Journal of Convergence for Information Technology
Vol. 9. No. 11, pp. 81-91, 2019
e-ISSN 2586-4440
DOI : https://doi.org/10.22156/CS4SMB.2019.9.11.081

트위터 사용자들의 감성을 이용한 사회적 이슈 분석

김한나[1], 정영섭[2*]
[1]순천향대학교 미래융합기술학과 석사과정, [2]순천향대학교 빅데이터공학과 조교수

Social Issue Analysis Based on Sentiment of Twitter Users

Hannah Kim[1], Young-Seob Jeong[2*]
[1]Student, Department of Future Convergence Technology, Soonchunhyang University
[2]Assistant Professor, Department of Bigdata Engineering, Soonchunhyang University

요 약 대중들의 소통의 창구로 자리매김 하고 있는 소셜 네트워크 서비스(SNS)에 작성된 글은 감성을 많이 포함하고 있다는 특징을 갖고 있다. 그 중 트위터는 공개 Application Programming Interface(API)를 통한 데이터의 수집이 편리하다는 장점을 지니고 있다. 본 논문에서는 트위터 상에 표현된 사용자들의 감성 정보를 통해 사회적 이슈를 분석하고 마케팅 분야 활용 가능성을 제시한다. 이는 국민 또는 소비자의 의견과 반응을 필요로 하는 정부, 기업 등에 도움이 될 수 있다. 본 논문에서는 최근 사회적 이슈에 대한 트위터 텍스트 데이터를 긍정 또는 부정으로 분류하여 질적 분석을 제공하였고, 각 트윗의 좋아요 수, 리트윗 수 등에 대한 상관관계 분석을 통해 양적 분석을 제공하였다. 질적 분석의 결과로 국민의 지지를 얻기 위해 관세정책을 홍보하고, 버즈 사용자에게는 기술적 편의를 제공할 것을 제안하였다. 양적 분석의 결과, 트위터 사용자들의 관심을 끌기 위해서는 긍정적인 트윗을 짧고 간단하게 작성해야 함을 밝혔다. 데이터의 수집 기간이 짧고, 단 두 가지의 키워드만을 분석하여 일반화 가능성이 떨어지는 한계를 가져 향후, 보다 긴 기간의 다양한 사회적 이슈를 분석할 예정이다.

주제어 : 소셜 네트워크 서비스, 트위터, 감성 분류, 사회적 이슈 분석, 합성곱 신경망

Abstract Recently, social network service (SNS) is actively used by public. Among them, Twitter has a lot of tweets including sentiment and it is convenient to collect data through open Aplication Programming Interface (API). In this paper, we analyze social issues and suggest the possibility of using them in marketing through sentimental information of users. In this paper, we collect twitter text about social issues and classify as positive or negative by sentiment classifier to provide qualitative analysis. We provide a quantitative analysis by analyzing the correlation between the number of like and retweet of each tweet. As a result of the qualitative analysis, we suggest solutions to attract the interest of the public or consumers. As a result of the quantitative analysis, we conclude that the positive tweet should be brief to attract the users' attention on the Twitter. As future work, we will continue to analyze various social issues.

Key Words : Social network service, Twitter, Sentiment classification, Social issues analysis, Convolutional neural networks

*This work was supported by the Soonchunhyang University Research Fund. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP; Ministry of Science, ICT & Future Planning) (No. 2019021348).

*Corresponding Author : Young-Seob Jeong(bytecell@sch.ac.kr)

Received July 19, 2019
Revised October 15, 2019
Accepted November 20, 2019
Published November 28, 2019

1. 서론

과거의 사용자들은 자신의 생각과 정보를 공유할 수 있는 개방된 온라인 채널인 소셜미디어(블로그, UCC 등)를 활용해왔다[1]. 이후 소셜미디어에 인맥관리가 서비스로 추가된 Social Networking Service(SNS)가 활성화되었으며, 모바일의 발달로 인해 접근성 또한 좋아져 많은 사람들이 쉽고 간단하게 서로 소통할 수 있게 되었다. 국내에서는 싸이월드, 아이러브스쿨 등이 크게 주목을 받았으며, 현재 SNS 시장을 주도하고 있는 페이스북과 트위터 이용자수는 이미 2018년에 18억 명을 돌파하였다[2]. SNS는 사람들이 텍스트, 이미지, 동영상과 같은 콘텐츠를 사용하여 자신의 생각을 표현하는 장이 되었다. 트위터의 경우 사용자들이 편하게 자신의 생각을 작성할 수 있는 공간이고, 공개 Application Programming Interface(API)를 사용하여 필요한 데이터를 수집할 수 있기 때문에 트위터 텍스트 데이터를 분석하는 연구가 활발하다[3-8]. 특히, 트위터 데이터가 가지는 감성(긍정/중립/부정 등)을 분류하는 것도 큰 의미를 갖는다. 감성분류를 통해 사용자들의 의견을 식별할 수 있다면 사용자들의 정치적, 사회적 선호도를 파악[3, 9]할 수 있을 뿐 아니라 사용자 맞춤형 서비스를 제공[4-6, 10]하거나 특정 SNS 계정의 활성화[7, 8] 등에도 사용될 수 있다. 따라서 본 논문에서는 감성분류기로 트위터 데이터를 감성(긍정 또는 부정) 분류하고 각 데이터가 갖는 특성을 분석하여 사회적 이슈를 고찰한다.

제2장에서는 트위터 분석 및 감성 분류 모델과 관련된 연구를 소개한다. 제3장에서는 데이터 수집 및 전처리 등과 같은 연구방법을, 제4장에서는 연구의 결과를 제시한다. 제5장에서는 결론과 향후 연구의 필요성을 시사한다.

2. 이론적 배경

SNS 사용자들은 프로슈머(prosumer)로써 정보의 생산과 소비의 중심에 서 있다. 즉, 생산과 소비를 동시에 주도하고 있으며, 모바일의 발달로 인해 그 정보들은 실시간으로 생산 및 공유가 가능해지고 있다[11]. 따라서 SNS를 통해 생산되는 정보들은 최근 사회의 동향 및 사용자들의 행동 패턴을 파악할 수 있는 근거가 된다[12]. 또한, 최근 사회적 실제를 이해하는 도구로써 각광받게 되었다[11]. 즉, SNS 에 포함된 사용자들의 의견 및 감성을 분석하는 것은 현재 사회에서 발생하는 이슈들에 대해 고찰할 수 있는 하나의 방법이 된다. 분석의 결과들은 사용자들의 특성과 관심사에 대한 맞춤형 정보를 제공하거나[13, 14], 새로운 비즈니스 기회를 창출하는 원천으로 활용할 수 있다[15]. 이미 외국에서는 국가 차원에서 '이슈 스캐닝'을 활용하여 현재 사회가 당면한 여러 이슈들에 대한 과학적인 파악을 수행하고 있다[16]. 이와 같이 사회 이슈를 분석하는 일은 현대 사회에서 매우 중요하기 때문에 본 논문에서는 트위터 데이터를 감성 분류하여 분석함으로써 사회 이슈 분석에 대한 가능성이 있음을 제시하였다.

2.1 트위터 분석

대부분의 트위터 데이터 분석 연구는 사람이 직접 감성 혹은 성별, 정보 유형(정보적, 교육적, 오락적 등), 성격 유형(성실성, 외향성, 친화성 등)에 따라 분류하여 질적 분석을 수행하거나, 분류기를 사용하여 높은 성능을 내는 양적분석을 수행한다. 본 논문에서는 트위터 데이터에 분류기를 적용하여 긍정 또는 부정으로 분류하는 질적 분석을 수행하였고, 트윗의 특성(트윗 길이, 특수문자 유무 등)과 트윗의 관심도(좋아요, 리트윗)의 상관관계를 수치적으로 해석하는 양적분석을 모두 수행했다.

Table 1. Use case of analysis result

Paper	Use case
[3, 9]	Identify users' political and social opinions
[4-6, 10]	Suggest marketing solutions
[7, 8]	Suggest SNS account activation solutions
Our paper	Analyze social issues and suggest the uses possibility of marketing

트위터 데이터 분석결과는 여러 분야에서 활용된다. 사용자들의 정치적, 사회적 의견을 파악하는데 활용되기도 하며, 개인화 서비스와 같은 마케팅 또는 특정 SNS 계정의 활성화에 활용되기도 한다. Table 1은 트위터 데이터 분석 결과 활용사례를 유형별로 분류하여 나타낸 것이다. 사용자들의 정치 및 사회적 의견을 파악하기 위해 트위터 데이터를 분석한 연구는 다음과 같다. Lee 등[9]은 약 100만개의 트윗을 수집하여 명사만을 추출해 분석에 사용하였다. 토픽모델링의 한 종류인 잠재 디리클레 할당(LDA)을 사용하여 15개의 토픽으

로 군집화하였다. 추출된 토픽과 사회현상을 비교한 결과 트위터 메시지에는 사회적 이슈를 나타내는 주제가 들어있음을 확인하였다. Jung 등[3]은 '세월호 사건'을 키워드로 한 트윗 560만 건의 본문, 리트윗 수, 사용자 계정을 수집하였다. 사용자 계정 정보를 사용하여 이용자 집단을 식별하고 집단 간 공통 단어를 추출하였다. Word2vec[17]을 사용하여 분석하고 동일 단어에 대한 집단 간 의미론적 차이를 분석하였다. 진보성향을 가진 사용자가 트위터에서 활발히 활동하고 더 큰 영향력을 가진다는 분석 결과를 보였다.

분석한 결과를 마케팅에 활용하고자 하는 연구는 다음과 같다. Hong 등[10]은 나이키 신제품 신발을 키워드로 한 78개의 트윗을 수집하였다. 텍스트의 유사성을 근거로 인지적 반응, 정서적 반응, 상호작용, 구매의도, 제품애착도의 5가지로 범주화하였다. 각 트윗에 대한 질적 분석을 수행하고 마케팅에 활용할 수 있는 방안을 제시하였다. Lee 등[4]은 전자제품 갤럭시 기어 S2를 키워드로 하여 출시 전 트윗 1,377개, 출시 후 트윗 3,426개를 수집하였다. 감성단어(좋다, 나쁘다)와 속성단어(배터리, 디자인)로 사람이 직접 구분하여 출현빈도가 높은 단어를 시각적으로 표현하였으며 제품에 대한 마케팅 전략을 제시하는데 도움을 주고자 하였다. Ahn 등[5]은 26,000여개의 트윗을 수집하고 167명의 성격정보를 설문으로 수집하였다. 주성분분석[18]을 통해 트윗을 10개의 요인으로 분류하였고, 사용자 성격을 경험개방성, 성실성, 신경증, 외향성, 친화성으로 분류하여 트윗과의 상관관계를 분석하였다. 사용자 성격에 따른 맞춤형 서비스나 정보를 제공하여 가치를 창출할 수 있다는 점에서 의의가 있는 연구이다. Hong 등[6]은 한국도로공사의 Voice of Customer (VOC) 데이터 1,031건과 트위터 데이터 138건을 수집하여 사용하였다. 수집한 데이터로 감성사전을 구축하고 사전을 활용하여 각 데이터의 감성을 판단하였고, 76.81%의 매칭률을 얻었다. 구축한 감성사전을 트윗에 적용하였고, 긍정, 부정 트윗의 감성 키워드를 출현빈도를 기준으로 하여 시각적으로 제공하여 유의미한 결과를 도출하였다. 비정형 교통데이터를 교통정보 전달 매체 및 사용자 감성 모니터링 매체로서 활용할 것을 제안하였다.

국가 SNS 계정의 활성화 방안으로 트위터 데이터 분석을 수행한 연구도 있었다. Gang 등[7]은 4개국의 국가기록관에서 작성한 트윗과 트윗 수, 리트윗 수, 좋아요 수를 수집하여 정보유형 분석 및 시계열 분석을 수행하였다. 미국의 국가기록관인 NARA와 영국의 TNA, 호주의 NAA, 우리나라의 국립기록원을 비교하였다. 뉴스 및 업데이트, 외부 정보 제공, 채용공고 등 6가지 정보 유형으로 분류하여 분석하였다. 또한 시간의 흐름에 따른 트위터 운용 현황 및 이용자의 반응 추이를 살펴보고자 시계열 분석을 진행하였다. 이미지 및 해시태그의 사용으로 국가기록원 계정이 활성화될 수 있다는 결론을 제공하였다. Gang 등[8]은 우리나라의 국가기록원과 대통령 기록관의 트윗과 트윗 수, 리트윗 수, 좋아요 수, 작성 일자를 수집하여 데이터로 사용하였다. 시기별 사회적 이슈와 트윗의 상관관계를 분석하여 이용자의 관심도를 밝혔다. 사회적 이슈를 포함한 게시글을 활발하게 업로드한다면 사용자들의 관심을 끌 수 있다는 결론을 제시하였다.

트위터 데이터에는 사용자들의 감성표현이 많이 포함되어 있으므로 감성을 분석한 결과는 맞춤형 서비스 제공과 같은 마케팅이나 SNS 계정 활성화 등 다양한 분야에 활용될 수 있다. 이상의 선행연구들은 하나의 키워드를 분석 주제로 삼아 다양한 유형(감성, 정보 유형, 성격 유형)으로 분류하는 질적 분석을 수행하였으나 본 논문에서는 질적 분석과 양적 분석을 모두 제공한다. 본 논문은 두 가지의 다른 키워드를 분석 주제로 삼아 긍정 또는 부정으로 감성분류하고 그 결과에서 보이는 사회적 이슈를 질적 분석하여 사용자들의 의사결정에 도움을 주는 결과를 도출한다. 또한 트윗의 특성(길이, 특수문자 유무 등)과 트윗의 관심도(좋아요, 리트윗)의 상관관계를 수치적으로 해석하여 그 결과를 홍보 및 마케팅에 적용시킬 수 있는 가능성을 제시한다. Hong 등[6]의 연구와 같이 감성 분류 모델을 훈련시키는데 사용한 데이터와는 유사한 성격을 가지는 다른 종류의 데이터를 적용시켜 분석하였다.

2.2 감성 분류 모델

사용자들이 SNS를 통해 표현하는 정서는 감정과 감성으로 구분할 수 있다. 감정은 기쁨, 행복, 만족 등으로, 감성은 긍정, 부정, 중립 등으로 분류된다. 또한 최근에 주로 사용하는 분류도구로는 머신러닝 또는 딥러닝이 있다. 머신러닝은 자질을 어떻게 정의하느냐에 따라 성능이 크게 변화하는 특성을 가지고 있어 자질 정의에 많은 시간과 노력이 요구된다. 딥러닝의 경우 자

동으로 임의의 자질을 추출하며 머신러닝과 비교했을 때 더 높은 성능을 보인다. 본 논문에서는 딥러닝을 사용하여 트위터 데이터를 감성(긍정, 부정)으로 분류하고, 질적 분석을 수행하여 사회적 이슈를 다룬다.

텍스트 처리에 Recurrent Neural Network (RNN)[19]과 Convolutional Neural Network (CNN)[20-22]을 사용한 여러 연구가 수행되어왔다. 텍스트 감성분류에 RNN을 변형한 모델을 사용한 연구는 다음과 같다. Tang 등[23]은 Gated Recurrent Neural Network (GRNN)를 제안하여 Yelp 2013 - 2015 data와 IMDB data에서 각각 66.6%와 45.3%의 정확도를 얻었다. Qian 등[24]은 Long Short-Term Memory (LSTM)[25]를 사용하여 영화리뷰 데이터를 긍정, 부정으로 이진분류하였고 82.1%의 정확도를 얻었다. Huang 등[26]은 Hierarchical Long Short-Term Memory (HLSTM)를 사용하여 웨이보 트윗 텍스트 데이터를 분류하였고 64.1%의 정확도를 얻었다. CNN을 분류에 사용한 연구들도 수행되었는데, Kim[20]은 한 층의 컨볼루션을 사용하여 7가지의 다른 종류 데이터를 분류하였고 최대 89.6%의 정확도를 얻었다. Zhang 등[27]은 Kim[20]의 CNN에 NB-SVM을 결합하여 더 나은 성능을 얻었고, 4가지 데이터셋에 대해 최대 97.2%의 분류 정확도를 얻었다. Kim 등[28]은 두 층의 컨볼루션와 한 층의 풀링을 결합한 네트워크를 제안하였고 [20, 27]의 모델과 이진 감성 분류 성능을 비교하였고, 영화리뷰 데이터에 대해 81%의 정확도를 얻었다.

RNN은 은닉층의 출력이 은닉층의 입력으로 다시 제공되는, 즉 학습이 반복적으로 이루어지는 특성을 가지고 있다. RNN이 텍스트 데이터에 적용되었을 때, 현재의 단어만으로 의미를 해석하는 것이 아니라 앞 또는 뒤 단어와의 관계성을 통해 현재 단어의 의미를 해석한다. 즉, 단어의 등장 순서를 고려하여 학습이 이루어지기 때문에 텍스트의 순차적인 패턴 분석에 효과적이라고 알려져 있다. CNN은 컨볼루션 필터가 움직이며 모든 데이터를 훑어 특징적인 패턴(자질)을 추출하는 특성을 가지고 있어 이미지 처리에 특화되어 있다고 알려져 있다. 그러나 입력 데이터가 1차원으로 주어진다면 CNN의 같은 기능을 텍스트에서도 사용할 수 있다. CNN을 텍스트에 적용하였을 때, 컨볼루션 필터가 모든 텍스트를 돌아다니며 지역적 정보를 저장하고 주요 자질을 추출한다. 따라서 단어 하나하나가 아닌 몇 가지 주요 키워드 조합으로 표현되는 감성분류에 활용되기 적합하다[28, 29]. 또한 CNN은 RNN에 비해 적은 양의 데이터로 학습이 가능하므로 본 논문에서는 최근에 수행된 Kim 등[28]의 CNN을 트위터 데이터를 분류하는데 사용하였다.

본 논문에서는 여러 딥러닝 분류기 중 텍스트 감성분류에 적합한 CNN을 사용하여 트위터 데이터를 감성분류하였다. 긍정, 부정 트윗에 대하여 각각 단어의 출현빈도순으로 시각적 정보를 제공하고, 좋아요(like) 수와 리트윗(retweet) 수에 대한 수치적 해석을 제공하여 사회적 이슈를 분석하는 질적 분석과 양적분석 모두를 제공한다.

3. 연구 모형 및 가설

본 논문에서는 SNS의 최근 게시글로부터 사용자들의 감성을 파악해 사회적 이슈를 분석한다. 또한 트윗과 사용자들의 관심 정도(좋아요, 리트윗)의 상관관계를 분석하여 홍보 및 마케팅에 활용될 수 있는 가능성을 제시한다. (1) 데이터 수집 및 전처리 단계, (2) 감성분류 단계, (3) 시각화 및 수치적 해석 단계로 구성된다.

3.1 데이터 수집 및 전처리

파이썬 모듈 tweepy[30]를 사용하여 6월 17일 ~ 6월 24일, 8월 23일 ~ 9월 6일 기간 내의 트윗 중에 특정 키워드를 포함하는 트윗을 크롤링하였다. 정치 분야 및 기술 분야에 대한 대중들의 의견을 알아보고 사회 이슈를 분석하기 위해 특정 키워드를 대상으로 실험을 수행하였다. 미국의 트럼프 대통령은 6천 만 명의 팔로워를 갖고 있고, 월 평균 153건의 트윗을 게시하는 등 트위터를 활발하게 사용하고 있다[1]. 이는 미국 트위터 사용자는 물론 한국의 많은 누리꾼들의 관심을 받고 있다. 따라서 정치 분야에서는 트럼프 정부에 대한 대중들의 감성을 파악하기 위해 "Trump"를 키워드로 사용하였다. 또한, 최신 기술 분야에서는 출시 전부터 인기 검색어에 오르며 큰 관심을 얻고 있는 삼성의 블루투스 기기 갤럭시 버즈를 키워드로 삼았다. 갤럭시 버즈는 출시 5일만에 품절현상을 겪고, 미국 유력 소비자 전문

1) http://www.newstof.com/news/articleView.html?idxno=727

매체 컨슈머리포트(CR)의 무선이어폰 평가에서 1위를 달성하는 등 하나의 큰 이슈가 되었다[2]. 따라서 갤럭시 버즈에 대한 소비자들의 평판을 알아보기 위해 "Galaxy Buds" 또는 "GalaxyBuds"를 키워드로 사용하였다. “Trump”를 포함한 트윗 20,689개의 본문, 좋아요수, 리트윗 수를 수집하였고, 중복트윗을 제외했을 때, 9,437개의 데이터를 얻었다. “Galaxy Buds”를 포함한 트윗 13,775개의 본문, 좋아요 수, 리트윗 수를 수집하였고, 중복 트윗을 제외했을 때, 4,142개의 데이터를 얻었다. 데이터에 대한 자세한 통계는 Table 2에 나타내었다.

Table 2. Data statistics

Keyword	Trump	Galaxy Buds
Number of tweet	9,437	4,142
Maximum number of words in a tweet text	792	73
Average number of like	0.27	3.87
Average number of retweet	505.74	6.68

데이터에 대한 전처리 과정은 다음과 같다. 모든 단어를 소문자와 했으며, http로 시작하는 구문은 [URL]로, 숫자는 [NUM], @로 시작하는 구문은 [NAME]으로 변환하였다. 특수문자는 [SPE]로 변환하였고 두 번이상의 공백 및 탭은 삭제하였다. 감성분석에서 not의 중요성이 크기 때문에 isn't를 is not으로, can't를 can not으로 변환하였다. 공백 기준으로 단어를 구분하고 불용어를 삭제하였다. 본 논문에서는 1-gram 사전을 사용하였으며, 사전을 구축하기 위해 영화리뷰 데이터의 모든 단어와 각 키워드에 해당하는 트위터 데이터의 모든 단어를 합하였다. Hong 등[6]은 한국도로공사의 Voice of Customer (VOC)와 트위터 데이터를 결합하여 감성사전을 구축하였으며, 이를 트위터 데이터에 적용시켜 유의미한 결과를 보였다. 이처럼 본 논문에서도 영화리뷰 데이터와 트위터 데이터를 결합하여 감성사전을 구축하였고, 영화리뷰 데이터로 훈련시킨 감성 분류기를 트위터 데이터에 적용하여 유의미한 결과를 도출하였다. 영화리뷰 데이터는 본 논문에서 감성 분류를 위해 사용한 CNN 모델을 학습시키는데 사용되었으며, Hong 등[6]에서 사용한 VOC 데이터와 트위터 데이터의 공통점과 같이, 작성자의 개인적인 의견 및 감성을 담고 있다는 점에서 트위터 데이터의 특성과 일치한다. 따라서 영화리뷰 데이터 사전에도 감성 단어가 다수 포함되어 있으므로 트위터 데이터 사전과 결합하여 사용하였다. 즉, 영화리뷰 데이터에 포함된 모든 단어 9,396개와 Trump를 키워드로 하는 데이터에 포함된 모든 단어 16,246개를 결합하고 중복된 단어는 하나만 남기도록 하여 총 20,166개의 단어로 사전을 구축하였다. 또한, 영화리뷰 데이터에 포함된 모든 단어 9,396개와 Galaxy Buds 또는 GalaxyBuds를 키워드로 하는 데이터에 포함된 모든 단어 6,368개를 결합하고 중복된 단어는 하나만 남기도록 하여 총 13,474개의 단어로 사전을 구축하였다.

3.2 감성 분류

수집 및 전처리를 수행한 데이터를 분류기에 적용하여 감성 분류를 하였다. Kim 등[28]이 사용한 CNN모델은 영화리뷰 데이터로 학습되었으며, 다음과 같이 구성된다. : 임베딩층 + 컨볼루션층 + 컨볼루션층 + 풀링층 + 완전연결층 (Fig. 1). 임베딩 차원은 25로 설정했으며, 컨볼루션 필터 크기는 각각 3, 필터 개수는 각각 16과 8로 지정하였다. 풀링 크기는 2, 필터 간격은 1로 지정하였다. 출력층을 제외한 모든 층의 활성화 함수로 Rectified Linear Unit (Relu)[31]를 채택하였고, 출력층의 활성화 함수는 Softmax를 사용하였다. 0.001의 초기 학습률을 가지는 Adam optimizer[32]를 사용하였고, 완전 연결층의 가중치 행렬에 0.001을 곱하여 전체 손실에 더하는 L2-regularization을 사용하였다. 위와 같은 모델로 트위터 데이터를 긍정, 부정으로 분류하였다.

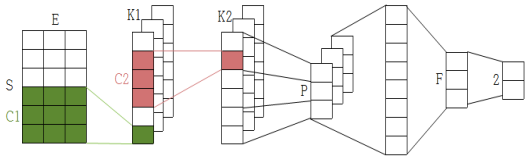

Fig. 1. Convolutional Neural Networks Architecture

3.3 시각화 및 수치적 해석

긍정으로 분류된 트윗과 부정으로 분류된 트윗에 대하여 워드클라우드로 나타낸 후, 빈도가 높은 단어를 포함하는 트윗을 골라내어 긍정 혹은 부정 트윗인

2) http://biz.chosun.com/site/data/html_dir/2019/08/15/2019081500410.html

지에 대한 질적 분석을 수행하였다. 또한 긍정 트윗과 부정 트윗의 좋아요 수, 리트윗 수 평균 값과 표준편차 값을 분석하여 부정 트윗의 경우 좋아요 수가 높은 반면 긍정 트윗의 경우 리트윗 수가 더 크다는 결과를 도출하였고, 이를 기반으로 SNS상에서 많은 관심을 얻기 위해서는 긍정적인 트윗을 작성해야 한다는 양적 분석 결과를 제공하였다.

4. 연구 결과

수집한 트위터 데이터를 Kim 등[24]의 CNN 모델을 사용하여 긍정 또는 부정으로 분류한 결과는 다음과 같다. “Trump”를 키워드로 한 데이터는 긍정 트윗이 5,812건, 부정 트윗이 3,625건으로 분류되었으며, “Galaxy Buds”를 키워드로 한 데이터는 긍정 트윗이 2,100건, 부정 트윗이 2,042건으로 분류되었다.

4.1 시각화

다음은 “Trump”를 키워드로 한 데이터에 대한 결과이다. Fig. 2는 부정 트윗에서 자주 등장한 단어로 워드클라우드를 그린 결과이다. Kill, devil과 같이 부정적인 단어도 존재하지만, 부정적인 성격을 띄지 않는 단어, 즉 break, remark와 같은 단어도 존재한다. 따라서 각 단어가 포함된 트윗이 실제로 부정적인 트윗인지 알아보기 위해 Table 3와 같이 빈도순으로 상위 10개의 부정단어를 포함한 트윗 중 사회 이슈를 포함하는 트윗들을 직접 감성 분석해보았다 . 트럼프에 대한 지지율을 높이기 위한 부정적인 여론조사에 불만을 가지는 경우도 있고, 수백만 명의 미등록 이민자의 대량 추방에 대한 언급도 존재했다. 또한, 국가 예산을 여가를 위해 사용하는 것을 불평하거나 무분별한 비난을 하는 경우도 있다. 대통령직을 잃을 위험 또는 이슬람 혐오증을 언급하며 비난하는 트윗도 존재했다. 분석 결과 단어 자체로는 부정적인 의미를 지니지 않는 단어들도, 이 단어를 포함하는 트윗들이 부정적인 성격을 띠는 경향이 있음을 알 수 있었다.

Fig. 3은 긍정 트윗에서 자주 등장한 단어로 워드클라우드를 그린 결과이고, Table 4은 빈도순으로 상위 10개의 긍정단어를 포함하는 트윗 중 사회 이슈를 포함하는 트윗을 보여준다. 트럼프 대통령이 약속을 잘 지키는 점을 칭찬하거나 관대한 억만장자라는 점을 언급하기도 한다. 또한 트럼프의 관세정책에 대해 칭찬하는 트윗도 존재했다. 즉, 긍정 트윗에 자주 등장하는 단어를 포함하는 대부분의 트윗은 긍정적인 트윗임을 알 수 있었다.

부정적인 트윗에서는 이민자 또는 이슬람인에 대한 차별을 비판하는 목소리가 드러났고, 긍정적인 트윗에서는 약속한 정책에 대한 수행을 칭찬하는 경우가 있었다. 따라서 Trump 정부가 이민자들과 이슬람인들에 대한 차별을 줄이고, 공약을 잘 지키는 점과 관세 정책에 대해 홍보함으로써 국민들로부터 좋은 평가를 받을 수 있을 것이라 예상된다.

Fig. 2. Frequent words in negative tweets about "Trump“

Table 3. Negative tweet samples about "Trump"

Words	Sentences
Approval	With so many negative polls for Trump you would think he would try to do things to drive up his approval rating but all he does is make Americans hate him more.
Break	BREAKING: Trump Tweet Leaks ICE Plans for Mass Deportations of 'Millions' of Undocumented Immigrants Starting Next Week.
Kill	All trump can do is tear down programs and kill plans and ideas.
Kill	Trump is homicidal/suicidal. He doesn't care if he kills billions of humans, insects, animals. In fact, he does not even care if he kills his own grandchildren. He's committed so many crimes, He is so deeply into his nazi craziness, that he is completely insane.
Budget	Trump just cut budget for food for 2.4 million poor Americans. He spends that per week when he goes to Mar-a-Lago
Defend	Wow. Jeremy Hunt has defended Trump's blatantly islamophobic attack on Sadiq Khan.
Risk	Rep. Alexandria Ocasio-Cortez. “I think we have a very real risk of losing the Presidency to Donald Trump.

Fig. 3. Frequent words in positive tweets about "Trump"

Table 4. Positive tweet samples about "Trump"

Words	Sentences
Promise Thank Accomplish	Thank you President Trump for keeping all of your promises! It's really unbelievable what you have to go threw everyday and still accomplish everything you said you would!
Liberal	But it's perfectly fine to make up stuff about President Trump. It's liberal billionaires they love.
Accomplish	Trump's Tariffs Accomplished More in Two Days Than Congress in 20 Years.

다음은 "Galaxy Buds"를 키워드로 한 데이터에 대한 결과이다. Fig. 4는 부정 트윗에서 자주 등장한 단어로 워드클라우드를 그린 결과이다. Suck, shit과 같이 부정적인 단어도 존재하지만, stay, issue와 같이 단어 자체가 부정적인 성격을 지니지 않는 단어도 존재한다. 따라서 각 단어가 포함된 트윗이 실제로 부정적인 트윗인지 알아보기 위해 Table 5와 같이 빈도순으로 상위 10개의 부정단어를 포함한 트윗 중 사회 이슈를 포함하는 트윗들을 직접 감성 분석해보았다 . 갤럭시의 보상 프로그램에 불만을 가지기도 했으며, 품절되어 구매할 수 없다는 의견도 있었다. 특히, 갤럭시 버즈가 잘 작동하지 않는 일에 대해 불만을 가지는 목소리가 많았다. 분석 결과 단어 자체로는 부정적인 의미를 갖지 않는 단어라도, 이 단어를 포함하는 트윗들이 부정적인 성격을 띄는 경향이 있음을 알 수 있었다.

Fig. 5은 긍정 트윗에서 자주 등장한 단어로 워드클라우드를 그린 결과이고, Table 6는 빈도순으로 상위 10개의 긍정단어를 포함하는 트윗 중 사회 이슈를 포함하는 트윗을 보여준다. 갤럭시 버즈가 에어팟에 비해서 안드로이드 사용자에게 편리하다는 의견과 버즈를 사용하게 되어서 기쁘다는 의견이 포함되어 있다. 즉, 긍정 트윗에 자주 등장하는 단어를 포함하는 대부분의 트윗은 긍정적인 트윗임을 알 수 있었다.

갤럭시 버즈에 대해 부정적인 의견을 나타낸 트윗은 고장과 보상서비스에 대해 비판했으며, 반면 안드로이드 사용자들에게는 긍정적인 평가를 받았다. 따라서 갤럭시 측은 안드로이드 사용자들에게 특히 어필할 수 있는 홍보물을 제작하거나, 갤럭시 버즈의 고장원인과 보상서비스를 수정하여 사용자들의 불편함을 덜어야 할 필요가 있다.

Fig. 4. Frequent words in negative tweets about "Galaxy Buds"

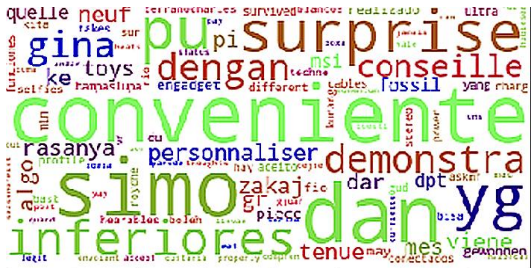

Fig. 5. Frequent words in positive tweets about "Galaxy Buds"

Table 6. Positive tweet samples about "Galaxy Buds"

Words	Sentences
Conveniente	Galaxy Buds, case gigante? Tem é um formato diferente. Ando com ela no bolso das calças na boa. Aceito que a da Apple, pelo formato, seja mais conveniente para andar no bolso
Surprise	Behold the most nerdy, awesome piece of jewelry I'll ever own. It was a bday surprise from Clayton, one of my bestest buds and favorite humans in the galaxy.
Mucha	Los airpods son muy buenos, pero claro, si tienes android pierdes muchas funciones... los galaxy buds no estan mal dicen
Simo	Hoy me he comprado los Galaxy Buds y son buenísimos, ahora para que sea profit, tengo que regresar a correr 3 veces por semana y aumentar masa; me hubiese comprado los AirPods 2 pero odio Apple

4.2 수치적 해석

Table 7과 Table 8은 “Trump”와 “Galaxy Buds”를 키워드로 추출된 트윗의 각 감성에 대한 좋아요(like) 평균 개수, 리트윗(retweet) 평균 개수를 나타낸 결과이다. Table 7은 트럼프의 계정에서 수집된 트윗도 포함되어있어 리트윗의 수가 매우 크다는 것을 알 수 있다. 그에 반해 Table 8은 좋아요 수, 리트윗 수가 40 내외로 크지 않다는 것을 알 수 있다. Table 7과 Table 8에서 공통으로 볼 수 있듯이, 긍정 트윗은 좋아요 수보다 리트윗 수가 높으며 부정 트윗은 리트윗 수보다 좋아요 수가 높다. 리트윗은 좋아요와 다르게 게시글을 자신의 뉴스피드에 게시하는 형식으로 이루어진다. 따라서 사람들은 부정 트윗에 대해 좋아요는 누르지만 자신의 뉴스피드에 게시하지 않는다는 성향이 있으며, 이 결과를 홍보 및 마케팅에 적용하기 위해서는 부정적인 트윗보다는 긍정적인 트윗을 게시함으로써 많은 사람들의 뉴스피드로 전달될 수 있도록 해야 한다.

Table 7. Average counts of like and retweet about “Trump”

	Like	Retweet
Positive	5.8	176
Negative	226.5	4.87

Table 8. Average counts of like and retweet about “Galaxy Buds”

	Like	Retweet
Positive	34.4	42
Negative	40.4	37.1

좋아요 수와 리트윗 수에 어떤 변수가 영향을 미치는지 알아본 결과는 다음과 같다. Table 9과 Table 10는 “Trump”와 “Galaxy Buds”를 키워드로 하는 트윗의 좋아요 수(like)와 리트윗 수(retweet), 트윗의 알파벳 개수(alphabet), 단어 개수(word), 특수문자 개수(special)의 상관관계를 분석한 결과이다. 먼저 트윗의 좋아요 수와 리트윗 수, 트윗의 알파벳 개수, 단어 개수, 특수문자 개수를 0에서 1 사이의 값으로 정규화하고, 각 변수들 간의 상관 정도를 -1에서 1 사이의 값으로 표현하는 피어슨 상관계수를 사용하여 상관관계를 분석하였다. 예를 들어, Table 9에서 트윗의 알파벳 개수(alphabet)와 리트윗 수(retweet)의 상관계수는 -0.949로 강한 음의 상관관계를 가지므로 트윗의 길이가 길어질수록 리트윗의 수는 적어짐을 나타내고, 단어 개수(word)과 트윗의 알파벳 개수(alphabet)의 상관계수는 0.95로 강한 양의 상관관계를 가지므로 단어의 개수가 많으면 트윗의 길이가 함께 길어짐을 나타낸다. 또한 상관계수가 0인 경우 아무런 상관관계를 갖지 않는다고 해석할 수 있다.

“Galaxy Buds”와 “Trump”을 키워드로 하는 트윗의 리트윗 수와 트윗의 알파벳 개수와의 상관관계는 각각 -0.949, - 0.800이다. 즉, 리트윗 수와 트윗 길이는 강한 음의 상관관계를 갖는다. 많은 사람들은 트윗의 길이가 짧은 경우 자신의 뉴스피드로 리트윗 하는 경향이 높다고 할 수 있다. 자신의 생각을 많이 표현하기 위해 트윗을 길게 쓰기 보다는 짧고 간단하게 표현하는 것이 많은 사람들에게 보일 수 있는 방법이다. 따라서 부정적인 의견보다는 긍정적인 의견을 함축하여 간단하게 표현한다면 큰 관심을 끌 수 있다고 볼 수 있다.

Table 9. Correlation analysis result about “Trump”

	Like	Retweet	Alphabet	Word	Special
Like	1.0000	-0.036	0.0020	0.0213	-0.051
Retweet	-0.004	1.0000	-0.949	-0.906	-0.429
Alphabet	0.0020	-0.949	1.0000	0.9500	0.4958
Word	0.0213	-0.9056	0.9500	1.0000	0.5124
Special	-0.051	-0.429	0.4958	0.5124	1.0000

Table 10. Correlation analysis result about “Galaxy Buds”

	Like	Retweet	Alphabet	Word	Special
Like	1.0000	0.0806	-0.479	-0.299	-0.061
Retweet	0.0806	1.0000	-0.800	-0.529	-0.035
Alphabet	-0.479	-0.800	1.0000	0.6479	0.0504
Word	-0.299	-0.529	0.6479	1.0000	0.0865
Special	-0.061	-0.035	0.0504	0.0865	1.0000

5. 결론

본 논문에서는 트위터에서 제공하는 공개 API를 통해 정치 및 최신 기술 분야에서 이슈가 되고 있는 두 가지 키워드 (Trump, Galaxy buds)를 포함한 트위터 데이터를 수집하여 사용자들의 의견을 식별하였다.

영화리뷰 데이터로 학습된 CNN 감성 분류기를 수집된 트위터 데이터에 적용하여 긍정 또는 부정으로 감성 분류를 수행하였다. 감성 별로 빈번하게 등장하는 단어를 워드클라우드를 통해 시각화하였고, 빈도가 높은 단어들을 포함하는 트윗 중 사회 이슈를 포함하는 트윗을 분석하였다. 즉, 긍정 트윗을 통해 추구해야하는 점을, 부정 트윗을 통해 개선해야 하는 점을 시사한다. 또한, 감성 별 트윗의 좋아요 수와 리트윗 수의 관계를 분석한 결과 긍정 트윗이 리트윗 수가 높으며, 부정 트윗이 좋아요 수가 높다는 결과를 보였다. 감성 외에도 좋아요, 리트윗 수에 영향을 미치는 요인을 알아보기 위해 트윗 길이, 특수문자 개수 등과 상관관계를 분석한 결과 트윗의 길이가 길수록 리트윗의 수가 적어진다는 결과를 얻었다. 또한 트럼프 정부 입장에서 국민들의 지지를 얻기 위해 어떤 태도를 취해야 하는지, 갤럭시 측에서 사용자들의 구매를 이끌기 위해 어떤 점을 개선해야 하는지에 대해 시사한다. 본 논문의 양적 분석 결과를 홍보 및 마케팅에 적용할 수 있는 방안은 다음과 같다. 좋아요에 비해 더 많은 사용자들에게 노출될 수 있는 리트윗의 수를 높게 하기 위해서는 긍정적인 의미를 담은 트윗을 짧고 간단하게 작성해야한다는 결론을 제공한다.

그러나 데이터의 수집 기간이 짧은 점과 단 두 가지의 키워드만을 분석했다는 점에서 제시된 결과가 일반화 될 수 없다는 한계점을 가지고 있다. 따라서 향후 연구로는 한 달 또는 그 이상의 기간 동안 쓰인 트윗을 데이터로 사용하며, 더 많은 키워드를 기준으로 데이터를 수집하여 폭넓은 연구를 수행할 예정이다.

REFERENCES

[1] C. W. Choi. (2011). Development Process and Major Cases of SNS. *Industrial Engineering Magazine, 18(1),* 20-23.

[2] Wikipedia Contributors. (2019, July 11). *Social networking service*. Retrieved July 16, 2019, from Wikipedia website: https://en.wikipedia.org/wiki/Social_networking_service

[3] H. Jung, J. Bae, S. Hong, C. Park, & M. Song. (2016). Analysis of Twitter Public Opinion in Different Political Views. *Korean Journal of Journalism and Communication Studies, 60(2),* 269-302.

[4] J. Lee & H. Park. (2016). Comparing Customer Reactions Before and After of a Smart Watch Release through Opinion Mining. *The Korean journal of bigdata, 1(1),* 1-7.

[5] H. J. Ahn. (2016). The Relationship between Personality and the Behavior of Using Twitter of Korean Users. *Journal of Korean Institute of Information Technology, 14(1),* 171-177.

[6] D. Hong, H. Jeong, S. Park, E. Han, H. Kim & I. Yun. (2017). Study on the Methodology for Extracting Information from SNS Using a Sentiment Analysis. *Journal of The Korea Institute of Intelligent Transport Systems, 16(6),* 141-155.

[7] J. Y. Gang, T. Y. Kim, J. W. Choi & H. J. Oh. (2016). A Study on the Vitalization Strategy Based on Current Status Analysis of National Archives. *Journal of the Korean society for information management, 33(3),* 263-285.

[8] J. Y. Gang, T. Y. Kim, J. W. Choi & H. J. Oh. (2016). A Study on Social Media Usage of Government Archival Services and Users' Interestedness: Focused on "National Archives of Korea" and "Presidential Archives". *Journal of the Korean society for information management, 33(2),* 135-156.

[9] R. D. Lee, J. M. Kim & J. S. Lim. (2016). Analysis of Twitter Topic using LDA. *The Journal of Korean Institute of Communications and Information Sciences,* 1010-1011.

[10] J. Hong & S. Choi. (2016). Consumers' Responses toward New Nike Product in Twitter Messages. *Journal Korea Society of Visual Design Forum, 50,* 73-84.

[11] J. H. Bae, J. E. Son & M. Song. (2013). Analysis of Twitter for 2012 South Korea Presidential Election by Text Mining Techniques. *Journal of Intelligence and Information Systems, 19(3),* 141-156.
DOI : 10.13088/jiis.2013.19.3.141

[12] C. H. Lee, J. Hur, H. J. Oh, H. J. Kim, P. M. Ryu & H. K. Kim. (2013). Technology trends of issue detection and predictive analysis on social big data. *Electronics and Telecommunications Trends, 28(1),* 62-71.

[13] M. Y. Chong. (2015). Selecting a key issue through association analysis of realtime search words. *Journal of Digital Convergence, 13(12),* 161-169.
DOI : 10.14400/jdc.2015.13.12.161

[14] Y. J. Ham, C. W. Ahn, K. H. Kim, G. B. Park, K. J. Kim, D. Y. Lee & S. M. Park. (2014). A Study on Policy Priorities for Implementing Big Data Analytics in the Social Security Sector : Adopting AHP Methodology. *Journal of Digital Convergence, 12(8),* 49-60.
DOI : 10.14400/jdc.2014.12.8.49

[15] M. Y. Chong. (2016). Extracting week key issues and analyzing differences from realtime search keywords of portal sites. *Journal of Digital Convergence, 14(12),* 237-243.
DOI : 10.14400/jdc.2016.14.12.237

[16] T. Y. Kim, Y. Kim & H. J. Oh. (2017). An Analysis of the Relationship between Public Opinion on Social Bigdata and Results after Implementation of Public Policies: A Case Study in "Welfare" Policy. *Journal of Digital Convergence, 15(3),* 17-25.
DOI : 10.14400/jdc.2017.15.3.17

[17] Google Code Archive - Long-term storage for Google Code Project Hosting. (2019). *word2vec.* Retrieved July 16, 2019, from Google.com website: https://code.google.com/p/word2vec/

[18] H. Abdi & L. Williams. (2010). Principal Component Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics, 2(4),* 433-459.

[19] S. Lai, L. Xiu, K. Liu & J. Zhao. (2015). Recurrent Convolutional Neural Networks for Text Classification. *In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (pp. 2267-2273). Austin, Texas : Association for the Advancement of Artificial Intelligence.

[20] Y. Kim. (2014). Convolutional Neural Networks for Sentence Classification. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1746-1751). Doha, Qatar : Association for Computational Linguistics.

[21] K. Nal, E. Grefenstette & P. Blunsom. (2014). A Convolutional Neural Network for Modelling Sentences. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 655-665). Baltimore, Maryland, USA : Association for Computational Linguistics.

[22] T. Lei, R. Barzilay & T. Jaakkola. (2015). Molding CNNs for text: non-linear, non-consecutive convolutions. *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1565-1575). Lisbon, Portugal : Association for Computational Linguistics.

[23] D. Tang, B. Qin & T. Liu. (2016). Document Modeling with Gated Recurrent Neural Network for Sentiment Classification. *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1422-1432). Lisbon, Portugal : Association for Computational Linguistics.

[24] Q. Qian, M. Huang, J. Lei & X. Zhu. (2016). Linguistically Regularized LSTM for Sentiment Classification. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (pp. 1679-1689). Vancouver, Canada. : Association for Computational Linguistics.

[25] J. Chung, C. Gulcehre, K. Cho & Y. Bengio. (2014). *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.* Retrieved from https://arxiv.org/pdf/1412.3555.pdf

[26] M. Huang, Y. Cao & C. Dong. (2016). *Modeling Rich Contexts for Sentiment Classification with LSTM.* Retrieved from https://arxiv.org/pdf/1605.01478.pdf

[27] Y. Zhang, Z. Zhang, D. Miao & J. Wang. (2019). Three-way enhanced convolutional neural networks for sentence-level sentiment classification. *Information Sciences, 477,* 55-64.
DOI : 10.1016/j.ins.2018.10.030

[28] H. Kim & Y. S. Jeong. (2019). Sentiment Classification Using Convolutional Neural Networks. *Applied Sciences, 9(11),* 2347.
DOI : 10.3390/app9112347

[29] M. K. Kwon & H. S. Yang. (2017). Performance Improvement of Object Recognition System in Broadcast Media Using Hierarchical CNN. *Journal of Digital Convergence, 15(3),* 201-209.
DOI : 10.14400/jdc.2017.15.3.201

[30] Tweepy. (2019). Retrieved July 16, 2019, from Tweepy.org website: https://www.tweepy.org/

[31] A. F. M. Agarap. (2019). *Deep Learning using Rectified Linear Units (ReLU).* Retrieved from https://arxiv.org/pdf/1803.08375.pdf

[32] D. P. Kingma. & J. L. Ba. (2015). *ADAM: A Method for Stochastic Optimization.* 3rd International Conference on Learning Representations (pp. 1-15), San Diego, CA, USA.

김 한 나(Hannah Kim) [학생회원]

· 2017년 8월 : 순천향대학교 수학과 (이학사)
· 2017년 9월 ~ 현재 : 순천향대학교 미래융합기술학과
· 관심분야 : 딥러닝, 자연어처리
· E-Mail : hannah@sch.ac.kr

정 영 섭(Young-Seob Jeong) [정회원]

· 2010년 2월 : 한양대학교 컴퓨터공학과(공학사)
· 2012년 2월 : 한국과학기술원 전산학과(공학석사)
· 2016년 2월 : 한국과학기술원 전산학과(공학박사)
· 2016년 2월 ~ 2016년 12월 : Naver
· 2017년 1월 ~ 현재 : 순천향대학교 빅데이터공학과 교수
· 관심분야 : 인공지능, 빅데이터
· E-Mail : bytecell@sch.ac.kr