J. Inf. Technol. Appl. Manag. 25(1): 19~32, March 2018
https://doi.org/10.21219/jitam.2018.25.1.019

ISSN 1598-6284 (Print)
ISSN 2508-1209 (Online)

텍스트 마이닝과 기계 학습을 이용한 국내 가짜뉴스 예측

윤태욱* · 안현철**

Fake News Detection for Korean News Using Text Mining and Machine Learning Techniques

Tae-Uk Yun* · Hyunchul Ahn**

Abstract

Fake news is defined as the news articles that are intentionally and verifiably false, and could mislead readers. Spread of fake news may provoke anxiety, chaos, fear, or irrational decisions of the public. Thus, detecting fake news and preventing its spread has become very important issue in our society. However, due to the huge amount of fake news produced every day, it is almost impossible to identify it by a human. Under this context, researchers have tried to develop automated fake news detection method using Artificial Intelligence techniques over the past years. But, unfortunately, there have been no prior studies proposed an automated fake news detection method for Korean news.

In this study, we aim to detect Korean fake news using text mining and machine learning techniques. Our proposed method consists of two steps. In the first step, the news contents to be analyzed is convert to quantified values using various text mining techniques (Topic Modeling, TF-IDF, and so on). After that, in step 2, classifiers are trained using the values produced in step 1. As the classifiers, machine learning techniques such as multiple discriminant analysis, case based reasoning, artificial neural networks, and support vector machine can be applied.

To validate the effectiveness of the proposed method, we collected 200 Korean news from Seoul National University's FactCheck (http://factcheck.snu.ac.kr). which provides with detailed analysis reports from about 20 media outlets and links to source documents for each case. Using this dataset, we will identify which text features are important as well as which classifiers are effective in detecting Korean fake news.

Keywords : Fake News Detection, Korean News, Machine Learning, Text Mining

Received : 2018. 01. 29. Revised : 2018. 03. 14. Final Acceptance : 2018. 03. 15.

* Master's Candidate, Graduate School of Business IT, Kookmin University, e-mail : ytu100@kookmin.ac.kr
** Corresponding Author, Associate Professor, Graduate School of Business IT, Kookmin University, 77, Jeongneung-ro, Seoungbuk-gu, Seoul, 02707, Republic of Korea, Tel : +82-2-910-4577, Fax : +82+2-910-4017, e-mail : hcahn@kookmin.ac.kr

1. 서 론

뉴스는 매스미디어(Mass Media) 매체를 통해 대중에게 아직 알려지지 않은 새로운 소식과 관련한 정보를 제공하는 언론 시장의 가장 중요한 매개체 중 하나이다. 대중들은 다양한 뉴스 가운데 관심 있는 것들만 주목하여 취사 선택하는데, 수동적 입장에 있는 대중으로서는 주체적 입장에 있는 매스미디어의 뉴스 보도 내용을 대부분 진실된 정보로 수용 할 수 밖에 없게 된다.

그러나, 가짜뉴스(Fake News)는 사실이 아닌 정보들을 포함하고 있어 뉴스 정보에 대한 신뢰를 구축하고 공유를 활성화하는데 있어 저해를 일으키고 있다. 버즈피드의 조사에 따르면 지난 2016년 제 45대 미국 대선 기간 중 가짜뉴스가 주요 매체의 실제 뉴스보다 더 많은 반응을 이끌어냈다[Institute for Korean Democracy, 2017]. 이는 정치적인 문제와 연결되어 전 세계적으로 많은 사람들에게 가짜뉴스의 심각성을 깨닫게 하는 계기가 되었다. 또한 현대경제연구원의 연구에 따르면 가짜뉴스로 인한 국내 경제적 비용은 약 30조 900억 원이 발생하는 것으로 추정하고 있다[Hyundai Research Institute, 2017]. 이와 같이 가짜뉴스는 현실세계에 직접적으로 심각한 정치적, 경제적 피해를 미치고 있다.

때문에 가짜뉴스를 탐지하여 구별해 내는 것은 매우 중요하다. 그러나, 하루에도 수많은 양의 가짜뉴스가 생성되고 확산되고 있어, 사람이 모든 것을 하나씩 구분해내는 것은 거의 불가능하다. 아울러, 가짜뉴스를 누구나 쉽게 구별해내기는 어려우며, 개인의 주관적인 견해가 반영 될 여지가 많기 때문에, 전문가들이 사실을 근거로 객관적인 검증을 통해 구별해 내야 한다.

이러한 이유로 자동화된 가짜뉴스 탐지 방법론 개발에 대한 정부와 산업계의 요구가 최근에 매우 커졌으며, 학계 역시 이에 부응하고자 많은 학자들에 의해 연구가 진행되고 있다. 그러나, 해외에서는 텍스트 마이닝(Text Mining)과 기계 학습(Machine Learning)을 이용하여 다양한 연구가 이루어지고 있는 반면 국내에서는 자동화된 가짜뉴스 탐지에 관련된 거의 논문이 전무한 상황이다. 이는 국내 뉴스에 대해 가짜와 사실로 레이블 되어 있는 데이터를 확보하는 것이 매우 어렵기 때문이다.

이러한 배경에서 본 연구는 자동화된 국내 가짜뉴스 예측 모형의 구축을 목표로 하여, 인공지능(Artificial Intelligence) 기법을 이용한 가짜뉴스 탐지 방법론을 제안한다. 본 연구의 제안 방법론은 텍스트 마이닝의 종류 중 하나인 토픽 모델링(Topic Modeling) 기반의 가짜뉴스 다분류(multiclass classification) 예측에 기계 학습 기법을 적용하도록 설계되어 있다.

본 연구에서는 제안 방법론을 수집한 데이터에 적용하여 그 효과를 확인하고, 이를 통해 활용되는 데이터의 정보 중 어떤 정보가 가짜뉴스를 예측하는데 더 큰 의미가 있는지 고찰해 보고자 하였다.

본 논문의 뒷부분은 다음과 같이 구성된다. 우선 제 2장 연구배경에서는 가짜뉴스, 텍스트 마이닝 등 본 연구의 기초가 되는 개념들의 이론적 배경과 자동화된 가짜뉴스 탐지 방법론에 대한 연구 현황을 간략히 살펴본다. 제 3장에서는 제안 방법론에 대해서 설명한다. 제 4장에서는 앞서 제시한 방법론의 유용성을 검증하기 위한 실험 데이터 및 설계 내용을 설명하고, 최종 산출된 실험결과를 종합적으로 정리해 제시한다. 끝으로 마지막 절에서는 결언으로 전반적인 연구 성과 및 의의를 설명하고, 본 연구의 시사점 및 한계점과 함께 향후 연구 방향에 대하여 논의한다.

2. 연구 배경

전술했듯이 본 연구에서는 텍스트 마이닝의 종류 중 하나인 토픽모델링 기반의 가짜뉴스 다분류 예측에 기계 학습 기법을 적용하는 새로운 형태의 방법론을 제안하는 내용을 담고 있다. 이에 본 장에서는 본 연구의 제안 방법론을 이해하는데 필요한 다양한 개념들의 이론적 배경을 살펴보고, 이어 본 연구와 마찬가지로 자동화된 가짜뉴스 탐지 방법론을 새롭게 개발하거나 성능을 개선하고자 시도한 기존 연구들에 대해 살펴본다.

2.1 가짜뉴스

가짜뉴스는 2010년대 이후로 인터넷이 발달하고 사회관계망 서비스가 급속도로 발달함에 따라 언론사가 아닌 개인들이 사실이 아닌 내용을 진짜 뉴스처럼 퍼뜨리는 사태가 많이 일어나면서 사회 문제로 대두되었는데, 2016년 미국의 대통령 선거를 기점으로 크게 확산되었다[Hong and Jung, 2017]. 가짜뉴스의 확산은 현실 세계에 심각한 피해를 가져 올 수 있다. 이에 따라, 최근들어 가짜뉴스로 인한 사회 정보 질서의 왜곡과 갈등을 막기 위해 다양한 대응방안에 대한 논의가 이루어지고 있다. 이러한 상황에서, 가짜뉴스는 허위사실 적시에서부터 패러디사이트까지 넓은 스펙트럼 안에서 정확한 구분 없이 혼랍스럽게 사용되고 있어 정의와 범위에 대한 의견이 다양하게 나뉘고 있다[Hwang and Kwon, 2017; Hong and Jung, 2017]. 따라서, 먼저 가짜뉴스의 개념을 명확히 한 후에 그에 알맞은 대응방안을 마련할 필요가 있다.

2017년 2월 14일 한국언론학회와 한국언론진흥재단 주최로 열린 '가짜뉴스 개념과 대응방안' 세미나에서는 가짜뉴스는 정치·경제적 이익을 위해 의도적으로 언론 보도의 형식을 하고 유포된 거짓 정보라고 개념을 정의했다. 이에 따라 가짜뉴스는 3가지 주요 특징을 가진다. 첫 번째, 가짜뉴스는 정치·경제적 이익을 주된 목표로 한다. 정치적으로 허위 사실을 제작·유포하여 상대 후보자나 정당을 폄하하고 상대적 이득을 얻으려 하거나 경제적으로 자신을 포함한 이해관계자들이 이득을 얻고 타인에게는 손실을 줄 것을 목적으로 한다[Han and Yoon, 2017]. 두 번째, 가짜뉴스는 거짓 정보를 담고 있다. 가짜뉴스의 거짓 정보는 정보 전달의 의도와 상관없이 부정확하거나 사실과 달라 옳지 않은 정보들을 지칭하는 잘못된 정보(misinformation)와 특정한 의도를 가지고 정밀하게 만들어진 정보들을 지칭하는 기만적 정보(disinformation)를 모두 포함한다[Hwang and Kwon, 2017]. 앞선 두 가지 특징은 가짜뉴스와 유사한 유형인 루머, 스팸, 찌라시 등과 같은 거짓 정보에도 해당되는 특징이다. 반면에 다음에 제시되는 특징은 가짜뉴스에만 해당하는 가장 차별화되는 특징이다. 세 번째, 가짜뉴스는 언론 보도의 형식을 취한다. 형식적인 면에서 일반적인 온라인 커뮤니티의 글과 같은 형식과 달리 완벽하지는 않더라도 기사로 오인될 수 있을 정도의 전문적인 기사 형식을 갖추어야 한다는 것이다[Han and Yoon, 2017].

가짜뉴스에 대한 대응방안 중 하나로써 가짜뉴스 탐지 방법론의 중요성은 커지고 있다. 가짜뉴스를 탐지하는 기법은 비기술적 접근 기법, 기술적 접근 기법, 하이브리드 분석 기법으로 크게 3가지로 나뉜다[Institute for Information and communications Technology Promotion, 2017]. 비기술적 접근 기법으로는 전문가 기반 기법, 집단지성 기반 기법 등이 있으며 기술적 접근 기법으로는 인공지능 기반 기법, 시맨틱 기반 기법, 이상 확산 패턴 탐지 기법 등이 있다. 비기술적 접근 기법은 사람의 물리적 개입이 많이 필요해 가짜 뉴스를 전수 탐지하기에는 어렵다는 한계점이 있다. 때문에, 오늘날에는 기술적 접근 기

법을 이용한 자동화된 가짜 뉴스 탐지 방법론에 대한 많은 연구들이 이루어지고 있다.

2.2 자동화된 가짜뉴스 탐지 방법론

전술했듯이, 자동화된 가짜뉴스 탐지 방법론에는 기술적 접근 기법을 이용하고 있으며 인공지능 기반 기법, 시맨틱 기반 기법, 이상 확산 패턴 탐지 기법이 이에 해당된다. 인공지능 기반 기법은 언어와 구문을 분석하기 위해 과거 문제가 된 가짜 뉴스에 자주 등장한 단어와 표현을 활용하여 기계 학습을 통해 가짜뉴스일 확률을 추정하는 기법이다[Institute for Information and communications Technology Promotion, 2017]. 자동화된 가짜뉴스 탐지 방법론을 위한 대부분의 연구에서는 신속한 분석 결과를 도출해 낸다는 장점으로 인해 인공지능 기반 기법을 활용하고 있다.

대표적으로 2016년 12월1일부터 2017년 6월 15일까지 진행된 Fake News Challenge의 사례를 볼 수 있다. Fake News Challenge는 가짜뉴스의 근절을 위해 Dean Pomerleau와 Delip Rao가 주도하여 전 세계 학회와 업계에서 100명이 넘는 자원봉사자와 71개 팀이 협력하여 진행된 프로젝트로써, 인공지능 기법을 이용하여 뉴스기사에 숨어 있는 조작이나 잘못된 보도를 구별해 내는 가능성을 찾아보는 것이 목적이다. 프로젝트 결과, 상위 3개 팀에서는 딥러닝(Deep Learning), 의사결정나무(Decision Tree), 멀티스레딩(Multi-threading), TF-IDF(Term Frequency-Inverse Document Frequency)와 같은 기법을 적용하였다.

한편, 보다 높은 정확도와 공신력을 위해 인공지능 기반 기법에 다양한 기술적 접근 기법을 결합하는 하이브리드 가짜뉴스 탐지 방법론에 대한 연구도 활발히 진행되었다.

Bajaj[2017]는 자연어처리(Natural Language Processing, NLP)의 관점에서 뉴스의 내용이 가짜인지 진짜인지 여부를 예측하는 방법론을 제안하였다. 제안 방법론의 검증을 위해 Kaggle 데이터세트, Signal Media News 데이터세트를 활용해 63,000개의 기사를 Tensorflow와 Python 등의 기계 학습 기법들을 사용하여 성능을 평가하였다. 이 연구는 방대한 데이터 하에서 실질적인 실험이 수행되었다는 점에서 의의가 있다. 하지만, 실험 안의 다른 비교 기법들에 비해 성과가 개선됨을 확인하였으나 그 차이가 과연 통계적으로 유의한 수준인지 확인하지는 못하다는 한계를 갖는다.

Conroy et al.[2015]은 언어적 접근법과 네트워크 분석을 결합한 가짜뉴스 탐지 방법론을 제안하였다. 언어적 접근법으로는 사기성 메시지 내용의 패턴을 가지고 SVM을 활용하여 분류하는 방법을 사용하였으며, 네트워크 분석으로는 데이터 네트워크에서 술어관계를 통해 연결하여 사실을 확인하는 방법을 사용하였다. 하지만 실험에 대한 결과는 없이 방법론에 대한 제안만 하였다는 한계점이 있다.

Salas-Zarate et al.[2017]은 감성분석에 의한 풍자인식과 자연어처리를 통해 텍스트를 전처리하여 텍스트의 특징을 추출한 뒤에 기계 학습 기법에 적용하는 방법론을 제안하였다. 이 연구는 인공지능 기반 기법에 시맨틱 기반 기법을 효과적으로 적용하였다는 점에서 의의가 있다. 반면, Jin et al.[2017]은 트위터에서 상충되는 관점을 발견하여 신뢰성 전파 네트워크를 구축하여 반복적 추론을 통해 가짜뉴스에 대한 최종 평가 결과를 생성하는 방법론을 제안하였다. 이를 위해 주제와 주제에 대한 관점들을 한 쌍으로 모델링하고 K-means 군집분석을 통해 두 개의 상충되는 관점으로 분리하였다. 또한, 트윗 간의 신뢰성 네트워크를 구축하여 지지와 반대에 따라 링크의 연결정도를 정의하였다. 이 연구는 인공지능 기반 기법에 이상 확산 패턴 탐지 기법을 효과적으로 적용하였다는 점에서 의의가 있다. 하

지만 이들의 접근법은 트위터 데이터만을 가짜뉴스의 데이터로 활용한다는 한계를 갖는다.

이처럼 자동화된 가짜뉴스 탐지 방법론에 대하여 주로 텍스트 마이닝과 기계 학습을 이용한 연구들이 해외에서 최근 활성화되고 있는 추세지만, 안타깝게도 국내에서는 거의 찾아보기 힘들다. 가짜뉴스에 대한 규율 및 규제 방안에 대한 Hwnag and Kwon[2017]과 Han and Yoon[2017]의 연구들이나 시맨틱 기반 기법과 이상 확산 패턴 탐지 기반 기법을 동시에 적용하는 Kwon et al.[2017]의 연구들이 있으나, 인공지능 기반 기법을 자동화된 가짜뉴스 탐지 방법론에 활용한 사례는 아직 발표되지 않은 상태이다.

3. 제안 방법론

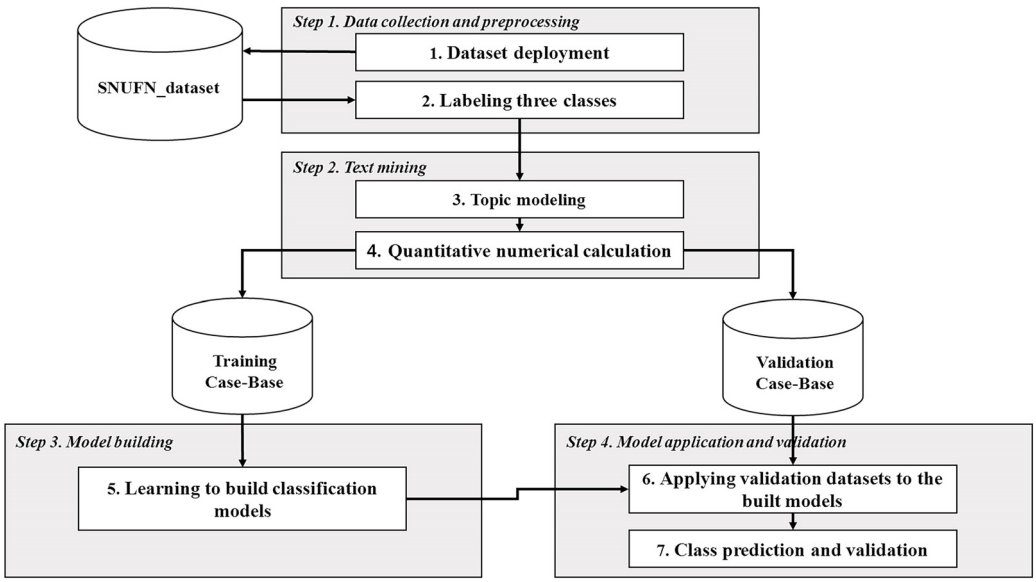
본 연구에서는 토픽모델링 기반의 가짜뉴스 다분류 예측에 기계 학습 기법을 적용하는 방법론을 제안한다. 다음의 <Figure 1>은 본 연구의 제안 벙법론이 작동하는 프로세스를 시각화하여 제시하고 있다. <Figure 1>에 제시된 바와 같이, 제안 방법론의 작동 과정은 다음과 같이 크게 4단계 절차에 의해 진행된다.

단계 1. 데이터 수집 및 전처리

제안 방법론의 첫 번째 단계는 데이터 수집 및 전처리 단계이다. 이 단계에서는 서울대학교 팩트체크(http://factcheck.snu.ac.kr/)로부터 데이터를 수집하여 SNUFN_dataset을 구축하고 분석의 용이성을 위해 전처리를 통하여 판정(Label)을 진실, 중립, 거짓의 3가지 클래스로 변환하며 메타 데이터들을 각각 이산형 변수 또는 연속형 변수로 변환한다.

단계 2. 텍스트 마이닝

이렇게 하여 데이터 수집 및 전처리 작업이 끝나고 나면, 2단계는 텍스트 마이닝 단계이다. 이 단계에서는 전처리된 SNUFN_dataset 중에 짧은 문장에 대하여 토픽모델링을 수행하여 다음의 <Figure 2>에 예시된 것과 같은 문서-토픽 가중치 행렬(Document-Topic Weight Matrix)을 도출하여 짧은 문장을 정량화된 특징값들로 변환한다. 문서-토픽 가중치는 일반적으로 TF-IDF를



〈Figure 1〉 Architecture of Proposed Methodology

이용하여 산출되며, 본 연구의 제안 방법론 또한 TF-IDF에 기반한 문서-토픽 가중치들을 사용한다[Jeon and Ahn, 2015].

단계 3. 모형 구축

이렇게 하여 짧은 문장이 정량화된 특징값들로 변환이 되면, 우선 SNUFN_dataset을 학습용과 검증용으로 나눈다. 그 후, 학습용 데이터셋에 기계학습을 적용하여 학습된 모델을 구축하는 3단계를 수행한다. 이러한 3번째 단계는 다양한 기계학습 기법들을 활용함으로써, 여러 개의 모델을 구축하게 된다.

단계 4. 모형 적용 및 검증

기계 학습을 통해 여러 개의 모델이 구축되면, 마지막 4단계로 검증용 데이터셋을 각각의 모델에 적용한다. 이 단계에서는 검증용 데이터셋에 대한 클래스를 예측하게 되고, 예측정확도의 모델 간 비교를 통해 예측성과를 최종 점검하여, 가장 우수한 모델을 선택하는 작업이 이루어진다. 이 과정 후에, 추가적으로 통계검정을 수행하여 제안 모형이 통계적으로 유의한 수준인지 확인한다.

4. 실증 분석

4.1 실험 데이터

제안 방법론의 검증을 위한 실험 데이터로는 SUNFN_dataset을 적용하였다. SUNFN_dataset은 본 연구를 위해 자체적으로 수집, 정리한 데이터셋으로서, LIAR_dataset을 벤치마킹하여 제작되었다. LIAR_dataset은 영어 가짜뉴스 데이터로 2007년부터 2016년까지 POLITIFACT.COM의 전문가들에 의해 직접 레이블 되어 있는 약 12,800개의 짧은 문장과 추가적으로 Speaker를 고려하여 해당인과 관련된 Party Affiliations, Current Job, Home State, TCHC(Total Credit History Count), Contexts/Venues, Subjects와 같은 메타 데이터를 수집하여 만들어졌다[Wang, 2017]. 국내에는 POLITIFACT.COM과 유사한 온라인 사이트로 서울대학교 FactCheck가 있는데, 이는 22개의 언론사와 대학이 협업하여 언론사들이 검증한 공적 관심사를 국민들에게 알리기 위해 서울대학교 언론정보연구소에서 비정치적, 비영리적으로 운영하는 정보서비스이다. SNUFN_dataset은 서울대학교 FactCheck에 게시되어 있는 2017년 3월 29일부터 9월 18일까지의 국내 가짜뉴스 데이터를 대상으로 하고 있다. 구체적으로 전문가들에 의해 직접 레이블 되어 있는 200개의 짧은 문장과 함께, 해당 뉴스의 주제, 관련인, 관련인 직업, 소속정당, TCHC, 언론사 등이 메타 데이터로 추가되었다. 다음의 <Figure 3>은 이러한 SNUFN_dataset의 구성 체계를 예시로 제시하고 있다.

	Topic 1	Topic 2	Topic 3	…	Topic m
Doc 1	0.389	0.230	0.000		-0.214
Doc 2	0.113	0.393	0.158		0.109
Doc 3	0.197	-0.091	0.154		0.095
…					
Doc n	0.319		0.187		0.531

〈Figure 2〉 Example of Document-Topic Weight Matrix

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1								Total credit history count									
2	ID	Label	Statement	주제	관련인	관련인 직업	소속정당	판단 유보	거짓	대체로 거짓	사실반 거짓	대체로 사실	사실	언론사			
3	1.snufc	대체로거짓	"77%로 미군 주둔비 부담 커...원본은 50%"	정치,19대	이재명	시장	더불어민주	0	0	1	0	0	0	중앙일보		Label	합계
4	2.snufc	사실반거짓	"더불어민주당도 국모닝 했다"	정치,19대	안철수	당대표	국민의당	2	9	6	5	1	3	중앙일보		판단 유보	14
5	3.snufc	사실반거짓	"공공부문 일자리 81만개 만들 수 있다"	정치,19대	문재인	대통령	더불어민주	5	7	13	11	10	1	중앙일보		거짓	50
6	4.snufc	거짓	세월호 희생자들이 천안함 희생자들보다 보상을 많	정치,19대	세월호	기타	소속정당없	0	1	0	0	0	0	SBS		대체로 거짓	45
7	5.snufc	사실	Q: 자유한국당 홍준표 경남지사가 위안부 합의에 (	정치,국제,	홍준표	당대표	자유한국당	1	15	8	12	5	4	매일경제		사실반 거짓반	41
8	6.snufc	대체로사실	"홍준표 지사가 국회 운영위원장 시절, 판공비 일부	정치,19대	김진태	국회의원	자유한국당	0	0	0	0	1	0	조선일보		대체로 사실	31
9	7.snufc	대체로사실	"대통령선거 본선에 나가기 직전에 사표를 제출하	정치,19대	홍준표	당대표	자유한국당	1	15	8	12	5	4	매일경제		사실	19
10	8.snufc	사실반거짓	추미애 더불어민주당 대표가 3월 8일 "사드를 배치	정치,국제,	추미애	국회의원,?	더불어민주	0	0	0	1	0	0	매일경제		총계	200
11	9.snufc	거짓	자유한국당 홍준표 대선후보가 경선 토론회에서 "'	정치,19대	홍준표	당대표	자유한국당	1	15	8	12	5	4	SBS			
12	10.snufc	대체로거짓	신연희 강남구청장이 대선 관련 가짜 뉴스가 담긴	정치,19대	신연희	구청장	자유한국당	0	0	1	0	1	0	SBS			
13	11.snufc	사실반거짓	심상정 정의당 대선 후보 "병사 한 명이 21개월 군	정치,사회,	심상정	국회의원	정의당	0	0	1	2	2	0	조선일보			
14	12.snufc	대체로사실	자유한국당 대선 후보로 선출된 홍준표 후보는 선	정치,19대	홍준표	당대표	자유한국당	1	15	8	12	5	4	MBN			
15	13.snufc	사실	홍준표 자유한국당 후보가 15대 중선 때 선거법 위	정치,사회,	홍준표	당대표	자유한국당	1	15	8	12	5	4	조선일보			
16	14.snufc	대체로거짓	2011년 서울대 정교수 자리로 안철수 부부가 모두	정치,19대	안철수	당대표	국민의당	2	9	6	5	1	3	SBS			
17	15.snufc	거짓	2006년 문재인의 민정수석실 "노무현 대통령의 사	정치,사회,	문재인	대통령	더불어민주	5	7	13	11	10	1	조선일보			
18	16.snufc	사실반거짓	1. 예비후보 등록 안해도 선거운동 할 수 있다. 기	정치,19대	홍준표	당대표	자유한국당	1	15	8	12	5	4	KBS			
19	17.snufc	사실반거짓	현재 지지율 1위 문재인 민주당 경선 후보 아들과	정치,19대	문재인	대통령	더불어민주	5	7	13	11	10	1	SBS,서울신문,한국일보			
20	18.snufc	대체로사실	경남도지사직 사퇴 시기를 늦춰서 보궐선거가 없도	정치,19대	홍준표	당대표	자유한국당	1	15	8	12	5	4	SBS			
21	19.snufc	대체로사실	문재인 더불어민주당 대선후보 선출 당일, 안희정	정치,19대	문재인	대통령	더불어민주	5	7	13	11	10	1	한국일보			
22	20.snufc	사실반거짓	유승민 바른정당 대선후보의 말처럼 "홍준표 지사	정치,19대	유승민	국회의원	바른정당	0	1	1	3	1	3	조선일보,한국일보			
23	21.snufc	사실반거짓	"안철수 국민의당 후보 부인 김미경 교수가 카이스	정치,19대	안철수	당대표	국민의당	2	9	6	5	1	3	SBS			
24	22.snufc	대체로사실	'3D 프린터'를 '스리디 프린터' 대신 '상디 프린터'로	정치,19대	문재인	대통령	더불어민주	5	7	13	11	10	1	서울신문			
25	23.snufc	대체로사실	문재인 전 더불어민주당 대표는 공무원 일자리 공	정치,사회,	문재인	대통령	더불어민주	5	7	13	11	10	1	매일경제			
26	24.snufc	대체로거짓	文 측 "안철수 장녀 재산 공개해야"는 사실?	정치,19대	문재인	대통령	더불어민주	5	7	13	11	10	1	KBS			
27	25.snufc	판단유보	문재인 민주당 후보 아들이 고용정보원에 낸 이력	정치,19대	문재인	대통령	더불어민주	5	7	13	11	10	1	SBS			
28	26.snufc	대체로거짓	문재인캠프 영입인사의 과거 이력에 관한 주장은 /	정치,19대	문재인	대통령	더불어민주	5	7	13	11	10	1	JTBC			

〈Figure 3〉 Example of SNUFN_Dataset

4.2 실험 설계

SNUFN_dataset의 판정은 크게 6가지의 클래스로 나뉘는데, 본 연구에서는 편의상 판단유보와 사실 반 거짓 반은 중립, 대체로 거짓과 거짓은 거짓, 대체로 사실과 사실은 사실로 표기하여 3가지의 클래스로 묶고, 랜덤샘플링(random sampling)을 통해 각 클래스에 대해 50개의 데이터를 추출한다. 이러한 작업을 하는 이유는 6가지 판정의 빈도가 불균형하고, 판단유보와 사실이 거짓에 비해 상대적으로 너무 부족하기 때문이다. 일반적으로 기계학습을 적용하여 학습된 모델을 구축할 때 불균형한 분포의 데이터를 활용하게 되면, 상대적으로 많은 빈도를 보이는 데이터에 과적합(overfitting)되어 좋은 모델을 구축할 수 없다[Ahn, 2014].

또한, SNUFN_dataset의 메타 데이터 중 주제, 관련인, 관련인 직업, 소속정당, 언론사는 전처리를 통하여 이산형 변수로 변환하였으며, TCHC는 전처리를 통하여 정규화된 연속형 변수로 변환하였다. 메타 데이터를 실험에 적용 할 때는 두 가지 방법으로 나누어 적용하였다. 첫 번째 방법은 모든 데이터를 일일이 변환하여 전체를 활용하였다. 두 번째 방법으로는 각각 임의로 설정한 기준에 따라 통합하여 적용하였는데, 관련인은 제 19대 대선주자, 기타, 관련인 없음으로 나누었으며, 관련인 직업은 정치인, 기타로 나누었으며, 소속정당은 진보, 보수, 중립으로 나누었으며, 언론사는 TV방송사, 신문사, 언론사 없음으로 나누었으며, TCHC는 판단유보, 거짓, 사실로 나누어 활용하였다. 특히, 주제는 이미 서울대학교 FactCheck에서 기준을 정하여 나누어둔 데이터이기 때문에 오로지 첫 번째 방법으로만 활용하였다.

토픽 모델링을 수행 할 때 보다 정확하고 유의미한 명사들을 구분해 내어 성능을 높이기 위해 서울대학교 꼬꼬마 형태소 분석기를 활용하여 Stoplist를 만들어 토픽 필터에 활용하였다.

모형의 구축을 위한 자료의 구분과 관련해서는 학습용 데이터로 각 클래스의 70%의 해당하는 데이터(총 105건)를 그리고 나머지 30%(총 45건)를 검증용 데이터로 사용한다. 입력변수의 후보군으로는 짧은 문장이 토픽모델링을 통해 변환되어 정량화된 특징값들을 변수들과 메타 데이터들을 사용한다.

본 연구에서 제안하는 방법론의 성능을 최적화하기 위해 다중판별분석(Multiple Discriminant Analysis, MDA), 사례기반추론(Case Based Reasoning, CBR), 인공신경망(Artificial Neural Networks, ANN), SVM(Support Vector Machines,

SVM)를 활용하여 분류 예측 모델을 구축하고 검증용 데이터셋에 대한 예측 정확도를 비교해 보고자 한다. 또한, 데이터 양이 충분하지 않아 방법론 성능측정의 통계적 신뢰도를 높이기 위해 적은 양의 데이터셋에서 보다 방대한 양의 데이터로 실험 해볼 수 있어 매우 유용하다고 알려진 k-fold 교차검증(cross validation)도 수행한다. 본 연구에서는 다음의 <Figure 4>에 예시된 것과 같은 방법으로 편의상 5-fold로 나누어 교차검증을 진행하였으며, 5-fold에서는 학습용 데이터로 각 클래스의 80%의 해당하는 데이터(총 120건)를 그리고 나머지 20%(총 30건)를 검증용 데이터로 사용한다.

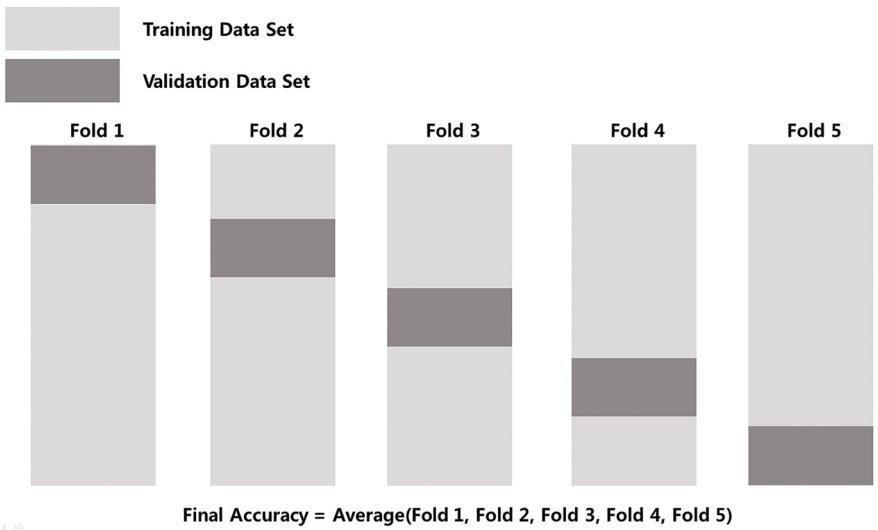
최종적으로 선택된 분류 모델이 가짜뉴스 예측에 가장 적합한 대안인지를 정밀하게 검증하기 위해, Two-sample test for proportions를 통해 모델 간 성능의 차이가 통계적으로 유의한 수준인지 확인한다[Noh and Ahn, 2017].

끝으로 본 연구의 실험을 위한 방법론은 토픽 모델링은 SAS를 활용하였으며 MDA, CBR은 SPSS를 활용하였으며 ANN은 NeuroShell을 활용하였으며, SVM은 LIBSVM을 활용하였다.

4.3 실험 결과

본 연구에서는 전문가에 의해서 판정이 되어진 기사에 대해서 제안한 모형으로 예상 판정을 도출한 다음, 실제 판정과 비교하여 예측 정확도를 확인해보는 방식으로 검증을 진행하였다. 이때, 예측 정확도는 검증용 데이터 셋에 대하여 얼마나 잘 예측하였는지를 기준으로 하였다.

우선 토픽 모델링에서 몇 개의 토픽을 생성하여 반영하는 것이 가장 우수한 예측 정확도를 보이는지 알아보기 위한 실험을 진행하였다. 데이터의 수가 200개로 매우 적고 토픽 분석에 활용되는 텍스트 또한 짧은 문장이기 때문에 일반적인 실험에 비해 상대적으로 적은 최소 7개, 최대 10개의 토픽을 추출 할 수 있었다. 따라서, 최소와 최대 사이의 범위가 매우 좁아 최소, 최대 두 가지의 경우에 대해서만 실험을 진행하였다. 기계 학습 기법 중에는 적은 양의 데이터에서도 높은 예측 정확도를 보인다고 알려진 SVM을 활용하였다. 아래에는 토픽 개수별 실험 결과가 <Table 1>에 제시되어 잇다. <Table 1>의 결과를 살펴보게 되면, 검증용 데이터셋에 대하여



〈Figure 4〉 Example of 5-Fold Cross Validation

토픽의 개수를 최대로 추출한 10개의 경우의 예측 정확도가 51.11%로 최소로 추출한 7개의 경우의 예측 정확도인 37.78% 보다 약 13% 더 우수한 예측 정확도를 산출하게 됨을 알 수 있다.

〈Table 1〉 Prediction Accuracy According to the Number of topics

# of topics	Training	Validation
Topic = 7	61.90%	37.78%
Topic = 10	53.33%	**51.11%**

그렇다면, 이번에는 기계 학습 기법 중에 SVM이 가장 우수한 정확도를 보이는지 알아보기 위해 <Table 1>의 결과에 따라 토픽의 개수를 10개로 정해둔 뒤 MDA, CBR, ANN과 비교해보는 실험을 진행하였다. 아래에는 기계 학습별 실험 결과가 <Table 2>에 제시되어 있다. <Table 2>의 결과를 살펴보게 되면, 검증용 데이터셋에 대하여 SVM을 적용한 경우의 예측 정확도가 51.11%로 가장 낮은 예측 정확도를 보이는 CBR을 적용한 경우의 예측 정확도인 44.40% 보다 약 7% 더 우수한 예측 정확도를 산출하며 다음으로 우수한 예측 정확도를 보이는 MDA를 적용한 경우의 예측 정확도인 48.90% 보다 약 2% 더 우수한 예측 정확도를 산출하게 됨을 알 수 있다.

한편 검증용 데이터셋에 대하여 각 기법 별로 거짓(총 15개)을 거짓, 중립, 사실로 예측한 빈도가 <Table 3>에, 검증용 데이터셋에 대하여 각 기법 별로 사실(총 15개)을 거짓, 중립, 사실로 예측한 빈도가 <Table 4>에 제시되어 있다. 이 표에서 볼 수 있듯이, 제안기법인 SVM은 거짓 음성(false negative) 발생빈도가 ANN과 함께 1건으로 나타나, 다른 기법 대비 가장 작은 수치를 나타냈다. 거짓 양성(false positive) 발생빈도 역시 SVM은 MDA와 함께 6건으로 나타나, 다른 기법 대비 가장 낮은 것으로 나타났다. 이상의 결과를 종합해 보면, SVM은 예측 정확도 뿐 아니라 오탐지에 따른 비용 측면에서도 타 기법을 압도하는 우수한 성능을 나타냈다고 할 수 있다.

〈Table 2〉 Prediction Accuracy of the Machine Learning Techniques

Model	Setting value	Training	Vaildation
MDA	Full entry method	51.40%	48.90%
CBR	Manhattan distance	30.50%	44.40%
ANN	Hidden layer = 13	79.04%	46.66%
SVM (Proposed method)	RBF kernel, C = 100, δ^2 = 100	53.33%	**51.11%**

〈Table 3〉 Frequencies of the False Negative Cases

Actual Label	Model	Predicted Label		
		False	Neutrality	True
False	MDA	8	4	3
	CBR	8	3	4
	ANN	7	7	1
	SVM	9	5	1

〈Table 4〉 Frequencies of False Positive Cases

Actual Label	Model	Predicted Label		
		False	Neutrality	True
True	MDA	6	2	7
	CBR	7	2	6
	ANN	9	2	4
	SVM	6	4	5

다음으로 <Table 1>과 <Table 2>의 결과를 합하여 토픽 10개를 추출하여 SVM을 적용한 모델을 기본 모델로 정하고 추가로 각 메타 데이터를 적용하는 실험을 진행하였다. 전술했던 것처럼 메타 데이터를 적용하는 두 가지 방법은 첫 번째, 전체를 모두 이용하는 방법과 두 번째 임의의 기준에 따라 통합하여 이용하는 방법으로 활용하였다. 또한, 모든 메타 데이터를 적용해 보았는데,

각 메타 데이터 별로 우수한 결과를 보이는 방법들을 조합하여 최적 조합을 만들어 활용하였다. 아래에는 메타 데이터별 실험 결과가 <Table 5>에 제시되어 있다. <Table 5>의 결과를 살펴보게 되면, 검증용 데이터셋에 대하여 기본 모델에 언론사 메타 데이터의 전체를 모두 이용하여 추가하는 경우의 예측 정확도가 60.00%로 다음으로 우수한 예측 정확도를 보이는 기본 모델에 각 메타 데이터 별 최적 조합으로 전체 메타 데이터를 추가하는 경우의 예측 정확도인 55.56%다 약 4% 더 우수한 예측 정확도를 산출하게 됨을 알 수 있다.

다음으로 <Table 5>의 결과에 따라 기본 모델에 언론사 메타 데이터의 전체를 모두 이용하는 모델에 MDA, CBR, ANN, SVM을 적용하여 다른 기법과 예측 정확도를 비교 분석해 보는 실험을 진행하였다. 전술했던 것처럼 데이터의 양이 150개로 충분하지 않아 비교 방법에는 5-fold 교차검증을 적용하였다. 아래에는 5-fold 교차검증 실험 결과가 <Table 6>에 제시되어 있으며 <Table 7>에는 기계 학습 기법에 따라 각 Fold 별 가장 높은 결과를 보이는 모형의 설정값이 제시되어 있다. <Table 6>의 결과를 살펴보게 되면, 검증용 데이터셋에 대하여 SVM을 적용한 경우의 평균 예측 정확도가 55.33%로 가장 우수한 예측 정확도를 산출하게 됨을 알 수 있다. 따라서, 본 연구의 실증분석에서는 토픽을 10개 추출하여 언론사 메타데이터를 추가한 모델에서 SVM > CBR > MDA > ANN의 순으로 예측 정확도가 산출됨을 확인할 수 있었다. 특히, ANN

〈Table 5〉 Prediction Accuracy by Meta Data

Model	How to apply	Setting value	Training	Validation
Base + Category	2	Linear kernel, C = 10	63.81%	53.33%
Base + Related person	1	Polynomial kernel, C = 1, d = 3	60.95%	44.44%
	2	Linear kernel, C = 100	61.90%	40.00%
Base + Occupation	1	RBF kernel, C = 10, δ^2 = 1	76.19%	40.00%
	2	RBF kernel, C = 10, δ^2 = 1	66.67%	48.89%
Base + Party	1	Polynomial kernel, C = 1, d = 3	49.52%	48.89%
	2	Polynomial kernel, C = 55, d = 1	56.19%	48.89%
Base + Media	1	RBF kernel, C = 10, δ^2 = 50	61.90%	**60.00%**
	2	Linear kernel, C = 10	53.33%	48.89%
Base + TCHC	1	Polynomial kernel, C = 33, d = 1	55.24%	46.67%
	2	Polynomial kernel, C = 10, d = 1	57.14%	48.89%
Base + ALL	Best combination	Linear kernel, C = 1	77.14%	55.56%

의 경우 학습용 데이터셋에 대하여 95.84%의 평균 예측 정확도를 보이는데 반해 검증용 데이터셋에 대하여는 44.67%의 평균 예측 정확도만 보이는 것을 보아 학습시 데이터가 과적화 되었기 때문에 성능이 가장 낮게 나온 것을 알 수 있다. 이는 일반적으로 ANN은 충분한 데이터가 필요함에도 불구하고 너무 적은 데이터가 주어졌기 때문으로 보여진다.

〈Table 6〉 Prediction Accuracy of 5-Fold Cross Validation

Dataset		MDA	CBR	ANN	SVM
Fold 1	Train	70.00%	46.70%	95.00%	74.17%
	Valid	43.30%	43.30%	50.00%	53.33%
Fold 2	Train	65.00%	35.00%	96.67%	60.83%
	Valid	60.00%	60.70%	36.67%	63.33%
Fold 3	Train	50.80%	53.30%	96.67%	69.17%
	Valid	40.00%	40.00%	36.67%	43.33%
Fold 4	Train	49.20%	45.80%	96.67%	63.33%
	Valid	46.70%	52.00%	46.67%	60.00%
Fold 5	Train	68.30%	45.00%	94.17%	63.33%
	Valid	43.30%	56.70%	53.33%	56.67%
Average	Train	60.66%	45.16%	95.84%	66.17%
	Valid	46.66%	50.54%	44.67%	**55.33%**

〈Table 7〉 Optimum Setting Value for Each Fold by Machine Learning Technique

Data-set	MDA	CBR	ANN	SVM
Fold 1	Full entry method	Manhattan distance	Hidden layer = 35	RBF kernel, C = 1, δ^2 = 1
Fold 2	Full entry method	Euclidean distances	Hidden layer = 18	RBF kernel, C = 10, δ^2 = 25
Fold 3	Step selection method	Euclidean distances	Hidden layer = 18	RBF kernel, C = 55, δ^2 = 50
Fold 4	Step selection method	Euclidean distances	Hidden layer = 52	RBF kernel, C = 33, δ^2 = 50
Fold 5	Full entry method	Manhattan distance	Hidden layer = 70	Polynomial kernel, C = 100, d = 2

마지막으로 모형 간 성과의 차이가 통계적으로 유의한 지를 검증하기 위해, 이표본 비율검정(two sample test for proportions)을 수행하였다. 본 연구에서 적용된 이표본 비율검정의 귀무가설 H0는 $p_A = p_B$, 대립가설 Ha는 $p_A > p_B$(p_A : 모형 A의 검증용 데이터셋에 대한 평균 예측정확도 비율)이다. 아래 <Table 8>는 이러한 이표본 비율검정의 결과를 나타내고 있다. 이 표에서 볼 수 있듯이, SVM 모형은 MDA는 물론 CBR과도 95% 신뢰수준 하에서 통계적으로 유의한 성과 차이를 보이는 것으로 나타났다. 또한, ANN에 대해서는 99% 신뢰수준 하에서 통계적으로 유의한 성과 차이를 보이는 것을 확인할 수 있었다.

〈Table 8〉 Z-Values of Two Sample Test for Proportions

	CBR	ANN	SVM
MDA	-0.116	0.348	**-1.501***
CBR		0.463	**-1.386***
ANN			**-1.848****

*statistical significant at 5%, **statistical significant at 1%.

5. 결 론

본 연구에서는 텍스트 마이닝을 활용해 짧은 문장을 정량화된 텍스트의 특징값들로 변환하여 기계학습에 적용하는 새로운 방법론을 제안하였다. 수집된 데이터를 활용하여 제안 모형의 성능을 검증한 결과, 토픽을 10개 추출하고 언론사 메타데이터를 활용하여 SVM 기법에 적용하는 모형이 다른 모형들과 비교해, 가장 우수한 결과를 보이며 통계적으로도 유의한 성과 차이를 확인할 수 있었다. 구체적으로 본 연구가 갖는 시사점을 고찰해 보면 다음과 같다.

첫째, 본 연구는 국내에서 그간 충분히 다루어지지 않았던 자동화된 가짜뉴스 탐지 방법론 실험을 인공 지능 기법에 접목할 수 있는 방안을 연구 했다는 점에서 의미가 있다. 특히 본 연구

의 제안 모형은 복잡하고 사람의 인위적 개입이 상대적으로 많이 요구되는 것이 아니라 가장 기초적인 수준의 텍스트 마이닝과 기계 학습 기법만으로 구동이 가능한 모형을 제안했다는 점에서 실무적으로 의미가 있는 연구라 사료된다.

둘째, 국내 가짜뉴스 데이터를 확보하는 방법을 제안하고 가능성을 확인했다는 점에서 학술적으로 의미가 있는 연구라 사료된다. 또한, 데이터 수집이 상대적으로 용이하다는 이유로 대다수의 가짜뉴스 탐지 방법론 연구에서 사용되어 온 영어 텍스트 데이터를 사용하지 않고, 오늘날 그 중요성이 더 크게 대두되고 있는 팩트체크에서 데이터를 따로 수집해, 검증에 사용하였다는 점도 기존 연구와 본 연구가 크게 차별화되는 부분이라고 할 수 있다.

반면 본 연구의 한계점들은 다음과 같다. 첫째, 국내 가짜뉴스에 대한 자동화된 예측 모형을 구축하고 분석하기 위해, 데이터 확보의 어려움에 따라 직접 데이터를 수집하였음에도 불구하고 데이터의 양이 충분하지 않으며 불균형한 분포를 보이고 있다. 따라서, 본 연구의 제안 모형이 보다 일반화된 성능 향상을 가져오는지 확인하기 위해서는 보다 많은 데이터를 확보하여 방대한 양의 데이터를 통한 실험이 이루어질 필요가 있다.

둘째, 현재 제안 모형의 경우, 뉴스 기사에 대하여 전문가에 의해 판정이 되어 있는 보고서의 짧은 문장을 텍스트 데이터로 활용하고 있는데, 아직 판정이 내려지지 않은 기사를 예측하기 어려울 가능성이 있다. 이에 따라, 판정이 내려지지 않은 기사를 활용하거나 새로운 데이터셋을 구축하는 방법론을 추가한 실험이 이루어질 필요가 있다.

끝으로 현재 제안된 연구 모형에서 명확하게 드러난 본 연구의 한계점을 개선하기 위해 다소 연구자 임의의 설정이 개입될 지더라도, 판단유보와 같은 중립적 데이터를 거짓이나 진실로 변환하는 방식을 적용하거나 혹은 서울대학교 Fact-Check 이외에서도 추가 데이터를 수집해 볼 수 있다. 또한, 데이터 부족이나 불균형을 효과적으로 개선 할 수 있다고 알려진 데이터 샘플링(Data sampling) 기법이나 판정이 내려지지 않은 기사를 추가하여 활용하는 비지도학습(semi-supervised learning)과 같은 다양한 기법을 적용해 보고, 그 성능을 확인해 보는 연구가 의미 있는 후속 연구의 주제가 될 것으로 예상된다.

References

[1] Ahn, H., “Optimization of Multiclass Support Vector Machine using Genetic Algorithm : Application to the Prediction of Corporate Credit Rating”, *Information Systems Review*, Vol. 16, No. 3, 2014, pp. 161-177.

[2] Bajaj, S., “The Pope Has a New Baby! : Fake News Detection Using Deep Learning”, *Technical Report*, Stanford Univ, 2017.

[3] Conroy, N. J., Rubin, V. L., and Chen, Y., “Automatic Deception Detection : Method for Finding Fake News”, Proceedings of the Association for Information Science and Technology, 2015.

[4] Han, G. and Yoon, C., “A Study on the Regulation of The Fake News”, *Science, Technology and Law*, Vol. 8, No. 1, 2017, pp. 59-90.

[5] Hong, S. Y. and Jung, E. C., “Fake News and Journalism’s Credibility Crisis-Phenomena and Alternatives-”, *Crisisonomy*, Vol. 13, No. 8, 2017, pp. 43-60.

[6] Hwang, Y. and Kwon, O., “A Study on the Conceptualization and Regulation Measures

on Fake News : Focused on self-regulation of internet service providers", *Journal of Media Law, Ethics and Policy Research*, Vol. 16, No. 1, 2017, pp. 53-101.

[7] Hyundai Research Institute, "Economic Cost Estimation and Implications of Fake News", *Weekly Economic Review,* Vol. 736, Available at http://hri.co.kr/board/reportView.asp?numIdx=27886&firstDepth=1&secondDepth=1(Accessed on March 25, 2018).

[8] Institute for Information & communications Technology Promotion, "Fake News Detection Technique Trends and Implications", *Weekly ICT Trends*, No. 1816, 2017, pp. 12-23.

[9] Institute for Korean Democracy, "Fake News and Democracy", *Issue & Review on Democracy*, No. 14, 2017.

[10] Jeon, B. and Ahn, H., "A Collaborative Filtering System Combined with Users' Review Mining : Application to the Recommendation of Smartphone Apps", *Journal of Intelligence and Information Systems*, Vol. 21, No. 2, 2015, pp. 1-18.

[11] Jin, Z., Cao, J., Zhang, Y., and Luo, J., "News Verification by Exploiting Conflicting Social Viewpoints in Microblogs", *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[12] Kwon, S., Cha, M., and Jung, K., "Rumor detection over varying time windows", *PloS one*, Vol. 12, No. 1, 2017, e0168344.

[13] Noh, H. and Ahn, H., "A study on the recommendation algorithm based on trust/distrust relationship network analysis", *Journal of Information Technology Applications & Management*, Vol. 24, No. 1, 2017, pp. 1-17.

[14] Salas, Z. M. d. P., Paredes, V. M. A., Rodriguez, G. M. Á., Valencia, G. R., and Alor, H. G., "Automatic detection of satire in Twitter : A psycholinguistic-based approach", *Knowledge-Based System*, Vol. 128, 2017, pp. 20-33.

[15] Wang, W. Y., "Liar, Liar Pants on Fire : A New Benchmark Dataset for Fake News Detection", Technical Report, Dept. of Computer Science, Univ of California, 2017.

저자소개

Tae-Uk Yun

Tae-Uk Yun is currently a master's program at Graduate School of Business IT, Kookmin University, where he also earned his bachelor degree in Management Information Systems. His primary research interests include data mining for business, Information Technology adoption and use, and Human-Computer Interaction.

Hyunchul Ahn

Hyunchul Ahn is an associate professor in the Graduate School of Business IT, Kookmin University, Seoul, Korea. He has a master's degree and a Ph.D. in management engineering from the Korea Advanced Institute of Science and Technology (KAIST) Graduate School of Management. His research interests include technical issues on intelligent information systems for marketing and finance, and behavioral issues on the adoption of information systems. His works has been published in Annals of Operations Research, Applied Soft Computing, Computers & Operations Research, Computers in Human Behavior, Expert Systems with Applications, International Journal of Electronic Commerce, Information & Management, and Technological Forecasting and Social Change.