KIPS Trans. Softw. and Data Eng.
Vol.11, No.4 pp.149~156
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2022.11.4.149

CoAID+: COVID-19 News Cascade Dataset for Social Context Based Fake News Detection

Soeun Han† · Yoonsuk Kang†† · Yunyong Ko††† · Jeewon Ahn†††† · Yushim Kim††††† · Seongsoo Oh†††††† · Heejin Park††††††† · Sang-Wook Kim†††††††

ABSTRACT

In the current COVID-19 pandemic, fake news and misinformation related to COVID-19 have been causing serious confusion in our society. To accurately detect such fake news, social context-based methods have been widely studied in the literature. They detect fake news based on the social context that indicates how a news article is propagated over social media (e.g., Twitter). Most existing COVID-19 related datasets gathered for fake news detection, however, contain only the news content information, but not its social context information. In this case, the social context-based detection methods cannot be applied, which could be a big obstacle in the fake news detection research. To address this issue, in this work, we collect from Twitter the social context information based on CoAID, which is a COVID-19 news content dataset built for fake news detection, thereby building CoAID+ that includes both the news content information and its social context information. The CoAID+ dataset can be utilized in a variety of methods for social context-based fake news detection, thus would help revitalize the fake news detection research area. Finally, through a comprehensive analysis of the CoAID+ dataset in various perspectives, we present some interesting features capable of differentiating real and fake news.

Keywords : Fake News Detection, Propagation, Coronavirus, Social Context Based Detection

CoAID+: 소셜 컨텍스트 기반 가짜뉴스 탐지를 위한 COVID-19 뉴스 파급 데이터

한 소 은† · 강 윤 석†† · 고 윤 용††† · 안 지 원†††† · 김 유 심††††† ·
오 성 수†††††† · 박 희 진††††††† · 김 상 욱†††††††

요 약

최근 전 세계적으로 COVID-19이 유행하는 상황 속에서 이와 관련된 가짜뉴스가 심각한 사회적 혼란을 야기하고 있다. 이러한 배경에서 가짜뉴스를 정확하게 탐지하기 위해, 뉴스가 소셜 미디어를 통해 파급되는 과정과 같은 소셜 컨텍스트 정보를 활용하는 소셜 컨텍스트 기반 탐지 기법들이 널리 사용되고 있다. 그러나 대부분의 기 구축된 가짜뉴스 탐지를 위한 데이터들은 뉴스 자체의 내용 정보 위주로 구성되어, 소셜 컨텍스트 정보를 거의 포함하지 않는다. 즉, 이 데이터들에는 소셜 컨텍스트 기반 탐지 기법을 적용할 수 없으며, 이러한 데이터의 한계는 가짜뉴스 탐지 연구 분야의 발전을 저해하는 방해 요소이다. 본 논문은 이러한 한계를 극복하기 위해, 기존의 저명한 가짜뉴스 데이터인 CoAID 데이터를 기반으로, 소셜 컨텍스트 정보를 추가적으로 수집하여, CoAID 데이터의 뉴스 내용 정보와 해당 뉴스들의 소셜 컨텍스트 정보를 모두 포함하는 CoAID+ 데이터를 구축한다. 본 논문에서 구축한 CoAID+ 데이터는 기존의 대부분의 소셜 컨텍스트 기반 탐지 기법들에 적용될 수 있으며, 향후 새로운 소셜 컨텍스트 기반 탐지 기법들에 대한 연구도 더욱 활성화시킬 수 있을 것으로 기대된다. 마지막으로, 본 논문은 다양한 관점에서 CoAID+ 데이터를 분석하여 진짜뉴스와 가짜뉴스의 파급 패턴 및 키워드에 따른 파급 패턴도 파악하여 소개한다.

키워드 : 가짜뉴스 탐지, 파급, 코로나바이러스, 소셜 컨텍스트 기반 탐지

1. 서 론

가짜뉴스란 사실 검증이 가능함에도 불구하고, 특수한 목적을 위해 허위 정보를 포함하여 생산 및 유통되는 기사를 의미한다[1]. 최근 코로나바이러스에 의한 감염병(COVID-19)이 전 세계적으로 유행하는 가운데, 2021년 9월 기준 사망자 약 460만 명을 포함하여 수많은 피해가 발생하고 있다. 이와 같이 COVID-19이 전 세계적으로 대유행하는 상황 속에서 감염병에 대한 가짜뉴스는 심각한 사회적 혼란을 야기할 수 있다. 실제로 고농도 알코올을 섭취하면 COVID-19를 치료할 수 있다는 가짜뉴스로 인해 60여명의 사람들이 사망하는 사례가 있었다.[1)]

※ 이 논문은 정부의 재원으로 정보통신기획평가원(No.2020-0-01373, 인공지능대학원지원(한양대학교)) 및 한국연구재단(No.2018R1A5A7059549)의 지원을 받아 수행된 연구임.
※ 이 논문은 2021년 한국정보처리학회 춘계학술대회의 우수논문으로 "COVID-19 가짜뉴스 탐지를 위한 전파 데이터셋"의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 한양대학교 컴퓨터소프트웨어학과 석사과정
†† 비 회 원 : 한양대학교 컴퓨테이셔널 사회과학연구센터 박사후연구원
††† 정 회 원 : 한양대학교 인공지능 혁신인재교육 연구단 박사후연구원
†††† 비 회 원 : 한양대학교 컴퓨터소프트웨어학과 석·박사통합과정
††††† 비 회 원 : Arizona State University 행정학과 부교수
†††††† 비 회 원 : 한양대학교 행정학과 교수
††††††† 비 회 원 : 한양대학교 정보통신학부 교수
††††††† 종신회원 : 한양대학교 정보통신학부 교수
Manuscript Received : September 24, 2021
Accepted : November 17, 2021
* Corresponding Author : Sang-Wook Kim(wook@hanyang.ac.kr)

이러한 배경에서 가짜뉴스를 정확하고 빠르게 탐지하는 것은 더욱 중요한 과제로 대두되고 있으며, 이에 따라 가짜뉴스 탐지에 대한 연구들이 활발하게 진행되고 있다[2,3]. 일반적으로 가짜뉴스 탐지는 크게 두 분류로 나뉜다. 먼저, 뉴스가 포함하고 있는 내용 정보를 기반으로 가짜 뉴스 여부를 판단하는 내용 정보 기반 탐지(content based detection)와, 뉴스가 소셜미디어에서 사용자들의 행동, 사용자간 관계 등과 같은 정보를 기반으로 하는 소셜 컨텍스트 기반 탐지(social context based detection)로 구분된다. 내용 정보 기반 탐지의 경우, 최근 가짜뉴스들이 진짜뉴스를 모방하여 매우 유사한 형태로 작성됨에 따라, 내용 정보만으로는 가짜뉴스를 정확하게 탐지하는 것이 어렵다는 한계를 가진다[2,4].

이러한 한계를 극복하기 위해, 소셜 컨텍스트 기반 탐지는 소셜미디어 상의 사용자 정보나 사용자간 관계 정보, 뉴스가 사용자들에게 어떻게 퍼져나가는 지에 대한 파급 정보 등, 단순 모방이 어려운 다양한 정보를 함께 활용하여 가짜뉴스를 탐지한다[2,3,5]. 특히, 소셜미디어 정보들 중 뉴스가 사용자들에게 어떻게 퍼져나가는지에 대한 '파급 정보'는 가짜뉴스 탐지에 있어서 가장 효과적인 정보 중 하나로 많은 탐지 기법들에서 널리 사용된다[3,6]. 이는 가짜뉴스와 진짜뉴스가 소셜미디어에서 파급되는 형태가 본질적으로 서로 다르며 — 일반적으로 진짜뉴스와 비교하여 가짜뉴스의 파급이 더 널리, 빠르게 진행된다 — 이를 분석하여 활용함으로써 효과적으로 가짜뉴스를 탐지할 수 있기 때문이다.

그러나, 최근 구축되고 있는 대부분의 COVID-19 관련 가짜뉴스 탐지를 위한 데이터들[7-10]은 뉴스의 내용 정보를 위주로만 구성되어, 뉴스 파급 정보와 같은 소셜미디어 정보는 거의 포함하지 않는다는 한계를 가진다. 따라서 소셜 컨텍스트 기반 탐지 기법들을 적용하기 어렵다. 예를 들어 CoAID (**CO**vid-19 he**A**lthcare m**I**sinformation **D**ataset)[7]는 2020년 펜실베니아 주립 대학교에서 구축된 데이터로, 3,921건의 기사들에 대해 'PolitiFact.com', 'FactCheck.org'과 같은 신뢰할만한 팩트 체크 매체를 통해 진위여부를 판별하여, 3,055건의 진짜뉴스와 866건의 가짜뉴스로 구분하고, 각 뉴스들을 인용한 트윗(tweet)과 그 트윗에 대한 답글(reply)과 같은 일부 소셜 정보를 포함한다. 그러나 각 트윗 별 정보만 포함하여, 해당 트윗이 어떻게 다른 사용자들에 리트윗(retweet)되어 파급되는 지에 대한 정보는 없으므로 소셜 컨텍스트 기반 탐지 기법을 적용할 수 없다.

본 논문은 이러한 동기로부터, CoAID 데이터에 포함된 뉴스들이 소셜미디어(Twitter)에서 어떻게 파급되는지에 대한 정보를 추가로 수집함으로써, CoAID 데이터의 확장 버전인 CoAID$^+$ 데이터를 구축하여 이를 공공 데이터로 제공한다[2]. 본 논문에서 구축한 CoAID$^+$ 데이터는 기존의 소셜 컨텍스트 기반 탐지 기법들에 적용될 수 있으며, 나아가 향후 새로운 소셜 컨텍스트 기반 탐지 기법 연구에도 긍정적인 영향을 미칠 수 있을 것으로 기대된다. 또한 본 논문은 CoAID$^+$ 데이터를 다양한 관점에서 분석하여, 진짜뉴스와 가짜뉴스의 파급 패턴과 뉴스의 키워드에 따른 파급 패턴 등 흥미로운 분석 결과를 제공한다.

본 논문의 주요 공헌은 다음과 같다.

- CoAID을 중심으로 기 공개된 COVID-19 관련 가짜뉴스 탐지를 위한 데이터의 한계점을 파악한다.
- CoAID 데이터를 기반으로 소셜 정보를 추가 수집함으로써 확장 데이터인 CoAID$^+$ 데이터를 구축하고, 이를 공공 데이터로 제공하여 가짜뉴스 탐지 연구 분야의 활성화에 이바지한다.
- 마지막으로, CoAID$^+$ 데이터를 다양한 관점에서 분석하여 진짜뉴스와 가짜뉴스의 파급 패턴 및 고유한 특징들을 파악한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구, 3장에서는 본 논문에서 확장한 CoAID$^+$ 데이터의 구축, 구조, 통계에 대한 내용을 자세히 설명한다. 4장에서는 CoAID$^+$를 다양한 관점에서 분석한 결과를 소개한다. 마지막으로 5장에서는 본 논문의 결론을 제시하였다.

2. 관련 연구

2.1 가짜뉴스 탐지 방법

내용 정보 기반 기법 (content based methods). 뉴스의 내용에서 나타나는 언어적 특징을 활용하여 가짜뉴스를 탐지하는 방법이다. [15]에서는 뉴스 내용에서 나타나는 구문(syntactic)과 의미론적(semantic) 특징을 이용하고, [16,17]은 어휘(lexical)의 빈도를 특징으로 이용한다. 또한 [18]에서는 문체(writing style)를 이용하는 방법을 소개하였다. 그러나 이러한 내용 정보 기반 기법은 가짜뉴스가 사람들을 속이기 위해 진짜뉴스를 모방하여 정교하게 작성되고, 뉴스를 해석하기 위해서는 사회적 맥락이나 상식에 대한 지식이 필요로 되기 때문에 뉴스의 내용만으로 가짜뉴스를 탐지하는데 한계가 존재한다.

소셜 컨텍스트 기반 기법 (social context based methods). 소셜미디어에서 사용자 정보, 사용자 간 관계 정보, '좋아요(like)'나 '리트윗(retweet)'과 같은 사용자들의 행동 정보를 활용하여 가짜뉴스를 탐지하는 방법이다. [11-13]에서는 소셜미디어 상의 사용자 특성 및 사용자 간 관계 정보를 이용하였으며, [2-4,14]는 뉴스의 파급 정보를 활용하여 가짜뉴스를 탐지한다. 이처럼 소셜 컨텍스트 기반 기법은 소셜미디어 상의 모방이 어려운 다양한 정보를 활용할 수 있기 때문에 최근 가짜뉴스 탐지 분야에서 더욱 활발하게 연구되고 있다.

1) https://www.yna.co.kr/view/AKR20200813129800009
2) https://github.com/sosilver2080/CoAID-plus

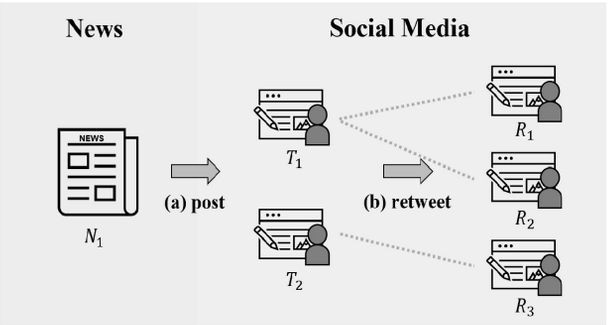

Fig. 1. The Propagation Process

2.2 CoAID 데이터

CoAID 데이터[7]는 COVID-19 관련 가짜뉴스 탐지 연구를 돕기 위해 2020년 펜실베니아 주립 대학에서 구축한 데이터로, 2019년 12월부터 2020년 7월까지 8개월간 총 9개의 신뢰할만한 미디어에 보도된 COVID-19 관련 뉴스들을 수집 대상으로 한다. 결과적으로 CoAID 데이터는 진위여부가 검증된 3,921건의 뉴스들과 이를 트위터 상에 게재한 트윗 150,002건으로 구성되며, 게재된 뉴스의 정보와 해당 트윗 아이디 정보로 구성되어 있다. CoAID 데이터에 대한 자세한 설명은 아래 링크에서 확인할 수 있다[3].

2.3 정보 파급 그래프(Casacde)

정보 파급 그래프(Cascade)란 정보가 사용자들을 통해 파급되는 과정을 나타내는 그래프로, 소셜미디어에서 어떤 뉴스가 사용자들을 통해 어떻게 파급되는지를 표현한다. 예를 들어, 미디어에 뉴스 N_1가 공개됐을 때 이를 사용자가 인용하여 소셜미디어에 게재하면 소셜미디어 게시물(tweet) T1이 생성된다(Fig 1(a)). 이후 소셜미디어의 다른 사용자 들이 해당 게시물 T_1을 공유(retweet)함으로써 R_1, R_2가 생성되며 뉴스가 소셜미디어를 통해 파급된다(Fig 1(b)). 해당 뉴스 N_1의 파급 과정은 정보 파급 그래프 $\langle N_1, T_1, T_2, R_1, R_2, R_3\rangle$로 표현된다.

Algorithm 1. Crawling process for news cascade data

Require: News article dataset D = {n_1, n_2, ..., n_n}, G $\in \mathbb{R}^n$
1: Initialize G ← ∅
2: **for** each news article n_i ∈ D **do**
3: T_i ← get_tweets(n_i)
4: **for** each tweet t ∈ Ti **do**
5: RT ← get_retweets_with_comment(t)
6: G[i].append(RT)
7: **end for**
8: **end for**
9: **Return** G

3) https://github.com/cuilimeng/CoAID

3. CoAID+ 데이터

본 논문에서는 기존 CoAID 데이터의 한계를 극복하기 위해 구축한 CoAID+ 데이터를 소개한다. 기존 CoAID 데이터에서는 COVID-19 관련 뉴스 정보와 이를 트위터 상에 처음 게재한 트윗 정보로 구성되어 있다. 즉, 뉴스가 실제 어떤 경로를 통해 파급되었는지 알 수 없기 때문에 기존 소셜 컨텍스트 기반의 가짜뉴스 탐지 기법을 이용할 수 없다. 이러한 기존 CoAID 데이터의 한계를 극복하기 위해, 본 논문에서는 트위터 상 뉴스의 파급 정보를 추가로 수집하여 파급 정보를 포함하는 CoAID+ 데이터를 소개한다. 해당 데이터는 공개된 링크에서 이용 가능하다.

3.1 데이터 구축

CoAID에서 제공된 총 3,921개 뉴스를 기반으로 하여 뉴스가 트위터 상 파급된 흐름을 추가로 수집하였다. 수집을 위해, 트위터에서 제공되는 Twitter API를 이용하여 Algorithm 1의 과정으로 수집하였다. 또한 트위터에서 리트윗은 방법에 따라 Retweet without comment와 Retweet with comment로 구분된다. 먼저 Retweet without comment의 경우, 원본 게시물을 그대로 게재하는 방식으로 어떤 사용자로부터 파급되었는지 알 수 없다. 따라서 뉴스의 정확한 파급 흐름을 알 수 없다. 반면에 Retweet with comment의 경우, 원본 게시물에 사용자의 의견을 덧붙여 게재하는 방식으로 어떤 사용자로부터 영향을 받아 파급된 것인지 명확하게 알 수 있다. 따라서 우리는 Retweet with comment를 파급 정보로 수집하였다. 이를 통해, 해당 뉴스 기사가 트위터 상에서 파급되는 전 과정의 파급 데이터를 수집하였고, 소셜 컨텍스트 기반의 가짜뉴스 탐지 기법에 적용할 수 있게 되었다. 또한, 우리는 파급 정보 뿐만 아니라, 〈tweet_id, user_id, tweet_text, create time, like/reply/retweet count〉를 트윗 정보로 수집하였다.

3.2 데이터 구축

뉴스가 파급된 흐름을 수집하였으므로 정보 파급 그래프로 데이터를 표현할 수 있다(Fig. 2). 따라서 루트노드를 뉴스로 설정하였고, 뉴스를 처음 게재한 게시물이 초기 트윗(tweet)이 된다. 이후 트윗이 사용자들 사이에서 리트윗(retweet)되어 뉴스가 파급되면서, 정보 파급 그래프를 형성한다. 즉, 하나의 뉴스 당 하나의 정보 파급 그래프가 형성된다. 또한 시간적 특징 분석을 위해, 트윗 게재 시간을 노드의 속성으로 설정하였다. 이렇게 형성된 정보 파급 그래프 분석을 통해, 구조적, 시간적인 측면에서 어떠한 파급 패턴을 갖는지 확인이 가능하다.

3.3 데이터 통계

Table 1은 CoAID+ 데이터의 통계를 나타낸다. 해당 데이

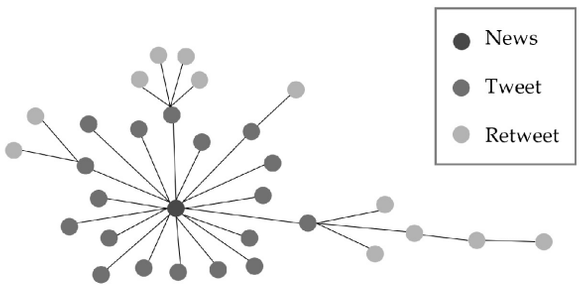

Fig. 2. The Propagation Process

Table 1. Dataset Statistics

CoAID+ Dataset			
# of true news	3,055	# of tweets	150,002
# of fake news	866	# of retweets	43,310
# of users	129,057	# of cascades	2,763

터는 총 3,921개의 뉴스와 이를 게재한 193,312개의 트윗으로 구성된다. 이 중 뉴스는 진짜뉴스 3,055개와 가짜뉴스 866개로 구분되며, 트윗은 150,002개의 초기 트윗과 이들을 공유한 43,310개의 리트윗으로 구성된다. 소셜미디어 상 뉴스의 파급에 참여한 사용자는 총 129,057 명이며, 이를 통해, 총 파급 그래프는 2,763개가 생성되었다.

4. 분 석

확장한 CoAID+ 데이터를 다양한 관점에서 분석함으로써, 우리 데이터가 어떠한 경향을 갖는지 분석하였다. 먼저 분석 특징으로는 이전 연구에서 사용한 정보 파급 그래프의 구조적, 시간적 특징들을 추출하여, CoAID+ 데이터의 정보 파급 그래프에서 구조적, 시간적 측면으로 분석하였다. 이때 트위터 상에 게재되지 않은 뉴스의 경우, 파급되지 않았다고 판단하여 분석에서 제외하였다. 즉, 파급이 발생된 1,096개의 뉴스로만 분석을 진행하였다.

4.1 분석 특징

1) 구조적 특징(Structural feature)

파급 그래프의 노드 간 연결의 구조적 패턴을 분석하는 것으로, 뉴스의 전반적인 파급 패턴을 확인할 수 있다. 구조적 특징으로 아래의 9가지에 대해서 분석하였다.

- (S1) Tree depth: 파급 그래프의 최대 깊이를 의미하며, 소셜미디어에서 사용자에 의해 정보가 얼마나 멀리 파급 되는가를 알 수 있다.
- (S2) Number of nodes: 뉴스가 파급된 사용자의 수를 의미하며, 얼마나 많은 사용자에게 파급되었는가를 알 수 있다.
- (S3) Max breadth: 파급 그래프의 거리 별 최대 트윗 수를 의미하며, 뉴스로부터 동일한 거리에 존재하는 트윗 수를 통해, 특정 거리에서 얼마나 많이 파급되었는가를 알 수 있다.
- (S4) Structural virality: 파급 그래프 상에서 모든 노드 쌍의 평균 거리를 의미하며, 파급의 형태를 알 수 있다.
- (S5) Maximum outdegree: 뉴스의 파급 과정에서 가장 많이 트윗/리트윗된 수를 의미하며, 파급에서 가장 영향력 있는 게시물을 알 수 있다.
- (S6) Number of tweets: 뉴스를 처음 트위터 상에 게재한 초기 트윗 수를 의미한다.
- (S7) Depth of node with maximum outdegree: 뉴스로부터 maximum outdegree까지의 깊이를 의미하며, 영향력 있는 노드까지의 거리를 알 수 있다.
- (S8) Number of tweets with retweets: 리트윗 된 초기 트윗의 수를 의미한다.
- (S9) Fraction of tweets with retweets: 리트윗 된 초기 트윗의 비율을 의미하며, 초기 트윗이 얼마나 리트윗 되어 파급되는가를 알 수 있다.

2) 시간적 특징(Temporal feature)

파급 그래프의 노드 간 연결의 시간적 패턴을 분석하는 것으로, 파급 과정의 빈도와 강도를 확인할 수 있다. 시간 속성으로는 각 트윗의 게재 시간을 초 단위의 절대 시간으로 환산하여 사용하였다. 시간적 특징으로는 아래의 9가지에 대해서 분석하였다.

- (T1) Average time difference between the adjacent retweet nodes: 인접한 리트윗 간의 평균 시간차를 의미하며, 얼마나 빨리 리트윗 되는가를 알 수 있다.
- (T2) Time difference between the first tweet and the last retweets: 파급 그래프 상에서 첫 번째 초기 트윗과 마지막 리트윗의 시간차를 의미하며, 파급 그래프의 수명을 알 수 있다.
- (T3) Time difference between the first tweet and the tweet with maximum outdegree: 첫 번째 초기 트윗과 maximum outdegree 까지의 시간차를 의미하며, 영향력 있는 노드까지의 도달 시간을 알 수 있다.
- (T4) Time difference between the first and last tweet posting news: 뉴스를 게재하는 첫 번째 초기 트윗과 마지막 초기 트윗의 시간차를 의미하며, 얼마나 오래 실제 뉴스가 게재되는가를 알 수 있다.
- (T5) Time difference between the tweet posting news and last retweet node in deepest path: 가장 멀

Table 2. The Extracted Structural and Temporal Features of News Articles in CoAID+

Features	Fake		True		Features	Fake		True	
	Mean	Median	Mean	Median		Mean	Median	Mean	Median
S1	2.59	2	2.47	2	T1	25,427	None	63,877	None
S2	164.53	63	159.23	53	T2	3,274,014	888,347	3,271,199	1,026,472
S3	127.77	58	126.21	45	T3	450,698	102,811	907,908	133,686
S4	2.25	2.18	2.20	2.12	T4	2,714,204	1,728,996	2,729,019	2,050,785
S5	18.33	3	19.44	2	T5	534,308	23,223	825,426	22,469
S6	118.94	48	117.57	43	T6	36,442	None	51,541	None
S7	1.04	1	1.04	1	T7	88,948	38,605	106,606	43,454
S8	10.28	3	8.35	2	T8	258,692	31,962	560,844	66,466
S9	0.17	0.10	0.10	0.06	T9	187,146	9,334	119,822	9,550

리 파급된 경로에서의 초기 트윗과 마지막 리트윗의 시간차를 의미하며, 가장 멀리 파급된 경로의 수명을 알 수 있다.

- (T6) Average time difference between the adjacent retweet nodes in the deepest path: 가장 멀리 파급된 경로에서 인접한 리트윗 간의 시간차를 의미하며, 얼마나 빈번하게 리트윗 되었는지를 알 수 있다.
- (T7) Average time between tweets posing news: 초기 트윗 사이의 평균 시간차를 의미하며, 초기 트윗 사이의 시간 간격을 알 수 있다.
- (T8) Time difference between the first tweet post time and first retweet time: 첫 번째 초기 트윗과 첫 번째 리트윗의 시간차를 의미하며, 뉴스가 처음 게재된 후, 얼마나 빨리 리트윗 되는가를 알 수 있다.
- (T9) Average time difference between the tweet post time and first retweet time: 각 초기 트윗과 첫 번째 리트윗간의 평균 시간차를 의미하며, 뉴스가 처음 게재된 후, 리트윗 되기까지 얼마나 걸리는지 알 수 있다.

4.2 진짜뉴스와 가짜뉴스의 파급 특징 비교

COVID-19 관련 진짜뉴스와 가짜뉴스가 소셜미디어 상에서 파급되는 패턴을 구조적, 시간적 특징으로 분석하였다. 이러한 분석을 통해 CoAID+가 어떠한 경향을 갖는지 확인하고자 한다. 분석 특징으로는 4.1절에서 소개한 구조적 특징과 시간적 특징을 사용하였으며, 총 18개의 특징을 분석한 결과를 Table 2로 보였다.

확장한 데이터의 구조적 특징에 대해 평균값(Mean)과 중앙값(Median)으로 분석 결과를 나타낸다. 분석 결과에서 알 수 있듯이, 평균값에서는 S5를 제외한 모든 특징에서 가짜뉴스의 값이 크고, 중앙값에서는 모든 특징에서 가짜뉴스의 값이 크거나 같다는 것을 알 수 있다. 즉, 가짜뉴스가 더 멀리, 더 많이 공유되었음을 알 수 있다. 이러한 구조적 특징 분석을 통해, 전반적으로 구조적 특징에서 가짜뉴스가 더 멀리, 더 많이 파급되는 경향을 확인하였다. 특히 S1의 경우, t-test를 통해서 통계적으로 유의미한 차이가 있음을 검증하였다.

확장한 데이터의 시간적 특징에 대해 평균값(Mean)과 중앙값(Median)으로 분석 결과를 나타낸다. 분석 결과에서 알 수 있듯이, 평균값에서는 T2, T9를 제외한 모든 특징에서 가짜뉴스의 값이 작고, 중앙값에서는 T5를 제외한 모든 특징에서 가짜뉴스의 값이 작다는 것을 알 수 있다. 즉, 가짜뉴스가 더 빨리, 더 빈번하게 공유되었음을 알 수 있다. 이러한 시간적 특징 분석을 통해, 시간적 특징에서 전반적으로 가짜뉴스가 더 빨리, 더 짧은 간격으로 파급되는 경향을 확인하였다.

이러한 분석 결과에서 시간적 특징인 T1과 T6의 중앙값이 none으로 나타나는 것을 확인할 수 있다. T1과 T6은 리트윗이 발생한 이후, 소셜미디어 상에서 얼마나 빠르게 파급되는지를 나타내는 특징이다. 즉, 뉴스 기사로부터 두 단계 이상의 파급이 진행되었을 때 측정되는 특징이며, 값이 작을수록 빠르게 파급되는 것을 의미한다. 분석 결과에서 알 수 있듯이, 50% 이상의 뉴스 기사가 두 단계 이상의 파급이 진행되지 않았으므로 T1과 T6의 특징 값이 측정 되지 않는다. 따라서 두 특징에서 중앙값이 none으로 나타나고 있다.

COVID-19 관련 진짜뉴스와 가짜뉴스의 파급 패턴을 구조적, 시간적 측면으로 분석함으로써, 가짜뉴스가 더 널리, 빠르게 파급되는 것을 확인하였다. 또한 이러한 결과는 일반적인 진짜뉴스와 가짜뉴스의 파급 패턴이 COVID-19 관련 뉴스에서도 동일하게 보인다는 것을 확인하였다.

4.3 키워드별 그룹의 파급 특징 비교

COVID-19에 대한 사람들의 관심이 높아짐에 따라, 특히 관심도가 높은 키워드인 vaccine과 face mask로 뉴스를 그룹화 하여 두 그룹의 파급 패턴을 구조적, 시간적 특징으로 분석하였다. 분석 특징으로는 4.1절에서 소개한 구조적 특징과 시간적 특징을 사용하였으며, 총 18개의 특징을 분석한 결과를 Table 3으로 보였다.

먼저 키워드별로 뉴스를 그룹화 하는 과정을 소개한다. 키

Table 3. The Difference Among the Structural and Temporal Features of News Groups (Vaccine, Face Mask)

Features	Vaccine		Face mask		Features	Vaccine		Face mask	
	Mean	Median	Mean	Median		Mean	Median	Mean	Median
S1	2.27	2	2.43	2	T1	None	None	None	None
S2	121.90	44	148.31	68.50	T2	None	None	None	None
S3	98.49	38	119.07	62.50	T3	478,803	151,116	1,161,135	488,712
S4	2.18	2.09	2.21	2.10	T4	3,951,686	2,001,884	5,592,663	3,970,103
S5	7.46	2	35.64	2	T5	None	None	None	None
S6	97.44	38	86.70	54	T6	None	None	None	None
S7	1	1	1.02	1	T7	92,246	32,016	126,239	47,389
S8	5.88	2	7.14	3	T8	448,289	61,047	812,032	395,289
S9	0.11	0.06	0.16	0.08	T9	289,059	4,234	194,553	20,781

Table 4. Keyword Group

Keyword	Related word
Vaccine	Vaccine, vaccination, vaccinating, nvaccinated, first dose, second dose, single-dose, dose, Pfizer, AstraZeneca, Oxford-AstraZenera, BioNTech, Pfizer- BioNTech, Moderna, Sputnik V, Johnson & Johnson, J&J, Janssen, Novavax, Ad26.COV2.S, Booster, side effect, generic, ChAdOx1, Convidecia, EpiVacCorona, Sinopharm, BBIBP-CorV, CoronaVac, Covaxin, WIBP- CorV, CoviVac, CanSinoBIO, Bovine vaccine, Pneumonia vaccine, COVID-19 vaccine, flu vaccine
Face mask	Face covering, facial covering, covering, mask, masks, face mask, kn95, n95, face shield, Bandana, three-layered, 3-layer, double masking, masking, wearing, breathers, cloth face covering, mask-wearing, nose-only, doubling up masks, triple-layer, triple-layered, face covering, face coverings, medical mask, surgical mask, protective

워드와 연관된 단어를 각각 추출하여 뉴스 제목에 해당 단어가 포함될 경우, 동일한 그룹으로 간주하였다. 따라서 Table 4와 같이 vaccine과 연관된 단어 37개, face mask와 연관된 단어 27개에 대하여 뉴스를 그룹화 하였다.

확장한 데이터의 구조적 특징에 대해 평균값(Mean)과 중앙값(Median)으로 분석 결과를 나타낸다. 분석 결과에서 알 수 있듯이, 평균값에서는 S6을 제외한 모든 특징에서 face mask 그룹의 값이 크고, 중앙값에서는 모든 특징에서 face mask 그룹의 값이 크거나 같다는 것을 알 수 있다. 즉, face mask 그룹이 더 멀리, 더 많이 공유되었음을 알 수 있다. 반면에 시간적 특징에 대한 분석결과를 보면, 평균값과 중앙값 모두에서 face mask 그룹의 값이 더 크다. 즉, face msk 그룹이 더 오랫동안 파급되는 경향을 확인하였다.

COVID-19 관련 키워드별 그룹을 비교 분석함으로써, face mask 그룹이 더 멀리, 더 많은 사람들에게 공유되는 것을 확인하였고, 파급이 더 오랫동안 유지되는 것을 확인하였다. 이러한 분석 결과는 데이터 수집 시기와 관계가 있을 것으로 예상된다. 데이터 수집 시기가 COVID-19 상황 초기이므로 face mask의 예방효과에 대한 관심이 높았었기 때문에 이러한 파급 패턴이 있었을 것으로 생각된다.

5. 결 론

본 논문에서는 기존 CoAID에는 포함되지 않은 소셜미디어 상의 파급 정보를 추가로 수집함으로써, 소셜 컨텍스트 기반 가짜뉴스 탐지가 가능한 CoAID$^+$ 데이터를 구축하였다. 이렇게 확장한 CoAID$^+$ 데이터의 구축방법, 구조, 통계 정보를 자세히 소개한다. 이후 CoAID$^+$ 데이터에서 정보 파급 그래프의 파급 패턴을 구조적, 시간적 측면으로 분석하였다. 이를 통해 일반적인 뉴스의 파급 패턴에서 확인된 바와 같이 COVID-19 관련 뉴스에서도 가짜뉴스가 더 멀리, 더 빨리 파급되는 것을 확인하였다. 또한, 사회적으로 관심이 많은 키워드로 뉴스를 그룹화 하여 비교 분석함으로써, COVID-19 발생 초기에 face mask에 대한 파급이 활발했음을 알 수 있다. 우리는 향후 이러한 분석 특징을 이용한 가짜뉴스 탐지 방법에 대한 연구를 진행하고자 한다.

References

[1] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," In *Proceeding of the ACM SIGKDD Explorations Newsletter*, Vol.19, No.1, pp.22-36, 2017.

[2] Y. Liu and Y. F. Wu, "Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks," In *Proceeding of the AAAI Conference on Artificial Intelligence*, 2018.

[3] K. Shu, D. Mahudeswaran, S. Wang, and H. Liu, "Hierarchical propagation networks for fake news detection: Investigation and exploitation," In *Proceeding of the International AAAI Conference on Web and Social Media*, Vol.14, pp.626-637, 2020.

[4] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, "Fake news detection on social media using geometric deep learning," *arXiv preprint arXiv:1902.06673*, 2019.

[5] K. Shu, S. Wang, and H. Liu, "Beyond news contents: The role of social context for fake news detection," In *Proceeding of the Twelfth ACM International Conference on Web Search and Data Mining*, pp.312-320, 2019.

[6] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, Vol.359, No.6380, pp.1146-1151, 2018.

[7] L. Cui and D. Lee, "Coaid: Covid-19 healthcare misinformation dataset," *arXiv preprint arXiv:2006.00885*, 2020.

[8] X. Zhou, A. Mulay, E. Ferrara, and R. Zafarani, "Recovery: A multimodal repository for covid-19 news credibility research," In *Proceeding of the 29th ACM International Conference on Information & Knowledge Management*, pp.3205-3212, 2020.

[9] G. K. Shahi and D. Nandini, "FakeCovid--A multilingual cross-domain fact check news dataset for COVID-19," *arXiv preprint arXiv:2006.11343*, 2020.

[10] M. Abdul-Mageed, A. Elmadany, E. M. B. Nagoudi, D. Paddi, K. Verma, and R. Lin, "Mega-cov: A billion-scale dataset of 100+ languages for covid-19," *arXiv preprint arXiv:2005.06012*, 2020.

[11] C. Castillo, M. Marcelo, and B. Poblete, "Predicting information credibility in time-sensitive social media," *Internet Research*, 2013.

[12] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier, "Tweetcred: Real-time credibility assessment of content on twitter," *International conference on social informatics*, Springer, Cham, 2014.

[13] Z. Jin, J. Cao, Y. Zhang, and J. Luo, "News verification by exploiting conflicting social viewpoints in microblogs," In *Proceeding of the AAAI Conference on Artificial Intelligence*, Vol.30. No.1, 2016.

[14] E. Tacchini, G. Ballarin, ML.Vedova, S. Moret and L. Alfaro, "Some like it hoax: Automated fake news detection in social networks," *arXiv preprint arXiv:1704.07506*, 2017.

[15] S. Badaskar, S. Agarwal, and S. Arora, "Identifying real or fake articles: Towards better language modeling," In *Proceeding of the Third International Joint Conference on Natural Language Processing: Volume-II*, 2008.

[16] B. Riedel, et al. "A simple but tough-to-beat baseline for the Fake News Challenge stance detection task," *arXiv preprint arXiv:1707.03264*, 2017.

[17] H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," *International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, Springer, Cham, 2017.

[18] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, "A stylometric inquiry into hyperpartisan and fake news," *arXiv preprint arXiv:1702.05638*, 2017.

[19] S. Han, Y. Kang, Y. Ko, J. Ahn, Y. Kim, S. Oh, H. Park, and S. Kim, "COVID-19 Cascade Dataset for Fake News Detection," *The KIPS Spring Conference 2021*, Vol.28, No.1, pp.312-313, 2021.

한 소 은

https://orcid.org/0000-0003-1007-8998
e-mail : sosilver@hanyang.ac.kr
2016년 덕성여자대학교 컴퓨터공학부(학사)
2020년 ~ 현 재 한양대학교
컴퓨터소프트웨어학과 석사과정
관심분야 : 데이터마이닝, 사회연결망 분석

강 윤 석

https://orcid.org/0000-0002-5892-2265
e-mail : dyskang@hanyang.ac.kr
2013년 한양대학교 컴퓨터공학(학사)
2013년 ~ 2022년 한양대학교
컴퓨터소프트웨어학과(석・박사)
2022년 ~ 현 재 한양대학교 컴퓨테이셔널
사회과학연구센터 박사후연구원
관심분야 : 사회연결망 분석, SSD에서의 데이터마이닝

고 윤 용

https://orcid.org/0000-0003-1283-4697
e-mail : koyunyong@hanyang.ac.kr
2013년 한양대학교 컴퓨터소프트웨어학과
(학사)
2013년 ~ 2021년 한양대학교
컴퓨터소프트웨어학과(석・박사)
2021년 ~ 현 재 한양대학교 인공지능 혁신인재교육 연구단
박사후연구원
관심분야 : 데이터마이닝, 분산 답 러닝, 사회 연결망 분석

안 지 원

https://orcid.org/0000-0002-3764-9917
e-mail : dkswldnjs@hanyang.ac.kr
2019년 한양대학교 물리학과(학사)
2019년 ~ 현 재 한양대학교 컴퓨터소프트웨어학과 석・박사통합과정
관심분야 : 데이터마이닝, 사회연결망 분석

김 유 심

https://orcid.org/0000-0003-4728-9588
e-mail : ykim@asu.edu
1994년 대전대학교 행정학과(학사)
1997년 서울대학교 행정대학원(석사)
2006년 Ohio State University 정책관리학(박사)
2007년 ~ 현 재 Arizona State University 행정학과 부교수
관심분야 : 공공정책분석, 정책평가, 환경정의 및 정책, 공공보건 관리, 시뮬레이션

오 성 수

https://orcid.org/0000-0003-0316-3852
e-mail : ohseongsoo@hanyang.ac.kr
1992년 한양대학교 행정학과(학사)
1999년 서울대학교 행정대학원(석사)
2009년 Georgia Tech-Georgia State University 정책학과(박사)
2011년 ~ 현 재 한양대학교 행정학과 조교수, 부교수, 교수
관심분야 : Public Management

박 희 진

https://orcid.org/0000-0002-8608-5994
e-mail : hjpark@hanyang.ac.kr
1994년 서울대학교 컴퓨터공학(학사)
1996년 서울대학교 컴퓨터공학(석사)
2001년 서울대학교 컴퓨터공학(박사)
2003년 ~ 현 재 한양대학교 정보통신학부 교수
관심분야 : 알고리즘, 정보보호, 생물정보학, 스트링매칭

김 상 욱

https://orcid.org/0000-0002-6345-9084
e-mail : wook@hanyang.ac.kr
1989년 서울대학교 컴퓨터공학(학사)
1991년 한국과학기술원 전산학과(석사)
1994년 한국과학기술원 전산학과(박사)
2006년 ~ 현 재 한양대학교 정보통신학부 교수
관심분야 : 데이터베이스 시스템, 저장시스템, 트랜잭션 관리, 데이터마이닝, 사회연결망 분석, 웹 데이터 분석