ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2015 June: 21(2): 69~92 http://dx.doi.org/10.13088/jiis.2015.21.2.69

텍스트 마이닝을 이용한 감정 유발 요인 'Emotion Trigger'에 관한 연구*

안주영
연세대학교 문과대학 문헌정보학과, 주저자
(juyoung228@gmail.com)

배정환
연세대학교 문과대학 문헌정보학과, 공동저자
(haruaki.pn@gmail.com)

한남기
연세대학교 문과대학 문헌정보학과, 공동저자
(hng88@naver.com)

송 민
연세대학교 문과대학 문헌정보학과 부교수, 교신저자
(min.song@yonsei.ac.kr)

최근 소셜 미디어의 사용이 폭발적으로 증가함에 따라 이용자가 직접 생성하는 방대한 데이터를 분석하기 위한 다양한 텍스트 마이닝(text mining) 기법들에 대한 연구가 활발히 이루어지고 있다. 이에 따라 텍스트 분석을 위한 알고리듬(algorithm)의 정확도와 수준 역시 높아지고 있으나, 특히 감성 분석(sentimental analysis)의 영역에서 언어의 문법적 요소만을 적용하는데 그쳐 화용론적·의미론적 요소를 고려하지 못한다는 한계를 지닌다. 본 연구는 이러한 한계를 보완하기 위해 기존의 알고리듬 보다 의미 자질을 폭 넓게 고려할 수 있는 Word2Vec 기법을 적용하였다. 또한 한국어 품사 중 형용사를 감정을 표현하는 '감정어휘'로 분류하고, Word2Vec 모델을 통해 추출된 감정어휘의 연관어 중 명사를 해당 감정을 유발하는 요인이라고 정의하여 이 전체 과정을 'Emotion Trigger'라 명명하였다. 본 연구는 사례 연구(case study)로 사회적 이슈가 된 세 직업군(교수, 검사, 의사)의 특정 사건들을 연구 대상으로 선정하고, 이 사건들에 대한 대중들의 인식에 대해 분석하고자 한다. 특정 사건들에 대한 일반 여론과 직접적으로 표출된 개인 의견 모두를 고려하기 위하여 뉴스(news), 블로그(blog), 트위터(twitter)를 데이터 수집 대상으로 선정하였고, 수집된 데이터는 유의미한 연구 결과를 보여줄 수 있을 정도로 그 규모가 크며, 추후 다양한 연구가 가능한 시계열(time series) 데이터이다. 본 연구의 의의는 키워드(keyword)간의 관계를 밝힘에 있어, 기존 감성 분석의 한계를 극복하기 위해 Word2Vec 기법을 적용하여 의미론적 요소를 결합했다는 점이다. 그 과정에서 감정을 유발하는 Emotion Trigger를 찾아낼 수 있었으며, 이는 사회적 이슈에 대한 일반 대중의 반응을 파악하고, 그 원인을 찾아 사회적 문제를 해결하는데 도움이 될 수 있을 것이다.

주제어 : 감정 유발 요인, Word2Vec, 감성분석, 텍스트 마이닝, 소셜 이슈

논문접수일 : 2015년 6월 5일 논문수정일 : 2015년 6월 17일 게재확정일 : 2015년 6월 18일
투고유형 : 국문급행 교신저자 : 송민

1. 서론

전 세계적으로 인터넷 상의 데이터가 급증함에 따라, 사람들의 의견이 표출된 텍스트로부터 특정 감정을 추출하고자 하는 연구가 활발히 진행되었다(Liu, 2010; Narayanan et al., 2009; Sadamitsu et al., 2008). 이에 따라 기술의 정확도와 수준 역시 성장해 왔으나, 대부분의 경우 알고리듬을 적용한 양적 접근이며, 이러한 접근법은 언어의 문법적 요소만을 다루므로 화용론적·

* 이 논문은 2012년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2012-2012S1A3A2033291)

의미론적 요소는 고려하지 못하고, 같은 문장 혹은 문단 내 특정 단어의 표면적 의미를 파악하는 수준의 분석만 이루어지고 있다는 한계를 지닌다. 본 연구는 이러한 한계를 보완하기 위해 기존의 알고리듬보다 의미 자질을 폭 넓게 고려할 수 있는 Word2Vec(Mikolov et al., 2013) 기법을 적용하였다.

또한 이를 위해 한국어 품사 분류 체계를 활용하여 '감정어휘'를 추출하였다. 감정어휘란 어떤 대상이나 일 또는 현상에 대해 느끼게 되는 기분의 상태를 나타내는 어휘를 말하는데(Jang, 2001), 본 연구에서는 이러한 감정어휘로 형용사를 선정하였다. 이후 Word2Vec 모델을 통해 추출된 감정어휘의 연관어 중 명사를 해당 감정을 유발하는 요소로 분류하고 이를 'Emotion Trigger'라 명명하였다. 이는 한국어 품사를 이용하여 어휘 간의 관계를 추론하는 방식으로, 감성분석이 활발하게 이루어지고 있는 상품평 분석 방식에서 차용한 방법이다(Song and Lee, 2011).

본 연구는 문헌 내에서 나타나는 특정 키워드와 관련된 감정을 분석함에 있어, 단순히 감정 상태만을 알아내는 기존의 연구들과 달리 Word2Vec 기법을 적용하여 특정 감정과 관련된 다른 요인들을 파악할 수 있다는 점에 그 의의가 있다. 여기에는 새로운 기법뿐만이 아니라 언어학적 요소 역시 결합되어 있으며, 그 과정에서 특정 감정을 유발하는 'Emotion Trigger'를 찾아낼 수 있었다. 본 연구에서는 사례 연구로 특정 사건들에 대한 뉴스, 블로그, 트위터 데이터만을 사용하였으나, 제시한 방법론은 특정 문헌의 종류에 한정되지 않기 때문에 대중들의 의견이 표출된 텍스트 데이터라면 무엇이든 대중들의 감정을 파악하고, 그 원인을 구체적으로 분석하는 데 이용할 수 있을 것이다.

2. 선행 연구

텍스트를 이용한 감성 분석은 알고리듬, 기계학습 등을 이용한 언어학적 방식과 감성 사전 등을 활용한 의미 분석 방식으로 나뉜다. 국내에서는 언어학적 방식의 감성 분석 연구가 활발히 이루어져 왔다(Kim and Seo, 2013; Kim et al., 2011; Lee et al., 2013; Hong et al., 2014; Kang et al., 2012). 이러한 언어학적 방식의 감성 분석에는 해당 텍스트 데이터의 품사나 어휘 기능이 이용되기도 하였다. Kim and Lee (2014)은 어휘에 대한 기능적, 감성적 정보를 이용하여 어휘 기능을 나타내는 특징을 추출한 후 추출된 특징을 은닉 마르코프 모델(Hidden Markov Model, HMM)에 적용하여 특징의 순서를 고려한 감성분석을 수행하였다. Kang et al. (2009)은 상품 리뷰에 대해 품사들의 조합과 긍·부정의 감성 레이블로 구성된 "패턴 DB"를 구축한 후 비교를 통해 상품 리뷰의 감성을 판단하였다. 이러한 연구들은 언어의 심층적 요소를 고려했다는 점에서 의의가 있으나, 한글이 아닌 영어 데이터를 사용한 것이 그 한계라 할 수 있다. 한글은 그 특성이 영어와 달라 분석이 어려우며 텍스트 분석의 대상이 되는 인터넷 공간상에서 사용되는 언어는 정제된 한글이 아니다. 따라서 한글에 특화된 분석이 이루어질 필요가 있다.

한편, 의미론적 접근 방법을 사용하여 감성 사전을 구축하는 감성 분석의 경우, 영어 텍스트를 기반으로 한 SentiWordNet 활용 연구들이 활발히 이루어져 왔다(Hamouda and Rohaim, 2011; Hung and Lin, 2013; Ohana and Tierney, 2009; Saggionα and Funk, 2010). 국내에서도 한글의 여러 문법적· 의미적 특성을 고려하여 감성 사전을 구축하려는 시도가 활발히 이루어져 왔다. Choi

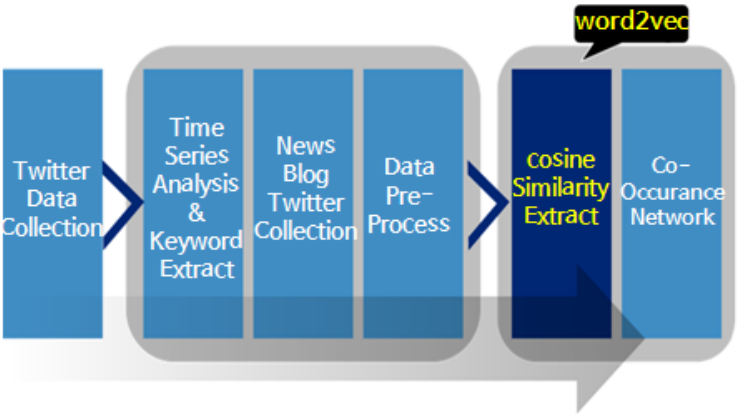
and Kwon (2014)은 문장에서 드러나는 감정 어휘를 발견하고, 이들의 정도값을 결정하여 한국어 SentiWordNet을 개발하고자 하였다. Seo et al. (2015) 은 한국 문법의 반의어 규칙을 적용한 오피니언 감성 사전을 설계하였으며, Jang et al. (2015)은 한국어의 문법적 요소 이외에도 인터넷 언어에서 사용되는 이모티콘, 특수 기호, 한글 초성의 감정 기호 등을 추출하여 인터넷 감정 기호 사전을 구축하였다. 이러한 최근의 여러 연구에도 불구하고, 한글의 경우, 외국에 비해 사전 구축에 대한 연구가 미흡하고 사전도 미흡한 실정이다(Seo and Ko, 2014).

3. 연구 설계

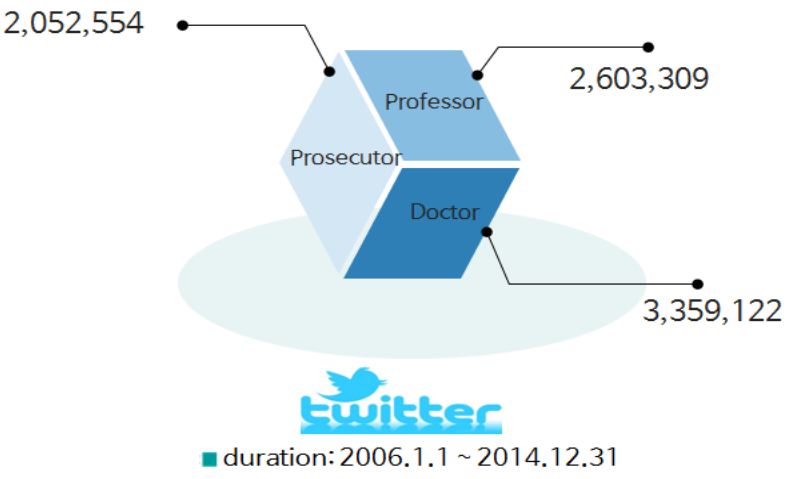
3.1. 연구 모형

본 연구는 모든 텍스트 데이터를 대상으로 분석을 수행하는 것이 이상적인 연구 방법이나, 시간과 자원의 한계를 고려하여 특정 주제를 선정하고 이에 대한 사례 연구를 진행하였다. 데이터 수집에서 가장 우선적으로 고려한 사항은, 대중들의 의견과 감정이 텍스트에 명확히 표현되는 주제여야 한다는 점이다. 다음으로는 유의미한 연구 결과를 보여줄 수 있을 정도의 데이터 규모를 고려하였다. 마지막 기준은 추후 연구를 위하여 넓은 시간 간격을 가지고 있어야 한다는 것이다. 또한, Emotion Trigger를 이용하여 감정 유발 요인을 찾아내는 것이 사회적으로 유의미한 연구가 되기 위해서는 해당 영역(domain)에 문제 상황이 존재해야 한다. 그래야만 사건에 대한 대중들의 부정적 감정의 원인을 찾아내어 문제 상황을 개선시킬 수 있기 때문이다. 이런 점들을 고려하여 사회적으로 커다란 이슈가 되고 있는 '전문가의 위상'이라는 주제를 전체적인 연구 대상으로 선정하였고, 전문가를 대표할 수 있는 직종인 '교수', '의사', '검사'에 대한 주요 사건을 구체적인 키워드로 하여 사례 연구를 진행하였다.

아래 <Figure 1>은 전체적인 연구 과정을 도식화한 것으로, 먼저 연구 주제의 키워드를 추출하기 위한 사전 분석을 수행하였다. 이를 통해 선정된 구체적인 키워드로 데이터를 다시 수집하여 뉴스·블로그 데이터에 대한 동시 출현 네트워



〈Figure 1〉 Research Model



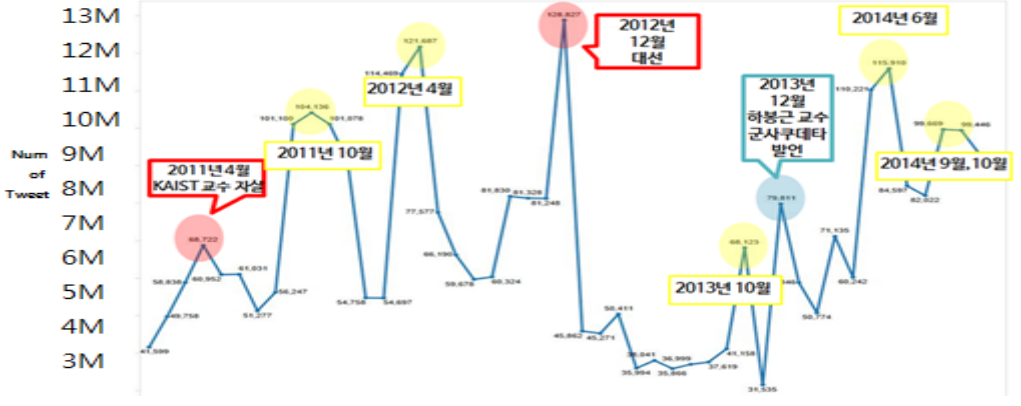
〈Figure 2〉 Advanced Data Collection

크 분석을 수행하였고, Word2Vec 분석은 트위터 데이터를 대상으로 하였다. 이는 트위터 250자에 화자의 감정이 잘 드러나기 때문으로 트위터를 대상으로 한 여러 감성 분석 연구가 있어 왔다(Go and Huang, 2009; Saif et al., 2012a; Kouloumpis et al., 2011; Saif et al., 2012b).

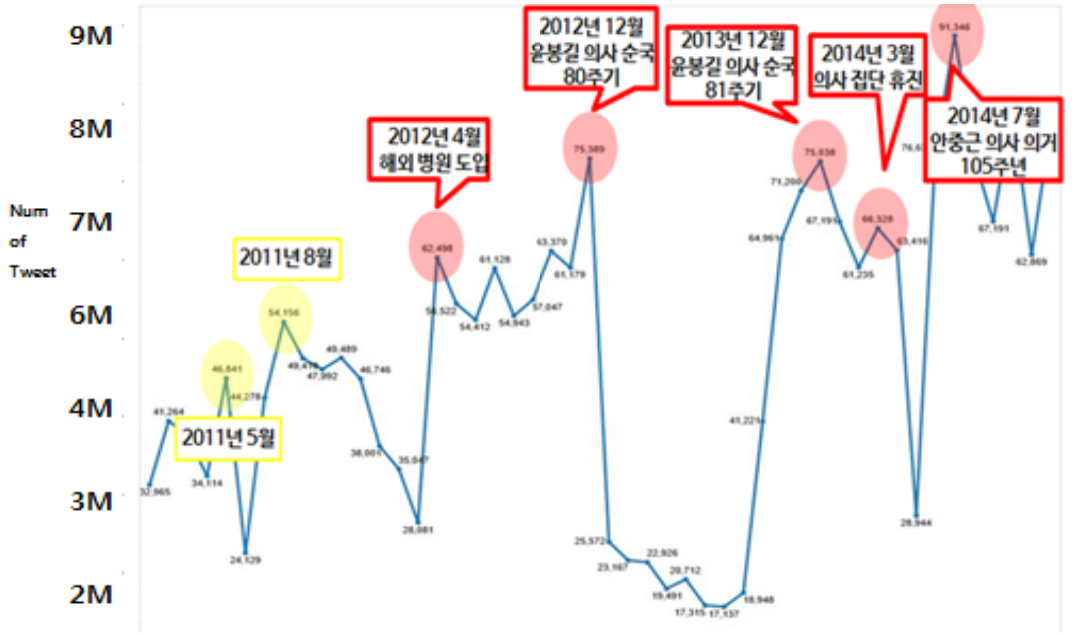
3.2. 1차 데이터 수집 및 분석

1차 데이터 수집은 본 연구의 구체적인 키워드를 선정하기 위한 사전 실험 단계로, 결과는 위의 <Figure 2>와 같다. '전문가의 위상'이라는 전체적인 연구 주제 하에 전문가를 대표할 수 있는 직종인 '교수', '의사', '검사'라는 단어를 검색어로 트위터에서 데이터를 수집하였고, 기간은 트위터가 시작된 2006년부터 2014년까지의 한국어 트윗(tweet)을 대상으로 하였다.

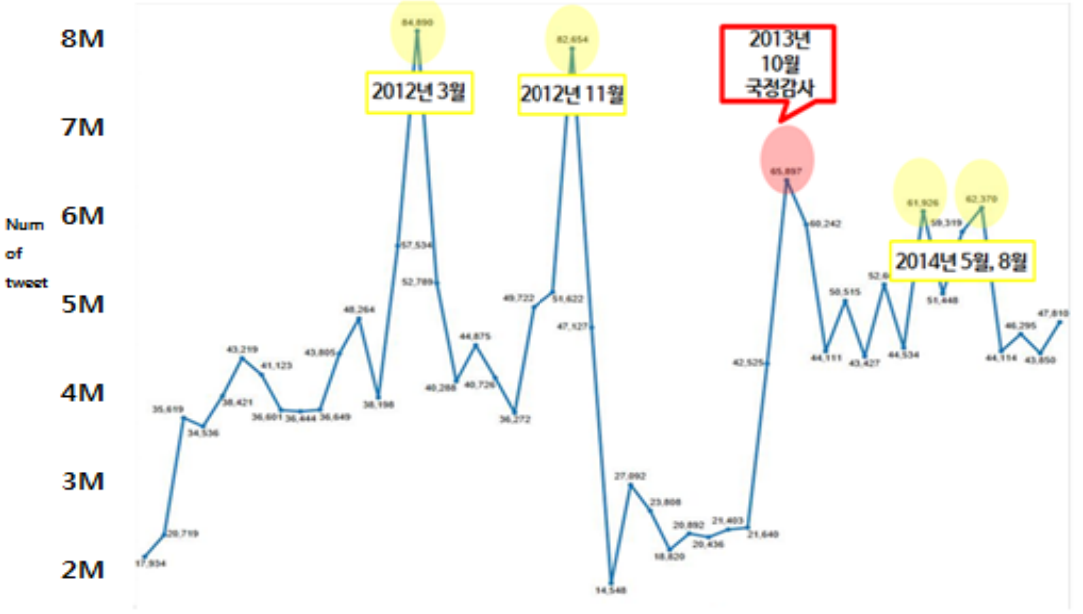
1차 수집 데이터에 대한 각 검색어(교수, 의사, 검사 순) 별 시계열 추이는 <Figure 3>, <Figure



〈Figure 3〉 'Professor'; Time Series of Twitter Search Result



〈Figure 4〉 'Doctor'; Time Series of Twitter Search Result



〈Figure 5〉 'Prosecutor'; Time Series of Twitter Search Result

4>, <Figure 5>와 같다. 그래프 상의 피크(peak)에 해당하는 시점의 실제 뉴스 기사를 비교하여 각 직업과 관련된 주요 사건을 찾고, 이를 통해 2차 수집을 위한 키워드를 선정하였다. 사건 설명이 표시된 지점들은 각 검색어의 동음이의어와 관련된 사건이거나, 교수·의사·검사의 위상에 초점이 맞추어진 사건이 아니므로 키워드 선정에서 제외하였다.

시계열 분석을 통해 얻은 각 직업 관련 구체적인 사건 키워드는 아래의 <Table 1>과 같고, 추후 실험은 이를 사용하여 수집한 2차 데이터를 대상으로 한다.

〈Table 1〉 Secondary Keywords for Data Gathering

Professor	교수 성추행
	교수 연구비 횡령
	교수 임용비리
	폴리페서
Doctor	신해철 스카이병원 의사
	음주수술 성형외과 의사
	의사 리베이트
Prosecutor	검사 음란행위
	스폰서 검사

3.3. 2차 데이터 수집 및 분석

1차 데이터 수집 및 분석을 통해 선정한 각 직업 관련 구체적인 사건 키워드로 뉴스, 블로그, 트위터에서 데이터를 다시 수집하였으며, 2차 데이터 수집 통계는 <Table 2>와 같다.

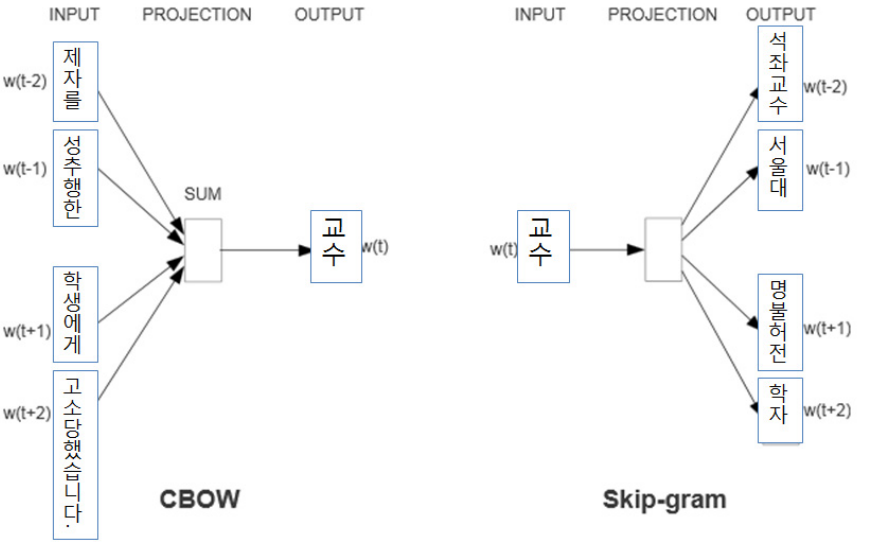
데이터 분석을 위해서 다음의 두 기법을 사용하였다. 먼저, 뉴스와 블로그 데이터에 대한 동시 출현 네트워크 분석을 수행하였다. 즉, 한 문장 내에서 동시에 출현한 단어들의 쌍을 추출한

〈Table 2〉 Secondary Data Collection

	Keyword	News	Blog	Twitter	Total
Professor	교수 성추행	328	3,178	8,798	25,720
	교수 연구비 횡령	1,499	36	1,627	
	교수 임용비리	2,106	27	396	
	폴리페서	1,229	63	6,433	
Doctor	신해철 스카이병원	3,208	1,408	8,763	35,110
	음주수술 성형외과 의사	808	2,030	439	
	의사 리베이트	9,951	2,435	6,068	
Prosecutor	검사 음란행위	2,763	982	1,566	43,225
	스폰서 검사	13,090	3,598	21,226	

것이다. 이는 각 키워드들과 연관된 일반 여론을 찾아내기 위한 방법인 동시에 Word2Vec이라는 새로운 기법과의 비교를 통해 Word2Vec만의 특징을 찾아내기 위한 것이다.

다음으로 수집 데이터에 Word2Vec 기법을 적용하여 각 키워드와의 연관어를 추출하고, 그 결과를 동시 출현 네트워크 분석 결과와 비교해서 차이점을 발견할 수 있었다. 2013년 등장한 Word2Vec(Mikolov et al., 2013) 기법은 신경망 분석 알고리듬의 한 유형이다. 신경망 알고리듬은 Harris의 분포가설(Harris, 1954)에서 비롯된 언어 추론 모델로서 특정 단어의 앞뒤에 위치한 단어의 분포를 벡터화(vectorization)하여 단어의 의미를 추론한다. 단어에서 벡터를 통한 수치를 이끌어내는 기법들 자체는 이전에도 많았지만 Word2Vec에서의 벡터는 단순한 수치적 의미가 아닌, 의미적 자질의 집합으로 표현됨으로써 일종의 '개념'을 나타내는 기능을 수행한다. 이를 위해 기존의 신경망 추론 모델인 CBOW(New Log-linear Models)와 Skip-gram 모델을 동시에 사용하여 더 빠르고 효과적으로 단어의 의미를



〈Figure 6〉 The Sample of Word2Vec Model(Mikolov, T., et al., 2013, p.5.)

추론한다. 위의 <Figure 6>은 Word2Vec의 설계 모델에 본 연구의 분석 결과에서 추출한 예시를 적용해 도식화한 것이다. CBOW는 특정 단어 주변의 단어들을 이용하여 단어에 대한 다차원 벡터를 형성하며, Skip-gram 모델은 해당 단어와 관련된 단어들을 예측한다.

감정 유발 요인 Emotion Trigger를 찾아내는 데에는 트위터 데이터를 사용하였다. 트위터 데이터는 이용자의 의견이 즉각적으로 표출되지만, 250자의 길이 제한으로 인해 텍스트가 매우 짧고 은어, 줄임말, 이모티콘 등이 빈번하게 등장한다. 따라서 기존 문헌을 대상으로 한 방법으로는 감성을 분석해내기 힘들다(Kim and Lee, 2014, p.734). 따라서 본 연구에서는 감성 분석을 위한 새로운 접근 방법을 제시한다.

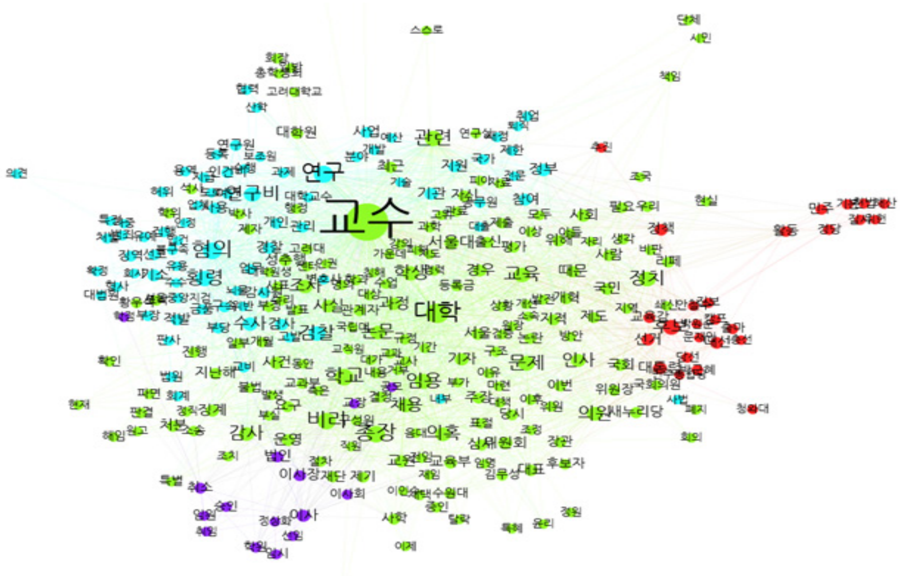
4. 연구결과 및 분석

4.1. 동시 출현 네트워크 분석

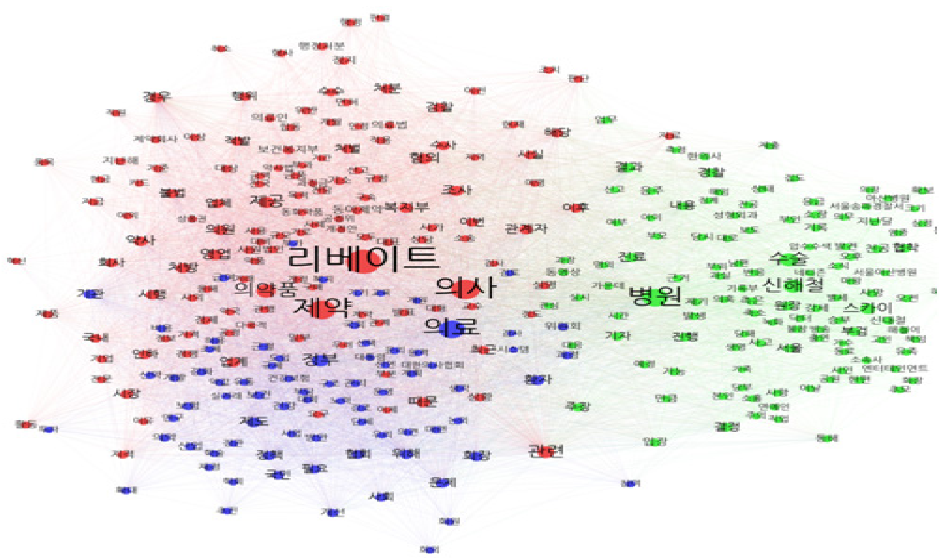
데이터 내에서 출현 단어의 단순 빈도 분석의 경우, 데이터 수집에 사용한 키워드들이 높은 빈도로 나타나 수집된 데이터의 양에 따라 그 빈도가 결정되는 경향을 보였고, 특별한 의미를 발견하기 힘든 일반 명사들이 순위 대부분을 차지하고 있었다. 따라서 단어들 간의 관계를 직관적으로 파악할 수 있는 동시 출현 네트워크 분석이 더 유의미하다고 판단하여 뉴스와 블로그 데이터를 대상으로 한 문장 내 함께 출현한 단어들의 쌍을 네트워크로 연결하여 그 분포와 빈도를 살펴보았다. 각 직업별 동시 출현 네트워크를 시각화 도구인 [1]Gephi를 이용해 <Figure 7>, <Figure 8>, <Figure 9>로 시각화하였고, 각 노드(node)의 크기는 동시 출현 빈도를 나타낸다. 또한, 각 단

1) http://gephi.github.io

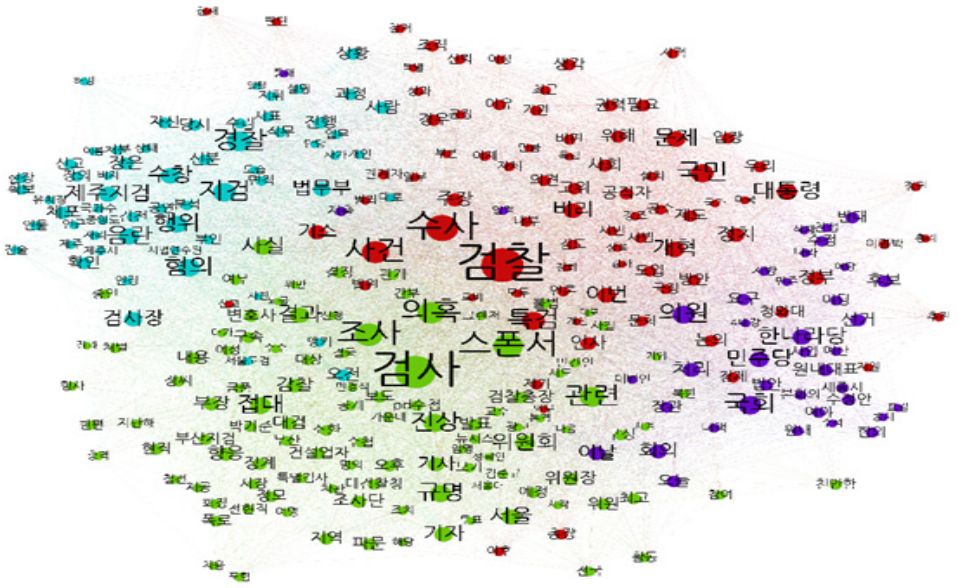
어 간의 관계가 네트워크에 반영되기 때문에 함께 자주 출현한 단어들의 경우 군집을 이루는 경향을 확인할 수 있다.



〈Figure 7〉 'Professor'; Co-Occurrence Network



〈Figure 8〉 'Doctor'; Co-Occurrence Network



〈Figure 9〉 'Prosecutor'; Co-Occurrence Network

4.2. Word2Vec 분석

아래의 <Table 3>은 '교수' 데이터의 Word2Vec 분석 결과이다. 교수의 경우 본문에 'O O 대학교 O O O 교수….'와 같은 기관명, 학과명, 인물명 등이 함께 등장하는데 Word2Vec 분석 결과 상위 랭크 단어들은 대부분 이에 속한다. 하지만 이러한 교수 관련어들을 제외한 기울임체로 표시된 키워드들을 살펴보면 데이터 수집에 쓰인 키워드가 아닌 새로운 단어들임을 확인할 수 있다. 이는 Word2Vec 기법이 단순 출현 빈도를 계산해 내는 것이 아니라, 단어 고유의 의미 자질을 효과적으로 선별해 내고 있음을 보여준다.

〈Table 3〉 'Professor'; The Result of Word2Vec Analysis

	Professor			
	Blog	Twitter	News	
1	서울대	서울대	서울대	*Word except institution name and department name*
2	김태일	창원대	중앙대	
3	동아대학교	***명불허전***	한반도선진화재단	
4	수학과	의예과	석좌교수	
5	성악	전주대	휘문고	
6	숙명여대	***부적격***	대학교수	
7	재직	석좌교수	버클리대	
8	경북대	경희대	한양대	

	Professor		
	Blog	Twitter	News
9	서울대학교	성신여대	대학원
10	이화여대	휴스톤	고려대
11	융합과학	전북대	도쿄대
12	***대필***	해당자	리페
13	법학전문대학원	성낙인	학과
14	***교습***	휴스턴	성균관대
15	겸임	어윤대	안경환
16	이상돈	의학사	정치학
17	원준	***후학***	***관료***
18	석좌교수	인천대	경영학
19	객원	인터넷신문	***합성어***
20	교수안	송호근	인사
21	심리학과	정치과학	김태윤
22	성낙인	창원대학교	출신
23	아주대	경남도립거창대학	가타
24	***특채***	***빈축***	한국은행
25	연세대	한양대	특전
26	중앙대	교육과학부	학자
27	자원학과	부학장	*밥통*
28	전북대	수학과	울대
29	한국교원대	김종인	이양희
30	한국방송통신대학교	정외과	가황

아래의 <Table 4>는 '의사' 데이터의 Word2Vec 분석 결과이다. 의사 관련 데이터 수집에 사용한 키워드는 '의사' 뿐이었지만, '한의사', '약사', '의약사', '의료인' 등 의사를 대체할 수 있는 유의어를 다수 확인할 수 있었다. 또한 '수수', '뒷돈', '몰지각', '죄의식' 등 부정적인 단어가 상위에 다수 출현하는 것을 통해 단순 출현 빈도에서는 추출되지 않던 의료계에 대한 부정적 인식을 많이 잡아냈다는 것을 알 수 있다.

〈Table 4〉 'Doctor'; The Result of Word2Vec Analysis

	Doctor			*Synonym with 'Doctor'*
	Blog	Twitter	News	
1	***한의사***	***한의사***	군산경찰서	
2	범죄자	현행법	***의약사***	
3	변호인단	***의료인***	***의료인***	

	Doctor			
	Blog	Twitter	News	
4	***의료인***	사사	제약회사	
5	수수	몰지각	***약사***	
6	뒷돈	사가	은의	
7	서울행정법원	병원	명제	
8	금품	금지법	***중의사***	
9	공판	윤리위	도처	Negative Word
10	***의약사***	제다	선급금	
11	유죄	극렬	수백	
12	수뢰	대한의원	공소시효	
13	죄명	진료실	정하나	
14	사면초가	만시지탄	특무	
15	번복	눈높이	공중보건	
16	일양약품	대구지법	죄의식	
17	서울중앙지법	사전통지	리베이트	
18	확정판결	본회	그대	
19	향응	수수죄	금품	
20	수금액	배임죄	향응	
21	고합	돈많이쳐먹는종자들이수록파업	댓가	
22	***약사***	직역	로타	
23	무더기	트리뷰	살포	
24	사기죄	나경	김지	
25	대한의원	공분	모제	
26	합의체	엄단	가로	
27	사전통지	공여자	거제경찰서	
28	목돈	소송비용	***한의사***	
29	개원	중의사	후원금	
30	공중보건	악법	십만	

아래의 <Table 5>는 '검사' 데이터의 Word2Vec 분석 결과이다. 본 결과의 경우, 데이터별로 상위 랭크 단어의 양상이 다르게 나타났다. 블로그와 뉴스에서는 공통적으로 고급차 이름들(그랜져, 벤츠 등)과 건설업계 관련 단어들(건설업, 강바닥 등)이 등장한 반면 트위터에서는 '스폰서 검사'를 보도했던 pd수첩 관련 단어들이 상위에 출현하고 있다.

〈Table 5〉 'Prosecutor'; The Result of Word2Vec Analysis

	Prosecutor			
	Blog	Twitter	News	
1	***유착***	***스폰서***	사건	***Word related with Sponser***
2	***스폰서***	진실	그랜저	
3	***곤욕***	검찰	***스폰서***	
4	그랜저	강의	벤츠	
5	건설업	PD 수첩	***곤욕***	
6	검찰	강바닥	부장	
7	건설업자	수첩	***소개비***	
8	그랜저	연출	강찬우	
9	***접대***	프로듀서	검찰	
10	이심전심	시대	샤넬	Word can be a target of a back-scratching alliance of government and businesses
11	빤스	목격담	어제오늘	
12	폭로	서사	그랜져	
13	용재	금정	상림	
14	***술대접***	***뒷이야기***	천벌	
15	섹검	수정안	내용증명	
16	***떡값***	창원지검	***비리***	
17	변호사들	화이팅	슬롯머신	
18	삼성특검	국무위원	최근	
19	***몰상식***	첨단	승용차	
20	***온정주의***	대강	***체면***	
21	스포	아자	머리끝	
22	기업인	간만	자라	
23	벤츠	이다	***만신창이***	
24	***비리***	보면	***추문***	
25	***커넥션***	오늘	***악습***	
26	동석	편의	***대납***	
27	부산경남	시상식	홍역	
28	본궤도	세종시	의기소침	
29	정씨	다큐	회색	
30	토착	흥미	수사	

4.3. 동시 출현 네트워크와 Word2Vec의 비교 분석

아래의 <Table 6>에서는 동시 출현 네트워크 분석 결과와 Word2Vec의 분석 결과를 정리·비교하였다. 주로 Word2Vec의 결과에서 부정적 단어의 빈도가 높은 것을 알 수 있다. 특히 검사와 의사의 경우 상위 20개 단어들 중 부정적인 단어의 수가 Word2Vec 분석 결과에서 3배가량 높게 나타났다. 아울러 전반적으로 동시 출현 네트워크 분석 결과의 경우 데이터 수집에 사용한 키워드가 상위 랭크 단어에 속하는 반면, Word2Vec은 이러한 키워드가 상위 랭크 단어에 존재하지 않았다. 이것은 Word2Vec의 의미 자질 벡터화의 특성이 나타난 결과로 보인다. 예를 들어, 의사 데이터 수집 시 '신해철', '음주' 등의 단어를 키워드로 사용하였으나, Word2Vec 분석결과 상위

〈Table 6〉 The Comparison of Co-Occurrence and Word2Vec

Common word / Negative Word

Professor	Blog	Co-occurrence	사람	사건	사회	성추행	자신
			대통령	생각	여성	우리	학생
			후보	교육	사실	학교	한국
			경제	피해자	협의	조사	기본
		Word2Vec	서울대	김태일	동아대학교	수학과	성악
			숙명여대	재직	경북대	서울대학교	이화여대
			융합과학	대필	법학전문 대학원	교습	겸임
			이상돈	원준	석좌교수	객원	교수안
	News	Co-occurrence	대학	연구	협의	총장	학교
			비리	횡령	정치	임용	교육
			후보	연구비	검찰	학생	의원
			의혹	감사	기사	논문	채용
		Word2Vec	서울대	중앙대	한반도 선진화재단	석좌교수	휘문고
			대학교수	버클리대	한양대	대학원	고려대
			도쿄대	리페	학과	성균관대	안경환
			정치학	관료	경영학	합성어	인사
Doctor	Blog	Co-occurrence	리베이트	의료	수술	병원	제약
			신해철	사람	치료	의약품	제공
			환자	생각	스카이	문제	처분
			정부	음주	처방	자신	건강
		Word2Vec	한의사	범죄자	변호인단	의료인	수수
			뒷돈	서울행정법원	금품	공판	의약사
			유죄	수뢰	죄명	사면초가	번복
			일양약품	서울중앙지법	확정판결	향응	수금액

Doctor	News	Co-occurrence	***리베이트***	제약	의료	병원	의약품
			신해철	수술	정부	제공	스카이
			혐의	영업	복지부	조사	처방
			제도	***약사***	의원	환자	업계
		Word2Vec	군산경찰서	의약사	의료인	제약회사	***약사***
			은의	명제	중의사	도처	선급금
			수백	공소시효	정하나	특무	공중보건
			죄의식	***리베이트***	그대	금품	향응
Prosecutor	Blog	Co-occurrence	***검찰***	프로젝트	사건	사람	***스폰서***
			수사	경찰	국민	생각	조사
			사실	관리	비리	사회	행위
			대통령	음란	문제	김수창	범인
		Word2Vec	유착	***스폰서***	곤욕	그랜저	건설업
			검찰	건설업자	그랜져	접대	이심전심
			빤스	폭로	용재	술대접	조영진
			섹검	주양	떡값	변호사	삼성특검
	News	Co-occurrence	***검찰***	수사	***스폰서***	조사	사건
			의혹	경찰	국회	특검	국민
			의원	진상	혐의	접대	규명
			행위	음란	대통령	김수창	개혁
		Word2Vec	***사건***	그랜저	***스폰서***	벤츠	곤욕
			부장	소개비	강찬우	***검찰***	샤넬
			어제오늘	상림	천벌	내용증명	비리
			슬롯머신	최근	승용차	체면	머리끝

20위권에 이 단어들은 등장하지 않는다. 이러한 단어들이 일반적으로 의사와 관련성이 떨어진다는 것을 Word2Vec가 계산해낸 것이다. 따라서 다음 장에서 소개할 Emotion Trigger에서 단순 감정 판정에서 한 단계 나아가 해당 감성과 어떤 키워드가 연관되어 있는지 밝혀내는데 동시 출현보다 Word2Vec이 더 효과적인 방법이라고 판단하였다.

4.4. Emotion Trigger

감정 유발 요인 Emotion Trigger를 추출하기 위한 실험 과정은 다음과 같다. 먼저 트위터 데이터를 대상으로 Komoran 형태소 분석기[2]를 사용하여 전처리 수행 후, 정제된 데이터를 사용해 Word2Vec 라이브러리로 Word Vector 모델을 생성하였다. Komoran 형태소 분석기를 사용한 이유는 해당 라이브러리가 Wikipedia 사전을 자체

2) http://www.shineware.co.kr/?page_id=835

내장하고 있어 다른 형태소 분석기보다 고유 명사에 더 유연하게 대응할 수 있기 때문이다. 이렇게 생성된 Word Vector 모델에서 연구 대상으로 삼은 교수, 의사, 검사를 '대상(target)어휘'로 선정, 이들과 코사인 유사도(cosine similarity)가 높은 형용사들을 추출하였다. 이 형용사들은 Word2Vec 알고리듬에 따라 대상어휘들과 의미·문맥적으로 유사한 단어들이라고 해석할 수 있는데, 본 연구에서는 이들을 대상어휘들에 대한 사람들의 감정을 대변하는 '감정어휘'라고 정의하였다. 다음의 <Table 7>, <Table 8>, <Table 9> 좌측 두 열은 이렇게 추출한 감정어휘들을 코사

〈Table 7〉 'Professor'; The Parts of 'Emotion Trigger' by Emotional Words

Emotional Word		Common Emotional Word's Emotion Trigger				
		역겹	어처구니없	우스꽝스럽	부끄럽	우습
keyword	cosine_sim	keyword	keyword	keyword	keyword	keyword
뒤늦	0.68363321	진부	재차	꿈속	앳되	되짚
당차	0.67052803	용납	스스럼없이	살다보면	행여	헛것
무겁	0.56709038	징징대	바바리맨	황홀	역지사지	그렇잖
징그럽	0.52498171	창피	얼버무리	류의	커밍아웃	명답
올바르	0.40982589	오지랖	구역질	헛것	스트라이크	몰라주
부끄럽	0.40430476	씨부리	글쎄	가나이	헤프	소린
섣부르	0.40314126	구기	별것	흑백	구역질	곰곰
짙다	0.39798585	고이	어이없	동정심	비웃	도통
잦다	0.36359223	무트	곤혹	정신력	돌멩이	별루
드물	0.35783699	따라다니	두둔	축수	문재인에게만 90 도로조아리는조국교수폴리페서도여직원집주소를	곱다
거세	0.353501	글쎄	떠밀리	되짚	왠지	네네
두렵	0.33409226	참모습	지참	변화무쌍	창피	이만큼
남다르	0.33170383	지겹	모욕감	통속	자책	버겁
늦다	0.33112454	얼렁뚱땅	흐지부지	무트	자폐증	안쓰럽
어처구니없	0.32294111	이신	분개	거란	이놈	애기
짧다	0.32171493	광주학생항일운동	구리	이신	우습	살다보면
배부르	0.31575917	일안	앙심	슬퍼하	안타깝	제모습
느닷없	0.31357509	타진요	가만있	적막	째려보	능청
안타깝	0.30508825	자식새끼	류의	몸서리치	앞뒤	여태껏
슬프	0.37062365	능청	밀치	여리	비참	후회

〈Table 8〉 ‘Doctor’; The Parts of ‘Emotion Trigger’ by Emotional Words

Emotional Words		Common Emotional Word's Emotion Trigger				
		역겹	어처구니없	우스꽝스럽	부끄럽	우습
keyword	cosine_sim	keyword	keyword	keyword	keyword	keyword
값싸	0.63698006	얼버무리	이욕	어리석	게으름	죄다
이르	0.614883	내키	이올린	한낱	두렵	되묻
약다	0.58393914	그래요	진위	형씨	욕심	맘대로
애꿎	0.58019662	썩이	내세	아유미	안식	그래요
끊임없	0.56680068	염치없	철이	파리대왕	빚다	요전
관계없	0.56530297	하오	팩트	피에타	동갑내기	마냥
뒤늦	0.54814659	태클	중정	이죄	겸손	하오
공교롭	0.53644299	이눔	유작	빌어먹	알아주	늘어놓
어처구니없	0.51244267	두문자	측도	나단	또박	딱하
걸맞	0.49386487	놀리	서상수	마메	법치국가	수작
올바르	0.48525069	탐나	정정	알아듣	와르르	곰곰이
거세	0.46366194	한결같이	누리꾼	짓뭉개	속상하	안온
가파르	0.46348361	선입관	주검	죄다	크리스틴	어린애
싸다	0.45571567	아하	경악	워즈워스	마메	한낱
늦다	0.45284704	가엾	낭심	바이런	하하하	얼버무리
비리	0.4504801	어리석	나루	빈틈	그분	늙은이
박하	0.44642818	비아냥거리	이어서	현모양처	허무	무식
손쉽	0.44575569	호호호	유능	표범	언사	그러니까
형편없	0.42350226	방금	심경	이세상	너그럽	곰곰
줄기차	0.42162727	작자	사실무근	불현듯	환갑	어쩐지

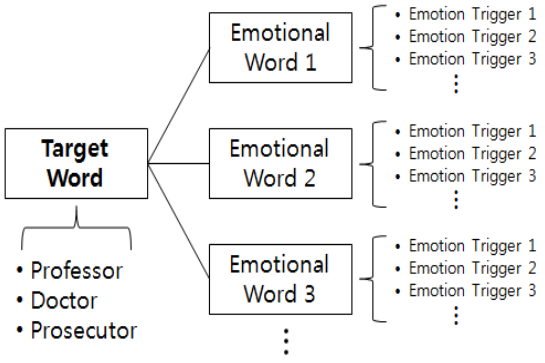
인 유사도가 높은 순으로 단어 전거 없이 상위 20위까지 제시한 결과이다. 즉, 각 직업을 나타내는 대상어휘와 실험 데이터 내에서 가장 의미적으로 유사한 감정어휘들인 것이다.

다음으로 이러한 감정어휘들이 출현하는 이유를 알아보기 위하여 다시 Word Vector 모델을 사용하여 감정어휘들과 의미·문맥적으로 유사한 명사들을 다시 추출하고 이들을 ‘감정유발요인’이라고 정의하였다. 그리고 ‘대상어휘 - 감정어휘 - 감정유발요인’으로 이어지는 전 과정을 ‘Emotion Trigger’라 명명하였다. <Table 7>, <Table 8>, <Table 9> 우측 열의 단어들은 세 대상어휘에서 공통적으로 출현한 감정어휘들 중 일부(역겹다, 어처구니없다, 우스꽝스럽다, 부끄

럽다, 우습다)와 코사인 유사도가 높은 감정유발 요인들을 단어 전거 없이 상위 20위까지 제시한 결과이다. 이를 통해 대상에 대해 동일한 감정을 표현하더라도 이를 유발하는 요인은 모두 다르게 나타남을 확인할 수 있다.

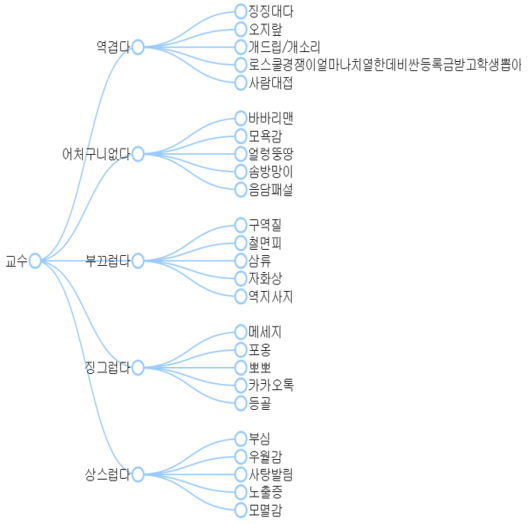
〈Table 9〉 'Prosecutor'; The Parts of 'Emotion Trigger' by Emotional Words

Emotional Word		Common Emotional Word's Emotion Trigger				
		역겹	어처구니없	우스꽝스럽	부끄럽	우습
keyword	cosine_sim	keyword	keyword	keyword	keyword	keyword
어이없	0.45265243	검열	남존여비	검진	유치	원내대표
독하	0.44372551	교류	빨갱이	빙고	으시대	기존
많다	0.42646456	저도	직선	바코드	검찰권력정권에갖다바치는정치검사들이활개칠때떡검섹검스폰서검	아깝
어렵	0.42613833	유해진	선진국	벗기	짓다	오랜만
흐리	0.42538325	도서	공평	출판사	수뇌부	영원
안타깝	0.41379119	여쭈	성은	흠흠	부끄러움	파마
호되	0.40438895	중단	원내대표	바라보	거렁뱅이	기소독점주의
우습	0.39945356	세다	차원	궁리	알다	농담
끄떡없	0.39511585	인증	청문회	앗다	새끼	숙이
못나	0.38827132	강자	국회의원	존경	모르	다이
가깝	0.38807737	송두리	멈추	가르	얻다	제발
정신없	0.38516587	떨리	실세	고통	모양	향후
바쁘	0.38108648	자비	업무	기존	뒷거래	무능
이르	0.37967955	두렵	의형제	보세	노릇	중심
드세	0.3788907	인용	증명	결혼	도스	두려워하
부럽	0.37585572	산타령	트다	심정	까닭	나도
강하	0.37579682	가볍	번거롭	모델데려와서수청들게한놈그거눈감아준놈특검으로다시눈감아준넘	통속	짐작
어처구니없	0.37336814	의인	일반인	알아보	승리	재경일보
괜찮	0.3730764	반전	갑과	굳다	숭상	변협
슬프	0.37062365	강요	용의	이안	아라한	두발

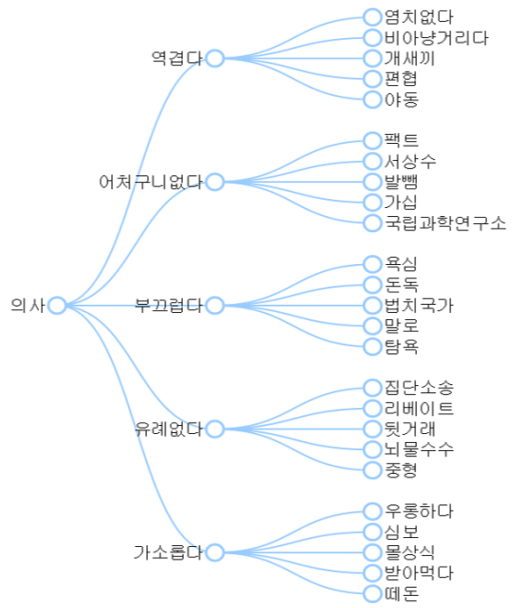


〈Figure 10〉 Emotion Trigger Model

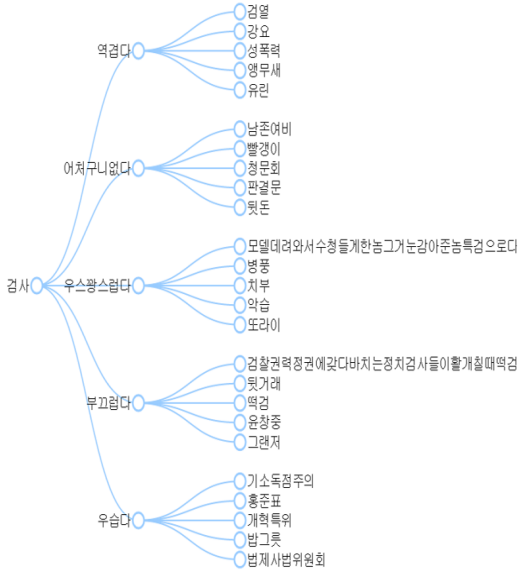
위의 <Figure 10>은 '대상어휘 - 감정어휘 - 감정유발요인'으로 연결되는 Emotion Trigger 과정을 도식화 한 결과이고, <Figure 11>, <Figure 12>, <Figure 13>은 각각 교수, 의사, 검사 대상어휘에 대한 Emotion Trigger 결과 중 유사도가 높고 부정적인 의미를 지닌 단어들을 일부 선정하여 시각화 한 결과이다. 시각화를 위한 단어 선정 시 부정적인 단어들을 위주로 한 이유는 본 연구가 사회적으로 부정적인 사건들을 대상으로 한 사



〈Figure 11〉 The Visualization of 'Professor'



〈Figure 12〉 The Visualization of 'Doctor'



〈Figure 13〉 The Visualization of 'Prosecutor'

례 연구이기 때문에 해당 사건들에 대한 대중들의 부정적인 감정을 파악하기 위해서이다.

5. 결론

본 연구는 키워드간의 관계를 밝힘에 있어 최근의 한계를 극복하고자, Word2Vec 기법을 적용하여 의미론적 요소를 결합하는 과정에서 대상에 대한 감정 유발 요인을 찾아내는 새로운 방법론 'Emotion Trigger'를 제시했다는 점에 커다란 의의가 있다. 소셜 미디어를 분석하는 기존 연구들에서는 사회적인 특정 이슈가 발생했을 때 그에 대한 대중들의 단순 반응 정도만을 파악할 수 있었지만, 본 연구의 Emotion Trigger를 활용한다면 '세월호 사건'이나 '메르스 사태'와 같이 주요한 사회 문제를 분석하는데 있어 소셜 데이터 속에서 나타나는 대중들의 실제 감정을 반영하고, 나아가 그 감정들이 어떤 요인으로 인해 유발되는지를 보다 깊게 파악하여 사회 분열이나 갈등을 조장하는 문제들을 해결하기 위한 방안을 제시할 수 있는 좋은 발판이 될 것으로 기대한다.

한편, Emotion Trigger의 가장 큰 한계는 형용사와 명사 품사만을 감정어휘와 감정유발요인의 분류 기준으로 사용함으로써 이 둘 간의 명확한 인과 관계를 증명하기 어렵다는 점에 있다. 이러한 한계는 향후 연구자 태깅(tagging)을 통한 텍스트 분류 결과와의 비교 등을 통해 보완할 예정이다. 특히 감정어휘를 형태소 분석기 상의 형용사로만 추출하였는데, 이 역시 추가적으로 감정 사전을 구축하여 보다 정교하게 해야 할 것이다. 또한, 트위터라는 소셜 데이터 특성 상 오타, 띄어쓰기, 비속어 등에 대한 한글 텍스트 처리 역시 보완해야 할 점이다.

추후 이러한 문제들을 보완한다면 Emotion Trigger의 역추적을 통해 어떤 사건들이 사회적으로 부정적인 감정을 야기하는지 그 원인을 찾아, 기존의 오피니언 마이닝에서 한 단계 더 나아간 심층적 오피니언 마이닝을 수행할 수 있을 것이다. 또한 본 연구에서는 트위터 데이터만을 사용하여 Emotion Trigger를 수행하였지만, 제시한 방법론은 특정 문헌의 종류에 한정되지 않기 때문에 향후 다양한 데이터 속에서 대중들의 감정을 파악하고 그 원인을 구체적으로 분석하는데 이용할 수 있을 것이다. 특히, 실시간으로 축적되는 시계열 데이터에 적용한다면 사건의 추이에 따라 대중들의 의견이 어떻게 바뀌어 나가는지 그 흐름을 추적하는 트렌드 분석 역시 가능할 것이다.

참고문헌(References)

An, J. K. and H. W. Kim, "Building a Korean Sentiment Dictionary and Applications of Natural Language Processing," *Korea Intelligent Information System Society*, (2014), 177~182.

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

Go, A., R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision", CS224N Project Report, Stanford, 2009, 1~12.

Hamouda, A., and M. Rohaim, "Reviews classification

using sentiwordnet lexicon," *World Congress on Computer Science and Information Technology*, 2011.

Harris, Zellig S., "Distributional structure," Word, 1954.

Hong, S. R., Y. O. Jeong, and J. H. Lee, "Semi-supervised learning for sentiment analysis in mass social media," *Journal of Korean Institute of Intelligent System*s, Vol. 24, No. 5(2014), 482~488.

Hung, C. and H. K. Lin, "Using objective words in SentiWordNet to improve word-of-mouth sentiment classification," *IEEE Intelligent Systems,* Vol. 28, No. 2(2013), 47~54.

Jang, H. J., "Classification System for Emotional Verbs and Adjectives," *Korea Society for Information Management,* (2001), 29~34.

Jang, K. A., S. H. Park, and W. J. Kim, "Automatic Construction of a Negative/positive Corpus and Emotional Classification using the Internet Emotional Sign," *Korean Institute of Information Scientists and Engineers*, Vol. 42, No. 4(2015), 512~521.

Kang, H. H., S. J. Yoo, and D. H. Han, "Design and Implementation of System for Classifying Review of Product Attribute to Positive/ Negative," *Korean Institute of Information Scientists and Engineers*, 36(2C), (2009), 1~6.

Kang, H. H., S. J. Yoo, and D. H. Han, "Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews," *Expert Systems with Applications*, Vol. 39, No. 5(2012), 6000~6010.

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

Kim, K. M. and J. H. Lee, "Sentiment Analysis of Twitter using Lexical Functional Information," *Korean Institute of Information Scientists and Engineers*, (2014), 734~736.

Kim, S. W. and N. Kim, "A Study on the Effect of Using Sentiment Lexicon in Opinion Classification," *Korea Intelligent Information System Society,* (2013), 121~128.

Kim, Y. S. and Y. H. Seo, "Journal of Korea Entertainment Industry Association," *Korea Entertainment Industry Association*, (2013), 206~210.

Kouloumpis, E., T. Wilson, and J. Moore, "Twitter sentiment analysis: The good the bad and the omg!," *ICWSM*, Vol. 11(2011), 538~541.

Lee, C. S., D. H. Choi, S. S. Kim, and J. W. Kang, "Classification and Analysis of Emotion in Korean Microblog Texts," *Korean Institute of Information Scientists and Engineers,: Database*, Vol. 40, No. 3(2013), 159~167.

Liu, B., "Sentiment analysis and subjectivity," *Handbook of natural language processing*, Vol. 2(2010), 627~666.

Mikolov, T., K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, arXiv preprint arXiv: 1301.3781.

Narayanan, R., B. Liu, and A. Choudhary, "Sentiment analysis of conditional sentences," *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing,* Vol. 1(2009), 180~189.

Ohana, B. and B. Tierney, "Sentiment classification of reviews using SentiWordNet," *9th. IT & T*

Conference, (2009), 13.

Sadamitsu, K., S. Sekine, and M. Yamamoto, "Sentiment Analysis Based on Probabilistic Models Using Inter-Sentence Information," *LREC*, (2008).

Saggionα, H., and A. Funk, "Interpreting Senti WordNet for opinion classification," *Proceedings of the seventh conference on international language resources and evaluation LREC10*, (2010), 1129~1133.

Saif, H., Y. He, and H. Alani, "Alleviating data sparsity for twitter sentiment analysis," *CEUR Workshop Proceedings* (CEUR-WS. org), (2012), 2~9.

Saif, H., Y. He, and H. Alani, "Semantic sentiment analysis of twitter," *The Semantic Web–ISWC 2012,* 2012b, 508~524.

Seo, J. H., H. J. Cho, and J. T. Choi, "Design for Opinion Dictionary of Emotion Applying Rules for Antonym of the Korean Grammar," *JKIIT,* Vol. 13, No. 2(2015), 109~117.

Seo, J. R. and C. Ko, "Big Data Analysis by Sensitivity Analysis," *Journal of The Society of Convergence Knowledge*, Vol. 2, No. 1(2014), 15~21.

Song, J. S., S. W. Lee, "Automatic Construction of Positive/Negative Feature-Predicate Dictionary for Polarity Classification of Product Reviews," *Korean Institute of Information Scientists and Engineers : Software and Application*, Vol. 38, No. 3(2011), 157~168.

Abstract

A Study of 'Emotion Trigger' by Text Mining Techniques

Juyoung An* · Junghwan Bae* · Namgi Han* · Min Song**

The explosion of social media data has led to apply text‐mining techniques to analyze big social media data in a more rigorous manner. Even if social media text analysis algorithms were improved, previous approaches to social media text analysis have some limitations. In the field of sentiment analysis of social media written in Korean, there are two typical approaches. One is the linguistic approach using machine learning, which is the most common approach. Some studies have been conducted by adding grammatical factors to feature sets for training classification model. The other approach adopts the semantic analysis method to sentiment analysis, but this approach is mainly applied to English texts. To overcome these limitations, this study applies the Word2Vec algorithm which is an extension of the neural network algorithms to deal with more extensive semantic features that were underestimated in existing sentiment analysis. The result from adopting the Word2Vec algorithm is compared to the result from co-occurrence analysis to identify the difference between two approaches. The results show that the distribution related word extracted by Word2Vec algorithm in that the words represent some emotion about the keyword used are three times more than extracted by co-occurrence analysis. The reason of the difference between two results comes from Word2Vec's semantic features vectorization. Therefore, it is possible to say that Word2Vec algorithm is able to catch the hidden related words which have not been found in traditional analysis. In addition, Part Of Speech (POS) tagging for Korean is used to detect adjective as "emotional word" in Korean. In addition, the emotion words extracted from the text are converted into word vector by the Word2Vec algorithm to find related words. Among these related words, noun words are selected because each word of them would have causal relationship with "emotional word" in the sentence. The process of extracting these trigger factor of emotional word is named "Emotion Trigger" in this study. As a case study, the datasets used in the study are collected by searching using three keywords: professor,

* Department. of Library and Information Science, College of Liberal Arts, Yonsei University
** Corresponding Author: Min Song
Department. of Library and Information Science, Yonsei University
120-749, 50 Yonsei-ro, Seodaemun-gu, Seoul, Korea
Tel: +82-2-2123-2405, Fax: +82-2-393-8348, E-mail: min.song@yonsei.ac.kr

prosecutor, and doctor in that these keywords contain rich public emotion and opinion. Advanced data collecting was conducted to select secondary keywords for data gathering. The secondary keywords for each keyword used to gather the data to be used in actual analysis are followed: Professor (sexual assault, misappropriation of research money, recruitment irregularities, polifessor), Doctor (Shin hae-chul sky hospital, drinking and plastic surgery, rebate) Prosecutor (lewd behavior, sponsor). The size of the text data is about to 100,000(Professor: 25720, Doctor: 35110, Prosecutor: 43225) and the data are gathered from news, blog, and twitter to reflect various level of public emotion into text data analysis. As a visualization method, Gephi (http://gephi.github.io) was used and every program used in text processing and analysis are java coding. The contributions of this study are as follows: First, different approaches for sentiment analysis are integrated to overcome the limitations of existing approaches. Secondly, finding Emotion Trigger can detect the hidden connections to public emotion which existing method cannot detect. Finally, the approach used in this study could be generalized regardless of types of text data. The limitation of this study is that it is hard to say the word extracted by Emotion Trigger processing has significantly causal relationship with emotional word in a sentence. The future study will be conducted to clarify the causal relationship between emotional words and the words extracted by Emotion Trigger by comparing with the relationships manually tagged. Furthermore, the text data used in Emotion Trigger are twitter, so the data have a number of distinct features which we did not deal with in this study. These features will be considered in further study.

Key Words : Emotion Trigger, Word2Vec, Sentimental Analysis, Text Mining, Social Issues

Received : June 5, 2015 Revised : June 17, 2015 Accepted : June 18, 2015
Type of Submission : Fast Track Corresponding Author : Min Song

저 자 소 개

안 주 영

연세대학교 문헌정보학과 국문학 학사를 졸업 후 문헌정보학 대학원 석사과정에 재학 중이다. 주요 관심분야는 텍스트 마이닝에 기반한 빅데이터 분석 및 정보 검색 분야이다.

배 정 환

연세대학교 문헌정보학 학사를 졸업 후 동 대학원 석사과정에 재학 중이다. 주요 관심분야는 텍스트 마이닝에 기반한 소셜 미디어 빅데이터 분석 및 정보 시각화 분야이다.

한 남 기

연세대학교 문헌정보학 학사를 졸업 후 동 대학원 석사과정에 재학 중이다. 주요 관심분야는 텍스트 마이닝에 기반한 빅데이터 분석, 정보 공학 및 정보 검색 분야이다.

송 민

Prof. Song has a background in Text Mining, Bioinfomatics, Information Retrieval and Information Visualization. Prior to Yonsei, he was an Associate Professor with tenure in the Department of Information Systems at New Jersey Institute of Technology (NJIT). At NJIT, he received several grants from NSF and IMLS and published a number of papers in the Text Mining research area. Before joining NJIT, Professor Song worked at Thomson Scientific (now Thomson Reuters). At Thomson, the major responsibilities were to develop Knowledge Management tools, middleware components, and the search engine for citation database. His recent work in Text Mining addresses automatic database selection, entity and relation extraction, high speed document filtering, algorithms that learn a person's information needs from experience, automatic analysis of gathered information. He is also involved in a variety of information visualization projects. Prof. Song is also interested in information and knowledge management in large organizations. He is currently interested in applying Text Mining algorithms to Bioinformatics and Social Media.