비정형 뉴스 데이터 분석을 통한 여론조사 지지율 도출 방안 연구

송종훈*, 최기현*, 구자환**, 김응모***
*성균관대학교 소프트웨어대학
**성균관대학교 사회과학대학
***성균관대학교 사회과학대학 소비자가족학과 / 소프트웨어대학

e-mail : {fullfill, gyunee, jhkoo, ukim}@skku.edu

A Study on Derivation of Approval Rating using Analysis of Unstructured News Data

Jong-Hun Song*, Gi-Hyeon Choi*, Ja-Hwan Koo**, Ung-Mo Kim***
*College of Software, Sungkyunkwan University
** College of Social Sciences, Sungkyunkwan University
*** Dept. of Consumer and Family Sciences, College of Social Sciences / College of Software, Sungkyunkwan University

요 약

현재, 대부분의 여론조사는 전통적 여론 조사 방식을 사용하고 있다. 그러나 이 방식은 온라인 상에서의 여론을 반영하지 못한다는 문제점이 존재한다. 따라서 이를 해결하고 온라인 상에서의 여론을 반영하기 위해, 비정형 뉴스 데이터를 이용한 지지율 분석 방안을 제안하고자 한다. 이 연구에서는 제안 방안을 알아보고 기존의 방식과 비교한 장단점, 시사점, 개선방안 등을 알아봄으로써 새로운 여론조사 방식의 제안을 목적으로 한다.

1. 서론

지금까지의 여론조사는 전통적 여론 조사 방식을 사용하였고 지지율이 매주 집계되었다. 그러나 지지율을 계산하는 방식이 단편적이고, 온라인 상에서의 여론을 반영하지 못하는 경우가 많았다. 따라서 온라인 상에서의 여론을 반영하기 위해 뉴스 데이터를 이용해 지지율을 분석하는 방법을 연구 목표로 삼고자 한다.

기존의 연구 중에는 전통적 여론 조사 방식의 표본 구성과 조사 방법론에 대한 문제점을 찾아낸 연구도 존재하고 이를 해결하기 위해 SNS 상에서의 정치 이슈를 분석함으로써 여론을 알아보려는 연구도 진행되었다.

이런 점들을 고려할 때 사람들이 많이 이용하는 뉴스 데이터를 이용하여 기존의 방식과는 다른 여론 조사 방법론을 개발할 필요성이 대두되었다. 따라서 이 연구에서는 뉴스 데이터를 수집한 후, 해당 뉴스에 분석 대상의 키워드가 포함되어 있는지를 판별한 뒤에 분석 대상이 포함된 뉴스 데이터만 추출하여 그 데이터를 가지고 여론을 파악하는 방법을 통해 연구를 진행하였다.

이 방법론은 기존 방식과 비교해 볼 때 분명한 장점이 존재하지만 부족한 점 또한 분명하다. 향후 이 연구의 데이터 추출과 분석 방식을 발전시킨다면 여론 조사 방법 중 하나로 사용이 가능할 것이다.

본 연구는 2 장에서는 관련 연구들을 알아보며 전통적 여론 조사 방식을 설명하고 문제점들에 대해서 알아본다. 또한 현재 연구되고 있는 비정형 데이터 분석을 통한 여론조사 방식도 알아본다. 3 장에서는 뉴스 데이터를 분석하여 여론조사 지지율을 도출해낼 수 있는 방법을 제안한다. 4 장에서는 이러한 방법으로 도출된 데이터들을 기존 방법들과 비교해보며 장점과 단점을 알아보고 개선방안에 대해서도 탐구해본다. 5 장에서는 전체적인 결론을 내리며 마무리한다.

2. 관련 연구

2.1 전통적 여론 조사 방식

여론 조사가 정치적, 사회적으로 큰 영향력을 미치고 있는 지금 정확한 조사에 대한 필요성이 점점 커지고 있다. 그에 따라 수많은 방식이 탄생하고 사라졌으며 지금 이 순간에도 새로운 방식이 생겨나고 있는 중이다.

* 이 논문은 2016 년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2016R1A6A3A11930831)

초기의 여론 조사는 조사원에 의한 대인 조사였으며 주로 인구 총조사에 의존해 표본 할당을 정했다. 그러나 표본의 편향으로 인해 할당표집 대신에 확률표본 이론에 근거한 표집 기법을 채택하기 시작하였고, 유선 전화의 보급이 시작되면서 전화번호를 기반으로 한 확률적 표본조사 방식으로 넘어가게 되었다.

유선 전화를 통한 방식을 사용하게 되어 조사비용은 저렴해지고, 편리성 또한 증가하였다. 그러나 대상자와 연락이 되지 않을 경우, 재접촉이 필요하고, 조사 대상을 대체해야 할 경우가 생겨 이 경우에는 2.5배 이상의 시간과 비용이 더 소요되는 것으로 밝혀졌다. 또한 2010 년 6 월 지방선거에서 여론 조사 결과가 크게 빗나가면서 기존의 전화번호 표본추출 방식에서 벗어나 임의번호 걸기 방식이 도입되었다.

임의번호 걸기 방식은 지역번호와 국번 이외의 4자리를 무작위로 생성하여 전화조사를 시도하는 방식으로써 전화번호부에 등재되지 않았던 미등재 계층까지도 조사에 반영할 수 있게 되었다. 이를 통해서 확률표집에 더 가까이 갈 수 있게 되었지만, 이 방식에도 문제점이 존재하였다. 이 방식은 없는 번호로 인한 시간적, 비용적 낭비가 상당히 큰 편이고 이에 따라 주어진 조사기간 내에 조사대상자를 충분히 확보하는 것이 어려웠다. 이러한 문제를 극복하기 위해 휴대전화, 또는 휴대전화와 유선전화 등을 조합한 이중 표집틀을 조사에 적용하는 방식이 탄생하였다.

현재 한국에서는 혼합 조사가 주로 이루어지고 있으며, 다양한 방식의 혼합을 통해 도달률의 오차를 줄여나가고 있는 중이다. 그렇지만 미등재 유선번호의 증가와 휴대전화 보급, 낮은 협조율 등으로 여론조사에 문제가 있음에도 불구하고 상대적으로 새로운 조사 방법 적용에는 소극적인 편이다. [1,2]

2.2 전통적 여론 조사 방식의 문제점

이러한 전통적 여론 조사 방식의 표본 구성과 조사 방법론에서 크게 3 가지의 문제점을 찾을 수 있다.

첫 번째 문제점은 표본의 세대별 대표성 문제이다. 전 세대가 비슷한 비중으로 표본이 취합되어야 하지만 대부분의 조사기관에서 고연령층이 과대 표집되어 표본의 불균형이 발생한다. 중앙선거여론조사공정심의위원회에 올라있는 자료들을 보면 세대별로 목표 표본의 수는 거의 비슷하지만 조사된 표본의 수는 5060 세대의 경우 2030 세대의 두 배를 넘는 경우를 쉽게 찾아볼 수 있다. 이러한 경우, 2030 세대의 표본에 두 배 이상의 가중치를 곱해 결과를 산출해야 하는데 표본의 성향이 일반적인 2030 세대를 대표하지 못할 경우 오차의 크기가 크게 증가함을 알 수 있다.

두 번째 문제점은 표본의 성별 대표성 문제 및 정치성향 가중치 부여 문제이다. 목표 표본의 경우에는 성별 간 비슷한 비중으로 표본을 구성하지만 실제 조사 결과를 보면 남성이 여성에 비해 과대 표집된 데이터가 상당수를 차지하고 있음을 알 수 있다. 이렇게 되면 성별 구성비를 보정하기 위해서 여성 측 표본에 가중치를 부여하게 되는데 이 경우에 여성 측 표본의 편차 정도에 따라 오차가 커짐을 알 수 있다. 일부 조사의 경우 정치성향 관련 변수 가중치까지 부여했다고 밝히고 있는데 이 경우에는 오차가 커질 확률이 더욱 더 증가하게 된다.

세 번째 문제점은 낮은 응답율 문제이다. ARS 조사의 경우 평균적으로 한 자리 수의 응답율을 기록하고 있는데, 낮은 응답율은 무작위성의 원칙을 훼손하기 때문에 문제가 된다. 무작위성의 원칙이란 뽑힌 표본들이 전체 모집단을 잘 대표할 수 있도록 무작위적으로 표본을 추출하는 것을 뜻한다. 최초 무작위로 추출한 대상을 접촉할 수 없을 경우 대체할 표본을 위해 다른 번호로 연결하게 되는데 이 경우에 정치적 관심이 높은 사람들이 표본이 될 가능성이 높아지고 전체 유권자 분포와는 오차가 커지게 된다. [3,4,5]

2.3 비정형 데이터 분석을 통한 여론조사 방식

2.2 절에서 알아본 전통적 여론 조사 방식의 문제점들을 보완해 줄 방식으로 현재 비정형 데이터 분석을 통한 여론조사 방식이 연구되고 있다. 여기에는 웹 마이닝 기법이 사용되는데 웹 마이닝 기법이란 데이터 마이닝 기법의 일종으로써 온라인 상에서 발생되는 다양한 데이터 중에서 유용한 정보를 찾고, 이를 분석하는 것을 의미한다. 웹 마이닝 기법은 정보 기술의 발전에 따라 검색 기능과 분석 기능이 향상되고 있으며 정확도 또한 올라가고 있다.

웹 마이닝 기술을 통해 온라인 상에서 발생하는 정치적 이슈에 대한 사람들의 선호도 조사가 가능하고 더 나아가 SNS 상에서의 정치적 이슈 트렌드를 분석하거나 온라인상에서 정치적 의견에 대해 분석하고 예측하는 것도 가능하다. 이런 연구들을 토대로 온라인 여론 조사가 가능할 것으로 보인다. [6,7]

3. 제안 방법

3.1 데이터 수집

2018 년 1 월 1 일부터 2018 년 5 월 31 일까지의 뉴스 데이터 27,180 건을 수집하였다. 이는 높은 뉴스 점유율을 차지하고 있는 네이버의 뉴스 데이터 중에서 조회수가 가장 높은 랭킹 뉴스 데이터를 수집한 것이다. 수집을 위해서 R 프로그램의 rvest 패키지를 이용하였으며 수집한 메타데이터는 다음과 같다.

<표 1> 수집한 메타데이터

레이블	설명
Section id	포함된 섹션(정치, 경제, 사회, 생활/문화, 세계, IT/과학 섹션 중 하나)
Title	기사 제목
Press	언론사
Url	기사 url
Date	발행일자
Content	기사 본문
Good	'좋아요' 개수
Warm	'훈훈해요' 개수
Sad	'슬퍼요' 개수
Angry	'화나요' 개수

<표 2> 여론조사 지지율과 수집한 데이터로부터의 지지율

날짜		여론조사 지지율(%)	수집한 지지율(%)	오차율 (%)
1 월	1 주차	71.5	78.6	10.0
	2 주차	70.2	58.7	16.3
	3 주차	65.3	70.2	7.5
	4 주차	60.5	44.7	26.1
	5 주차	62.8	58.8	6.3
2 월	1 주차	62.6	54.2	13.4
	2 주차	63.0	51.6	18.1
	3 주차	64.7	63.2	2.3
	4 주차	64.4	75.4	17.0
3 월	1 주차	66.7	89.1	33.5
	2 주차	70.5	84.9	20.4
	3 주차	67.8	71.7	5.8
	4 주차	69.9	67.4	3.6
4 월	1 주차	70.8	58.1	17.9
	2 주차	68.9	57.0	17.2
	3 주차	68.4	61.1	10.6
	4 주차	69.5	83.2	19.8
5 월	1 주차	78.2	74.9	4.2
	2 주차	76.1	93.7	23.2
	3 주차	74.3	64.8	12.8
	4 주차	73.1	55.8	23.7
	5 주차	73.0	59.6	18.4
오차율 평균(%)				14.9

<표 3> α 값에 따른 값의 변화와 오차율의 변화

날짜		α=0.5 (%)	α=0.6 (%)	α=0.7 (%)	α=0.8 (%)	α=0.9 (%)
1 월	1 주차	75.0	74.3	73.6	72.9	72.2
	2 주차	66.9	68.1	69.1	70.1	70.8
	3 주차	68.5	68.9	69.5	70.1	70.8
	4 주차	56.6	59.2	62.0	65.0	68.2
	5 주차	57.7	59.1	61.1	63.8	67.2
2 월	1 주차	56.0	57.1	59.0	61.9	65.9
	2 주차	53.8	54.9	56.8	59.8	64.5
	3 주차	58.5	58.2	58.7	60.5	64.4
	4 주차	66.9	65.1	63.7	63.5	65.5
3 월	1 주차	78.0	74.7	71.3	68.6	67.8
	2 주차	81.4	78.8	75.4	71.8	69.5
	3 주차	76.6	76.0	74.3	71.8	69.8
	4 주차	72.0	72.5	72.2	70.9	69.5
4 월	1 주차	65.1	66.8	68.0	68.4	68.4
	2 주차	61.0	62.9	64.7	66.1	67.2
	3 주차	61.1	62.2	63.6	65.1	66.6
	4 주차	72.2	70.6	69.5	68.7	68.3
5 월	1 주차	73.5	72.3	71.1	70.0	69.0
	2 주차	83.6	80.9	77.9	74.7	71.4
	3 주차	74.2	74.4	74.0	72.7	70.8
	4 주차	65.0	67.0	68.5	69.3	69.3
	5 주차	62.3	64.0	65.8	67.4	68.3
오차율 평균(%)		8.7	6.9	5.2	3.9	4.1

3.2 지지율 공식

지지율(Approval Rating) 공식은 다음과 같다.

$$AR(\%) = \frac{1}{n}\sum_{k=1}^{n}\frac{(L_k + W_k)}{(L_k + W_k) + (S_k + A_k)} \times 100$$

(L 은 '좋아요', W 는 '훈훈해요', S 는 '슬퍼요', A 는 '화나요')

L 과 W 는 기사에 대한 긍정적인 반응을 나타내며 S 와 A 는 부정적인 반응을 나타낸다. 긍정적인 반응의 합을 전체 반응의 합으로 나눈 것이 지지율이라고 할 수 있으며 해당 기사에 대한 선호도를 파악할 수 있는 공식이다. 본 연구에서는 검색한 정치인에 대한 지지율을 찾는 것이므로 같은 기간에 쓰여진 전체 기사들의 지지율과 비교하여 실제 지지율을 계산하는 과정이 필요하다. 기준점은 50%로 하였고 AR 값들의 변화에 따라 실제 AR 값은 변하게 된다. 이 식은 다음과 같다.

$$\text{실제 AR}(\%) = \frac{\text{검색어가 포함된 AR}}{\text{전체 AR}} \times 50$$

3.3 데이터 분석

3.1 절에서 수집한 데이터를 토대로 3.2 절의 지지율 공식을 사용해서 데이터 분석을 시행하였다. 검색어는 문재인 대통령이었고 관련된 줄임말들(예 : 文, 문대통령 등)을 고려하여 진행하였다.

먼저 일별로 검색어에 대한 지지율을 구한 다음, 해당 일의 전체 지지율을 구한다. 이 두 값을 통해 실제 AR 값을 구할 수 있게 되고 주간 평균을 구해서 각 주차의 값으로 사용하게 된다. 이후에는 얻어낸 지지율 값과, 여론조사 대표기관 3 곳이 조사한 지지율 값들을 비교하였다. 또한 지수평활법을 적용하여 α의 값에 따라 변하는 오차율을 구하였다.

4. 분석 결과

4.1 데이터 분석 결과

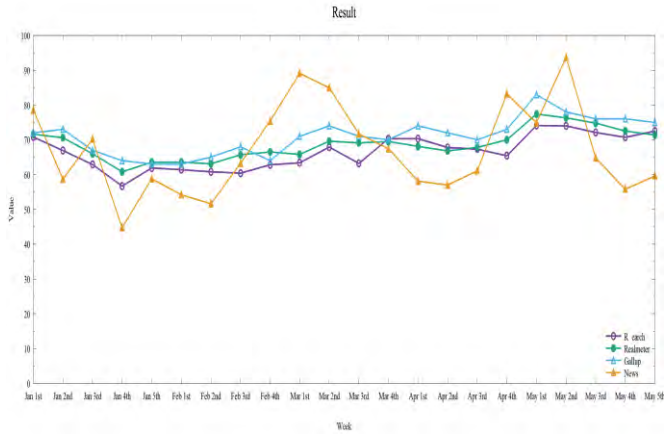
4.1.1 표를 통한 분석

데이터 분석을 통해 나온 값들을 토대로 표 2 와 표 3 을 만들어 보았다. 원래는 일별로, 주별로, 달별로 표현이 가능하지만 전체적인 흐름을 보기 쉽게 하기 위하여 주별로 결과값을 나타내보았다. 여론조사 지지율의 경우에는 여론조사 3 대 기관으로 평가받는 알앤서치, 리얼미터, 갤럽의 여론조사 결과값의 평균을 사용하였다.

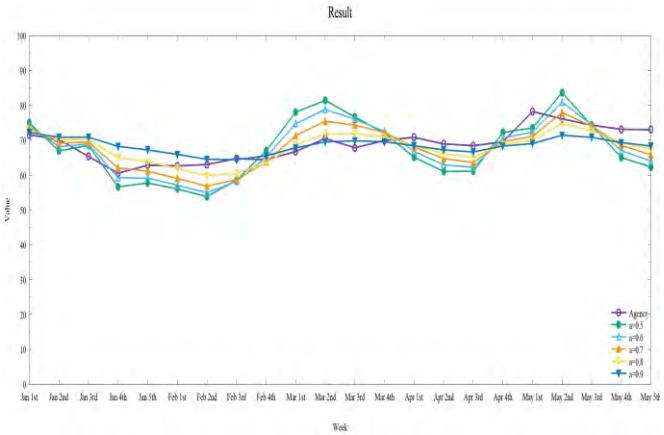
지수평활법 적용전 수집한 지지율은 3.2 절의 지지율 공식을 통해 얻어진 실제 AR 값으로써 주마다 변화폭이 큰 것을 볼 수 있다. 이는 정치적인 사건에 따라서 매주마다 여론이 바뀌게 되는데 이것을 반영한 것이라고 할 수 있다. 이 때에는 오차율이 여론조사 지지율과 비교해 크게 차이나는 것을 볼 수 있다.

이러한 오차율을 줄이고 과거의 값들도 반영하기 위해서 지수평활법을 사용하였다. 일반적으로 과거의 지지율은 현재의 지지율에 영향을 미치기 때문이다. 여기에 적용한 지수평활법은 이전의 값들이 현재의 값들에도 영향을 주는 것을 가정하여 만들어진 방법론으로서 α값을 설정해서 구하게 된다. 이전의 값들에는 α값을 곱하고 현재의 값에는 (1-α)를 곱함으로서 현재의 값에 어느정도 과거의 값들이 반영되게 하는 방법론이다. 1 월 첫째주의 초기값의 경우, 과거 데이터는 여론조사 기관들의 1 월 첫째주 평균값 데이터를 사용하였다.

지수평활법을 사용한 후에는 오차율이 눈에 띄게 감소한 것을 볼 수 있고 여론조사값들에 거의 근접한 값들을 보이고 있다. 또한 α값이 증가함에 따라 오차율이 감소하는 경향을 보이기는 하지만 꼭 비례하는 것은 아니고 α값이 0.8 일 때 최소가 되는 것을 볼 수 있다.



(그림 1) 여론조사 3사와 뉴스 데이터로부터의 지지율 그래프



(그림 2) α 값의 변화에 따른 여론조사 3사 평균값과의 비교 그래프

4.1.2 그래프를 통한 분석

4.1.1 절의 표를 통한 분석값을 토대로 그래프를 만들어 보았다. 그림 1 의 경우에는 여론조사 3 사(알앤서치, 리얼미터, 갤럽)의 여론조사 발표값과 뉴스 데이터로부터 추출해낸 지지율 그래프를 비교해보았다. 3 사의 경우에는 전반적으로 서로 비슷한 그래프를 나타내고 있는 반면에 뉴스 데이터 기반 그래프는 주마다 큰 변동폭을 보이고 있다. 이는 매주 발생하는 정치적 사건에 대해 여론이 크게 변동하는 것으로 해석해볼 수 있다. 그러나 지지율의 경우 과거의 영향이 분명하게 존재하고 과거의 정치인의 이미지가 현재의 지지율에도 영향을 미친다. 따라서 과거의 값을 어느정도 반영할 필요성이 있으며 이에 따라 α 을 적용한 지수평활법을 적용해보았다.

이러한 큰 변동폭을 가졌던 뉴스데이터를 지수평활법을 적용해 그래프로 나타낸 것이 그림 2 이다. 이 그래프에서는 3 사의 평균값 그래프와 α 값의 변화에 따른 그래프를 비교해보고 있다. α 값이 커짐에 따라 변동폭이 작아지고 과거의 지지율 반영도가 높아지는 것을 볼 수 있다. 이를 통해 α 값이 커지면 여론조사의 평균값 그래프와 오차율이 적어진다고 생각할 수 있으나 실제로는 α=0.8 인 부분에서 오차율은 가장 작게 발생하였다. 적절한 α 값을 찾음으로써 여론조사의 정확도를 향상시킬 수 있으며 오차율을 줄일 수 있다는 것을 알 수 있다.

4.2 시사점

기존의 여론 조사와 비슷한 추이를 따라가지만 변화의 크기는 훨씬 폭넓은 것으로 보아 여론의 반영도가 높다고 생각할 수 있다. 또한 기존의 방식은 상당수의 투자하는 시간과 인력이 필요하지만 이 방법의 경우에는 시간과 인력을 크게 줄일 수 있다. 그러면서도 α 값을 조정함에 따라 기존의 여론조사값과 비슷한 값을 도출해낼 수 있다는 점이 장점이다.

물론 이 방법도 약점이 존재한다. 정확도의 측면에서 다양한 조사 루트를 가지고 있는 기존 여론 조사에 비해 약점을 가지고 있다. 따라서 두 방식에서 장점만을 찾아 섞어서 사용하는 편이, 보다 정확성을 높이는 데에 도움이 될 것으로 보인다.

4.3 개선 방안

지수평활법만이 아닌 다른 방법론도 적용해 볼 필요성이 존재하며 현재는 기사 본문만 사용하였지만 댓글도 반영한다면 보다 정확할 것으로 예상된다. 또한 현재는 검색어가 포함만 되어 있으면 데이터를 추출해내었지만 향후에는 다른 분석 방법을 통해 기사의 내용을 이해하고 일정 비중 이상 검색어가 나오지 않으면 관련도가 없다고 판단하고 취사선택하는 방법론도 적용해볼 필요가 있다.

5. 결론

현재 제안한 방식으로는 부족한 점이 많이 존재하지만 기존 여론 조사 방식보다 나은 점이 있다는 것을 분명하게 알 수 있었다. 향후 이 방법론을 계속해서 수정하고 보완하여 기존의 여론조사와 함께 사용한다면 의미 있는 방법론이 될 수 있을 것이다.

참고문헌

[1] 김지윤, 강충구, 여론조사의 대표성, 고려대학교 평화와민주주의연구소, 2014
[2] 배종찬, 선거 여론조사 이해를 위한 5 가지 조건, 관훈저널, 2017
[3] 정한울, 외주민주주의 시대의 선거여론조사, 동아시아연구원, 2015
[4] 허순영, 이수철, 전화 선거여론조사에서 무응답률 증가로 인한 편의와 응답률 제고 방안, 한국데이터정보과학회지, 2016
[5] 고길곤, 김대중, 선거 여론조사 응답오차의 원인에 관한 연구, 한국행정학회, 2016
[6] 김동성, 김종우, 소셜 미디어 상에서의 여론 분석을 위한 사회 네트워크 분석 활용 방안, 대한산업공학회, 2015
[7] 이수범, 김용준, 이선정, 소셜 빅데이터를 활용한 제 19 대 대통령선거 TV 토론의 수용자 반응 연구 연관어 분석과 감정어 분석을 중심으로, 방송문화진흥회, 2018