한국컴퓨터정보학회 동계학술대회 논문집 제25권 제1호 (2017. 1)

SNS 감정 분석을 이용한 선거 후보자 순위 예측 시스템

*문유진, O이한수, 박혁, 이재영, 김선국
한국외국어대학교 경영정보학과
e-mail: yjmoon@hufs.ac.kr, sursurhansur@gmail.com, ph_dolphin@nate.com,
wodud9079@naver.com, seonkuk_kim@hotmail.com

System Implementation of Winner Forecasting for Election Cadidates Utilizing SNS Emotion Analysis

*Yoo-Jin Moon, OHansoo Lee, Hyuk Park, Jaeyoung Lee, Sunguk Kim
Dept. of Management Information Systems, Hankuk University of Foreign Studies

● 요 약 ●

대한민국 20대 총선, 영국의 유럽연합 탈퇴인 브렉시트, 트럼프와 힐러리의 대결인 미국 대선, 이 셋의 공통점은 언론의 예측과 다른 투표 결과가 나왔다는 점이다. 이러한 일련의 사건들로 인해, 각종 언론사에서 실시하고 있는 표본조사의 신뢰도에 대한 근본적 재검토의 필요성이 제기되고 있는 실정이다. 본 논문에서는 선거 후보자 지지율을 효율적이며 효과적으로 분석하기 위하여 SNS 감정분석을 제안한다. SNS 감정분석은 기존의 표본을 구하고 분석하는 방식보다 더 빠르게 표본 수집 및 분석이 가능하다. 또한 R프로그램과 구글을 이용하여 처리하기 때문에 기존 방식에 비하여 매우 저렴하다. 현재 언론사의 예측이 빗나가고 있는 시점에서 SNS 감정분석이 훌륭한 대안이 될 수 있을 것이다. 본 연구에서의 트래픽*감정분석 점수를 보았을 때, SNS 감정분석이 여론을 더 정확히 반영한다는 것을 증명한다.

키워드: SNS 데이터(SNS data), 감정 분석(emotion analysis), 데이터베이스 시스템(database system), 표본 수집(sample collection)

I. Introduction

대한민국 20대 총선, 영국의 유럽연합 탈퇴인 브렉시트, 트럼프와 힐러리의 대결인 미국 대선, 이 셋의 공통점은 언론의 예측과 다른 투표 결과가 나왔다는 점이다 ([1], [2], [3]). 이러한 일련의 사건들로 인해, 각종 언론사에서 실시하고 있는 표본조사의 신뢰도에 대한 근본적 재검토의 필요성이 제기되고 있는 실정이다. 이러한 문제점을 해결하기 위한 방법으로 SNS의 글들을 분석하는 정성적 방법이 존재한다. 여기서는 차기대선 후보자 A, B, C, D, E를 SNS 감정분석을 통하여 분석해보고자 한다. SNS에는 매우 방대한 글들이 올라오므로 그것들을 다 분석하기란 어려운 일이다. 그래서 여기서는 정치인 A, B, C. D, E를 미시적, 거시적으로 분류해서 분석해보고자 한다. 거시적으로는 후보자들의 구글링을 통한 트래픽 조사를, 미시적으로는 R프로그램을 통하여 SNS(Twitter) 감정분석을 실시할 것이다 ([4], [5]).

II. Preliminaries

이 연구에서 Facebook은 사용하지 않고 Twitter만을 이용할 것이다. 그 이유는 Facebook은 Twitter에 비하여 자신의 정보가 더 많이 드러나기에 부정적 감정보다 긍정적 감정의 비율이 높기 때문이다.

현재 국내 정치는 최순실 게이트 및 대통령 탄핵사태로 인하여 매주 토요일마다 집회가 열리고 있으며 대통령 지지율은 4%로 역대 최저치를 찍고 있다 ([6]). 이러한 상황 속에서 앞으로 누가 지도자가 될지 혹은 어느 정당이 여당이 될지가 대한민국의 안정 및 미래 발전을 위하여 그 어느 때보다 중요해지고 있다.

조사 당시의 각 후보들의 지지율은 2016년 11월 11일 당시 Fig. 1에서 보여주듯이 문재인, 반기문, 이재명, 안철수, 박원순 순으로 각각 23.3%, 16.7%, 9.6%, 8.7%, 5.9%이었다. 그리고 각 차기대선 후보자 문재인, 반기문, 박원순, 이재명, 안철수를 A, B, C, D, E로 명하겠다.

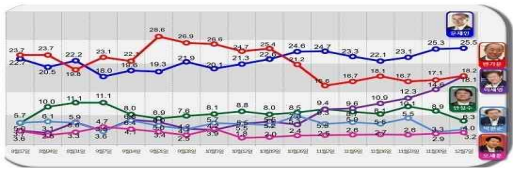
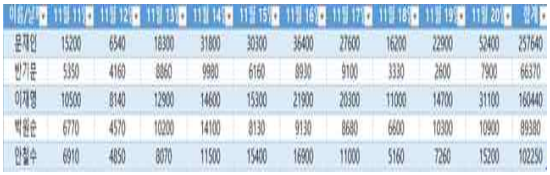


Fig. 1. 대선 후보 지지율 (2016년 11월 11일)



이름/날	11월 11일	11월 12일	11월 13일	11월 14일	11월 15일	11월 16일	11월 17일	11월 18일	11월 19일	11월 20일	합계
문재인	15200	6540	18300	31800	30300	36400	27600	16200	22900	52400	257640
반기문	5350	4160	8860	9980	6160	8930	9100	3330	2600	7900	66370
이재명	10500	8140	12900	14600	15300	21900	20300	11000	14700	31100	160440
박원순	6770	4570	10200	14100	8130	9130	8680	6600	10300	10900	89380
안철수	6910	4850	8070	11500	15400	16900	11000	5160	7260	15200	102250

Fig. 3. 후보별 트래픽 양 (11월 11~20일)

III. The Proposed Scheme

구글을 이용하여 11월 11일부터 11월 20일까지 후보자들의 트래픽을 조사하였다. 각 후보들의 트래픽은 다음과 같다.

트래픽 기준으로 순위는 A > D > E > C > B 순 이였다. 이후 트위터에서 4명의 후보(A, B, C, D)의 이름이 언급된 글을 각 후보당 매일 10개씩 찾아 열흘 동안 100개, 총 400개의 데이터를 수집하였다. 그 후 수집한 데이터를 강한 긍정, 긍정, 중립, 부정, 강한 부정의 감정을 분류하였다. 또한 긍정이라면 왜 긍정인지, 부정이라면 왜 부정인지를 알기 위하여 감정을 더 세부적으로 나누어 보았다.

긍정은 행복, 안심, 긍지, 재미, 만족 5가지로, 부정은 분노, 혐오, 슬픔, 공포, 불만족 5가지로 분류하였다. Fig. 2은 E-R Diagram의 Entity 구성을 보여준다.

MySql에 데이터를 넣고 긍정, 부정 비율을 살펴 보았다. 긍,부정 비율은 (강한 긍정 + 긍정/ 전체[강한 긍정 + 긍정 +중립 + 부정 + 강한 부정])로 구하였다. SQL 문을 돌려본 결과 C, D 후보의 긍,부정율이 92, 51 %로 절반이 넘었다.

그 다음 강한 긍정 2점, 긍정 1점, 중립 0점, 부정 -1점, 강한 부정 -2점으로 5점 척도를 구성하였다. 이 후 Fig. 3에서 보여준 각 후보의 total reactivity반응을 합산하여 앞서 구한 긍,부정 비율과 곱해주었다. 그 결과 D후보의 경우에만 0.61로 양수를 보였고 나머지는 음수를 보였다. 앞서 긍,부정율이 절반을 넘었던 C 후보의 경우 강한 부정 쪽의 의견이 많았기에 오히려 음수를 보였다. Fig. 3의 구글에서 찾은 트래픽과 후보들에 대한 반응 총합을 곱해 보았다.

Fig. 4에서, 이 시스템 활용 결과 D > C > B > A순으로 나왔다. 즉 이 연구 시작점인 2016년 11월 11일의 대선 후보 지지율과는 판이하게 다른 결과였다. 그러나 2016년 12월 14일 이후의 후보 지지율은 이 시스템의 결과와 동일하게 여론이 움직이고 있음을 각 언론에서 공표하였다 ([6]).

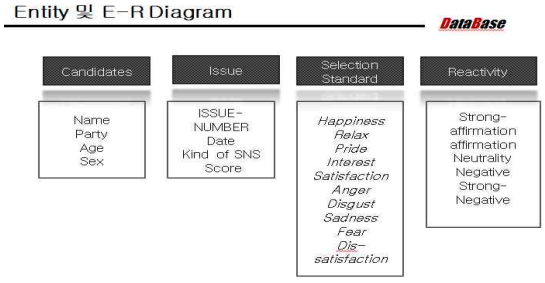
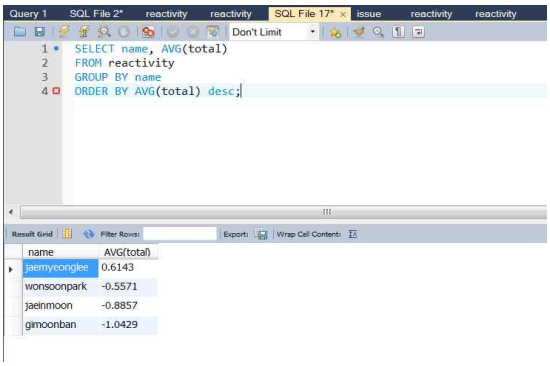


Fig. 2. Entity 구성



```
SELECT name, AVG(total)
FROM reactivity
GROUP BY name
ORDER BY AVG(total) desc;
```

name	AVG(total)
jaemyeonglee	0.6143
wonsoonpark	-0.5571
jaeinmoon	-0.8857
gimoonban	-1.0429

Fig. 4. 각 후보당 감정 반응 평균(점수)

IV. Conclusions

표본조사의 신뢰성이 의심되어 SNS 감정분석과 트래픽을 이용하여 조사하였다. 조사 결과는 언론에서 보여준 후보 지지율에 비하여 D후보가 압도적으로 높은 것을 확인할 수 있었다. D 후보를 제외한 나머지 A, B, C 후보들이 모두 음의 점수가 나와 우열을 가리기 어렵지만 D > C > B > A 순이었다. 한 달이 지난 지금 차기 대선 후보 지지율을 보면 D후보가 높은 것을 볼 수 있다.

SNS 감성분석과 트래픽을 곱했을 때의 결과에 맞게 D후보의 지지율이 올라갔으며, 본 연구가 여론을 거의 정확하게 예측하였다.

비교적 간단하게, 차기 대선 후보 지지도를 알 수 있었으며 심층적으로 어느 감정에서 기인한 긍정, 부정을 알 수 있었다. 적은 표본의 문제를 해결하기 위해서 R 프로그램에서 twitterR패키지를 이용하여 트위터의 글을 추출해서 분석한다면 자료의 신뢰성을 높일 수 있을 것이다.

References

[1] http://myhostis.tistory.com/89

[2] https://www.viewsnnews.com/article?q=139593

[3] http://www.fnnews.com/news/201611251038317486

[4] David Kroenke and David Auer, “Database Concepts”, Pearson, 2015.

[5] Prabhanjan N. Tattar, Suresh Ramaiah, B. G. Manjunath, “A Course in Statistics with R”, Willey, 2016.

[6] http://naver.com