Journal of KIIT. Vol. 18, No. 6, pp. 107-115, Jun. 30, 2020, pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2020.18.6.107

단어 임베딩 및 언어심리학적 변인을 이용한 온라인 뉴스 댓글에 대한 반응 분석

안 형 준*

Analysis of the Reaction to User Comments of Online News with Word Embedding and Psycho-Linguistic Variables

Hyung Jun Ahn*

본 연구는 2016년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2016S1A5A2A01926914)

요 약

디지털 기기의 일상적 사용이 보편화됨에 따라 뉴스의 소비도 디지털 채널로 옮겨가고 있다. 많은 독자들은 뉴스 기사에 대한 댓글을 통해 여론 형성에 참여하고 있다. 따라서 온라인 뉴스에 대한 댓글을 분석하는 것은 새로운 매체 사용 행태를 이해하는데 매우 중요하다고 할 수 있다. 본 연구는 인공 신경망 기반 임베딩 및 언어심리학적 변인들을 이용하여 댓글 텍스트 및 그에 대한 반응을 분석하고자 하였다. 구체적으로는 임베딩을 이용하여 댓글에 포함된 정치, 경제, 생활, 문화 등 주제 범주 어휘들을 파악하고, 또한 언어심리학적 특징과 관련된 변수들을 이용하여 댓글에 대한 반응을 살펴보았다. 분석 결과, 언어심리학적 변수들 및 주제 어휘들의 유의한 영향, 뉴스 특성과 댓글 특성 간의 유의한 상호작용을 확인하였다.

Abstract

With the increasing daily use of digital devices, consumption of news is also shifting to digital channels. Many readers are involved in the formation of public opinion through comments on news articles. Therefore, analyzing comments on online news can be said to be very important for understanding new media usage behavior. The purpose of this study was to analyze the comment texts and the response using the word embeddings based on artificial neural network, along with some psycho-linguistic variables. Specifically, we used the embeddings to identify the vocabulary of subject categories such as politics, economy, life and culture, and also examined the responses to comments using psycho-linguistic variables. As the result of the analysis, we confirmed the significant influences of the psycho-linguistic variables and topic vocabulary, and the significant interaction between news characteristics and comment sentiments.

Keywords
online news, text mining, embedding, psycho-linguistics

* 홍익대학교 경영학과 교수
- ORCID: http://orcid.org/0000-0003-1431-7159

· Received: Mar. 18, 2020, Revised: May 25, 2020, Accepted: May 28, 2020
· Corresponding Author: Hyung Jun Ahn
College of Business Administration, Hongik University, Korea
Tel.: +82-2-320-1730, Email: Hjahn@hongik.ac.kr

Ⅰ. 서 론

스마트폰 등 모바일 디지털 기기의 사용이 보편화됨에 따라 뉴스의 소비도 종이 지면이 아닌 디지털 기반의 매체로 옮겨가고 있다. 전 세계적으로 신문 구독자 수는 감소하는 한편, 반대로 온라인 뉴스의 소비는 증가하고 있다. 사용자들은 컴퓨터 혹은 모바일 기기를 이용해서 온라인 뉴스를 소비하는데, 이때 주로 페이스북과 같은 소셜네트워크서비스, 혹은 검색 포털이 제공하는 뉴스 서비스를 많이 이용하게 된다[1][2]. 한국에서는 특히 포털 사이트의 웹페이지나 모바일 앱을 통한 뉴스의 소비가 많은 편인데, 이때 많은 독자들이 뉴스 기사에 대한 댓글을 통해 여론 형성에 참여하고 쌍방향 커뮤니케이션을 수행한다. 따라서 온라인 기사의 댓글을 분석하는 것은 이러한 여론 형성 과정을 이해하는데 매우 중요하다고 할 수 있다.

그러한 중요성에도 불구하고 국내의 온라인 뉴스 댓글 분석에 대한 학술적 연구는 아직까지 충분치 않은 편이다. 일부 선행 연구에서 온라인 댓글 텍스트의 일부 특징을 분석하거나, 혹은 특정한 이슈에 대한 댓글의 내용을 분석한 바가 있으나, 댓글 텍스트의 다양한 특성과 댓글에 대한 반응 간의 관계에 대해 분석한 연구는 많지 않다.

이와 같은 배경에서 본 연구는 최근 점점 많이 활용되고 있는 인공 신경망 기반 임베딩(Embedding)과, 사회과학 분야에서 많이 연구되어 온 언어심리학적 변인들을 이용하여 댓글 및 그에 대한 반응을 분석하고자 하였다. 구체적으로는 임베딩 방식을 이용하여 댓글에 포함된 정치, 경제, 생활, 문화 등 주제 범주 어휘들을 파악하고, 또한 언어심리학적 특징과 관련된 변수들을 이용하여, 그러한 특징들이 댓글에 대한 반응에 어떠한 영향을 끼치는지 살펴보았다. 추가적으로 댓글에 대한 반응이 뉴스 기사의 특성과 그에 대한 댓글의 특성 간의 상호 관계에 어떠한 영향을 받는지도 분석하였다. 웹 크롤링을 이용하여 국내의 주요 포털 사이트 중 하나의 인기 뉴스 기사들의 댓글들을 수집하고, 이를 Python 기반의 한국어 분석 도구를 이용하여 처리한 후 통계 분석을 실시하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대한 요약이 제시되며, 3장에서는 연구의 목적 및 방법을 설명한다. 4장은 분석의 결과 및 토론을, 5장은 결론 및 향후 연구 과제를 제시한다.

Ⅱ. 관련 연구

2.1 온라인 뉴스 댓글 분석

뉴스의 소비에서 온라인 채널의 중요성은 점점 커지고 있다. 국내에서 이제까지 수행 된 온라인 뉴스 댓글에 대한 연구들은 크게 두 가지 범주로 구분될 수 있다. 첫째 범주의 연구에서는 주로 온라인 댓글의 특성들에 초점을 맞추고 이를 통계적, 계량적으로 분석한다. 예를 들어 영어권과 한국을 대상으로 악성 댓글에 드러난 언어적 특성을 비교한 연구[3], 소셜 댓글과 일반 댓글의 특성 간 차이를 분석한 연구[2], 온라인 토론의 댓글 트리 구조를 정량적으로 분석한 연구[4] 등이 이에 포함된다. 두 번째 범주의 연구에서는 특정한 주제에 초점을 맞추고 그에 관련된 뉴스 기사들의 댓글을 분석한다. 이러한 연구들은 해당 주제에 대한 뉴스의 댓글에 나타나는 사용자들의 주된 관심사항을 파악하는 것이 목적이라고 할 수 있다. 예를 들어 한 연구에서는 월드컵에 대한 기사들을 분석하여 댓글에 드러난 독자들의 주요 관심사, 관심을 받은 경기 혹은 선수 등을 분석하였다[5]. 또 다른 연구에서는 댓글 분석을 통해 신고리 원전 5·6호기에 대한 주요 이슈들과 이들의 시간별 변화를 분석한 바 있다[6].

본 연구는 첫 번째 범주의 연구에 해당한다고 볼 수 있으며, 특히 온라인 댓글에 포함된 언어심리학적 특성을 활용한다는 점, 또한 인공신경망 기반 임베딩을 활용하여 댓글의 주제어휘의 영향을 함께 분석한다는 점에서 기존 연구와 차별화 된다. 기존 연구 중 언어심리학적 변인을 활용한 연구가 일부 존재하기는 하나[3], 본 연구와 같이 언어심리학적 특성이 댓글에 대한 반응에 끼치는 영향을 분석한 연구는 찾아보기 힘들다.

2.2 컴퓨터 기반 텍스트 분석

텍스트 분석은 정보 검색 분야에서 오랫동안 연

구되어 온 주제이며 구글과 같은 검색 엔진에 널리 활용되고 있다. 전통적인 텍스트 분석은 텍스트를 분류하거나, 혹은 텍스트를 대표하는 키워드를 추출하고, 이 키워드를 통해 원하는 문서를 빠르고 효과적으로 검색하는 데 그 목적이 있다. 최근에는 그러한 용도 외에도 텍스트의 감정 분석, 주요 키워드 간의 관계 도출, 텍스트 요약, 기계 번역 등 분석의 목적과 기법이 매우 다양해지고 있다.

한편 심리학의 세부 분과인 언어심리학 분야에서도 컴퓨터를 이용해 텍스트를 분석하고 이를 통해 사람들의 심리를 이해하고자 오랫동안 연구해 온 바 있다. 언어심리학 분야에서는 사람들의 말 혹은 글에 드러난 특징들이 각 개인의 심리적 특성을 반영한다고 보고 이를 사회과학적으로 연구해 왔다. 예를 들어 특정한 대명사 군이나 특정한 시제의 활용 빈도가 개인의 고유한 성격 유형 및 특성과 관련 있음이 제시된 바 있다[7].

본 연구에서는 그러한 언어심리학적 변수 중 구체적으로 '사회적 어휘', '인지적 어휘', 두 가지 특징에 초점을 맞추어 댓글을 분석하고자 하였다. 이때 사회적 언어는 사람들의 사회적 관계 혹은 행위와 관련된 어휘로, 예를 들어 '친구', '가족', '대화' 등이 이에 해당한다고 볼 수 있다. 인지적 어휘는 사람들의 인지적 사고 과정과 관련된 어휘들로, 주로 판단, 추론, 영감, 분석 등과 관련된 어휘이다. 예를 들어, '이유', '왜냐하면', '만약', '생각' 등이 이에 해당한다고 할 수 있다[8]. 연구에 의하면 각 개인의 특성과 지향에 따라 이러한 어휘들의 사용 빈도가 달라지는 것으로 알려져 있다[7][9].

최근에는 급속히 발전한 인공지능 연구의 결과로 텍스트 분석 기법 또한 많이 발전하고 있으며, 특히 텍스트의 구성 요소를 인공신경망에 바탕을 둔 임베딩으로 표현하는 연구들이 많이 등장하여 널리 활용되고 있다[10][11]. 임베딩은 단어, 문장 등의 텍스트 구성 요소를 숫자로 구성된 벡터로 표현하는 것을 의미하며, 이는 곧 각 요소들이 다차원 공간의 좌표로 표현될 수 있음을 의미한다. 이때 임베딩 벡터들은 각 텍스트 구성 요소의 잠재적 의미(Latent semantics) 및 문법적 특징을 반영하게 되며, 따라서 텍스트 분석에 매우 효과적인 수단이 된다. 예를 들어, 단어 임베딩을 사용하면 "France - Paris + England = X"와 같은 방정식 연산을 통해 "X = London"이라는 결과 값을 도출해 낼 수 있다[12]. 본 연구에서도 Word2Vec 임베딩 방식 중 스킵그램(Skip-gram) 기법을 활용하였다[13].

III. 연구의 목적 및 방법

3.1 연구의 목적

본 연구의 목적은 온라인 뉴스의 댓글을 분석하여 댓글의 어떠한 특징들이 댓글에 대한 반응에 영향을 끼치는 지 분석하는 것이다. 구체적으로 본 연구가 답하고자 하는 질문들은 다음과 같다. 첫째, 온라인 뉴스 댓글의 특성, 특히 언어심리학적 어휘 및 특정 주제 범주의 어휘 활용은 댓글에 대한 긍정, 부정 반응에 어떠한 영향을 끼치는가? 둘째, 온라인 뉴스 댓글의 특성과 기사의 특성 간의 상호관계가 댓글에 대한 반응에 영향을 끼치는가? 본 연구는 뉴스 댓글 텍스트와 그 반응에 대한 데이터를 수집해 분석함으로써 이러한 질문에 답하고, 온라인 뉴스 사용자들의 여론 형성 과정을 이해하고자 하였다.

3.2 연구 방법

구체적인 연구의 방법은 다음과 같다. 첫째, 많은 사용자들이 관심을 가지고 있는 주요 기사들에 대한 댓글을 분석하기 위해 국내 대표적 포털 사이트가 제공하는 뉴스 서비스를 분석 대상으로 하였다. 이를 위해 해당 포털 사이트에서 제공하는 뉴스 중, 클릭한 횟수를 기준으로 매일 특정 시점에 가장 인기 있는 뉴스 기사 20개에 대해, 각각 총 500개까지의 댓글을 1개월 동안 수집하였다. 댓글은 웹 크롤링 방식을 활용하여 많은 사람들이 공감한 비율 순으로 상위 순위부터 수집 하였다. 기사의 댓글이 500개가 넘는 경우에는 500개까지만, 그 미만일 경우 모든 댓글을 수집하였다. 이렇게 하여 1개월, 즉 31일 간 총 256,890개의 댓글을 수집하였고 이는 1개의 기사 당 평균 약 414.3개의 댓글이 수집되었음

을 의미한다. 이때 어절 수가 적은 경우 댓글에서 충분한 특징을 추출할 수 없어 본 연구의 취지에 부합하지 않는다고 보고, 어절 수 20개 미만인 경우를 제외한 후 총 123,972개의 데이터를 실제 분석에 활용하였다. 표 1에는 수집된 데이터에 대한 요약 정보가 제시되어 있다.

표 1. 수집 데이터 요약
Table 1. Summary of collected data

수집 기간	31일
수집 대상 기사	매일 특정 시점 포털의 인기 기사 상위 20개(클릭 수 기준)
수집 대상 댓글	기사 당 최대 500개 (공감 비율 순)
총 댓글 수	256,890개
평균 댓글 수	414.3개
분석 대상 댓글 수	123,972개 (어절 20개 미만 제외 후)

표 2. 분석에 사용된 변수
Table 2. Variabels used for analysis

	Variable	Code
분석변수	언어심리변수 - 사회적 어휘(어절 중 비율) - 인지적 어휘(어절 중 비율) 긍정 감정 어휘(어절 중 비율) 부정 감정 어휘(어절 중 비율) 주제별 어휘 비율(어절 중 비율, 정치/경제/생활/문화)	 Soc Cog Pos Neg Pol, Econ, Life, Cul
종속변수 (댓글 반응)	댓글 좋아요(log 값) 댓글 싫어요(log 값) 댓글 종합(좋아요 + 싫어요, log 값)	RLike RHate RPlus
통제변수	어절 수 기사 긍정 반응 합계(log 값) 기사 부정 반응 합계(log 값) 댓글 작성 시간 (기사 작성 시간과의 초단위 차, log 값)	eoj APos ANeg TimeDiff

표 3. 각 범주 별 단어들의 예
Table 3. Examples of words for each category

정치	경제	생활	문화
정치인	내수	사회생활	유교
공작	남북관계	직장	매너
구태	세계경제	내무	한류
양당	파탄	휴식	예술
이념	위축	여가	융성
보복	제조업	출퇴근	영화
...	...	...	...

수집된 댓글 텍스트의 처리를 위해 한국어 처리 공개 소프트웨어인 Python 기반의 KoNLPy를 활용하였다[14]. KoNLPy의 태거(Tagger)를 활용하여 모든 단어를 형태소로 변형한 후, 역시 Python 기반 공개소프트웨어인 gensim을 활용하여 임베딩으로 변환하였다[15]. 구체적으로는 앞서 소개된 스킵그램 방식의 임베딩을 활용하였으며, 벡터의 차원은 200으로 설정하였다.

다음으로, 댓글에 포함된 언어심리학적 변수들과 댓글에 대한 반응 간의 관계를 분석하기 위해 회귀분석을 실시하였다. 회귀분석을 위해 사용된 변수들은 표 2에 정리되어 있다.

우선, 분석 대상 변수로는 앞서 2절에서 소개된 두 가지 언어심리학적 변수인 사회적 어휘와 인지적 어휘, 감정분석에서 많이 활용되는 감정 어휘(긍정 및 부정 각각), 임베딩 기반의 주제별 어휘 등을 사용하였다. 사회적 어휘, 인지적 어휘, 감정 어휘들의 경우 LIWC를 바탕으로 제작된 한국어 분석 도구인 HLIWC를 이용하여 추출된 값을 활용하였다[9].

주제별 어휘 변수의 경우 뉴스 기사들이 크게 '정치', '경제', '생활', '문화' 등으로 구분된다는 점을 바탕으로 이와 유사하게 네 가지로 구성하였다. 이를 위해 앞서 생성된 임베딩 벡터를 활용하여 댓글의 각 어휘들과 네 가지 주제어와의 코사인(Cosine) 기반의 유사도를 계산하고, 그 값이 .3 이상인 경우 각 주제 범주와 관련이 있는 것으로 간주하였다. 각 주제 범주 별 어휘들의 예가 표 3에 제시되어 있다. 그 다음 해당 어휘들의 유사도 합을 각 댓글에 대해 각각의 주제를 대표하는 변수로 활용하였다.

종속변수는 댓글에 대한 반응으로 좋아요, 싫어요, 그리고 이 둘을 합한 값 등 총 세 가지를 활용하였다. 통제변수로는 전체 어절 수, 기사 자체에 대한 긍정 혹은 부정 반응 수, 댓글 작성 시간과 기사 작성 시간 간의 차이 등을 활용하였다. 이 중 기사에 대한 긍정 반응은 포털사이트가 각 기사에 대해 집계해 제공하는 '좋아요', '훈훈해요', '후속기사 원해요'를 합한 값을 활용했으며, 부정 반응은 '슬퍼요', '화나요' 반응을 합한 값을 활용하였다.

Ⅳ. 연구 결과 및 분석

4.1 댓글 반응에 대한 영향 요인 분석

표 4에는 주요 변수들의 기술 통계가 요약되어 있다. 분석 대상 댓글들의 평균 어절 수는 약 49개이며, 최댓값은 300이다. 이때 최댓값은 해당 포털 사이트가 허용하는 댓글의 길이 제한에 의한 것이다. 인지적 어휘는 댓글 당 평균 2.4개, 최대 25개가 등장하는 것을 알 수 있으며, 사회적 어휘는 평균 1.48개, 최대 30개가 등장하는 것을 알 수 있다. 주제 관련 단어의 경우 경제 및 정치에 관한 어휘들이 평균적으로 가장 많이 포함된 것으로 나타났다. 댓글에 대한 '좋아요'의 수는 평균 56.5개, 최대 약 38,000개임을 알 수 있다. '싫어요'의 개수는 평균 4.4개, 최대 3,234개로 '좋아요' 반응보다는 대체로 훨씬 적음을 알 수 있다. 기사에 대해서는 평균적으로 부정 반응이 긍정 반응보다 약 2배 정도 많은 것으로 나타났다. 댓글이 작성된 시간은 기사 작성 후 최대 하루 정도, 평균 4시간 정도인 것으로 나타났다.

표 5에 회귀분석 결과가 정리되어 있다. '좋아요'가 종속변수인 경우 R^2 값이 약 27 퍼센트, '싫어요'의 경우 약 12 퍼센트로 나타나, 연구 모델이 '좋아요' 반응을 더 잘 설명함을 알 수 있다. 두 경우 모두 전체 모델의 F 값은 높은 수준에서 유의한 것으로 나타나고 있다.

언어심리학적 변수들의 경우 '좋아요' 및 '싫어요' 반응 모두에 대해 유의한 영향을 끼치는 것으로 나타났다. 이때 인지적 어휘의 '싫어요' 반응의 경우만 부의 방향으로, 나머지는 모두 정의 방향으로 유의한 영향을 끼침을 알 수 있다.

표 4. 주요 변수들의 기술 통계
Table 4. Descriptive statistics of key variables

	Min	Max	Average
Soc	0	30	1.48
Cog	0	25	2.40
Econ	0	46	1.65
Life	0	24	.53
Pol	0	69	1.42
Cul	0	72	1.13
eoj	20	300	49.19
RLike	0	38482	56.54
RHate	0	3234	4.40
APos	0	9654	484.98
ANeg	0	23176	954.17
TimeDiff	0	86340sec (23.98 time)	17340.19sec (4.82 time)

표 5. 댓글 반응 '좋아요' 및 '싫어요'에 대한 분석
Table 5. Analysis of 'like' or 'hate' reaction to comments

	Variable	댓글 좋아요(log)			댓글 싫어요(log)		
		B	t	p	B	t	p
	Constant	3.093	89.683	.000***	3.260	121.398	.000***
통제 변수	APos	.217	76.378	.000***	.059	26.655	.000***
	ANeg	.259	109.991	.000***	-.063	-34.500	.000***
	TimeDiff	-.403	-130.412	.000***	-.278	-115.767	.000***
	eoj	.000	4.275	.000***	.001	12.936	.000***
독립 변수	Pos	-.051	-.234	.815	-.020	-.120	.904
	Neg	-.337	-1.984	.047*	.028	.209	.835
	Soc	1.180	12.166	.000***	.353	4.668	.000***
	Cog	.335	4.020	.000***	-.155	-2.382	.017*
	Econ	3.335	13.557	.000***	-.143	-.748	.455
	Life	4.820	15.148	.000***	-2.002	-8.081	.000***
	Pol	.019	.088	.930	5.574	33.711	.000***
	Cul	-2.666	-9.918	.000***	-2.173	-10.384	.000***
	모형	R^2=.266, F=3739.85 (p=0.000)			R^2=.122, F=1434.98 (p=0.000)		

(*: $p < 0.05$, **: $p < 0.01$, ***: $p<0.001$)

감정 어휘의 경우 부정 감정 어휘만 낮은 유의수준에서 '좋아요'에 부의 영향을 끼치는 것을 알 수 있으며, 나머지의 경우는 모두 유의하지 않은 것으로 나타나고 있다. 주제어들의 경우에도 대체로 유의한 영향을 끼침을 확인할 수 있다. 그러나 정치 주제의 경우 '좋아요' 반응에, 경제 주제의 경우 '싫어요' 반응에 유의한 영향을 끼치지 않음을 알 수 있다. 또한 문화 주제의 경우 '좋아요', '싫어요' 반응 모두에, 사회 주제의 경우 '싫어요' 반응에 부의 영향을 끼치며, 나머지는 모두 정의 영향을 끼침을 알 수 있다.

통제변수들의 경우도 대부분 유의한 영향을 끼치는 것으로 나타나고 있다. 대상 기사에 대한 반응이 클수록 반응의 긍정, 부정 여부와 무관하게 댓글에 대한 '좋아요' 반응도 커짐을 알 수 있다. 그러나 기사에 대한 부정 반응이 커지면 댓글에 대한 '싫어요' 반응은 오히려 적어짐을 알 수 있다. 댓글이 작성된 시점이 늦을수록 '좋아요' 및 '싫어요' 모두에 부의 영향을 끼치는 것으로 나타나고 있어서, 빨리 작성된 댓글일수록 그에 대한 반응이 큼을 알 수 있다.

다음으로 표 6에는 '좋아요'와 '싫어요' 반응을 합한 종합 반응, 즉 전체 반응 수준에 대한 분석 결과가 정리되어 있다.

표 6. 종합 댓글 반응에 대한 분석
Table 6. Analysis of overall reaction to comments

	Variable	B	t	p
	Constant	3.795	110.866	.000***
통제 변수	APos	.194	68.830	.000***
	ANeg	.206	88.352	.000***
	TimeDiff	-.417	-136.090	.000***
	eoj	.001	7.584	.000***
독립 변수	Pos	-.105	-.485	.628
	Neg	-.267	-1.583	.113
	Soc	1.169	12.139	.000***
	Cog	.218	2.637	.008**
	Econ	2.578	10.559	.000***
	Life	4.000	12.665	.000***
	Pol	1.425	6.760	.000***
	Cul	-2.706	-10.146	.000***
	모형	R^2=.238, F=3228.22 (p=0.000)		

(*: p < 0.05, **: p < 0.01, ***: p<0.001)

언어심리학적 어휘들은 모두 댓글 반응에 정의 영향을 끼침을 알 수 있다. 감정 어휘는 모두 유의한 영향을 끼치지 않으며, 반대로 주제어들은 모두 유의한 영향을 끼치지만 그 중 문화 주제의 경우에만 부의 영향을 끼침을 알 수 있다.

4.2 기사 특성과 댓글 감정의 상호작용 분석

다음으로, 기사의 특성과 감정 어휘 간의 상호작용 효과에 대한 분석을 실시하였다. 앞선 분석에서 댓글의 긍정 및 부정 감정 어휘의 비율은 대부분 댓글 반응에 유의한 영향을 끼치지 않는 것으로 나타났다. 그러나 그 영향은 기사 자체의 특성, 즉 기사에 대한 전반적 긍정 반응 및 부정 반응의 정도에 따라 다를 수 있다고 보고, 두 특성 사이의 상호작용 효과를 분석하였다. 이를 위해 기사 자체에 대한 긍정 및 부정 반응에 대한 변수와, 댓글의 긍정 및 부정 감정 어휘 비율에 대한 변수를 서로 곱하여 총 네 개의 상호작용 항목 변수를 생성하였다. 이때 생성된 상호작용 변수의 다중공선성 문제를 제거하기 위해 관련 변수들은 모두 정규화 한 Z 값을 이용하였다.

표 7에 분석 결과가 요약되어 있다. 제시된 표에는 상호작용 항목만 포함되어 있으며, 생략된 나머지 변수들의 계수 부호 및 유의확률은 앞선 종합 분석과 동일하였다.

표 7. 기사 특성과 댓글 감정 어휘 상호 작용
Table 7. Interaction between comment sentiments and article characteristics

	댓글 좋아요(log)			댓글 싫어요(log)		
	B	t	p	B	t	p
기사긍정 × 댓글긍정	.015	3.046	.002**	.016	3.988	.000***
기사긍정 × 댓글부정	.007	1.438	.151	.005	1.246	.213
기사부정 × 댓글긍정	-.015	-3.027	.002**	.000	-.120	.905
기사부정 × 댓글부정	-.011	-2.257	.024*	-.017	-4.476	.000***

(*: p < 0.05, **: p < 0.01, ***: p<0.001)

분석 결과를 보면, ‘좋아요’ 반응의 경우 기사에 대한 긍정 반응이 클수록 긍정적 어휘 사용에 정의 영향을, 반대로 기사에 대한 부정 반응이 클수록 긍정 혹은 부정 어휘 사용에 모두 부의 영향을 받는 것으로 나타났다. ‘싫어요’ 반응의 경우 기사의 긍정 반응이 클수록 긍정적 감정 어휘에 정의 영향을, 기사의 부정적 반응이 클수록 부정적 감정 어휘에 부의 영향을 받는 것으로 나타났다.

4.3 주요 분석 결과 요약

분석을 통해 다음과 같은 주요 사항들이 파악되었다. 첫째, 모형에 포함된 언어심리학적 변수인 사회적, 인지적 어휘들이 모두 댓글 반응에 유의한 영향을 끼치는 것으로 나타났다. 이는 이제까지 많은 한국어 텍스트 분석에서 간과되어 왔던 언어심리학적 변인들의 중요성을 확인해 주는 결과라고 할 수 있다.

둘째, 임베딩을 활용해 분석한 댓글의 주제 범주 어휘들의 영향도 대체로 유의한 것으로 나타났다. 특히 정치적인 주제의 경우 ‘싫어요’ 반응을 증가시키며, 경제, 생활 주제의 경우 ‘좋아요’ 반응을 증가시키는 것을 확인할 수 있었다. 정치적 주제에 대한 부정적 반응은 대체로 한국의 여론이 정치 관련한 전반적인 주제들에 대해 매우 비판적이기 때문인 것으로 보인다. 한편 문화적 주제들은 긍정 혹은 부정 댓글 반응을 모두 유의하게 감소시키는 것으로 나타는데, 이는 국내에서 온라인 뉴스의 소비 및 댓글 활동에 대한 참여 동기가 대체로 문화적인 이슈보다는 정치, 경제, 생활 등과 더욱 관련이 있기 때문으로 추측할 수 있다.

셋째, 기본 모형의 회귀분석 결과에서는 감정 어휘의 사용이 긍정, 부정 유무에 무관하게 댓글 반응에 대부분 영향을 끼치지 않는 것으로 나타났으나, 추가적인 상호작용 효과의 분석 결과, 원래 기사의 특성에 따라 감정 어휘의 사용이 유의한 영향을 줄 수 있는 것으로 나타났다. 이러한 결과는 그림 1에 요약되어 있다. 즉, 원래 기사에 대한 반응이 전반적으로 긍정적, 혹은 부정적인지에 따라 댓글에 포함된 감정 어휘의 영향이 달라질 수 있음을 의미한다. 또한 긍정 반응이 큰 기사일수록 댓글에 포함된 감정 어휘들이 댓글에 대한 반응을 증가시키는 반면, 부정 반응이 큰 기사일수록 댓글의 감정 어휘들이 댓글에 대한 반응을 반대로 감소시키는 것을 확인할 수 있다.

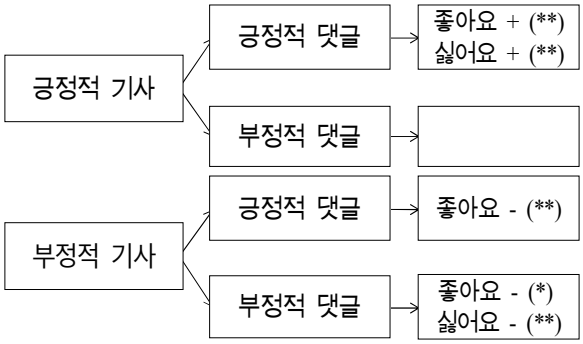

그림 1. 상호작용 효과 분석 결과 요약
Fig. 1. Summary of the analysis of the interaction effect
(*: p<0.05, **: p<0.01)

Ⅴ. 결론 및 연구의 한계

본 연구는 인공 신경망 기반의 임베딩 및 언어심리학적 변인들을 이용하여 온라인 뉴스의 댓글 특성과 그에 대한 반응 간의 관계를 분석하였다. 이를 위해 웹 크롤링을 통해 수집한 댓글에서 관련된 특징들을 추출하고, 그러한 특징들이 댓글에 대한 긍정 혹은 부정 반응에 어떠한 영향을 끼치는지 분석하였다.

본 연구의 주된 결과는 다음과 같다. 첫째, 언어심리학적 특징인 사회적 어휘, 인지적 어휘의 사용은 댓글 반응에 유의한 영향을 끼치는 것으로 나타났다. 둘째, 임베딩을 활용하여 추출한 정치, 경제, 생활, 문화 등의 주제 범주 어휘들도 댓글에 대한 반응에 다양한 방식으로 유의한 영향을 끼치는 것으로 나타났다. 특히 정치적 주제어가 많을수록 부정 반응이 증가하고, 경제 및 생활 관련 주제어가 많을수록 긍정 반응이 증가하는 것으로 나타났다. 반면 문화 관련된 주제어는 댓글 반응을 전반적으로 감소시키는 것으로 나타났다. 셋째, 감정어휘의 사용은 기사의 특성에 따라 상이한 반응을 가져오는 것으로, 즉 기사 특성과의 상호작용 효과가 존재하는 것으로 나타났다. 이는 본래의 기사에 대한 전

반적 반응이 긍정적인지, 혹은 부정적인지에 따라 댓글의 감정 어휘 사용에 대한 반응이 달라질 수 있음을 의미한다.

본 연구의 주된 의의는 국내의 온라인 뉴스 댓글에 대한 연구가 많지 않은 상황에서 신경망 기반의 임베딩 및 언어심리학적 변인을 활용한 뉴스 댓글 분석의 사례를 제시했다는 데 있다. 이와 같은 분석 방법은 향후 온라인 뉴스 뿐 아니라 다른 분야의 사용자 텍스트 분석에 다양하게 응용될 수 있다. 또한 분석 결과 언어심리학적 변수들 및 주제 어휘들의 유의한 영향, 뉴스 특성과 댓글 특성 간의 유의한 상호작용 등이 확인되었으며, 이를 바탕으로 한 향후 연구에 대한 가능성도 제시했다고 할 수 있다.

본 연구의 한계는 다음과 같다. 첫째, 본 연구의 분석 대상 텍스트는 제한된 기간 동안 특정한 포털 사이트의 인기 뉴스로 국한되어 있다. 따라서 본 연구의 결과를 다른 특성을 가진 온라인 뉴스에 대해 적용하는 경우는 주의를 요한다. 둘째, 한국어 텍스트를 분석하는 구체적 방식에 따라 연구 결과는 달라질 수 있다. 예를 들어 임베딩을 구축하거나 이용하는 구체적인 방식, 주제 범주 어휘의 포함 범위 등이 분석 결과에 영향을 끼칠 수 있다.

References

[1] W. Kim, H. I. Jo, and B. G. Lee, "Analyzing the Characteristics of Online News Best Comments", Journal of Digital Contents Society, Vol. 19, No. 8, pp. 1489-1497, Aug. 2018.

[2] S. Kim & S. Yang, "A Comparative Analysis between General Comments and Social Comments on an Online News Site", The Journal of the Korea Contents Association, Vol. 15, No. 4, pp. 391-406, Apr. 2015.

[3] Y. Kim, Y. Kim, Y. Kim, and K. Kim, "The Characteristics of Malicious Comments: Comparisons of the Internet News Comments in Korean and English", The Korea Contents Association, Vol. 19, No. 1, pp. 548-558, Jan. 2019.

[4] S. H. Kim, H. Tak, and H. G. Cho, "User Characterization from Replying Comment Structures in Online Discussion", The Korea Contents Association, Vol. 18, No. 11, pp. 135-145, Nov. 2018.

[5] S. Park, G. Won, and S. Lee, "Web News Comment-based Sentiment Analysis of the South Korean National Team Members in the 2014 Brazil World Cup", Journal of Korean Society for Sport Management, Vol. 20, No. 2, pp. 13-28, Apr. 2015.

[6] C. Kim and Y. Choi, "A Study on Analysis of the Issues on the Public Deliberation Committee on Shin-Gori No. 5 & 6 through Comments of Online News", Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology, Vol. 9, No. 5, pp. 317-326, May 2019.

[7] Y. R. Tausczik and J. W. Pennebaker, "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods", J. Lang. Soc. Psychol., Vol. 29, No. 1, pp. 24-54, Jan. 2010.

[8] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: LIWC", in Mahway: Lawrence Erlbaum Associates, 71, 2001.

[9] J. Lee and H. J. Ahn, "Impact of the Psycho-linguistic Features of the Facebook Brand Page Posts on User Reaction", The Journal of internet electronic commerce research, Vol. 16, No. 1, pp. 37-56, Feb. 2016.

[10] D. Ganguly, D. Roy, M. Mitra, and G. J. Jones, "Word embedding based generalized language model for information retrieval", in Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, Santiago, Chile, pp. 795-798, 2015.

[11] L. Vilnis and A. McCallum, "Word representations via gaussian embedding", ArXiv Prepr. ArXiv14126623, 2014.

[12] K. Ethayarajh, D. Duvenaud, and G. Hirst,

"Towards Understanding Linear Word Analogies", ArXiv181004882 Cs, 8 2019, http://arxiv.org/abs/1810.04882, [accessed, Feb 04, 2020]

[13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality", in Advances in Neural Information Processing Systems 26 (NIPS 2013), Nevada, USA, pp. 3111-3119, 2013.

[14] E. L. Park and S. Cho, "KoNLPy: Korean natural language processing in Python", in Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology, Chuncheon, Korea, pp. 133-136, 2014.

[15] R. Rehurek and P. Sojka, "Gensim-python framework for vector space modelling", NLP Cent. Fac. Inform. Masaryk Univ. Brno Czech Repub., Vol. 3, No. 2, 2011.

저자소개

안 형 준 (Hyung Jun Ahn)

2004년 2월 : KAIST
경영공학(공학박사)
2008년 3월 ~ 현재 : 홍익대학교
경영대학 교수
관심분야 : 지능정보시스템,
인공지능, 빅데이터, 문화예술과
IT 등