ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2015 June: 21(2): 49~67 http://dx.doi.org/10.13088/jiis.2015.21.2.49

집단지성을 이용한 한글 감성어 사전 구축*

안정국
연세대학교 정보대학원
(jace@yonsei.ac.kr)

김희웅
연세대학교 정보대학원
(kimhw@yonsei.ac.kr)

최근 다양한 분야에서 빅데이터의 활용과 분석에 대한 중요성이 대두됨에 따라, 뉴스기사와 댓글과 같은 비정형 데이터의 자연어 처리 기술에 기반한 감성 분석에 대한 관심이 높아지고 있다. 하지만, 한국어는 영어와는 달리 자연어 처리가 어려운 교착어로써 정보화나 정보시스템에의 활용이 미흡한 실정이다. 이에 본 연구는 감성 분석에 활용이 가능한 감성어 사전을 집단지성으로 구축하였고, 누구나 연구와 실무에 사용하도록 API서비스 플랫폼을 개방하였다(www.openhangul.com). 집단지성의 활용을 위해 국내 최대 대학생 소셜네트워크 사이트에서 대학생들을 대상으로 단어마다 긍정, 중립, 부정에 대한 투표를 진행하였다. 그리고 집단지성의 효율성을 높이기 위해 감성을 '정의'가 아닌 '분류'하는 방식인 폭소노미의 '사람들에 의한 분류법'이라는 개념을 적용하였다. 총 517,178(+)의 국어사전 단어 중 불용어 형태를 제외한 후 감성 표현이 가능한 명사, 형용사, 동사, 부사를 우선 순위로 하여, 현재까지 총 35,000(+)번의 단어에 대한 투표를 진행하였다. 본 연구의 감성어 사전은 집단지성의 참여자가 누적됨에 따라 신뢰도가 높아지도록 설계하여, 시간을 축으로 사람들이 단어에 대해 인지하는 감성의 변화도 섬세하게 반영하는 장점이 있다. 따라서 본 연구는 앞으로도 감성어 사전 구축을 위한 투표를 계속 진행할 예정이며, 현재 제공하고 있는 감성어 사전, 기본형 추출, 카테고리 추출 외에도 다양한 자연어 처리에 응용이 가능한 API들도 제공할 계획이다. 기존의 연구들이 감성 분석이나 감성어 사전의 구축과 활용에 대한 방안을 제안하는 것에만 한정되어 있는 것과는 달리, 본 연구는 집단지성을 실제로 활용하여 연구와 실무에 활용이 가능한 자원을 구축하여 개방하여 공유한다는 차별성을 가지고 있다. 더 나아가, 집단지성과 폭소노미의 특성을 결합하여 한글 감성어 사전을 구축한 새로운 시도가 향후 한글 자연어 처리의 발전에 있어 다양한 분야들의 융합적인 연구와 실무적인 참여를 이끌어 개방적 협업의 새로운 방향과 시사점을 제시 할 수 있을 것이라 기대한다.

주제어 : 감성어 사전, 한글자연어처리, 감성 분석, 집단지성, 폭소노미

논문접수일 : 2015년 5월 20일 논문수정일 : 2015년 6월 12일 게재확정일 : 2015년 6월 13일
투고유형 : 국문급행 교신저자 : 김희웅

1. 서론

최근 데이터의 기하급수적인 증가와 소셜미디어의 급성장으로 빅데이터에 대한 관심이 많아지면서, 기업들은 다양하고 좋은 데이터를 어떻게 확보하느냐와 이를 어떻게 활용하는지에 대해 관심의 초점을 맞추고 있다(Prentice, 2011). Laney (2001)는 빅데이터를 3V (volume, velocity, and variety)로 정의하였으며, 최근에는 가치(value)를 추가한 4V로 정의를 하기도 한다(McAfee, 2012). 다시 말해 빅데이터 환경에서는 데이터의 사이즈, 처리/분석 속도, 다양성, 가치

* 본 논문은 2012년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원 (NRF-2012-2012S1A3A2033291)을 받아 수행된 연구임

창출이 매우 중요하며, 이는 기업의 경쟁우위 확보에 중요한 부분을 차지할 수 있다는 의미이다.

국내의 경우 정부가 주도적으로 공공데이터를 개방하고 있으며(Lee and Yoon, 2012), 기업들도 다양한 부가가치를 창출하기 위하여 데이터 개방과 더불어 매쉬업(mashup)과 같은 효과적인 웹서비스의 융합을 시도를 하고 있다(Lee, 2012). 이러한 정부와 기업들의 적극적 주도하에 데이터 공유와 협업을 기반으로 하는 생태계를 만들어 가는 분위기가 조성됨에 따라, 최근에는 고급 데이터 분석 기법인 자연어 처리 기술을 기반으로 한 감성 분석이 각광을 받고 있다(Pang and Lee, 2008).

감성 분석은 주로 주관적인 데이터를 분석하여 사람들의 성향 분석, 선별, 예측, 판단 등을 가능하게 하며, 최근 들어 SNS의 급부상과 함께 다양한 분야에 오피니언 마이닝의 개념으로도 적용이 되고 있다(Khan et al., 2014; Pang and Lee, 2008). 기존에 주로 사용했던 설문조사와 인터뷰는 시간과 비용이 드는 반면, 소셜미디어에서의 고객 선호도 조사는 비용절감효과 및 실시간 분석이 가능한 장점이 있다(Kim et al., 2011). Jang et al. (2015)는 사람들의 SNS의 사용에 대한 다양한 동기 요인에 대한 연구를 하였으며, Cho et al. (2014)는 영화 개봉 첫 주의 온라인 리뷰를 바탕으로 영화 흥행성 예측을 위한 연구를 진행하였고, Jang(2009)은 온라인 쇼핑몰에서의 상품평을 분석하여 의견을 긍정과 부정으로 판단하는 자동 분석 알고리즘을 소개하였다. 이러한 소셜미디어의 감성 분석은 기업적인 측면에서의 수익 증대, 마케팅 전략, 서비스 개선에 다양한 기회를 주었으며, 정치 캠페인과 같은 다양한 분야에서도 많은 효율성 증대를 가져왔다. 실제로 2012년 미국 대선의 오바마 캠프에서의 소셜미디어를 활용한 성공적인 여론 분석은 선거비용 절감과 맞춤형 선거전략을 가능하게 하였다(Bollen et al., 2009). 하지만 현재까지는 한글 감성 분석에 쓰이는 개방된 감성어 사전이 없으며, 기존 연구들이 방법을 제안하는 것에만 한정이 되어 있어 한글 감성 분석에 대한 연구가 활발하지 못한 실정이다.

이에 본 연구의 목적은 신뢰도가 높은 한글 감성어 사전을 구축하여 누구나 연구나 실무에 활용이 가능한 자원을 마련함으로써, 한국어 자연어 처리 연구의 개방형 협업(Open collaboration)을 위한 플랫폼을 구축하는데 있다. 이를 위하여 국내 대학생들의 집단지성을 이용하여 단어의 감성의 깊이 표현이 가능하게 하였으며, API (Application Programming Interface) 서비스 플랫폼을 구축을 하여 실시간으로의 접근성이 용이하도록 개방하였다.

2. 개념적 배경

2.1 감성 분석

기존의 설문이나 인터뷰와는 달리 텍스트에서 사람들의 주관적인 성향과 의견 등을 분석하는 자연어 처리기술을 감성 분석이라 한다(Pang and Lee, 2008). 최근 빅데이터와 소셜미디어를 활용한 다양한 감성 분석이 이루어지고 있으며, 영어의 경우에는 WordNet의 각 어휘에 긍정, 중립, 부정을 태깅한 SentiWordNet이 대표적인 사례이다(Baccianella et al., 2010). 정확한 감성 분석을 위해서는 신뢰도가 높은 감성어 사전이 사용되어야 하는데, 언어의 사용 과정에서 발생하는 다양한 의미의 변이와 동태적 활용, 동음이의

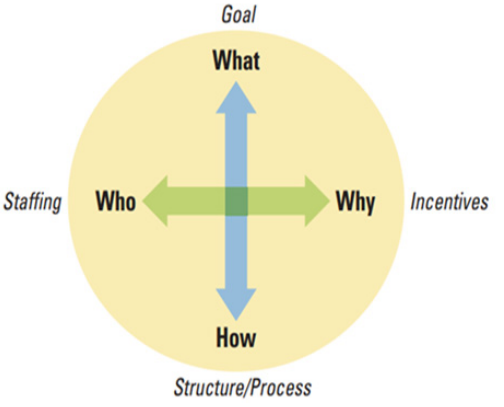
어가 분석에 있어 어려운 작업으로 인식되고 있다(Lee, 2011; Jung et al., 2008). 따라서 이러한 복잡한 상황을 고려하여 신뢰도가 높은 감성어 사전을 구축하고, 이를 통해 감성 평가의 정확성을 향상시키는 연구가 이루어지고 있다(Baccianella et al., 2010; Ban and Jung, 2001; Taboada et al., 2011).

Taboada et al(2011)는 상품, 영화, 도서평 등의 다양한 의견 텍스트의 감성을 분석하기 위하여 감성어 사전 구축에 대한 연구를 진행하였는데, 어휘의 감성을 정량적인 지표로 표현하기 위해 각 어휘에서 사용되는 단어의 긍정과 부정 값의 차이를 이용하였고, Kim et al(2012)는 감성 분석을 통해 기업의 성과의 방향성을 예측하는 연구를 하였다.

하지만 기존의 연구들은 문서의 극성 판별에 가장 큰 영향을 미치는 감성어 사전의 구축에 대한 명확한 언급이 없으며, 신뢰도가 높은 한글 감성어 사전을 사용하지 못했다는 한계점을 가지고 있다. 이에 본 연구는 한글 감성 분석에 활용될 수 있는 신뢰도가 높고 감성의 깊이 표현이 가능한 감성어 사전을 구축한다.

2.2. 집단지성 (Collective Intelligence)

집단지성은 다수의 개체들이 서로 협력 혹은 경쟁을 통해 얻게 되는 지적 능력의 결과물로 정의되며 사회학, 컴퓨터 공학 등의 다양한 분야에서 연구가 되어왔다(Bonabeau, 2009). 집단지성의 개념은 분야에 따라 조금씩 다르며, Lévy(1997)는 과학 기술을 이용한 인류의 공동 지적 능력과 자산이 소통을 통해 집단지성으로 구축되었다고 하였고, Sulis(1997)는 집단지성을 참여자들 간의 확률적으로 갖는 모임으로 인식을 하기도 하였다, Boder(2006)는 집단지성이 기업의 경영에 있어서 중심이 되는 역할로써, 기업의 다양한 활동에 집단지성을 적용할 필요가 있다고 주장하였다. 기존의 문헌에서는 집단지성의 개념이 현실에서 가장 잘 반영된 사례로 위키피디아를 들고 있다(Cachia et al., 2007; Black, 2008). 위키피디아가 구현한 집단지성의 경우, 개방된 형태에서 생성되어 진화되는 특징을 갖고 있으며 집단이라는 개념을 전문가들이 아닌 아닌 일반적 개념의 '대중'으로 정의하였다. 이러한 대중의 지성의 합은 개인의 지성보다 현명하다는 인식에서 제기된 개념으로, 참여자들의 다양한 경험이 현명한 판단과 좋은 결과물을 이끌어 낼 수 있다(Malone et al, 2010).



〈Figure 1〉 Elements of Collective Intelligence (Malone et al., 2010)

Malone et al.(2010)는 집단지성을 구성하는 요소를 크게 WHO, WHY, HOW, and WHAT으로 정의를 하였으며, 성공적인 집단지성을 이끌어내기 위해서는 누가(WHO), 어떠한 동기요인(WHY)으로 참여를 하며 어떠한 방식(HOW)으로 무엇(WHAT)을 만들 것인가에 대한 전체적인 프레임워크가 있어야 한다고 하였다(Figure 1 참조).

본 연구에서는 한글 감성어 사전 구축(WHAT)을 위해, 기본적인 교육적 소양을 가진 대학생(WHO)들을 선택을 하였고, 이들의 참여를 유도하기 위해 가입승인 및 등급향상과 같은 보상(WHY)을 주었으며, 오류를 방지하여 신뢰도를 높이는 시스템의 투표 방식을 선택하였다(HOW). 본 연구는 집단지성을 연구의 설계에 직접적으로 적용하여 실질적으로 연구와 실무에 도움이 되는 자원을 창출하였고, 이를 기반으로 향후 연구들에게 새로운 시사점을 주는데 큰 의의가 있다고 본다.

1.1. 폭소노미 (Folksonomy)

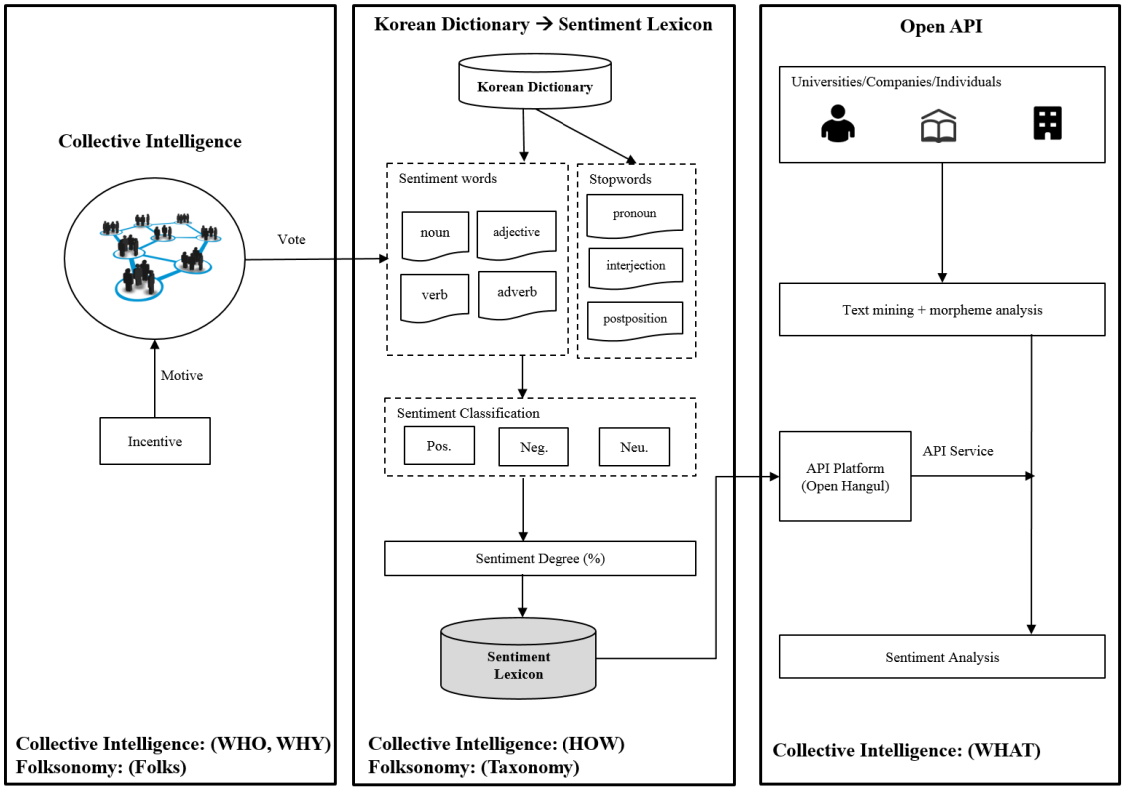
폭소노미는 Folks와 Taxonomy의 합성어로 '사람들에 의한 분류법'이라는 의미를 내포하고 있으며, 2005년에 시스템 설계사인 Thomas (2007)에 의해 처음 소개가 되었다. 크게 협업적 태깅(collaborative tagging), 소셜분류(social classification), 소셜색인(social indexing), 소셜태깅(social tagging)의 의미로 나뉘며 웹2.0, 블로그, 소셜커뮤니티 사이트들이 확산과 더불어 각광을 받게 되었다. 폭소노미의 대표적인 예로는 블로그의 글들에서 흔히 볼 수 있는 태그(tag)를 들 수가 있으며, 이는 기존의 카테고리가 아닌 새로운 카테고리의 구분을 확장시키는 장점이 있다(Russell, 2005; Thomas, 2007). 예를 들면, 블로그에서 '삼성 야구팀'에 대한 글은 다른 축구나 농구와 관련된 기사들과 같은 '스포츠' 카테고리이다. 하지만 문서에 '야구,' '삼성,' '스포츠', '이승엽'과 같은 다양한 태그를 달면, 태그들이 새로운 카테고리들을 생성을 하게 되는 것이다. 즉 기존에는 다른 카테고리에 있던 문서들이 태그를 활용함으로써 새로운 카테고리로 구분되어 수직적 카테고리가 아닌 수평적 카테고리로써의 무한한 확장성을 가진다. 이 외에도 폭소노미 기반으로 메타데이터(Metadata)를 이용한 소셜 북마크 온톨로지 구축을 제안하는 연구(Ohmukai et al., 2005)와 온톨로지를 폭소노미에 활용하는 연구가 있었다(Medelyan and Legg, 2008; Echarte et al., 2007; Ohkura et al., 2006).

그러나 폭소노미는 다양한 정보를 얻을 수 있는 장점이 있는 반면, 태깅이 무작위이므로 신뢰성이 떨어지는 단점이 있다(Gruber, 2007; Hwang and Kang, 2008), 이러한 한계점으로 인해 폭소노미에 대한 관심과 연구는 2007년 이후로 많이 감소된 추세이다. 하지만 본 연구의 연구자들은 폭소노미의 특성인 참여에 의한 협업과 분류가 빅데이터 환경의 시스템에 근본적인 틀과 방향성을 줄 수 있을 것이라고 확신을 하였고, 폭소노미의 단점인 낮은 신뢰성을 극복하기 위해 앞서 언급한 집단지성을 활용하기로 하였다. 본 연구에서는 폭소노미와 집단지성을 결합하여 개인적인 선입견이 들어가지 않는 집단들에 의한 신뢰도가 높은 감성어 사전을 구축하며, 이러한 시도는 폭소노미에 대한 재조명을 하는 점에 있어 중요한 의의가 있다.

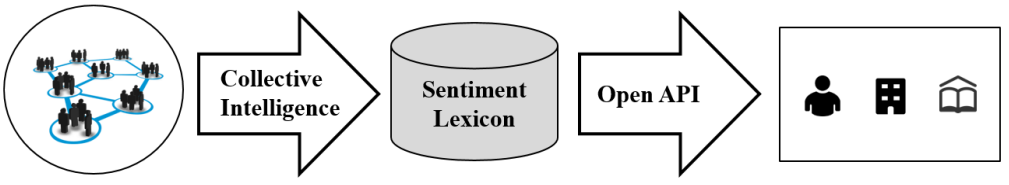
3. 연구 설계

3.1. 연구 절차

본 연구는 크게 (1)집단지성, (2)감성어 사전을 구축, (3)오픈 API 구축의 3단계 프레임워크를 가지고 있으며(Figure 3 참조), (Figure 2)에서는 세부적인 단계들을 살펴볼 수 있다. 1단계에서는 집단지성을 이용하여 단어들의 감성에 대한 투



〈Figure 2〉 Research Design



〈Figure 3〉 Research Workflow

표를 하게 하며, 집단지성의 요소 중에 WHO, WHY, 그리고 폭소노미의 Folks(사람들에 의한)의 개념이 적용이 된다. 2단계에서는 감성어 사전을 구축하기 위한 전처리 작업을 하여 국립국어원의 사전을 기반으로 총 517,178(+)개 단어를 데이터베이스에 생성하였다. 불용어를 제외한 단어들에 대해서 투표를 랜덤으로 하게 하며, 이는 집단지성의 HOW와 Folksonomy의 Taxonomy(분류법)의 개념이 적용이 되는 단계이다. 그리고 3단계에서는 감성 분석을 하는 연구자, 기업, 개인들에게 오픈API로 제공을 하기 위한 서비스 플랫폼을 구축한다.

〈Table 1〉 Lexicon Structure

Index			Decomposition	Word type
ㄱ	거	거머릿과	ㄱㅓㅁㅓㄹㅣㅅㄱㅘ	Noun
ㄱ	거	거머먹다	ㄱㅓㅁㅓㅁㅓㄱㄷㅏ	Verb
ㄱ	거	거머멀쑥	ㄱㅓㅁㅓㅁㅓㄹㅆㅜㄱ	
ㄱ	거	거머멀쑥이	ㄱㅓㅁㅓㅁㅓㄹㅆㅜㄱㅇㅣ	Adverb
ㄱ	거	거머멀쑥하다	ㄱㅓㅁㅓㅁㅓㄹㅆㅜㄱㅎㅏㄷㅏ	Adjective
ㄱ	거	거머무트룩하다	ㄱㅓㅁㅓㅁㅜㅌㅡㄹㅜㄱㅎㅏㄷㅏ	Adjective
ㄱ	거	거머무트름	ㄱㅓㅁㅓㅁㅜㅌㅡㄹㅡㅁ	
ㄱ	거	거머무트름하다	ㄱㅓㅁㅓㅁㅜㅌㅡㄹㅡㅁㅎㅏㄷㅏ	Adjective
ㄱ	거	거머무트름히	ㄱㅓㅁㅓㅁㅜㅌㅡㄹㅡㅁㅎㅣ	Adverb
ㄱ	거	거머번드르	ㄱㅓㅁㅓㅂㅓㄴㄷㅡㄹㅡ	
ㄱ	거	거머번드르하다	ㄱㅓㅁㅓㅂㅓㄴㄷㅡㄹㅡㅎㅏㄷㅏ	Adjective

3.2. 사전 데이터 설계

기본 색인 구축에 있어서 1차로는 단어의 'ㄱㄴㄷ'의 모음 순으로 하였고, '가나다'의 자모 결합 순으로 2차로 나누었다(Table 1 참조). 추가로 자모를 분리하여 저장하였는데, 이는 단순히 분류나 나열을 위한 것이 아니라 검색의 최적화, 유사도, 오타인식, 기본형 추출을 위한 것으로 향후 다양한 확장이 가능한 장점이 있다.

본 연구에서는 삭제/삽입/수정의 횟수의 총합인 '편집거리'를 이용하는 레빈쉬타인 거리 매트릭스를 검색에 사용하며, 이는 기존의 순차적으로 문자를 인식하는 방식과는 다른 효율적이며 논리적인 단어인식을 가능하게 한다(Levenshtein, 1966). 예를 들면 'Setting와 'Katten' 의 편집거리는 3이다(Table 2 참조). 하지만 레빈쉬타인 거리는 알파벳처럼 문자당 1byte인 경우만 가능하므로 글자당 초성, 중성, 종성의 결합형인 2byte인 한글은 그대로 사용할 수가 없다. 따라서 레빈쉬타인 거리 매트릭스를 한글에 적용할 수 있게 자음과 모음을 각각 1byte로 인식하도록 분리하여 데이터베이스에 입력하였다(Table 1 참조). 즉 커서의 이동단위를 자모 단위로 가능하도록 하여 레빈쉬타인 거리 계산을 용이하게 하는 것이다.

〈Table 2〉 Levenshtein Distance
(Setting vs. Katten)

		K	A	T	T	E	N
	0	1	2	3	4	5	6
S	1	1	2	3	4	5	6
E	2	2	1	2	3	4	5
T	3	3	2	1	2	3	4
T	4	4	4	3	1	2	3
I	5	5	4	3	2	2	3
N	6	6	5	4	3	3	2
G	7	7	6	5	4	4	3

3.3. 우선 순위 단어의 형태 선별

총 517,178(+)개 단어 중에서 감성의 표현이 가능한 단어의 형태들을 선별하였으며 조사, 감탄사, 관형사 등의 단어들은 제외하고 명사, 형

〈Table 3〉 Classification of Words

Word type	Total	Sentiment	Word type	Total	Sentiment
interjection	682	N	assistant verb	14	N
interjection·noun	85	N	assistant adjective	17	N
determiner	207	N	adverb	17,425	Y
determiner·interjection	11	N	adverb·interjection	3	N
determiner·noun	1,267	Y	numeral	60	N
pronoun	382	N	numeral·determiner	195	N
pronoun·interjection	5	N	numeral·determiner·noun	3	N
pronoun·determiner	3	N	ending	6	N
pronoun·adverb	1	N	bound noun	913	N
verb	68,370	Y	bound noun·postposition	13	N
verb·adjective	2	Y	affix	209	N
noun	337,659	Y	postposition	300	N
noun·adverb	109	N	adjective	16,562	Y
Total words	517,178		Total sentiment words	441,283	

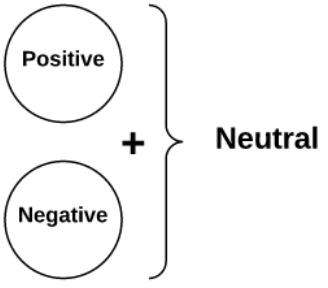
용사, 동사, 부사들을 우선 순위로 두었다(Table 3 참조).

동음이의어의 경우에는 첫 번째 순위에 있는 의미가 단어의 중요도와 사용빈도가 높으므로 우선 순위로 두었다. 또한 불용어로 분류된 단어 형태에도 간혹 감성 단어들이 있지만 전체적인 변별력을 높이기 위해 감성어 대상에서 제외하였고, 이는 감탄사의 경우만 하더라도 문장의 성격과는 무관하게 주변 단어의 강조를 위해 쓰이는 경우가 많기 때문이다. 총 517,178개의 단어 중 불용어로 처리된 단어들은 75,895개이다(Table 3 참조). (Figure 2)의 연구설계 2단계를 보면 감성어 단어로 분류된 형태들(명사, 형용사, 동사, 부사)을 대상으로 집단지성의 참여자들이 투표를 하게 되고, 나머지 불용어 형태의 단어들(감탄사, 관형사, 대명사, 수사, 의존명사, 접사, 조사 등)은 투표의 대상에서 제외를 하였다.

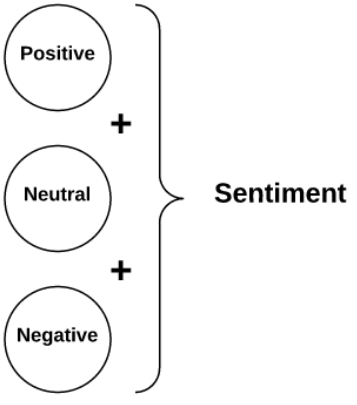
3.4. 단어들의 감성 태그 및 점수화 알고리즘

태그는 성격에 따라 유형의 구분이 된다. Xu et al. (2006)는 태그를 Content-based, Context-based, Attribute, Subjective, Organizational로 5가지로 분류를 하였다. 본 연구에서의 태그는 단어의 내용에 대해 표현하는 '내용 기반형(Content-based)'이며, 긍정, 중립, 부정의 감성의 극성 판별과 깊이나 확률이 태그가 된다. 참고로 본 연구에서는 폭소노미의 태그를 긍정, 중립, 부정의 투표로 변환을 하였다. 이는 폭소노미의 자율적인 태깅(주관식 방식)과는 다른 선택적 태깅(객관식 방식)이며, 기존의 폭소노미의 단점인 태깅의 남용(Gruber, 2007)으로 인한 낮은 신뢰도와 정확도를 극복하기 위한 것이다. 각 단어마다 사람들이

긍정, 중립, 부정을 선택하게 하여 이를 기반으로 감성의 깊이나 확률을 정량적으로 계산한다. 즉 세 가지의 제한된 태깅을 하지만 표현은 개인의 판단이 아닌 집단지성의 합이므로 단어마다 감성의 깊이 표현을 개인으로써 하는 것이 아니고 집단의 구성원으로써 하게 되는 것이다. 예를 들면 A라는 사람이 '시원하다'의 단어에 긍정으로 선택한다면 100% 긍정어로 태깅이 되는 것이 아니라 기존의 집단지성으로 계산된 값에 추가가 되어 A의 긍정 태깅이 부분적으로 반영이 되는 것이다. 이런 방식의 장점은 개인이 결과물을 독단적으로 조작하는 태깅의 남용을 방지하여 신뢰성을 높일 수 있다는 것이다.



〈Figure 4〉 Sentiment Polarity



〈Figure 5〉 Sentiment Score

또한 우리가 쓰는 언어는 시간에 따라 조금씩 변하는 자연어라고 하는데, 본 연구에서 구축한 감성어 사전은 이러한 자연어 특성을 고려하기 위해 과거와 현재의 투표를 모두 반영하여 언어의 시간적 변화를 점차적으로 수용한다. 감성 점수를 계산하는 방법은 중립을 어떻게 처리하느냐에 따라 두 가지 방법을 단계적으로 사용한다.

감성어 판별은 긍정과 부정의 점수만 계산하여 중립을 자동으로 분류하여 극성을 구분하는 방법(Figure 4 참조)과 긍정, 부정, 중립을 독립적으로 계산하여 감성 점수를 계산하는 방법(Figure 5 참조)이 있다. 전자의 경우에는 중립의 투표를 제외하는데 그 이유는 사람들이 단어의 뜻을 모를 경우에도 중립으로 선택을 하기 때문이다. 그래서 (Table 3)에서 1차적으로 분리를 했던 불용어들을 2차적으로 분리할 수 있는 장점이 있다. 점수화 계산은 단어의 긍정 또는 부정일 확률이 0% ~ 60% (threshold=60%)이면 자동적으로 중립으로 분류가 되며 긍정이나 부정으로 판별이 되기 위해서는 최소한의 60% 이상의 확률이 있어야 한다. 후자인 경우인 (Figure 5)의 긍정, 중립, 부정을 독립적으로 점수화한 방식은 비감성어도 중립어로 인식을 하게 하였으며 실질적으로 비감성어와 중립어의 차이는 있지만 텍스트 마이닝 분석에 있어서는 큰 차이는 없다. 본 연구에서 쓰인 감성어 점수 알고리즘의 첫 단계에서는 단어의 중립을 판단하기 위해 전자(Figure 4 참조), 그리고 두 번째 단계에서는 긍정과 부정의 확률을 계산하기 위해 후자의 방법을 쓴다(Figure 5 참조).

본 연구에서의 각 단어에 대한 감성 점수화 방식은 다음과 같다(Figure 6 참조). 1단계에서는 긍정, 중립, 부정의 투표의 결과를 바탕으로 중립으로 투표한 사람들이 긍정이나 부정이라고

```
// for neutral    (Step 1)
if($data[neutral] > $data[positive]    && $data[neutral] > $data[negative]) {
            ${score.$i} = 100 *    $data[neutral]/(abs($data[positive] - $data[negative]) + $data[neutral]);
            ${sentiment.$i} = ' Neutral ';
}
// for    positive (Step 2)
elseif($data[positive]    > $data[negative]) {
            ${score.$i} = 100 *    $data[positive]/($data[positive] + $data[negative]);
            ${sentiment.$i} = 'Positive';
}
// for    negative
elseif($data[positive]    < $data[negative]) {
            ${score.$i} = 100 *    $data[negative]/($data[positive] + $data[negative]);
            ${sentiment.$i} = 'Negative';
}
// when    positive = neutral = negative (Step 3)
else {
            ${score.$i} = '100';
            ${sentiment.$i} = ' Neutral';
}
```

〈Figure 6〉 Sentiment Algorithm

생각하는 사람들보다 많을 경우, 중립으로 판별을 하고 추가적으로 중립어로서의 확률을 계산을 한다. 즉 집단지성의 관점에서는 집단이 단어를 중립으로 결정했다는 의미이며, 폭소노미의 관점에서는 단어가 사람들에 의해 중립으로 태깅/분류가 되었다는 것을 의미한다. 2단계에서는 긍정과 부정의 투표수를 비교하여 긍정이나 부정의 확률을 %로 계산을 한다. 마지막으로 3단계에서는 긍정, 부정, 중립의 수가 같아 1단계와 2단계에 해당이 되지 않는 경우는 100% 중립어로 계산을 한다.

3.5. 데이터 수집

앞에서 설명한 집단지성의 네 가지 구성 요소 중에서 참여자들(WHO)을 선택함에 있어서, 본 연구자들은 기본적인 교육적 소양을 가진 다양하고 평범한 사람들이 본 프로젝트에 적절한 집단이라고 생각하였다. 우선적으로 평균 연령과 교육 수준을 고려했으며, 다양한 집단들로 나누어 봤을 때 대학생들이 가장 적절하다고 생각하였다. 이에 국내 대학생 소셜네트워크 사이트에서 객관식 설문문항 형식으로 투표를 실시하였으며, 2014년 8월 15일부터 2015년 5월 15일까지 총 35,000번의 단어에 대한 투표가 진행되었으며 참가자들은 주어진 단어들에 대해 긍정, 중립, 부정의 답변 중에 하나를 선택을 하였다(Figure 7).

객관식 설문문항 형식에서는 오답방지를 위해 답이 확연한 단어 두 개를 삽입하였다. 예를 들면 '좋다'와 '멍청하다'와 같이 확연한 긍정이거나 부정인 단어를 넣어, 참가자가 두 문항을 동시에 틀릴 경우 나머지 문항의 답변들도 오류로 인식을 하여 투표에 반영을 하지 않았다. 웹 서비스의 특성을 고려하여 편의성, 그리고 답변의 정확도를 높이기 위해 응답시간이 1분 내외로 걸리도록 하였으며 한 명당 총 10 단어씩 답변을 하게 하였다.

〈Figure 7〉 Voting System for Sentiment

4. 연구결과 및 실무 활용

4.1. 감성어 사전 API

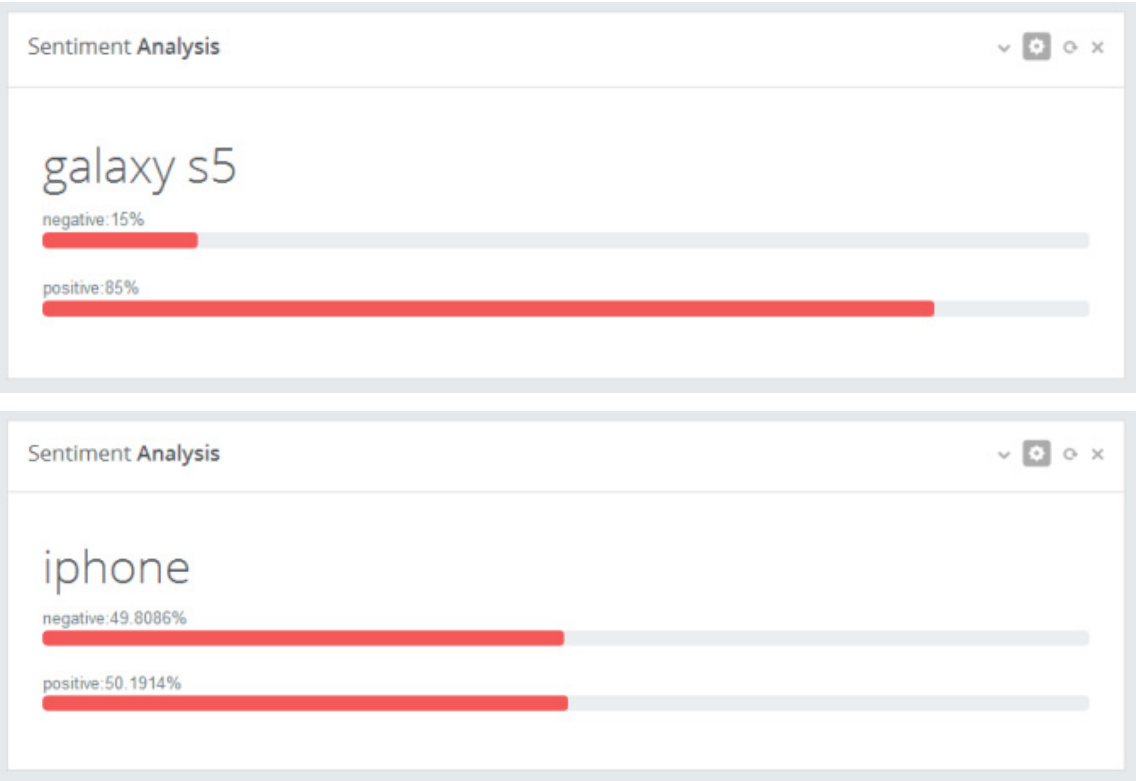
본 연구에서 구축한 한글 감성어 사전은 빅데이터 분석을 수행하는 회사나 기관들이 API를 활용할 수 있도록 플랫폼을 구축하였다. 플랫폼에서 제공하는 API서비스는 감성에 대한 질의 외에도 기본형 추출, 카테고리 추출 등과 같은 다양한 기능도 제공한다(오픈한글, www.openhangul.com). 예를 들면, 사용자가 특정단어를 API로 질의하면, 플랫폼에서는 단어에 대한 형태, 감성판별, 감성의 깊이에 대한 응답을 준다. 감성의 깊이에 대한 표현은 본 연구의 차별화된 부분이며, 이는 집단지성을 활용해 감성어 사전을 구축하였기 때문에 가능한 것이다.

4.2. 감성 분석을 활용한 평판 분석

감성어 사전 API를 확장하여 활용하면 기업의 관점에서는 자사 제품, 브랜드, 경쟁사 평판 분석 등에 적용이 가능하다. 또한 시간적 관점에 따른 다각적인 분석도 가능한데 기업이 신제품의 출시 전후에 대한 평판을 분석할 수도 있고, 반대로 자사 제품의 출시가 경쟁사의 평판에도 영향을 미치는 지에 대한 분석도 가능하다.

출시 전에 대한 분석은 기대치를 반영하는 preview, 출시 후는 고객의 반응을 반영하는 review가 될 수 있다. 예를 들어 (Figure 8)을 보면 갤럭시 S5에 대한 사람들의 반응은 85%, 아이폰은 50%가 긍정으로 나온다. 감성 분석을 한 시점은 아이폰 6가 나오기 전의 시점이므로 사람들이 아이폰 6에 대한 기대치가 낮다는 의미가 될 수 있다. 이를 이용하여 확장적인 분석도 가능한데, 예를 들면 아이폰 6가 출시 된 이후의 시점(사용후기)을 출시 전(기대치)과의 비교를 통해 '별로지만 기대했던 것보다는 덜 별로이다'와 같은 주관적이며 상대적인 분석도 가능하다.

(Figure 9, 10)은 글의 내용을 세부적으로 분석



〈Figure 8〉 Sentiment Analysis (iPhone vs Galaxy)

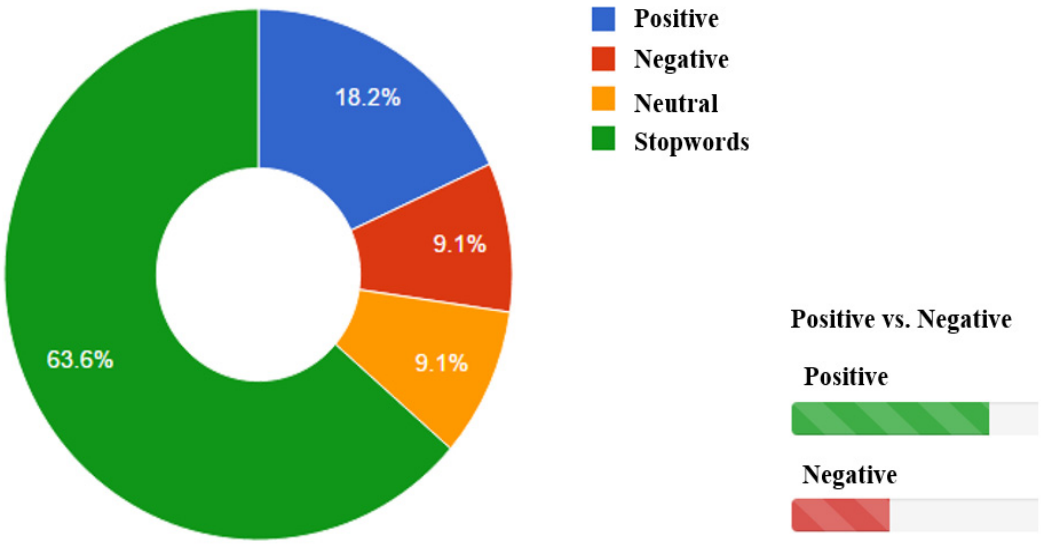
하는 과정을 직관적으로 표현을 하여 이해를 도운 것이다. 예를 들면 한 고객이 스마트폰을 구매를 한 후 페이스북이나 트위터에 'OOO를 샀는데 무게는 가벼운 듯 한데 기능은 좋다. 색상이 밝은데 화질이 나쁘지는 않아서 나름 좋은 것 같아' 라는 글을 남겼다면, 문장을 각각의 최소 형태의 단어로 분리하는 작업(토큰화)을 하여 감성 분석을 하고 시각적으로 표현을 할 수가 있다.

(Figure 9)에서는 우선적으로 문장에 대한 토큰화 작업을 하여 감성어 단어 분류를 하고 빈도수 계산을 한다. 빈도수 계산은 단어의 중복출현을 의미하는데 이는 빈도수에 따른 가중치를 적용하기 위한 것이다(Hwang and Ko, 2009). 또한 (Figure 9)에서 계산된 빈도와 확률을 기반으로 개발자들이나 연구자들이 다양한 알고리즘을 구축할 수 있으며, (Figure 10)처럼 텍스트를 토큰화하여 비감성어나 중립어의 비율을 알고리즘에 포함하거나 제외하여 해석할 수 있다. 참고로 이러한 알고리즘에 대한 고려는 감성 분석을 하는 개발자들이나 연구자들의 주관적인 논리가 적용되어 확장이 되는 부분이다. (Figure 10)에서는 문장 전체에서의 감성어 비율과 긍정, 부정의 정도를 계산을 하여 정량적으로 분석한 것뿐만 아니라 직관적인 시각화로 표현하였다.

▸ Token Analysis

Token	Type	Freq.	Pos.	Neg.	Neu.	Score	Weighted Score	%
좋다	Adjective	2 times	26	0	0	Pos.: 100%	Pos.: 200%	
가볍다	Adjective	1 times	4	0	17	Neu.: 81%	Neu.: 80.95%	
나쁘다	Adjective	1 times	0	20	0	Neg.: 100%	Neg.: 100%	
밝다	Verb	1 times	14	0	2	Pos.: 100%	Pos.: 100%	

〈Figure 9〉 Sentiment Probability by Token



〈Figure 10〉 Visualization of Sentiment Analysis

5. 연구결과에 대한 기대효과 및 활용 방안

본 연구에서 객관적인 감성 분석을 위한 사전 구축의 필요성을 인식하게 된 배경은 최근 소셜 미디어가 방대한 양의 이용자들의 의견을 실시간으로 축적함에 있다(Lipsman et al., 2012). 페이스북, 트위터와 같은 SNS 등에서 수집되는 데

이터들은 특성상 해당 국가의 언어적 특성에 매우 큰 영향을 받게 되고, 이로 인해 분석방법에 차이가 존재한다. 한글은 어미와 조사가 발달한 교착어이므로 자연어 처리가 어렵고, 자연어 처리를 위한 감성어 사전과 같은 자원이 부족해 감성 분석에 대한 연구가 활발하게 진행되지 못하였다.

집단지성의 특징은 참여자들의 수가 누적이 될수록 좀 더 나은 결과물이 나온다는 것이다(Malone et al., 2010). 위키피디아의 경우를 보더라도 다수의 사람들이 참여한 영어나 다른 언어의 버전의 경우는 한국어 버전과 비교를 하면 데이터의 양적인 측면이나 내용의 질적인 측면에서 많은 차이를 보이고 있으며(Hwang and Choi, 2010), 이는 집단지성의 결과물이 참여자들의 누적효과에 많은 영향을 받는다는 것을 의미한다. 본 연구는 시간을 축으로 참여자들의 지성의 누적으로 인한 감성어 사전의 신뢰도가 높아지는 시스템으로 구축이 되었으며, 이로 인해 향후 한국어 자연어 처리에 이바지하는 유용한 자원이 되기를 기대하고 있다.

본 연구에서 구축된 감성어 사전과 오픈 API 서비스는 다양한 영역에서 활용이 가능하며(Table 4 참조), 누구나 참여가 가능한 개방형 협업의 기초를 제공하여 한글 감성어 사전을 활용한 연구와 실무에의 활발히 쓰이기를 기대한다.

6. 연구의 한계 및 향후 연구 방향

본 연구에서 제시한 한글 감성어 사전 활용에 있어 한계점은 다음과 같다. 감성어 사전은 감성 분석에 쓰이는 도구이며 감성 분석을 자체적으로 하는 알고리즘이 아니므로, 감성 분석을 함에 있어서 문장에서의 맥락, 문맥의 해석과 동음이의어의 복잡성 문제는 감성어 사전이 해결을 해주는 영역이 아니다. 즉 감성어 사전은 사전의 고유의 목적에 맞게 문맥을 고려하지 않고 단어를 독립적인 상태로서의 감성을 표현해 주기 때문에 복잡한 문맥의 감성 분석은 분석자가 알고리즘에서 해결해야 한다(Nasukawa and Yi, 2003). 또한 API의 사용을 위해서는 형태소 분석

〈Table 4〉 Applicable Fields

Field	Use	Example
Business	Decision Making	Establishing a decision making model based on the customers' opinions on social media
	Brand Monitoring	Monitoring customers' product reviews and previews through ex-ante and ex-post monitoring process
	Reputation Management	Preventing a negative publicity by building a strategy for positive images
Accounting/ Finance	Corporate Transparency	Utilizing financial statements footnotes to find patterns by analyzing the high frequency words
Investment	Social Media Analysis	Establishing an intelligent investment decision model by analyzing companies' online rumors and news
Marketing	Effectiveness	Measuring an effectiveness of marketing campaigns by conducting ex-ante and ex-post monitoring
	Market Segmentation	Utilizing market segmentation strategies by analyzing customers' preferences based on their demographics
	Competitor Monitoring	Monitoring competitor's online reputation

과 같은 전처리 작업이 필요하다. 감성어 사전의 단어들은 국어사전처럼 기본형 단어들만 있기 때문에 분석할 단어들의 기본형 변환을 위한 형태소 분석과 토큰화(Tokenization)가 필요하다 (Lee, 2011).

이러한 한계점을 해결함과 동시에 향후 계획하는 연구방향은 크게 두 가지로 분류할 수 있다. 첫째, 문맥적인 문제를 해결하기 위해 감성 분석에서는 사전에 텍스트의 카테고리를 구분하는 작업이 필요하며, 감성어 사전에서는 카테고리별로 추가적으로 사용되는 단어들의 온톨로지 구축이 필요하다. 예를 들면, 경제나 스포츠의 카테고리를 인식한 후, 기본적인 감성어 사전에 추가적으로 경제나 스포츠 관련 감성어 사전을 적용시켜 특정 카테고리에서 다른 의미로 쓰일 수 있는 단어들의 정확한 감성어 분석을 가능하게 할 수 있다. 둘째, 토큰화와 형태소 분석에 대한 기본적인 지식과 경험이 없는 사용자들을 위해 플랫폼에서 이 모든 전처리 작업들을 해결하는 것이 필요하다. 따라서 향후 연구에서는 감성어 사전과 같은 추가적인 '자원적인' 측면과 더불어 '프로세스적인' 측면으로의 확장된 연구를 진행을 할 예정이다.

참고문헌(References)

Baccianella, S., A. Esuli, and F. Sebastiani, "Senti WordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," *LREC,* Vol. 10(2010), 2200~2204.

Ban, S. B. and C. S. Jung, "A neural network model for recognizing facial expressions based on perceptual hierarchy of facial feature points," *Korean journal of cognitive science,* Vol.12, No.1/2(2001), 77~89.

Black, E. W., "Wikipedia and academic peer review: Wikipedia as a recognised medium for scholarly publication?," *Online Information Revie,w,* Vol. 32, No. 1(2008), 73~88.

Boder, A., "Collective intelligence: a keystone in knowledge management," *Journal of Knowledge Management*, Vol. 10, No. 1(2006), 81~93.

Bollen, J., A. Pepe, and H. Mao, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," *arXiv preprint arXiv*:0911.1583. (2009).

Bonabeau, E., "Decisions 2.0: The power of collective intelligence," *MIT Sloan management review*, Vol. 50, No.2(2009), 45~52.

Cachia, R., R. Compañó, and O. D. Costa, "Grasping the potential of online social networks for foresight," *Technological Forecasting and Social Change*, Vol. 74, No. 8(2007), 1179~1203.

Cho, S. Y., H.-K, Kim, B. Kim, and H. -W. Kim, "Predicting Movie Revenue by Online Review Mining: Using the Opening Week Online Review," *Information Systems Review,* Vol. 16, No. 3(2014), 111~132.

Echarte, F., J. J. Astrain, A. Córdoba, and J. E. Villadangos, "Ontology of Folksonomy: A New Modelling Method," *SAAKM*, 289, 36(2007).

Gruber, T., "Ontology of folksonomy: A mash-up of apples and oranges," *International Journal on Semantic Web and Information Systems (IJSWIS),* Vol. 3, No. 1(2007), 1~11.

Hwang, J. S. and S. Y. Choi, "Analysis of Participants' Features in Different Collective Intelligence Models: Comparative Analysis between Korea and U.S.A.," *Journal of*

Cybercommunication, Vol.27, No.4(2010), 257~301.

Hwang, J. W., and Y. J. Ko, "A Document Sentiment Classification System Based on the Feature Weighting Method Improved by Measuring Sentence Sentiment Intensity," *Journal of KIISE* Vol.36, No.6(2009), 491~497.

Hwang, S. H., and Y. K. Kang, "Hierarchical Triadic Context Analysis for Folksonomy-Based Web Applications," *JDCTA*, Vol.2, No.1(2008), 20~27.

Jang, J.-Y.,"A Sentiment Analysis Algorithm for Automatic Product Reviews Classification in On-Line Shopping Mall," *The Journal of Society for e-Business Studies*, Vol.14, No.4 (2009), 19~33.

Jang, Y., E. Cho, and H. Kim, "An Exploratory Study on Online Prosocial Behavior," *Knowledge Management Research*, Vol.16, No.1(2015), 225~242.

Jung, Y. C., Y. J. Choi, and S. H. Myaeng, "A Study on Negation Handling and Term Weighting Schemes and Their Effects on Mood-based Text Classification," *Korean journal of cognitive science*, Vol.19, No.4 (2008), 477~497.

Khan, F. H., S. Bashir, and U. Qamar, "TOM: Twitter opinion mining framework using hybrid classification scheme," *Decision Support Systems*, Vol.57(2014), 245~257.

Kim, J. O., S. Lee, and H. S. Yong, "Automatic Classification Scheme of Opinions Written in Korean," *Journal of KIISE*: *Database*, Vol. 38, No.6(2011), 423~428.

Kim, Y., N. Kim, and S. R. Jung, "Stock-Index Invest Model Using News Big Data Opinion Mining," *Journal of Intelligence and Information Systems* Vol.18, No.2(2012), 143~156.

Laney, D., "3D data management: Controlling data volume, velocity and variety," META Group, 2001.

Lee, J. S., "Three-Step Probabilistic Model for Korean Morphological Analysis," *Journal of KIISE* Vol.38, No.5(2011), 257~268.

Lee, S., and H. Yoon, "The Study on Strategy of National Information for Electronic Government of S. Korea with Public Data analysed by the Application of Scenario Planning," *The Journal of The Korea Institute of Electronic Communication Sciences* Vol.7, No.6(2012), 1259~1273.

Lee, Y.-J, "A Semantic-Based Mashup Development Tool Supporting Various Open API Types," *Journal of Internet Computing and Services* Vol.13, No.3(2012), 115~126.

Levenshtein, V. I., "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet physics doklady*, Vol. 10, No. 8(1966), 707~710.

Lévy, P., *Collective intelligence*, Plenum/Harper Collins, 1997.

Lipsman, A., G. Mudd, M, Rich, and S. Bruich, "The power of "like": How brands reach (and influence) fans through social-media marketing," *Journal of Advertising research*, Vol. 52, No. 1(2012), 40.

Malone, T. W., R. Laubacher, and C. Dellarocas, "The collective intelligence genome," *IEEE Engineering Management Review*, Vol.38, No.3(2010), 21~31.

McAfee, A., and E. Brynjolfsson, "Big data: the management revolution," *Harvard business review*, Vol. 90, No.10(2012), 61~67.

Medelyan, O., and C. Legg, "Integrating Cyc and Wikipedia: Folksonomy meets rigorously defined common-sense," *Proceedings of the WIKI-AI: Wikipedia and AI Workshop at the AAAI'08 Conference*, Chicago, US, (2008).

Nasukawa, T., and J. Yi. "Sentiment analysis: Capturing favorability using natural language processing," *Proceedings of the 2nd international conference on Knowledge capture*, ACM, (2003), 70~77.

Ohkura, T., Y. Kiyota, and H. Nakagawa, "Browsing system for weblog articles based on automated folksonomy," *Proceedings of the WWW 2006 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics,* at WWW, Vol. 2006(2006).

Ohmukai, I., M. Hamasaki, and H. Takeda, "A proposal of community-based folksonomy with RDF metadata." *Proceedings of the 4th International Semantic Web Conference* (ISWC2005), (2005).

Pang, B., and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval* Vol.2, No.1-2(2008), 1~135.

Prentice, S., "CEO Advisory:'Big Data'Equals Big Opportunity," *Gartner*, March 31, 2011.

Russell, T., "Contextual authority tagging: Cognitive authority through folksonomy," *Unpublished manuscript. Retrieved,* Vol. 11, No.16(2005).

Sulis, W., "Fundamental concepts of collective intelligence," *Nonlinear Dynamics, Psychology, and Life* Science, Vol. 1, No.1(1997), 35~53.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational linguistics*, Vol. 37, No. 2(2011), 267~307.

Thomas, V. W., "Folksonomy," *online posting*, 2007.

Xu, Z., Y. Fu, J. Mao, and D. Su, "Towards the semantic web: Collaborative tag suggestions," *Collaborative web tagging workshop at WWW2006*, Edinburgh, Scotland, (2006).

Abstract

Building a Korean Sentiment Lexicon Using Collective Intelligence

Jungkook An* · Hee-Woong Kim**

Recently, emerging the notion of big data and social media has led us to enter data's big bang. Social networking services are widely used by people around the world, and they have become a part of major communication tools for all ages. Over the last decade, as online social networking sites become increasingly popular, companies tend to focus on advanced social media analysis for their marketing strategies. In addition to social media analysis, companies are mainly concerned about propagating of negative opinions on social networking sites such as Facebook and Twitter, as well as e-commerce sites. The effect of online word of mouth (WOM) such as product rating, product review, and product recommendations is very influential, and negative opinions have significant impact on product sales. This trend has increased researchers' attention to a natural language processing, such as a sentiment analysis. A sentiment analysis, also refers to as an opinion mining, is a process of identifying the polarity of subjective information and has been applied to various research and practical fields. However, there are obstacles lies when Korean language (Hangul) is used in a natural language processing because it is an agglutinative language with rich morphology pose problems. Therefore, there is a lack of Korean natural language processing resources such as a sentiment lexicon, and this has resulted in significant limitations for researchers and practitioners who are considering sentiment analysis. Our study builds a Korean sentiment lexicon with collective intelligence, and provides API (Application Programming Interface) service to open and share a sentiment lexicon data with the public (www.openhangul.com). For the pre-processing, we have created a Korean lexicon database with over 517,178 words and classified them into sentiment and non-sentiment words. In order to classify them, we first identified stop words which often quite likely to play a negative role in sentiment analysis and excluded them from our sentiment scoring. In general, sentiment words are nouns, adjectives, verbs, adverbs as they have sentimental

* Graduate School of Information, Yonsei University
** Corresponding author: Hee-Woong Kim
Graduate School of Information, Yonsei University
134 Shinchon, Seodaemun, Seoul 120-749, Korea
Tel: +82-2-2123-4195, E-mail: kimhw@yonsei.ac.kr

expressions such as positive, neutral, and negative. On the other hands, non-sentiment words are interjection, determiner, numeral, postposition, etc. as they generally have no sentimental expressions. To build a reliable sentiment lexicon, we have adopted a concept of collective intelligence as a model for crowdsourcing. In addition, a concept of folksonomy has been implemented in the process of taxonomy to help collective intelligence. In order to make up for an inherent weakness of folksonomy, we have adopted a majority rule by building a voting system. Participants, as voters were offered three voting options to choose from positivity, negativity, and neutrality, and the voting have been conducted on one of the largest social networking sites for college students in Korea. More than 35,000 votes have been made by college students in Korea, and we keep this voting system open by maintaining the project as a perpetual study. Besides, any change in the sentiment score of words can be an important observation because it enables us to keep track of temporal changes in Korean language as a natural language. Lastly, our study offers a RESTful, JSON based API service through a web platform to make easier support for users such as researchers, companies, and developers. Finally, our study makes important contributions to both research and practice. In terms of research, our Korean sentiment lexicon plays an important role as a resource for Korean natural language processing. In terms of practice, practitioners such as managers and marketers can implement sentiment analysis effectively by using Korean sentiment lexicon we built. Moreover, our study sheds new light on the value of folksonomy by combining collective intelligence, and we also expect to give a new direction and a new start to the development of Korean natural language processing.

Key Words : Sentiment Lexicon, Korean Natural Language Processing, Sentiment Analysis, Collective Intelligence, Folksonomy

Received : May 20, 2015 Revised : June 12, 2015 Accepted : June 13, 2015
Type of Submission : Fast Track Corresponding Author : Hee-Woong Kim

저 자 소 개

안 정 국
현재 연세대학교 정보대학원에서 박사과정 재학 중이며, 주요 연구분야는 Big Data Analytics, Natural Language Processing, Data Mining, Text Mining 등이다.

김 희 웅
National University of Singapore 정보시스템학과에서 근무 후, 현재 연세대학교 정보대학원 교수로 근무 중이다. 주요 관심분야는 디지털 비즈니스, 정보시스템 관리 및 활용 등이다. 관련 연구들은 MIS Quarterly, Information Systems Research, Journal of Management Information Systems, Journal of the Association for Information Systems, IEEE Transactions on Engineering Management, Journal of Retailing, European Journal of Operational Research, Communications of the ACM 등에 40여편의 논문이 게재되었다.