「정보시스템연구」 제29권 제3호
한국정보시스템학회
2020년 9월, pp. 237~251

http://dx.doi.org/10.5859/KAIS.2020.29.3.237

SNS대상의 지능형 자연어 수집, 처리 시스템 구현을 통한 한국형 감성사전 구축에 관한 연구*

이 종 화**

〈목 차〉

Ⅰ. 서 론

공간, 소통의 도구, 언어 그리고, 직업은 서로 다르지만 공통의 관심사만 있다면 언제든 소통할 수 있는 세대를 z세대라 한다(강주연 등, 2020). z세대는 네트워크 환경에 익숙하여 자료나 정보, 콘텐츠의 저장(save) 개념이 없다. 즐기고 소통하는 모든 콘텐츠는 네트워크 환경의 클라우드(cloud) 공간에서 제공되며 대면하지 않은 언택트(untact)를 이미 그들은 즐기고 있기 때문이다. 온라인에서 관계를 맺어가는 z세대는 소셜 커뮤니티에서 팔로워, 구독자, 이웃 등의 친구를 찾고 관계를 유지하며 영상 및 화상 통화로 소통하는 것을 선호한다. 또한, 디지털 네이티브(digital native) 세대이며 네트워크 세상인 디지털 환경과 물리적 현실 세계의 관계를 구분하지 않고 자연스럽게 받아들이고 온라인 게임을 함께 즐기며 특정 콘텐츠 채널을 좋아하는 모습을 확인할 수 있다(고홍석 · 신중현, 2018). 이러한 z세대의 특징은 국내 네트워크 접속 환경인 5G와 모바일 보급의 팽창으로 전 세대로 확장되고 있다. 또한, 포스트코로나(post-corona)시대를 맞아 국내 기존 ICT 산업의 인프라가 K방역의 원동력이 되었으며 데이터 기반 언택트 서비스의 퍼스트 무버(first mover)로서 역할을 하고 있다.

2020년 과학기술정보통신부 업무계획에 따

* 이 논문 또는 저서는 2018년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임. (NRF-2018S1A5B5A07072061)
** 동의대학교 e비즈니스학과, jhlee6050@deu.ac.kr

르면 D.N.A 기반(Data, Network, AI) 디지털 선도 국가 전략을 계획하고 있다. 데이터 규제 빗장을 열어 2019년은 1,458종의 데이터를 개방하였으며 2020년은 3,094종으로 2배 이상의 데이터를 개방하여 새로운 서비스의 원유를 제공할 예정이다. 데이터 기반의 인공지능(AI) 관련 펀드 조성에 3,000억 원을 조성하여 차세대 핵심기술과 인재 양성의 초석을 마련한다는 계획이다(과학기술정보통신부, 2020). 이렇듯 정부, 기업, 모든 산업에서 데이터에 중심을 두고 있으며 미래 산업의 새로운 비즈니스 모델을 제시할 중요한 원유 역할을 하고 있다. 3차 산업혁명은 물리적인 제품을 소프트웨어로 바꿈으로써 콘텐츠의 개념을 만들었다. 아날로그 시대 음반은 카세트테이프에서 디지털 시대로 넘어오면서 코드화 되어 CD나 DVD 등의 저장매체로 소유하게 되었다. 그리고 온라인 시대에 공유 가능한 형태의 제품이나 서비스를 탄생시켰다. 즉, 아날로그, 디지털, 온라인 과정을 거쳐 제품이나 서비스가 진화되며 파괴적 혁신(disruptive innovation)의 굴레를 이루고 있었다. 하지만, 데이터 기반의 새로운 산업의 물결이 시작되면 기존 방식과는 다르게 역으로 서비스가 생성되고 있다. 방대한 데이터를 활용하여 패턴을 찾고 새로운 가치를 창출하므로 물리적인 제품이나 콘텐츠 서비스를 만들어 내고 있다. 또한, 먼저 경험한 고객들의 댓글, 후기 등은 제품이나 서비스를 리뉴얼하게 만드는 동력으로 작용하여 네트워크내의 상호작용의 가치가 고객의 협상력을 높이고 있다. 이렇듯 디지털 기술이 기업과 사회 구조를 변화시키는 디지털 트랜스포메이션(digital transformation)이 진행되면서 고객과의 상호작용이 일어날 것이다(권종원, 송태승, 2019; 이진수, 2017).

코로나19 이후 비대면 환경에서 소통의 근본 목적인 상대방과의 의미 이해 과정이 더욱 정교하게 요구되고 있다. 모바일을 활용한 서비스 플랫폼 시장이 팽창되면서 z세대 소통은 다양한 미디어로 표현하고 있다. 비대면 환경은 비정형 데이터의 발생이 팽창되면서 감정 분석의 의미가 커지고 있다. ‘행복’의 감정 세트 내에서도 감정 시드는 표현 언어에 따라 다양한 단어들의 감정시드로 구성되어 있다. 하지만 ‘행복’의 감정을 표현하는 키워드 속에서 감정의 깊이인 가중치를 추출한다면 비정형 데이터 분석이 정교한 정형 데이터로 변환 될 것이다. 이러한 발상으로 본 연구는 이전 연구의 결과를 배경으로 진행되었다. Plutchick(1980)의 감정 세트를 기준으로 감정 단어들의 가중치를 추출하여 한글로 재해석하고 SNS상의 함축적 의미가 있는 감정 단어를 활용하고자 한다. 다차원 감성 사전 구축과 입체적 감성 분류를 위하여 1) 7가지 감성 세트 구성을 위한 감성 시드(Seed)를 선정하고, 2) 각 감성 시드에서 파생된 SNS 해시태그를 수집하여 감성 단어사전을 구축, 3) 같은 감성 내 단어들의 우선순위를 구분하기 위하여 TF-IDF(Term Frequency Inverse Document Frequency)를 이용한 가중치 추출, 4) NMF(Non-negative Matrix Factorization) 알고리즘을 통한 감성 단어의 특성치를 이용한 감성 차원 축소를 적용하고자한다. 또한, 벡터 간의 거리 측정은 특정 지점에서 두 벡터 간 벌어진 정도인 각도를 활용하여 유사도 측정을 진행하는 코사인 거리(Consine distance)로 감정 단어 지수를 개발하고자 한다.

본 논문의 2장은 TF-IDF분석과 NMF 그리

고, 코사인 유사도에 관하여 선행 연구를 진행하였다. 3장에서는 감정 단어 간 유사도 측정을 위한 프레임워크, 4장은 실현을 통한 7가지 감정 단어의 가중치를 공유하였으며 마지막 장에는 연구의 결론과 시사점 및 향후 연구를 제시하고자 한다.

Ⅱ. 선행 연구

2.1 TF-IDF(term frequency-inverse document frequency)

미디어의 발달로 뉴스를 접하는 방법도 다양해지고 있다. 신문, 잡지, TV 등 여러 매체가 많지만 바쁜 현대인들이 가장 빠르게 이용할 수 있는 것이 인터넷기사이다. 하지만 매일 쏟아지는 기사의 홍수 속에서 내용이 긴 기사를 끝까지 다 읽는 다는 것이 쉽지만은 않은 일이다. 의미를 전달하기 위해서는 무수히 많은 단어들로 그 문장을 기술한다. 문서 내 특정 단어의 빈도가 높거나 다수 문서 사이에서 자주 등장하는 높은 빈도의 단어 등 문서의 의미를 분석하기 위해서 단어 빈도 또한 중요한 이슈이다. TF-IDF는 TF와 IDF를 곱한 값으로 정의되어 있으며 d-문서, t-단어, 문서 n개의 개수를 이용하여 TF, DF, IDF, TF_IDF를 정의하면 다음과 같다.

TF(Term Frequency)는 문서 내 특정 단어의 등장 빈도를 의미하며 문장을 단어로 나누고, 전체 단어 빈도에 따라 특정 단어가 얼마나 사용되었는지를 파악하여, (식1)를 이용하여 해당 문서의 성질을 파악하는 지표로 사용된다 (Qaiser and Ali, 2018; Salton and Buckley, 1988). 즉, 문서별로 해당 단어가 얼마나 등장하는지에 대한 빈도가 된다. 문서 내 단어가 한 번 이라도 있으면 1, 아니면 0으로 표현하는 불린 빈도를 적용하는 방법과 단어의 빈도 조절을 위한 로그 스케일 빈도로 구분한다.

$$tf(t,d) = \log(f(t,d)+1) \text{(식1)}$$

DF(Document Frequency)는 특정 단어가 나타나는 문서 수를 나타내는 가중치이다. 단어가 주체인 TF와는 다르게 문서가 주체가 되는 방법이며, (식2)를 이용하여 연구 문서에 해당 단어의 빈도를 확인할 수 있다(Salton and Buckley, 1988). 즉, 문서 내 다수의 같은 단어의 빈도를 보이더라도 DF는 1로 처리된다.

$$df(t,d) = \frac{|d \in D : t \in d|}{D} \quad \text{(식2)}$$
$$= \frac{\text{단어} t\text{가 포함된 문서의 수}}{\text{전체문서의 수}}$$

IDF(Inverse Document Frequency)는 특정 단어 t가 모든 문서에 등장하는 흔한 단어라면 가중치를 낮추어주어야 한다(이종화 등, 2019). 즉, DF값이 클수록 빈도의 가중치에 영향을 줄 수 있으므로 DF값의 역수를 취하여 제한하는 기법을 갖고 있다. 역수를 취하게 되면 전체 문서의 수가 많아질수록 IDF의 값이 기하급수적으로 커지게 되므로, IDF 또한 로그를 취하하여 (식3)과 같이 표현한다(Salton and Buckley, 1988).

$$idf(t,D) = \log\frac{|D|}{|d \in D : t \in d|} \quad \text{(식3)}$$
$$= \log\frac{\text{전체 문서의 수}}{\text{단어 } t\text{가 포함된 문서의 수}}$$

TF-IDF는 정보검색이나 문서분류에 가중치를 구하는 알고리즘이며 문서 간의 비슷한 정도나, 문서 내 어떤 단어가 중요한지의 척도를 계산하거나 문서 내 핵심 단어를 추출하기 위한 척도가 되거나 검색엔진에서 검색결과의 순위를 결정에 활용된다(이종화 등, 2019; Qaiser and Ali, 2018; Christian et al., 2016; Salton and Buckley, 1988).

$$tf-idf(t,d,D) = tf(t,d) \times idf(t,D) \text{(식4)}$$

(식4)는 TF와 IDF를 곱한 것으로 특정 문서 내에서 단어 빈도가 높거나, 전체 문서들엔 그 단어를 포함한 문서가 적을수록 TF-IDF 가중치가 높아진다. 따라서 TF-IDF를 이용하여 연구 문서에 나타나는 흔히 사용되는 단어를 제거하거나, 이슈 단어가 얼마나 중요한 가를 측정할 때에도 사용된다.

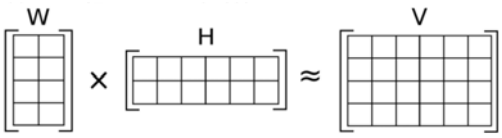
2.2 NMF(Non-negative Matrix Factorization)

머신비전산업에서 인공지능 기술인 머신러닝, 딥러닝이 빠르게 확산되고 있다. 인공지능 자동화 기술을 통하여 사물 인식의 비전 기술(vision technology)의 상용화가 가능해질 뿐만 아니라 수 많은 데이터의 학습 기술로 빠르고 정확하며 신뢰성과 유연성을 갖춘 비전 인식 기술을 가능하게 했다(Gunasekaran, 1996). 이런 기계학습은 데이터 마이닝(data mining) 또는 패턴 인식(pattern recognition)이라고도 불린다(Cheng et al., 2011). 기계학습은 크게 지도학습(supervised learning)과 비지도학습(unsupervised learning)으로 분류를 하고 있다(Written et al., 2016; Balahur et al., 2012; Walker et al., 2012). 하지만 경우에 따라 강화학습(reinforcement learning)을 추가하여 3가지로 분류를 하는 사람들도 있다(Ye et al., 2003).

지도학습(supervised learning)이란 어떤 입력에 대해서 어떤 결과가 나와야 하는지 사전 지식을 갖고 있는 경우에 해당 입력에 대해 특정 출력이 나오도록 하는 규칙을 찾아내며 회귀 방법이 여기에 해당한다(Balahur et al., 2012; Walker et al., 2012). 비지도학습(unsupervised learning)이란 입력은 있지만 정해진 출력이 없는 경우를 말하며, 순수하게 데이터들이 갖고 있는 속성들을 이용해 그룹으로 나누는 경우를 말하며, 지도학습이 회귀 방법을 사용하는 것과 달리, 분류(clustering)에 해당이 된다(Written et al., 2016). 가령, 인구 통계학적인 기본적인 자료인 성별, 나이, 학력, 출생 지역, 기타 등의 데이터 조합을 이용하여 어떤 정당이나 단체를 지지하는지 살펴보는 것도 해당된다. 일반적으로는 컴퓨터 클러스터링, 그림이나 동영상에 대한 자동 인식(auto tagging), 시장 조사 등에 사용이 된다. 강화학습(reinforcement learning)은 로봇의 학습 등에 사용할 수 있으며, 자신과 환경과의 상호 관계에 따라 자신의 행동을 개선해 나가는 학습법을 말한다(Kaelbling et al., 1996).

NMF는 비지도 학습에 속하며 행렬 차원 감소를 이용하기 때문에 비음수 행렬 인수분해라

한다. 본 연구의 목적에 맞게 문서 분류 분석이나 컴퓨터 시각 처리에 널리 사용된다(박선 등, 2013; 김철원 · 박선, 2011; Seung and Lee, 2001). 주성분분석(PCA)이나 산점도를 이용하는 t-SNE는 원본 데이터 특징을 추출하여 새로운 특징 셋으로 표현을 하지만 새로운 특징 셋이 원본 특징과 어떤 연관 관계를 가지는지는 해석이 불가능하다. 하지만 NMF의 경우 새로운 특징 셋이 어떻게 원본 데이터와 관계를 가지는지 확인이 가능하여 본 연구에 적용하고자 한다(Seung and Lee, 2001). Seung and Lee(2001)의 모델인 NMF는 행렬 인수 분해 알고리즘을 이용하였다. 원본 행렬 V를 인수 분해를 통하여두개의 행렬로 분리하고자 한다. <그림 1>은 원본 행렬 V = W * H 로 분리한 예이다.

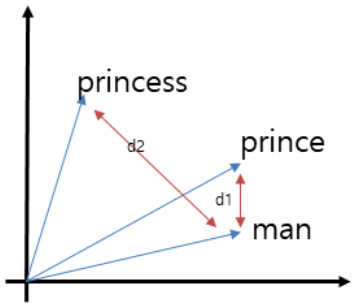


<그림 1> NMF의 행렬 인수 분해 구조

W는 가중치 행렬로 특정 카테고리의 특성과의 관계를 나타내며 H는 특성 행렬로 기존 특성에 대비하여 새로운 특성에 대한 관계를 도출되는 행렬로 활용된다.

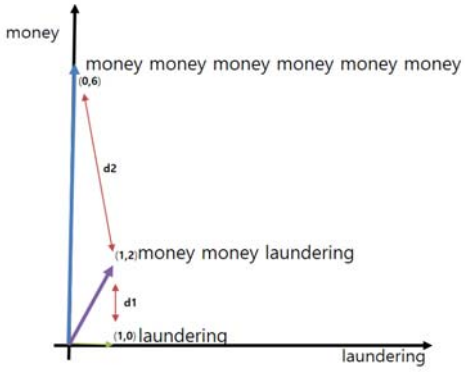
2.3 Cosine Similarity

유사도 측정은 다양하지만 자연어 처리에 많이 사용되는 유클리드 거리(Euclidean distance)와 코사인 유사도(Cosine Similarity)를 살펴본다(Ye, 2015; Danielsson, 1980).



<그림 2> 유클리드 측정 구조

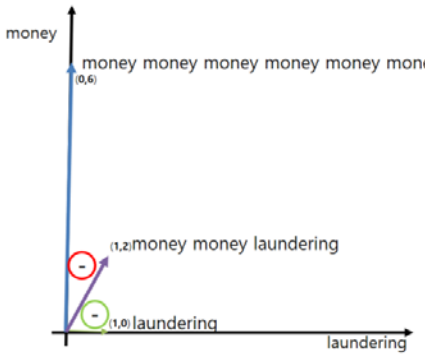
유클리드 거리는 두 단어의 벡터 크기를 반영하여 두 점과의 거리를 피타고라스의 정리에 의하여 거리를 측정한다(Danielsson, 1980). <그림 2>는 man과 prince 거리인 d1이 d2보다 가까운 것을 확인할 수 있으면 man과 prince가 유사도가 더 높다고 볼 수 있다.



<그림 3> 빈도를 반영한 유클리드 모형

<그림 3>의 예를 보면 d1과 d2의 거리를 비교해보면 d1이 유사도가 더 높게 측정되는 것을 확인할 수 있다. 이는 벡터의 크기(magnituder)를 반영하여 money의 빈도가 높을수록 거리는 점점 더 멀어지므로 유사도가 더 떨어지는 것을 확인 할 수 있었다.

자연어 처리에서 자주 등장하는 단어인 경우 유클리드 거리를 이용한 유사도 측정의 한계를 처리하기 위하여 코사인 유사도를 이용한다.



<그림 4> 코사인 유사도 구조

<그림 4>는 특성 값을 나타내는 벡터 A와 B가 있을 때, 두 벡터의 각도가 같은 선에 가까울수록 즉, 작은 각도일수록 두 개의 특성이 유사하다고 판단하기로 한다. 즉 A와 B의 각도 θ가 최소일 수 록 값이 유사하다고 판단하면 된다. 벡터의 크기를 무시 하여 (식4)와 같이 코사인 유사도를 측정한다(Ye, 2015; Tata and Patel, 2007).

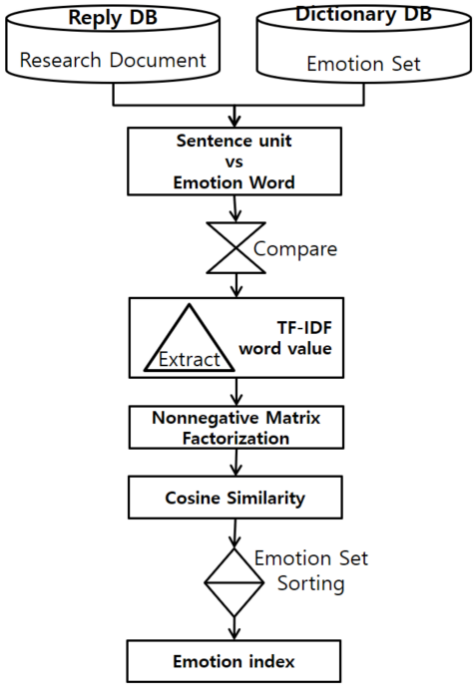
$$\cos(\theta) = \frac{A \cdot B}{\| A \| \| B \|} \quad \text{(식4)}$$

본 연구의 감정 지수는 TF-IDF 과정을 통하여 행렬로 변환하여 NMF 알고리즘을 활용하여 각 문서의 특성을 추출한 후 베겉 거리 측정 방법으로 코사인 거리 알고리즘으로 감정 세트 내 감정 단어의 유사도를 결정 하고자 한다.

Ⅲ. 연구방법과 프레임워크

본 연구는 Plutchick(1980)의 감정 세트를 기준으로 감정 단어들의 가중치를 추출하여 한글 감정 세트 내의 감정 랭킹을 추출하며 감정 분석(sentiment analysis)의 신뢰도를 높이고자 한다.

본 연구자의 이전 연구인 “감성 단어 일반화 연구”, “감정 단어 실시간 수집 시스템”의 이전 연구 결과인 7가지 감정 세트 내 감정 시드 단어를 본 연구에 반영되었다(이종화 등, 2018; 이종화, 2018).



<그림 5> 본 연구의 프레임워크

<그림 5>는 본 연구를 위한 프레임워크이다. TF-IDF 방식으로 문장 내 감정 단어를 벡터

를 추출하여 NMF를 이용하여 각 문장과 감정 단어의 행렬을 인수분해 과정의 차원 축소 방식으로 감정 단어의 특성을 추출하였으며 추출된 벡터간 거리는 빈도를 제한을 고려한 코사인 유사도를 통하여 감정 세트 내 감정 단어의 가중치를 추출하고 한다.

본 연구는 데이터 분석에 앞서 연구 재료가 되는 뉴스 기사는 본 연구자의 선행연구로 수집 시스템이 구축되어 있다.

“Python을 이용한 SNS 크롤링 시스템 구축”을 이용하여 매일 새롭게 등장되는 뉴스 콘텐츠에 대해 명사, 형용사 등 네티즌들의 감성을 전처리 과정을 거쳐 DB에 저장한다(이종화, 2018). 이종화(2018) 연구는 파이썬 웹드라이버의 가상 웹브라우저를 이용하여 인스타그램, 트위터, 유튜브, 인터넷 기사와 댓글 등 소셜 데이터를 실시간 수집 가능한 시스템을 연구하였다. 뉴스 기사가 인터넷을 통하여 업로드가 완료되면 1차 전처리 과정과 비표준어 사전을 활용한 댓글 전처리 과정을 거쳐 함께 DB에 저장하였다.

“감성 단어 일반화 연구”의 이전 연구는 일곱 가지 감정 세트 내 단어의 일반화 연구를 진행하여 190여개의 시드 단어를 발굴하였다.

이러한 이전 연구 결과를 활용하여 문장을 구성하는 단어의 벡터 과정을 거쳐 행렬 변환하는 과정에서는 TF-IDF를 이용한 가중치 추출하며, 감정 Set 내 행렬의 차원 감소 기법은 NMF 알고리즘을 통한 감정 단어의 특성치를 이용한 감성 차원을 해결하고자 한다. 감정 Set 내 감정 단어의 유사도를 측정하여 벡터간의 거리 측정 방법으로 Cosine Distance 알고리즘을 이용하여 감정 Set 내 감정 단어의 순위를 개발하고 네티즌의 감정의 깊이를 읽을 수 있는 특별한 기준으로 본다.

Ⅳ. 연구실험과 결과

본 연구는 네티즌의 의견 속에서 감정어의 우선순위를 결정하고자 한다. 모든 문장에 빈번히 나타난 키워드를 제한하여 의미 있는 감정 단어를 부각하기 위한 TF_IDF를 사용하였다. 각 문장의 특성값을 추출하기 위한 방법으로 NMF의 차원 축소를 진행하였으며 축소된 차원의 벡터사이 위치 값을 각도로 환산한 코사인으로 적용하여 빈도의 차이에 영향을 최소화하였다.

<표 1> 감정 세트 실험 데이터 및 시드, 특성

Emotion	Reply	Generalization	Features
Angry	44,799	56	55
Bad	20,020	8	7
Disgusted	17,434	9	8
Fearful	27,462	13	12
Happy	50,029	50	49
Sad	31,459	43	42
Surprised	10,280	11	10

<표 1>은 감정 세트별 감정 단어는 이종화(2018)의 “SNS 해시태그를 이용한 감정 단어 일반화 연구” 결과인 190여 개의 감정 단어 시드의 통계이며 <표 2>의 나열은 감정 단어 리스트이다.

본 실험에서는 20만 개의 댓글을 분석하여 결과를 도출되었으며 차원 축소를 위한 NMF

의 특성(<표 1>의 features)은 분석 데이터와 해당 감정 단어의 빈도를 적용하였다. <그림 1>의 V = W(n*r) * H(r*m)의 행렬에서 (n + m)r < n*m의 조건을 만족하여 최대한 특성을 도출하여 본 연구에 적용하였다.

<표 2> 해시태그를 이용한 감정 일반화

angry		happy	
굴욕	0.977	무료	0.968
적대	0.976	즐거운	0.967
분노	0.975	친한	0.965
복수	0.975	믿는	0.965
억울	0.975	자랑	0.964
실망	0.975	호기심	0.964
부담	0.975	재미	0.963
핏대	0.974	화사	0.962
비난	0.974	희망	0.961
통제	0.974	자신감	0.960
난폭	0.974	자유로운	0.956
보복	0.974	안전	0.955
씁쓸	0.974	반가운	0.954
더러운	0.973	평화로운	0.950
반항	0.972	흥분	0.949
비열	0.972	만족	0.946
귀찮	0.971	흥미	0.945
피하고	0.971	기대	0.944
원망	0.970	행복	0.944
무시	0.969	귀중	0.944
질투	0.969	따사로운	0.943
끔찍	0.968	사랑	0.943
불안	0.968	용감	0.942
혐오	0.967	감사	0.941
분한	0.967	날아갈	0.935
욕하는	0.967	신난	0.934
배신	0.966	기분좋	0.931
불편	0.965	흐뭇	0.928
성급	0.965	고마운	0.925
시끄	0.964	황홀	0.924
비판	0.963	발랄	0.924
불만	0.963	더없이	0.922
모욕	0.962	아늑	0.920
고함	0.962	좋은	0.918
구역질	0.959	벅찬	0.918
지겨운	0.959	온화	0.916
미운	0.958	기쁜	0.913
좌절	0.957	인정	0.913
얄미운	0.956	시원한	0.911
심술	0.955	후련	0.911
미칠	0.955	상쾌	0.908
짜증	0.943	정다운	0.901
무례	0.941	잘만드는	0.899
지루	0.937	포근	0.896
반대	0.926	느긋	0.895
혼난	0.916	화창	0.893
열받	0.915	살맛나는	0.882
성질	0.910	상큼	0.874
따분	0.909	끝내주는	0.870
성난	0.908	짜릿	0.853
속상	0.904	sad	
화난	0.898	후회	0.967
언짢	0.896	취약	0.966
불쾌	0.893	상처	0.965
공격적	0.889	신음	0.965
못마땅	0.886	충격	0.965
bad		위축	0.964
바쁜	0.841	서운	0.964
스트레스	0.823	고립	0.964
통제	0.797	단절	0.964
무관심	0.795	고독	0.964
나쁜	0.718	하위	0.963
지루	0.713	허탈	0.961
피곤	0.672	좌절	0.961
졸린	0.672	안타까운	0.961
disgusted		외로	0.960
소름	0.857	허전	0.959
실망	0.836	절망	0.959
무서운	0.832	민감	0.959
불쾌	0.822	버림받	0.958
혐오	0.795	불행	0.957
구역질	0.795	불만족	0.955

비판	0.785	처량	0.953
망설	0.752	저지른	0.950
어리둥	0.688	슬픈	0.947
fearful		고통	0.935
약한	0.892	더러운	0.935
학대	0.877	적적	0.933
불안	0.775	뭉클	0.932
협박	0.670	죽고싶	0.929
걱정	0.606	부끄	0.924
거절	0.529	비참	0.921
의미없	0.495	낭패	0.920
제외	0.462	애절	0.920
거부	0.436	시무룩	0.917
신경질	0.429	쓸쓸	0.913
위협	0.420	무기력	0.909
놀란	0.394	애틋	0.894
두려운	0.380	자포자기	0.891
surprised		눈물겨운	0.888
흥분	0.890	애처로운	0.886
깜짝	0.878	허약	0.866
혼란	0.852	공허	0.864
두려운	0.842	우울한	0.845
당황	0.833	끝	
놀란	0.827	열망	0.749
갈망	0.785	충격	0.718
활기	0.764	난처	0.679

본 연구는 심리학자 Plutchick(1980)의 일곱 가지 감정 세트를 한글로 변환하여 적용하였다. 감정 세트 내 감정 단어는 본 연구자의 감정 단어 일반화 연구의 결과를 이용하여 선정되었으며 차원 축소를 통한 코사인 유사도를 이용하여 가중치를 행렬로 구성하였다. ‘Angry'는 일곱 가지 감정 세트 중 가장 많은 56개 시드를 갖고 있으며 4만 4천여 개의 댓글과 55개의 특성을 고려하여 차원을 축소하였다. 유사도 결과 ’굴욕‘, ‘적대’, ‘분노’, ‘복수’, ‘억울’, ‘실망’, ‘부담’등의 감정 단어순으로 결과가 나타났으며 상대적으로 ‘성난’, ‘속상’, ‘화난’, ‘언짢’, ‘불쾌’, ‘공격적’, ‘못마땅’ 등의 감정 단어는 낮은 유사도를 보이고 있다.

‘Bad’는 일곱 가지 감정 세트 중 가장 적은 8개 시드를 갖고 있으며 2만 여개의 댓글과 7개의 특성을 고려하여 차원을 축소하였고 유사도 결과 ‘바쁜’, ‘스트레스’, ‘통제’, ‘무관심’, ‘나쁜’, ‘지루’, ‘피곤’, ‘졸린’순으로 감정 세트 내 감정 순위를 보이고 있다.

‘Disgusted’는 만 7천여 개의 댓글을 활용하여 감정 단어 순서를 결정하였고 9개의 시드를 이용하여 8개의 특성을 고려하여 감정 내 유사도를 측정하였다. ‘소름’, ‘실망’, ‘무서운’, ‘불쾌’, ‘혐오’, ‘구역질’, ‘비판’, ‘망설’, ‘어리둥’의 순으로 감정 랭킹이 나타났다.

‘Fearful’는 2만 7천여 개의 댓글을 이용하여 ‘약한’, ‘학대’, ‘불안’, ‘협박’, ‘걱정’, ‘거절’, ‘의미없’, ‘제외’, ‘거부’, ‘신경질’, ‘위협’, ‘놀란’, ‘두려운’의 순으로 우선순위가 결정되었으며 ‘약한’과 ‘놀란’의 거리차이가 같은 간정 내에서 가장 먼 거리차이를 보인다.

‘Happy’는 5만 여개의 가장 많은 댓글로 분석 되었으며 ‘무료’, ‘즐거운’, ‘친한’, ‘믿는’, ‘자랑’, ‘호기심’의 감정 단어가 높은 유사도를 보이며 상대적으로 ‘살맛나는’, ‘상큼’, ‘끝내주는’, ‘짜릿’의 감정 단어는 낮은 유사도를 보이고 있다. 하지만 ‘Happy’ 감정 세트 내에서 코사인 거리는 0.1의 차이로 감정 단어 간 거리는 조밀한 것으로 나타났다.

‘Sad’는 3만 천여 개 댓글을 43개의 시드로 분석 되었으며 ‘후회’, ‘취약’, ‘상처’, ‘신음’, ‘충격’이 감정 내 높은 유사도를 보였으며 ‘애처로운’, ‘허약’, ‘공허’, ‘우울한’은 낮은 유사

도를 보였다.

마지막 감정 세트인 ‘Surprised’는 만 여개의 댓글에서 실험을 준비하였으며 ‘흥분’, ‘깜짝’, ‘혼란’, ‘두려운’, ‘당황’, ‘놀란’, ‘갈망’, ‘활기’, ‘열망’, ‘충격’, ‘난처’의 순으로 감정 단어의 순위를 결정할 수 있었다.

Ⅴ. 결론 및 향후 연구과제

전 세계적으로 한 네트워크 속에서 데이터 전쟁을 하고 있다. worldometers에 따르면 인터넷 사용자(1일 46억 명 이상), 구글 검색(1일 52억 건 이상), 트윗(1일 5억 6천만 건 이상) 그리고, 매일 매일 발행되는 3억 4천 건 이상의 신문의 통계를 볼 수 있다(worldometers.info, 2020. 08.). 빅데이터 시대에 방대한 데이터가 있을수록 패턴을 찾을 수 있는 확률은 높아진다. 하지만 원하는 것만을 빨리 찾아주는 큐레이션(curation)이 없다면 빅데이터의 가치를 느끼진 못할 것이다. 본 연구는 방대한 댓글 속에서 감정을 찾아주는 큐레이션으로 감정 시드를 본 연구자의 기존 연구를 확장하여 감정 단어에 감정 색인(index)을 만들고자 하였다. 감정의 지표는 그 시대의 표현 방법에 따른 변화이다. 동시에 감정 단어의 변화를 얼마나 빠르게 반영하느냐에 따라 분석 결과의 신뢰성도 좌우될 것이다.

비정형적 데이터 분석은 신뢰성 있는 데이터 확보와 관리 측면에서 본 연구가 시사점은 다음과 같다. 데이터 전처리 과정에 TF-IDF 알고리즘을 활용하므로 전체 문서에 높은 빈도의 단어는 특징에서 배제하고, 문서에 특정 단어가 많이 언급될 수 록 그 단어의 가중치 값을 크게 하여 문서의 특징을 찾고자하였다. 또한, t-SNE나 PCA는 특성을 해석하기에 불명확함의 한계점을 있는 것을 감안하여 본 연구는 NMF 알고리즘을 이용하여 특성 해석이 가능하며 행렬로 압축되어 있는 특성을 기반으로 유사도를 측정하였다. 자연어 처리에서 자주 등장하는 단어인 경우 유클리드 거리를 이용한 유사도 측정의 한계를 처리하기 위하여 코사인 유사도를 적용하여 감성 단어 간 거리에 적용하였다.

본 연구에 앞서 선행된 연구들에서 데이터를 활용하였다. ‘Repay DB’는 실시간으로 수집되고 있는 인터넷 뉴스 기사의 댓글 수집 시스템을 활용하였다. 본 연구자의 기존 연구 결과물인 “Python을 이용한 SNS 크롤링 시스템 구축”을 이용하여 연구 자료가 실시간으로 수집되고 있다. ‘Dictionary DB’는 연구 대상에서 감정어만을 추출하기 위하여 “SNS 해시태그를 이용한 감정 단어 일반화 연구”의 결과물을 이용하여 연구의 연결성을 확보하였다.

분석은 Linux CentOS 6.x 서버 환경에서 오픈 소스 소프트웨어인 Linux R를 이용하여 진행되었으며 R 패키지를 활용하여 연구를 진행하였다. 연구 문서와 감정 사전을 비교한 결과를 TF-IDF의 가중치를 활용하여 연구 설계를 하였다. 특히, 차원 감소를 위한 NMF기법을 활용하여 가중치 행렬과 특성 행렬로 나누어 추출된 특성 행렬로 해당 특징이 어떤 감정 단어들에 의해서 반응하는지를 벡터 처리하여 그 거리를 각도 측정을 통하여 그 유사도를 측정하였다.

본 연구를 위하여 감정 세트 내 일반화된 감정 단어 190개를 이용하여 20만 개의 댓글에서

감정 지수의 순위를 결정하였다. 첫 번째 과정은 높은 빈도의 단어가 중요 단어에 영향을 줄 수 있으므로 TF-IDF 알고리즘을 통하여 1차 적으로 감정 단어의 벡터 변환을 진행하였다. 두 번째 과정은 문서의 특징을 추출하기 위하여 행렬 차원을 축소하여 감정 단어에 대비한 새로운 특성에 대한 관계를 NMF 알고리즘을 이용하여 측정하였다. 이러한 측정치를 이용하여 세 번째는 감정 단어가 거리를 직접적인 유클리드 거리를 사용한 것이 아니라 두 벡터 간 떨어진 각을 이용하여 거리를 코사인 유사도로 측정하였다. 이러한 과정을 통하여 감정 단어 지수를 도출하였다.

일곱 가지 감정 세트 중 감정 단어 내 유사도 차이를 가장 뚜렷한 세트는 'Fearful'이며 감정 키워드 간 유사도 차이가 미미한 감정 세트는 'Angry' 감정으로 나타났다. 'Fearful'를 대표하는 키워드들은 '약한', '학대', '불안' 등이 있다. 'Surprised' 세트는 '흥분', '깜짝', '혼란'의 단어가 대표하는 키워드로 나타났으며, 'Disgusted' 세트는 '소름', '실망', '무서운', '불쾌'등이 대표 키워드로 상위에 있었다. 'Bad' 세트는 '바쁜', '스트레스'의 단어가 대표적이며, 'Sad' 세트는 '후회', '취약', '상처', '신음', '충격' 등의 감정어가 감정을 이끄는 것으로 나타났다. 'Happy' 세트는 '무료', '즐거운', '친한', '믿는' 등의 키워드가 감정을 이끌었다. 마지막으로, 'Angry' 세트는 '굴욕', '적대', '분노', '복수', '억울', '실망', '부담' 등이 감정을 주도하고 있었다.

본 연구의 감정 단어의 순위결정은 빈도 중심 연구를 벗어나 특성이 있는 단어에 비중을 두고 연구를 한다면 고객의 니즈 해석에 신뢰도가 높아질 것으로 본다. 하지만 변화하는 고객의 니즈분석을 통하여 지속적인 감정 단어의 관리가 필요하다. 물론, 본 연구의 단순 실험으로 감정 지수를 결정하는 것은 속단이다. 특정 뉴스 댓글은 언론사의 뉴스 논조에 따라 감정을 표현하는 네티즌의 의견이 다양하게 댓글로 표현되기 때문이다. 제각각 다른 네티즌의 모든 감정을 반영해야하는 감성 지수는 그 실험 횟수와 기간은 상당할 것으로 본다. 네트워크의 지속적 변화는 새로운 연구 환경을 제공한다고 볼 수 있다. 하지만 시시각각 증가하는 고객들의 반응과 후기, 평가 등을 시각화하여 다양한 정보를 한 눈에 볼 수 있도록 사용자 인터페이스를 개발한다면 실시간으로 고객의 니즈를 판단할 수 있는 대시보드가 될 것이다.

특히, 단어를 숫자인 벡터로 변환하거나 문장 내 감정 단어의 연관성을 수치로 표현 할 때, 단어 벡터의 차원을 작은 차원으로 축소할 때 즉, 다양한 워드 임베딩(word embedding)기법을 활용하여 순위 결정을 해야 할 것이다.

참고문헌

강주연, 이이든, 김지수, "텍스트 마이닝을 활용한 'Z 세대'관련 뉴스데이터 의미연결망 분석," 미래청소년학회지, 제17권, 2020, pp. 25-48.

고홍석, 신중현, "디지털 네이티브 세대의 미디어 이용행태에 관한 탐색적 연구," 한국콘텐츠학회논문지, 제18권, 제3호, 2018, pp. 1-10.

권종원, 송태승, "제조 혁신 위한 플랫폼 기반의

디지털 트랜스포메이션 추진 동향," 전자공학회지, 제46권, 제12호, 2019, pp. 34-46.

김철원, 박선, "의미특징과 워드넷 기반의 의사연관 피드백을 사용한 질의기반 문서요약," 한국정보통신학회논문지 제15권, 제7호, 2011, pp. 1517-1524.

박선, 김경준, 김경호, 이성로 "의미특징 기반의 용어 가중치 재산정을 이용한 문서군집의 성능 향상," 한국정보통신학회논문지, 제17권, 제2호, 2013, pp. 347-354.

이종화, "Python을 이용한 SNS 크롤링 시스템 구축," 한국산업정보학회논문지, 제23권, 제5호, 2018, pp. 61-76.

이종화, 이문봉, 김종원, "TF-IDF 를 활용한 한글 자연어 처리 연구," 정보시스템연구 제28권, 제3호, 2019, pp. 105-121.

이종화, 이윤재, 이현규, "SNS 의 해시태그를 이용한 감정 단어 수집 시스템 개발," 정보시스템연구, 제27권, 제2호, 2018, pp. 77-94.

이종화, "SNS 해시태그를 이용한 감정 단어 일반화 연구," 인터넷전자상거래연구 제18권, 제4호, 2018, pp. 53-63.

이진수, "데이터 사이언스 기반의 디지털 트랜스포메이션," 방송과 미디어 제22권, 제4호, 2017, pp. 18-25.

Balahur, A., Hermida, J. M., and Montoyo, A., "Detecting implicit expressions of emotion in text: A comparative analysis," *Decision Support Systems*, Vol. 53, No. 4, 2012, pp. 742-753.

Cheng, F., Shen, J., Yu, Y., Li, W., Liu, G., Lee, P. W., and Tang, Y., "In silico prediction of Tetrahymena pyriformis toxicity for diverse industrial chemicals with substructure pattern recognition and machine learning methods," *Chemosphere*, Vol. 82, No. 11, 2011, pp. 1636-1643.

Christian, H., Agus, M. P., and Suhartono, D., "Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF)," *ComTech: Computer, Mathematics and Engineering Applications*, Vol. 7, No. 4, 2016, pp. 285-294.

Danielsson, P. E., "Euclidean distance mapping," *Computer Graphics and image processing*, Vol. 14, No. 3, 1980, pp. 227-248.

Gunasekaran, S., "Computer vision technology for food quality assurance," *Trends in Food Science & Technology*, Vol. 7, No. 8, 1996, pp. 245-256.

Kaelbling, L. P., Littman, M. L., and Moore, A. W., "Reinforcement learning: A survey," *Journal of artificial intelligence research*, Vol. 4, 1996, pp. 237-285.

Plutchik, R., *A general psychoevolutionary theory of emotion*, In Theories of emotion, 1980.

Qaiser, S., and Ali, R., "Text mining: use of TF-IDF to examine the relevance of words to documents," *International Journal of Computer Applications*, Vol.

181, No. 1, 2018, pp. 25-29.

Salton, G., and Buckley, C., "Term-weighting approaches in automatic text retrieval," *Information processing & management*, Vol. 24, No. 5, 1988, pp. 513-523.

Seung, D., and Lee, L., "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, Vol. 13, 2001, pp. 556-562.

Tata, S., and Patel, J. M., "Estimating the selectivity of tf-idf based cosine similarity predicates," *ACM Sigmod Record*, Vol. 36, No. 2, 2007, pp. 7-12.

Walker, M. A., Anand, P., Abbott, R., Tree, J. E. F., Martell, C., and King, J., "That is your evidence?: Classifying stance in online political debate," *Decision Support Systems*, Vol. 53, No. 4, 2012, pp. 719-729.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J., *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, 2016.

Ye, C., Yung, N. H., and Wang, D., "A fuzzy controller with supervised learning assisted reinforcement learning algorithm for obstacle avoidance," *IEEE Transactions on Systems*, Vol. 33, No. 1, 2003, pp. 17-27.

Ye, J., "Improved cosine similarity measures of simplified neutrosophic sets for medical diagnoses," *Artificial intelligence in medicine*, Vol. 63, No. 3, 2015, pp. 171-179.

이 종 화 (Lee, Jong-Hwa)

부경대학교 경영학 박사학위를 취득하고, 현재 동의대학교 정보경영학부 e비즈니스학 전공 교수로 재직 중이다. 주요 관심분야는 BigData, Mining, Content Analysis, Sentiment Analysis 등이다.

<Abstract>

Research on Designing Korean Emotional Dictionary using Intelligent Natural Language Crawling System in SNS

Lee, Jong-Hwa

Purpose

The research was studied the hierarchical Hangul emotion index by organizing all the emotions which SNS users are thinking. As a preliminary study by the researcher, the English-based Plutchick (1980)'s emotional standard was reinterpreted in Korean, and a hashtag with implicit meaning on SNS was studied. To build a multidimensional emotion dictionary and classify three-dimensional emotions, an emotion seed was selected for the composition of seven emotion sets, and an emotion word dictionary was constructed by collecting SNS hashtags derived from each emotion seed. We also want to explore the priority of each Hangul emotion index.

Design/methodology/approach

In the process of transforming the matrix through the vector process of words constituting the sentence, weights were extracted using TF-IDF (Term Frequency Inverse Document Frequency), and the dimension reduction technique of the matrix in the emotion set was NMF (Nonnegative Matrix Factorization) algorithm. The emotional dimension was solved by using the characteristic value of the emotional word. The cosine distance algorithm was used to measure the distance between vectors by measuring the similarity of emotion words in the emotion set.

Findings

Customer needs analysis is a force to read changes in emotions, and Korean emotion word research is the customer's needs. In addition, the ranking of the emotion words within the emotion set will be a special criterion for reading the depth of the emotion. The sentiment index study of this research believes that by providing companies with effective information for emotional marketing, new business opportunities will be expanded and valued. In addition, if the emotion dictionary is eventually connected to the emotional DNA of the product, it will be possible to define

the “emotional DNA”, which is a set of emotions that the product should have.

Keyword: Big Data, TF-IDF, Non-negative Matrix Factorization, Text Mining, Emotional Word, Sentimental Analysis

* 이 논문은 2020년 8월 19일 접수, 2020년 9월 11일 1차 심사, 2020년 9월 29일 게재 확정되었습니다.