The Korean Journal of Applied Statistics (2018) DOI: https://doi.org/10.5351/KJAS.2018.31.5.655
31(5), 655–665

Prediction improvement of election polls by unstructured data analysis

Sunbin Park[a] · Myung Joon Kim[a,1]

[a]Department of Business Statistics, Hannam University

(Received August 8, 2018; Revised August 27, 2018; Accepted August 27, 2018)

Abstract

Social network services (SNS) have become the most common tool for the communication of public and private opinions as well as public issues; consequently, one may form or drive public opinions to advocate by spreading positive content using SNS. Controversy for survey data based opinion poll accuracy continues in relation to response rate or sampling methodology. This study suggests complementary measures that additionally consider the sentiment analysis results of unstructured data on a social network by data crawling and sentiment dictionary adjustment process. The suggested method shows the improvement of prediction accuracy by decreasing error rates.

Keywords: sentiment analysis, social network, data crawling, sentiment dictionary

1. 서론

소셜 미디어는 최근 전 세계적인 커뮤니케이션 도구로 자리 잡은 페이스북이나 트위터 등을 칭하며 사용자 개개인이 정보를 생산하고 전파하며 소비하는 특성을 가지고 있다. 또한 정보의 생산자와 소비자간 양방향 소통을 가능하게 함으로써 일방적으로 정보를 전달하는 기존의 매스컴과는 다른 형태로 그 영향력 및 사용도가 점차 증가하고 있다. 페이스북과 트위터는 전 세계에서 15억 명의 사용자를 보유한 소셜 네트워크 서비스(social network service; SNS)로 모바일 기반의 높은 접근성, 다양한 어플리케이션 응용이 가능한 플랫폼의 개방성과 피처폰에서 스마트폰으로 변화하는 패러다임 전환과 맞물려 폭발적인 사용자의 증가를 이루었다. 그 중 트위터는 사용자 간의 네트워크 형성으로 인한 높은 실시간 전파성으로 소셜 분석에 가장 적합한 환경과 특성을 가지고 있다. 최근 특정 조사기관에서는 소셜 미디어 상에 존재하는 텍스트 데이터를 분석하여 투표 결과를 예측하는 시스템들을 구축하려는 노력들을 보이기도 하고 있는데, 이는 소셜 미디어가 가지는 소통의 힘이 단순 의견 교류에서만 멈추지 않고 정서적인 감정까지 전염시킬 정도로 크다는 것이며, Kramer 등 (2014)은 이를 검증하는 연구결과를 제시하기도 하였다.

이러한 특성을 고려하여 Kim과 Hwang (2014)은 트위터의 타임라인에 열거된 단어들의 감성분석을 통해 정치적 성향을 파악하기도 하였으며, Bae 등 (2013)은 2012년 대선 결과와 관련한 트위터 내용을 분석하기 위한 텍스트 마이닝 기법을 제안하기도 하였다. 국내 뿐만 아니라 해외에서도 선거 결과를 예측

[1]Corresponding author: Department of Business Statistics, Hannam University, 70 Hannamro, Daedeok-Gu, Daejeon 34430, Korea. E-mail: mkim@hnu.kr

하는데 있어 이러한 소셜 미디어에 대한 분석결과를 활용하는 것이 필요하다는 연구 결과들이 제안되었으며, Wegrzyn-Wolska와 Bougueroua (2012)가 프랑스 대선 결과에 접목하여 제시한 연구, 미국의 의회 선거결과에 소셜 미디어의 내용을 고려하는 것이 필요한지 여부에 대하여 Williams와 Gulati (2008)가 제시한 연구 결과들이 대표적인 사례라 할 수 있다. 이러한 연구들의 특징은 특정 정당 및 후보관련 단체들에 대한 시민들의 민심에 대한 실시간 모니터링과 비정형 데이터 분석을 통한 정치 흐름과 득표율을 예측하는 데 활용하는 것에 초점을 두고 있다. 이는 소셜 미디어 상에 존재하는 데이터는 실시간으로 분석이 가능하며, 기존에 활용하던 전화, 현장조사에 비해 빠르고 효율적이며 경제적인 이점 또한 존재한다는 것에 기인한다.

국내의 경우 여론조사를 통한 선거 결과 예측에 대한 연구는 활발히 진행되어 왔으며, 이는 주로 표본의 대표성 문제, 표본 수집 방식의 개선 등에 초점을 두고 있다. Lee 등 (2006)은 표본조사에 대한 결측치에 대한 대체 방안을, Kim과 Huh (2009)와 Kim과 Jung (2017)은 전화조사의 편향성과 무선전화 조사 비율에 따른 결과의 변화에 대한 연구 결과들을 통하여 문제점을 보완하는 방법을 제안하였으며, 또한 Kim과 Kwon (2009)는 방문 조사 시 응답거부와 부재율이 미치는 영향에 관한 연구들을 제안하기도 하였다. 이러한 연구들은 기존의 표본 조사가 가지는 문제점과 한계점의 보완을 통하여 예측력을 개선하고자 하는 연구들이라 할 수 있다. 하지만 소셜 미디어상에 실시간으로 존재하는 다양한 의견을 반영하여 예측의 정확도를 향상하고자 하는 연구는 제한적인 것이 사실이다.

따라서 본 논문의 목적은 이러한 소셜 미디어가 가지는 소통력의 이점과 개개인의 의견을 마음대로 피력할 수 있다는 장점을 이용하여 기존 진행하던 선거여론조사 데이터와 소셜 네트워크상에 있는 비정형 데이터에 오피니언 마이닝을 적용한 것을 추가하여 선거 여론조사 예측 정확도를 향상시키는데 있다. 기존의 여론조사가 대부분 온라인 설문, ARS, 전화조사로 수집된 데이터들에만 의존해 왔다면, 본 연구에서는 기존 조사방법들과 더불어 소셜 미디어 상의 비정형 데이터 분석 결과를 결합하여 기존 방식이 가지는 예측의 한계점을 극복하고 보다 정교한 방식을 제안하는 것에 연구에 의의를 둘 수 있으며, 19대 대선 여론 조사 결과를 대상으로 연구를 진행하였다.

본 논문의 구성은 2장은 비정형 데이터 분석과 관련하여 감성 사전의 추가적인 구축 과정에 대한 결과를 제시하였으며, 3장에서는 실제 온라인 공간에서 데이터를 크롤링하고 전처리(pre-processing)하는 과정을 포함한 비정형 데이터를 활용하여 예측력을 개선하는 방안에 대한 연구결과를 제시하였다. 이어 4장에서는 주요 결론 및 본 연구가 가지는 문제점과 한계점을 진단하고, 이를 통하여 향후 추가적으로 진행 가능한 연구 주제들을 제안하였다.

2. 비정형 데이터 분석

데이터는 크게 정형데이터, 비정형데이터로 나누어진다. 정형데이터란 시스템의 테이블과 같이 고정된 열과 행에 저장되는 데이터로 각각의 변수명마다 데이터가 지정된 것을 뜻하며 시험 문제의 답, 설문조사 결과 등 특정한 형태에 맞춰진 데이터로써 일반적으로 진행되는 대부분의 조사와 실험들이 여기에 포함된다.

비정형 데이터는 정형데이터를 제외한 모든 데이터로 정의할 수 있으며, 특정 형식과 형태로 정리되지 않은 모든 데이터로 이해할 수 있다. 오늘날 하루 평균 2억개 이상이 발생하는 트윗으로 대표되는 텍스트 형식의 데이터와 매일 온라인 상에 업로드되는 2억 5천만장 이상의 사진과 동영상 등으로 대표되는 이미지 및 영상 데이터 등이 이에 포함된다고 할 수 있다.

이러한 비정형 데이터를 통하여 일상 생활에서 활발하게 적용되고 보급되어진 비정형 데이터 분석의 사례로는 GPS 정보 분석으로 통하여 이동 경로를 안내해 주는 실시간 데이터 분석 등이 대표적이라 할 수

있겠다. 이 외에도 Lee와 Lee (2015)는 로그 정보로 남겨지는 실시간 비정형 데이터들을 분석하여 일어날 상황을 예측하는 방안에 대한 연구 결과를 제시하기도 하였으며, Choi 등 (2011a)은 시스템 보안 및 웹 보안을 더욱 강화하기 위한 대용량 보안 로그 분석 방식을 제안하였다. 또한 음원 자료에 대한 분석도 Choi 등 (2011b)에 의하며 진행되었으며, GPS 경로 추적을 이용하여 주행 중인 차량의 카메라로부터 영상을 실시간으로 수집, 포장된 도로의 손상을 파악하여 도로 관리자에게 전달해 도로 상태를 관리할 수 있게 하도록 Park 등 (2018)에 의해 제안된 연구는 비정형 데이터 중 동영상 자료를 활용한 사례이다.

이처럼 광범위하게 퍼져있으며, 다양한 형식으로 존재하는 비정형 데이터를 바탕으로 어떠한 분석 방식으로 접근하여 어떻게 활용할 것인지에 대한 연구가 다양한 분야에서 다양한 각도로 진행되어지고 있다.

2.1. 소셜 미디어를 이용한 데이터 분석

소셜 미디어는 온라인에서 구축된 사회적 네트워크 서비스를 전반적으로 칭하는 말이며 Choi (2009)는 사람들이 자신의 의견이나 생각 또는 직접 제작한 콘텐츠 등을 다른 사람들과 공유하기 위해 사용하는 온라인 틀과 플랫폼으로 정의하였다. 이러한 소셜 미디어는 사용자들의 콘텐츠를 매개로 하여 의견이나 관점, 경험들을 공유하는 활동들로 사용자간 심리적 거리감을 축소시켜 유대관계를 형성시키게 된다.

이러한 유대관계의 형성은 공유 및 상호작용 과정, 유사성과 자기개방 등으로 더욱 촉진하는 것으로 Park 등 (2014)의 연구에 의하여 나타나기도 하였다. 개인적인 의견을 피력하면서 나타난 감정의 공유에 대하여 실질적인 접촉이 없어도 네트워크를 통해 감정이 전염된다는 것이며 이는 국내에서 뿐만 아니라 Kramer 등 (2014)의 연구 결과에 의하면 해외에서도 나타나는 동일한 현상으로 이해할 수 있다.

이러한 소셜 미디어중 하나인 트위터는 다른 소셜 미디어보다 높은 전파성을 띠는 것으로 Kim 등 (2014)에 의하여 연구 결과가 제시되었으며, 따라서 정치, 사회, 연예, 사회 분위기 등 전반적인 흐름을 파악하기에 용의한 미디어로 판단 가능하다. 트위터는 타임라인에 게시된 글들을 이용해 트위터 사용자의 정치적 성향을 파악 할 수 있으며 이러한 소셜 미디어를 통하여 선거 결과를 예측하고자 하는 연구는 전 세계적으로 활발하게 진행되고 있다.

Bae 등 (2013)은 2012년 한국 대선과 관련하여 트위터 분석 결과를 제시하였고, Hyun (2010)은 2010년 지방선거와 관련한 연구 결과를 제시한 사례들이 대표적인 활용 연구라고 할 수 있다. 이러한 연구들에 의하면 특정 선거 후보자를 지지하는 페이스북 팬의 수는 실제로 그 후보에게 투표하는 투표자 수와 연관이 있으며 유권자들이 선호하는 후보를 예측하는데 유용하게 활용 가능하다는 것이다.

해외에서도 이러한 이점을 활용하여 트위터 REST API를 이용하여 데이터를 수집하고 semantic orientation from point-wise mutual information (SO-PMI) 기법을 활용해 선거 경향을 분석하는 연구 결과들이 제시된 사례들이 있다. 선거 뿐만 아니라 사회적인 현상에서도 소셜 미디어를 이용한 연구 결과들이 활발하게 진행 중이며, Jho와 Kim (2012)가 반값 등록금 이슈의 확산과 소멸에 대한 분석도 이에 해당한다 할 수 있겠다. 또한 정치, 사회 분야 이외에도 트위터 댓글 등을 이용하여 개인이 느끼는 감정에 따라서 음악을 추천하는 서비스를 상용화하는 것들도 이러한 소셜 미디어의 데이터 분석에 기반한 것이라 하겠다.

2.2. 오피니언 마이닝을 이용한 감성 분석

오피니언 마이닝은 글에 나타나 있는 작성자의 감정을 분석하는 기법으로 감성 분석(sentiment analysis)으로 불리며, 이 감성분석은 사람이 텍스트 내용의 전부를 일일이 확인하기 힘든 빅데이터 분야에서 주로 이용되고 있다. 감성 분석은 자료의 전처리 과정(data pre-processing)과 감성 사전(sentiment

dictionary)을 이용해 해당 글이 갖는 감성을 자동으로 파악해 주는 분석 과정으로 이해할 수 있다.

즉, 사용자가 작성한 글의 성향을 분석해 문장에 대한 의견이 긍정인지 부정인지를 파악하는 분석 과정으로 일반적으로 상품평, 영화평과 같이 평점이 존재하거나 좋아요 와 같은 감성을 나타낼 수 있는 주관적인 정보가 들어있는 문장들에 적용된다. 감성분석은 데이터를 크롤링하고, 이를 통하여 수집된 자료의 전처리 과정을 거쳐 분석 가능한 형태소별로 분해하고 분해된 형태소들에 각각 점수를 부여한 후 문장별로 산출된 점수들을 통하여 해당 문장의 특성을 파악하는 일련의 과정을 포함한다.

이러한 분석 방식은 감성을 느낄 수 있는 모든 분야에 적용 가능하며 특히 고객의 만족도 수준의 파악이나 신상품 출시에 따른 고객의 반응에 대하여 실시간으로 빠르게 파악하고자 하는 경우 이러한 분석이 유용하게 활용될 수 있다. Kim과 Jeong (2013)은 고객과의 상담 및 민원 업무 분야에서 발생하는 음성을 감정 상태와 사용 어휘에 따라 분류하고자 이러한 분석 방식을 적용하였으며, Park 등 (2004)은 사진의 이미지와 색상에 따라 느껴지는 감성을 구조화 할 수 있는 모델을 연구 결과로 제시하기도 하였다.

또한 Chang (2009)는 소비자들의 상품구매 욕구를 끌어내고 소비자가 원하는 정보를 효율적으로 정리하여 제공하기 위하여 온라인 쇼핑몰에 게시되어 있는 상품평을 자동으로 분류하여 긍정, 부정의 의견을 요약된 결과로 제공하는 모델을 만들어 활용하는 방안에 대하여 제안하였다.

앞서 살펴본 바와 같이 감성 분석은 텍스트 수준을 넘어 음성, 이미지 등 다양한 형태의 데이터에 적용하여 현재의 상태에 대하여 실시간으로 피드백해 줌으로써 이를 판단하고 활용하고자하는 사용 주체가 합리적인 의사 결정을 빠르게 진행할 수 있는 유용한 분석 도구로 그 범위를 점차 확대해 나가는 분야라고 할 수 있다.

본 논문도 기존에 진행되고 있는 여론조사를 보다 정교하게 예측하기 위하여 기존에 제시되어 왔던 여론조사의 기간 조정, 표본 방식의 개선, 조사 방법론의 개선이 아닌 새로운 데이터 수집 및 분석 방식 결과를 적용함으로써 여론조사의 정확도 향상에 연구의 목적을 두고 있다. 따라서 소셜 미디어들 중 트위터와 페이스북에서 데이터를 크롤링하고, 수집된 데이터에 대한 감성분석을 통하여 긍정문의 변화를 파악하고 기존 여론조사 결과에 감성분석 결과를 보완하는 방안을 본 논문에서 제시하고자 한다.

2.3. 감성 사전 추가 구축 및 결과

인간의 감성은 크게 기쁨, 화남, 슬픔, 즐거움과 같은 총 4가지로 나누어진다. 본 연구에서는 기쁨, 즐거움을 긍정, 화남, 슬픔을 부정으로 하여 감정사전 DB를 구축하였으며 기존 사용되던 긍정어, 부정어 사전에 크롤링된 데이터의 일부를 분석하여 정치적인 성향에 있어 긍정과 부정으로 많이 활용될 수 있는 단어들을 감성사전에 추가하여, 분석의 정확도를 향상시킬 수 있는 방안을 추가적으로 사전에 진행하였다. 감성 사전의 적절성은 Table 2.1에 정의된 분류 기준과 이를 다음과 같이 정의된 지표에 적용하여 판단할 수 있다.

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}},$$
$$\text{Recall} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}},$$
$$\text{F-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

위의 수식에 Precision은 분석결과 긍정으로 분류된 단어의 실제값이 긍정인 경우로 판단한 비율을 나타낸 것으로 얼마나 정확한 분류를 했는지 나타내 주는 지표라고 할 수 있다. Recall은 실제 긍정(true)인 단어들 가운데 분석 결과값을 긍정으로 해석해주는 비율로써 전체 긍정 데이터 중 분석에 활용된 데이터

Table 2.1. Positive and negative classification

Class	Real value	
	True	False
Positive	True positive	False positive
Negative	False negative	True negative

Table 2.2. Comparison of sentiment dictionary performance

Class	Precision		Recall		F-score	
	Current	New	Current	New	Current	New
A candidate	79.26	80.89	86.66	96.00	82.80	87.80
B candidate	65.21	65.93	98.36	98.36	78.43	78.94
C candidate	59.52	58.62	94.33	96.22	72.99	72.85
Average	67.99	68.48	93.11	96.86	78.07	79.86

의 비율을 말해준다. F-Score는 Precision과 Recall의 조화평균으로 Precision으로 평가되는 정확도와 Recall로 평가되는 재현율을 통해 사전의 정확성을 측정하는 지표이다.

연구에 새롭게 적용하고자 하는 사전의 성능을 분석하기 위해 분석 대상이 되는 대선 후보자들을 대상으로 하여 100개의 트윗을 무작위 추출하고 실제 사람이 느끼는 감성에 대한 사실을 문맥 파악을 통하여 실제적으로 파악하고, 이에 대한 일치성 여부를 감성분석 결과와 비교하였다. Table 2.2는 기존의 감성 사전에 정치적인 긍정과 부정을 판단할 수 있는 일부 단어들을 사전에 추가하여 진행한 결과로 수정된 사전에서 긍, 부정 단어 추출률이 다소 높아진 것을 확인할 수 있으며, Recall의 경우 상승폭이 상당 부분 증가함을 확인할 수 있었다. 또한 F-Score의 경우 전반적으로 높아진 것을 확인 할 수 있었지만 특정 후보의 경우 다소 감소하는 현상이 나타났으나, 이는 다른 후보들에서 부정적인 단어로 취급되는 비속어들을 해당 후보가 직접 언급하는 경우가 많아 수식어처럼 붙어 다니는 경우에 기인한 것으로 나타났다.

새롭게 구축된 감성 사전의 전체적인 정확도가 기존의 감성 사전을 사용하는 것보다 개선되는 효과가 있는 것으로 판명되는 바, 새롭게 구축된 사전의 활용이 예측력 개선에 유의미한 도구가 될 수 있을 것으로 판단하였다.

3. 데이터 분석 과정 및 분석 결과

기존에 실행되었던 여론조사에서는 온라인 공간에서 생성되는 데이터가 활용되지 않았다는 점, 전화 ARS 조사 또한 젊은층보다 장년층이 더 많이 분포해 표본의 층화추출이 원활하게 이루어지지 않는다는 점과 이를 보정하고자 하는 가중치 반영 방식 등의 차이들로 인하여 여론조사의 정확도는 조사업체별로 작게는 1%에서 크게는 약 10%까지 차이가 나는 것으로 확인되었다. 이러한 여론조사 현황을 개선하기 위하여 본 논문에서는 기존 여론조사 결과와 SNS에서 추출한 대선 후보와 관련된 자료들을 수집하고 감성 분석 결과를 기존 여론조사에 보완하는 여론조사 개선방안을 도출하였다. 기존의 방식을 보완하는 과정을 요약하여 도식화 한 것이 Figure 3.1이다.

이번 장에서는 분석 대상이 되는 데이터의 선정과정에서부터 데이터의 수집 및 전처리 과정을 포함한 사전 분석 과정과 본 논문에서 제안하고자 하는 비정형 데이터 분석 결과의 활용 방안에 대하여 도식화된 Figure 3.1의 과정을 자세하게 순차적으로 설명하며, 기존의 여론 조사 결과 반영을 위한 자료는 19대 대선 여론조사를 실시한 사이트 중 매출액이 높은 업체 3곳인 메트릭스, 한국갤럽, 한국리서치의 예상 득표율의 평균을 활용하였다.

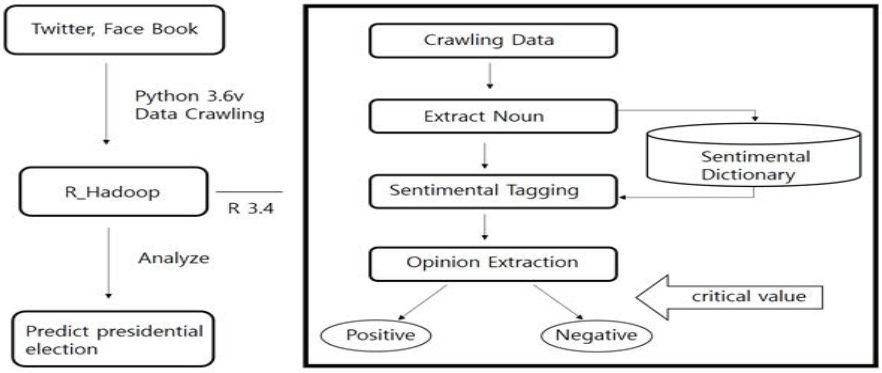

Figure 3.1. Research process flow.

3.1. 분석 데이터의 선정 및 수집 과정

대중들의 인식을 조사하는 방법으로는 뉴스와 트위터, 페이스북, 인스타그램 등 여러 종류의 SNS가 사용될 수 있다. 이러한 소셜 네트워크 서비스 중 본 논문에서는 기존 여론조사에서 젊은층이 비교적 부족한 것을 고려해 젊은층이 많이 사용하는 SNS를 선택하였으며 그 중 비교적 데이터의 조작이 힘들고 많은 사람의 의견을 얻을 수 있으며, 정치적 성향의 글이 비교적 많이 올라오는 트위터와 페이스북으로 최종 선택하였다. 따라서 트위터와 페이스북에서 데이터를 수집한 후 일정 수준의 데이터 처리과정을 거쳐 감성분석 결과를 반영하여 최종 예상 지표를 산출하게 된다.

데이터 수집을 위하여 Python 3.6을 사용하였고 데이터 수집 기간은 선거 당시의 여론을 파악하기 위하여 선거 유세 기간인 2017년 4월 17일부터 공표 금지기간 전날인 2017년 5월 7일까지 진행하였다. 트위터의 경우 수집 키워드는 19대 대선과 관련된 키워드인 '대통령', '후보' 2개의 단어와 당선 가능성이 높은 후보 3명의 이름을 포함하여 총 5개의 키워드이며 페이스북은 해당 기간 동안 각 후보의 공인 페이지에 게시된 게시글의 댓글을 종합하였다.

데이터의 수집을 진행한 이후, 전처리 과정에서 R 프로그램에서 하둡 시스템을 사용하는 RHADOOP의 rhdfs와 rmr2 패키지를 사용하여 비정형 데이터에 대한 분석을 진행하였다. 이는 분석 과정에 포함되는 50만여개에 이르는 비정형 데이터를 분석하는데 효율적으로 활용 가능한 방식이며, 이에 대한 효율성 확인을 위한 하둡의 성능을 파악하기 위해 1부터 900만까지 제곱을 하는 기초적인 연산을 진행한 결과 단일 컴퓨터에서 진행한 시간인 33초보다 1/3이 감소한 21초가 소요되었다. 본 연구의 분석과정은 단일 머신에서도 분석 진행이 가능하기는 하나, 대용량 자료에 대한 분산 처리과정이 가지는 효용성에 대한 참고를 위하여 사례 결과를 제시하는 바이다.

3.2. 데이터의 전처리 과정

트위터에 게시된 트윗들은 사용자 자신의 의견을 포함한 트윗을 올리는 경우가 대부분이지만 의미없는 트윗이나 조작 가능성이 있는 트윗, 리트윗(retweet)을 통하여 게시하기도 한다. 따라서 본 연구에서는 의미 없는 트윗과 중복되는 트윗, 원문을 그대로 사용한 리트윗을 분석에서 제외하는 전처리 과정을 포함하였다.

분석에서 제외한 대상 중 조작 가능성이 있는 트윗은 하루 동안 같은 글이 여러 개가 등록될 시 이러한 글들을 의도적으로 다수의 계정이 한꺼번에 게시물을 올린 경우로 간주하여 해당 데이터는 연구에서 제

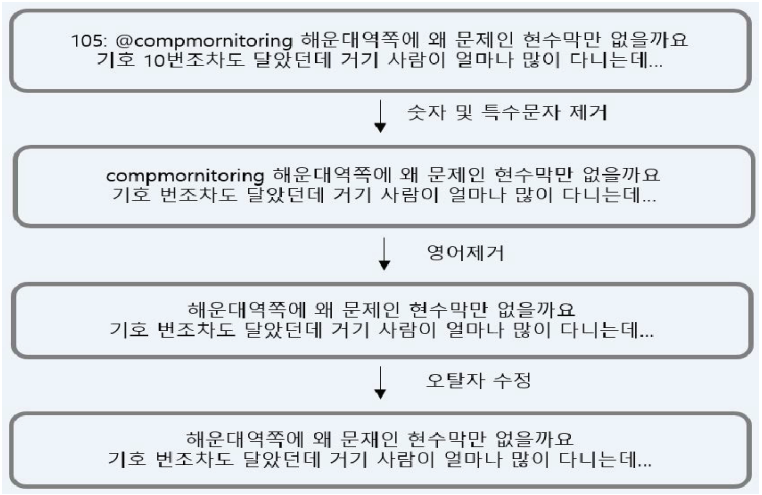

Figure 3.2. Pre-processing data step example.

외하였다. 또한 리트윗의 경우 사용자가 직접 작성한 것이 아닌 팔로우한 계정의 트윗을 그대로 공유하여 타인의 글을 인용한 것이기 때문에 전처리과정에서 제외하였다.

수집된 데이터들을 통한 감성 분석을 진행하기 위하여 (1) 해석 불가능한 불용어 및 분석에 불필요한 숫자 제거, (2) 혐오 발언과 관련된 비속어 통폐합, (3) 영어 단어와 오탈자 수정의 단계별 전처리 과정을 수행하였으며, 이에 대한 구체적인 예시는 Figure 3.2와 같다. 이러한 전처리 과정을 수행한 후 문장들에서 명사 단위로 데이터를 추출하기 위해 KoNLP 패키지를 이용하여 각 형태소로 분류하였다.

3.3. 분석 방법 및 결과

감성분석은 사용자가 작성한 글의 성향을 분석해 문장에 대한 의견이 긍정인지 부정인지를 나타내는 분석으로 일반적으로 상품평, 영화평과 같이 평점이 존재하거나 '좋아요'와 같은 감성을 나타낼 수 있는 주관적인 정보가 들어있는 문장들에 적용된다. 감성분석은 전처리된 문장들을 사용하여 각각 분석 가능한 형태소별로 분해하고 분해된 형태소들에 해당하는 점수를 부여한 후 문장별로 산출된 점수들을 합하여 문장별 특성을 파악하였다.

후보의 득표율을 예측하려면 부정적인 글보다 득표율에 영향이 있는 긍정적인 글을 파악하는 것이 중요하기 때문에 전체 긍정문서중 후보별 긍정 문서의 비율을 계산하여 후보별 긍정 문장의 비율을 파악하였으며, 이에 해당하는 결과가 Figure 3.3과 같다. 또한 후보별 투표은 교집합이 존재하지 않으므로 이를 기존 예측 투표율의 보완지표로 활용하기 위하여 후보별 긍정문서 비율 결과를 토대로 전체 긍정 문서중 후보별 긍정문서 비율로 재산출하였다.

또한 연령대별로 SNS를 사용하는 비율이 상이한 바, 이를 예측 투표율에 활용하기 위한 연령대별 가중치를 산정하였으며, 가중치는 연령별 SNS 이용자 비율과 연령별 투표율을 고려한 가중평균을 적용하였다. 연령대별 SNS 이용자 비율은 정보통신 정책연구원에서 발간한 SNS 이용추이 및 이용형태 분석 보고서 3페이지 제시된 2017년 이용률 수치를 인용하였으며, 산출된 결과는 Table 3.1과 같다.

본 연구에서는 지금까지 진행한 모든 결과들을 반영하여 다음 수식과 같은 수정 예측 투표율을 제안한다. 이는 기존 전화 조사 등으로 이루어지는 예측 득표율에 비정형 데이터 분석 결과를 반영하여 새롭게

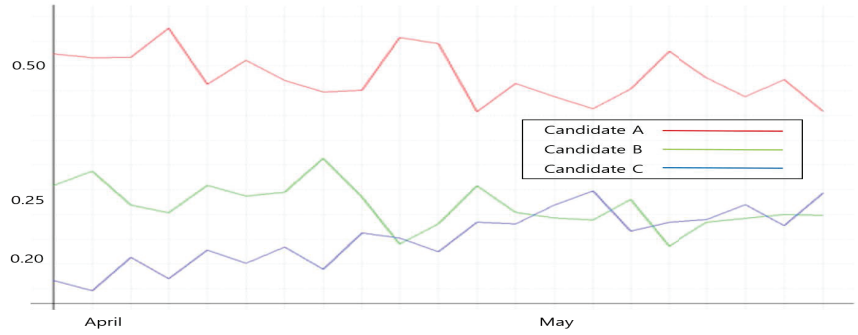

Figure 3.3. SNS positive wording rate for the each candidate.

Table 3.1. Weighted average calculation result

Age	SNS usage rate	Turnout rate	Weighted average
20–29	69.3	76.1	0.527
30–39	46.9	74.2	0.347
40–49	28.8	74.9	0.215
50–59	12.1	78.6	0.095
60 above	1.7	74.0	0.012

Table 3.2. Adjusted prediction result

Class	Real	CR	Current error	Positive rate	$\hat{P}$	New error
A candidate	41.1	39.2	1.9	50.6	41.9	0.8
B candidate	21.4	22.6	1.2	24.0	22.9	1.5
C candidate	24.0	14.5	9.5	25.4	17.0	2.6

계산되는 예측 득표율로 두 결과의 가중 평균 형태로 이해할 수 있다.

$$\hat{P}_i = \sum_j w_j \times \mathrm{UR}_{ij} + \sum_j (1 - w_j) \times \mathrm{CR}_{ij}.$$

수식에서 $\hat{P}_i$는 최종적으로 제안하는 후보별 예측 득표율을 나타내며, UR은 각 후보별로 비정형데이터 분석으로 예측된 긍정 비율을, CR은 기존 여론 조사에서 예측하는 후보별 득표율을 나타낸다. w_j Table 3.1에서 산출된 연령대별 가중치를 의미한다.

본 연구에서 제안하고자 하는 득표율 예측 수정 개선 방안을 적용한 결과 Table 3.2와 같은 결과를 도출하였으며, 기존의 여론 조사 결과보다 일정 수준 향상되는 모습을 나타내었으며, 이는 득표율 예측에 있어 비정형 데이터의 분석 결과를 보완하는 방법에 대하여 보다 심도있는 연구가 필요하다는 것을 의미한다고 할 수 있겠다.

4. 결론 및 제안

기존의 여론 조사 방법에 비정형 데이터 분석 결과를 반영하여 새롭게 예측률을 추정하고자 본 논문에서 제시한 방법을 적용한 결과 A 후보는 오차가 1.9%에서 1.1%로 감소하였으며, C 후보 역시 상당 수준의 오차가 감소하는 것을 확인할 수 있었다. 하지만 B 후보의 경우에는 예측률의 오차가 다소 증가하는

결과가 나타났다.

이는 후보별로 쓰이는 문장의 차이에 따라 나타난 오차로 예측률이 개선된 두 후보와 관련된 데이터에는 직설적인 표현이 많았던 것에 비해 다른 한 후보의 경우에는 중의적인 어구와 비꼬는 문장이 타 후보보다 많이 나타났으며 이러한 이중적인 의미를 지닌 문장에 대한 재해석이 필요함을 알 수 있었다.

이는 비정형 데이터 분석이 가지는 한계점으로 복합명사의 경우 명사 추출과정에서 본래의 의미가 소멸되는 문제, 후보별 별명과 애칭 등이 지역별 특색에 따라 다르게 나타나는 문제 등 사전 구축에 대한 제한사항을 첫번째로 꼽을 수 있겠다. 또한 한국어가 가지는 중의적인 표현의 특수성을 효과적으로 반영해야 하는 문제점 및 온라인 상에서의 비정형 데이터의 경우 여론 조사와는 달리 복수의 대상에게 동일한 의견을 나타낼 수 있다는 점 등 분석 과정에서의 제한 사항이 추가적인 고려 대상이 될 수 있다.

그러나 이러한 비정형 데이터 분석이 가지는 제한 사항에도 불구하고 간단한 비정형 데이터의 분석 결과를 보완하는 방식을 적용하여 오차를 감소시키는 효과가 실증된 본 논문의 시도는 향후 비정형 데이터에 대한 심도 있는 추가적인 연구 진행이 매우 큰 의미를 가질 수 있음을 시사한다. 따라서 앞서 제시한 사전과 과정의 문제점을 개선하는 방안 및 핵심 명사와의 거리를 계산해 명사별로 점수를 줄 수 있는 단어 거리분석(N-Gram) 등 문장의 특성을 추출하는 거리가중치 알고리즘에 대한 연구를 추가적으로 제안하는 바이다.

References

Bae, J. H., Son, J. E., and Song, M. S. (2013). Analysis of twitter for 2012 South Korea presidential election by text mining techniques, *Journal of Intelligence and Information Systems*, **19**, 141–156.

Chang, J. Y. (2009). A sentiment analysis algorithm for automatic product reviews classification in on-line shopping mall, *The Journal of Society for e-Business Studies*, **14**, 19–33.

Choi, D. S., Mun, G. J., Kim, Y. M., and Noh, B. N. (2011a). An analysis of large-scale security log using MapReduce, *Korean Institute of Information Technology*, **9**, 125–132.

Choi, H., Tak, Y., and Hwang, E. (2011b). Music recommendation scheme based on twitter analysis. In *Proceedings of The 38th KIISE Fall Conference*, **38**, 279–282.

Choi, M. and Yang, S. (2009). Internet social media and journalism report, Korea Press Foundation, 2009-1

Hyun, K. (2010). Election polling, what is problem?, *Kwanhun Journal*, **116**, 9–17.

Jho, H. and Kim, J. (2012). Political communication and civic participation through blogs and twitter, *Journal of Cybercommunication Academic Society*, **29**, 95–130.

Kim, J. H. and Jung, H. (2017). Causal study on the effect of survey methods in the 19th presidential election telephone survey, *The Korean Journal of Applied Statistics*, **30**, 943–955.

Kim, S. Y. and Huh, M. H. (2009). Systematic bias of telephone surveys: meta analysis of 2007 presidential election polls, *The Korean Journal of Applied Statistics*, **22**, 375–385.

Kim, S. Y. and Kwon, S. P. (2009). The effect of survey refusal and noncontact on nonresponse error: for economically active population survey, *The Korean Journal of Applied Statistics*, **22**, 667–676.

Kim, S. and Hwang, B. (2014). Propensity analysis of political attitude of twitter users by extracting sentiment from timeline, *Journal of Korea Multimedia Society*, **17**, 43–51.

Kim, Y. and Jeong, S. R. (2013). Intelligent VOC analyzing system using opinion mining, *Journal of Intelligence and Information Systems*, **19**, 113–125.

Kim, W., Lee, J., Park, J., and Choi, J. (2014). A technique of the approval rating analysis for political party using opinion mining, *The Journal of Korean Institute of Information Technology*, **12**, 133–141.

Kramer, A. D. I., Guillory, J. E., and Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks, *PNAS*, **111**, 8788–8790.

Lee, J. H., Kim, J., and Lee, K. J. (2006). Missing imputation methods using the spatial variable in sample survey, *The Korean Journal of Applied Statistics*, **19**, 57–67.

Lee, S. and Lee, D. (2015). Real time predictive analytic system design and implementation using Bigdata-

log, *Journal of The Korea Institute of Information Security and Cryptology*, **25**, 1399–1410.

Park, C., Lim, S., Cha, S., Lee, I., and Kim, J. (2014). Formation of weak ties in social media, *The Korea Contents Association*, **14**, 97–109.

Park, J., Lee, H., Kang, K., and Kim, B. (2018). Real-time pavement damage detection based on video analysis and notification service, *KIISE Transactions on Computing Practices*, **24**, 59–66.

Park, S. J., Jung, W. H., Han, J. H., and Shin, S. J. (2004). Analysis of affective words on photographic images and the effects of color on the images, *Korean Journal of the Science of Emotion and Sensibility*, **7**, 41–49.

Wegrzyn-Wolska, K. and Bougueroua, L. (2012). Tweets mining for French presidential election, *Computational Aspects of Social Networks*, 2012 Fourth International Conference, 138–143.

Williams, C. and Gulati, G. (2008). What is a social network worth? Facebook and vote share in the 2008 presidential primaries, *American Political Science Association*, Annual Meeting, 1–17.

비정형 데이터 분석을 통한 선거 여론조사 예측력 개선 방안 연구

박선빈[a] · 김명준[a,1]

[a]한남대학교 비즈니스통계학과

(2018년 8월 8일 접수, 2018년 8월 27일 수정, 2018년 8월 27일 채택)

요 약

소셜 네트워크 서비스(social network service; SNS)는 개개인의 의견을 공유하거나 소통하는 일반적인 도구로 사용되고 있으며, 특히 정치적인 이슈의 전파 과정에서 타인과의 공유를 통하여 자신이 지지하는 후보에 대한 긍정적인 홍보 등을 통해 여론을 형성 또는 확장한다. 기존의 여론 조사 결과는 응답률, 표본 수집의 방식 등과 관련하여 예측의 정확성에 대한 끊임없는 논란이 되어왔다. 본 논문은 이러한 소셜 네트워크 서비스 상에 존재하는 수많은 비정형 데이터의 감성 분석을 통하여 여론조사의 예측력을 개선, 보완하는 방안을 제시하고자 한다. 제시하고자 하는 연구 내용은 비정형 데이터 크롤링 및 기존에 사용되던 감성 사전에 대한 추가적인 보정 과정을 포함하고 있으며, 이를 통하여 본 논문에서 제안하는 방식은 오차의 감소를 통하여 예측력을 개선하는 결과를 나타냈다.

주요용어: 감성 분석, 사회 관계망, 데이터 크롤링, 감성사전

[1]교신저자: (34430) 대전광역시 대덕구 한남로 70, 한남대학교 비즈니스통계학과. E-mail: mkim@hnu.kr