텍스트 마이닝을 활용한 대선 관련 SNS 분석

권영우* · 정덕길*

*동의대학교.

SNS Analysis Related to Presidential Election Using Text Mining

Young-Woo Kwon* · Deok-Gil Jung**

*Dong-eui University

E-mail : kyu2369@deu.ac.kr · dgjung@deu.ac.kr

요 약

최근 소셜 미디어의 이용률이 폭발적으로 증가함에 따라, 방대한 데이터가 네트워크로 쏟아져 나오고 있다. 이들 데이터는 기존의 정형 데이터뿐만 아니라 이미지, 동영상 등의 비정형 데이터가 있으며, 이들을 포괄하여 빅데이터라고 불린다. 이러한 빅데이터는 오피니언 마이닝, 테스트 마이닝 등의 기술적인 분석 기법과 빅데이터 요약 및 효과적인 표현방법에 대한 시각화 기법에 대하여 활발한 연구가 이루어지고 있다. 이 논문은 인기 있는 사회연결망 서비스인 Twitter의 트윗을 수집하고, 빅데이터 분석 기법인 텍스트 마이닝을 활용하여 2017년 대선에 대하여 분석하였다. 또한 분석된 자료의 효과적인 전달을 위해 워드 클라우드 진행하였다. 이 논문을 위하여 인기 있는 SNS인 Twitter의 최근 7일간 트윗(tweet)을 수집하고 분석하였다.

키워드

대선, 텍스트 마이닝, 텍스트 분석, SNS-빅데이터

Ⅰ. 서 론

최근 스마트폰은 빠른 보급을 바탕으로 사용자의 생활 속 깊이 자리 잡고 있다. 세계 주요 50개국의 스마트폰 보급률 평균은 70%이며, 국내 스마트폰 보급률은 91%로 육박하고 있다[1]. 이러한 스마트 폰은 타인과 소통하는 중요한 수단이 되었으며, 그 수단을 사회연결망 서비스(Social Network Service)라 부른다. 국내 사회연결망 서비스는 카카오스토리, 밴드, 페이스북, 인스타그램, Vingle, Twitter 등이 있으며, 인기 있는 SNS인 페이스북과 트위터의 사용자수는 각각 1,440만명, 530만이 사용되고 있다[2].

SNS 사용자들은 각기 다른 생활환경에서 일어나는 일들을 등록하여 데이터들을 생산하고 있으며, 이렇게 SNS상에 축척된 데이터를 SNS-빅데이터라 부른다. 빅데이터의 형태는 정형화의 정도에 따라 정형, 반정형, 비정형 데이터로 분류 된다. 정형 데이터란 단순히 고정된 필드에 저장된 데이터를 말하며, 반정형 데이터는 고정된 필드에 저장되어 있지는 않지만 메타데이터, 스키마 등을 포함하는 데이터나 XML, HTML 텍스트 등을 말한다. 마지막으로 비정형 데이터는 이미지, 동영상, 스프레드시트, e-mail, 인터넷 페이지 등을 예로 들 수 있다[3]. Twitter의 트윗은 비정형 데이터 포함된다.

텍스트 마이닝(Text Mining)이란 비정형 데이터를 수집, 처리, 추출, 분석의 과정(그림 1)을 거쳐 분석 가능한 데이터로 가공하는 과정을 말한다[4].

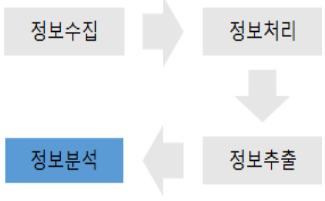

그림 1. 텍스트 마이닝 데이터 분석 4단계

텍스트 마이닝 분석은 고객 맞춤형 마케팅, 질병 예측, 행동패턴 분석 등 다양한 분야에서 활용 가능하며, 이를 위해 대학, 기업에서 정부까지 분석하기 위해 많은 노력을 하고 있다. 정부에서는 미래의 경쟁력 우위를 좌우하는 중요한 자원이라

고 판단하고 있으며, 빅데이터 전문 인재양성, 활용방안 등에 대해 투자를 하고 있으며, 기업에서는 SNS의 고객정보, 소비패턴, 요구사항 등의 빅데이터 분석 결과를 마케팅전략 수립에 적극 활용하고 있다[5].

이 논문에서는 Twitter의 트윗을 수집대상으로 하고 있으며, 수집된 트윗를 바탕으로 텍스트 마이닝 분석, 시각화 하였다. 이 논문에서 제안하고자 하는 범위는 트윗을 수집하는 부분, 텍스트 마이닝 부분, 시각화 부분으로 구성되어 있다.

II. 연구 방법 및 설계

Twitter의 트윗을 수집하기 위해 개발자용 서비스 홈페이지(http://dev.twitter.com/)에서 제공하는 Twitter API와 빅데이터 통계 분석 도구인 RStudio에서 연동 하였다[6]. 이 논문에서는 그림 2와 같이 트윗 수집, 텍스트 마이닝, 데이터 시각화를 구현하였다.

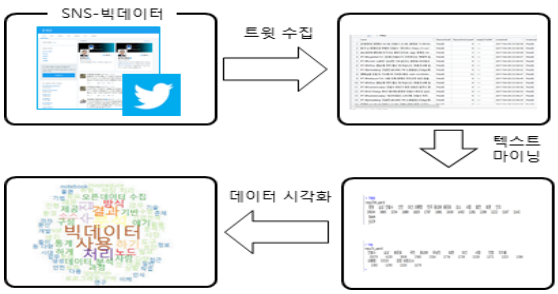

그림 2. 구현 내용

2.1 데이터 수집 방법

Twitter의 트윗의 정보를 사용하기 위해 Twitter API와 RStudio를 연동하였으며, API의 package 내부에 존재하는 함수인 searchTwitter 함수를 활용하여 트윗을 수집하였다.

searchTwitter 함수는 searchString, since, until, locale 등 검색을 위한 속성 값을 가지며, 속성 값에 따라 원하는 결과를 수집할 수 있다[7]. 이 논문에서는 searchString의 속성을 활용하여 "안철수", "문재인" 2가지 키워드를 중심으로 트윗을 수집하였다.

2.2 텍스트 마이닝

수집된 트윗에 대하여 불필요한 자음 또는 모음만으로 이루어진 텍스트나, 특수기호 등 사전에 정의되지 않은 문자 제거, 문자 분리, 문자 카운트를 통한 데이터 분석을 위해 필터링하였다. 문자 제거를 위하여 gsub, unlist, Filter 등의 함수를 사용하였으며, 문자 분리 및 추출하기 위하여 KoNLP 패키지의 extractNoun 함수를 사용하였다. 마지막으로 빈도분석를 위하여 table 함수를 사용하였다[8].

2.3 데이터 시각화

텍스트 마이닝을 통해 구조화된 데이터는 R 패키지를 통해 분석한다. R은 선형 및 비선형 모델, 통계적 테스트, 시계열 분석, 클러스터링 등 복잡한 통계 분석과 Graphics(Visualization) 분야에 특화된 언어와 통계 분석 패키지로 구성된 오픈소스 소프트웨어이다. R은 기본적인 통계 기법부터 모델링 등 다양한 기법이 구현 가능하며, 구현된 결과는 다양한 시각 패키지를 활용하여 사용자에게 보여준다[9]. 이 논문에서는 수집된 데이터를 중심으로 빈도분석을 실시하였고, 비정형 데이터 분석에 많이 활용되는 워드 클라우드를 통하여 시각화하였다.

III. 결 론

이 논문에서는 II장에서 언급한 기술을 이용하여 2017년 04월 20일~2017년 04월 26일까지 생성된 트윗을 수집하였고, 빈도분석을 진행하였다. 위 기간 동안 키워드 "안철수"가 언급된 트윗은 약 90만 건 이였으며, 키워드 "문재인"이 언급된 트윗은 약 170만 건 이였다. 키워드에 대하여 수집된 트윗은 그림 3과 같이 수집된 순서대로 정리되어 RStudio에 출력한다.

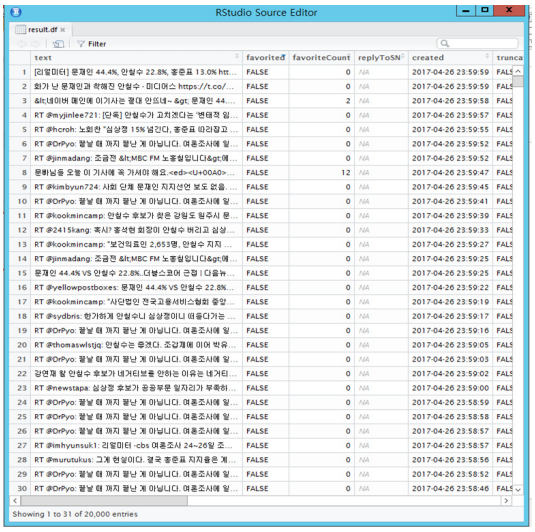

그림 3. 수집 내용

3.1 수집된 트윗 분석

키워드 "안철수", "문재인"을 중심으로 수집된 트윗을 텍스트 마이닝 과정을 통해 의미 없는 단어를 제외하였다. 표 1에서는 각 키워드 별로 추출된 단어들의 빈도수를 측정한 후 상위 15개를 순서대로 나열한 결과를 표시하고 있다.

표 1. 각 키워드 별 상위 15개의 단어 검색 결과

구분 / 빈도 순위	키워드 "안철수"	키워드 "문재인"
1	안철수	문재인
2	문재인	심상정
3	심상정	안철수
4	홍준표	선언
5	국민	대선
6	동성애	대통령
7	유승민	민주
8	토론	동성애
9	대선	홍준표
10	사람	성소수자
11	안랩	사람
12	지지율	발언
13	대통령	토론
14	지지자	Daum
15	경영	당선

질의 결과 중 눈에 띄는 단어로 "동성애"와 "성소수자"가 있다. 실제 두 후보는 질의 기간인 2017년 04월 20일~2017년 04월 26일 사이에 대선 토론에서 동성애에 대해서 언급된 적이 있다.

그림 4. 키워드 "안철수"의 워드 클라우드 결과

그림 5. 키워드 "문재인"의 워드 클라우드 결과

검색 결과의 가시성을 높이기 위해 표 1의 자료를 바탕으로 워드 클라우드를 진행하였다. 자료를 시각화를 위해 R packages인 wordcloud 내부의 wordcloud 함수를 사용하였으며, 함수 내부의 속성인 시각화 사이즈, 폰트 속성 등을 이용하여 표현하였다.

3.2 키워드별 분석 결론

본 연구에서는 2017년 04월 20일~2017년 04월 26일 동안에 주요 키워드 포함하는 트윗을 중심으로 분석을 실시하였다. 본 연구는 인기 있는 SNS를 텍스트 마이닝을 통해 수집하고 R패키지를 통해 분석하였다는 점에서 의미가 있다. 또한 최근 사회적으로 큰 이슈가 되고 있는 2017년 대선에 대하여 분석하고 관찰하였다. 해당 후보에 대한 이슈가 되었던 단어는 빠르게 업데이트가 되고 있음을 확인할 수 있었다. 하지만 본 연구는 특정기간에 한정되어 데이터를 수집했다는 점에서 한계점을 갖는다. Twitter에서 지원하는 API는 최근 트윗을 검색할 수 있는 searchTwitter 함수가 존재하지만 최근 7일간의 데이터만 제공하기 때문에 정확한 데이터 검색에 어려움이 있었다. 또한 검색 키워드는 후보의 이름을 중심으로 검색되기 때문에 후보에 대한 속어나 은어에 대해서는 검색되지 않았다. 그렇기 때문에 분석된 데이터가 정확한 예상 결과라고 단정 짓기 어렵다.

최근 SNS의 분석은 감성분석을 통한 사용자의 심리 분석까지 요구하고 있다. 다음 연구에서는 SNS 뿐만 아니라 뉴스, 저널 등을 활용한 데이터의 확장과 감성분석을 통해 Twitter 사용자의 심리 상태까지 관찰한다면 보다 정확한 데이터가 될 수 있을 것으로 기대 된다.

참고문헌

[1] http://www.dt.co.kr/contents.html?article_no=2016070102100151780001
[2] 양민혁 외, "SNS 데이터를 활용한 국내대학 인식 및 선호도", 한국빅데이터서비스학회, 1권 1호, pp.1-13, 2014.
[3] 조성우, "Big Data 시대의 기술", 중앙연구소, kt종합기술원, pp.1-8, 2011.
[4] 정근하, "텍스트마이닝과 네트워크 분석을 활용한 미래예측 방법 연구", 한국과학기술기획평가원, 2010.
[5] 김유영 외, "영화 리뷰 감성분석을 위한 텍스트 마이닝 기반 감성 분류기 구축", 지능정보연구, pp.71-89, 2016.
[6] http://dev.twitter.com/
[7] Jeff Gentry, "Package twitteR", 2016.8
[8] https://cran.r-project.org/web/packages/KoNLP/index.html
[9] 이종화 외, "Data Dictionary 기반의 R Programming을 통한 비정형 Text Mining Algorithm 연구", 한국정보산업학회논문지, pp.113-124, 2015.