ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2016 March: 22(1): 83~105 http://dx.doi.org/10.13088/jiis.2016.22.1.083

주관적 웰빙 상태 측정을 위한 비정형 데이터의 상황기반 긍부정성 분석 방법*

최석재
경희대학교 빅데이터 연구센터
(sjchoi@khu.ac.kr)

송영은
경희대학교 일반대학원 경영학과
(dudeun13@khu.ac.kr)

권오병
경희대학교 경영학과
(obkwon@khu.ac.kr)

의료IT 서비스의 유망 분야인 정신건강 증진을 위한 주관적 웰빙 서비스(subjective well-being service) 구현의 핵심은 개인의 주관적 웰빙 상태를 정확하고 무구속적이며 비용 효율적으로 측정하는 것인데 이를 위해 보편적으로 사용되는 설문지에 의한 자기보고나 신체부착형 센서 기반의 측정 방법론은 정확성은 뛰어나나 비용효율성과 무구속성에 취약하다. 비용효율성과 무구속성을 보강하기 위한 온라인 텍스트 기반의 측정 방법은 사전에 준비된 감정어 어휘만을 사용함으로써 상황에 따라 감정어로 볼 수 있는 이른바 상황적 긍부정성(contextual polarity)을 고려하지 못하여 측정 정확도가 낮다. 한편 기존의 상황적 긍부정성을 활용한 감성분석으로는 주관적 웰빙 상태인 맥락에서의 감성분석을 할 수 있는 감정어휘 사전이나 온톨로지가 구축되어 있지 않다. 더구나 온톨로지 구축도 매우 노력이 소요되는 작업이다. 따라서 본 연구의 목적은 온라인상에 사용자의 의견이 표출된 비정형 텍스트로부터 주관적 웰빙과 관련한 상황감정어를 추출하고, 이를 근거로 상황적 긍부정성 파악의 정확도를 개선하는 방법을 제안하는 것이다. 기본 절차는 다음과 같다. 먼저 일반 감정어휘 사전을 준비한다. 본 연구에서는 가장 대표적인 디지털 감정어휘사전인 SentiWordNet을 사용하였다. 둘째, 정신건강지수를 동적으로 추정하는데 필요한 비정형 자료인 Corpora를 온라인 서베이로 확보하였다. 셋째, Corpora로부터 세 가지 종류의 자원을 확보하였다. 넷째, 자원을 입력변수로 하고 특정 정신건강 상태의 지수값을 종속변수로 하는 추론 모형을 구축하고 추론 규칙을 추출하였다. 마지막으로, 추론 규칙으로 정신건강 상태를 추론하였다. 본 연구는 감정을 분석함에 있어, 기존의 연구들과 달리 상황적 감정어를 적용하여 특정 도메인에 따라 다양한 감정 어휘를 파악할 수 있다는 점에서 독창성이 있다.

주제어 : 주관적 웰빙, 텍스트 마이닝, 감성분석, 상황적 긍부정성, 비정형데이터

논문접수일 : 2016년 2월 22일 논문수정일 : 2016년 3월 1일 게재확정일 : 2016년 3월 3일
원고유형 : 일반논문 교신저자 : 권오병

1. 서론

정신건강 증진을 위한 주관적 웰빙 서비스(subjective well-being service, SWB Service)는 의료 IT 분야의 하나로서 건강을 포함하여 자신의 삶의 질을 개선하고자 하는 욕구가 증가하는 현대인들을 대상으로 최근 활성화되기 시작했다. 여기서 주관적 웰빙(SWB)이란 개인이 그 동안

* 이 논문은 일부 2016년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구이며 (No.R0126-15-1007,맞춤형 개인 행복 증진을 위한 큐레이션 커머스용 글로벌 오픈 마켓 구축 기술개발), 또한 일부 2014년 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구(NRF-2014S1A5B8060940)되었음.

겪어온 경험 등에 근거하여 자신의 생애에 대한 주관적 평가를 뜻한다(Diener, 2009). 주관적 웰빙 서비스 구현의 전형적인 프로세스는 첫째 개인의 주관적 웰빙에 관련된 데이터를 수집하고, 둘째 주관적 웰빙 상태를 측정한 다음, 셋째 측정 결과에 따라 적절한 건강관리법을 추천하거나 서비스를 구동하는 것이다 (Qi et al., 2015). 따라서 수집된 데이터로 주관적 웰빙 상태를 정확하고 편리하게 측정하는 것이 매우 중요하다 (Kwon and Choi, 2011).

주관적 웰빙 상태를 측정하기 위해서 다양한 측정 방법 역시 제안되고 있는데, 대부분 직접 사용자가 작성한 설문 결과나 인체 부착 센서를 통해 얻어진 데이터에 의한 측정에 의존하고 있다. 그러나 설문지의 경우 적지 않은 항목에 대해서 사람이 직접 입력해야 하는데 사용자의 동적인 건강 상태를 파악하려는 경우 수시로 설문을 시행한다는 한계점을 지닌다. 한편, 부착 센서를 통한 접근법은 설문조사에 비해 더욱 정확하고 동적인 정보를 확보할 수 있다. 하지만 아직까지 개인 의료기기의 측정 가능 항목이 맥박수, 혈당, 혈압, 호흡수 등으로 제한되어 있으며 측정의 정확도를 보장받으려면 고가의 의료 장비를 사용해야 하는데 주관적 웰빙 서비스처럼 의료기관이 아닌 장소에서 서비스를 받아야 하는 경우에는 사용하기 불가능하다. 그리고 무엇보다도 위의 두 방법, 즉 설문지법과 센서 접근법은 구속적(obtrusive)이라는 문제점을 가지고 있다. 이에 무구속적(unobtrusive)으로 처리하기 위한 대안으로 연령이나 성별 등과 같은 사용자 프로필 데이터를 가지고 데이터 마이닝 등의 방법을 사용하여 측정하는 방법이 있다. 하지만 이 방법도 사용자의 주관적 웰빙 상태를 파악하기에는 정보가 불충분하여 결국 측정 정확도를 개선하기 어렵다는 한계점을 가진다. 따라서 사용자 프로필 데이터뿐 아니라 사용자가 무구속 상태에서 제공한 정보를 추가할 수 있다면 측정 정확도를 개선함과 동시에 효율적이고 동적인 측정이 가능한 방법이 될 것이다.

사용자가 무구속 상태로 생성하는 데이터로는 개인이 공개한 블로그나 SNS 등을 통해 만들어지는 비정형데이터인 텍스트가 있다. 이를 가지고 텍스트마이닝 기법을 활용한다면 그 텍스트에 담겨진 글쓴이의 감정이나 어떤 대상에 대한 긍부정성을 추론할 수 있어 관련 연구가 진행 중이며, Affective Norms for English Words (ANEW)와 Facebook Gross National Happiness (FGNH), The Positive and Negative Affect Schedule (PANAS) 등은 그 중 가장 대표적인 시도이다 (Dodds and Danforth, 2010; Dodds et al., 2011; Watson et al., 1988; Qi et al., 2015). 특히 Qi et al. (2015)의 연구에서는 사용자들 사이의 온라인 대화 속에서 남긴 텍스트를 가지고 그 텍스트에 담긴 개인의 감정(emotion) 상태를 계산하는 방법을 제안하였는데, 계산의 주요 근거는 감정어의 등장 횟수이며, 이를 위해 사전에 감정어 사전을 확보하였다. 감정어를 사전에 확보하는 것은 감성분석의 필수 요건이다 (Dodds and Danforth, 2010; Kim and Kim, 2014). 최근 텍스트마이닝 분야에서는 상황적으로 긍부정성(contextual polarity)을 띄는 감정어를 추출하는 연구를 진행하여 감정어 목록의 풍부성을 꾀하고 상황적 긍부정성의 모호성을 극복(contextual polarity disambiguate)하려는 연구도 진행 중이다 (Agarwal et al., 2015).

하지만 위의 텍스트 기반 웰빙 상태 연구는 중요한 한계점이 두 가지 있는데 그것은 첫째, 주관적 웰빙에서 사용한 감정이 증오, 사랑, 즐거움, 슬픔, 분노, 기대, 걱정, 놀람의 여덟 가지로 경직되어 있다는 것이다. 주관적 웰빙 서비스 활

용으로는 스트레스라든가 우울 등 다른 응용 영역이 있음에도 불구하고 정해진 여덟 가지 감정만으로는 이러한 요인을 측정할 수 없다. 두 번째 한계점으로는 사전에 구축된 감정어휘사전 또는 상식수준(common-sense)의 상황적 감정어휘에 등록된 단어만 활용했다는 것이다(Liu and Singh, 2004; Havasi et al., 2007). 특히 특정 단어가 감정어휘가 될지 안될지는 상황에 따라 다르다. 예를 들어 “구직”이라는 단어는 기존 감정어휘 사전에는 등록될 수 없는 단어이고 일반적 관점에서도 감정어라고 볼 수 없으나, 스트레스 측정 관점에서는 중요한 단서가 될 수 있다. 이렇게 관점에 따라 특정 주관적 웰빙 분야에서 감정어가 되기도 하고 아니기도 하는 단어를 ‘주관적 웰빙 요인 상황감정어(SWB element related contextual polarity)’라고 칭할 수 있다. 즉, 원래 감정어를 나타내는 어휘는 아니지만 주관적 웰빙 관점에서 감정어로 분류할 수 있는 어휘를 말한다. 하지만 상황감정어를 고려한 감성분석을 통해 스트레스나 우울 등과 같은 주관적 웰빙 상태를 텍스트 기반으로 측정하는 연구는 존재하지 않았다.

따라서 본 연구의 목적은 온라인 상에 사용자의 생각이 공개적으로 표출된 비정형 텍스트로부터 각 주관적 웰빙 요인별 상황 감정어를 추출하고, 이를 근거로 주관적 웰빙 상태 측정을 개선하는 방법을 제안하는 것이다. 본 연구는 감정을 분석함에 있어, 일관된 규칙 적용 가능성만을 고려하는 기존의 연구들과 달리 상황적 감정어를 적용하여 특정 도메인에 따라 다양한 감정 어휘를 파악할 수 있다는 점에 의의가 있다. 본 연구에서 제안한 방법으로 개인 맞춤형으로 주관적 웰빙 상태 측정의 정확도를 높인다면 삶의 질 개선에 도움이 되는 서비스를 개발할 수 있을 것이다.

2. 문헌연구

2.1. 주관적 웰빙 측정

주관적 웰빙(SWB)이란 개인이 그 동안 겪어온 경험 등에 근거하여 자신의 생애에 대한 주관적 평가를 뜻한다 (Diener, 2009). 이러한 주관적 웰빙의 구성 요소는 삶에 대한 만족, 긍정적 정서, 부정적 정서로 구성되며(Diener, 1984; Diener et al.,1999), 특히 정서적 측면에 의미를 가지고 있다 (Kang, 2015). 따라서 주관적 웰빙 상태를 측정하기 위해서 개인의 정서상태를 파악하려는 정신건강 측정 방법들이 다양하게 연구되고 있다. <Table 1>에는 주관적 웰빙 측정을 위한 개인의 정서상태 측정 방법과 특징을 정리하였다.

주관적 웰빙 측정의 전통적 방법인 직접 설문은 제시된 측정 척도에 대하여 응답자가 일정 시

〈Table 1〉 Conventional Methods of Measuring SWB

Method	Property
Questionnaire	Self-report about one’s intra-state for the measurement questionnaire in the traditional way
Figure and Image	Compare to the questionnaire semantic measure is possible, but having difficulties about measuring the dynamic state of emotion
User Profile	Measuring method using the methods such as data mining with a user life-log data such as the age, sex and medical history
Sensors	Measurement is made using a variety of user data sensor via wearable sensors such as smartphone. It can be monitored in real-time measurement.

점 동안의 정서 경험 정도에 대해 응답하는 방식이다 (Park et al., 2012b). 주로 스트레스, 우울, 불안, 분노 등 부정적인 감정에 대한 문항에서 지난 일주일을 떠올리며 자신이 가지고 있는 감정에 대해 '예' 혹은 '아니오'로 응답하게 되어 있다 (Kim et al., 2015). 그리고 심리적 상태를 측정한 일반 건강 설문지(The General Health Questionnaire, GHQ) 지표는 응답자가 일정기간 (예: 지난 2~3 주간) 동안의 변화된 심리적 상태를 일정기간 이전의 상태와 비교하여 현재 상태의 문제점을 파악하도록 고안되었다 (Shin, 2001; Park et al., 2012b). 하지만 이러한 측정 방법은 기억을 회상해야 하기 때문에 정서의 변화를 모호하게 만들거나 잘못 해석하게 된다(Jang et al., 2007). 기억 과정에서 오는 오류를 최소화하기 위해서, 매 순간마다의 상황에 대한 정서를 조사하는 방법인 경험표집방법(Experience Sampling Method, ESM)과 전날 하루 일상에 대한 보고를 하고 각 활동에 대한 정서경험을 평가하는 방법인 일상재구성법(Day Reconstruction Method, DRM)을 사용할 수 있다(Park et al., 2012a; Jang et al., 2007). 하지만, 경험표집방법과 일상재구성법 설문 측정 방법은 주로 심층적으로 연구하기 위한 목적으로 사용되며 적지 않은 항목 문항을 가지고 있어 응답자에게 부담도 매우 큰 방법이다. 그리고 직접 설문 측정의 특성상 사람이 직접 입력해야 하는 불편함을 감수해야 하고, 정적인 특징을 가지고 있기 때문에 수시로 설문으로 사용자의 상태를 파악해야 하며, 이는 비용이 많이 들여야 한다는 한계점과 구속적이라는 문제점을 지닌다.

2.2. 감성 컴퓨팅 및 상황적 감정어

텍스트 기반으로 감성 분석하는 이른바 감성 컴퓨팅(Sentic Computing) 분야의 연구들은 일반적으로 특정 대상에 대한 견해를 파악하는 것이 주된 관심이기 때문에 감성의 긍정과 부정의 양극성의 강도를 표현하고 있는 어휘를 인식하여 감성 사전을 구축한다 (Cambria, 2016). 대표적인 예로 SenticNet 은 개념 수준에서의 감성 분석을 위해 공개된 RDF/XML 포맷의 감성 단어 목록으로 다중 단어로 구축된 표현방법도 감정어휘로 포함되는 것이 특징이다. 최근 개발된 SenticNet 에는 30,000 종류의 개념(concept)에 대한 감성값을 등재하고 있다. 자연어 표현 중에서 감성적 극성(polarity)를 띄는 어휘를 추출하기란 쉬운 일이 아니다. 그래서 SenticNet 의 경우 상식 수준의 온톨로지가 구축되면 이를 연결함으로써 감정어휘 목록을 확장할 수 있게 보완했다 (Cambria et al., 2014). 또한 SentiWordNet (Esuli and Sebastiani, 2006)은 <Table 2>와 같이 형태소(part of speech tag)와 해당 단어에 대한 극성값이 부가된 10 만개 이상의 영어 단어들로 구성되어 있으며 텍스트 파일의 형태로 제공되고 있다. 이렇게 구축된 SentiWordNet 은 뉴질랜드 고유어 (Medagoda et al., 2015), 한국어 (Choi and Kwon, 2014) 및 베트남어 (Vu et al., 2014) 버전의 SentiWordNet 구축을 위한 원천으로도 사용되고 있다. 이외에 Affective Lexicon (Ortony et al., 1990), linguistic annotation scheme (Wiebe et al., 2005), WordNet-Affect (Strapparava and Valitutti, 2004) 등이 널리 활용되는 감정어휘 사전들이다. <Table 2>는 감정어휘사전인 SentiWordNet 의 구조를 보여준다.

이러한 감정어휘 사전들은 궁극적으로 각 개인별로 차이가 나는 감성 표현 및 그 극성값을 정확히 인식하는 데 사용된다. 이를 위해서는 상황적 긍부정성을 고려해야 한다. 상황적 긍부정성(contextual polarity)은 사전에 정의된 또는 context-free 상태에서의 감정어와는 달리 글의

〈Table 2〉 Examples of Emotional Dictionary : SentiWordNet

	POS	Sense	Positive	objectivity	Negative	Example
1	Adjective	short#1	0	1	0	a short story
		short#2	0	1	0	short skirts
		short#3	0.125	0.125	0.75	short in stature
	Adverb	short#6	0	0.75	0.25	I was caught short
	Verb	short#2	0	1	0	create a short circuit in
2	Adjective	long#1	0	1	0	a long time
		long#3	0.25	0.625	0.125	looked out the long French windows
		long#9	0.125	0.5	0.375	long on brains
	Adverb	long#1	0	1	0	a promotion long overdue

문맥에 의하여 특정 단어가 긍부정성을 가지는 것을 말한다. 그러나 사실 문맥을 사전에 정의하기란 어려운 일이다. 따라서 온톨로지 또는 의미망(semantic net)을 활용하여 일정 문맥에 대한 도메인을 공급자가 결정하고 그 후에 이 문맥에서 등장하는 감정어를 상황적 감정어로 인식하는 방법이 제안되고 있다 (Agarwal et al., 2015). 이때 온톨로지 또는 의미망이 얼마나 일반적인지가 이슈가 될 수 있다. 일반적인 의미망 구축을 위해서 고려할 수 있는 것이 상식(common sense)에 대한 의미망을 사전에 구축하여 사용하는 것인데, ConceptNet은 일반인들에 의하여 인터넷 상에서 협력적으로 구축된 의미망으로서 상식에 의한 의미망의 대표적인 형태라고 할 수 있다 (Liu and Singh, 2004; Havasi et al., 2007).

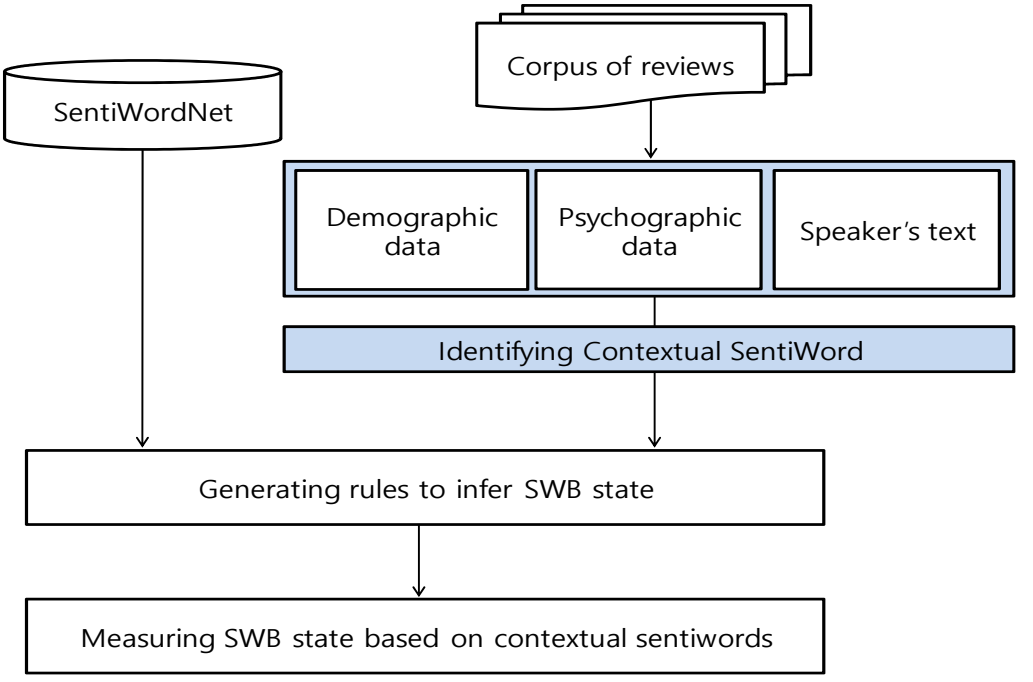
그러나 지금까지의 노력과 제안 방법은 긍부정성, 즉 극성값(polarity)만을 적용했을 뿐이며 주관적 웰빙과 같은 특정 심리상태에서의 긍부정성을 판별하지는 못한다. 예를 들어 스트레스 관점에서의 긍부정성과 우울 관점 내지는 피로감 관점에서의 긍부정성은 다르다. 하지만 주관적 웰빙은 Common sense를 다루는 ConceptNet에도 아직 구축되어 있지 못하여 자동적으로 감성분석을 할 수 없는 단계이다. 따라서 contextual polarity는 다시 이러한 심리상태에 독립적인 문맥과 특정 심리상태에 종속적인 문맥으로 다시 나뉘어 더욱 정교하게 분석되어야 한다.

3. 상황적 감정어 구축을 통한 주관적 웰빙상태 측정

3.1.전체적인 프로세스

본 연구에서 제안하는 주관적 웰빙 관련 상황적 감정어를 추출하여 텍스트 작성자의 주관적 웰빙상태를 측정하는 방법으로 <Figure 1>과 같은 프로세스를 제안한다. 먼저 일반 감정어휘 사전을 준비한다. 본 연구에서는 가장 대표적인 디지털 감정어휘 사전인 SentiWordNet을 사용하기로 한다. SentiWordNet은 극성값이 부가된 감성사전이다 (Esuli and Sebastiani, 2006). SentiWord Net은 POS 태그 된 단어 명사, 동사, 형용사, 부사의 네 개 품사, 약 2백만 단어에 대한 극성값이 구성되어 있다.

둘째, 정신건강지수를 동적으로 추정하는데 필요한 비정형 자료인 Corpora를 확보한다. 이를



〈Figure 1〉 Proposed System Framework

위해 온라인 서베이를 할 수 있다.

셋째, Corpora로부터 세 가지 종류의 자원을 확보한다. 먼저 인구동태적 정보(예: 성별, 나이, 학력, 소득수준)와 심리동태적 정보(예: 평소의 심리 상태)가 화자의 감성에 영향을 주는지를 파악한다. 두 번째로 화자가 제공한 텍스트의 구조적인 특징(예: 글의 길이, 특수 문자의 사용 유무)이 영향을 주는지 확인한다. 다음으로 각 정신건강 요소(예: 스트레스, 우울, 피로, 분노, 불행감 등)의 각 수준(정상, 비정상)별로 상대적으로 빈번하게 등장하는 단어를 추출하여 상황적 감정어로 인식한다.

넷째, 자원을 입력변수로 하고 특정 정신건강 상태의 지수값을 종속변수로 하는 추론 모형을 구축하고 추론 규칙을 추출한다. 다섯째, 추론 규칙으로 정신건강 상태를 추론한다.

3.2. 주관적 웰빙 관련 상황적 감정어 사전 구축

3.2.1. 기본 감정어휘 사전 구축

먼저 일반 감정어 사전을 준비한다. 일반 감정어는 현존하는 감정어휘사전을 활용한다. 그러나 현재 한글에서는 잘 구축된 감정어휘사전이 없으므로 세계적으로 널리 쓰이는 영문 감정어휘사전인 SentiWordNet을 활용한다. SentiWordNet에는 3만여 어휘에 대한 극성값이 할당되어 있다. 이를 구글번역기를 활용하여 한글 단어로 번역한다. 이때 두 개 이상의 SentiWordNet 포함 단어가 동일한 한글로 번역되어 중의성 문제가 발생할 수 있다. 중의성 문제는 기계가 자연어를 온전히 이해하기 어렵게 하는 부분으로서 해결이 필요하다 (Jeong and Park, 2015). 본 연구에서

〈Table 3〉 Examples of Reviewer's Corpus

No.	Reviews about Unpaid Wages	Reviews about Car Insurance
1	저 소식이 안타깝고 이런 현실이 답답하다. (It is a very frustrating reality to hear that news)	어이없다. 경기가 어렵다고 하는데도…아닌 것 같다. (Confused. Didn't they hear their argument? They say economic situation is very difficult. It is not right)
2	안타깝다. 나도 힘들지만 그내들도 많이 힘들것같다.… 더욱 맘이 아플것같다. (Regretful. I am also very hard but they are also seem to be tough… It would be very hurting for them)	열 받는다. 당연히 보험료를 차등적으로 적용해야 한다. (Very upset. Of course, They have to be applied differential insurance premium)
3	서민만 힘듦. (Only the common people is hard)	보험료 개편 필요. (This insurance premium policy should be reformed)
4	안타까운 내용이지만 주변에서 흔히 접할 수 있는 것 같다. (it seems a pity, but frequently we can encounter this situation)	하루 빨리 개선되어야 한다고 봄. 특히 외제차 바가지 수리비. (It should be reformed in a short time. Especially for the foreign car's repair costs)
5	실컷 부려먹고도 몇달째 급여를 안주고 있다니..머하는 사장인지~~보험비용도 천차만별~~ (He exploits the employee but they do not give the payment for a few months. What he does!)	보험비용도 천차만별 (The insurance cost varies widely)
6	우리 사회가 안보이고 … 관심을 가져야 되겠습니다. (Our society is not seen… We should be concerned for this)	애국으로 국산차를 이용하는데..너무 사회적 격차가 다양한 곳에서 발생하네요..!! (I use a domestic car for a patriotic reason… But social gap takes places in various places…!!)

는 한글로 번역된 단어를 다시 구글번역기를 통해 영어로 번역하여 등장하는 단어의 극성값을 취한다. 이렇게 하는 이유는 구글번역기가 그 한글의 단어에 대한 대표 의미를 가지는 영어 단어를 내기 때문이다. 이렇게 한 다음 사용자가 발화한 텍스트에 해당 단어들의 등장한 정도에 의하여 감성값을 도출한다.

3.2.2. 상황적 감정어 추출을 위한 말뭉치 확보

온라인 뉴스 기사에 대한 댓글은 어떤 사안에 대해 일반인들의 감정 상태, 특히 본 연구에서 관심을 가지는 스트레스나 우울, 분노감을 자발적이고 솔직하며 객관적으로 표출하는 가장 대표적인 공개 자료이다. 이에 착안하여 주관적 웰빙 관련 상황적 감정어 추출을 위한 말뭉치 확보를 위해 온라인 뉴스 기사를 복수 개 선정하고 이에 대한 일반인들의 댓글을 확보한다. 다만 온라인 상에 올라있는 일반의 댓글에는 실제로 그 사람이 느낀 스트레스나 우울, 분노감의 정도에 대한 평가를 하지 않았기 때문에 이 부분은 온라인 설문을 실시한다. 즉, 온라인 설문 상에서 특정 기사를 노출시킴으로써 해당 기사에 대해 댓글을 쓰게 하고 본인이 댓글을 작성할 때의 주관적 웰빙 값(특히 스트레스, 우울, 분노 상태)을 5점 척도로 받았다. 이렇게 하여 댓글 말뭉치 및 각 댓글에 대한 주관적 웰빙 값들을 수집할 수 있다. 댓글 말뭉치를 확보한 예는 <Table 3>과 같다. 단, 오탈자와 띄어쓰기 오류는 수정하지 않고 원문 그대로 표기했다.

3.2.3. 전처리

문장을 각각의 최소 형태의 단어로 분리하기

〈Table 4〉 Results by Classification Index

Morphlogy	Frequency	POS	Emotional Degree	Number of Reviews	Freq./Number of Reviews
가- (go)	6	Aux. predicate	5	297	0.020202
가정 (family)	3	Common Noun	5	297	0.010101
가족 (family)	9	Common Noun	5	297	0.030303
가지- (take)	3	Common Verb	5	297	0.010101
갑 (higher position)	4	Common Noun	5	297	0.013468

위해 형태소 분석기인 RHINO 2.0을 이용하였다. 형태소 분석 단계에서 상황적 감정어 처리와 무관한 품사들(예: 전치사 등)과 관련 품사(명사, 동사)들 중에서도 '나, 우리, 하다, 되다'와 같이 stop words 에 속하는 단어들은 사전에 제거한다.

3.2.4. 상황적 감정어의 인식과 극성 계산

상황적 감정어 사전을 구축하기에 앞서 이전 절의 말뭉치 구성과 형태소 분석을 토대로 어휘 정보를 추출하고, 추출된 어휘 정보의 결과들을 테이블로 만들었다. 이는 <Table 4>와 같이 형태소, 빈도, 품사, 감정정도, 댓글수, 빈도/댓글수로 구성되어 있다. 여기서 동일한 형태소와 품사에 대한 '빈도 대비 댓글 수'의 비율(%)이 가장 크게 나온 값(MLE 기준)을 추출하면 1차 단어사전이 만들어진다. 다음으로 두 명의 인코더가 각각 감정어휘 여부를 판단하였다. 서로 만장일치할 경우에는 상황적 감정어로 선별하였고, 한 명의 인코더만 감정어휘라고 본 경우는 보류하여 상황적 감정어 저장소(inventory)에 저장하였다. 이로써 스트레스, 우울, 정신적 피로, 분노 네 가지의 상황적 감정어 저장소와 각각의 지수가 조합되어 구축되었다.

3.3. 말뭉치를 활용한 인과관계 도출

정신상태에 관련하여 구축한 말뭉치를 활용하여 정신건강 상태를 종속변수로 하는 인과모형을 개발한다. 말뭉치에는 다음과 같은 정보가 존재한다. 첫째, 인구통계학적 정보(예: 성별, 나이, 학력, 소득수준), 둘째, 심리통계학적 정보(예: 평소의 심리 상태), 셋째, 화자가 제공한 텍스트의 구조적인 특징(예: 글의 길이, 특수 문자의 사용 유무)

위의 요소들을 독립변수로 하고 회귀분석을 수행하여 인과모형을 개발한다. 실제로 특정 기사에 대해서 사람들이 느끼는 감정에 대한 설문조사한 결과를 말뭉치로 구축하였다. 설문 조사는 2015년 11월에 실시하였으며 온라인 서베이 기관의 패널 중에서 1,605명으로부터 유효응답을 받은 것이다. 이에 스트레스, 우울, 피로에 대해서 다중회귀분석을 수행하였다. 그 결과 <Table 5>에서 나타난 대로 4대 정신건강요소인 스트레스, 우울, 분노, 피로감에 대해 영향을 주는 요인은 대체로 성별, 연령, 소득, 학력과 같은 인구통계학적 요인과 텍스트의 길이, 그리고 본인의 심리상태에 대한 심리통계적 자료이며, 그 외 문장 부호 등(느낌표, 이모티콘의 사용 등)은 정신건강요소 값에 영향을 주지 않는 것으로 나

〈Table 5〉 Results of Regression Analysis

	Model 1	Model2	Model3	Model 4
dependent variable	Stress	Depression	Anger	Fatigue
(const)	3.751**	3.327**	3.888**	2.684**
gender	-.164**	.028	-.056	-.113*
age	.002	.003	-.001	.009**
income	-.050**	-.044**	-.035	-.034
scholar	-.101**	-.034	-.098**	-.072*
length	.004**	.003**	.003**	.002**
quest	.124**	.147**	.110**	.183**
emoticon	-.016	.039	.007	-.077
exclamation mark	-.066	-.001	.144	-.232
question mark	.080	-.085	.032	-.044
ending	-.051	.017	-.010	-.080
dot	.026	-.025	-.031	-.053
semicolon	.408	-.446	-.071	-.502
F-Value	9.988**	8,466**	6.144**	8.045**
Adjusted R2	.063	.053	.037	.050

Note1) * p<0.05, ** p<0.15
Note2) 'quest' indicates the level of SWB state by self-report right before being exposed by the online news to upload reply

타났다. 따라서 문장 부호는 stop word에 포함하여 처리해도 무방하다는 결론을 도출하였다.

도출한 결과를 요약하면 성별, 연령대, 소득수준, 학력은 상황적 감정어의 결정에 영향을 주는 요인이며 댓글의 길이와 질문형 문장은 화자의 감성 정도를 예측하는 데 유의한 정보이다. 반면 이모티콘의 활용여부 등 나머지 특성은 화자의 감성 정도를 예측하는데 유의하게 사용되지 않는다.

3.4. 상황적 감정어 구축

본 논문에서 의미하는 상황은 개인의 인구통계학적 특성(예: 성별, 나이, 학력, 소득수준), 심리통계학적 특성(예: 평소의 심리 상태) 및 화자가 제공한 텍스트의 주제 등이며 이 세 가지 상황에 따라 어떤 단어는 감정어가 되기도 하고 되지 않기도 하는 것이 상황적 감정어다.

상황적 감정어는 다시 주제(주관적 웰빙 요인)에 독립적인 상황적 감정어와 주제(주관적 웰빙 요인)에 종속적인 상황적 감정어로 나뉜다. 주제(주관적 웰빙 요인)에 독립적인 상황적 감정어란 상황적 감정어 중에서 여러 주관적 웰빙 요인에 공통적으로 출현하는 것이다. 예를 들어 '힘들다', '못된' 등은 어떤 주관적 웰빙 요인에서도 등장이 가능한 상황적 감정어다. 이에 비해 주제(주관적 웰빙 요인)에 종속적인 상황적 감정어는 주관적 웰빙 요인에 따라서 어떤 경우는 감정어 어떤 주관적 웰빙 요인에서는 감정어가 되기 어려운 경우이다. 예를 들어 우울에 대한 주관적 웰빙 요인에서는 '무전유죄', '불평등', '불편' 등의 단어가 감정어가 되나 스트레스라는 웰빙 분류에서는 '악덕 (기업주)', '임금체불' 등이 감정어가 될 수 있다. 상황적 감정어 여부를 파악하기 위하여 각 텍스트를 다음과 같이 분류하였다.

텍스트 = <일련번호, 댓글 내용, 스트레스 수준, 우울감 수준, 분노심 수준, 불행감 수준, 피로감 수준, 성별, 연령, 소득수준, 학력, 단어의 수, 의문문 유무 >

텍스트로부터 상황적 감정어 추출은 coder들을 통해 작업할 수 있다. 만약 3명의 coder를 사용한다면, 그들 각자에게 댓글의 긍부정성 파악에 도움이 되는 단어들을 추출하였다. 그리고 난 후에 단어들이 일치하는지 점검하고 일치하지 않는 것들에 대해서는 다시 감정어인지를 각자 다시 판단하게 한다. 2차 검사 때에도 감정어로 인정하기 어려운 경우 감정어 목록에서 제외한다. 이러한 순서를 통해 파악된 감정어는 형태소 분석을 통하여 어간 정보로 변환될 수 있다.

어간 형식으로 변환된 상황적 감정어는 두 주제(또는 기사)에 대해서 공통적으로 나타난 경우 주관적 웰빙 요인독립적 상황 감정어, 한 쪽의 주제에서만 등장한 경우 주관적 웰빙 요인종속적 감정어로 판단한다. 따라서 상황적 감정어에는 다음과 같은 분류 태그가 적용되었다.

상황적 감정어 = <감정어, 돗수, 주관적 웰빙 요인독립성, 주관적 웰빙 요인>

이렇게 하여 115가지의 주관적 웰빙 요인 독립적 감정어와 두 가지의 주제, 곧 주제1 및 주제2에 대한 각각 333개, 445개의 주관적 웰빙 요인 종속적 감정어 등 총893개의 상황적 감정어가 확보되었다. <Table 6>은 확보된 상황적 감정어의 저장 예이다.

3.5. 상황적 감정어 기반의 주관적 웰빙상태 측정

상황적 감정어 사전이 구축되면 기본적인 감정어 사전 (예: SentiWordNet, HowNet 등)과 더불어 감성분석을 수행한다. 이에 다음과 같은 프로세스를 반영한 프로그램을 구축하였다.

단계1: 테스트용 댓글을 입력
단계2: 댓글에 대해 형태소 분석
단계3: 감정어 사전 및 유사감정어 사전에 속한 단어와 일치하는 단어 검색. 검색이 되면 그 단어의 감성값(긍정 1, 부정 -1)을 댓글의 긍부정성 합에 합산

〈Table 6〉 Examples of Contextual Polarity Words

Sentiword	Frequency	Independency of SWB Type	Related SWB Type
답답 (Hidebound)	35	Independent	Stress, Depression
도대체 (Why)	4	Independent	Stress, Depression
……..			
갑질 (Power trip)	2	Dependent	Stress
사업주 (Boss)	8	Dependent	Stress
……..			
형평성 (Equity)	19	Dependent	Depression
차별 (Irrespective)	6	Dependent	Depression
……..			

단계4: 긍부정성 합을 댓글의 단어 수로 나눈 후 그 값을 Sigmoid function에 입력하여 난 결과값이 일정 상한선 수준 이상이면 긍정적 글로 판정하고, 일정 하한선 수준 이하이면 부정적 글로 판정함. N개의 단어를 가진 댓글의 긍부정성 값인 x는 각각 기본적인 감정어 사전을 기반으로 한 j번째 단어의 감성값 σ_j^B과 주관적 웰빙 요인독립적 상황 감정어 사전을 기반으로 한 j번째 단어의 감성값 σ_j^I, 그리고 주관적 웰빙 요인종속적 상황 감정어 사전을 기반으로 한 j번째 단어의 감성값 σ_j^D인의 weighted average로 구해진다. 한편 C는 텍스트 및 사용자 프로필에 의하여 결정되는 상수이며, C를 결정하는 데에는 글 작성자의 성별이나 나이, 소득 수준과 같은 사용자 프로필과 글의 길이(단어 수)와 의문문 사용여부가 사용된다.

$$x = \frac{\sum_{j=1}^{N}\left(\beta_B\sigma_j^B + \beta_I\sigma_j^I + \beta_D\sigma_j^D\right)}{N} + C$$

단, $0 \le \beta_B, \beta_I, \beta_D \le 1$ 이며 $\beta_B + \beta_I + \beta_D = 1$ 이다.

$$y = \frac{1}{1 + e^{-\alpha x}}$$

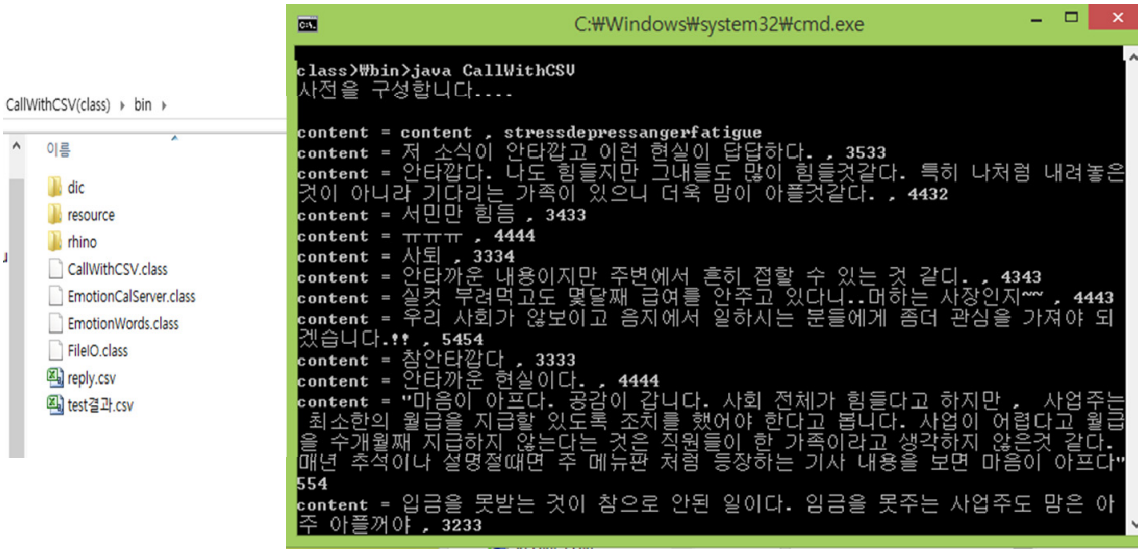
$$\text{if} \begin{cases} y \ge \theta_U, & p(y) = 1 \\ y \le \theta_L, & p(y) = -1 \\ \text{otherwise}, & p(y) = 0 \end{cases}$$

$$0 \le \theta_L \le \theta_U \le 1$$

4. 실험 및 결과

4.1. 검증 방법

본 연구에서 제안한 방법의 감성분석 성능을 분석하기 위해 다음과 같은 방법들과 비교하였다.



〈Figure 2〉 SWB State Inference System

방법1) 기존 감정어휘사전인 SentiWordNet을 활용한 감성분석
방법2) 주관적 웰빙 요인 독립적 상황적 감정어를 추가한 감성분석 (SentiWordNet + topic independent sentiwords)
방법3) 주관적 웰빙 요인 종속적 상황적 감정어도 추가한 감성분석 (SentiWordNet + topic independent sentiwords + topic dependent sentiwords)

SentiWordNet에서 감정어를 추출하는 것은 프로그램을 응용하였으며, 전체적인 감성분석은 SentiWordNet 개발팀(http://sentiwordnet.isti.cnr.it)이 공개한 오픈 소스를 기반으로 하여 Java application으로 별도 개발하였다. 개발된 프로그램으로 감성분석을 실시하는 예제 화면은 <Figure 2>와 같다. 한편 성능 비교를 위해 사용되는 측정치는 전반적인 정확도(overall accuracy)를 사용하였다.

4.2. 자료 수집

실험에 사용할 온라인 기사 및 그 댓글에 대해서 다음과 같은 방법으로 수집하였다. 먼저 상황적 감정어를 구축하기 위해 두 가지의 전혀 다른 주제에 대해서 특정 기사를 수집하고 데이터 수집기관의 성년 패널들 1,632명을 대상으로 이에 대하여 본인의 감정을 표현하는 댓글을 온라인상으로 작성하게 하였다. 댓글 작성 후에는 자신의 감성 상태를 다차원적으로 기록하게 하였다. 이는 감성을 긍정성/부정성으로 단순화하여 표현하는 것보다 다차원적인 특성을 가지기 때문이다. 이에 스트레스, 우울감, 불행감, 분노심, 피로감의 다섯 가지 차원으로 각각 댓글을 작성할 때의 상황을 5점 척도(1: 전혀 없음, 5: 매우 심각함)로 질문하였다. 이를 통해 주제1과 주제2에 대해서 각각 1,632개의 댓글을 다음과 같이 수집하였다. 단어수는 주제1(스트레스) data set의 경우 12,966단어, 주제2(우울감) data set의 경우 11,104단어로 구성되어 있다. 이 중에서 상황적 감정어 추출을 위해 사용한 댓글은 각각 1,430개이며, 나머지 댓글들은 테스트용으로 활용하였다. 최종적으로 구축된 상황적 감정어 목록의 예는 Appendix A와 같다.

4.3. 결과

테스트용 댓글 집합은 매 번마다 전체 집합에서 30%를 랜덤하게 선택하여 생성하였으며 이를 120회 반복 실험하였다. 판정은 긍정, 중립, 부정으로 하였으며 따라서 랜덤하게 판정할 경우 33.3%의 정확도를 기대할 수 있다.

세 방법의 성능 분석 결과 다음 <Table 7>과 같은 결과가 나왔다. 이를 통계적으로 확인하기 위하여 pairwise T-Test를 수행한 결과 <Table 8>과 같이 방법3 > 방법2 > 방법1의 순서로 좋은 성능이 나오는 것으로 파악되었다. 결론적으로 본 연구에서 제안한 방법인 상황을 고려하지 않

〈Table 7〉 Comparison Result of Performance

Methods	Theme 1 (related to Stress)	Theme 2 (related to Depression)
Method 1	50.88 (5.59)	41.53 (2.05)
Method 2	65.93 (5.15)	61.66 (1.83)
Method 3	82.31 (4.22)	81.77 (1.51)

은(out of context) 사전에 정의된 감정어 목록에 주관적 웰빙 요인 독립적 상황적 감정어와 주관적 웰빙 요인 종속적 상황적 감정어를 모두 고려한 감성분석이 통계적으로 유의하게 가장 우수한 긍부정 분석 정확도를 보였다.

5. 토의 및 결론

5.1. 학술적 의의

본 연구의 의의는 먼저 화자의 감정 상태 파악에 상황적 감정어를 적극적으로 사용하여 실험 결과의 성공률을 높였다는 점이다. 한국어의 감정어 목록에 관하여는 기존에 연구된 바가 있었으나 그 수가 500개를 넘지 못하는 한계가 있었다 (Ahn et al., 1993; Park, 2001; Gim, 2004). 목록의 수가 매우 제한적이었기 때문에 자신의 감정을 매우 직접적으로 제시하는 일부의 경우 외에는 실제 사용된 문장에서는 감정어가 잘 발견되지 않는 문제가 발생하였다. 그러나 본 연구에서는 상황적 감정어를 감정 판정에 도입함으로써 화자의 감정을 판별할 가능성을 크게 높였다. 상황적 감정어는 일반적인 감정어처럼 언어학적 조건을 일관되게 유지하지는 못하지만 대부분의 상황에서 감정을 간접적으로 표현하는 데 사용되어 활용도는 매우 높다. 이들의 활용 가능성에 주목하고 감정 판정에 적극적으로 도입한 것이 본 연구의 첫 번째 차별점이다.

둘째, 상황적 감정어의 구성 방법을 제안한 점이다. 감정은 기본적으로 정의를 내리기가 매우 어려운 속성을 가지고 있다 (Fehr and Russell, 1984). 따라서 일반적인 감정어는 물론 상황적 감정어는 아직 그 목록이 확정되지 않았으며, 선정을 위한 방법도 아직 확정되지 않았다. 접속사의 성질을 이용하는 방법(Hatzivassiloglou and McKewon, 1997), 연어 관계에 있는 어휘를 이용하는 방법(Turney and Littman, 2003), WordNet의 속성을 이용하는 방법(Kamps et al., 2004; Baccianella et al., 2010) 등 아직까지는 다양한 방법을 시도하고 있는 추세이다. 본 연구에서는 설문조사로 댓글을 쓰게 함으로써 상황적 감정어의 후보들을 얻는 방법을 택하였다. 원하는 상황에 관련된 텍스트를 일반 화자에게 제시하고, 그에 대한 댓글을 확보함으로써 해당 상황에서 쓰일 수 있는 단어의 집합들을 얻었다. 그리고 복수의 coder가 판정하여 상황적 감정어 목록을 얻었다. 이 과정에서 상황과 관계 없이 항상 감정을 전달한다고 여겨지는 것과 특정 상황에서만 감정을 전달하는 것을 구분하여 주관적 웰빙 요인 독립적 감정어와 주관적 웰빙 요인 종속적 감정어도 분류하였다.

5.2 실무적 의의

실무적 관점, 특히 주관적 웰빙 서비스 구현과

〈Table 8〉 Comparison Result of Pairwise T-Test

	Method 1	Method 2
Method 2	15.05 (t=21.69***) 20.13 (t=80.19***)	0 0
Method 3	31.43 (t=49.11***) 40.24 (t=173.29***)	16.38 (t=26.95***) 20.11 (t=92.84***)

Note1) *p<0.01, **p<0.001, ***p<0.0001
Note2) Numbers in the parenthesis indicates the t-value

관련되어 본 연구는 다음과 같은 의의를 갖는다. 첫째, 주관적 웰빙 서비스를 구현하고자 할 때 획득해야 할 기초 정보에 대해 밝힌 점이다. 성별과 연령에 따라 느끼는 감정의 정도가 달라진다는 연구 (Diener et al., 1985)나 성별과 연령이 글의 문체 영향을 준다는 연구 (Schler et al., 2006)가 있었지만 본 연구에서는 성별과 연령은 물론 소득수준, 학력 등의 개인 정보도 감정 및 감정 어휘의 사용에 영향을 준다는 점을 발견하였다. 또한 사용자가 작성한 댓글의 길이와 문장의 종류는 감정을 예측하는 데 유의하다는 것을 알 수 있었다. 따라서 주관적 웰빙 상태의 측정에는 개인의 선천적인 정보는 물론, 후천적인 정보와 텍스트의 형식적인 정보도 같이 수집해야 정밀한 예측이 가능하다는 것을 알 수 있다.

둘째, 주관적 웰빙 상태 측정을 위한 무구속적 방안을 제시하였다는 점이다. 기존의 방법은 매번 사용자에게 많은 항목에 걸친 설문조사를 하거나(Christensen et al., 2003; Kahneman et al., 2004) 안면 인식, 심장 박동 인식과 같은 센서를 이용하는 방법이었다(Yasunari et al., 2000; Sommerer and Laurent, 2011). 하지만 침실, 욕실, 작업실 등 일상의 다양한 공간에서 웰빙 상태의 측정 및 서비스가 이루어지기 위해서는 즉시적인 측정이 이루어져야 하고, 그러려면 의료기관에서나 사용할 수 있는 고가의 센서는 활용하기가 어렵다. 본 연구에서 제안한 방안은 사용자의 발화를 인식하는 것만으로 웰빙 상태의 측정이 가능하다. 스마트폰에 입력되는 발화는 물론, 일상생활에서도 음성인식기만 갖추면 사용된 단어를 분석하여 웰빙 상태를 측정할 수 있다. 따라서 사용자로서는 측정이 이루어지고 있다는 것을 거의 의식하지 않게 된다.

5.3. 연구의 한계점

본 연구의 결과가 실제 서비스로 이어지기 위해서는 해결해야 할 몇 가지 과제가 남아 있다. 먼저 고려되어야 할 상황 도메인이 결정되어야 한다. 상황 감정어는 기본적으로 해당 상황에서만 감정어로 기능하는 것이므로 다른 상황에서는 감정어로서의 기능을 보장할 수 없다. 따라서 인지되어야 할 상황이 미리 결정되고, 해당 상황에서 쓰일 수 감정어의 목록이 확보되어야 한다. 그런데 상황이라고 하는 것은 매우 다양하므로 하나의 서비스에서 모든 상황에 대한 준비를 할 수는 없다. 인지되어야 할 상황을 결정하는 논리적 판단 방법이 필요하다.

둘째, 상황 감정어를 판정하는 자동화된 방안이 필요하다. 상황 도메인이 결정되었다고 하더라도 그 수가 매우 많으면 본 연구에서 취한 방법과 같이 일일이 사람의 판단으로 감정어를 판정하기란 어려운 일이다. 본 연구에서는 두 가지 상황을 염두에 두었기 때문에 가능했지만, 실제에 있어서는 많은 상황에 대한 대처가 필요할 것이기 때문에 보다 자동화된 방법으로 감정어를 판정하고, 후에 복수의 coder 합의로 보정하는 방법이 필요하다. 상황적 감정어는 고도의 인지능력을 필요로 하는 일이므로 이 작업이 쉬운 일은 아니겠지만 실용화를 위해서는 필요한 부분이다.

5.3. 결론

본 논문은 개방된 비정형 데이터를 통해서 주관적 웰빙 상태를 자동 측정할 수 있는 텍스트 분석 방법을 제안하는 것이었다. 이를 위해 상식기반의 감정어만을 사용하지 않고 사전적으로는 감정어가 아니지만 주관적 웰빙 상태와 관련된 글에서는 감정어로 변경되는 상황적 감정어를

추가적으로 고려했을 때 감성분석의 정확도가 현격하게 개선되는 것을 실험을 통해 입증하였다. 본 연구는 상황적 극성값을 지니는 감정어를 개발하려는 감성 컴퓨팅 분야의 일부 시도에 부합하는 결과를 보인 것이며 학술적으로도 또한 실무적으로도 중요한 의미를 지님을 확인하였다. 이에 본 논문에서는 가장 관심이 높은 주관적 웰빙 상태인 스트레스와 우울감의 두 가지만으로 입증하였으나 추후 더 다양한 주관적 웰빙 상태에 대한 상황정 감정어를 구축하는 것이 필요하다고 본다.

참고문헌(References)

Agarwal, B., N. Mittal, P. Bansal, and S. Garg, "Sentiment Analysis Using Common-Sense and Context Information," *Computational Intelligence and Neuroscience*, Vol.2015(2015), Article ID 715730, 1~9.

Ahn, S.H., S.H. Lee, and O.S. Kwon, "Activation Dimension : A Mirage in the Affective Space?," *The Korean Journal of Social and Personality Psychology*, Vol.7, No.1(1993), 107~123.

Baccianella, S., A. Esuli and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," *Proceedings of the 7th Conference on International Language Resources and Evaluation(LREC)*, Vol.10(2010), 2200~2204.

Cambria, E., D. Olsher, and D. Rajagopal, "SenticNet 3: A Common and Common-Sense Knowledge Base for Cognition-Driven Sentiment Analysis," *Twenty-eighth AAAI Conference on Artificial Intelligence*, (2014), 1515~1521.

Cambria., E., "Affective Computing and Sentiment Analysis," *IEEE Intelligent Systems*, Vol.31, No.2(2016), 1~7.

Choi, S. and O. Kwon, "The Study of Developing Korean SentiWordNet for Big Data Analytics: Focusing on Anger Emotion," *Journal of Society for e-Business Studies*, Vol.19, No.4 (2014), 1~19.

Christensen, T.C., L.F. Barrett, E. Bliss-Moreau, K. Lebo, and C. Kaschub, "A Practical Guide to Experience-Sampling Procedures," *Journal of Happiness Studies*, Vol.4, No.1(2003), 53~78.

Diener, E., "Subjective Well-Being," *Psychological Bulletin*, Vol.95, No.3(1984), 542~575.

Diener, E., *The Science of Well-Being*, Springer Netherlands, 2009.

Diener, E., E. M. Suh, R. E. Lucas, and H.L. Smith, "Subjective Well-Being: Three Decades of Progress," *Psychological Bulletin*, Vol.125, No.2(1999), 276~302.

Diener, E., E. Sandvik, and R.J. Larsen, "Age and Sex Effects for Emotional Intensity," *Developmental Psychology*, Vol.21, No.3(1985), 542~546.

Dodds, P. S., K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth, "Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter," *PLOS One*, Vol.6, No.12:e26752 (2011), 1~26.

Dodds, P.S. and C.M. Danforth, "Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents," *Journal of Happiness Studies*, Vol.11, No.4(2010), 441~456.

Esuli, A. and F. Sebastiani, "SentiWordNet: A Publicly Available Lexical Resource for

Opinion Mining," *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Vol.6(2006), 417~422.

Fehr, B. and J.A. Russell, "Concept of Emotion Viewed from a Prototype Perspective," *Journal of Experimental Psychology*: *General,* Vol.113, No.3(1984), 464~486.

Gim, E. Y., *A Study on the Korean Emotion*, PhD Thesis, Chonnam National University, 2004.

Hatzivassiloglou, V. and K. R. McKeown, "Predicting the Semantic Orientation of Adjectives," *Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics*, (1997), 174~181.

Havasi, C., R. Speer, and J. Alonso, "ConceptNet 3: A Flexible, Multilingual Semantic Network for Common Sense Knowledge," *Recent Advances in Natural Language Processing*, (2007), 27~29.

Jang, J. Y., K. Ryu, E. K. Suh, and I. C. Choi, "Quality of Life of Working Men, Women, and Housewives Measured by the Day Reconstruction Method (DRM)," *Korean Journal of Social and Personality Psychology*, Vol.21, No.2(2007), 123~139.

Jeong, H. J. and B. H. Park, "Korean Word Sense Disambiguation using Dictionary and Corpus," *Journal of Intelligent Information Systems*, Vol.21, No.1(2015), 1~13.

Kahneman, D., A. B. Krueger, D. A. Schkade, N. Schwarz, and A. A. Stone, "A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method," *Science*, Vol.306, No.5702(2004), 1776~1780.

Kamps, J., M. J. Marx, R. J. Mokken, and M. D. Rijke, "Using WordNet to Measure Semantic Orientation of Adjectives," *Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation*, Vol.4(2004), 1115~1118.

Kang, S.P., *The Effects of Self-Leadership on Psychological, Subjective Well-being: Perceived Organizational Justice A Moderator,* Master's Thesis, Chosun National University, 2015.

Kim, S., E.H. Lee, S.T. Hwang, S.H. Hong, and K. Lee, "Reliability and Validity of the Korean Version of the Beck Hopelessness Scale," *Journal of Korean Neuropsychiatric Association*, Vol.54, No.1(2015), 84~90.

Kim, S.W. and N.G. Kim, "A Study on the Effect of Using Sentiment Lexicon in Opinion Classification," *Journal of Intelligence and Information Systems*, Vol.20, No.1(2014), 133~148.

Kwon, O.B. and S.J. Choi, "A Methodology of Measuring Degree of Contextual Subjective Well-Being Using Affective Predicates for Mental Health Aware Service," *Journal of Intelligence and Information Systems*, Vol.17, No.3(2011), 1~23.

Liu, H. and P. Singh, "ConceptNet — a Practical Commonsense Reasoning Tool-Kit," *BT Technology Journal*, Vol.22, No.4(2004), 211~226.

Ortony, A., G. L. Clore, and A. Collins, *The Cognitive Structure of Emotions,* Cambridge University Press, 1990.

Park, J. E., S. J. Shim, and H. G. Lee, "The Method of Measuring Subjective Quality of Life," *Daejeon Statistical Research Institute*, (2012a), 143~214.

Park, J. I., Y. J. Kim, and M. J. Cho, "Factor Structure of the 12-item General Health Questionnaire in the Korean General Adult Population," *Journal of Korean Neuropsychiatric Association*, Vol.51(2012b), 178~184.

Park, I. J., *The Analysis of Korean Affective Terms: Listing Affective Terms and Exploring Dimensions in the Affective Terms*, PhD Thesis, Seoul National University, 2001.

Qi, J., X. Fu, and G. Zhu, "Subjective Well-Being Measurement Based on Chinese Grassroots Blog Text Sentiment Analysis," *Information & Management*, Vol.52, No.7(2015), 859~869.

Schler, J., M. Koppel, S. Argamon, and J. Pennebaker, "Effects of Age and Gender on Blogging," *Proceedings of the 2006 AAAI spring symposium*, Vol.6(2006), 199~205.

Shin, S. I., "The Validity and Reliability of the Korean Version of the General Health Questionnaire: KGHQ-20 & KGHQ-30," *Korean Journal of Social Welfare*, Vol.46(2001), 210~230.

Sommerer, C. and M. Laurent, "Mobile Feelings-Wireless Communication of Heartbeat and Breath for Mobile Art," in *The Mobile Audience Media Art and Mobile Technologies*, M. Rieser(eds.), Rodopi Publications, 2011, 271~275.

Strapparava, C. and A. Valitutti, "WordNet-Affect: An Affective Extension of WordNet," *Language Resources and Evaluation,* Vol.4 (2004), 1083~1086.

Turney, P.D. and M.T. Littman, "Measuring Praise and Criticism: Inference of Semantic Orientation from Association," *ACM Transactions on Information Systems*, Vol.21, No.4(2003), 315~346.

Vu, X. S., H. J. Song, and S. B. Park, "Building a Vietnamese SentiWordNet using Vietnamese Electronic Dictionary and String Kernel," *13th Pacific Rim Knowledge Acquisition Workshop*, (2014), 223~235.

Watson, D., L. A. Clark, and A. Tellegen, "Development and Validation of Brief Measures of Positive and Negative Affect: the PANAS Scales," *Journal of Personality and Social Psychology*, Vol.54, No.6(1988), 1063~1070.

Medagoda, N., S. Shanmuganathan, and J. Whalley, "Sentiment Lexicon Construction Using SentiWordNet 3.0," *Proceedings of the 11th International Conference on Natural Computation*, (2015), 802~807.

Wiebe, J., T. Wilson, and C. Cardie, "Annotating Expressions of Opinions and Emotions in Language," *Language Resources and Evaluation*, Vol.39, No.2(2005), 165~210.

Yasunari, Y., S. Kim., T. Kawano, and T. Kilazoe, "Effect of Sensor Fusion for Recognition of Emotional States Using Voice, Face Image and Thermal Image of Face," *Proceedings of the 2000 IEEE International Workshop on Robot and Human Interactive Communication*, (2000), 178~183.

Appendix A. 상황적 감정어 추출 예

단 아래에 색으로 표시한 부분은 스트레스 관련 글과 우울 관련 글 모두에 동시에 등장하는 주관적 웰빙 요인독립적인 상황적 감정어이며, 그 외의 것은 주관적 웰빙 요인종속적인 상황적 감정어이다.

주제 1	감정어	중요도
임금체불	--	1
임금체불	!!	2
임금체불	…..	1
임금체불	각박	1
임금체불	강제	1
임금체불	개선	1
임금체불	거참	1
임금체불	걱정	6
임금체불	겠네요	1
임금체불	고생	3
임금체불	고소	1
임금체불	공감	1
임금체불	관심	1
임금체불	근로	1
임금체불	급여	11
임금체불	긴급	1
임금체불	끔찍	1
임금체불	나쁜	2
임금체불	너무	3
임금체불	노동	1
임금체불	놈	1
임금체불	답답	5
임금체불	당하다	1
임금체불	대책	1
임금체불	도대체	1
임금체불	막막	1
임금체불	말도안됨	1
임금체불	망해	1
임금체불	멱살	1
임금체불	모양	1
임금체불	몰지각	1
임금체불	못	3
임금체불	무겁	1
임금체불	무슨	1
임금체불	무엇	1

주제 2	감정어	중요도
자동차보험료	--	1
자동차보험료	!!	3
자동차보험료	~	2
자동차보험료	…	17
자동차보험료	가진자	4
자동차보험료	각성	1
자동차보험료	갑	3
자동차보험료	개떡	1
자동차보험료	개선	2
자동차보험료	개한민국	1
자동차보험료	게다가	1
자동차보험료	공정치	1
자동차보험료	괘씸	1
자동차보험료	그렇지	1
자동차보험료	꼴	1
자동차보험료	너무	1
자동차보험료	놈	3
자동차보험료	단념	1
자동차보험료	답답	2
자동차보험료	답없	1
자동차보험료	당연히	1
자동차보험료	당하다	1
자동차보험료	도대체	1
자동차보험료	돈	3
자동차보험료	등골	1
자동차보험료	따위	1
자동차보험료	따져	1
자동차보험료	때려	1
자동차보험료	땜빵	1
자동차보험료	말이안됨	2
자동차보험료	망정	1
자동차보험료	무전유죄	1
자동차보험료	문제	7
자동차보험료	뭐	1
자동차보험료	미친	1

주제 1	감정어	중요도
임금체불	문제	3
임금체불	뭐	3
임금체불	미안	1
임금체불	밀린급여	1
임금체불	부당	2
임금체불	부디	1
임금체불	불공평	1
임금체불	불만	2
임금체불	불행	1
임금체불	불황	1
임금체불	빈부격차	2
임금체불	빡빡	1
임금체불	빨리	1
임금체불	사태	1
임금체불	상실	1
임금체불	슬프	2
임금체불	시급	2
임금체불	시켜	1
임금체불	신고	6
임금체불	실망	1
임금체불	실컷	1
임금체불	심하	1
임금체불	쓰레기	1
임금체불	쓸쓸	3
임금체불	아닐까	2
임금체불	아프	10
임금체불	악덕	2
임금체불	안된	4
임금체불	안타깝	32
임금체불	않나요?	1
임금체불	양심	1
임금체불	어려워	1
임금체불	언제까지	1
임금체불	오죽	1
임금체불	왜	1
임금체불	요망	1
임금체불	우울	3
임금체불	원망	1
임금체불	월급	4

주제 2	감정어	중요도
자동차보험료	바보	1
자동차보험료	봉	7
자동차보험료	부당	5
자동차보험료	부유층	1
자동차보험료	분노	2
자동차보험료	분통	2
자동차보험료	불공평	5
자동차보험료	불쌍	1
자동차보험료	불쾌	1
자동차보험료	불편	1
자동차보험료	불평등	1
자동차보험료	불합리	5
자동차보험료	빈부	1
자동차보험료	빈부격차	3
자동차보험료	빈익빈	4
자동차보험료	빨리	1
자동차보험료	사회격차	1
자동차보험료	속상	1
자동차보험료	시정	2
자동차보험료	심하	1
자동차보험료	쓰레기	1
자동차보험료	씁쓸	1
자동차보험료	아닐까	1
자동차보험료	안된	1
자동차보험료	암담	1
자동차보험료	약자	1
자동차보험료	어긋나	1
자동차보험료	어이	3
자동차보험료	억울	3
자동차보험료	언제까지	1
자동차보험료	에효	1
자동차보험료	열	3
자동차보험료	옛말	1
자동차보험료	왜	4
자동차보험료	울리	1
자동차보험료	유감	1
자동차보험료	유전무죄	1
자동차보험료	이러니	3
자동차보험료	이런	1

주제 1	감정어	중요도
임금체불	이게	1
임금체불	인가?	1
임금체불	인간적	1
임금체불	임금	6
임금체불	임금체불	10
임금체불	잘못	1
임금체불	재정난	1
임금체불	저러다	1
임금체불	적어도	1
임금체불	절망	1
임금체불	조차	2
임금체불	좋	2
임금체불	직장	1
임금체불	참다	1
임금체불	처벌	3
임금체불	체불	6
임금체불	최소한	2
임금체불	취직	1
임금체불	커녕	1
임금체불	피곤	1
임금체불	하락	1
임금체불	해결	1
임금체불	헬조선	3
임금체불	현실	2
임금체불	화	3
임금체불	횡포	1
임금체불	힘	1
임금체불	힘내	4
임금체불	힘들	19
임금체불	ㅜ	1
임금체불	ㅠ ㅠ	6

주제 2	감정어	중요도
자동차보험료	이모양	1
자동차보험료	이상	2
자동차보험료	있죠	1
자동차보험료	잘나다	1
자동차보험료	잘못	1
자동차보험료	재검토	1
자동차보험료	정말	1
자동차보험료	정의	1
자동차보험료	제대로	1
자동차보험료	제발	1
자동차보험료	조장	1
자동차보험료	좀	1
자동차보험료	죽	1
자동차보험료	짜증	4
자동차보험료	차별	1
자동차보험료	ㅋㅋㅋ	1
자동차보험료	피해	1
자동차보험료	필요	4
자동차보험료	하루이틀	2
자동차보험료	한심	1
자동차보험료	해결	1
자동차보험료	헐	2
자동차보험료	헬조선	4
자동차보험료	형평성	9
자동차보험료	호구	1
자동차보험료	홀대	1
자동차보험료	화	9
자동차보험료	황당	2
자동차보험료	횡포	3
자동차보험료	힘들	2
자동차보험료	ㅠ ㅠ	1

Abstract

Analyzing Contextual Polarity of Unstructured Data for Measuring Subjective Well-Being

Sukjae Choi* · Yeongeun Song** · Ohbyung Kwon***

Measuring an individual's subjective wellbeing in an accurate, unobtrusive, and cost-effective manner is a core success factor of the wellbeing support system, which is a type of medical IT service. However, measurements with a self-report questionnaire and wearable sensors are cost-intensive and obtrusive when the wellbeing support system should be running in real-time, despite being very accurate. Recently, reasoning the state of subjective wellbeing with conventional sentiment analysis and unstructured data has been proposed as an alternative to resolve the drawbacks of the self-report questionnaire and wearable sensors. However, this approach does not consider contextual polarity, which results in lower measurement accuracy. Moreover, there is no sentimental word net or ontology for the subjective wellbeing area. Hence, this paper proposes a method to extract keywords and their contextual polarity representing the subjective wellbeing state from the unstructured text in online websites in order to improve the reasoning accuracy of the sentiment analysis.

The proposed method is as follows. First, a set of general sentimental words is proposed. SentiWordNet was adopted; this is the most widely used dictionary and contains about 100,000 words such as nouns, verbs, adjectives, and adverbs with polarities from -1.0 (extremely negative) to 1.0 (extremely positive). Second, corpora on subjective wellbeing (SWB corpora) were obtained by crawling online text. A survey was conducted to prepare a learning dataset that includes an individual's opinion and the level of self-report wellness, such as stress and depression. The participants were asked to respond with their feelings about online news on two topics. Next, three data sources were extracted from the SWB corpora: demographic information, psychographic information, and the structural characteristics of the text (e.g., the number of words used in the text, simple statistics on the special characters used). These were considered to adjust the level of a specific SWB. Finally, a set of reasoning rules was generated for each wellbeing

* Humanitas BigData Research Center, Kyung Hee University
** School of Management, Kyung Hee University
*** Corresponding author: Ohbyung Kwon
School of Management, Kyung Hee University
26 Kyungheedae-ro, Dongdaemun-gu, Seoul 130-701, Korea
Tel: +82-2-961-2148, Fax: +82-2-961-0515, E-mail: obkwon@khu.ac.kr

factor to estimate the SWB of an individual based on the text written by the individual.

The experimental results suggested that using contextual polarity for each SWB factor (e.g., stress, depression) significantly improved the estimation accuracy compared to conventional sentiment analysis methods incorporating SentiWordNet. Even though literature is available on Korean sentiment analysis, such studies only used only a limited set of sentimental words. Due to the small number of words, many sentences are overlooked and ignored when estimating the level of sentiment. However, the proposed method can identify multiple sentiment-neutral words as sentiment words in the context of a specific SWB factor. The results also suggest that a specific type of senti-word dictionary containing contextual polarity needs to be constructed along with a dictionary based on common sense such as SenticNet. These efforts will enrich and enlarge the application area of sentic computing.

The study is helpful to practitioners and managers of wellness services in that a couple of characteristics of unstructured text have been identified for improving SWB. Consistent with the literature, the results showed that the gender and age affect the SWB state when the individual is exposed to an identical queue from the online text. In addition, the length of the textual response and usage pattern of special characters were found to indicate the individual's SWB. These imply that better SWB measurement should involve collecting the textual structure and the individual's demographic conditions. In the future, the proposed method should be improved by automated identification of the contextual polarity in order to enlarge the vocabulary in a cost-effective manner.

Key Words : Subjective Well-Being, Text mining, Sentiment Analysis, Contextual Polarity, Unstructured Data

Received : February 22, 2016 Revised : March 1, 2016 Accepted : March 3, 2016
Publication Type : Regular Paper Corresponding Author : Ohbyung Kwon

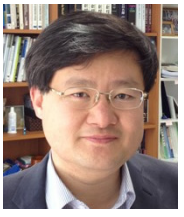
저 자 소 개

최 석 재

현재 경희대학교 후마니타스 빅데이터 연구센터 전임연구원으로 재직 중이다. 고려대학교 국어국문학과에서 국어학으로 학사, 석사, 박사 학위를 취득하였고, 미국 Carnegie Mellon Univ. Language Technologies Institute 객원연구원, 중국 연변과학기술대학 언어공학연구소 실장, 고려대학교 BK21 연구교수, 성신여자대학교 국어국문학과 초빙교수 등을 거쳤다. 한국어 정보를 전산화하는 방안과 비정형 빅데이터의 분석에 관심이 있다.

송 영 은

현재 경희대학교 일반대학원 경영학과에서 빅데이터 경영전공 석사과정에 재학 중이며, 목원대학교 정보컨설팅학과에서 학사학위를 취득한 바 있다. 그리고 Caitech연구소에서 행복지수 프로젝트에 참여 중이다. 주요 관심 분야는 텍스트 마이닝에 기반한 빅데이터 분석, 감성 분석이다.



권 오 병

현재 경희대학교 경영학과 교수로 재직 중이다. 1988년 서울대학교 경영학과 (경영학사), 1990년 한국과학기술원 경영과학과 (공학석사), 1995년 한국과학기술원 경영과학과 (공학박사)를 졸업하였다. 2001년~2002년에는 카네기멜론대학 전산학부에서 방문과학자로 근무한 바 있으며 2009년~2011년에는 샌디에고주립대학 경영정보학과의 겸직교수로 재직한 바 있다. 관심분야는 빅데이터분석, 사물인터넷, 의사결정지원시스템 등이다.