J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

온라인 공간에서 관심집단 대상 비정상 정보의 특징 분석과 탐지☆

Characterization and Detection of Opinion Manipulation on Common Interest Groups in Online Communities

이 시 형1*
Sihyung Lee

요 약

인터넷 포털과 사회관계망 서비스(SNS) 등의 온라인 공간에서 사용자 간의 의견 공유가 활발해짐에 따라 이를 악용하여 특정 개인이나 집단의 이익을 위해 유포되는 비정상 정보도 증가하고 있다. 특히 비정상 정보가 정치적인 목적으로 유포되면 선거 결과뿐 아니라 다양한 사회 정책과 시민 생활에도 영향을 미친다. 이러한 비정상 정보는 불특정 다수에 대한 유포에서 시작하였으며 이들의 특성을 분석하고 탐지하기 위한 기존 연구도 이러한 불특정 다수 대상 유포에 초점을 맞추었다. 하지만 최근에는 더욱 효과적으로 영향을 미치기 위해 공통 관심사를 가진 집단(예: 부동산에 관심 있는 사람들의 모임)을 대상으로 내용과 형식을 조정한 맞춤형 정보를 유포하고 있다. 본 논문에서는 이러한 관심 집단을 대상으로 한 비정상 정보의 특성을 분석하고 이를 탐지하는 방법을 제시한다. 이를 위해 선거 전후에 10개의 공통 관심 집단에 게시된 의견을 수집하여 분석하였다. 그 결과, 각 집단에 맞춤화된 정보가 실제 유포되고 있으며 선거일이 가까워짐에 따라 점차 증가함을 보였다. 또한, 비정상 정보를 탐지하기 위한 시스템을 제안하였는데, 이 시스템은 개별 의견에서 보이는 특징뿐 아니라 의견 게시자의 전반적인 행위 및 게시자와 협력한 사용자의 특성을 종합적으로 분석한다. 제안한 시스템을 수집한 데이터에 적용한 결과 90% 이상의 정확도로 비정상 의견을 탐지하였으며 다수의 사용자가 조직적으로 비정상 의견을 유포한 정황을 발견하였다. 제안한 시스템으로 관심 집단에 게시된 의견을 주기적으로 검사한다면 비정상 정보의 유포를 더 빠르게 차단하고 영향을 줄일 수 있을 것이다. 또한, 탐지에 활용한 특징은 정치적인 목적 이외의 비정상 정보 판별에도 활용될 수 있을 것이다.

☞ 주제어 : 비정상 정보, 온라인 공간, 관심 집단, 정치적 목적 의견 조작

ABSTRACT

As more people share their opinions in online communities, such as Internet portals and social networking services, more opinions are manipulated for the benefit of particular individuals and groups. In particular, when manipulations occur for political purposes, they influence election results as well as government policies and the quality of life. This type of manipulation has targeted the general public, and their analysis and detection has also focused on such manipulation. However, to more efficiently spread propaganda, recent manipulations have targeted common interest groups(e.g., a group of those interested in real estate) and propagated information whose content and style are customized to those groups. This work characterizes such manipulations on common interest groups and proposes method to detect manipulations. To this end, we collected and analyzed opinions posted on 10 common interest groups before and after an election. As a result, we found that manipulations on common interest groups indeed occurred and were gradually increasing toward the election date. We also proposed a detection system that examines individual opinions, their authors, and their collaborators. Using the collected opinions, we demonstrated that the proposed system can accurately classify more than 90% of manipulated opinions and that many of these opinions were posted by multiple collaborators. We believe that regular audits of opinions using the proposed system can quickly isolate manipulations and decrease their impact. Moreover, the proposed features can be used to identify manipulations in domains other than politics.

☞ keyword : Opinion Manipulation, Online Community, Interest Group, Political Manipulation

1. 서 론

인터넷에서 사용자들이 서로 의견을 공유하며 상호작용하는 공간을 온라인 공간(online community)이라 하며, 포털 사이트와 사회관계망 서비스(SNS)를 모두 포함한다. 이러한 온라인 공간은 언제 어디서나 편리하게 접속 가능하다는 장점으로 인해 많은 사용자가 일상적으로 사용하며 생각과 정보를 공유

1 School of Computer Science and Engineering, Kyungpook National University, Daegu 41566, Korea
* Corresponding author (sihyunglee@knu.ac.kr)
[Received 16 July 2020, Reviewed 23 September 2020, Accepted 13 October 2020]
☆ 이 연구는 2020년 (재)동일문화장학재단 학술연구조성비 지원에 의해 수행되었음

한다[1]. 이에 따라 온라인 공간에 게재되는 의견이 사용자들의 가치관, 행동, 선택에 점차 더 많은 영향을 미치고 있다[2,3].

온라인 공간의 영향력이 커짐에 따라 이를 악용하여 비정상 정보를 유포하는 사례도 증가하고 있다[4,5,6]. 비정상 정보란 온라인 공간의 원래 목적에 맞는 정상적인 의견 게재에 반해서 특정 개인이나 집단의 이익을 위해 의도적으로 유포되는 정보를 말하며, 허위 상품 평이나 정치적 선동 의견이 이러한 의견의 예이다. 특히 비정상 정보가 정치적인 의도를 갖고 유포되면 선거 결과에 영향을 미칠 뿐 아니라 이에 대한 후속 효과로 시민 경제나 일상생활에도 변화를 가져오게 된다[7,8]. 본 논문에서는 이와 같은 정치적인 목적의 비정상 정보를 분석하고 이를 탐지하는 시스템을 제안한다.

정치적인 목적의 비정상 정보는 온라인 공간의 불특정 다수를 대상으로 유포되기 시작하였으며, 이에 따라 이를 분석하고 탐지하는 연구도 이러한 불특정 다수를 대상으로 유포되는 의견에 초점을 맞추었다[9,10]. 하지만 최근에는 온라인 공간의 사용자를 관심도에 따라 여러 다른 공통 관심 집단으로 분류한 후, 각 집단의 특성에 맞게 의견의 내용과 형식을 조정하여 유포하는 사례가 발생하고 있다[11]. 예를 들어 특정 연령대와 성별로 구성된 집단에게는 이들과 연관된 정책에 대한 의견을 유포하면서 이 집단이 선호하는 사진이나 제목을 사용하는 것인데, 이러한 방식이 더 큰 공감대를 형성하기 때문이다.

본 논문에서는 이처럼 특정 관심 집단을 목표로 유포되는 비정상 의견의 특징을 분석하고 이를 탐지하는 시스템을 제안한다. 이를 위해 서로 다른 관심 분야를 가진 10개의 인터넷 카페에 게재된 의견을 수집하여 분석하였다. 특히 2020년 4월 국회의원 선거 전후에 게재된 의견을 분석하였는데, 선거 전후로 정치적인 목적의 비정상 정보 유포가 대규모로 발생해 왔기 때문이다[4,5]. 이러한 분석의 주요 결과 및 의미를 정리하면 다음과 같다.

- 공통 관심 집단을 대상으로 한 정치적 목적의 비정상 정보가 다수 존재하며 선거일이 다가옴에 따라 그 수가 증가함을 보였다. 또한, 이들은 각 집단의 특성에 맞게 내용이 조정되었음도 보였다.
- 의견 게시자의 행위를 분석함으로써 비정상과 정상 정보를 구분할 수 있는 103개의 특징을 제안하였다. 이들은 인터넷 카페의 고유한 특성을 고려해 선정되었으며, 정치적인 목적 이외의 비정상 정보 판별에도 활용될 수 있다.
- 제안한 특징을 활용한 비정상 정보 탐지 시스템을 구현하고 수집한 데이터를 사용해 검증한 결과 90% 이상의 정확도로 비정상 의견을 분류함을 보였다. 또한, 관심 집단을 대상으로 한 비정상 의견 유포가 여러 사용자에 의해 조직적으로 일어나고 있음도 보였다.

본 논문은 다음과 같이 구성된다. 2장에서는 본 연구의 배경을 설명하고 관련 연구와의 차이점을 비교한다. 3장에서는 인터넷 카페에서 의견을 수집한 과정을 기술하며 4장에서는 이러한 의견 중 비정상 의견을 구분할 수 있는 특징을 제안한다. 5장에서는 제안한 특징을 사용한 탐지 시스템을 구현하고 정확도를 측정한다. 마지막으로 6장에서는 향후 연구 과제를 소개하며 결론을 맺는다.

2. 관련 연구

온라인 공간에서 정치적인 목적의 비정상 의견 유포는 남미[12], 미국[6], 이탈리아[3], 인도[13], 사우디아라비아[14]를 포함해 전 세계적으로 발생하고 있으며, 선거에서 특정 후보에게 유리한 분위기를 조성하거나 반정부 시위의 여파를 최소화하는 것을 목적으로 수행되었다. 이들은 주로 트위터, 페이스북, 인스타그램 등의 온라인 공간을 통해 불특정 다수에게 정보를 전파하였다. 최근에는 페이스북에서 사용자들이 좋아요(like) 버튼을 클릭한 사례를 분석하여 이들의 관심사를 자동으로 분석한 후 관심 집단으로 분류한 연구가 있었는데[15], 이러한 결과를 기반으로 각 관심 집단에 대한 맞춤형 정보를 전파함으로써 보다 효과적으로 투표 결과에 영향을 주었다[11].

이러한 비정상 의견의 특징을 분석하고 탐지하는 연구는 주로 상업적인 목적의 의견(예: 상품에 대한 허위 사용 후기)에 대해 수행되었다. 특히 5개의 별점을 사용한 상품 후기(5-star rating system)에 대해 (1) 각 사용자의 별점이 다른 사용자들의 평균 별점과 어느 정도 차이가 나는지[16]와 (2) 한 사용자가 같은 상품에 대해 여러 번 연속해 매우 높거나(5점) 낮은(1점) 별점을 남겼는지[17] 등을 활용해 비정상 의견을 탐지하였다. 정치적인 목적의 비정상 의견은 별점을 사용하지 않으므로 이러한 특징을 바로 적용하기는 어렵다.

일부 비정상 의견에 관한 연구가 정치적인 목적의 의견을 분석하였는데, 이들 연구는 특정 관심 집단에 유포되는 의견보다는 트위터와 포털 사이트에서 불특정 다수에게

유포되는 의견을 분석하였다[9,10]. 특히 소수의 사용자가 유사한 내용을 반복해 언급하는 것을 비정상 사용자의 중요한 특징으로 보았다. 이러한 특징은 본 논문에서 다루는 관심 집단에 게시되는 비정상 의견 탐지에도 활용할 수 있다. 본 논문에서는 이에 더해서 관심 집단에서만 보이는 유포 방식을 분석하고(예: 글을 게시하는 관심 집단의 일원인 것처럼 신분을 속이고 이러한 집단에서 자주 사용하는 어투를 사용해 공감대 형성), 관심 집단의 특성을 고려한 탐지 방법을 제안한다(예: 인터넷 카페와 같은 관심 집단은 트위터나 포털 답글과 다르게 의견 길이의 제한이 엄격하지 않고 이미지와 동영상을 자유롭게 사용 가능하므로 이미지·동영상 사용에서 보이는 비정상적인 특성 활용). 또한, 이 방법을 실제 데이터에 적용함으로써 관심 집단에 유포되는 비정상 의견 탐지에 용이함을 보인다.

3. 데이터 수집과 기본 특징 분석

3.1 분석대상 데이터 수집

정상과 구분되는 비정상 정보의 특징을 분석하기 위해서는 정상과 비정상으로 분류된 데이터가 필요하다. 이러한 데이터는 또한 탐지 시스템의 정확도를 검증하기 위해서도 필요하다. 이를 위해 네이버 카페[18]와 다음 카페[19] 중 가입자 수가 50만 명 이상인 10개의 카페에 게재된 54,786개의 의견을 수집하였다. 이들 카페는 표 1과 같이 공통 관심사를 가진 집단이다. 이러한 공통 관심사에 대한 각자의 생각을 글로 게재함으로써 다른 가입자들과 공유한다. 일반적으로 카페의 관심사와 관련된 글이 다수이지만 선거가 다가오면 정치와 관련된 글의 게재가 증가한다. 특히 운영진이 정치 관련 글을 게시하지 않도록 지침을 공지하였음에도 불구하고 이러한 글이 지속해서 게재된다[20]. 이처럼 카페의 관심사에서 벗어나며 규정에 반하여 게재된 정치와 관련된 글을 비정상 정보로 분류하였다.

표 1의 카페에 게재된 글에 대한 수집은 제21대 국회의원 선거(2020.04.15.) 전후 약 6개월간, 2019.11.04.에서 2020.05.14.까지 진행되었다. 보다 영향력이 큰 글을 분석하기 위해 베스트 게시글을 수집하였는데, 이들은 최근 1주일 내 게재된 글 중 조회수와 댓글 수가 가장 높은 100개의 글을 의미한다. 베스트 게시글은 게시판 상위에 정렬되므로 더 많은 사용자가 읽게 된다.

표 2는 수집한 의견의 예를 보여준다. 각 의견은 항목 1~10까지 10개의 정보로 구성된다. 이 중 항목 1은 의견이 게재된 인터넷 카페의 이름이다. 항목 2는 의견의 제목이며, 게시한 내용에 대한 요약으로 볼 수 있다. 항목 3~4는 각각 글을 게재하는데 사용한 ID와 게재한 시간이다. 항목 5는 의견의 내용으로 임의의 길이로 작성할 수 있으며 그림이나 영상을 포함할 수 있다. 항목 6은 다른 사용자들이 이 글을 읽은 횟수이며 항목 7은 의견에 대한 응답으로 게재된 답글의 개수이다. 이들 항목 6~7은 시간에 따라 변하는 값으로 변화 과정을 관찰하기 위해 10분 간격으로 수집되었다. 항목 6의 조회수가 568→2238처럼 짧은 시간에 급격히 증가한다면 자동화된 도구와 다수의 도용된 ID를 사용해 조작되었을 가능성이 있다[21]. 이러한 주기적인 수집은 의견 게재 시점부터 시작하여 수집된 값의 변화가 거의 없어지는 1주일 후까지 수행되었다. 마지막 세 개의 항목 8~10은 이 글에 대한 응답으로 게재된 답글 각각에 대한 정보를 기록한다. 하나의 글에 여러 개의 답글이 게재될 수 있으며, 표 2는 이러한 답글 2개를 예로 보여준다. 답글의 내용은 최대 3,000자까지 작성할 수 있다.

(표 1) 의견을 수집한 인터넷 카페 (OO 부분은 익명 처리)
(Table 1) Internet communities (cafes) where opinions were collected

카페 이름	관심 주제	가입자 수
레몬OO	인테리어, 요리	≥2,900,000
맘스OO	임신, 출산, 육아	≥2,800,000
파OO룸	화장품, 미용, 쇼핑	≥1,900,000
OO카페	패션	≥1,600,000
OO사커	축구	≥1,000,000
부동산OO	부동산 투자	≥1,000,000
OO매니아	헬스, 패션	≥1,000,000
OO격투기	격투기, 헬스	≥900,000
OO10	재테크, 투자	≥700,000
도OO	게임	≥500,000

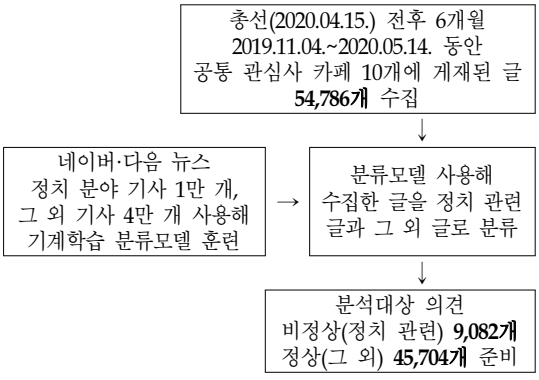
3.2 데이터 분류

수집한 의견 중 정치와 관련된 글을 선별해 비정상 정보로 분류하였다. 표 3은 이러한 분류 과정의 개요를 보여준다. 분류하는 시간을 단축하고 정확도를 높이기 위해 기계학습 모델인 Naive Bayes classifier[22]를 훈련해 활용하였다. 모델 훈련을 위한 학습 데이터(training set)로 네이버와 다음의 뉴스 기사[23,24]를 사용하였는데, 이들이 주제에 따라 정치 분야와 그 외 분야로 분류되어 있기 때문이다. 이 중 정치 분야의 기사 10,000개를 정치 관련 글

(표 2) 수집한 의견의 예 (OO 부분은 익명 처리)
(Table 2) Example of collected opinions

ID	항목	값
1	카페	OOOO
2	제목	정부의 마스크 판매 몰아주기 수혜 유통 담당 기업 OO가 OO당과 관련이
3	사용자명	User-01
4	게재시간	2020-0309, 10:10:00
5	내용	기업 OO만 유통채널로 선정해 독점적 특혜를 주었다는 … (중략) … OO당 OO 후보는 기업 OO에서 고문으로 활동하고 있다. … (그림) … 또한 기업 OO의 대표의 남편 OO씨는 OO당의 선거 캠프 출신으로 … (영상) …
6	조회수	0, 122, **568, 2238**, 2886, 3434, …, **3652**
7	답글수	0, 0, 2, 7, **15, 34**, 51, 67, 82, …, **144**
<답글 #1>		
8	사용자명	User-02
9	게재시간	2020-0309, 10:12:00
10	내용	내로남불 엄청나네요 중국 공산당과도 관련 있다고 … (중략) …
<답글 #2>		
8	사용자명	User-03
9	게재시간	2020-0309, 10:15:00
10	내용	정말 파도 파도 끝없이 나오네요. 촛불 시위하고 탄핵해야 … (중략) …
… (이후 답글은 생략) …		

(표 3) 의견 수집 및 분류 과정 개요
(Table 3) Overview of opinion collection and classification

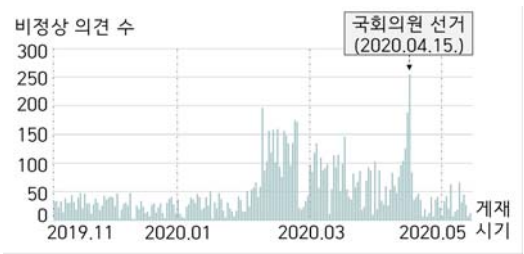


에 대한 학습 데이터로 사용하였으며, 그 외 4개 분야(문화, IT, 경제, 사회)의 기사 40,000개를 정치와 관련되지 않은 글에 대한 학습 데이터로 사용하였다. 이들 기사는 의견을 수집한 6개월 동안 게재된 기사이며, 의견 교환이 활발히 일어난 기사를 선정하기 위해 답글이 1,000개 이상인 기사를 사용하였다. 기사 본문에 포함된 명사를 특징(feature)으로 사용해 각 기사를 모델링하였으며 단어가 사용된 횟수를 이러한 특징의 값으로 사용하였다. 예를 들어 '정부'가 2회, '정책'이 1회 사용된 기사는 {'정부'=2, '정책'=1}로 모델링된다. 학습 데이터에서의 명사 추출은 KoNLPy[25] 패키지를 사용해 구현하였으며 이를 사용한 Naive Bayes 모델의 훈련은 Weka 패키지[26]를 사용해 구현하였다. 학습 데이터에 대해 10-fold cross validation을 수행하였을 때 분류의 정확도는 91.45% AUC(Area Under ROC Curves)였다.

이처럼 훈련한 모델을 사용해 수집한 의견을 정치 관련 글과 그렇지 않은 글로 분류하였으며, 분류한 결과를 연구자가 다시 한번 검토하여 잘못 분류된 경우를 수정하였다. 최종적으로 54,786개의 의견 중 9,082개(16.58%)를 정치 관련 글(비정상 의견)로 분류하였다.

3.3 비정상 의견 개수와 내용에서 보이는 특징

비정상 의견을 분류한 후에는 이들에 나타나는 기본적인 특징을 분석해 보았다. 먼저 비정상 의견의 수가 시기에 따라 어떻게 변하는지를 분석하였으며, 다음으로 비정상 의견의 내용에서 보이는 특징을 분석하였다. 이러한 분석은 비정상 의견이 정확히 분류되었음을 검증하는 역할을 하며, 탐지 시스템 개발에 들어가기에 앞서 데이터의 특성을 파악하는 데도 도움이 된다.



(그림 1) 비정상 의견 수의 변화 추이
(Figure 1) Number of manipulated opinions over time

그림 1은 비정상 의견이 게시된 수를 시기별로 보여준다. 총선 약 2개월 전인 2020년 2월부터 비정상 의견의 수가 전반적으로 증가하였다. 이러한 추세는 총선 전인 4월 중순까지 계속되었으며, 총선 직전에 가장 많이 증가했다

가 총선 후에는 급감하였다. 이는 비정상 의견이 선거 결과에 영향을 미치기 위해 게재되었음을 보여준다. 또한, 정치적인 논쟁을 일으킬 수 있는 이슈가 발생할 때 이를 활용한 비정상 의견의 수가 증가하였으며(예: 코로나-19가 급격히 퍼진 2020년 2월 중순[27], 공적 마스크 유통업체를 선정한 직후인 2020년 3월 초[28] 등), 베스트 게시글의 30% 이상을 비정상 의견이 점유하기도 하였다.

비정상 의견의 내용을 분석해 보면 게시된 카페에 따라 그 주제와 어투가 다름을 볼 수 있었다. 표 4는 두 개의 서로 다른 카페에 게시된 의견 및 이에 대한 응답으로 게재된 답글의 예를 보여준다. 카페 #1은 부동산이 관심 분야인데, 이에 게재된 비정상 의견의 내용도 집값과 전·월세 등 관심 분야를 반영하였다. 카페 #2는 미용이 관심 분야이며 '룸메', 'ㅠㅠ', 'ㅋㅋ' 등의 표현이 자주 사용되는데, 비정상 의견에서도 이러한 어투를 볼 수 있었다. 정리하면 비정상 의견은 그 내용과 어투를 각 관심 집단의 특성에 맞추어 조정하는데, 이는 사용자들의 공감을 얻고 영향력을 높이는 데 도움이 된다[11].

(표 4) 카페에 따른 의견 내용과 어투의 차이
(Table 4) Differences in opinion contents and style in different Internet communities

항목	카페 #1	카페 #2
카페 특성	부동산이 주된 관심 분야임	미용이 관심 분야 'ㄹ메', 'ㅠㅠ', 'ㅋㅋ' 등이 자주 사용됨
비정상 의견 내용	OOO 당선 후 현시점기준 업적 뭐 있나요? 궁금해서요. 딱히 기억 나는 게 없어서 … (중략) …	ㄹ메님들 북한이 OOO 왜 싫어하는 거예요? ㅠㅠ 지금까지 OOO만큼 북한 좋아하고 사랑이 넘치는 … (중략) … ㅋㅋㅋㅋㅋㅋ
답글 #1	집값 상승 ㄴ 에잇 무슨 상승이에요. 집값 폭등이죠.	OOO가 처음부터 북한 옹호했으니 … (중략) … ㅋㅋㅋㅋㅋ
답글 #2	전월세가 상승	ㄹ메님 말씀처럼 OOO 탄핵 안되는 게 신기해요. ㅋㅋㅋㅋ … (중략) … 밀어준 우리들이 OOO죠 ㅠㅠ

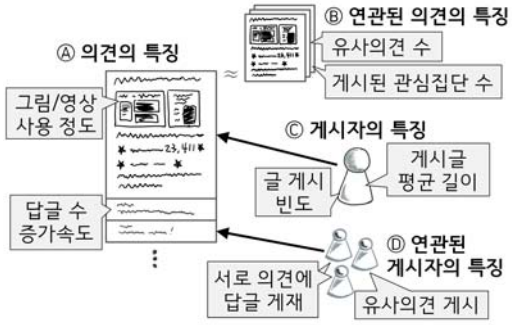
4. 행위기반 모델링 및 특징 분석

4.1 행위기반 특징 개요

관심 집단을 대상으로 유포되는 비정상 정보를 탐지하기 위해, 탐지에 활용할 수 있는 다양한 특징을 분석하였다. 이 중 정상 의견과 비정상 의견을 구분하는 데 도움이 되는 103개의 특징(feature)을 선별하였으며, 이러한 특징을 사용하여 각 의견을 모델링하고 탐지하였다. 이 장에서는 이러한 특징을 차례로 제시한다.

이 특징들은 의견에서 사용한 어휘보다는 의견을 게재하는 행위를 주로 나타낸다. 어휘는 시기별 주요 이슈의 변화(예: 코로나-19 → 마스크 → 선거) 및 카페의 관심 분야(예: 부동산, 미용, 헬스)에 따라 달라지지만, 행위에서 보이는 특징(예: 그림과 영상을 자주 사용, 여러 사용자가 조직을 이루어 의견 게재)은 여러 다른 시기나 관심 분야에서 널리 공통으로 발견되기 때문이다. 따라서 행위에 기반을 둔 특징을 사용하면 시기에 따른 영향이 적으며 여러 관심 집단에 대해 활용 가능한 탐지 시스템을 개발할 수 있다.

선별된 특징은 그림 2와 같이 4개의 범주 Ⓐ~Ⓓ로 나누어진다. 범주 Ⓐ는 각 의견 O_i에서 보이는 특징으로, 의견의 길이, 그림의 사용 정도, 답글 수, 게시한 시간 등을 포함한다. 범주 Ⓑ는 의견 O_i와 연관된 다른 의견을 기술하는 특징으로, O_i와 내용이 유사한 의견의 수, 이러한 의견이 게시된 관심 집단의 수 등을 포함한다. 범주 Ⓒ는 의견 O_i를 게시한 사용자 U_i가 보이는 특징으로, U_i가 글을 게시하는 빈도, 게시한 글의 평균 길이 등을 포함한다. 범주 Ⓓ는 사용자 U_i와 연관된 다른 사용자를 기술하는 특



(그림 2) 의견을 기술하는 특징의 4가지 범주
(Figure 2) Four categories of features that characterize opinions

징으로, U_i와 서로의 의견에 답글을 게재한 사용자의 수, 이러한 사용자의 비정상 정도 등을 포함한다.

이렇게 4가지 범주의 특징을 함께 활용하면 의견 O_i에 대해 더 잘 이해할 수 있으므로 비정상 여부를 판단하는 데 도움이 된다. 예를 들어 범주 Ⓐ의 특징을 활용해 O_i의 답글 수가 비정상적으로 급격히 증가함을 발견했다면, 범주 Ⓑ의 특징을 이용해 이와 유사한 글이 다수의 다른 관심 집단에도 게재되었음을 알 수 있다. 또한, 범주 Ⓒ의 특징을 사용해 이 글의 게시자 U_i가 O_i와 비슷한 특징을 가지는 글을 자주 게시해 왔음을 알 수 있으며, 범주 Ⓓ의 특징을 활용해 O_i에 답글을 게재한 많은 사용자가 U_i와 서로 답글을 게재하며 협력하는 관계임도 알 수 있다.

표 5는 범주 Ⓐ~Ⓓ에 속하는 103개의 특징을 보여준다. 비정상 의견과 정상 의견을 구분하는 데 도움이 되는 특징을 사용하기 위해 information gain 지표가 0.1 이상인 특징을 선정하였다. 이어지는 4.2~4.5 장에서는 이러한 특징을 범주별로 기술한다.

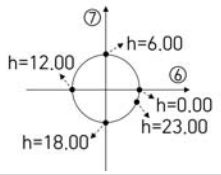
4.2 범주 Ⓐ: 의견에서 보이는 특징

범주 Ⓐ는 각 의견 O_i에서 볼 수 있는 특징이다. 먼저 표 5의 특징 ①~⑤는 의견 O_i의 본문이 보여주는 특징으로, 의견의 길이 및 그림·영상·링크·특수문자·숫자의 사용 정도를 나타낸다. 비정상 의견은 주장하는 바에 신빙성을 더하기 위해 그림과 영상, 외부 자료에 대한 링크, 보고서에 실린 숫자 등을 근거로 사용한다. 또한, 특수문자를 사용해 여러 근거를 목록으로 정리해 제공하거나 특정 부분을 강조하기도 한다. 결과적으로 비정상 의견은 이러한 근거자료로 인해 길이가 길어지는 경우가 많다.

특징 ⑥~⑦은 의견 O_i가 게시된 시간을 나타낸다. 시간의 주기적인(cyclical) 성질을 나타내기 위해 cosine과 sine 두 가지 값의 조합을 사용한다. 예를 들어 O_i가 00시 00분에 게재되었다면 h=00.00이며, 특징 ⑥~⑦에 의해 (cosine(2π×00.00÷24), sine(2π×00.00÷24)) = (1.00, 0.00)으로 표현된다. 따라서 (0.00, 1.00)으로 표현되는 06시 00분보다는 (0.97, -0.26)로 표현되는 23시 00분에 더 가까움을 알 수 있다. 이러한 표현 방법을 사용하면 비정상 의견이 특정 시간대에 자주 게재되는지를 확인할 수 있으며, 여러 다른 사용자가 협력해 유사한 시간대에 의견을 게재하는 것도 포착할 수 있다.

특징 ⑧~⑮는 의견 O_i의 조회수 및 답글수에서 보이는 특징이다. 특징 ⑧~⑨는 최종적으로 수집된 값으로 표 2의 의견의 경우 각각 3652와 144이다. 특징 ⑩~⑪은 10분 간격으로 수집한 값의 최댓값으로 표 2의 의견의 경우 각각 1670(568→2238로 +1670 했을 때 가장 많이 증가)과 19(15→34로 +19 했

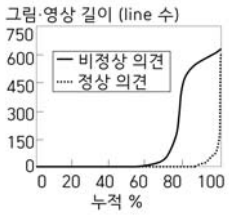
(표 5) 4가지 범주 Ⓐ~Ⓓ에 속하는 특징
(Table 5) Features in four categories Ⓐ~Ⓓ

세부 유형	특징의 표현 방법과 단위
범주 Ⓐ: 의견 O_i의 특징	
의견 내용	① 의견 길이: 기본 폰트 사용 시 줄(line) 수 ② 그림·영상 길이: 그림·영상이 차지하는 줄 수 ③ URL 수, ④ 특수문자 수 ⑤ 숫자 수: 연속된 숫자는 하나로 셈
게재 시간	⑥ cosine(2π×h÷24) ⑦ sine(2π×h÷24) h: O_i가 게시된 시간을 00.00~23.99 사이 값으로 표현 
노출 정도	⑧~⑨ 조회수, 답글수: 최종 수집 값 ⑩~⑪ 2개 값의 최대 증가량 Δ: 10분 단위 ⑫~⑬ 2개 값의 최대 증가시간 −의견게재시간 ⑭ 조회수 1,000 도달 시간 −의견게재시간 ⑮ 답글수 50 도달 시간 −의견게재시간
범주 Ⓑ: O_i와 연관된 다른 의견에서 보이는 특징	
유사 의견	⑯ O_i와 내용이 유사한 의견 수
	⑰ 유사한 의견이 게시된 관심 집단 수
	⑱~⑲ 유사한 의견의 비정상 score의 합과 평균
범주 Ⓒ: 의견 O_i의 게시자 U_i의 특징	
의견 내용	⑳~㊴ 게시한 의견에서 특징 ①~⑤ 값의 분포*
게재 시간	㊵~㊼ 게시한 의견에서 특징 ⑥~⑦ 값의 분포*
노출 정도	㊽~㊾ 게시한 의견에서 특징 ⑧~⑮ 값의 분포*
게시 빈도	㊿ 의견을 게시한 관심 집단 수 (81) 유사의견을 게시한 관심 집단 수 (82) 비정상 의견을 게시한 관심 집단 수 (83)~(84) 게시한 의견의 총수와 월평균 (85)~(88) 관심 집단별 게시한 의견 수의 분포* (89) 게시한 의견 중 유사의견 수 (90)~(91) 게시한 의견의 비정상 score의 합과 평균
범주 Ⓓ: U_i와 연관된 다른 사용자의 특징	
협력 정도	(92) U_i와 서로 의견에 답글 게재한 사용자 수 (93) U_i가 게시한 글과 유사의견 게시한 사용자 수 (94)~(101) 특징 (92) 또는 (93)에 해당하는 사용자의 특징 ⑳~(91) 값의 분포* (102) O_i와 유사의견 게시한 사용자 수 (103) 특징 (92)나 (93)에 해당하는 사용자가 O_i에 게재한 답글의 수
분포*: 최댓값·최솟값·중간값·평균 4개 값	

을 때 가장 많이 증가)이다. 특징 ⑫~⑬은 최댓값을 보인 상대 시간으로 의견 게재 30분 후에 최댓값을 보였다면 0.50시간이 된다. 이와 유사하게, 특징 ⑭~⑮는 조회수와 답글수가 각각 1,000과 50에 도달하기까지 걸린 상대 시간이다.

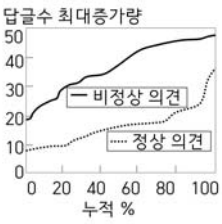
그림 3은 특징 ②, ⑪, ⑮에 대한 누적 퍼센티지를 보여준다. 세로축은 각 특징의 값을 나타내며, 가로축은 세로축이 보여주는 값을 가지는 의견의 누적 퍼센티지를 보여준다. 그림 3-(a)를 예로 들면, 실선으로 표시된 비정상 의견의 약 60%는 그림·영상의 길이가 0이므로 그림·영상을 사용하지 않음을 알 수 있으며, 나머지 40%는 그림·영상을 사용함을 알 수 있다. 이에 반해 정상 의견은 약 12~13%만이 그림·영상을 사용한다. 정리하면 비정상 의견에서 그림·영상을 더 자주 사용하며, 따라서 그림·영상의 사용 여부가 비정상 의견을 구분할 수 있는 한 가지 단서가 될 수 있음을 알 수 있다. 5장에서는 이를 포함한 특징 ①~⑱을 종합적으로 고려하여 의견의 정상과 비정상 여부를 판별한다. 그림 3-(b)는 비정상 의견에서 사용한 그림의 예를 보여주는데, 기사나 칼럼 등을 캡처하여 근거로 제시한 경우가 많았다.

그림 3-(c)와 (d)는 비정상 의견에 대한 답글이 정상 의견에 비해 빠르게 게재되며, 50개의 답글이 게재되는데 50분 미만이 걸림을 보여준다. 특히 비정상 의견 게시 후 먼저 게재되는 50개의 답글은 다수가 비정상 의견에 찬성하는 내용이었다. 즉 비정상 의견 게시 후 최대한 빨리 답글을 게재하며 이에 동의하는 분위기를 조성하여 베스트 게시글에 들어갈 수 있도록 한다.

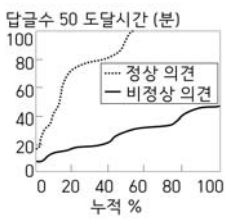


(a) 특징② 그림·영상 길이

(b) 그림을 사용한 예



(c) 특징⑪ 답글수 최대증가량



(d) 특징⑮ 답글수 50 도달 시간

(그림 3) 범주 Ⓐ에서 선별된 특징에 대한 누적 퍼센티지

(Figure 3) Cumulative percentage of selected features in category Ⓐ

4.3 범주 Ⓑ: 연관된 의견에서 보이는 특징

범주 Ⓑ는 의견 O_i와 연관된 다른 의견에서 볼 수 있는 특징으로, 연관된 의견은 O_i와 내용이 유사한 의견을 의미한다. 유사한 비정상 의견을 다수 유포하는 것은 더 많은 사용자에게 영향을 미치기 위해 자주 사용되는 전략이다[9,10]. 따라서 유사의견과 관련된 특성은 비정상 의견을 판별하기 위한 단서로 사용될 수 있다.

표 5의 특징 ⑯~⑰은 의견 O_i와 유사한 의견의 개수 및 이들이 게시된 서로 다른 관심 집단의 개수이다. 의견 O_j가 두 가지 기준 중 하나 이상을 만족하는 경우 O_i의 유사 의견으로 본다. 첫 번째 기준은 아래와 같은 Jaccard Coefficient(JC)[22]가 임계치 T_{JC} 이상인 경우이다.

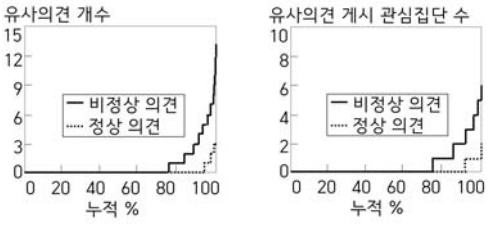
$$JC(O_i, O_j) = \frac{| W(O_i) \cap W(O_j) |}{| W(O_i) \cup W(O_j) |}$$

$W(O_i)$는 의견 O_i에 사용된 단어의 집합이다. 따라서 JC 값은 두 의견 O_i와 O_j가 공통된 단어를 사용한 정도를 나타낸다. 이러한 지표를 사용함으로써 의견 전체를 완전히 복사한 경우뿐 아니라 일부 단어를 유사한 단어로 대체한 때도 탐지할 수 있다. 유사의견을 판별하는 두 번째 기준은 의견 본문에 같은 그림이나 영상을 사용한 경우이다. 이 기준을 사용하면 텍스트 없이 그림이나 영상 파일로만 이루어진 의견도 서로 유사한지 확인할 수 있다.

특징 ⑱~⑲는 의견 O_i와 유사한 의견 O_j의 비정상 score의 합과 평균이다. O_j의 비정상 score는 비정상 의견을 판별하는 분류기(classifier)의 출력값으로 0~1 사이의 값을 가진다. 이 값이 0에 가까울수록 O_j가 정상에 가까움을 나타내며, 1에 가까울수록 O_j가 비정상에 가까움을 나타낸다. 분류기의 구현은 5장에서 기술한다. 만약 유사한 의견이 없다면 특징 ⑱~⑲의 값은 0이다. 이러한 특징을 사용함으로써 비정상인 유사의견을 가지는 의견 또한 비정상으로 분류될 가능성이 커진다.

그림 4는 특징 ⑯~⑰에 대한 누적 퍼센티지를 보여준다. 이로부터 비정상 의견이 정상 의견보다 많은 유사의견을 가지며, 더 많은 관심 집단에 게시됨을 알 수 있다.

특히 20% 이상의 비정상 의견이 유사의견을 가졌으며, 많은 경우 10개 이상까지 유사의견이 게시되었다. 본 연구에서 분석한 10개의 관심 집단 외에 다른 관심 집단도 분석한다면 더 많은 유사의견이 발견될 것으로 추정된다.



(a) 특징⑯ 유사의견 개수 (b) 특징⑰ 유사의견 게시된 관심 집단 개수

(그림 4) 범주 Ⓑ에서 선별된 특징에 대한 누적 퍼센티지
(Figure 4) Cumulative percentage of selected features in category Ⓑ

4.4 범주 Ⓒ: 의견 게시자에서 보이는 특징

범주 Ⓒ는 의견 O_i를 게시한 사용자 U_i에서 볼 수 있는 특징으로, U_i가 여러 의견을 게시하는 행위를 종합적으로 보았을 때 어떤 특성을 보이는지를 기술한다. 이러한 특징을 활용해 U_i가 일반적으로 비정상적인 행위를 많이 해왔음이 드러난다면, U_i가 게시한 의견 O_i도 비정상으로 판별될 가능성이 커진다. 이렇게 일반적인 특징을 나타내기 위해 4개의 통계량(최댓값·최솟값·중간값·평균)을 사용하며 이를 표 5에서는 분포*로 표기하였다.

표 5의 특징 ⑳~㊴는 U_i가 게시한 의견 전체를 보았을 때 특징 ①~⑮ 값이 보이는 분포를 나타낸다. 예를 들어 첫 4개의 특징 ⑳~㉓은 의견의 길이를 나타내는 특징 ① 값의 분포로 각각 U_i가 게시한 의견 길이의 최댓값·최솟값·중간값·평균을 나타낸다. 마찬가지로 다음 4개의 특징 ㉔~㉗은 특징 ② 값의 분포로 각각 U_i가 게시한 의견에서 사용한 그림·영상 길이의 최댓값·최솟값·중간값·평균을 나타낸다. 이렇게 4개의 특징씩 차례로 특징 ①~⑮ 값 각각의 4가지 통계량에 대응된다.

특징 ㊵~㊶은 U_i가 의견을 게시한 빈도수와 관련된 특징이다. 이 중 특징 ㊵~㊷는 U_i가 의견을 게시한 서로 다른 관심 집단의 수를 나타내며, 특징 ㊸~㊾는 U_i가 게시한 의견의 개수와 관련된 통계량이다. 특징 ㊿~(51)은 U_i가 게시한 의견의 전반적인 비정상 정도를 보여준다.

비정상 의견과 정상 의견에 대해 범주 Ⓒ의 특징을 비교해본 결과 비정상 의견의 게시자가 더 많은 의견을 게시하였으며, 더 많은 관심 집단에 글을 게시함을 알 수 있었다. 또한, 한 게시자가 게시한 여러 글을 비교해보면 표 6과 같이 서로 다른 신분을 가장하며 그에 맞는 서로 다른 어투를 사용하는 것도 볼 수 있었다. 이처럼 글을 게시한 관심 집단에 맞게 그 집단의 일반적인 구성원인 것으로 가장함으로써 글의 신뢰도를 높인다.

(표 6) 같은 게시자가 서로 다른 신분을 가장한 예
(Table 6) Cases where the same user disguises multiple different identities

항목	의견 #1	의견 #2
글을 게시한 카페 특성	40대 이상 남성이 다수	20~30대가 다수
비정상 의견 내용	안녕하세요 저는 OO에 사는 50대 직장인이고 두 아이의 아버지입니다. 지금 OO 상황이 얼마나 심각한지 알리려고 … (중략) …	저도 대학생인데요, 등록금 때문에 아르바이트 … (중략) … 국민 세금으로 무조건 퍼주기 … (중략) …

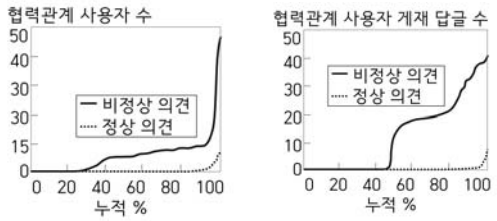
4.5 범주 Ⓓ: 연관된 사용자에서 보이는 특징

범주 Ⓓ는 의견 O_i를 게시한 사용자 U_i와 연관된 다른 사용자에서 볼 수 있는 특징이다. 연관된 사용자란 협력의 가능성이 있는 사용자로, 서로의 의견에 답글을 게재하며 지원해 주거나 유사한 의견을 게시한 사용자를 의미한다. 비정상 의견 게시자들은 더 효과적인 유포를 위해 조직을 이루어 유포하며[16], 따라서 협력과 관련된 특성은 비정상 의견 판별을 위한 단서로 사용될 수 있다.

표 5의 특징 (52)~(53)은 의견 O_i의 게시자 U_i와 협력 관계에 있는 사용자의 수를 세며, 특징 (54)~(60)은 이러한 사용자들이 게재한 글이 어느 정도로 비정상에 가까운지를 측정한다. 특징 (62)~(63)은 의견 O_i에서 볼 수 있는 협력의 흔적을 나타내며 각각 O_i와 유사한 의견을 게시한 사용자 수 및 O_i에 게재된 답글 중 협력 관계에 있는 사용자가 게재한 답글 수이다.

그림 5는 특징 (52)+(53) 및 특징 (63)에 대한 누적 퍼센티지를 보여준다. 이로부터 비정상 의견의 게시자는 정상 의견의 게시자에 비해 다른 사용자와 협력했음을 보여주는 단서가 훨씬 많음을 알 수 있다. 많은 경우 50명에 가까운 사용자와 협력 관계를 볼 수 있었으며 답글 중 최대 40개 이상이 협력 관계의 사용자가 비정상 의견에 동의하며 게재한 답글이었다. 이러한 답글 대부분이 비정상

의견 게시 1시간 이내에 게재되었는데, 이는 최대한 빨리 베스트 게시글에 들어갈 수 있도록 하기 위함이다.



(a) 특징㉒+㉓ 협력 관계 사용자 수 (b) 특징㉛ 협력 관계 사용자가 게재한 답글 수

(그림 5) 범주 Ⓓ에서 선별된 특징에 대한 누적 퍼센티지
(Figure 5) Cumulative percentage of selected features in category Ⓓ

5. 탐지 시스템 구현 및 분석 결과

5.1 탐지 시스템 개요와 Iterative Classification

4장에서 기술한 네 가지 범주의 특징을 종합적으로 활용하는 비정상 의견 탐지 시스템을 구현하고 정확도를 측정하였다. 4장에서 각 범주의 특징을 개별적으로 분석하였다면 5장에서는 이들을 함께 조합해 사용할 때 볼 수 있는 특성을 분석한다. 이렇게 네 가지 범주 Ⓐ~Ⓓ의 특징을 종합해 사용하는 것은 Ⓐ 비정상이 의심되는 의견 O_i에서 시작하여 Ⓑ O_i와 유사한 의견의 특성을 확인하고 Ⓒ O_i의 게시자 U_i의 행위를 확인한 후 Ⓓ U_i와 자주 상호작용하는 사용자의 행위를 검사해 최종적으로 O_i가 비정상임을 확증하는 과정이라고 볼 수 있다.

비정상 의견 탐지 시스템은 지도학습(supervised learning)에 기반을 둔 분류기(classifier)를 사용해 구현했다. 3장에서 수집한 의견 각각을 범주 Ⓐ~Ⓓ에 속하는 103개 특징으로 모델링한 후 이를 사용해 분류기를 훈련하고 검증하였다. 이 과정에서 한 가지 유의할 부분은 의견을 모델링하는데 사용한 일부 특징이 다른 연관된 의견의 분류 결과를 필요로 한다는 것이다(다른 의견의 비정상 score와 관련된 특징 ⑱~⑲, ⑨⓪~⑨①, ⑨④~⑩⓪). 이로 인해 의견 O_i의 특징값을 결정하고 분류하려면 먼저 이와 연관된 의견 O_j를 분류해야 하고, 반대로 O_j의 특징값을 결정하고 분류하려면 먼저 이와 연관된 O_i를 분류해야 하는 모순되는 상황이 발생한다. 이러한 상호 참조 문제를 해결하기 위해 Iterative Classification Algorithm(*ICA*)[29]을 사용했다.

*ICA*는 분류를 여러 번 반복해 수행(multiple iterations)함으로써 점진적으로 최종적인 분류 결과를 도출하는 기법을 말하며, 본 연구에서와 같이 여러 분류 대상의 분류 결과가 서로 연관되어 있을 때 사용한다. 이 방법에서는 표 7과 같이 먼저 각 의견의 비정상 score를 0.5(정상과 비정상의 중간값)로 초기화한다. 그 후 매 iteration 마다 분류기로 의견을 다시 분류하며 비정상 score를 결정한다. 이렇게 직전 iteration에서 결정한 비정상 score를 사용해 의견을 모델링한 후 분류하고, 분류 결과에 따라 비정상 score를 조정하는 것을 반복한다. 각 의견의 비정상 score가 더 이상 변하지 않는 수렴 상태에 이르렀거나 또는 지정된 횟수만큼 분류를 반복했다면 iteration을 종료한다. 본 연구에서는 대부분의 경우 10번의 iteration 이내에 결과가 수렴하는 것을 볼 수 있었다. 이렇게 iteration이 종료한 후에는 비정상 score가 0.5를 초과하는 의견은 비정상으로, 그렇지 않은 의견은 정상으로 분류한다.

(표 7) 반복 분류 알고리즘
(Table 7) Iterative classification algorithm

O_i.score: 의견 O_i의 비정상 score ∈ 0(정상)~1(비정상)
i_{max}: 최대 iteration 횟수

```
00 for each O_j
01    O_j.score = 0.5      // 중간값으로 초기화
02
03 for i=0;i≤i_max;i++   // 반복 분류
04     for each O_j
05         classify O_j to determine O_j.score
06     if O_j.score remains same for all j, then break
```

5.2 실험 환경 구성

분류기를 위한 지도학습 모델은 Adaptive Boosting with Decision Stump[30], Convolutional Neural Network[31], Random Forest[32], Support Vector Machine[33]을 사용하였다. 모델에 따른 분류의 정확도에는 큰 차이가 없었으나 Random Forest의 정확도가 가장 높았다. 이를 사용한 결과를 5.3장에서 기술한다. Random Forest 모델은 Weka 패키지[26]를 사용해 구현하였으며 3.4GH CPU(Intel Core i7-6700)와 16GB RAM을 장착한 PC에서 훈련하고 실험을 진행하였다.

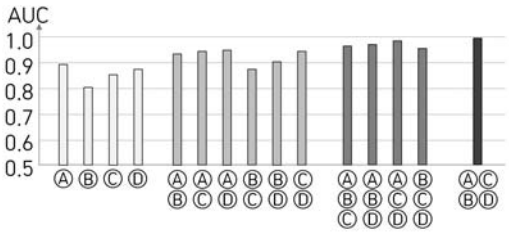
수집한 의견을 103개의 특징으로 모델링 할 때는 각 특징을 0~1 사이 값으로 정규화(normalization)하여 변화 정도가 큰 일부 특징에 의해 분류 결과가 좌우되지 않도록 하였다. 또한, 특징 ⑯~⑲, ⑧①, ⑧⑨, ⑨③~⑩③의 값을 결정하기 위해

서는 의견 간의 유사도를 판단해야 하는데, 이때 필요한 임계치 T_{JC} 값으로 0.5를 사용하였다. 이는 텍스트의 2/3 이상이 중복되는 경우 유사한 의견으로 보았음을 의미한다.

수집한 의견 54,786개 중 70%인 38,350개를 지도학습 모델의 훈련에 사용하였고(training set) 나머지 30%인 16,436개를 훈련한 모델의 정확도 측정에 사용하였다(test set). 본 연구에서 사용한 데이터에는 정상 의견(45,704개)이 비정상 의견(9,082개)보다 많으므로 이 비율 그대로 훈련에 사용하면 정상 의견 쪽으로 편향된 모델이 학습될 가능성이 크다[34]. 정상 의견으로 분류하면 맞을 확률이 높기 때문이다. 이러한 편향을 줄이기 위해 re-sampling[35] 기법을 적용해 정상 의견의 수와 비정상 의견의 수를 유사하게 조정한 후 모델을 훈련하였다.

5.3 분류 결과

그림 6은 네 가지 범주 Ⓐ~Ⓓ의 특징을 다르게 조합해 사용했을 때 분류의 정확도를 보여준다. 그림 가장 왼쪽 네 개의 막대는 각 범주를 개별적으로 사용했을 때의 결과를 보여주며 오른쪽으로 갈수록 더 많은 범주를 조합해 사용한 결과를 보여준다. 각 범주를 개별적으로 사용할 때는 80~90%의 정확도를 보이다가 더 많은 범주를 사용할수록 정확도가 증가하며 네 개의 범주를 모두 함께 사용하면 98.3%에 이른다. 이는 서로 다른 범주의 특징을 조합해 사용하면 상호 보완됨을 의미한다. 예를 들어 의견에서 보이는 특징(범주 Ⓐ)에 더해 의견의 게시자에서 보이는 특징(범주 Ⓒ)을 함께 분석하면 비정상 의견을 더 정확하게 분류할 수 있다.



(그림 6) 4개 범주의 특징을 다르게 조합해 사용했을 때 분류의 정확도

(Figure 6) Classification accuracy for when different combinations of four feature categories were used

비정상 의견에 대해 네 가지 범주 Ⓐ~Ⓓ의 특징을 종합적으로 분석해 본 결과 유사한 패턴으로 함께 활동하는 계정의 그룹을 다수 확인할 수 있었다. 이들은 (1) 비슷한 시간대에 (2) 같은 관심 집단에 (3) 유사한 글을 게시하였으며, (4) 서로의 게시글에 동의하는 답글도 게재하였다. 일부 계정은 (5) 유사한 문자열로 이루어져 있었는데(예: iiiijjjj, iiijjjjj, iijjjjjj 등), 이들은 불법으로 취득한 개인정보를 사용해 함께 생성되었을 가능성이 있다[36]. 이렇게 유사한 행위를 하는 그룹은 많은 경우 40개 이상의 계정으로 이루어지기도 하였는데, 여러 사용자가 조직적으로 활동하는 것일 수도 있으며 한 사용자가 여러 계정을 번갈아 가며 사용하는 것일 수도 있다. 이처럼 유사한 행동 패턴을 보이는 계정을 클러스터링(clustering) 등을 사용해 군집(cluster)으로 묶어 분석하는 것도 비정상 행위를 탐지하는 좋은 방법일 것이다.

표 8은 네 가지 범주 Ⓐ~Ⓓ의 특징을 모두 함께 사용할 때의 분류 결과를 분류별로 보여준다. 세로축은 test set에서의 의견의 분류를 나타내며, 가로축은 기계학습 모델을 사용해 분류한 결과를 나타낸다. 이로부터 비정상 의견의 92.85%는 비정상으로 정확히 분류되었으며 7.15%는 정상으로 잘못 분류(false negative)되었음을 알 수 있다. 또한, 정상 의견의 95.59%는 정상으로 정확히 분류되었으며 4.41%는 비정상으로 잘못 분류(false positive)되었음을 알 수 있다.

(표 8) 4개 범주의 특징을 사용한 분류 결과

(Table 8) Classification results by utilizing features in four categories

		분류 결과	
		비정상	정상
Test set	비정상 의견 (2,725개)	**92.85%**	7.15%
	정상 의견 (13,711개)	4.41%	**95.59%**

비정상 의견 중 정상으로 잘못 분류된 7.15%는 정치와 관련된 의견이지만 범주 Ⓐ~Ⓓ의 특징에서는 명확한 이상 징후를 보이지 않았다. 특히 이들 의견의 게시자 대부분은 1~2개 미만 소수의 글만 게시하였는데, 일정 시간 일회성으로 사용하는 목적으로 구매해 불법으로 활용한 계정으로 추측된다[36]. 본 연구에서는 수집하지 못한 각 계정에 대한 자세한 접속 정보(예: 접속한 IP 주소와 지역, 접속 시간 등)를 활용한다면 이러한 계정을 탐지하는 데 도움이 될 것이다.

정상 의견 중 비정상으로 잘못 분류된 4.41%는 정치와

관련되지 않은 의견이지만 범주 Ⓐ~Ⓓ의 특징에서는 이상 징후를 보여 비정상으로 탐지되었다. 이들 대부분은 특정 제품이나 서비스를 광고하는 상업적인 글로 유사한 내용이 여러 번에 걸쳐 여러 다른 관심 집단에 게시된 것을 볼 수 있었다. 많은 카페에서 이러한 광고 글을 게시하지 않도록 권고하고 있으므로, 이들을 탐지하는 것은 비정상 의견 제거에 도움이 될 것이다.

6. 결 론

본 논문은 온라인 공간에서 공통 관심 집단에 유포되는 정치적 목적의 비정상 정보의 특징을 분석하고 이를 탐지하는 시스템을 제시하였다. 국회의원 선거 전후에 공통 관심사를 가진 인터넷 카페에 게시된 글을 수집하여 분석한 결과, 각 관심 집단의 특성을 고려한 비정상 정보가 실제로 유포되고 있음을 확인하였으며 다수의 사용자에 의해 조직적으로 유포되고 있음도 보였다. 이를 탐지하기 위해 비정상 정보와 정상 정보를 구분할 수 있는 103가지 특징을 제안하였는데, 이들은 시기에 따라 자주 변할 수 있는 특정 어휘(예: 시기별 주요 정치 이슈)에 기반하기보다는 시기에 따른 영향이 더 적은 게시자의 행위(예: 다른 사용자와의 협력 정황)에 기반한 특징이다. 이러한 각 특징의 가중치를 학습하는 비정상 정보 탐지 시스템을 구현하고 수집한 데이터에 적용한 결과 90% 이상의 정확도로 비정상 정보를 분류해 냄을 보였다.

본 논문에서 제안한 비정상 정보의 특징과 탐지 시스템은 인터넷 카페 외에도 공통 관심 집단을 나타내는 다른 다양한 온라인 공간(예: 유투브 채널)에도 적용해 볼 계획이다. 또한, 제안한 방법이 정치적인 목적 외에 다른 목적으로 유포되는 비정상 정보(예: 허위 상품 후기 등 상업적인 목적의 비정상 정보)의 탐지에도 효과적인지 살펴보고자 한다.

참고문헌(Reference)

[1] M. Fraser and S. Dutta, “Throwing sheep in the boardroom: how online social networking will transform your life, work and world,” Wiley, 2008. https://books.google.co.kr/books?id=BlfPVTcFPyQC

[2] D. Centola, “The spread of behavior in an online social network experiment,” Science, vol. 329, no. 5596, pp. 1194-1197, 2010. https://doi.org/10.1126/science.1185231

[3] D. Mocanu, L. Rossi, Q. Zhang, M. Karsai, and W. Quattrociocchi, “Collective attention in the age of (mis)information,” Computing in Human Behavior, Vol. 51, Part B, pp. 1198-1204, 2015. https://doi.org/10.1016/j.chb.2015.01.024

[4] S. Choe, “Prosecutors detail attempt to sway South Korean election,” The New York Times, 2013. https://www.nytimes.com/2013/11/22/world/asia/prosecutors-detail-bid-to-sway-south-korean-election.html

[5] S. Choe, “Ally of south Korean leader conspired to rig online opinion, inquiry finds,” The New York Times, 2018. https://www.nytimes.com/2018/08/27/world/asia/moon-jae-in-online-scandal.html

[6] S. Shane and M. Mazzetti, “Inside a 3-year Russian campaign to influence U.S. voters,” The New York Times, 2018. https://www.nytimes.com/2018/02/16/us/politics/russia-mueller-election.html

[7] R. Bond, et al., “A 61-million-person experiment in social influence and political mobilization,” Nature, vol. 489, no. 7415, pp. 295-298, 2012. https://www.nature.com/articles/nature11421

[8] R. K. Garrett and B. E. Weeks, “The promise and peril of real-time corrections to political misperceptions,” ACM CSCW, pp. 1047-1058, 2013. https://dl.acm.org/doi/10.1145/2441776.2441895

[9] J. Ratkiewicz, M. D. Conover, M. Meiss, B. Goncalves, A. Flammini and F. Menczer, “Detecting and tracking political abuse in social media,” AAAI ICWSM, 2011. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2850

[10] S. Lee, “Detection of political manipulation in online communities through measures of effort and collaboration,” ACM Transactions on the Web, Vol. 9, No. 3, Article No. 16, 2015. https://doi.org/10.1145/2767134

[11] H. Grassegger and M. Krogerus, “The data that turned the world upside down,” VICE Media Group, 2017. https://www.vice.com/en_us/article/mg9vvn/how-our-likes-helped-trump-win

[12] J. Robertson, M. Riley, and A. Willis, “How to hack

an election," Bloomberg, 2016. https://www.bloomberg.com/features/2016-how-to-hack-an-election/

[13] K. Kohli, "Congress vs. BJP: the curious case of trolls and politics," The Times of India, 2013. https://timesofindia.indiatimes.com/india/Congress-vs-BJP-The-curious-case-of-trolls-and-politics/articleshow/23970818.cms

[14] B. Collins and S. Wodinsky, "Twitter pulls down bot network that pushed pro-Saudi talking points about disappeared journalist," NBC News, 2018. https://www.nbcnews.com/tech/tech-news/exclusive-twitter-pulls-down-bot-network-pushing-pro-saudi-talking-n921871

[15] W. Youyou, M. Kosinski, and D. Stillwell, "Computer-based personality judgments are more accurate than those made by humans," PNAS, 2015. https://doi.org/10.1073/pnas.1418680112

[16] M. B. Khalifa, Z. Elouedi, and E. Lefevre, "Evidental group spammers detection," IPMU, pp. 341-353, 2020. https://dx.doi.org/10.1007/978-3-030-50143-3_26

[17] V. Gupta, A. Aggarwal, and T. Chakraborty, "Detecting and characterizing extremist reviewer groups in online product reviews," IEEE Transactions on Computational Social Systems, Vol. 7, No. 3, pp. 741-750, 2020. https://doi.org/10.1109/TCSS.2020.2988098

[18] "Naver cafe," Naver, 2020. https://cafe.naver.com

[19] "Daum cafe," Daum, 2020. http://cafe.daum.net

[20] "Bulletin board posting guidelines," Administrator of real-estate cafe in Naver, 2020. https://cafe.naver.com/jaegebal/258720

[21] J. Lee, "Manipulation of recommendation counts by military and government agencies," Media Today, 2013. http://www.mediatoday.co.kr/news/articleView.html?idxno=112725

[22] B. Liu, "Web data mining: exploring hyperlinks, contents, and usage data," Springer, 2011. https://doi.org/10.1007/978-3-642-19460-3

[23] "News articles in Naver news," Naver, 2020. https://news.naver.com

[24] "News articles in Daum news," Daum, 2020. https://news.daum.net

[25] E. L. Park and S. Cho, "KoNLPy: Korean natural language processing in Python," HCLT, pp. 133-136, 2014. https://konlpy.org/

[26] "Weka: the workbench for machine learning," 2020. https://www.cs.waikato.ac.nz/ml/weka/

[27] S. Choe, "Shadowy church is at center of coronavirus outbreak in South Korea," The New York Times, 2020. https://www.nytimes.com/2020/02/21/world/asia/south-korea-coronavirus-shincheonji.html

[28] H. K. Lee, "South Korea takes new measures to have enough face masks domestically amid coronavirus," ABC News, 2020. https://abcnews.go.com/International/south-korea-takes-measures-face-masks-domestically-amid/story?id=69254114

[29] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad, "Collective classification in network data," AI Magazine, Vol. 29, No. 3, pp. 93-106, 2008. https://doi.org/10.1609/aimag.v29i3.2157

[30] R. E. Schapire, "A brief introduction to boosting," IJCAI, Vol. 2, pp. 1401-1406, 1999. https://dl.acm.org/doi/10.5555/1624312.1624417

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Communications of the ACM, Vol. 60, No. 6, pp. 84-90, 2017. https://dl.acm.org/doi/10.1145/3065386

[32] L. Breiman, "Random forests," Springer Machine Learning, Vol. 45, No. 1, pp. 5-32, 2001. https://doi.org/10.1023/A:1010933404324

[33] V. N. Vapnik, "The nature of statistical learning theory," Springer, 2000. https://doi.org/10.1007/978-1-4757-3264-1

[34] G. H. Nguyen, A. Bouzerdoum, and S. L. Phung, "A supervised learning approach for imbalanced data sets," ICPR, pp. 1-4, 2008. https://doi.org/10.1109/ICPR.2008.4761278

[35] G. Dupret and M. Koda, "Boostrap re-sampling for unbalanced data in supervised learning," Elsevier European Journal of Operational Research, Vol. 134, No. 1, pp. 141-156, 2001. https://doi.org/10.1016/S0377-2217(00)00244-7

[36] Y. Kim, "We can do everything with Naver IDs illegally sold at only 800 won," Money Today, 2020. https://v.kakao.com/v/20200205133018153

◐ 저 자 소 개 ◑

이 시 형(Sihyung Lee)
2000년 KAIST 전자전산학과 졸업(공학사)
2004년 KAIST 전자전산학과 졸업(공학석사)
2010년 Carnegie Mellon University 전자컴퓨터공학과 졸업(공학박사)
2010년~2011년 IBM TJ Watson 연구소(박사후연구원)
2011년~2019년 서울여자대학교 정보보호학과 교수
2019년~현재 경북대학교 컴퓨터학부 교수
관심분야: 인터넷 의견분석, 기계학습, 컴퓨터 네트워크, 네트워크 보안
E-mail : sihyunglee@knu.ac.kr