빅데이터, 여론조사 대체할 수 있나

빅데이터 만능 아니고 여론조사도 언론 활용하기 나름

정일권 ／ 광운대 미디어영상학부 교수

지지율 위주의 현행 선거 여론조사 보도는 여러 가지 폐단을 낳지만 긍정적인 측면이 전혀 없는 것은 아니다. 최소한 박빙의 판세에서는 여론조사가 선거에 대한 유권자의 관심을 불러일으켜 투표율 상승에 기여한다. 〈저작권자 ©연합뉴스, 무단복제 및 무단사용 금지〉

우리나라는 물론이고 세계 여러 나라에서 국민투표 혹은 선거 결과에 대한 여론조사의 예측이 빗나가는 경우가 속출하고 있다. 트럼프의 승리로 끝난 미국 대선, 영국의 브렉시트 찬반 국민투표, 그리고 2016년 우리나라 국회의원 선거 결과 예측이 모두 틀렸다. 사실 선거 여론조사의 예측이 빗나간 것이 어제오늘의 일도 아니다. 이제는 투표 결과의 예측 수단으로서 여론조사의 효용성에 의문을 표하는 사람들도 늘어나는 실정이다.

유권자, 후보자, 언론 모두가 원해

선거 여론조사 결과를 보도하는 것이 옳은지 그른지에 대한 논란은 여론조사 결과에 대한 언론사의 보도량이 본격적으로 늘어난 1980년대 이후 줄기차게 이어져 오고 있다. 그럼에도 불구하고 지금까지 선거와 관련해 수많은 여론조사 결과 관련 뉴스가 쏟아져 나오는 이유는 후보자 간 지지율 차이와 변화에 대한 수용자의 높은 관심에 있다. 수용자의 관심은 언론사의 경제적 이익과 사회적 영향력의 토대이기에 언론사 입장에서는 뉴스로서의 가치와 무관하게 이를 뉴스화하지 않기는 어렵다.

한편, 지지율 위주의 현행 선거 여론조사 보도는 여러 가지 폐단을 낳지만 긍정적인 측면이 전혀 없는 것은 아니다. 우선 유권자에게 이전에는 얻을 수 없었던 새로운 유형의 정보를 제공한다. 비록 한계를 지니지만 선거 판세에 대해 주변 사람의 주관적 평가가 아니라 과학적 방법으로 측정된 한층 객관화된 정보를 통해 유권자 스스로 선거 판세를 읽을 수 있게 해준다. 그리고 유권자들은 선거 여론조사 결과를 흥미롭게 받아들이고 다른 사람의 입을 통해서 전달받는 정보보다 더 많이 신뢰한다. 이런 이유에서 일관되지는 않지만 최소한 박빙의 판세에서는 여론조사가 선거에 대한 유권자의 관심을 불러일으켜 투표율 상승에 기여한다.

지지율 위주의 선거 여론조사는 후보자와 언론사에도 도움이 된다. 우선 후보자에게 적절한 캠페인 전략을 세울 수 있도록 정보를 제공하는데, 후보자는 여론조사 결과에서 드러난 지지율의 변화를 분석한 결과를 자신의 발언, 행보 그리고 공약을 결정하거나 수정하는 데 활용할 수 있다. 다음으로 언론사 입장에서 보면, 선거 여론조사 결과는 선거 여론과 관련한 다른 정보와 비교할 때 상대적으로 객관적이기에 공정성 시비에서 한결 자유로울 수 있다. 그뿐만 아니라 더 많은 수용자, 광고주 그리고 투자자를 유인할 수 있기 때문에 수익성의 차원에서도 가치가 높다.

이와 같이 지지율을 나타내는 정보를 보도할

필요성을 충족하면서 예측의 부정확성이라는 문제를 해결하는 방안으로 최근에는 선거 여론조사를 대체할 수 있는 방법에 대한 고민이 깊어지고 있다. 이 와중에서 급격히 부각되는 것이 '빅데이터 분석'이다. 인터넷과 모바일 미디어 환경이 만들어지고 일상적인 대화가 기록되는 소셜미디어의 이용이 늘어나면서 개인의 행위에 대한 정보가 체계적으로 저장되고 있다. 그리고 이 자료를 활용해 사람들이 행동뿐만 아니라 생각까지 엿볼 수 있는 시대가 됐다. 바로 빅데이터 분석이다. 본래 빅데이터 분석은 여론을 조사하는 수단으로 개발된 것은 아니다. 빅데이터가 활용되는 여러 분야 중에 여론조사가 포함된 것일 뿐이다.

빅데이터 분석, 무조건 정확할까?

빅데이터 분석이 여론조사와 구별되는 가장 뚜렷한 방법론적 차이는 분석의 1차 자료가 대상자의 생각이 아니라 행동이라는 점이다. 여론조사는 "누구를 지지하는가?"에 대한 질문에 조사 대상자가 자신이 마음속으로 생각하는 바를 드러내는 방식이지만, 빅데이터 분석은 대상자가 온라인 공간에 남긴 이용 흔적을 분석해 대상자의 생각을 읽어내는 방식이다. 따라서 선거 결과의 예측 정확성 면에서 빅데이터 분석이 여론조사보다 우월하기 위해서는 행동 데이터와 그러한 데이터를 통해 생각을 읽어내는 해석 과정의 우수성이 입증되어야 한다.

우선 빅데이터의 분석 대상인 행동 데이터의 속성을 살펴보자. 선거 결과 예측과 관련한 빅데이터 분석에서 활용되는 데이터는 블로그, 게시판, 뉴스 댓글, 소셜네트워크서비스(SNS) 등의 온라인 공간에 매일 쌓이는 이용자의 이용 흔적이다. 여기에는 특정 후보의 이름이 검색되거나 등장한 빈도, 그 후보를 다룬 기사의 클릭 수와 댓글 등이 포함된다. 빅데이터 분석을 통해 좀 더 정확한 선거 결과 예측이 가능하다고 주장하는 사람들은 이런 행동 데이터가 기존의 여론조사에서 활용되던 데이터가 지닌 문제를 완화한다고 주장한다. 즉 낮은 응답률과 속내를 감춘 응답의 문제를 개선해 더 많은 사람의 진실한 응답을 수집할 수 있는 방법이라는 것이다. 그러나 이 주장은 몇 가지 전제 조건이 충족됐을 때만 타당하다.

여론조사 대비 빅데이터 분석이 결과적으로 응답률이 높다는 주장은 여론조사 담당자들이 집전화나 휴대전화로 전화를 걸면 조사에 응하지 않고 전화를 끊어버리는 60~80%의 사람들도 온라인 공간에 선거에 대한 자신의 생각이 담긴 흔적을 남긴다는 점에 근거한다. 그러나 이는 데이터 이용 흔적을 남긴 사람의 숫자가 늘어난 것을 의미하지, 전체 투표 예정자 중 빅데이터를 통해 분석 가능한 사람의 비중이 높아진 것을 의미하지는 않는다. 애초에 표본으로 선정된 사람들 중 10%만이 조사에 응해 결과적으로 2,000명을 조사한 여론조사와 전체 유권자 중 10%가 남긴 2억 개의 빅데이터 정보를 조사하는 경우를 비교해 보면, 분석하는 데이터의 수는 2,000대 2억이지만 응답률은 10%로 동일하다. 1936년 미국 대통령 선거 당시에 앞서 네 번의 미국 대선 결과를 맞힌 '리터러리 다이제스트(The Literary Digest)'는 무려 237만 명의 응답자를 분석했고 당시 신생 여론조사 회사이던 갤럽은 불과 5,000명만 조사했다. 그러나 결과적으로 갤럽만 프랭클린 루스벨트 대통령의 당선을 정확히 예측했다. 이 사례에서 알 수 있듯이 중요한 것은 응답자 혹은 데이터의 수가 아니라 데이터의 질인 표본의 대표성이다. 응답률이 같은 경우라면 데이터 수가 많다고 빅데이터 분석이 여론조사보다 우수하다고 볼 수는 없다.

반면에 숨은 의도를 찾아낼 수 있다는 주장, 즉 데이터의 진실성 측면에서 우월하다는 주장은 상대적으로 더 설득력이 있다. 이것은 빅데이터

분석의 경우 한 개인이 남긴 여러 단서를 토대로 예측하는 것이기에 결과적으로 한 사람에게 동일한 문항을 여러 번 묻는 것과 같은 효과를 지니기 때문이다. 사람들이 자신의 의중을 아무리 감추려고 해도 여러 차례 질문을 받게 되면 은연중에 자신의 속마음을 내비치기 마련이다. 그러나 여론조사에서 침묵하는 사람들, 예를 들어 조사원의 전화를 아예 안 받거나 받았지만 끊어버리는 사람의 투표 의도를 알 수 없듯이 온라인에서 검색도, 댓글 달기도 하지 않는 사람들, 거짓으로 게시글과 댓글을 남기는 사람들의 투표 의도를 알아내기는 힘들다. 결국 빅데이터 분석을 통해 유권자의 숨은 의도를 찾아낼 수 있지만 이는 매우 제한적일 수밖에 없다.

빅데이터 분석도 그때그때 달라

다음으로 빅데이터 분석이 자료 해석의 측면에서 여론조사에 비해 우수한지를 살펴보자. 선거와 관련해서 여론조사는 후보자별 지지율이나 당선 예정자를 직접적으로 물어보는 방식을 취하기 때문에 조사된 자료가 해석에 따라 의미가 달라질 가능성이 높지 않다. 그러나 빅데이터 분석은 애초부터 데이터 자체를 수집, 저장, 처리하는 과정에 더해 분석가들에 의한 의미 분석 과정을 포함하고 있기 때문에 이 과정에서 데이터의 의미가 달라질 가능성이 높다. 예를 들어, 인터넷에선 어떤 후보를 지지하고 응원하는 글도 있지만 다른 후보 지지자가 '악플'을 다는 경우도 있기 때문에 후보자 관련 글의 검색 수나 게시글과 댓글에서 후보자가 언급된 횟수를 그대로 지지율로 치환하는 분석은 옳지 않다.

물론 이런 방식으로 투표 결과를 예측하는 빅데이터 분석이 없는 것은 아니지만, 많은 경우에 다양한 논리와 방법으로 분석가들이 빅데이터의 의미를 추출하는 과정을 거친 후 선거 결과를 예측한다. 예를 들어 2016년 미국 대선과 관련해 네이트 실버가 운영하는 '파이브서티에이트'는 최신의 데이터에 가중치를 부여했고, 국내 대학에 재직 중인 한 교수는 여론조사와 관련된 3,000개의 빅데이터를 △숨은 표가 없다고 가정했을 때, △1.0%포인트의 가중치를 둘 때, △2.0%포인트의 가중치를 둘 때로 나눠 득표 예측치를 발표하기도 했다. 이와 같이 빅데이터 분석은 어떠한 데이터를 수집하는지, 어떤 데이터에 가중치를 얼마나 두는지에 따라 그 결과가 달라지고 이 과정은 전적으로 개별 분석가의 데이터에 대한 가정과 분석 알고리즘에 따라 결정된다. 따라서 빅데이터 분석은 일관될 수 없다. 그러나 동일한 사안에 대해 여러 다른 빅데이터 분석 결과가 존재할 때 사전적으로 어느 것이 더 예측 정확성이 높은지를 알 수 있는 길은 없다.

이론적으로 빅데이터 분석의 우수성이 입증되지 않음에도 불구하고 어떤 사람들은 '최근의 선거 사례를 보면 여론조사는 틀렸지만 빅데이터 분석은 맞힌 경우가 많지 않으냐'고 반박하기도 한다. 실제로 2016년 미국 대선에서는 미국 주요 언론이 대부분 힐러리 클린턴(이하 힐러리) 전 국무장관의 승리를 예측한 가운데 국내 한 대학의 경영학부 교수는 빅데이터를 분석해 도널드 트럼프(이하 트럼프) 후보를 당선자로 예측했다. 앨런 리트먼 아메리칸대 교수와 LA타임스, 인도의 인공지능(AI) 스타트업 제닉AI(Genic AI)의 선거 예측 프로그램 '모그IA(Mog IA)'도 빅데이터를 분석해 트럼프 당선을 정확히 예측했다. 또한 국내에서는 이미 2014년에 한 업체가 개발한 '한국형 빅데이터 선거 분석 사이트–초이스 2014'가 그해 6월 10일 치러진 지방선거 및 교육감 선거에서 각 후보의 당선 여부를 맞히기도 했다. 그러나 이러한 사례를 빅데이터 분석을 통한 선거 결과 예측의 성공 가능성을 보여주는 것으로 해석하는 것은 다음의 두 가지 이유에서 무리다.

첫째, 같은 사안에 대해 모든 빅데이터 분석이 결과를 정확히 예측한 것은 아니다. 2016년

미국 대선의 경우 무디스 애널리틱스, 로센버그 & 곤살레스 폴리티컬 리포트, 뉴욕타임스 등이 빅데이터를 분석했지만 결과를 맞히지 못했다. 그리고 2015년 영국 총선에서 일부 빅데이터 분석가들이 페이스북과 트위터 등 SNS 분석을 통해 노동당의 집권을 예견했지만 결과는 정반대로 나타났다. 빅데이터를 분석해 결과를 맞혔을 경우 업체 스스로 이를 적극적으로 알리는 경우가 대부분임을 감안하면 아마도 맞힌 경우보다는 틀린 경우가 더 많을 것이다. 위에서 나열한 빅데이터 성공 사례는 빅데이터 분석이 여론조사 방식보다 우수함을 보여주는 것이 아니라 일부 빅데이터 분석을 통한 예측이 여론조사보다 정확했음을 보여줄 뿐이다.

둘째, 당선자를 맞히는 것이 예측의 성공 기준으로 타당하지 않을 수 있다. 2016년 미국 대선 결과를 보면, 트럼프 후보가 선거인단 수를 306명(56.9%), 힐러리 후보가 232명(43.1%)을 얻었고 전체 유권자 지지율에서는 힐러리 후보가 근소하게 앞섰다. 그런데 위에서 성공 사례로 소개된 일부 빅데이터 분석은 트럼프의 승리는 예측했지만, 선거인단 수나 지지율 면에서 여론조사보다 정확하지 않은 것들도 있다. 당선자를 맞혔으나 득표수나 선거인단 확보 수에서는 실제와 더 차이 났던 결과가 당선자를 못 맞혔으나 득표수와 선거인단 예측 면에서 더 정확했던 조사 결과보다 정확하다고 말할 수 있을까? 성공의 기준을 실제 유권자 지지율 혹은 선거인단 수로 바꾼다면 위에서 성공한 예측이라고 한 사례에 대한 평가가 달라질 수 있는 것이다.

사생활 침해와 데이터 조작 위험

빅데이터 분석이 서베이 형식의 기존 선거 여론조사보다 우월하다는 주장은 이론적 측면과 실증적 측면 모두에서 만족할 만한 수준에 이르지 못한다. 게다가 빅데이터 분석으로 여론조사를 대처하려면 몇 가지 위험 요소를 안아야 하는 점이 부담으로 작용한다.

첫째, 사생활 침해 위험이 높아진다. 선거와 관련된 빅데이터 정보의 대표성을 확장하기 위해서는 API 정보 이용이 자유로운 트위터 외의 다른 플랫폼에서도 개인의 행동 정보와 이용자의 프로파일 정보를 보다 자유롭게 얻을 수 있어야 한다. 이를 위해서 빅데이터 분석가들이 주장하는 방법이 수집된 개인 정보를 제3자에게 제공하면서 당사자가 제공 거부 의사를 밝히면 그제야 중단하는 '선(先)동의 후(後)거부 방식'의 채택이다. 이와 같이 이용자의 선택적 동의의 획득 과정에서 옵트인(Opt-in) 방식이 아닌 옵트아웃(Opt-out) 방식을 적용하면 개인정보자기결정권을 침해할 소지가 있고 결과적으로 사생활 침해의 가능성이 높아진다.

다음으로 온라인 공간에서 행동 정보의 양을 결정하는 것은 결국 화제성이므로 선거의 경우 후보자 진영에서 인위적으로 화제성을 키워 데이터가 조작될 가능성이 있다. 특정 후보에 대한 관심, 인지도, 혹은 기대 때문에 후보가 자주 거명되는 것이 아니라 후보 진영에서 제공하는 거짓 정보와 허위 사건(Pseudo Event)에 의해 인위적으로 특정 정보의 양이 많아질 수 있다. 이 경우 데이터 분석을 통해 알 수 있는 것은 선거 관련 여론이 아니라 각 후보 진영의 캠페인 활동성일 뿐이다.

마지막으로 여론조사에 비해 빅데이터 분석은 활용 범위가 제한적이다. 우선, 대선과 달리 총선의 경우 특정 선거구의 데이터를 별도로 확보하기가 쉽지 않다. 지명도 있는 후보가 출마한 지역의 경우 온라인 공간에서도 충분한 양의 정보가 축적될 수 있지만 그렇지 않은 지역의 경우 절대량도 부족하고 편향의 정도도 더 심하기 때문에 이를 통해 결과를 정확하게 예측할 수 없다. 또한 빅데이터 분석을 통해 실시간으로 변하는 여론의 흐름과 방향성은 비교적 쉽게 보여줄 수 있지만 '왜 그렇게

> 빅데이터를 분석해 결과를 맞혔을 경우 업체가 이를 적극적으로 알리는 경우가 대부분임을 감안하면 맞힌 경우보다 틀린 경우가 더 많을 것이다. 빅데이터 성공 사례는 여론조사보다 우수함을 보여주는 것이 아니라 일부 빅데이터 분석을 통한 예측이 정확했음을 보여줄 뿐이다.

변하는지'에 대한 해석의 근거를 찾을 수 없다. 특정 후보의 이름이 등장하는 빈도가 온라인 공간에서 급증할 때 데이터만으로 이 이유를 설명할 수는 없기 때문에 이 과정에서 분석가의 주관이 개입되는 것을 피할 수 없다.

이상과 같은 이유에서 선거 결과를 예측하기 위해 빅데이터 분석으로 여론조사를 대체하는 것은 적절하지 않다. 그러나 그렇다고 지금과 같이 낮은 응답률에 의존하는 여론조사를 그대로 활용하는 것도 문제다. 비록 완전하지는 않지만 숨은 표심을 일부라도 보여준다는 점에서 빅데이터는 여론조사의 결과를 보완할 수는 있다. 선거 결과를 예측할 필요가 있다면 빅데이터로 전체적인 흐름과 실시간 변화를 살피면서 원인 분석이나 해석의 근거를 확보하기 위해 여론조사를 활용하는 상호 보완식 이용법을 고려할 필요가 있다. 그리고 더욱 근본적으로는 지지율을 조사하는 두 방법을 어떻게 활용할 것인지를 고민해야 한다.

무엇보다 언론의 보도가 바뀌어야

미디어의 역할과 관련된 오래된 논쟁, 즉 여론을 '정확하게 반영해야 하는가?' 혹은 '올바른 여론의 형성에 기여해야 하는가?'는 선거 중 지지율 조사 보도에 대한 사회적 기능을 이해하는 데 통찰을 제공한다. 후보자별 지지율 조사를 통해 언론이 특정 시점의 선거 판세를 보여주는 것에만 그 역할을 한정한다면 언론은 민주주의 체제의 유지 기관으로서의 의무를 제대로 수행할 수 없다. 게다가 앞에서도 살펴본 것처럼, 여론조사나 빅데이터 분석을 통해서는 조사와 분석 방법을 아무리 정교하게 설계하고 객관적으로 분석하더라도 선거 결과를 정확하게 예측할 수 없다. 언론은 여론조사와 빅데이터 분석을 활용해 후보자와 유권자 사이의 정치적 소통을 활성화하고 이를 통해 올바른 선거 여론이 형성되는 데 기여할 수 있어야 한다.

언론이 선거 여론조사를 활용해 민주주의 발전에 기여할 수 있는 구체적인 방법을 생각해 보면, 크게 뉴스 내용과 보도 방식의 변화로 나눠 볼 수 있다. 여기에서 선거 여론조사 뉴스 내용의 변화란 두 가지 의미를 내포한다. 첫째는 보도의 소재인 여론조사 항목에 지지율 이외의 내용을 추가하는 것이고, 둘째는 조사한 항목 이외의 내용을 보도 기사에 추가해 수치로 나타난 민심의 의미를 깊이 있게 해석하는 것이다.

다음으로 보도 방식의 변화와 관련해서는 여론조사 결과가 중심인 뉴스가 아니라 보조적인 수단이 되는 뉴스를 만드는 것으로 요약할 수 있다. 달리 말하면, 여론조사자가 조사한 판세라는 정보를 수용자에게 전달하는 뉴스가 아니라 언론이 여론조사자를 활용해 정치인과 유권자에게 전달할 메시지의 정당성 혹은 근거로서 조사 결과를 활용하는 뉴스를 생산해야 한다. 이를 통해 여론조사든 빅데이터 분석이든 판세에 관한 정보를 담은 뉴스가 '누가 1위이고 지지율이 몇%인지'를 보여주기보다 사회적 쟁점과 해결책을 후보자에게서 끌어내고 유권자의 바람을 전달하는 목적을 수행할 수 있는 방향으로 뉴스가 바뀌어야 한다.