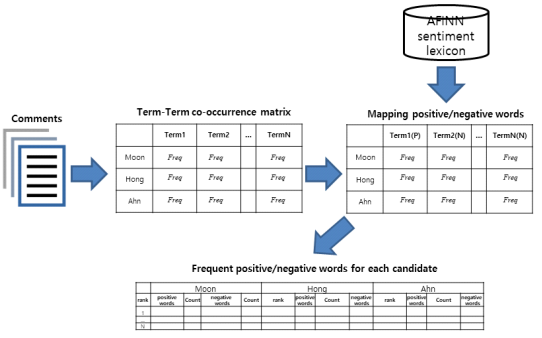
〈Figure 2〉 Frequency analysis by day

3.4 연관 감성어 분석

연관어분석은 두 키워드가 문헌에서 동시에 출현하였을 때 두 키워드가 표현하는 연구 주제가 서로 관련이 있다고 보는 것이다(Liu et al., 2012). 이는 언어학적 의미에서 의미의 근접성을 가리키며, 단어의 연결을 찾는 데 활용되고 있다. 하지만 단순 연관어분석은 키워드간의 의미 연결성을 일일이 해석 해야하는 번거로움이 있고 해석시 주관성이 많이 개입되어 객관성이 떨어진다. 이에 연관어분석 결과를 사전기반 감성분석과 결합하여 연관 키워드의 감성을 도출하였다.

본 연구는 주요 3명의 대선후보(‘문재인’, ‘홍준표’, ‘안철수’)에 대한 연관어들을 추출한 후, 각 연관어들을 감성사전인 AFINN을 이용해 긍

정, 부정으로 분류하였다. 동일 문장 안에 3개의 키워드(‘문재인’, ‘홍준표’, ‘안철수’)와 함께 쓰여진 단어들은 해당 키워드와 연관된 연관어로 간주하였다. 각 연관어들은 구글 번역 API를 이용하여 한글로 변환된 AFINN 감성사전과 대조되어 긍정, 부정어로 분류되었다. 최종적으로 각 후보 별 연관 긍정어, 부정어를 빈도 순으로 나타내어 후보별로 대두되는 긍정이슈, 부정이슈를 살펴보았다. <Figure 3>에서 본 연구에서 수행한 연관 감성어 분석 절차를 나타내었다.



〈Figure 3〉 Associative emotional word analysis

3.5 토픽 감성 분석

토픽 모델링은 문서나 텍스트(Coupus) 내에서 일정한 패턴을 찾아 잠재적으로 의미있는 토픽을 발견하는 절차적 확률 분포 모델이다(Steyver and Griffiths, 2007). 즉, 문헌을 구성하는 단어들이 독립적이지 않다는 가정하에 확률적으로 계산하여 결과 값을 토픽에 해당할 가능성이 높은 단어들의 집합으로 추출하는 알고리즘이다(Lee, 2018). 토픽 모델링 기법 중 하나인 LDA는 Blei(2003)에 의해 제안된 기법으로 데이터의 차원을 축소하는데 유용하며, 의미적으로 일관성이 있는 주제들을 생산한다는 장점을 가지기 때문에 텍스트 분석에서 인기 있는 모델로서 사용되어 왔다(Mimno et al., 2008).

토픽 모델링은 정형화되지 않은 대량의 데이터 처리 및 분석이 요구되면서 다양한 분야에서 잠재적으로 결과를 찾기 위해 감성분석과 더불어 자주 활용되고 있다(Xianghua et al., 2013). 감성분석은 문서나 텍스트 내에서 사람들의 태도, 의견, 성향 등 감정과 같은 주관적인 데이터를 분석함으로써 실증적으로 연구하는 자연어 처리 기술로(Liu, 2012), 기존의 오프라인 여론조사에 비해서 시간과 비용을 줄이고 사람들의 의견을 쉽게 파악하고 예측할 수 있어 그 활용도가 매우 높은 분야로 평가 받고 있다(Jang, 2014). 사람이 최종적으로 판단한 자료를 기반으로 분석을 진행하기 때문에 글의 맥락에 따라 감성사전을 구축하여 수행하는 어휘기반 접근방식과 기계학습 기반의 방법론들이 주로 사용된다(Tan et al., 2009; Mullen and Nigel, 2004).

본 연구는 토픽모델링으로 도출된 대선 후보 관련 토픽들이 긍정적인 내용인지 부정적인 내용인지 또는 어느 정도 긍정적이고 부정적인지 그 정도를 알기 위해 LDA 토픽모델링과 감성분석을 결합한 토픽 감성 분석 수행하였다. 토픽 감성 분석을 위해서 우선적으로 수집한 댓글을 LDA 토픽 모델링한다. 그 후 AFINN의 단어별 감성 점수를 채택하여 각 토픽별 감성 점수를 합산한다. 본 연구는 각 토픽별 단어의 확률 값을 단어 가중치로 설정하였다. 본 연구에서 토픽 감성 점수를 <Formula 1>과 같이 산출하였다.

$$Topic\,A\,\text{감성점수} = \sum_i \text{Prob}(word_i|Topic_A) \times AFINN(words)\,\text{점수}$$

〈Formula 1〉

토픽별 감성점수는 AFINN 감성점수와 각 토픽의 단어 확률 값을 곱하여 모두 더한 값으로 계산된다. 같은 단어가 여러 토픽에 서로 다른 확률 값으로 존재할 수 있다. 따라서 만약 토픽별 단어셋이 같더라도 분포 확률이 다르기 때문에 감성점수는 다르게 산출된다. 예를 들어 <Table 3>을 살펴보면 '희망'이라는 단어는 토픽1, 토픽2에 모두 포함되었지만 각 토픽에 존재할 확률이 다르다. 토픽1과 토픽2의 전체 감성점수는 각각 -0.111, 0.2289로 토픽1은 부정적 토픽, 토픽2는 긍정적 토픽으로 분류되었다.

〈Table 3〉 Example of calculating topic emotion score

term	probab-ility of topic 1	probab-ility of topic 2	sentim-ent score	sentim-ent score of topic 1	sentim-ent score of topic 2
보수	0.0322		-1	-0.0322	
거짓말	0.0301		-4	-0.120	
희망	0.0103	0.0150	4	0.0412	0.06
개혁		0.0240	3		0.072
자유		0.0323	3		0.0969
total score				-0.111	0.2289

3.5 득표율 예측

선거에서 항상 중심이 되는 이슈는 각 후보의 득표율을 예상하는 것이다. 따라서 각 기관은 시시각각 여론조사 결과를 발표하며 선거 전 각 후보의 득표율을 예측한다. 하지만 선거 직전 기간은 여론조사 공표금지 기간이어서 이 기간의 여론조사는 금지되어있다. 후보자들간의 경합이 치열할때에는 선거일 직전까지의 예측 득표율이 초미의 관심사이며 후보자들도 예측 득표율을 막판 선거캠페인 전략에 반영할 수 있을 것이다. 따라서 본 연구는 선거 직전 기간의 텍스트 데이터를 이용하여 각 후보의 득표율을 예측하였다. 여태까지 대부분의 연구는 언급 빈도 또는 팬의 수, 즉 데이터 볼륨만으로 후보자의 인기를 판단하고 이에 기반하여 득표율을 예측하였다. 예를 들어 후보자별 페이스북 팬의 수나 트윗에서의 언급 정도만을 파악할 뿐이었다(Williams and Gulati., 2008; Chung and Mustafaraj, 2011). 물론 인기 있는 후보에 대한 데이터 볼륨이 인기 없는 후보에 비해서 높을 수 있다. 하지만 인기와 무관하게 부정적 이슈를 지니고 있거나 보편적인 후보들과는 다른 특이한 이력을 지닌 이슈메이커 후보들도 데이터 볼륨이 높게 나올 수 있다. 따라서 데이터 볼륨과 함께 텍스트 데이터 자체의 긍정적, 부정적 정도도 득표율 예측의 가중치로 추가해야 한다. 소셜미디어에서의 데이터 볼륨과 감성정도는 사회적 실제 및 이슈를 반영한다는 기존의 연구는 이러한 사실을 뒷받침한다. (Vergeer et al., 2011). 따라서 본 연구는 후보별 댓글의 분포와 감성점수를 결합하여 예상 득표율을 산출하였다. 각 후보의 득표율은 다음 <Formula 2>와 같이 산출하였다. 전체 후보는 TV 대선 토론에 참여한 문재인, 홍준표, 안철수, 유승민, 심상정 후보로 한정하였다.

$$\text{예상후보}A\text{득표율} = \frac{\text{후보}A\text{ 댓글수} \times \frac{\text{후보}A\text{긍정감성점수}}{|\text{후보}A\text{부정감성점수}|}}{\text{전체후보 댓글수} \times \frac{\text{전체후보긍정감성점수}}{|\text{전체후보부정감성점수}|}} \times 100$$

〈Formula 2〉

4. 데이터 분석 결과

4.1 단어 빈도 분석 결과 및 해석

수집 기간 총 9일간 출현한 단어의 빈도 분석 결과는 <Table 4>와 같다. 분석 결과 '문재인' 후보가 다른 후보에 비해 전체적으로 뉴스기사 댓글에서 가장 많이 언급된 단어로 나타났다.

비교해보면 '문재인'은 9일간 꾸준히 선두권을 지키고 있으며 '홍준표'는 2일 전까지는 순위가 오르다가 2일 이후로 내려갔다 올라오는 양상을 보이고 있다. '심상정'은 2일까지 순위가 가파르게 상승하다가 2일 이후로는 순위권 밖으로 밀려났다. '유승민'은 꾸준히 10위권을 유지하며 그 추세를 이어가지만 5월 2일 대선토론일에 한번 급격한 상승을 한다. 한편 '안철수'는 순위 변동 폭이 크고 그 방향이 계속 바뀌고 있다. 날짜별 단어 빈도 분석 결과는 2일 이전과 이후로 패턴이 변화하는 경향이 있다. 2일은 후보자간 마지막 대선 후보 토론이 방영된 날로 그 날의 토론 결과에 따라 분석결과의 패턴이 변화하는 것을 알 수 있다.

4.2 연관 감성어 분석 결과 및 해석

날짜 별로 주요 후보에 대한 연관 감성어 분석을 진행하였다. 도출된 전체 연관된 감성어 중 빈도수 상위 10개를 <Table 5>와 같이 나타내었다. 문재인 후보는 긍정단어 리스트에 유력 당선 후보로서 '지지', '성공', '승리'와 같이 대선당선

〈Table 4〉 Word frequency analysis result

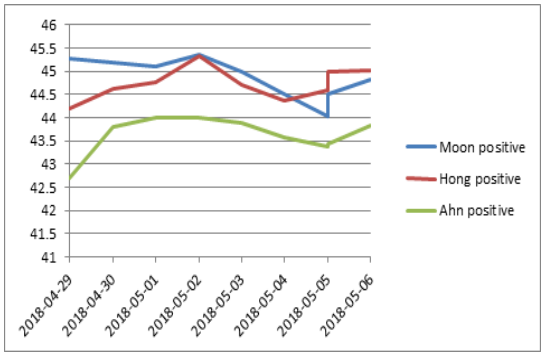
	4/29	4/30	5/1	5/2	5/3	5/4	5/5	5/6	5/7
1	대통령	대통령	문재인	홍준표	대통령	투표	투표	안철수	대통령
2	문재인	홍준표	대통령	대통령	홍준표	안철수	대통령	대통령	문재인
3	사람	문재인	홍준표	유승민	문재인	대통령	사람	문재인	안철수
4	촛불	사람	안철수	문재인	사람	문재인	문재인	홍준표	홍준표
5	후보	안철수	투표	보수	후보	사람	보수	사람	사람
6	국민	국민	사람	안철수	유승민	홍준표	국민	후보	후보
7	안철수	후보	국민	국민	안철수	국민	안철수	국민	국민
8	진짜	투표	후보	생각	생각	후보	홍준표	생각	투표
9	생각	생각	박근혜	박근혜	국민	보수	후보	진짜	보수
10	투표	나라	나라	심상정	진짜	대한민국	진짜	대한민국	유승민
11	심상정	대한민국	진짜	투표	수능	생각	대한민국	투표	대한민국
12	홍준표	진짜	생각	대한민국	투표	진짜	생각	나라	생각
13	나라	보수	대한민국	지지율	보수	유승민	나라	보수	나라
14	박근혜	박근혜	여론조사	나라	박근혜	박근혜	박근혜	공약	우리
15	우리	국민들	보수	토론	공약	나라	유승민	국민들	진짜
16	아들	지지율	공약	국민들	대한민국	사전투표	사전투표	유승민	박근혜
17	국민들	공약	지지율	정치	나라	대구	국민들	심상정	아들
18	대한민국	당선	국민들	대선	국민들	정치	투표율	박근혜	당선
19	공약	북한	심상정	여론조사	세월호	미래	정치	문준용	국민들
20	정치	안철수	당선	공약	정치	심상정	당선	당선	정치

과 관련된 단어가 주로 분포하였다. 반면 부정단어 리스트에는 '조작', '가짜'와 같이 문재인 후보의 진실성과 이중성에 대한 의혹이 주로 분포하였다. 홍준표 후보는 긍정리스트에 '보수', '능력'과 같이 능력있는 보수 후보라는 점이 긍정리스트에 분포하였다. 반면 부정단어 리스트에는 '쓰레기', '강간', '뇌물'과 같이 사회적 논란거리였던 홍준표 후보 자서전의 대학시절 이야기와 성완종 리스트 사건이 주로 분포하였다. 안철수 후보는 긍정단어 리스트에 '진정', '진심', '정의'와 같이 진실되고 깨끗한 후보라는 점이 긍정 리스트에 분포하였다. 반면 부정단어 리스트에는 '거짓말', '가짜', '실망'과 같이 안철수 후보 부인의 교수임용 의혹과 관련하여 부정적인 단어가 분포한 것으로 풀이된다.

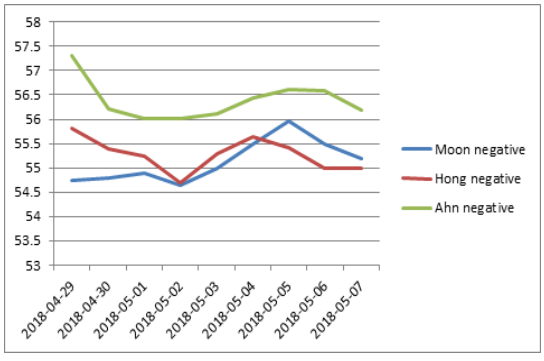
〈Table 5〉 Associative emotion list by candidate

candi-date	Moon Jae-in		Hong Joon-pyo		Ahn Cheol-soo	
senti-ment	positive	negative	positive	negative	positive	negative
1	지지	조사	지지	문제	지지	조사
2	사람	문제	사람	조사	사람	이상
3	지지자	반대	보수	죄인	지지자	거짓말
4	성공	조작	자유	쓰레기	진정	선동
5	정의	거짓말	정신	거짓말	진심	걱정
6	자유	부정	지지자	조작	능력	부족
7	승리	전투	이해	범죄	희망	가짜
8	이해	선동	정의	잘못	기회	사기
9	필요	가짜	능력	강간	승리	실망
10	세상	잘못	기회	뇌물	정의	배신

<Figure 4>, <Figure 5>에서는 날짜별로 분리된 결과에서 긍정과 부정에 해당하는 감성어 빈도를 그래프로 나타내었다. 홍준표 후보는 2일 토론회 이 후 이틀정도 패턴양상이 잠깐 바뀌었지만 그 외 날짜에서는 다른 후보에 비해 긍정어 비율이 높다. 홍준표 후보가 선거 전 여론조사 때에는 안철수 후보에게 뒤지다가 실제 대선에서 안철수 후보를 이긴 것을 볼 때에, 홍준표 후보에 대한 대중의 긍정적 반응 증가가 실제 투표결과에 영향을 미친 것으로 보인다. 특이한 점은 세 후보 모두 9일 전일에 걸쳐 부정어 비율이 긍정어 비율보다 많다. 이는 익명성이 보장된 소셜미디어상에서 자극적인 단어를 이용하여 후보자들을 비방하는 글이 많기 때문에 이러한 결과가 나온 것으로 해석된다.



〈Figure 4〉 Trend of positive words



〈Figure 5〉 Trend of negative words

4.3 토픽 감성 분석 결과 및 해석

<Table 6>은 토픽 감성 분석 결과를 나타내었다.

긍적적인 토픽은 크게 2가지가 도출되었다. '적폐청산' 토픽에서는 유력 대선 후보의 당선 기대와 새로운 정권에 대한 희망을 언급하고 있다. '공약' 토픽에서는 대선 후보들의 공약 내용을 다루고 있다. 특히 '일자리', '정규직', '노동자'와 같이 실업률 해소에 대한 공약이 많은 부분을 차지하고 있다.

부정적인 토픽은 크게 4가지가 도출되었다. '문재인후보 아들 취업 특혜 논란' 토픽에서는 아들 문준용씨의 취업 특혜에 대해서 다루고 있다. '외교/안보' 토픽에서는 미국, 중국, 북한에 둘러싸인 불안한 한반도 상황을 다루고 있다. '홍준표 후보 자서전 내용 논란' 토픽은 홍준표 후보의 대학 시절 성폭행 미수 의혹에 대하여 다루고 있다. '사전투표용지 논란' 토픽에서는 사전투표용지가 용지마다 조금씩 다르게 인쇄되어 부정투표 의혹이 짙어졌던 논란을 다루고 있다. 부정 감성점수를 살펴보면 '문재인 후보 아들 취업 특혜 논란'과 '홍준표 후보 자서전 논란'이 부정지수 1, 2위를 하고 있음을 알 수 있다. 이를 통해 각 후보들의 비도덕성이 대중들에게는 매우 안좋게 인식되고 있음을 알 수 있다.

〈Table 6〉 Topic emotional analysis result

topic	sentiment	sentiment score	key words
change regime	positive	0.116	문재인, 압도적, 1번, 정권교체, 적폐청산, 승리, 이명박, 지지, 압승, 당선, 이명박근혜
election promise	positive	0.005	세금, 공무원, 서민, 공약, 일자리, 교육, 복지, 정규직, 경제, 증세, 노동자
controversial issue on Moon	negative	-0.247	아들, 문준용, 취업, 공기업, 특혜, 해명, 죄인, 귀걸이, 아버지, 의혹, 채용, 이력서
Diplomacy / Security	negative	-0.094	북한, 미국, 전쟁, 김정은, 중국, 주적, 개성공단, 불안, 사드, 공산당, 적화통일
controversial issue on Hong	negative	-0.099	대구, 쓰레기, 돼지발정제, 경북, 양아치, 가짜보수, 여성부, 자한당
controversial issue on advance ballot	negative	-0.014	투표용지, 개표, 선관위, 투표함, 출구조사, 부정개표, 수개표

4.4 득표율 예측 결과 및 해석

<Table 7>은 득표율 예측 결과를 나타내었다. 본 연구에서 제안한 방법으로 득표율을 예측한 결과 실제 득표율과 후보 순위가 동일하였다. 득표율도 문재인후보가 실제보다 다소 적게 예측되었고 대신에 유승민 후보가 실제보다 다소 높게 예측된 것을 제외하고는 다른 후보들은 실제 득표율과 매우 유사하였다. 유승민 후보가 예측 득표율에서 높게 측정된 이유는 마지막 대선 토론회에서 선전하면서 긍정적 댓글 수가 급격히 증가했기 때문이라고 생각한다.

〈Table 7〉 Estimation results

rank	1	2	3	4	5
predicted rate	Moon (33.94%)	Hong (25.65%)	Ahn (20.85%)	Yoo (13.39%)	Sim (6.15%)
true rate	Moon (41.08%)	Hong (24.03%)	Ahn (21.41%)	Yoo (6.76%)	Sim (6.17%)

5. 결론

5.1 연구결과 토의

소셜미디어를 이용해 선거의 결과를 예측하거나 유권자의 정치적인 성향을 파악하여 선거 캠페인 도구로 활용하려는 노력이 전세계적으로 활발하게 진행되고 있다(Tumasjan et al., 2010). 그러나 기존연구들은 주로 트위터 데이터를 수집하여 연구를 수행하였으며 토픽 모델링 기법을 이용한 개괄적인 수준에서의 분석에 그쳤다.

이에 본 연구는 대한민국 최대 인터넷 포털 사이트 네이버의 대선관련 특집 기사의 댓글을 수집하였다. 그리고 단어 빈도 분석, 연관 감성어 분석, 토픽 감성 분석, 득표율 예측을 복합적으로 수행하여 주요 대선 후보자들에 대한 사회적 이슈, 민심을 자세히 살펴보았다. 2017년 4월 29일부터 2017년 5월 7일까지 9일간 총 226,447건의 댓글을 수집, 분석하여 다음과 같은 결과를 도출할 수 있었다.

첫째, 마지막 TV 대선 토론회였던 5월 2일의 토론이 뜨거운 관심을 받았고 그 결과에 따라 후보자 단어빈도와 연관 감성어 트랜드 패턴이 변화하였다. 대선 토론회 결과 유승민 후보는 A, 문재인 후보는 B+, 안철수 후보는 C, 홍준표, 심상정 후보는 C 평점을 받았다(중앙일보 라이브팩트체크팀, 2017). 평점 A를 받은 유승민 후보는 5월 2일 단어 빈도 분석 결과 후보자들 두번째에 위치하여 대중의 뜨거운 관심을 증명하였다. 이와 같은 결과는 TV 대선 토론이 직접적으로 여론에 큰 영향을 끼친다고 해석할 수 있다. 하지만 대선 토론회 이 후 곧 다시 이전과 같은 패턴으로 돌아가는 것을 볼 때에 그 영향력이 시간적으로 오래가지 않음을 알 수 있다. 즉 TV 대선 토론회가 대중들의 뜨거운 관심을 받고 토론회 결과에 따라 각 후보자들에 대한 민심이 요동치지만 대중들의 본질적인 각 후보자들에 대한 생각은 바뀌지 않는다고 해석할 수 있다.

둘째, 각 후보자들의 도적적인 측면에 대한 논란이 후보자 자신들에게 치명적일 수 있다는 것을 확인하였다. 문재인 후보는 아들 문준용씨의 취업 특혜를 필두로 한 거짓말 논란, 홍준표 후보는 자서전에 쓰여진 성폭력 미수 혐의가 토픽 감성 분석에서 대두되었다. 대중들이 각 후보자들을 평가할 때에 다른 측면 보다 도덕적 문제 측면에 대해서 민감하게 반응한다는 것을 확인하였다. 각 후보자들이 과거 저지른 잘못이 선거기간 내내 꼬리표처럼 달라붙어 치명적인 약점으로 작용하였다.

셋째, 새 정권에 대한 기대와 각 후보자들의 공약관련 댓글들도 많이 분포하였다. 새 정권에 대한 기대는 ‘국민’, ‘나라’, ‘대한민국’과 같은 단어들로 표현되었으며 공약 관련 단어는 ‘일자리’, ‘정규직’, ‘서민’, ‘공무원’과 같은 단어들로 표현되었다. 이는 새 정권이 열 새로운 시대에 대한 대중들의 기대와 희망이 응축된 것으로 해석된다.

넷째, 19대 한국대선 당선자를 추측할 수 있다. 댓글의 멘션수와 감성점수를 결합하여 5명의 후보의 득표율을 예측한 결과 실제 득표 결과와 순위가 일치함을 밝혀내었다. 특표율을 자세히 살펴보아도 문재인, 유승민 후보를 제외한 나머지 후보들은 실제 득표율과 1%이내의 오차를 보이며 매우 좋은 결과를 나타내었다.

5.2 학술적 실무적 시사점

본 연구는 텍스트 마이닝 분석을 통해 19대

대선후보들에 대한 정치, 사회, 문화적인 이슈 트래킹을 수행하고 그에 대한 민심을 파악하는 것을 목적으로 하고 있다. 앞서 제시한 연구 결과를 토대로 다음과 같은 시사점을 제시할 수 있다.

학술적 시사점으로는 첫째, 국내 최대 포털사이트인 네이버 뉴스의 댓글을 분석했다는 점에서 의의가 있다. 기존 텍스트 마이닝 연구들은 대부분 트위터, 페이스북에서 데이터를 수집하여 분석하였다(He et al., 2013). 하지만 국내 인터넷 플랫폼 점유율을 고려하였을 때, 네이버 뉴스기사가 인터넷 뉴스 이용점유율 55.4%로 압도적인 1위를 차지하고 있다(Opinion Concentration Investigation Committee, 2016). 따라서 네이버의 뉴스기사 댓글을 수집하여 분석하는 것이 훨씬 많은 사람들의 의견을 살펴볼 수 있어 분석결과의 타당성을 뒷받침할 수 있다.

둘째, 주요 대선 후보자와 관련된 주요 이슈 및 여론을 시간에 따라 트래킹 하였다. 기존 연구들은 대부분 토픽 모델링 기법을 이용한 개괄적인 수준에서의 분석에 그쳤다(Bae et al., 2013). 본 연구는 주요 대선 후보자들을 중심 연구 대상으로 선정하고 그들에 대한 이슈와 여론을 집중적으로 살펴보았다.

셋째, 각 이슈를 정량적인 지표와 연결하여 객관적인 관점에서 민심을 살펴보았다. 기존 연구 토픽 모델링 기법들은 같은 토픽에 속한 단어뭉치 선별과 각 문서들이 각 토픽에 속하는 확률값을 추론하고 있다. 하지만 토픽 모델링 결과에 대한 해석은 여전히 인간의 영역으로 남아있다(Chang et al., 2009). 이에 본 연구는 각 주제마다 감성점수를 산출하여 보다 객관적인 관점에서 각 토픽별 민심을 정량적 지표로 살펴보았다.

넷째, 텍스트 마이닝 기반으로 각 후보자들의 득표율을 예측하였다. 선거직전기간은 여론조사 공표 금지기간과 맞물려 있다. 하지만 이 시기가 가장 투표일에 맞닿아 있어 선거결과 예측에 큰 의미를 줄 수 있으며 대중들의 관심도 가장 높은 시기이다. 따라서 본 연구는 이 시기의 댓글을 수집하여 득표율을 예측하여 대중들이 가장 궁금해하는 점을 보여주었으며 그 결과도 실제 결과와 매우 유사하게 나왔다.

실무적 측면에서의 시사점으로는 첫째, 유권자들이 각 후보자들에 대한 이슈와 여론을 객관적으로 살펴볼 수 있도록 하여, 투표할 때에 보다 합리적인 결정을 내리도록 돕는다. 국정의 최고 책임자를 선출하는 대통령선거에서 유권자에게 주어지는 선거관련 정보는 매우 중요하다. 정보가 충분하지 않다면 유권자는 특정 후보자나 정당에 의해 제시되는 일방적인 정보에 의해 지지 후보의 선택이 좌우될 수 있기 때문이다(Yoo, 2008). 하지만 유권자들은 주로 자신의 주위에 있는 가족이나 친구 등과 같은 지인들의 후보자에 대한 평판이나 추천에 의해 후보자 결정에 영향을 받는다(Jung, 2014). 본 연구 결과를 투표전에 참고한다면 빅데이터로부터 대중들의 의견을 볼 수 있어 보다 객관적인 안목으로 후보자들을 평가하고 투표할 수 있을 것이다.

둘째, 후보자들 각각 자신들에 대한 여론을 객관적으로 파악하고 이에 따라 선거 전략을 짤 수 있을 것이다. 특정 후보의 개인적, 정책적 요인이나 캠페인 전략의 성패와 같은 주체적 요인은 지지율 변화의 원인으로 제시하는 경우도 있고, 경쟁 후보의 실수나 공격적 캠페인 같은 환경적 요인이 그 원인으로 제시하는 경우도 있다(Song et al., 2008). 후보자들은 자신들에 대한 대중들의 의견을 겸허히 받아들여 긍정적 이슈들은 더욱 적극적으로 선거 전략에 활용하고 부정적 이

슈는 보완, 해결해야 할 것이다. 1982년부터 1990년까지의 미국 의회 선거결과를 분석한 결과, 도덕적 문제로 평판에 타격을 입은 후보자들은 평균적으로 득표할 수 있는 득표율에서 약 10%의 손실을 보는 것으로 나타났다(Welch and Hibbing, 1997). 따라서 특히 도적적 측면에서의 문제점은 선거기간 내내 치명적인 약점이 될 수 있음을 인지하고 그에 대한 대처를 분명히 해야 할 것이다.

셋째, 국민의 정치 참여 행위로써 의의가 있다. 자신들의 정치적 견해를 표현하는 시민의 능력은 다양한 방식으로 나타날 수 있다(Ferber et al., 2005). 특히 현실세계에서는 거의 기회를 갖지 못하는 사회적 약자들에게 사이버 공간은 분명 정치과정에 참여할 수 있는 새로운 문화가 될 수 있다(Ha, 2006). 후보자들이 빅데이터 분석을 참조하여 선거캠페인을 벌인다면 국민들은 자신들이 원하는 바가 후보자들에게 피력, 반영된다는 것을 인지하고 웹상에서 더욱 적극적인 활동을 할 것이다.

5.3 연구 한계점 및 향후 연구 방향

본 연구는 네이버 뉴스의 댓글들을 수집한 후, 텍스트 마이닝 기법을 이용하여 19대 대선 후보자들에 대하여 집중적으로 살펴보았다. 하지만 수집한 데이터와 분석에 대하여 다음과 같은 한계점을 지니며, 이는 향후 연구에서 보완 되어야 할 것이다.

첫째, 네이버 뉴스 댓글 기사에 댓글을 다는 작성자들이 전체 선거 유권자를 대표한다고 볼 수 없다. 네이버 뉴스 댓글 작성자 통계를 살펴보면 남성이 80.9%, 여성이 19.1%를 보여 큰 차이를 보였다. 또한 연령대별로도 30대가 32.0%, 40대가 27.3%, 20대가 19.7%를 보였다(Yonhap News, 2016). 즉 네이버 댓글을 이용한 대선 관련 연구는 표본 선택에 있어서 표본이 모집단을 올바르게 대표한다고 볼 수 없다. 향후 연구는 데이터 수집시 인구통계학적 요소를 고려하여 표본을 추출하는 것이 필요하다.

둘째, 데이터 수집 기간이 9일에 그쳤다. 물론 그 9일이 여론조사 공표 금지 기간과 마지막 대선 후보 TV 토론과 맞물려있는 기간이라는 점에서 의의가 있다. 하지만 보다 긴 기간의 데이터를 수집했더라면 각 후보별 더욱 다양한 이슈를 살펴볼 수 있었을 것이다. 또한 여론조사결과와 비교하며, 각 후보별 이슈가 실제로 여론에 어떠한 영향을 끼치는지 확인할 수 있어 더욱 실용적인 연구가 되었을 것이다. 향후 연구에서는 충분한 기간의 댓글을 수집하는 것이 필요하다.

참고문헌(References)

Bae. J. H., J. E. Son, and M Song, "Twitter analysis of 2012 presidential elections using text mining", Intelligence Information Research, Vol. 19, No.3(2013), 141-156.

Balota, David A., and James I. Chumbley. "Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage," Journal of Experimental Psychology: Human perception and performance, Vol. 10, No. 3(1984), 340.

Balota, David A., and James I. Chumbley. "The locus of word-frequency effects in the pronunciation task: Lexical access and/or production?.," Journal of Memory and Language, Vol. 24, No. 1(1985), 89-106.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning, research 3(2003), 993-1022.

Breitzman. A. F, and Mogee. M. E, "The many applications of patent analysis", Vol. 28 (2002), 187-205.

Castro, Rodrigo, Leonardo Kuffó, and Carmen Vaca., "Back to# 6D: Predicting Venezuelan states political election results through Twitter," eDemocracy and eGovernment (ICEDEG), 2017 Fourth International Conference(2017), 148-153.

Chumbley, James I., and David A. Balota, "A word's meaning affects the decision in lexical decision," Memory and Cognition, Vol. 12, No. 6(1984), 590-606.

Chakrabarti, Soumen. "Mining the Web: Discovering knowledge from hypertext data.", Elsevier(2002).

Chakraborty, Goutam, Murali Pagolu, and Satish Garla. "Text mining and analysis: practical methods, examples, and case studies using SAS.", SAS Institute(2014).

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M., "Reading tea leaves: How humans interpret topic models.," In Advances in neural information processing systems, (2009), 288-296.

Cho. G. H, Lim. S. Y, and Hur. S, "An Analysis of the Research Methodologies and Techniques in the Industrial Engineering Using Text Mining", Journal of the Korean Institute of Industrial Engineers, vol. 40, No. 1(2014), 52-59.

Cho. S. G, and S. B. Kim, "Finding Meaningful Pattern of Key Words in IIE Transactions Using Text Mining", Journal of the Korean Institute of Industrial Engineers, Vol. 38, No. 1(2012), 67-73.

Choi, Y. J., and S.S. Park, "Interplay of text mining and data mining for classifying web contents.", Korean Journal of Cognitive Science Vol. 13, No. 3(2002), 33-46.

Chung, Jessica Elan, and Eni Mustafaraj., "Can collective sentiment expressed on twitter predict political elections?.," AAAI. Vol. 11 (2011), 1770-1771.

DMC Media, "2017 Social Media Usage Behavior and Ad Contact Attitude Analysis Report", DMC Report, 2017.07.10.

Fenoll, Vicente, and Lorena Cano-Orón. "Citizen engagement on Spanish political partie's Facebook pages: Analysis of the 2015 electoral campaign comments," Communication and Society, Vol. 30, No. 4 (2017).

Ferber, Paul, Franz Foltz, and Rudy Pugliese. "The internet and public participation: state legislature web sites and the many definitions of interactivity," Bulletin of Science, Technology and Society, Vol. 25, No. 1(2005), 85-93.

Ha. J. W., "A Study on Internet Politics Participation of College Students", Korean Press Information(2006), 369-405.

He, Wu, Shenghua Zha, and Ling Li, "Social media competitive analysis and text mining: A case study in the pizza industry." International Journal of Information Management, Vol. 33, No. 3(2013), 464-472.

Holton. C, "Identifying disgruntled employee systems fraud risk through text mining: A simple solution for a multi-billion dollar

problem.", Decision Support Systems, Vol. 46, No. 4(2009), 853-864.

Inhoff, and Albrecht Werner. "Two stages of word processing during eye fixations in the reading of prose," Journal of verbal learning and verbal behavior, Vol. 23, No. 5(1984), 612-624.

Jang. P. S., "Research on the main emotional analysis of social data", Journal of the Korea Computer Information Society, Vol. 19, No. 12(2014), 49-56.

Jung. I. T., "The Effect of Voter's Use of Social Media on the Determinants of Voting", Journalism Research, Vol. 18, No. 4(2014), 239-278.

Just, Marcel A., and Patricia A. Carpenter. "A theory of reading: From eye fixations to comprehension," Psychological review, Vol. 87, No. 4(1980), 329.

Kam. J. S, Kim. M. W, and B. H. Hyun, "A Study on Analysis of Patent Information Based Biotechnology Research Trend and Promising Research Themes", The Korea Society for Innovation Management and Economics, Vol. 21, No. 2(2013), 25-56.

Kang. B. G, M. Y. Huh, and S. B. Choi, "Performance analysis of volleyball games using the social network and text mining techniques.", Journal of the Korean Data and Information Science Society Vol. 26, No. 3(2015), 619-630.

Kim. H. Y, "Analysis of an Inaugural Address of Korean Presidents Based on Network", Korea Content Association, Vol. 3, No. 2(2013), 67-68.

Kim. M, Notkin. D, Grossman. D, and Wilson. G, "Identifying and summarizing systematic code changes via rule inference", Software Engineering, IEEE Transactions on, vol. 39(2013), 45-62.

Lee, S. G., "Study on the Improvement of e-Learning Satisfaction based on Text Mining", Yonsei Univ Master Thesis(2018).

Lee, Y, N., E. J. Choi, and M. J. Kim, "Analysis of the effects of presidential candidates' SNS reputation on election results", Digital fusion research, Vol. 16, No. 2(2018), 195-201.

Liu, B., "Sentiment analysis and opinion mining," Synthesis lectures on human language technologies, Vol. 5, No. 1(2012), 1-167.

Liu，G. Y.，Hu, J. M., and Wang, H. L., "A co-word analysis of digital library field in China. Scientometrics,", Vol. 91, No. 1(2012), 203-217.

Livne, A., Simmons, M. P., Adar, E., and Adamic, L. A., "The Party Is Over Here: Structure and Content in the 2010 Election," ICWSM, Vol. 11(2011), 17-21.

Mimno, David, Hanna Wallach, and Andrew McCallum." Gibbs sampling for logistic normal topic models withgraph-based priors," NIPS Workshop on Analyzing Graphs. Vol. 2008(2008).

Mullen, Tony, and Nigel Collier. "Sentiment analysis using support vector machines with diverse information sources," Proceedings of the 2004 conference on empirical methods in natural language processing. 2004.

O'Connor, Brendan, Routledge, B. R., and Smith, N. A, "From tweets to polls: Linking text sentiment to public opinion time series," Icwsm, Vol. 11, No.122-129(2010), 1-2.

Opinion Concentration Investigation Committee,

"Opinion concentration survey results", 2016.01.21.

Pai. M. Y, Chen. M. Y, Chu. H. C, and Chen. Y. M, "Development of a semantic-based content mapping mechanism for information retrieval", Expert Systems with Applications, vol. 40, No. 7(2013), 2447-2461.

Park. H, Seo. W, Coh. B, Lee. J, and J. Yoon, "Technology Opportunity Discovery Based on Firms' Technologies and Products", Journal of the Korean Institute of Industrial Engineers, Vol. 40, No. 5(2014), 442-450.

Rayner, and Keith. "Visual attention in reading: Eye movements reflect cognitive processes," Memory and Cognition, Vol. 5, No. 4(1977), 443-448.

Rebholz-Schuhmann, Dietrich, Harald Kirsch, and Francisco Couto, "Facts from text—is text mining ready to deliver?." PLoS biology, Vol. 3, No. 2(2005).

Shin. S. W., and Y. W. Lee, "Noun and Keyword Extraction for Korean Information Processing", Journal of the Korea Computer Information Society, Vol. 14, No.3(2009), 51-56.

Song. H. J., H. S. Kim, and W. J. Lee, "The Impact of Cognitive Appraisal and Emotional Response on Political Behavior", Korean Media Scholarship, Vol. 51, No. 4(2008), 353-376.

Steyvers, Mark, and Tom Griffiths., "Probabilistic topic models," Handbook of latent semantic analysis, Vol. 427, No. 7(2007), 424-440.

Tan, and Ah-Hwee., "Text mining: The state of the art and the challenges," In: Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases. Vol. 8(1999), 65-70.

Tan, S., Cheng, X., Wang, Y., and Xu, H., "Adapting naive bayes to domain adaptation for sentiment analysis," European Conference on Information Retrieval(2009), 337-349.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M., "Predicting elections with twitter: What 140 characters reveal about political sentiment," Icwsm, Vol, 10, No. 1(2010), 178-185.

Vergeer, M., Hermans, L., and Sams, S. "Is the voteronly a tweet away? Micro blogging during the 2009 European Parliament election campaign in the Netherlands." First Monday, Vol, 16, No. 8(2011).

Wang, H., Can, D., Kazemzadeh, A., Bar, F., and Narayanan, S., "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle," In Proceedings of the ACL 2012 System Demonstrations, Association for Computational Linguistics(2012), 115-120.

Welch, Susan, and John R. Hibbing., "The effects of charges of corruption on voting behavior in congressional elections, 1982-1990.," The Journal of Politics, Vol. 59, No. 1(1997), 226-239.

Williams, Christine B., and Girish Gulati., "The political impact of Facebook: Evidence from the 2006 midterm elections and 2008 nomination contest," Politics and Technology Review 1.1. (2008), 11-24..

Xianghua, F., Guo, L., Yanyan, G., and Zhiqiang, W., "Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon," Knowledge-Based Systems, Vol. 37(2013), 186-195.

Yonhap News, "Men in their 30s spend the most time for commenting in Naver news.", 2016.05.29.

Yoo. H. J., "An Empirical Study on the Effect of Information Environment on Voter Choice in Election", Korean Political Science Bulletin, Vol. 42, No. 4(2008), 155-188.

Abstract

Issue tracking and voting rate prediction for 19th Korean president election candidates

Dae-Ho Seo* · Ji-Ho Kim** · Chang-Ki Kim***

With the everyday use of the Internet and the spread of various smart devices, users have been able to communicate in real time and the existing communication style has changed. Due to the change of the information subject by the Internet, data became more massive and caused the very large information called big data. These Big Data are seen as a new opportunity to understand social issues. In particular, text mining explores patterns using unstructured text data to find meaningful information. Since text data exists in various places such as newspaper, book, and web, the amount of data is very diverse and large, so it is suitable for understanding social reality. In recent years, there has been an increasing number of attempts to analyze texts from web such as SNS and blogs where the public can communicate freely. It is recognized as a useful method to grasp public opinion immediately so it can be used for political, social and cultural issue research.

Text mining has received much attention in order to investigate the public's reputation for candidates, and to predict the voting rate instead of the polling. This is because many people question the credibility of the survey. Also, People tend to refuse or reveal their real intention when they are asked to respond to the poll.

This study collected comments from the largest Internet portal site in Korea and conducted research on the 19th Korean presidential election in 2017. We collected 226,447 comments from April 29, 2017 to May 7, 2017, which includes the prohibition period of public opinion polls just prior to the presidential election day. We analyzed frequencies, associative emotional words, topic emotions, and candidate voting rates. By frequency analysis, we identified the words that are the most important issues per day. Particularly, according to the result of the presidential debate, it was seen that the candidate who became

* Graduate School of Information, Yonsei University
** Division of Industrial Management Engineering, Korea University
*** Corresponding Author: Changki Kim
Graduate School of Information, Yonsei University
#418 NMH, 50 Yonsei-ro Seodaemun-gu, Seoul, 03722 Korea
Tel: +82-10-3473-3153 Fax: +82-50-4234-3153 E-mail: niceck7@daum.net

an issue was located at the top of the frequency analysis. By the analysis of associative emotional words, we were able to identify issues most relevant to each candidate. The topic emotion analysis was used to identify each candidate's topic and to express the emotions of the public on the topics. Finally, we estimated the voting rate by combining the volume of comments and sentiment score.

By doing above, we explored the issues for each candidate and predicted the voting rate. The analysis showed that news comments is an effective tool for tracking the issue of presidential candidates and for predicting the voting rate. Particularly, this study showed issues per day and quantitative index for sentiment. Also it predicted voting rate for each candidate and precisely matched the ranking of the top five candidates.

Each candidate will be able to objectively grasp public opinion and reflect it to the election strategy. Candidates can use positive issues more actively on election strategies, and try to correct negative issues. Particularly, candidates should be aware that they can get severe damage to their reputation if they face a moral problem.

Voters can objectively look at issues and public opinion about each candidate and make more informed decisions when voting. If they refer to the results of this study before voting, they will be able to see the opinions of the public from the Big Data, and vote for a candidate with a more objective perspective.

If the candidates have a campaign with reference to Big Data Analysis, the public will be more active on the web, recognizing that their wants are being reflected. The way of expressing their political views can be done in various web places. This can contribute to the act of political participation by the people.

Key Words : 19th president election, comments, text mining, Big data analysis

Received : May 28, 2018 Revised : September 6, 2018 Accepted : September 12, 2018
Publication Type : Regular Paper Corresponding Author : Chang-Ki Kim

저 자 소 개

서 대 호
한양대학교 정보시스템학과에서 학사학위를 취득 후 한양대학교 산업공학과에서 석사학위를 취득하였다. 그 후, 한국과학기술원, 모비젠, 전자부품연구원에서 연구원으로 재직하였다. 현재 연세대학교 정보대학원 박사과정에 재학 중이다. 관심분야는 텍스트 마이닝, 딥러닝, 프로세스 마이닝, 이상탐지이다.

김 지 호
서울과학기술대학교 글로벌융합산업공학과에서 학사 학위를 취득 후 고려대학교 산업경영공학부에서 석박사 통합과정으로 재학중이다. 관심분야는 데이터마이닝, 텍스트마이닝, 기계학습이다.

김 창 기
연세대학교 정보대학원 정보시스템 석사 학위 취득 후 SK 하이닉스 경영진단팀, 모바일 마케팅팀, 신세계 아이앤앤씨에서 연구원으로 재직하였다. 현재 연세대학교 정보대학원 박사과정에 재학중이며 NADOgive(나도기브) 대표이사로 재직하고 있다. 관심분야는 정보보호, 빅데이터, 인공지능이다.