ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2022 September: 28(3): 161~183 http://dx.doi.org/10.13088/jiis.2022.28.3.161

유튜브 데이터를 활용한 20대 대선 여론분석*

강은경
경희대학교 일반대학원 빅데이터응용학과
(luckiness1@khu.ac.kr)

양선욱
경희대학교 일반대학원 빅데이터응용학과
(seonuk.yang@khu.ac.kr)

권지윤
경희대학교 일반대학원 빅데이터응용학과
(kwon31403@khu.ac.kr)

양성병
경희대학교 경영대학 경영학과 & 빅데이터응용학과
(sbyang@khu.ac.kr)

여론조사는 유권자들의 투표행위를 예측하고, 그 행위에 영향을 준다는 점에서 선거운동의 강력한 수단이자, 언론의 가장 중요한 기사거리로 자리잡고 있다. 하지만, 여론조사가 활발할수록 후보자들의 공약과 정책을 검증하기 보다 당선 가능성이나 지지도에 관한 조사만 반복적으로 실시하는 등 선거 캠페인에 관한 효과 측정에서 유권자들의 마음을 제대로 반영하지 못하는 경우가 많다. 여론조사의 선거 결과에 대한 부실한 예측이 언론사의 권위를 실추시켰다 하더라도, 어느 후보가 최종 승리할지에 대해 인간의 본능적인 궁금증을 풀어줄 명백한 대안이 없기 때문에 사람들은 여론조사에 대한 관심을 쉽게 놓지 못한다. 이에, 온라인 빅데이터를 통해 인사이트를 발굴하는 환경을 제공하는 썸트렌드의 '유튜브 분석' 기능을 활용하여 20대 대선에 대한 여론을 회고적으로 파악해 보고자 한다. 본 연구를 통해 간단한 유튜브 데이터 분석 결과만으로도 실제 여론(혹은 여론조사 결과)에 근접한 결과를 쉽게 도출하고, 성능이 좋은 여론 예측모형을 구축할 수 있음을 확인하였다.

주제어 : 여론분석, 유튜브 데이터, 20대 대선, 머신러닝, 여론 예측모형

논문접수일 : 2022년 8월 29일 논문수정일 : 2022년 9월 8일 게재확정일 : 2022년 9월 13일
원고유형 : Regular Track 교신저자 : 양성병

1. 서론

선거 전 인터넷 조사나 ARS(automatic response system), 유무선 전화면접 등과 같은 방식으로 진행되는 정치 관련 여론조사는(공윤엽 등, 2022), 유권자들의 투표행위를 예측하고 그 행위에 영향을 준다는 점에서 중요성이 강조되어 왔다(홍원식 등, 2009; Patterson, 2005). 권혁남(2001)의 연구에 의하면, 각 정당은 여론조사 결과를 공천의 주요 기준으로, 후보들은 선거운동의 전략이자 수단으로 활용하고 있으며, 유권자들은 당선 가능성 여부에 대한 정보를 제공받아 경쟁력 있는 후보를 선택하는데 활용하고 있다. 이처럼 여론조사는 선거운동의 강력한 수단이자, 언론의 가장 중요한 기사거리로 자리잡고 있다(권혁남, 2001).

한편, 선거 캠페인에 관한 효과 측정으로 다양한 여론조사가 시도되고 있지만, 최근 국·내외

* 본 연구는 ㈜바이브컴퍼니의 지원을 받아 데이터를 수집하였음.
이 논문은 2020년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2020S1A5B8103855).

선거에서 유권자들의 마음을 제대로 반영하지 못하는 결과들이 다수 나타나고 있다(김찬우 등, 2017). 여론조사 업체가 난립할수록 후보자들의 공약과 정책에 대한 검증보다는 당선 가능성이나 지지도에 관한 부정확한 변화 추이만을 경마식 보도 형태로 보여주게 되므로, 유권자들의 판단을 흐리게 만들 가능성 또한 높다(김양원, 송경재, 2021). 한 예로, 2022년 2월 21일, 같은 날에 실시된 제20대 대통령 선거(이하 20대 대선)의 두 여론조사 결과만 보더라도, TBS-KSOI의 조사 결과는 더불어민주당 이재명 후보(이하 후보 1)가 1.5%p 오차범위 내에서 우세하다고 주장한 반면, JTBC-글로벌리서치의 결과는 국민의힘 윤석열 후보(이하 후보 2)가 8.3%p 오차범위 밖에서 우세한 것으로 나타난 바 있다(안지현, 2022). 이와 같이 다른 결과가 나타나는 이유로, 안지현(2022)은 정치 관심도에 따라 조사방법(KSOI: 녹음된 음성 청취 후 ARS 응답, 글로벌리서치: 전화 면접원 직접 문답)을 대하는 응답자들의 반응이 달라지는 점을 지적하였다. 김양원, 송경재(2021)는 2021년 11월 1일부터 10일까지 10일 간 실시된 전국 단위 여론조사가 총 40회로, 하루에 네 건 꼴로 여론조사가 진행되었으며, 같은해 11월 11일 현재 '중앙선거여론조사심의위원회'에 등록된 여론조사 기관이 총 81개에 달해, 이러한 여론조사의 과잉은 유권자들에게 정보를 주기보다는 혼란만을 야기할 수 있다고 주장하였다.

선거 결과를 제대로 예측하지 못한 여론조사 결과로 인해 언론사의 권위가 실추되었다 하더라도 여론조사에 대한 관심을 놓지 못하는 이유는 선거 때마다 어떤 후보가 앞서 있고 어떤 후보가 최종적으로 승리할지 등에 대한 인간의 본능적인 궁금증을 명확하게 풀어줄 대안이 없기 때문이다(권혁남, 2001). 따라서, 이러한 여론조사의 한계를 보완하기 위한 노력도 최근 활발해지고 있는데, 여론조사 메타분석(이하 메타분석)이나 검색어 트렌드 분석, 소셜 분석 등 전통적인 방식을 벗어나 다양한 방식으로 여론의 움직임을 확인하려는 방법도 시도되고 있다(임예인, 2022). 우선, 메타분석은 MBC나 SBS 등에서 동적선형모형을 이용하여 여론조사 결과를 종합하는 것으로, 전체적인 여론흐름을 종합적으로 살펴볼 수 있다는 장점이 있음에도 불구하고, 만약 여론조사 결과 자체가 부정확한 것이라면 메타분석 결과 자체의 신뢰성도 떨어질 수 있다는 근본적인 약점이 존재한다. 한편, 구글, 네이버, 카카오 등 검색엔진을 이용한 검색어 트렌드 분석은 단순 키워드의 검색량(키워드 서치 및 클릭)으로만 여론을 예측하기 때문에 긍·부정 여론을 구분하지 못하는 한계가 있으며(임예인, 2022), 트위터, 블로그, 커뮤니티, 인스타그램 등 소셜미디어 데이터를 활용한 소셜 분석은 인플루언서들의 일방성과 폐쇄성으로 인해 개인적인 의견(특히, 부정적인 의견)이 많고, 특정 정당이나 후보를 지지하는 용도로 사용되는 경우가 많으며, 동일한 데이터가 다수 생성(예: 리트윗)되는 문제로 자료수집 및 처리에 한계가 존재한다(하상현, 노태협, 2020). 이에, 세계 최대 동영상 공유 플랫폼이자 국내 인터넷 이용자의 80%가 사용하고 있는 유튜브의 일부 동영상 및 관련 댓글을 활용하여 여론을 분석하려는 시도가 소수 있었으나(<표 1> 참조), 유튜브의 빅데이터를 활용해 전체 여론을 예측한 연구는 많지 않은 실정이다.

이에 본 연구에서는 ㈜바이브컴퍼니(VAIV)의 썸트렌드 비즈(이하 썸트렌드) 서비스 내 '유튜브 분석' 기능을 활용하여 유튜브 데이터를 수집한 후, 20대 대선의 여론분석을 회고적으로 파악

〈표 1〉 유튜브 분석 관련 주요 선행연구

저자(연도)	연구 방법	연구내용
Krishna et al. (2013)	감성분석(Naïve Bayes 분류 기법 및 Weka 활용)	웹 '버즈', 주식시장 동향, 박스오피스 결과 및 정치적인 선거와 관련된 400 만 개 이상의 각 유튜브 동영상에서, 동영상당 1,000 개의 댓글을 수집하고, 'Federer', 'Nadal', 'Obama'와 같은 특정 키워드에 대한 데이터와 timestamp, 작성자 이름 수집 후, Naïve Bayes 분류 기법을 사용하여 감성분석 진행 ⇒ 키워드 관련 감성의 추세와 실제 이벤트 간의 상관관계 식별
Shevtsov et al. (2020)	감성분석	2020 년 미국 대통령 선거 기간(2020 년 7 월 ~ 9 월까지 6 가지 다른 시점) 중 가장 인기 있는 해시태그를 이용 750 만개의 트위터 데이터와 유튜브 동영상 및 메타데이터(좋아요, 댓글, 작성자 등)를 추출하여 감성분석 실시 ⇒ 트럼프에 대한 긍정적인 감성(트위터: 45.7%, 유튜브: 14.55%)이 바이든에 대한 긍정적인 감성(트위터: 33.8%, 유튜브: 8.7%) 보다 높게 나타남
김찬우 등 (2017)	댓글망 분석, 댓글 단어 빈도 분석, 단어쌍분석	2017 년 대통령 후보자 이름과 후보수락 연설문을 검색어로 설정하여 선별한 유튜브 동영상 중 각 후보별로 조회수나 댓글수가 가장 많은 동영상을 각 2 편씩 선정(단, 홍준표 후보는 수집 과정에서 두번째 영상이 삭제되어 한 편만 분석)하여 댓글을 수집하고, 댓글망 분석, 댓글 단어 빈도 분석, 단어 쌍 분석을 실시 ⇒ 대선 메시지가 댓글상에 나타나는 것이 중요, 유튜브 연구 및 선거캠페인 분석에 새로운 분석지표와 연구방향 제시
송화영 등 (2020)	오피니언 마이닝(감성분석), 단어빈도 분석, 의미연결망 분석	2020 년 총선 공식 선거 운동 기간(4/2~4/14) 중에 4 개 정당(더불어민주당, 미래통합당, 더불어시민당, 미래한국당)의 업로드 된 유튜브 선거 캠페인에 대한 대중의 반응을 영상목록, 조회 수, 댓글 수 등으로 수집하여 오피니언 마이닝(감성분석), 단어빈도 분석, 의미연결망 분석 등을 실시 ⇒ 유권자들의 반응과 앞으로의 정치적 선택, 행동 변화 등을 파악

해 보고자 한다. 구체적으로 다음의 절차를 따라 연구를 진행하였다. (1) 대선 후보의 이름을 키워드로 한 유튜브 데이터를 '유튜브 분석' 기능을 활용해 추출하여, (2) 키워드와 관련한 유튜브의 긍정/부정/중립 언급량 데이터와 동영상/조회/좋아요/댓글 수 데이터를 전처리하고, (3) 유튜브 분석, 소셜 분석, 트렌드 분석 데이터에 대해 머신러닝 기법을 활용하여 메타분석 결과를 학습시켜, (4) 학습된 모형으로 여론조사 공표금지 기간(이하 깜깜이 기간)과 대선결과를 예측하여 실제 결과와 비교한 후, (5) 각 데이터 셋의 특성과 학습 데이터로서의 성능에 대해 검증해 보고자 한다. 본 연구를 통해 유튜브 데이터를 활용한 간단한 분석 결과만으로도 실제 여론(혹은 여론조사 결과)에 근접한 결과를 쉽게 도출하고, 성능이 좋은 여론 예측모형을 구축할 수 있다는 점을 보여주고자 한다.

2. 이론적 배경 및 선행연구

2.1. 여론분석

여론조사는 유권자에게 여론의 추이를 파악할

수 있는 정보를 제공하고, 선거 후보자에게는 성공적인 선거전략을 마련하기 위한 근거를 제공한다(하승태, 2012). 1936년 과학적인 표본추출 방법에 근거한 조지 갤럽(George Gallup)의 여론조사 이후, 정치 맥락에서의 여론조사는 정확성과는 별개로 선거 결과를 예측하는 중요한 수단으로 인식되어 왔다(하승태, 2012). 우리나라에서는 1987년 제13대 대통령 선거 기간에 처음 도입된 이후(정일권, 김상연, 2021), 선거 판세 분석에 유리하고 비교적 저렴한 비용으로 쉽게 뉴스를 생산하여 대중의 주목을 끌 수 있다는 장점 때문에 많은 관심을 받아왔다(하승태, 2012).

그러나, 부실 여론조사로 인한 부정확한 결과 예측과 뉴스 내용의 편향성, 전략적인 투표를 유도할 수 있다는 점 등으로 인해 종종 후보자나 유권자들로부터 비난의 대상이 되기도 한다(정일권, 김상연, 2021). 이때문에, 최필선, 민인식(2013)은 여론조사가 홍수를 이루고 있다고 여겨질 정도로 자주 그리고 많이 이뤄지고 있지만, 조사결과의 신뢰성에 의문을 제기하며, 같은 날 발표된 여론조사라 하더라도 조사기관에 따라 다른 결과를 보여주므로 최종 선거 결과를 모집단의 지지율로 보고 표본조사 결과인 여론조사 결과의 정확성을 사후에 비교·평가할 필요가 있다고 주장하였다.

2.2. 소셜미디어를 활용한 여론분석

특정 후보를 지지하는 유권자의 비율이 설문에 참여한 집단의 비율보다 많거나 적어서 발생한 편향이나, 평균 응답률이 25%에도 미치지 못하는 낮은 응답률로 인해 정확한 선거 결과를 예측하기 어려울 수 있다(이예나 등, 2018). 또한, 침묵의 나선이론(spiral of silence theory)에 따라 많은 응답자가 속마음을 잘 드러내지 않고 침묵하는 경향을 보임으로써 선거 결과 예측이 어려울 수 있으므로, 이를 극복하기 위한 대안으로 소셜미디어(social media) 여론분석을 활용하고 있다(이예나 등, 2018). 배정환 등(2013)의 연구에 의하면, SNS(social network service)는 전문적인 지식이나 기술 없이도 콘텐츠를 실시간으로 생산 및 공유할 수 있는 세계적인 커뮤니케이션 도구로, 이용자 간의 소통뿐만 아니라 정치적인 신념과 사회적인 이슈에 관한 개인적인 의견을 댓글로 표현할 수 있는 새로운 여론으로 자리 잡았으며, 이러한 기능은 특히 사전선거 기간에 중요한 역할을 할 수 있다(Shevtsov et al., 2020).

정치적인 소통과 여론 형성의 도구가 기술의 발전에 따라 미디어를 활용한 TV(television)와 신문에서 인터넷과 모바일을 활용한 SNS로 빠르게 넘어가고 있다(이수범, 강연곤, 2013). 대중들은 SNS를 중심으로 한 다양한 미디어 플랫폼을 통해 방대한 양의 데이터를 주도적으로 생산하고 공유하고 있으며, 대중으로부터 생성된 소셜 빅데이터는 사회적인 이슈를 분석하기 위한 좋은 원천으로 주목받고 있다(공윤엽 등, 2022). 또한, 이서영, 권상집(2019)은 지지세력을 결집하고 조직화하는데 있어 기존의 미디어를 뛰어넘을 뿐 아니라 현실 정치를 바꿀 수 있는 새로운 소통의 장으로 SNS를 평가한 바 있다. 이러한 SNS의 영향력을 2008년 오바마 대선후보의 미국 대통령 당선 과정에서 확인하게 되면서(이서영, 권상집, 2019), 국내에서도 2010년을 기점으로 트위터, 페이스북, 인스타그램 등의 SNS에서 개인의 직·간접적인 빅데이터를 분석함으로써 유권자 개개인의 성향을 파악하고 그 성향에 맞춰 공약을 제시하는 선거운동 전략이 전개되었다(하상현, 노태협, 2020).

한편, 최근에는 대표적인 동영상 콘텐츠 공유 웹사이트이자 소셜미디어 서비스인 유튜브로 검색 시장이 이동하면서 여론 선도의 패권이 바뀌어 가고 있으며(김종훈, 2019; 박병언, 임규건, 2015), 동영상에 달린 댓글의 내용적 특성과 댓글망의 패턴은 선거캠페인의 효과 측정에 큰 가치를 지니게 되었다(김찬우 등, 2017). 2005년에 탄생한 유튜브는 매일 10억 시간 이상의 동영상이 시청되고, 월 20억 명 이상이 이용하는 세계 최대 소셜미디어 플랫폼이다(박중현, 2021). 콘텐츠 소비가 전파 중심에서 미디어 플랫폼 중심으로 넘어가는 현상에 따라 50대 이상에서는 재테크와 부동산, 정치적 성향의 콘텐츠 이용이 대폭 늘어났고, MZ세대에서는 '스낵컬쳐' 현상으로 텍스트보다는 이미지와 동영상을 선호하는 경향이 크게 나타났다(박상현 등, 2020).

Mehrabian(1981)에 의하면, 효과적인 의사소통은 비언어적 요소인 시각정보와 청각정보 93%에 의해 이루어지는데, 사람은 대개 눈으로 본 것 80%, 읽은 것 20%, 들은 것 10%를 기억하는 것으로 나타났다(Lee, 2014). 이에 따라, 유튜브 정치·시사 채널의 높은 이용률에 주목하여 이용자 관점에서 현상을 분석할 필요가 있는데(박상현 등, 2020), 대표적인 유튜브 정치·시사 채널은 다음과 같다. 2022년 8월 현재 신혜식 독립신문 대표가 운영하는 '신의 한수'는 구독자 145만명, 김어준 딴지일보 편집장이 운영하는 '딴지방송국'은 구독자 102만명, 인천시 계양구을 국회의원인 후보 1이 운영하는 '이재명'은 구독자 64.5만명, 제20대 대통령인 후보 2가 운영하는 '윤석열'은 46.4만명의 구독자를 거느리고 있다. 특히, 후보 1은 지난 3월 대한민국 정치인 중 유일하게 유튜브 채널 누적조회수 1억 뷰(view)를 돌파한 바 있다(이광춘, 2022).

이와 같은 흐름에 따라, 학계에서도 유튜브의 대표 동영상 및 관련 댓글을 활용해 여론을 분석하려는 연구가 일부 있어 <표 1>에 정리하였으나, 유튜브의 빅데이터를 활용해 전체 여론을 예측한 연구는 전무한 상황이다. 이에, 본 연구에서는 유튜브 데이터를 이용한 간단한 분석 결과만으로도 20대 대선의 회고적인 여론조사 결과와 비교하여 실제 여론(혹은 여론조사 결과)에 근접한 결과를 얻을 수 있는지, 또한 성능이 좋은 여론 예측모형을 구축할 수 있는지를 검증해 보고자 한다.

2.3. 머신러닝 기법을 활용한 여론분석

가장 널리 사용되는 동영상 공유 서비스인 유튜브는 비디오를 공유하는 것 이외에도 비디오 채널을 구독하고 댓글을 통해 다른 사용자와의 상호작용이 가능한 특징이 있다(Krishna et al., 2013). 특히, 전통적인 선거운동 방식에 관심이 없던 젊은 세대로부터 잠재적인 정치적 담론을 이끌어 낼 수 있다는 점에서 효율적인 민주화 수단으로 각광받고 있다(강은경 등, 2022; Chen and Wang, 2022; Ridout et al., 2010). Shevtsov et al. (2020)에 의하면, 유튜브에서는 매일 수백만 명의 사용자들이 자신의 정치적 신념이나 정당에 대한 생각 등을 동영상이나 댓글의 형태로 표현하고, 이러한 담론의 분석을 통해 유권자의 성향과 선호도를 파악함으로써 선거 결과를 예측할 수 있다. 또한, 유튜브는 언어적인 요소와 함께 후보자의 태도나 목소리, 행동 등 비언어적인 요소도 전달할 수 있는 중요한 수단이다(김찬우 등, 2017).

정치인들은 다른 소셜미디어 채널에 비해 유권자와 보다 직접적이고 적극적인 소통을 할 수

있는 채널로 유튜브를 선택하고 있으므로, 유권자들이 실시간으로 남긴 댓글에 빅데이터 분석 기법을 적용하여 그들의 정치적인 선택 및 행동 변화를 파악할 필요가 있다(송화영 등, 2020). 이와 같이 대용량 데이터를 분석하고 활용하는 것은 의사결정 과정에서 통찰력을 제공하고 급변하는 미래에 대한 효과적인 대응을 하기 위함이다(정지선 등, 2015). 예를 들어, 미국 대선의 결과는 미국내 시장뿐만 아니라 전 세계 경제에도 영향을 미치는 요인이므로, 미국 대통령 선거를 모델링하고 예측의 정확도를 높이는 일은 중요하다(Zolghadr et al., 2018). 이러한 이유로 Zolghadr et al. (2018)은 신뢰할 수 있는 예측모형을 찾기 위해 학습 알고리즘 기반의 인공신경망(artificial neural network: ANN)과 서포트 벡터 회귀(support vector regression: SVR)를 지난 세 번의 미국 대선(2004년, 2008년, 2012년) 결과에 적용함으로써, 선형회귀모형에 비해 더 정교한 결과를 도출하는데 성공하였다.

이처럼 예측의 정확도를 높여주는 학습 알고리즘 기반의 머신러닝 기법으로 선형회귀모형과 릿지 선형회귀모형(ridge linear regression model), K-최근접 이웃(K-nearest neighbor: K-NN), 의사결정나무(decision tree), 서포트 벡터 머신(support vector machine: SVM), ANN, 앙상블 모형(ensemble) 등이 있다(김형수, 2020). 선형회귀모형은 다른 머신러닝 모형과의 성능 비교를 위해, 릿지 선형회귀모형은 변수들의 중요도에 따른 가중치를 조절하기 위해, K-NN은 사전 정보가 부족한 데이터를 소수의 데이터만으로 예측하기 위해 주로 사용된다. 또한, 의사결정나무는 분류 순서에 따른 데이터의 중요도를 확인하기 위해, SVM은 여러 유형의 데이터에 적용시키기 위해, ANN은 비선형을 학습하기 위해, 앙상블은 다양한 기법을 결합하여 더 높은 성능을 예측하기 위해 사용된다. 이에, 본 연구에서는 일곱 가지 주요 머신러닝 기법을 활용하여 깜깜이 기간 중 후보 1과 후보 2의 지지율을 예측해 보고자 한다.

〈그림 1〉 역대 대선 선거 100일전 여론조사 및 득표율 재구성(박영석, 2021)

3. 연구방법

3.1. 데이터 수집기간

지난 14대 대선부터 19대 대선까지 'D-100일'을 전후한 여론조사에서 1위를 달린 후보의 여론조사 추이를 분석해보면, 여섯 번의 대선에서 다섯 번이나 최종 승리를 한 것으로 나타났다(박영석, 2021; 전창훈, 2021)(<그림 1> 참조). 2022년 3월 9일 치러진 20대 대선의 100일 전 날짜를 계산하면 2021년 11월 29일이 이에 해당하고, 해당일로부터 일정기간 동안은 모든 후보자에게 주요 이슈가 없었으며, 수집기간으로 산정한 100일은 분석에 필요한 데이터를 확보하기에 충분한 기간이라 판단하였다. 이에, 본 연구에서는 2021년 11월 29일부터 100일 간인 2022년 3월 8일까지가 민심의 향배를 가늠하기에 유의미한 기간으로 보고 이를 데이터 수집기간으로 설정하였다.

3.2. 데이터 수집대상

본 연구는 20대 대선을 후보 1과 후보 2의 양강구도로 보고, 이를 전제로 진행하였다. 또 한 명의 유력 후보였던 안철수 후보(이하 후보 3)는 대선기간 중 지지율을 최대 17%까지 확보하며 본 연구 결과에 유의한 영향을 미칠 수 있었으나, 2022년 3월 3일 후보 2와 막판 단일화를 선언하며 20대 대선에서 사퇴한 바 있다. 고정애(2022)의 자료에 의하면, 후보 3의 지지율 대부분이 단일화를 거친 후 후보 2에게 넘어갈 것이라는 예상과 달리, 20대부터 50대까지의 표심이 후보 1(38.3%)과 후보 2(37.7%)에게 양분되었음을 알 수 있다. 후보 3에 대한 60대 이상의 지지가 높지 않다고 추정한다면, 후보 3의 단일화 발표는 후보 2에게 긍정적인 영향을 미치지 못하고 양강구도만 더욱 공고하게 만든 것으로 풀이된다(고정애, 2022). 또한, KBS가 진행한 '2022 대통령 선거 심층출구조사 분석'에서 대통령 후보를 선택한 이유로 후보 1의 경우는 '후보 개인의 자질과 능력이 뛰어나서', '공약 및 정책이 마음에 들어서'가 주를 이룬 반면, 후보 2의 경우 '소속 정당이 좋아서', '이념 성향이 나와 맞아서'가 가장 높은 비율을 차지하였다(이승종, 2022). 이에 따라 후보 3의 지지자들 중에서 후보 1에게 표심을 준 이유로는 후보의 자질과 능력을, 후보 2는 동일한 정치 성향을 꼽은 것으로 보인다. 후보 1~3을 제외한 나머지 후보들은 20대 대선 기간동안 최대 5%에도 미치지 못하는 지지율에 그쳤기에 이들의 지지율은 영향이 크지 않다고 판단하고 본 연구에서 제외하였다. 결과적으로, 본 연구에서는 20대 대선을 후보 1과 후보 2의 양강구도로 전제하고, 두 후보를 제외한 나머지 후보에 대한 지지율 및 무효표를 두 후보의 지지율에 맞추어 균등 배분하여 100%로 보정하는 방법을 사용하였다.

3.3. 데이터 수집 및 전처리

본 연구에서는 유튜브에 게시된 정치·시사 채널의 동영상을 활용하여 20대 대선의 회고적 여론을 분석하고자 하였다. 본 연구는 유튜브 분석을 활용하여 양강구도인 두 후보에 대한 여론분석을 진행함으로써, 20대 대선결과를 회고적으로 예측하고, 도출된 예측결과를 소셜 분석, 검색어 트렌드 분석, 메타분석 결과와 비교함으로써 유튜브 분석에서 유의미한 결과를 도출할 수 있는지를 확인하고자 하였다. 유튜브 분석은 빅데이터 분석 서비스인 썸트렌드를 활용하여 유

튜브의 영상 제목과 영상의 설명 글을 수집한 후, 분석을 진행하였으며, 소셜 분석 역시 썸트렌드를 활용하여 뉴스, 트위터, 블로그, 커뮤니티, 인스타그램 등의 게시물을 수집한 후, 분석하였다. 검색어 트렌드 분석은 구글 트렌드, 네이버 데이터랩, 카카오 데이터 트렌드의 웹 검색량을 활용하여 분석을 진행하였고, 메타분석은 MBC의 여론조사 통합 사이트인 '[여론M] 여론조사를 조사하다'(MBC, 2022)의 여론조사 통합 데이터를 활용하였다. 좀 더 구체적으로, 본 연구에서는 데이터를 수집하기 위해 온라인 데이터를 통해 인사이트를 발굴하는 환경을 제공하는 SNS 빅데이터 분석 서비스인 썸트렌드를 활용하였다(박석봉, 정주호, 2021). 썸트렌드는 트위터와 뉴스, 그리고 약 7,000개 이상의 커뮤니티 게시판(예: 뽐뿌, 네이트판, 중고나라 등)과 블로그, 인스타그램 등에서 SNS 빅데이터를 수집하여 국내 최다 데이터를 보유하고 있을 뿐 아니라, 인공지능 기술을 이용하여 분석하고자 하는 검색어의 추이/연관어/감성/비교/랭킹 분석 등의 기능을 제공하며, 분석 데이터도 쉽게 다운받을 수 있도록 서비스를 제공하고 있다(바이브컴퍼니, 2022).

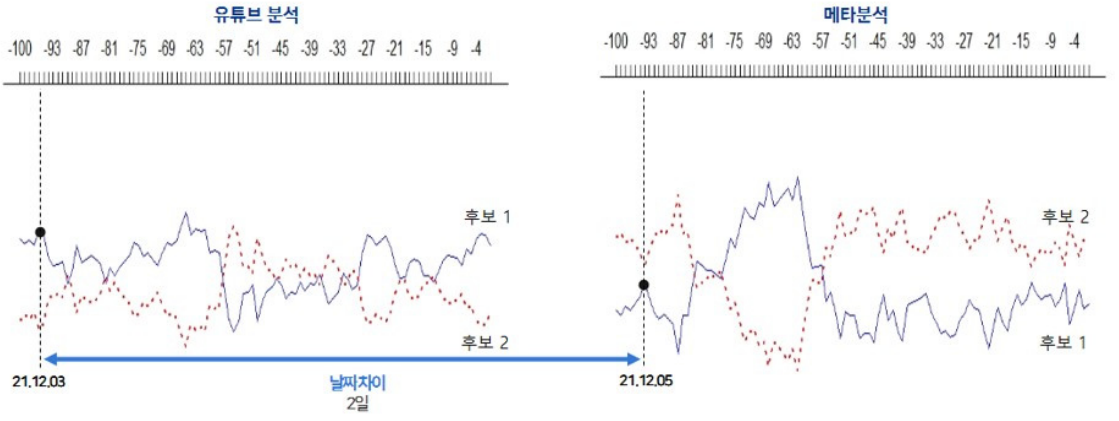
본 연구에서는 썸트렌드에서 양강구도로 진행되었던 20대 대선 두 후보의 이름을 키워드로 하여 검색한 후, 키워드 관련 유튜브의 긍정/부정/중립 언급량 데이터와 동영상/조회/좋아요/댓글 수 데이터를 분석에 각각 활용하였다. 해당 후보 키워드 관련 유튜브 동영상 수, 조회 수, 좋아요 수, 댓글 수가 반드시 해당 후보에게 긍정적으로 작용한다고 볼 수 없기 때문에, 각 후보자별 유튜브의 키워드 언급량 긍·부정 비율에 맞춰 유튜브 동영상/조회/좋아요/댓글의 긍·부정 개수를 계산하였으며, 중립 언급량은 긍·부정 언급량에 1:1 비율로 배분하였다. 이를 통해, 긍·부정 비율에 맞춰 나눠진 데이터 중 긍정적인 동영상, 조회, 좋아요, 댓글은 해당 후보자에 대한 지지를, 부정적인 동영상, 조회, 좋아요, 댓글은 해당 후보자에 대한 반대를 의미하는 것으로 설정하였다.

후보 1의 경우, 분석기간 동안 동영상 수, 조회 수, 좋아요 수, 댓글 수는 각각 33,682개, 1,422,688,047개, 94,898,418개, 7,265,534개로 집계되었으며, 키워드 언급량 긍·부정 비율은 62.58: 37.42로 파악되었다. 각 동영상, 조회, 좋아요, 댓글 수 데이터에 긍·부정 비율을 적용한 결과, 후보 1의 긍정 데이터 수는 동영상 21,077개, 조회 890,247,045개, 좋아요 59,382,485개, 댓글 4,546,408개로, 부정 데이터 수는 동영상 12,605개, 조회 532,441,002개, 좋아요 35,515,733개, 댓글 2,719,126개로 집계되었다. 후보 2의 경우, 분석기간 동안 동영상 수, 조회 수, 좋아요 수, 댓글 수는 각각 35,886개, 1,712,950,048개, 107,069,632개, 9,818,077개로 집계되었으며, 키워드 언급량 긍·부정 비율은 58.22: 41.78로 파악되었다. 각 동영상, 조회, 좋아요, 댓글 수 데이터에 긍·부정 비율을 적용한 결과, 후보 2의 긍정 데이터 수는 동영상 20,891개, 조회 997,193,930개, 좋아요 62,330,587개, 댓글 5,715,594개로, 부정 데이터 수는 동영상 14,995개, 조회 715,756,118개, 좋아요 44,739,045개, 댓글 4,102,483개로 집계되었다.

특히, 깜깜이 기간 예측을 위한 데이터 셋은 입력 데이터와 메타분석 데이터로 구분하였다. 입력 데이터는 머신러닝 모형 학습에서 독립변수에 해당하는 것으로 2021년 11월 29일부터 2022년 3월 7일까지 수집된 후보 1과 후보 2의 유튜브 분석, 소셜 분석, 검색어 트렌드 분석 데

〈표 2〉 유튜브 분석 및 메타분석 상관분석 결과

상관계수(r)	유튜브 분석	유튜브 분석+1 일 결과	**유튜브 분석+2 일 결과**	유튜브 분석+3 일 결과
메타분석	0.4642	0.5110	**0.5200**	0.4933

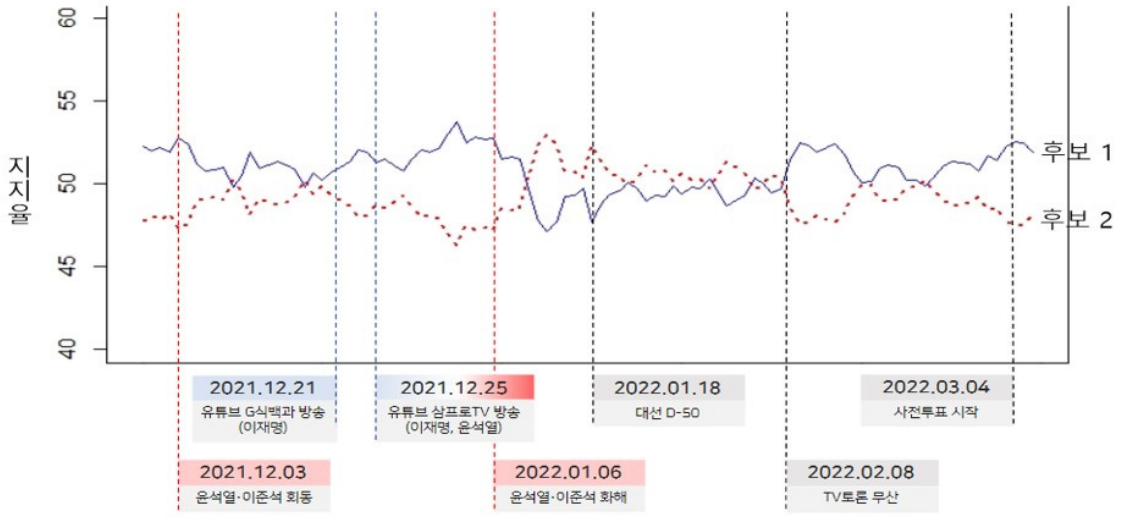


〈그림 2〉 유튜브 분석 vs. 메타분석

이터이고, 메타분석 데이터는 머신러닝 모형 학습에서 종속변수에 해당하는 것으로 2021년 11월 29일부터 2022년 3월 2일까지 수집된 데이터이다. 또한, 머신러닝 모형을 통해 예측한 깜깜이 기간과 대선결과의 예측값을 실제값과 비교하기 위해 여론조사 기관인 리얼미터에서 2022년 3월 3일부터 2022년 3월 8일까지 수집한 깜깜이 기간의 여론조사 자료를 사용하였고(김가현, 2022), 20대 대선결과는 중앙선거관리위원회가 발표한 자료를 사용하였다(중앙선거관리위원회, 2022).

회고적 예측을 목적으로 하는 본 연구의 특성상 유튜브 분석 지지율과 실제 대선결과 지지율 간의 차이를 조정할 필요성이 제기되었다. 보정을 위한 기준일을 정하기 위해 본 연구의 유튜브 분석 지지율과 '[여론M] 여론조사를 조사하다'(MBC, 2022)의 메타분석 결과를 비교해 보면(<그림 2> 참조), 두 후보에 대한 2021년 12월 3일의 유튜브 분석 결과와 당해 12월 5일 이후 메타분석의 지지율이 일정기간 유사한 추세를 보이고 있는 것을 확인할 수 있다. 또한, 2021년 12월 3일에 발생한 20대 대선결과에 영향을 미칠 정도의 이슈를 확인한 결과, 후보 2와 후보의 정당 대표 간 회동 이후 두 후보자의 지지율이 급변하기 시작한 것을 알 수 있다(<그림 3> 참조).

유튜브 분석 결과와 메타분석 결과를 비교해 보면(<그림 2> 참조), 유튜브 분석 결과가 메타분석 결과보다 2일 정도 선행하는 것을 관찰할 수 있다. 이는 여론조사는 대개 1일~3일 간 조사를 시행한 후 다음날 공표하게 되는 시간적인 지



〈그림 3〉 유튜브 분석 결과: 두 후보의 지지율

연이 있는 반면, 유튜브는 즉각(real-time)적으로 여론 반영이 가능하기 때문으로 판단된다. 또한, 메타분석 결과와 유튜브 분석 결과, 그리고 해당일 대비 1일에서 3일 이후의 유튜브 분석 결과 간 상관분석 결과를 비교한 결과, 2일을 더하여 보정하였을 때의 상관계수 결과가 가장 좋은 것을 확인할 수 있다(<표 2> 참조). 이를 근거로 본 연구에서는 머신러닝 모형에 사용될 데이터는 독립변수에 해당하는 유튜브 분석, 소셜 분석, 검색어 트렌드 분석 데이터의 기간을 2021년 11월 29일부터 2022년 2월 28일로 지정하였고, 종속변수에 해당하는 메타분석 데이터의 기간은 이로부터 2일을 더한 2021년 12월 1일부터 2022년 3월 2일로 지정하였다. 또한, 메타분석 데이터의 경우 후보 1과 후보 2 지지율의 합산이 100%가 되도록 조정하였다. 이렇게 선정된 총 변수의 수는 92개이며 해당 데이터 셋을 학습용 데이터 70%와 평가용 데이터 30%로 무작위로 분할하여 모델의 학습과 평가를 진행하였다.

4. 유튜브 분석을 활용한 20대 대선결과 예측모형 개발

4.1. 여론조사 공표금지 기간

깜깜이 기간은 선거일 6일 전부터 투표마감 시각까지이며, 공표금지 기간에 실시된 여론조사 결과를 인용하여 보도할 수 없다(중앙선거관리위원회, 2017). 1992년 대통령선거법 제65조 제1항이 제정되었을 당시 여론조사는 가능하였지만, 선거기간 내내 여론조사 결과를 공표하는 것이 금지되었다. 그후 1994년 선거관련 법제가 공직선거법으로 통합되면서 깜깜이 기간은 선거일 60일 전으로 단축되었고, 최종적으로 2005년 선거법 108조가 개정되며 현재와 같은 선거일 6일 전으로 확립되었다.

이렇게 깜깜이 기간을 두고 결과를 공표하지 않는 이유는 여론조사 결과가 유권자로 하여금 당선 가능성이 높은 후보쪽으로 표가 쏠리게 되

는 밴드왜건 효과(bandwagon effect)와 반대로 지지율이 낮은 후보자를 동정하여 지지하게 되는 언더독 효과(underdog effect)를 차단함으로써 공정한 선거를 진행하기 위해서이다(최문열, 2022). 그렇지만 깜깜이 기간으로 인해 선거 후보자는 여론조사 결과에 따른 유동적인 선거 운동이나 신속한 대응을 하지 못하게 되고, 유권자들은 정보가 차단된 상태에서 어려운 선택을 해야 하는 문제점도 보고되고 있다(최문열, 2022). 20대 대선에서 후보 3의 경우, 2022년 3월 3일 깜깜이 기간이 시작되자마자 후보2와 단일화를 선언하였으므로, 그 효과가 어느 후보에게 긍정적인 영향을 미쳤는지 알 수 없다. 이에 따라 후보자와 지지자들은 전략적인 선거 운동을 위해 깜깜이 기간 동안 지지율을 유추할 수 있는 수단을 찾을 필요가 있다.

4.2. 머신러닝 방법론

본 연구에서는 깜깜이 기간 중 후보 1과 후보 2의 지지율을 예측하기 위해 수치예측이 가능한 머신러닝 기법 중 선형회귀모형, 릿지 선형회귀모형, K-NN, 의사결정나무, SVM, ANN, 앙상블 모형 등 총 일곱 가지 기법을 사용하였다. 선형회귀모형은 다른 머신러닝 모형과의 성능비교를 위해 사용하였고, 릿지 선형회귀모형은 변수들의 중요도에 따라 가중치를 조절할 수 있어서 사용되었다. 다만, 분석 시점에서는 학습 데이터들 간의 중요성을 알지 못하므로 규제를 통해 가중치를 0까지 축소시킬 수 있는 라쏘 선형회귀모형은 배제하였다. K-NN은 데이터에 대한 사전 정보가 부족할 때 소수의 이웃 데이터만을 이용하여 예측이 가능하다는 장점이 있으므로, 학습 데이터의 특징에 대한 정보가 부족한 본 연구의 상황을 고려하여 사용하였다. 의사결정나무는 분류 순서에 따라 데이터의 중요도를 알 수 있고 수치예측이 가능하므로 사용하였다. SVM은 여러 유형의 데이터에 적용시킬 수 있어 다양한 용도로 사용이 가능하며, SVM 모형의 일종인 SVR에 수치 예측을 적용할 수 있으므로 본 연구의 머신러닝 기법으로 사용하였다. 또한, 본 연구에서 ANN은 비선형 학습이 가능하다는 점에 주목하여 선형학습 방식의 다른 머신러닝 기법들보다 높은 성능을 보여주리라 판단하였다. 마지막으로, 다양한 머신러닝 기법을 결합한 앙상블 모형은 더 높은 성능을 보여줄 수 있을 것으로 예상하여 이를 예측모형 개발에 사용하였다.

4.3. 모형구축 및 학습

본 연구는 깜깜이 기간과 대선결과를 예측하기에 적합한 데이터와 머신러닝 기법을 찾고, 유튜브의 정보가 다른 플랫폼의 정보보다 여론을 잘 반영한다는 단서를 찾는 것에 목적이 있다. 따라서, 총 네 종류의 데이터 셋을 일곱 개의 머신러닝 기법으로 학습한 후 머신러닝 기법 별로 성능과 예측결과를 비교하였다. 학습에 사용된 데이터 셋은 유튜브 분석, 소셜 분석, 유튜브 검색을 제외한 검색어 트렌드 분석, 유튜브 검색을 포함한 검색어 트렌드 분석 데이터이다. 사용된 머신러닝 기법은 선형회귀모형, 릿지 선형회귀모형, K-NN, 의사결정나무, SVM, ANN, 앙상블이다. 단일 종속변수만 학습이 가능한 선형회귀모형과 SVM은 각 후보별로 모형을 학습하게 하였고, 다중 종속변수도 학습이 가능한 릿지 선형회귀모형, K-NN, 의사결정나무, ANN은 두 후보를 함께 입력하여 모형을 학습하였다.

머신러닝 기법들은 학습 효율에 직접적인 영

향을 미치는 다양한 파라미터(parameter)가 존재하며, 이러한 파라미터들은 어떠한 성능평가 지표를 기준으로 삼는가에 따라 값이 달라질 수 있다. 본 연구에서는 최적의 모형을 학습하기 위해 평가의 기준이 되어줄 지표로 수식 (3)의 평균제곱근 오차(root mean square error: RMSE)를 사용하였다. RMSE를 기준으로 최고의 성능을 내는 모형의 파라미터를 채택하였고, 앙상블 모형의 경우 RMSE가 가장 낮은 모형 세 가지를 선정하여 학습을 진행하였다. 유튜브 분석, 소셜 분석, 유튜브를 제외한 검색어 트렌드 분석, 유튜브를 포함한 검색어 트렌드 분석의 각 데이터를 독립변수로 사용할 때 최적의 파라미터는 <표 3>과 같다.

$$\text{RMSE}^{*}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y_i})^2} \quad (3)$$

* y_i: measured value, $\hat{y_i}$: forecast value,
n: number of data

〈표 3〉 분석별 최적의 파라미터

구분	유튜브 분석		소셜 분석		검색어 트렌드 분석 (유튜브 제외)		검색어 트렌드 분석 (유튜브 포함)	
학습모형	RMSE	파라미터	RMSE	파라미터	RMSE	파라미터	RMSE	파라미터
선형회귀 모형	(1) 1.9573	-	(1) 2.6943	-	(1) 2.3583	-	(1) 2.1067	-
	(2) 1.9573		(2) 2.6943		(2) 2.3583		(2) 2.1067	
릿지 선형회귀 모형	(3) 1.9390	alpha: 1	(3) 1.8474	alpha: 10	(3) 2.2531	alpha: 10	(3) 2.0667	alpha: 1
K-NN	(3) 1.6056	n_neighbors: 3	(3) 1.8028	n_neighbors: 11	(3) 2.2658	n_neighbors: 9	(3) 1.7581	n_neighbors: 3
의사결정 나무	(3) 2.2820	max_depth: 4	(3) 2.2115	max_depth: 5	(3) 2.2685	max_depth: 9	(3) 1.7192	max_depth: 4
SVM	(1) 1.9679	C: 20, epsilon: 0.1	(1) 1.9686	C: 10, epsilon: 0.01	(1) 2.2783	C: 20, epsilon: 0.05	(1) 2.2133	C: 5, epsilon: 0.1
	(2) 1.9291	C: 20, epsilon: 0.1	(2) 1.9726	C: 10, epsilon: 0.01	(2) 2.2495	C: 5, epsilon: 0.05	(2) 2.1157	C: 5, epsilon: 0.1
ANN	(3) 1.9135	Alpha: 1, hidden_layer_sizes: [50,50], activation: identity, max_iter =1,000	(3) 1.8655	Alpha: 0.01, hidden_layer_sizes: [50,50], activation: identity, max_iter =1,000	(3) 2.2513	Alpha: 0.01, hidden_layer_sizes: [50,50], activation: identity, max_iter =1,000	(3) 1.9157	Alpha: 1, hidden_layer_sizes: [100,100], activation: identity, max_iter =1,000

구분	유튜브 분석		소셜 분석		검색어 트렌드 분석 (유튜브 제외)		검색어 트렌드 분석 (유튜브 포함)	
학습모형	RMSE	파라미터	RMSE	파라미터	RMSE	파라미터	RMSE	파라미터
앙상블	(1) 1.6920	Ridge, K-NN, ANN*	(1) 2.0067	Ridge, K-NN, ANN*	(1) 2.2845	Ridge, K-NN, ANN*	(1) 1.7486	K-NN, 의사결정나무, ANN*
	(2) 1.6747	K-NN, SVM, ANN*	(2) 1.9276	Ridge, K-NN, ANN*	(2) 2.2751	Ridge, SVM, ANN*	(2) 1.7057	K-NN, 의사결정나무, ANN*

(1) 후보1을 종속변수로 학습한 모형
(2) 후보2를 종속변수로 학습한 모형
(3) 후보1과 후보2를 함께 종속변수로 학습한 모형
* 앙상블 모형의 파라미터 항목은 학습에 사용된 모형들을 의미함

4.4. 분석 결과

각 데이터 셋에서 최고의 성능을 발휘하는 모델을 학습한 후 2022년 3월 1일부터 2022년 3월 7일까지의 데이터를 사용하여 2022년 3월 3일부터 2022년 3월 9일까지의 지지율을 예측하였다. 각 데이터셋의 자세한 분석 결과는 <표 4>부터 <표 7>까지 제시하였다. 표에서 '모형 RMSE' 항목은 각 모형이 학습 단계에서 보여준 최고의 RMSE를 의미한다. '3/3 ~ 3/9' 항목은 모형이 예측한 해당 기간의 지지율이며, 후보 1과 후보 2의 지지율의 합이 100%가 되도록 조정하였다. '예측 RMSE' 항목은 3/3 ~ 3/9 기간 동안 예측된 지지율과 실제 지지율과의 차이로 계산한 RMSE 값으로 해당 지표가 낮을수록 성능이 더 좋음을 의미한다.

유튜브 분석 데이터로 학습한 모형에서는 K-NN(RMSE= 1.6056)이 가장 성능이 높은 것으로 나타났으며, 앙상블 모형(후보 1: RMSE = 1.6920, 후보 2: RMSE = 1.6747)이 뒤를 이었다. 예측 결과가 가장 좋은 모델은 K-NN(후보 1: 예측 RMSE = 0.9735, 후보 2: 예측 RMSE = 0.9751)이었으며, ANN의 3월 4일(3/4) 예측값을 제외하고는 모든 모형의 예측값이 후보 지지율의 순위를 올바르게 예측하였다. 자세한 예측 결과는 <표 4>와 같다.

〈표 4〉 유튜브 분석 데이터로 예측한 깜깜이 기간과 대선결과

학습모형	모형 RMSE	구분	3/3	3/4	3/5	3/6	3/7	3/8	3/9	예측 RMSE
선형회귀 모형	1.9573	후보 1	45.73	49.68	46.93	46.98	47.48	46.99	48.52	2.0700
	1.9573	후보 2	54.27	50.31	53.07	53.02	52.52	53.01	51.48	2.0707
릿지 선형회귀 모형	1.9390	후보 1	45.89	49.96	47.28	46.92	47.74	47.41	48.32	2.0314
		후보 2	54.11	50.04	52.72	53.07	52.26	52.58	51.68	2.0324
K-NN	**1.6056**	후보 1	47.86	48.57	48.55	48.28	48.57	48.40	48.55	**0.9735**
		후보 2	52.14	51.43	51.45	51.72	51.43	51.60	51.45	**0.9751**
의사결정 나무	2.2820	후보 1	48.22	49.01	48.22	48.22	48.22	48.22	48.22	1.0353
		후보 2	51.78	50.99	51.78	51.78	51.78	51.78	51.78	1.0372

학습모형	모형 RMSE	구분	3/3	3/4	3/5	3/6	3/7	3/8	3/9	예측 RMSE
SVM	1.9679	후보 1	46.87	49.42	46.66	46.54	47.72	47.43	48.24	1.8390
	1.9291	후보 2	53.13	50.58	53.34	53.46	52.28	52.57	51.76	1.8401
ANN	1.9135	후보 1	45.93	50.49	47.70	46.74	48.11	47.99	48.12	2.1274
		후보 2	54.07	49.51	52.30	53.26	51.89	52.01	51.88	2.1284
앙상블	**1.6920**	후보 1	46.59	49.29	47.43	47.30	47.89	47.53	48.46	1.6356
	1.6974	후보 2	53.41	50.71	52.57	52.70	52.11	52.47	51.54	1.6367
메타분석/ 대선결과		후보 1	49.52	47.27	48.20	49.11	48.08	48.40	49.62	
		후보 2	50.47	52.72	51.80	50.88	51.91	51.59	50.37	

소셜 분석 데이터로 학습한 모형에서는 K-NN(RMSE= 1.8028)이 가장 높은 학습 성능을 보여주었으며, 이어서 릿지 선형회귀모형(RMSE = 1.8474)과 ANN(후보1: RMSE = 1.8655)이 근사한 성능을 보여주었다. 예측 결과는 K-NN(후보 1: 예측 RMSE = 0.7641, 후보 2: 예측 RMSE = 0.7660)이 가장 좋았으며, 이는 유튜브 분석 데이터로 학습한 K-NN보다 정확한 예측 결과이다. 다만, 소셜 분석 데이터의 대부분의 모형은 예측값에서 후보의 지지율 순위를 맞추지 못하였으며, 각 모형의 예측 RMSE 편차가 크게 나타났다. 자세한 예측 결과는 <표 5>와 같다.

〈표 5〉 소셜 분석 데이터로 예측한 깜깜이 기간과 대선결과

학습모형	모형 RMSE	구분	3/3	3/4	3/5	3/6	3/7	3/8	3/9	예측 RMSE
선형회귀 모형	2.6943	후보 1	46.96	49.59	62.70	59.94	52.20	49.99	51.68	7.2011
	2.6943	후보 2	53.04	50.41	37.30	40.06	47.80	50.01	48.32	7.2007
릿지 선형회귀 모형	**1.8474**	후보 1	48.77	48.72	55.44	52.43	49.79	48.59	49.75	3.1389
		후보 2	51.23	51.28	44.56	47.57	50.21	51.41	50.25	3.1389
K-NN	**1.8028**	후보 1	48.64	48.47	48.47	48.47	48.64	48.64	48.61	**0.7641**
		후보 2	51.36	51.53	51.53	51.53	51.36	51.36	51.39	**0.7660**
의사결정 나무	2.2115	후보 1	48.55	48.55	48.55	48.55	48.55	48.55	48.55	0.7933
		후보 2	51.45	51.45	51.45	51.45	51.45	51.45	51.45	0.7952
SVM	1.9686	후보 1	47.18	50.15	52.50	55.97	50.67	50.42	47.70	3.6550
	1.9726	후보 2	52.82	49.85	47.50	44.03	49.33	49.58	52.30	3.6557
ANN	**1.8655**	후보 1	49.17	49.43	58.61	53.73	50.12	47.78	50.27	4.4613
		후보 2	50.83	50.57	41.39	46.27	49.88	52.22	49.73	4.4611

학습모형	모형 RMSE	구분	3/3	3/4	3/5	3/6	3/7	3/8	3/9	예측 RMSE
앙상블	2.0067	후보 1	48.01	48.96	55.65	53.68	50.20	49.01	50.14	3.5178
	1.9276	후보 2	51.99	51.04	44.35	46.32	49.80	50.99	49.86	3.5175
메타분석/ 대선결과		후보 1	49.52	47.27	48.20	49.11	48.08	48.40	49.62	
		후보 2	50.47	52.72	51.80	50.88	51.91	51.59	50.37	

유튜브 검색을 제외한 검색어 트렌드 분석 데이터로 학습한 모형에서는 ANN(RMSE = 2.2513)이 가장 성능이 좋았으며, 다른 모형들도 비슷한 학습 성능을 보여주었으나, 유튜브 분석 데이터와 소셜 분석 데이터로 학습한 모형들보다 학습 성능이 대체로 저하된 것으로 나타났다. 검색어 트렌드 분석 데이터로 학습한 모형 중 예측 결과가 가장 좋은 경우는 K-NN(후보 1: 예측 RMSE = 0.8330, 후보 2: 예측 RMSE = 0.8347)이었으며, 자세한 예측 결과는 <표 6>과 같다.

모형 학습에 사용된 머신러닝 기법들은 유튜브 검색을 제외한 검색어 트렌드 분석 데이터로 학습했을 때보다 유튜브 검색을 포함한 검색어

〈표 6〉 검색어 트렌드 분석 데이터(유튜브 검색 제외)로 예측한 깜깜이 기간과 대선결과

학습모형	모형 RMSE	구분	3/3	3/4	3/5	3/6	3/7	3/8	3/9	예측 RMSE
선형회귀 모형	2.3583	후보 1	49.45	50.99	49.67	47.61	46.27	47.32	48.15	1.8864
	2.3583	후보 2	50.55	49.01	50.33	52.38	53.73	52.68	51.85	1.8875
릿지 선형회귀 모형	2.2531	후보 1	49.05	49.88	51.00	48.86	47.56	47.85	49.00	1.5067
		후보 2	50.95	50.12	49.00	51.14	52.44	52.15	51.00	1.5072
K-NN	2.2658	후보 1	48.54	48.51	49.21	49.21	48.65	48.65	48.65	**0.8330**
		후보 2	51.46	51.49	50.79	50.79	51.35	51.35	51.35	**0.8347**
의사결정 나무	2.6285	후보 1	48.79	49.57	46.78	46.78	48.23	48.79	48.23	1.4833
		후보 2	51.21	50.43	53.22	53.22	51.77	51.21	51.77	1.4847
SVM	2.2783	후보 1	48.65	49.73	50.28	47.87	46.70	47.05	48.74	1.5692
	2.2495	후보 2	51.35	50.27	49.72	52.13	53.30	52.95	51.26	1.5700
ANN	**2.2513**	후보 1	48.91	49.99	50.23	48.03	46.64	47.17	48.56	1.5961
		후보 2	51.09	50.01	49.77	51.97	53.36	52.83	51.44	1.5970
앙상블	2.2845	후보 1	49.99	49.79	50.00	50.28	50.39	50.28	50.09	1.6967
	2.2751	후보 2	50.01	50.21	50.00	49.72	49.61	49.72	49.91	1.6963
메타분석/ 대선결과		후보 1	49.52	47.27	48.20	49.11	48.08	48.40	49.62	
		후보 2	50.47	52.72	51.80	50.88	51.91	51.59	50.37	

〈표 7〉 검색어 트렌드 분석 데이터(유튜브 검색 포함)로 예측한 깜깜이 기간과 대선결과

학습모형	모형 RMSE	구분	3/3	3/4	3/5	3/6	3/7	3/8	3/9	예측 RMSE
선형회귀 모형	2.1067	후보 1	49.62	50.52	48.95	47.49	47.36	48.92	50.05	1.4509
	2.1067	후보 2	50.38	49.48	51.05	52.51	52.64	51.08	49.95	1.4505
릿지 선형회귀 모형	2.0667	후보 1	49.34	49.99	49.47	47.89	47.66	48.85	50.20	1.2684
		후보 2	50.66	50.01	50.53	52.11	52.34	51.15	49.80	1.2678
K-NN	1.7581	후보 1	48.52	48.82	47.89	47.89	47.89	48.88	48.88	**0.9108**
		후보 2	51.48	51.18	52.11	52.11	52.11	51.12	51.12	**0.9120**
의사결정 나무	**1.7192**	후보 1	48.53	49.57	49.57	49.57	48.53	48.53	48.53	1.1822
		후보 2	51.47	50.43	50.43	50.43	51.47	51.47	51.47	1.1836
SVM	2.2133	후보 1	48.77	49.53	49.71	47.84	47.49	48.12	49.55	1.1961
	2.1157	후보 2	51.23	50.47	50.29	52.16	52.51	51.88	50.45	1.1962
ANN	1.9157	후보 1	48.42	48.86	48.27	48.74	49.10	48.76	48.66	0.9223
		후보 2	51.58	51.14	51.73	51.26	50.90	51.24	51.34	0.9238
앙상블	1.7486	후보 1	48.89	49.61	48.88	48.38	48.18	48.98	49.38	1.0190
	1.7057	후보 2	51.11	50.39	51.12	51.62	51.82	51.02	50.62	1.0194
메타분석/ 대선결과		후보 1	49.52	47.27	48.20	49.11	48.08	48.40	49.62	
		후보 2	50.47	52.72	51.80	50.88	51.91	51.59	50.37	

트렌드 분석 데이터로 학습했을 때 모든 모형에서 학습 성능이 증가함을 확인하였다. 깜깜이 기간을 예측한 결과 역시 유튜브 검색을 제외한 검색어 트랜드 데이터보다 유튜브 검색을 포함한 검색어 트렌드 분석 데이터에서 두 후보의 지지율 순위를 더 정확히 예측하였다. 또한 대선결과의 예측 RMSE는 전반적으로 감소한 것을 확인할 수 있었고, 이에 대한 자세한 예측 결과는 <표 7>과 같다.

5. 토론 및 결론

5.1. 연구결과 요약

본 연구는 유튜브 분석, 소셜 분석, 검색어 트렌드 분석 데이터를 기반으로 머신러닝 기법을 학습하여 깜깜이 기간과 대선결과를 예측하였다. 우선, 유튜브 영상의 긍·부정 동영상 수, 조회수, 좋아요 수, 댓글 수 변수만을 활용한 유튜브 분석 결과는 기존의 대선결과를 예측하는 방법들의 결과를 보완할 수 있었으며, 유튜브 분석 결과만으로도 실제 선거 결과를 성공적으로 예측할 수 있었다. 유튜브 분석과 소셜 분석 결과의 경우, 메타분석 결과와 비슷한 패턴(높은 상관관계)과 일치도(낮은 RMSE)를 보이며 기존의

여론조사 결과보다 성능이 좋은 여론 예측모델을 구축할 수 있었다. 유튜브 분석 결과는 실제 여론의 흐름을 실시간(real-time)으로 알려주는 반면, 여론조사 결과는 약 2일 전의 여론을 공표하므로 두 결과에는 약 2일정도의 시간적 보정이 필요함을 확인하였다. 유튜브 채널과 SNS 채널은 실제 여론(혹은 여론조사 결과)에 비해 상대적으로 진보쪽으로 편향된 반면, 포털 서비스 채널(검색어 트렌드 분석)은 실제 여론(혹은 여론조사 결과)에 비해 상대적으로 보수쪽으로 편향되었음을 파악할 수 있었다.

본 연구의 분석 결과, 본 연구의 데이터 셋인 유튜브 분석과 소셜 분석, 유튜브 데이터를 제외한 검색어 트렌드 분석, 유튜브 데이터를 포함한 검색어 트렌드 분석의 데이터는 각각 다른 결과를 보여주었다. 유튜브 분석 데이터의 경우, 전반적으로 학습 성능이 우수하였고, 예측 결과에서도 후보의 지지율 순위를 가장 정확하게 예측하였다. 반면, 소셜 분석 데이터의 경우, 유튜브 분석 데이터보다 높은 학습 성능을 보여주었으나, 후보의 지지율 순위를 정확히 예측하지는 못하였다. 또한, 다른 데이터 셋에 비해 모형간 예측 RMSE의 편차가 가장 크게 나타났으며, 학습 성능이 좋지 못했던 모형의 예측 성능이 가장 좋게 나타나는 등 학습 성능만 가지고 예측 모형을 선택하기에는 다소 신뢰성이 떨어지는 것으로 나타났다. 검색어 트렌드 분석 데이터의 경우, 유튜브 데이터를 제외하였을 때는 유튜브 분석과 소셜 분석 데이터에 비해 학습 성능, 예측 성능이 모두 저하되었으나, 유튜브 데이터를 포함하였을 때는 학습 성능과 예측 성능이 모두 상승하는 결과를 보여주었다. 이를 바탕으로 검색어 트렌드 분석 데이터에 유튜브 데이터가 결합될 경우 모형의 예측 성능을 높일 수 있음을 확인하였다.

결과를 종합해 보면, 전통적인 여론조사가 가지고 있는 단점을 보완하기 위해 SNS나 유튜브를 활용한 여론조사 방법이 여론을 결집하고 예측할 수 있는 중요한 수단으로 자리를 잡았다는 사실을 알 수 있다. 특히, 최근 정치 맥락에서의 홍보 활동이 유튜브로 옮겨간 사실을 반영하듯 유튜브 데이터를 분석에 반영하였을 때 학습 성능이 올라가고 정확한 결과를 예측하는 것을 알 수 있다.

5.2. 연구의 시사점

본 연구의 이론적 의의는 다음과 같다. 첫째, 본 연구는 유튜브를 이용한 정치 활동이 본격적으로 활성화된 이후 치러진 대선에서 유튜브의 정치 관련 컨텐츠와 이에 대한 반응이 국내 정치 여론에 어떠한 영향을 주는지를 최초로 확인하였다는 점에서 의의를 가진다. 특히, 전통적인 여론조사의 대안으로 활용된 SNS와 검색어 트렌드를 비교하여 유튜브가 여론 형성에 중요한 요인으로 작용한다는 사실을 밝혀내었다는 점에서 이론적 의의를 가진다. 둘째, 본 연구는 국내 정치 맥락의 선거에서도 유튜브를 활용하여 여론조사를 예측할 수 있다는 가능성을 보여주었다. 본 연구에서는 유튜브 등 동영상 플랫폼의 비중이 점차 커지는 시점에서 간단한 유튜브 분석 결과만으로도 실제 여론(혹은 여론조사 결과)에 근접한 결과를 쉽게 얻을 수 있음을 확인하였는데, 이는 정치를 포함한 현대인의 정보를 습득하는 방식이 TV, 라디오와 같은 전통적인 미디어 매체에서 SNS, 동영상 플랫폼과 같은 인터넷 미디어 매체로 옮겨가고 있는 현상과 연관 지어 생각할 수 있다. 셋째, 국내 정치권 이슈에 대해

유튜브 분석, 소셜 분석, 검색어 트렌드 분석 데이터가 가지고 있는 특징을 확인하고, 이를 학습한 머신러닝 기법의 성능을 비교하였다는 점에서 이론적 의의를 가진다. 유튜브 분석, 소셜 분석, 검색어 트렌드 분석은 모두 현대인이 정치적 관심과 정보를 습득할 수 있는 수단이지만 각 플랫폼을 이용하는 목적이 상이하여 데이터 간 차별적인 특징이 존재함을 알 수 있다. 특히, 본 연구에서는 유튜브 분석 데이터의 포함 여부에 따른 예측의 정확성에 차이가 드러났다는 점에서 의의를 가진다고 할 수 있다.

한편, 본 연구의 실무적 의의는 다음과 같다. 첫째, 본 연구의 결과는 선거 후보자 및 지지자의 관점에서 유튜브 분석 데이터를 활용하여 실시간으로 여론을 확인할 수 있다는 점을 보여준다. 전통적인 여론조사는 조사와 집계 기간으로 인해 발표 시점의 여론과 실제 여론의 시점 간에 차이가 나타났다. 하지만, 유튜브 분석 데이터를 활용한다면 현시점의 여론을 실시간으로 확인할 수 있고, 이를 바탕으로 즉각적인 여론 대응이 가능할 것으로 판단된다. 둘째, 본 연구는 선거 후보자 및 지지자의 관점에서 유튜브 분석 데이터를 바탕으로 깜깜이 기간 동안에도 지지율 예측이 가능함을 밝혔다. 본 연구에서 제안한 예측모형을 바탕으로 여론조사가 공표되지 못하는 깜깜이 기간에도 지지율에 영향을 주는 정치적 이슈에 대한 민첩한 대응과 보다 적극적이고 전략적인 선거활동이 가능할 것으로 기대한다. 셋째, 언론기관 및 기타 여론조사 기관의 관점에서 유튜브 분석 데이터가 소셜 분석과 검색어 트렌드 분석 데이터보다 학습 성능이 좋았다는 분석 결과를 근거로 기존의 선거 결과 알고리즘에 유튜브 분석 데이터를 추가적으로 활용하여 보다 정확한 결과를 예측할 수 있을 것으로 기대한다.

5.3. 연구의 한계점 및 향후 연구방향

본 연구의 한계점 및 향후 연구방향은 다음과 같다. 첫째, 본 연구는 대선 후보를 두 명으로 축소하여 분석을 진행하였다는 점에서 한계점을 가진다. 앞으로의 선거에서 유의미한 지지율을 가지는 제3의 후보가 등장하거나 현재와 같은 양당 체제가 약화된다면 본 연구에서 고안된 대선결과 예측 방법론을 사용할 수 없게 된다. 따라서, 향후 연구에서는 세 명 이상의 후보를 상정한 방법론을 제안하여 보다 현실적인 예측이 가능할 수 있기를 기대한다. 둘째, 유튜브 분석 데이터의 경우 영상의 내용과 댓글이 반영되지 않았기 때문에 여론의 방향성은 예측할 수 없다. 이는 유튜버가 시청자의 이목을 끌기 위해 영상의 제목을 과도하게 자극적으로 설정한 경우라면, 영상의 내용에 대한 비판과 반박 댓글이 남겨져 다양한 상황에 대한 대응이 어려울 수 있기 때문에 향후 연구에서는 이 점을 보완하여 보다 정확한 예측을 가능하게 할 필요가 있다. 셋째, 유튜브 분석 과정에서 메타분석 결과와 비교하여 유튜브 분석 결과에 2일을 더하여 보정하는 방법을 사용하였다. 이는 메타분석 기술의 발달에 따라 상황이 변할 경우, 결과의 신뢰성을 위협할 수 있는 가정으로 향후 연구에서는 가정의 변화에 대한 추가적인 민감도 분석(sensitivity analysis)을 진행하여 결과의 견고성(robustness) 및 일반화 가능성을 제고할 필요가 있다. 넷째, 본 연구는 선거 결과를 제대로 예측하지 못하는 여론조사 결과의 부정확함에서 비롯되었으나, 유튜브 채널 이용자가 전체 유권자를 대표하기에는 무리가 있다. 따라서, 유튜브 분석 결과에

대한 면밀한 보정과 대표성을 확보할 수 있는 방안이 추가로 필요할 것으로 판단된다. 다만, 기존의 분석 방법 및 데이터에 본 연구에서 제안하는 방법론과 유튜브 데이터가 결합한다면, 좀 더 정확한 결과를 예측할 수 있을 것으로 기대한다.

참고문헌(References)

[국내 문헌]

강은경, 정연식, 양선욱, 권지윤, 양성병. (2022). MIS Quarterly 연구동향 탐색: 토픽모델링 및 키워드 네트워크 분석 활용. *지능정보연구*, 28(2), 207-235.

고정애, *안철수 지지층 38.3%는 이재명, 37.7%는 윤석열 찍었다*, 중아일보, 2022, Available at https://www.joongang.co.kr/article/25058749.

공윤엽, 김동하, 최현성, 황재호. (2022). 소셜 빅데이터 기반의 20 대 대선후보 키워드 분석을 통한 선거전략 수립. *인문사회 21*, *13*(1), 3191-3206.

권혁남. (2001). 16 대 총선 여론조사의 문제점 및 개선방안: 출구조사와 무응답률 문제를 중심으로. *언론과학연구*, *1*(1), 46-74.

김가현, *'깜깜이 기간' 尹으로 기운 포심… 지지 격차는 0.9%~5.2%P 요동*, 서울신문, 2022, Available at https://www.seoul.co.kr/news/newsView.php?id=20220310006005.

김양원, 송경재, *[열린라디오 YTN] 여론조사 과잉시대…대선후보 여론조사 믿어도 될까*, YTN 라디오 FM 94.5(20:20~21:00), 2021, Available at https://www.ytn.co.kr/_ln/0101_202111150859285538.

김종훈, *연령불문 인기 1위 '유튜브' 50대 증가하는 까닭은?*, 조선일보, 2019, Available at http://www.dizzotv.com/site/data/html_dir/2019/05/16/2019051680230.html.

김찬우, 박효찬, 박한우. (2017). 2017 년 대통령 후보수락 연설 유튜브 동영상의 댓글망과 의미망 분석. *Journal of The Korean Data Analysis Society (JKDAS)*, *19*, 1379-1390.

김형수. (2020). *Step by Step 비즈니스 머신러닝 in 파이썬*. 서울: 프레딕스.

바이브컴퍼니. *Sometrend Biz*, 2022, Available at https://biz.some.co.kr/intro.

박병언, 임규건. (2015). 일반영향요인과 댓글기반 콘텐츠 네트워크 분석을 통합한 유튜브(Youtube) 상의 콘텐츠 확산 영향요인 연구. *지능정보연구*, 21(3), 19-36.

박상현, 김성훈, 정승화. (2020). 유튜브 정치· 시사 채널 이용이 정치사회화에 미치는 영향. *한국콘텐츠학회논문지*, *20*(9), 224-237.

박석봉, 정주호. (2021). 빅데이터 분석기술을 활용한 병역제도 여론 연구: 인터넷 뉴스, SNS를 중심으로. *한국국가안보국민안전학회지*, 13, 7-30.

박영석, *[그래픽] 역대 대선 여론조사 및 득표율*, 연합뉴스, 2021, Available at https://www.yna.co.kr/view/GYH20211127000600044.

박중현, *유튜브의 인기 비결… 무엇이 현대인을 사로잡았나*, 똑똑, 2021, Available at https://www.dokdok.co/post/youtube-3.

배정환, 손지은, 송민. (2013). 텍스트 마이닝을 이용한 2012 년 한국대선 관련 트위터 분석. *지능정보연구*, 19(3), 141-156.

송화영, 박세정, 박한우. (2020). 2020 년 국회의원 선거 기간의 유튜브 빅데이터 분석. *Journal of The Korean Data Analysis Society*, *22*(5), 2063-2074.

안지현, *'오락가락' 여론조사 결과 왜?.. '정치 저관여층'에 달렸다*, JTBC, 2022, Available at https://news.v.daum.net/v/20220223190803040?s=print_news.

이광춘, *유튜브에 쌓인 데이터들… 대선 결과를 예측할까*, 오마이뉴스, 2022, Available at http://www.ohmynews.com/NWS_Web/View/at_pg_w.aspx?CNTN_CD=A0002816615.

이서영, 권상집. (2019). 트위터 메시지 분석을 통한 선거 결과 예측 고찰: 18 대 대선을 중심으로. *한국콘텐츠학회논문지*, *19*(4), 174-186.

이수범, 강연곤. (2013). 국내 일간지의 트위터 이슈에 관한 보도 프레임 분석: 정치적 소통과 여론 형성이라는 관점을 중심으로. *한국언론학보*, *57*(1), 28-53.

이승종, *[심층출구조사 분석]①49.3% "후보 만족스럽지 않지만 투표"*, KBS뉴스, 2022, Available at https://news.kbs.co.kr/news/view.do?ncd=5412126.

이예나, 최은정, 김명주. (2018). 대선후보의 SNS 평판이 선거결과에 미치는 영향 분석-19 대 대선을 중심으로. *디지털융복합연구*, *16*(2), 195-201.

임예인, "2022 대선, 여론조사의 패배를 묻다: 괴골 인터뷰 2/2", ㅍㅍㅅㅅ, 2022.05.11., Available at https://ppss.kr/archives/254019.

전창훈, *선거 100일전 웃은 후보가 결국 이겼다*, 부산일보, 2021, Available at http://www.busan.com/view/bstoday/view.php?code=2021112817571734900.

정일권, 김상연. (2021). 해석적 선거 여론조사 보도: 기사 작성과 취재 실천 방안. *방송문화연구*, *33*(2), 9-46.

정지선, 김동성, 김종우. (2015). 온라인 언급이 기업 성과에 미치는 영향 분석: 뉴스 감성 분석을 통한 기업별 주가 예측. *지능정보연구*, *21*(4), 37-51.

중앙선거관리위원회, *제20대 대통령선거(재보궐 포함) 투표구별 개표결과*, 중앙선거관리위원회 선거1과, 2022, Available at https://www.nec.go.kr/site/nec/ex/bbs/View.do?cbIdx=1129&bcIdx=175508.

중앙선거관리위원회, *공직선거법 제108조(여론조사의 결과공표금지 등)*, 중앙선거관리위원회 법제과, 2017, Available at https://www.law.go.kr/%EB%B2%95%EB%A0%B9/%EA%B3%B5%EC%A7%81%EC%84%A0%EA%B1%B0%EB%B2%95/%E7%AC%AC108%E6%A2%9D.

최문열, *[이슈 따라잡기] 3일부터 여론조사 공표금지, 배경과 이유*, UPDOWNNEWS, 2022, Available at http://www.updownnews.co.kr/news/articleView.html?idxno=300555.

최필선, 민인식. (2013). 18 대 대통령 선거 여론조사의 기관별 정확성 측정 및 비교. *조사연구*, *14*(2), 1-18.

하상현, 노태협. (2020). SNS 기반 여론 감성 분석. *The Journal of the Convergence on Culture Technology (JCCT)*, *6*(1), 111-120.

하승태. (2012). 선거여론조사와 후보별 보도량 분석: USA Today의 미대선 경선 보도를 중심으로. *지역과 커뮤니케이션*, *16*(2), 115-140.

홍원식, 한균태, 서영남. (2009). 대선보도와 여론조사-여론조사 공표 금지 조항 개정을 중심으로. *정치커뮤니케이션 연구*, *12*, 245-277.

MBC. *[여론M] 여론조사를 조사하다*, 2022, Available at http://poll-mbc.co.kr/.

[국외 문헌]

Chen, Y., & Wang, L. (2022). Misleading political advertising fuels incivility online: A social network analysis of 2020 US presidential election campaign video comments on YouTube. *Computers in Human Behavior*, *131*, 107202.

Krishna, A., Zambreno, J., & Krishnan, S. (2013, December). Polarity trend analysis of public sentiment on YouTube. In *Proceedings of the 19th international conference on management of data* (pp. 125-128).

Lee, K., *9 Informative Infographics To Guide Your Visual Content Marketing*, Buffer, 2014, Available at https://buffer.com/resources/infographics-visual-content-marketing/.

Mehrabian, A. (1981). *Silent messages: Implicit communication of emotions and attitudes*, Wadsworth Pub.

Patterson, T. E. (2005). Of polls, mountains: US journalists and their use of election surveys. *Public Opinion Quarterly*, *69*(5), 716-724.

Ridout, T. N., Franklin Fowler, E., & Branstetter, J. (2010, August). Political advertising in the 21st century: The rise of the YouTube ad. In *APSA 2010 Annual Meeting Paper*.

Shevtsov, A., Oikonomidou, M., Antonakaki, D., Pratikakis, P., & Ioannidis, S. (2020). Analysis of Twitter and YouTube during USelections 2020. *arXiv preprint arXiv:2010.08183*.

Zolghadr, M., Niaki, S. A. A., & Niaki, S. T. A. (2018). Modeling and forecasting US presidential election using learning algorithms. *Journal of Industrial Engineering International*, *14*(3), 491-500.

Abstract

Analysis of public opinion in the 20th presidential election using YouTube data

Eunkyung Kang* · Seonuk Yang* · Jiyoon Kwon* · Sung-Byung Yang**

Opinion polls have become a powerful means for election campaigns and one of the most important subjects in the media in that they predict the actual election results and influence people's voting behavior. However, the more active the polls, the more often they fail to properly reflect the voters' minds in measuring the effectiveness of election campaigns, such as repeatedly conducting polls on the likelihood of winning or support rather than verifying the pledges and policies of candidates. Even if the poor predictions of the election results of the polls have undermined the authority of the press, people cannot easily let go of their interest in polls because there is no clear alternative to answer the instinctive question of which candidate will ultimately win. In this regard, we attempt to retrospectively grasp public opinion on the 20th presidential election by applying the 'YouTube Analysis' function of Sometrend, which provides an environment for discovering insights through online big data. Through this study, it is confirmed that a result close to the actual public opinion (or opinion poll results) can be easily derived with simple YouTube data results, and a high-performance public opinion prediction model can be built.

Key Words : Opinion analysis, YouTube data, 20th presidential election, machine learning, public opinion prediction model

Received : August 29, 2022 Revised : September 8, 2022 Accepted : September 13, 2022
Corresponding Author : Sung-Byung Yang

* Department of Big Data Analytics, Kyung Hee University
** Corresponding author: Sung-Byung Yang
Department of Business Administration/Big Data Analytics, Kyung Hee University
26 Kyungheedae-ro, Dongdaemun-gu, Seoul 02447, Korea
Tel: +82-2-961-9548, E-mail: sbyang@khu.ac.kr

저 자 소 개

강은경
경희대학교 의료경영학과 석사학위 취득 후 동대학교 일반대학원 빅데이터응용학과 박사과정에 재학 중이며, O2O 서비스 품질이 고객만족도 및 고객충성도에 미치는 영향에 관한 연구를 진행한 바 있다. 주요 관심분야는 지식경영, IT 경영, 비대면 마케팅, 공유경제, 비즈니스 애널리틱스, 인과추론 등이다.

양선욱
경희대학교 일반대학원 빅데이터응용학과에서 석사학위를 취득하였으며, 유튜브 중간광고 설정이 영상이탈의도와 프리미엄 구매의도에 미치는 영향에 관한 연구를 진행한 바 있다. 주요 관심분야는 빅데이터 분석, 소비자 행동 등이다.

권지윤
경희대학교 일반대학원 빅데이터응용학과 석사과정에 재학중이며, 유튜브 실시간 방송 시청자의 지속시청의도 및 유료후원의도에 영향을 미치는 요인에 관한 연구를 진행중에 있다. 주요 관심분야는 데이터 분석, 데이터 사이언스 등이다.

양성병
경희대학교 경영학과/빅데이터응용학과 교수로 재직 중이며, 주요 관심분야는 빅데이터 분석, 온라인 리뷰, 고객관계관리, 지식경영, 온라인 커뮤니티, 전자상거래, 스마트 투어리즘 등이다.