JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 23, No. 6, pp. 1115-1122, Jun. 2022

Check for updates

기계학습 기반 유튜브 악플 분석: “사이버렉카”에 달린 댓글의 어휘적 특성

이 신 행[1*] · 이 주 연[2] · 조 민 정[2] · 박 태 강[3]
[1]*중앙대학교 미디어커뮤니케이션학과 조교수
[2]중앙대학교 미디어커뮤니케이션학과 석사
[3]네덱스 대표

Machine Learning-Based Analysis of Malicious Comments on YouTube: Lexical Features of Comments on “Cyber Wrecker”

Shin Haeng Lee[1*] · Ju Yeon Lee[2] · Min-Jeong Jo[2] · Tae-Kang Park[3]

[1]*Assistant Professor, Department of Media and Communication, Chung-Ang University, Seoul 06974, Korea
[2]Master of Arts, Department of Media and Communication, Chung-Ang University, Seoul 06974, Korea
[3]CEO, NedX, Seoul 06211, Korea

[요 약]

본 연구는 특정 유명인에 대한 선정적 유튜브 콘텐츠로 혐오를 조장하고 악성 댓글(이하 악플)을 확산시키는 일명 “사이버렉카” 채널에 주목해 여기에 달린 댓글을 수집한 후 기계학습 알고리즘으로 악플을 분류하여 그 어휘적 특성을 분석했다. 이를 위해 로지스틱 회귀 모델을 기계학습 알고리즘으로 사용하고 예측 성능을 높이기 위해 과적합을 방지하는 정규화 과정을 거쳤다. 그 결과, “사이버렉카” 콘텐츠는 욕설이나 비속어보다는 외모 비하나 조롱 목적의 멸칭과 모욕적 상징이 함축된 고유 명사가 사용되는 악플을 양산하고 있었고 이 과정에서 다양한 언어적 변이가 일어나고 있음을 발견했다. 이러한 결과를 바탕으로 기계학습의 방법을 이용한 악플 탐지의 가능성을 진단하고 그 한계를 극복하는 방안을 논의했다.

[Abstract]

Considering the so-called “cyber wrecker,” which spreads hatred with sensational YouTube content about celebrities, this study collected comments posted on its channels, classified malicious comments with a machine learning algorithm, and analyzed their lexical characteristics. To this end, a logistic regression model was used as the algorithm and a regularization process was applied to improve prediction performance by preventing overfitting. As a result, we found that “cyber wrecker” content produced malicious comments using proper nouns, which connoted a derogatory or insulting meaning for mocking purposes, rather than swear words or slang. Also, various linguistic variations were found in the posting of malicious comments. Based on these results, we discussed the machine learning method for detecting malicious comments and ways to overcome its limitations.

색인어 : 악성 댓글, 유튜브 콘텐츠, 사이버렉카, 기계학습, 텍스트 마이닝

Keyword : Malicious comments, YouTube contents, Cyber wrecker, Machine learning, Text mining

http://dx.doi.org/10.9728/dcs.2022.23.6.1

Received 19 April 2022; **Revised** 23 May 2022
Accepted 09 June 2022

***Corresponding Author; Shin Haeng Lee**

Tel: +82-2-820-5174
E-mail: shinlee@cau.ac.kr

Ⅰ. 서 론

인터넷과 스마트폰의 폭넓은 확산으로 인해 누구나 능동적으로 미디어를 변형하고 재생산하는 활동에 참여하는 컨버전스 문화(convergence culture)는 디지털 시대의 미디어 콘텐츠의 원리를 이해하기 위한 주요 현상이다[1]. 그리고 여기에는 유튜브(YouTube)와 같은 소셜미디어 플랫폼에 기반한 개인 미디어 콘텐츠의 생산과 소비가 자리하고 있다. 실제로 유튜브는 콘텐츠 크리에이터의 약 82.2%가 사용하고 있는 것으로 나타나 개인 미디어 시대의 대표적 플랫폼으로 떠오르고 있다[2]. 하지만 유튜브를 통한 개인 미디어의 대중화는 동시에 자극적이고 선정적인 콘텐츠의 생산은 물론, 이로 인한 악성댓글(이하 악플)의 확산이라는 문제를 야기하고 있다. 특히, 유튜브에서 최근 빠르게 성장하고 있는 콘텐츠인 “사이버렉카”는 특정 이슈와 인물에 대한 논란을 소개하고 더 나아가 조롱, 비하, 멸시 등의 모욕성 표현으로 혐오 정서를 조장할 뿐만 아니라 차별과 배제를 선동하는 내용으로 대상에 대한 악플을 생산하는 주된 경로로 지목받고 있다.

“사이버렉카”란 사회, 문화, 정치, 연예 등 다양한 영역에서 이슈 혹은 논란이 된 각종 사건·사고들에 대해 관련 정보를 재구성해 전달하거나 비판하는 영상 콘텐츠를 제공해 유튜브 사용자의 시청과 참여를 유도함으로써 상업적 이득을 취하는 크리에이터(주로 이슈 유튜버)와 커뮤니티 사용자들을 일컫는 멸칭이다. 이는 마치 교통사고가 일어나면 사후 처리를 위해 경쟁적으로 출동하는 견인차(렉카)와 유사하다는 의미로 사용된다. 물론 이렇게 유명인에 대한 가십거리를 선정적이고 경쟁적으로 보도하여 경제적 이득을 취하는 행태는 새로운 현상이 아니다. 황색언론으로 대표되는 선정적 보도 행태는 경쟁적 신문 시장의 도래와 함께 나타나 저널리즘의 상업화로 인한 문제로 비판받아온 지 오래다. 그러나 “사이버렉카”의 문제점은 익명성에 기대 콘텐츠를 선정적으로 구성하여 자칫 여론의 왜곡을 불러올 수 있다는 것에 그치는 것이 아니라 그로 인해 파생된 악의적이고 공격적인 댓글이 혐오를 조장해 사회적 갈등과 분열을 심화시킬 수 있다는 점이다.

그렇다면 유튜브에서 빠르게 성장 중인 “사이버렉카” 채널에 달리는 악플 공격의 주요 대상은 누구이고 피해의 정도는 어떠한가? 과거에는 악플 공격은 소위 연예인과 같은 유명인들에게 국한된 문제로 여겨졌다. 하지만 SNS(social networking service)와 개인 미디어 콘텐츠 산업의 확장으로 일반인들 또한 사이버 범죄의 상황에 직접적으로 노출되고 있는 추세다[3]. 특히, 악플은 대상이 되는 개인에게 심리적·정서적으로 부정적인 영향을 미칠 뿐만 아니라 온라인 공간에서의 활동을 위축시키고 집단 구성원으로서의 존엄성까지 훼손시킬 수 있다[4]. 더욱이 개인 미디어 콘텐츠를 생산하는 유튜브 크리에이터들의 경우 불특정 다수의 시청자들과의 소통이 중요하기 때문에 악플에 대한 노출 빈도와 민감도가 상대적으로 높을 수 밖에 없다. 반면에 유튜브와 같은 개인 미디어 플랫폼을 통한 콘텐츠 소비와 소통이 빈번해지며 악플 공격과 같은 사이버폭력의 가해 경험 역시 증가하고 있는 추세이고 여기에는 크리에이터에 의한 욕설이나 비방, 자극적인 표현이 심각한 영향을 주고 있는 것으로 나타났다[5].

결국, “사이버렉카”로 인한 악플이 가져올 수 있는 사회적 문제는 이러한 악플의 현황과 특성을 파악해 적절한 대응방안을 마련할 필요성을 제기한다. 물론 유튜브를 비롯한 온라인 플랫폼 기업들이 콘텐츠 크리에이터를 보호하고 건전한 온라인 소통의 장을 조성하기 위한 노력을 기울이고 있는 것은 사실이다. 한 예로, 대다수의 포털 웹사이트와 SNS에서는 사용자가 불쾌감을 느낄 수 있는 메시지에 대해 직접 신고할 수 있는 기능을 제공하고 있다. 하지만 신고 메시지와 가이드라인 위반 사항에 대해 즉각적인 피드백이 반영되고 처벌의 과정이 이뤄지기까지는 많은 자원과 시간이 소요돼 효과적인 대응 방안이 될 수 없다는 지적이다. 이에 대다수의 웹 플랫폼에서는 기계학습(machine learning)을 활용해 악플을 즉각적이고 항시적으로 탐지하는 방안을 적극 활용하고 있다. 그러나 인공지능 기반 악플 탐지 모델은 댓글의 대상과 맥락에 따라 그 성능이 크게 달라진다는 한계를 보인다[6]. 더욱이 댓글을 분류하고 차단하는 기능의 인공지능은 사용자 간의 소통 과정에 개입함으로써 담론 형성에 영향을 미칠 수 있다는 점에서 이러한 분류 방식과 결과를 면밀하게 검토할 필요가 제기된다. 따라서 본 연구는 유튜브의 “사이버렉카” 채널에 달린 악플을 기계학습의 방법으로 분류하고 이들을 구성하는 언어적 특성을 파악하여 보다 정교하고 신뢰할 수 있는 자동화 악플 분류기의 개발과 적용에 기여하고자 한다.

Ⅱ. 이론적 논의

유튜브를 필두로 한 개인 미디어 플랫폼의 사용 범위와 빈도가 빠르게 성장함에 따라 기존에 없던 방식의 새로운 유명 콘텐츠 제작자들이 두각을 드러내게 되었다. 이들은 유튜버라는 이름으로 활동하며 각자의 영역에서 특화된 콘텐츠를 제공하는 채널을 운영하는데, 그 종류는 뉴스와 교육과 같은 정보성 채널은 물론 오락과 뷰티 등의 생활 방식을 다루는 채널까지 매우 다양하다. 그리고 사회적 이슈와 유명인에 대한 논란거리를 다룸으로써 흥미 위주의 콘텐츠 소비를 유도하는 “사이버렉카” 또한 유튜브에서 빠르게 성장한 형태의 채널 중 하나다. 그러나 “사이버렉카”는 선정성을 무기로 한 콘텐츠 제작으로 특정 집단이나 개인에 대한 고정관념과 편견을 증폭시키고 혐오를 조장하는 문제를 낳고 있다는 지적이 제기되고 있다. 특히 사회적 소수자나 약자에 대한 혐오적 표현이 악플을 통해 게시되고 확산될 경우 이들에 대한 부정적 관념과 차별이 심화될 뿐만 아니라 악플의 대상자 역시 정신적 피해는 물론 온라인 공간에서의 활동이 위축될 수 있다는 점에서 문제의 심각성이 대두되고 있다.

그렇다면 유튜브와 같은 온라인 미디어 플랫폼을 통해 악플이 생산되고 확산되는 이유는 무엇일까? 여기에는 인터넷

에 기반한 소통 방식의 구조적 특성이 주요한 원인으로 작용한다. 초창기 인터넷은 사회적 이슈에 대한 논의에 있어 개인이 자유로운 방식으로 정보를 생산하고 소비하며 쌍방향의 소통이 가능하다는 점에서 건전한 여론 형성의 창구로서의 역할로 기대되며 급속한 성장을 이뤄왔다[7]. 그러나 그 과정에서 온라인의 비대면적이고 익명성의 특징으로 인해 악플과 같은 사이버 폭력이 빈번하게 발생하며 사회 구성원 간의 반목과 갈등을 심화시킬 수 있다는 문제의식 역시 생겨나게 되었다[8]. 특히 온라인 환경에서 이용자는 실제 자신의 삶, 그리고 정체성과는 무관하게 행동하는 경향이 있는데, 이는 오프라인 환경에서의 사회적 관계와 지위가 미치는 영향력이 최소화되고 가해 행동에 대한 처벌 위험이 낮다고 인식되며 공격적이고 폭력적인 성향이 쉽게 표출될 수 있다[9]. 더욱이 온라인 공간은 에코챔버(echo-chamber)효과와 같이 자신의 신념 혹은 의견과 비슷한 목소리에만 선택적으로 노출되어 확증 편향이 쉽게 일어나기 때문에 집단극화로 인한 반목과 혐오가 증가하고 있다[10].

그렇다면 유튜브에서 확산되고 있는 악플은 어떻게 정의될 수 있을까? 그동안의 악플에 대한 개념적 논의를 살펴보면 우선, 김민기와 이진로(2008)은 타인에 대한 욕설과 비방, 사생활 침해, 폭력, 음담패설 등으로 개념화한다[11]. 악플의 범위와 유형을 분석한 안태형(2013)의 연구에선 상대를 저주·협박하는 내용이나 사회 통념에 위배되는 내용으로 타인에게 피해를 주는 댓글로 악플을 정의한다[8]. 또한 혐오의 관점에서 악플을 바라보면 사회적 소수자로서의 속성을 가진 개인과 집단에 대한 차별, 적의, 폭력을 표출하고 선동하는 표현으로 정의할 수 있다[12]. 이상의 논의를 정리해보면 악플은 악의를 바탕으로 타인에 대한 언어적 폭력을 행사하는 개념임을 알 수 있다. 그러나 이러한 악플의 개념화만으로는 유튜브에서 최근 심각한 문제로 대두되고 있는 유명인은 물론 개인 유튜버를 대상으로 생산되고 확산되는 악플의 특성을 파악하기에는 무리가 따른다. 따라서 본 연구는 “사이버렉카” 채널에 주목해 이들 영상에 달리는 댓글에서 악플을 기계학습으로 분류해 그 언어적 특성을 포착하고자 한다.

Ⅲ. 본 론

본 연구는 유튜브에서 유명인에 대한 정보 전달과 담론 형성에 있어 “사이버렉카” 콘텐츠가 악플 생산과 확산에 주된 역할을 담당한다는 전제하에 유명 채널의 영상에 달린 댓글을 수집하고 악플을 분류했다. 그리고 기계학습의 방법으로 댓글에서 자동으로 악플을 판별하는 분류기를 다음의 절차로 만들었다. 우선, 수집된 댓글 전체에 대해서 악플을 표시하는 라벨링 작업을 진행했다. 그리고 악플 여부가 라벨링 된 댓글 데이터로부터 어휘적 특성을 추출하고 이를 바탕으로 악플을 예측하기 위한 최적의 알고리즘을 도출했다. 또한 악플 예측을 위한 최적의 알고리즘을 시험하고 그 결과를 검증함으로써 기계학습 기반 유튜브 악플 분류의 가능성과 한계를 진단했다.

3-1 데이터 수집

본 연구의 데이터는 다음의 절차로 수집했다. 우선, “괴인협회”와 “악인전 – 인물소개” 등 대표적 인물 관련 “사이버렉카” 채널을 선정했다. 2022년 3월 24일 “괴인협회”와 “악인전 – 인물소개” 채널에 업로드된 영상은 각각 216개와 67개이고 이들 영상에 대한 총 조회수는 각각 199,433,611회와 71,484,906회였다. 그리고 각 채널의 최신 콘텐츠 75개의 영상에 달린 총 59,999개의 댓글을 웹 크롤링으로 수집했다. 이렇게 수집된 콘텐츠는 2021년 5월 7일 “악인전 – 인물소개” 업로드된 “아이린, 당신이 몰랐던 14가지 사실”에서 2021년 11월 29일 “괴인협회”에 업로드된 “유병재, 당신이 몰랐던 15가지 사실”에 걸쳐있었다. 물론 본 연구의 댓글 데이터는 인물에 대한 채널에 집중된 만큼 유튜브 내 모든 “사이버렉카” 채널에서 생성되는 댓글을 대표하기에는 무리가 있다는 점이 고려될 필요가 있다.

3-2 데이터 라벨링

“사이버렉카” 영상에 달린 유튜브 댓글에서 악플을 자동적으로 판별하는 알고리즘 개발을 위해 악플의 특성을 학습시키기 위한 훈련용 데이터셋(training dataset)과 판별 결과를 검증하기 위한 시험용 데이터셋(test dataset)을 생성했다. 여기에는 악플과 그렇지 않은 댓글을 구분하는 라벨링 작업이 선행됐다. 그리고 악플은 욕설과 비속어는 물론 선정성과 폭력성, 조롱 및 차별 등이 포함된 표현 등으로 정의해 구별했다. 이러한 라벨링 작업에는 3인의 미디어커뮤니케이션 대학원의 석사과정 재학생이 참여했는데 이들은 선행 연구를 통해 악플에 대한 개념화와 조작적 정의에 기초한 악플 분류를 신뢰할 수 있는 수준으로 도출했다. 실제 코더 간 신뢰도 측정을 위해 무작위로 추출한 921건의 댓글에 대해 악플 분류를 진행해 일치도를 살펴본 결과, 크론바흐 알파(Cronbach’s alpha)값이 약 0.78인 것으로 나타나 신뢰할 만한 수준임이 드러났다. 또한 악플 여부에 대한 판단이 어려운 댓글에 대해서는 코더들의 상의 후 다수의 코더가 동의한 결과값을 부여했다. 그 결과, 코더에 의해 라벨링된 댓글 총 59,999건 중 악플로 분류된 댓글의 수는 2,851건이었다. 즉, 전체 댓글에서 악플이 차지하는 비중은 4.75%였다. 이러한 불균형 데이터셋은 기계학습에 의한 분류 알고리즘의 성능에 영향을 줄 수 있기 때문에 악플이 아닌 댓글에서 일부만 임의로 표집하는 균형화 작업으로 총 5,702건의 댓글로 구성된 최종 데이터셋을 이용했다.

3-3 어휘 특성 추출

본 연구는 어휘에 기반해 악플을 예측해 분류하는 기계학습 모델을 구현하기 위해 다음의 과정을 거쳤다. 우선, 준비된 데이터셋을 구성하는 댓글로부터 어휘적 특성을 추출하기 위해 자연어처리 알고리즘을 이용한 형태소 분석을 실시했다. 즉, 한국어 댓글에서 악플을 예측하기 위해 언어를 구성하는 기본 단위인 형태소를 파악하고 의미를 전달하는 체언, 용언, 관형사 및 부사 등의 품사 어휘들만 추리는 과정을 거쳤다. 특히 댓글의 문장을 형태소 단위로 분리하는 토큰화와 각 토큰의 품사를 태그하는 작업은 형태소 분석기에 따라 성능 차이가 두드러지는 만큼 지능형 한국어 형태소 분석기로써 범용적으로 뛰어난 성능을 보이는 Kiwi(Korean Intelligent Word Identifier)를 사용했다. 이 분석기는 '세종계획 말뭉치'와 '모두의 말뭉치'를 사용해 모델을 학습해 웹 텍스트와 문어 텍스트 분석에 있어 각각 약 87%와 94%의 높은 정확도를 보여주고 있어 댓글 대상 형태소 분석에 적합하다고 판단했다. Kiwi를 R 환경에서 구현하고자 "elbird" 패키지를 사용해 형태소 분석을 실시한 결과 5,702건의 댓글이 총 139,631개의 토큰(형태소)으로 구분되었으며 이 중 체언, 용언, 관형사 및 부사의 어휘만 추린 결과 총 9,131개의 어휘가 76,872회 사용되고 있는 것으로 드러났다. 결국 댓글에서 평균 13.48개의 어휘가 추출되었는데 가장 많은 어휘가 추출된 댓글에서는 88개의 어휘가 가장 적은 어휘가 추출된 댓글에서는 하나의 어휘가 추출되었다.

3-4 기계학습 모델링

형태소 분석으로 댓글에서 추출된 어휘 특성을 학습하고 이를 이용해 새로운 댓글이 악플에 해당하는지에 대한 여부를 분류하는 모델을 도출하기 위해 다음의 절차로 기계학습 모델을 적용했다. 우선, 기계학습으로 도출된 모델의 성능을 평가하기 위해 5,702건의 데이터셋을 훈련용 데이터셋(4,276건)과 평가용 데이터셋(1,426건)으로 구분했다. 그리고 훈련용 데이터셋을 이용해 악플 분류 모델을 학습시키기 위해 추출한 어휘 특성 중 빈도수 기준 상위 3,000개의 어휘에 대한 TF-IDF(Term Frequency - Inverse Document Frequency) 가중치를 악플 분류를 위한 독립 변인으로 사용했다. TF-IDF 가중치는 문서 전체에서의 중요도를 나타내는 단어 빈도수(Term Frequency)가 문서 간의 언어적 특성 차이를 보여주는 한계를 극복하기 위한 방안이다. 즉, 특정 단어가 많은 문서에서 등장할수록 해당 단어는 문서 간의 차이를 파악하는데 효과적이지 않기 때문에 다른 문서에서는 많이 등장하지 않지만 특정 문서에서 많이 등장하는 단어의 중요도를 높이기 위해 역문서 빈도수(Inverse Document Frequency)를 포함해 가중치를 산출한다.

이후 어휘 특성을 학습해 악플을 분류하기 위한 알고리즘으로 로지스틱 회귀(Logistic Regression, 이하 LR) 모델을 선택했다. LR 모델은 독립 변인으로 예측하는 종속 변인을 선형적 관계에 기초해 문서의 범주를 예측해 결과에 대한 이해가 쉬울 뿐만 아니라 간편하지만 효과적인 모델이 가능하다. 여기에서 간편하지만 효과적인 모델이란 데이터의 구조를 간단하게 설정함으로써 학습을 통해 얻어진 결과가 이상적 모델과의 차이가 커지게 되지만(강한 편향), 데이터로부터의 학습 결과의 편차는 작아지는(약한 분산) 경향을 띈다. 결국, 기계학습으로 정확한 악플 분류를 기대하기 위해서는 편향이 크지 않아 학습 데이터에 대한 설명력이 확보되면서도 모델이 너무 복잡하게 구성되지 않아 분산이 작게 나오는 최적화가 요구된다.

이를 위해 LR 모델을 이용한 학습시 필요 이상으로 복잡해지지 않도록 조절하는 라쏘(lasso) 정규화가 필요하다. 라쏘 정규화란 독립 변인인 다수의 어휘 특성 중에 중요하지 않은 변인들을 제외함으로써 예측에 불필요한 특성까지 학습해 새로운 데이터에 대한 분류 성능이 떨어지는 과적합을 방지하는 기법이다. 따라서 최적의 정규화를 위한 LR 모델을 결정하기 위해 튜닝(tuning) 과정을 거쳤다. 여기서 튜닝이란 독립 변인인 어휘 특성의 개수와 함께 LR 모델의 정규화를 어느 정도의 강도로 적용할지 규정하는 패널티(penalty)값을 조정해가며 가장 좋은 성능의 모델을 찾는 방법을 의미한다.

그리고 훈련용 데이터셋에 기반한 학습 결과의 일반화, 즉 모델의 평가용 데이터셋에 대해 기대되는 성능을 정확하게 판단하기 위해 훈련 과정에서 10겹 교차검증을 이용했다. 10겹 교차검증은 훈련용 데이터셋을 10개의 하위 데이터셋으로 나눈 뒤 9개의 학습 데이터셋에서 학습된 모델을 나머지 하나의 평가 데이터셋을 이용해 평가하는 과정을 순차적으로 10번 반복하여 평가 결과의 평균값으로 모델의 성능을 비교해 가장 좋은 분류 결과를 보인 모델을 선택했다. 이때 모델의 성능 평가는 학습 데이터셋으로 도출된 모델이 평가 데이터셋의 댓글 중 악플로 맞게 분류한 건수의 비율인 정확도에 기초했다. [그림 1]에서 제시하는 LR 모델의 튜닝 과정에 따르면 패널티값이 0.0001에서 증가할수록 어휘 특성의 개수가 모델 성능에 미치는 영향이 미미해지고 약 0.01의 값이 되며 정확도가 급격히 떨어지고 있음을 알 수 있다. 결국, 3,000개의 어휘 특성과 0.00113의 패널티값으로 학습된 LR 모델이 훈련 데이터셋으로 도출된 최종 악플 분류 모델로 선정됐다.

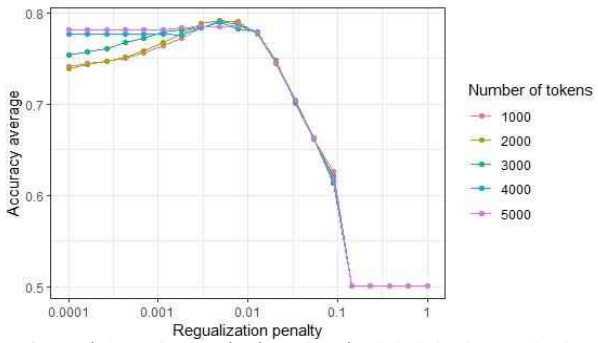


그림 1. 어휘 특성 개수와 라쏘 정규화 패널티에 따른 모델 성능
Fig. 1. Model performance across tokens and lasso regularization penalties

3-5 악플 분류 모델의 성능 시험

훈련용 데이터셋을 이용한 학습과 튜닝 과정으로 도출된 LR 모델이 악플 분류에 어떠한 성능을 보이는지를 시험용 데이터셋을 이용해 다음과 같이 평가했다. LR 모델이 시험용 데이터셋의 댓글(1,426건)에 대해 악플로 분류한 결과를 코더가 실제 악플로 분류한 결과와 비교해 봄으로써 모델의 성능을 평가하기 위한 지표인 정확도(accuracy), 정밀도(precision), 재현율(recall), 그리고 F1 점수(F1 score)를 계산했다 (표 1 참조). 정확도는 시험용 데이터셋의 전체 댓글 건수 중 LR 모델이 분류한 악플이 코더가 분류한 악플과 일치한 비율을 나타낸다.

하지만 정확도 지표만으로는 타당한 모델 성능 평가가 이루어지지 않을 수 있다. 가령, 댓글에서 악플이 출현하는 비율이 매우 낮을 경우라면 모델이 어떠한 댓글도 악플로 분류하지 않더라도 높은 정확도 지표를 보일 수 있기 때문이다. 이렇게 불균형한 데이터가 갖는 문제를 보완하기 위한 지표로 LR 모델이 분류한 악플 중 코더가 분류한 악플이 차지하는 비율인 정밀도와 코더가 분류한 악플 중 LR 모델이 분류한 악플이 차지하는 비율인 재현율, 그리고 이 두 지표의 조화평균인 F1 점수를 사용해 모델 성능을 평가했다. 그 결과, 모든 평가 지표가 약 78%로 나타나 소셜미디어에서의 혐오표현 혹은 악플을 기계학습으로 분류한 선행 모델의 정확도와 비슷한 성능을 보여 외적 타당성이 확보됐다고 판단했다[13]-[14].

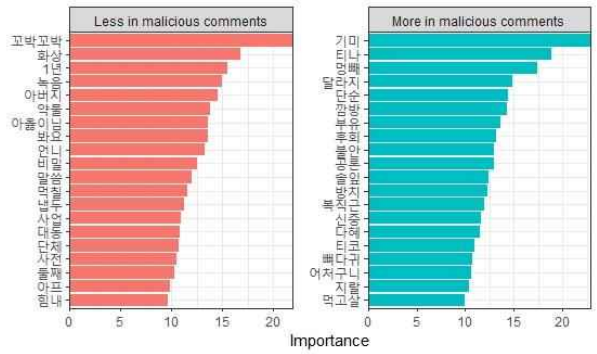
더욱이, [표 2]의 혼동행렬 역시 최종 LR 모델의 준수한 성능을 보여준다. 즉, 시험용 데이터셋의 댓글 중 약 78%의 댓글이 악플과 악플이 아닌 댓글로 옳게 분류됐을 뿐만 아니라 악플이 아닌데 악플로 분류된 댓글 사례 수(false positive: 159건)와 악플인데 악플로 분류되지 않은 댓글 사례 수(false negative: 156건)의 비율이 어느 한쪽으로 치우쳐지지 않음을 보여준다. 이는 악플 분류에 있어 편향이 발생할 가능성이 적다는 것으로 LR 모델이 좋은 성능을 보여주고 있음을 나타낸다.

표 1. 시험용 데이터셋에 대한 LR 분류기의 성능 지표
Table 1. Performance of our LR classifier on the test dataset

Algorithm	Accuracy	Precision	Recall	F1 score
Logistic Regression (Lasso regularized)	0.779	0.779	0.776	0.777

표 2. 시험용 데이터셋에 대한 악플 분류 혼동행렬
Table 2. Confusion matrix on the test dataset

		Truth	
		Non-malicious	Malicious
Prediction	Non-malicious	550	156
	Malicious	159	561



* Word features are displayed in Korean because it delivers their meanings and subtle nuances precisely.

그림 2. 유튜브의 ‘사이버렉카’에 달린 악플 분류에 중요한 어휘 특성

Fig. 2. The most important features in predicting whether a comment on YouTube clips from ‘cyber wreck car’ channels is malicious or not

Ⅳ. 악플 특성 분석 결과

LR 모델 기반의 악플 분류기를 통해 “사이버렉카” 채널의 댓글에 대한 성능 평가와 함께 분류 결정에 어휘적 특성이 어떠한 작용을 했는지를 평가했다. 이러한 결과를 바탕으로 한 논의에 앞서 본 연구는 악플 분석을 목적으로 하고 있어 욕설이나 비속어 등 불쾌감을 줄 수 있는 표현들이 언급되고 있음을 주의 사항으로 미리 밝힌다. 단, 특정인에 대한 멸칭과 비하의 표현이 포함된 악플을 언급하는 것은 그 자체로 모욕감을 재생산할 수 있어 연구 윤리의 차원에서 사회적 책임성을 고려해 결과에서 생략했다. 그리고 이러한 악플에 대한 연구의 결과로써 욕설과 비속어 등의 언급은 선행연구의 방식을 참고했다[6].

[그림 2]는 유튜브 댓글을 악플과 그렇지 않은 댓글로 분류함에 있어 중요한 영향을 미친 각각의 상위 20개의 어휘 특성을 보여준다. 우선, 악플로 분류된 댓글에서 두드러지게 관찰된 어휘적 특성들은 “기미”, “티나”, “멍빼”, “달라지”, “뼈다귀”와 같이 외모나 행동을 지적하기 위한 어휘들로 나타났다. 즉, 이러한 표현들은 “사이버렉카”로 파생된 댓글이 악플로 분류되는데 중요한 작용을 한 것으로 외모에 대한 지적과 비하가 악플의 주된 내용을 구성하고 있음을 보여준다. 반면, 악플로 분류되지 않는 댓글에서 나타난 주요한 어휘적 특성은 “아버지”, “아옳이님”, “언니”와 같은 호칭과 존칭이나 “봐요”나 “말씀”과 같이 존대의 표현이었다. 이것은 표현의 내용은 물론 형식이 악플의 주요한 특성으로 작용하고 있음을 의미한다. 흥미로운 점은 “지랄” 이외에는 욕설이 악플 분류의 두드러진 요인으로 등장하지 않았다는 것이다. 이는 특정 키워드에 기반한 분류보다는 내용의 맥락과 존댓말과 같은 언어의 표현 방식을 고려한 악플 탐지가 보다 효과적임을 의미한다.

또한, “사이버렉카”에 따른 악플을 본 연구의 LR 모델이 악플로 분류하지 못한 대표적 사례를 살펴봄으로써 유튜브 댓글의 특성과 이를 바탕으로 한 악플 분류기의 한계를 살펴보고자 한다. [표 3]은 LR 모델이 악플로 예측에 실패한 댓글 중 악플로의 분류 확률이 가장 낮았던 10개의 댓글을 제시한다. 즉, 인간 코더에 의한 판단 결과에 따라 악플로 분류되어야 했음에도 기계학습 분류기는 매우 낮은 확률로 악플 예측을 한 사례들을 정리했다. 이 댓글 사례들은 악플 분류의 정확도가 저하된 이유를 제시한다. 가령, “ㅈ(같)은”이나 “기래기”와 같이 비속어나 멸칭이 맞춤법에서 벗어난 형태로 사용되거나 “할자씨”와 “존라도” 같은 은어가 표현하는 악의성이 기계학습으로는 충분히 반영되지 못했음을 알 수 있다. 또한, 특정 어휘에 기반하지 않더라도 타인의 지능이나 외모, 행적, 출신지 등의 개인적 속성을 토대로 조롱하는 표현들에 대해 악의성 탐지가 민감하게 이뤄지지 않고 있음을 보여준다.

표 3. LR 분류기가 악플로의 예측에 실패한 악플 사례
Table 3. False negative cases of malicious comments, predicted by our LR classifier

	Prob.	Comment
1	12.37%	지능이 딸리면 결말은 항상 같은 곳이다
2	16.73%	또 좋은추억이 ㅈ(같)은추억으로
3	19.61%	기래기가 문제구만.. 잡아서 종신 독방형 때릴 수 있었는데 너무 편안하게 죽었군
4	23.56%	근데 진짜 너무못생겼다 와...
5	24.81%	팩트) 존라도에선 별일 아닌 일이다.
6	28.21%	좋게만 봤는데 아주 치졸한 할자씨구만
7	29.74%	괴인은 무슨.. 니 자신을 영상으로 만들어 보세요
8	30.15%	누가 보면 원숭이가 구걸하는줄
9	33.46%	옷입는게 유치원생 수준같냐? ㅋㅋㅋㅋㅋ
10	35.38%	부모가 일찍 잘 버렸네

* Comments are presented in Korean to deliver their meanings and subtle nuances precisely.

마지막으로, LR 모델이 악플이 아님에도 악플로 분류한 대표적 댓글 사례를 살펴보고자 한다. [표 4]의 사례들은 인간 코더는 악플로 판단하지 않았으나 기계학습 악플 분류기는 매우 높은 확률로 악플로 예측한 상위 10개의 댓글이다. 그러나 이 사례들은 인간 코더가 악플 탐지에 실패한 댓글들을 악플로 분류한 기계학습모델의 우수한 민감도를 보여준다. 가령, “멀쩡하게 생겨놓고 왜저런다니”나 “관상은 과학이야~~”와 같이 외모를 대상으로 한 비하와 조롱의 댓글은 인간 코딩의 오류로 악플로 분류되지 못했으나 기계학습의 결과로 악플 예측에 성공한 사례다. 또한 “성추행범”이나 “악인”, “인생 나락”, “즐라인” 등 타인을 경멸하거나 모욕하는 멸칭이 사용된 댓글이 악플로 예측된 것 역시 기계학습의 정교한 성능을 보여준다. 더 나아가, “삼청교육대가 존재했다면 젤 먼저보내야될 인간이다.”나 “가정교육과 피임의 중요성”와 같이 직접적인 비속어나 욕설, 혹은 멸칭이 사용되지 않았으나 표현의 맥락과 뉘앙스가 조롱과 비난의 악의적 의도를 포함한 댓글을 악플로 예측한 결과는 인공지능에 의한 악플 탐지와 차단의 경쟁력을 제시한다.

표 4. LR 분류기가 악플이 아님에도 악플로 예측한 댓글 사례
Table 4. False positive cases of malicious comments, predicted by our LR classifier

	Prob.	Comment
1	96.72%	멀쩡하게 생겨놓고 왜저런다니
2	96.00%	성추행범이 왜 유튜브와 방송으로 돈을 버는지 모르겠네.
3	94.42%	관상은 과학이야~~
4	93.71%	세상좋아졌지 이런 애들을 돈벌게 해주는 유튜브라는게 생기다니
5	92.41%	힘들게 사는애한테 악인이라는 낙인을 찍어버리는... 니가 제일 악인임 ㅇㅇ
6	92.33%	삼청교육대가 존재했다면 젤 먼저보내야될 인간이다.
7	90.72%	단순 개그맨이 꿈이였던 잘생긴 청년이 인생 나락으로 떨어진 사례
8	88.53%	자살하는 사람들은 징조가 있고 조용히 한다..저러게 하고 살고있는 사람들은 돈문제 꼭 엮여있지..
9	88.08%	가정교육과 피임의 중요성
10	86.80%	즐라인만 안만나도 인생반은 성공

* Comments are presented in Korean to deliver their meanings and subtle nuances precisely.

V. 결 론

유튜브를 통한 개인 미디어 콘텐츠의 홍수 속에서 사이버 폭력의 확산은 우리 사회의 신뢰 가능한 공동체 지속에 크나큰 위협이 되는 문제다. 특히, “사이버렉카”로 대표되는 유튜브 콘텐츠 채널은 익명의 크리에이터가 개인(유명인)에 대한 선정적이고 편향된 정보를 제공해 악플을 조장하고 선동할 수 있을 뿐만 아니라 이로 인한 인신공격과 혐오를 확산시킬 수 있다는 우려를 낳고 있다. 이에 본 연구는 특정 인물에 대한 “사이버렉카” 콘텐츠에 달린 댓글을 수집하고 악플 기준에 부합하는 댓글을 분류한 후, 이를 알고리즘으로 예측함으로써 기계학습의 방법을 이용한 악플 탐지와 차단의 가능성과 한계를 진단했다. 그리고 이러한 분석 결과를 바탕으로 갈수록 이용량과 영향력이 증가하고 있는 유튜브와 같은 개인 미디어

플랫폼에서 발생 가능한 악플의 유형과 특성에 대한 이해를 확장해 악플 탐지와 차단 시스템 개발에 기여하고자 했다.

이상의 목적을 바탕으로 본 연구의 결과가 제시하는 함의는 다음과 같다. 첫째, “사이버렉카” 콘텐츠로 인해 발생하는 악플의 유형은 외모 비하나 행위에 대한 조롱 등이 많이 발생하고 있다는 점이다. 또한 모욕적 멸칭은 악플의 중요 특성으로 나타난 반면 존칭은 악플이 아닌 댓글에서 두드러진 특성인 점을 고려할 때 댓글에서의 호칭이 사용되는 방식이 효과적인 악플 탐지를 위해 유의미한 작용을 할 수 있음을 알 수 있다. 둘째, “사이버렉카”는 외모와 행동과 관련한 조롱과 비하의 악플 뿐만 아니라 성별과 출신지는 물론 나이와 직업과 같은 개인의 속성에 대한 차별과 비하의 혐오 표현을 생산하고 있었다. 그러나 조롱하는 형태의 혐오 표현은 “존라도”, “즐라인”, “삼청교육대”와 같은 집단 명칭의 고유명사로도 나타났는데, 이는 그 명칭에 함축된 모욕적 의미와 상징이 반영된 결과로 알고리즘에 의한 악의성 탐지가 개체명의 인식을 넘어 이해 단계의 자연어처리로 극복해야 할 과제임이 드러났다. 셋째, 알고리즘 기반 악플 예측은 수많은 언어적 변이가 매우 빈번하게 일어나는 온라인 플랫폼의 특성이 쉽게 반영될 수 없다는 점이다. 이는 유튜버를 중심으로 한 개인 미디어 콘텐츠의 대중화와 더불어 특정인에 대한 인신공격성 모욕과 조롱 등의 악플이 표현되는 매우 다양한 방식과 변칙이 탐지되기 위한 데이터셋의 필요성을 역설한다. 그리고 여기에는 “사이버렉카”로 생산된 댓글이 유효한 자료로 활용될 수 있다고 판단된다.

끝으로, 본 연구의 한계는 다음과 같다. 첫 번째로 본 연구의 댓글은 모든 “사이버렉카” 채널에서 수집된 것이 아니라는 점이다. 악플 여부에 따라 분류된 댓글은 특정 개인(유명인)을 대상으로 한 “사이버렉카” 콘텐츠 제작에서 대표적인 두 채널에 국한된 것으로 이에 결과 해석의 일반화에는 각별한 주의가 요구된다. 두 번째로 데이터 라벨링 과정에서 인간 코더 간 신뢰도가 높지 않았다는 점이다. 이것은 여러 차례 반복된 연습과 코더 간 논의를 거친 결과였는데, 결국 댓글의 악의성 여부를 판단할 시 인간 코더의 주관성이 개입되어 일관된 기준 적용이 어려워졌음을 의미한다. 그리고 잘못 분류된 댓글 사례들이 제시하듯 일관되지 못한 악플 라벨링이 기계학습에 따른 악플 분류기의 성능에도 악영향을 미친 것으로 판단된다. 따라서 향후 연구에서는 라벨링의 신뢰성을 개선하기 위해 악플의 개념화와 다양한 사례에 기반한 조작적 정의를 보다 정교화해야 할 필요성을 제언한다.

감사의 글

본 논문은 2020년 대한민국 교육부와 한국연구재단의 인문사회분야 신진연구자지원사업의 지원을 받아 수행된 연구로서(NRF-2020S1A5A8046843), 관계부처에 감사드립니다.

참고문헌

[1] H. Jenkins, *Convergence Culture: Where Old and New Media collide,* New York, NY: New York University Press, 2006.

[2] S. Kim, A survey on individual media contents creator in Korea, Korea Creative Content Agency, 2021.

[3] J. Hong and E.-K. Na, “Online Hate Speech Diffusion Network Analysis : Issue-Specific Diffusion Patterns, Types and Intensity of Verbal Expression on Online Hatred,” *Korean Journal of Journalism and Communication Studies*, Vol. 60, No. 5, pp. 145-175, October 2016. https://doi.org/10.20879/kjjcs.2016.60.5.006

[4] K.-H. Kim, Y.-H. Cho, and J.-A. Bae. “Exploratory Study on Countering Internet Hate Speech: Focusing on Case Study of Exposure to Internet Hate Speech and Experts’ in-depth Interview,” *The Journal of the Korea Contents Association*, Vol. 20, No. 2, pp. 499-510, February 2020. https://doi.org/10.5392/JKCA.2020.20.02.499

[5] B. M. Jung, P. G. Nam, and H. Choi, A survey on cyber-bullying, National Information Society Agency, 2020.

[6] S. H. Lee. “Biased Artificial Intelligence: Analyzing the Types of Hate Speech Classified by ‘Cleanbot’, NAVER AI for Detecting Malicious Comments,” *Journal of Cybercommunication Academic Society*, Vol. 38, No. 4, pp. 33-75, December 2021. https://doi.org/10.36494/JCAS.2021.12.38.4.33

[7] K.-H. Cho. “Text Typological Study of Internet ‘Comments’,” *Text Linguistics*, Vol. 23, No. 23, pp. 203-230, December 2007. http://doi.org/10.22832/txtlng.2007.23.23.009

[8] T. H. An. “A Study on the Scope and Types of Vicious Replies on the Internet Bulletin Board,” *Urimal*, Vol. 32, pp. 109-131, April 2013.

[9] J. Suler. “The Online Disinhibition Effect,” *CyberPsychology & Behavior*, Vol. 7, No. 3, pp. 321-326, July 2004. https://doi.org/10.1089/1094931041291295

[10] S. Kim, “Misogynistic Cyber Hate Speech in Korea,” *Issues in Feminism*, Vol. 15, No. 2, pp. 279-317, October 2015.

[11] M. Kim and J. R. Lee, “The Publicness of the Internet and Solutions for Trolling,” *Journal of Political Communication,* No. 9, pp. 5-50, June 2008. http://doi.org/10.35731/kpca.2008..9.001

[12] S. S. Hong, A study on the actual conditions of hate speech and the regulatory measures, National Human Rights Commission of Korea, 2016.

[13] C. J. van Rijsbergen, *Information Retrieval,* London, UK: Butterworths, 1979.
[14] B. Vidgen and T. Yasseri, “Detecting Weak and Strong Islamophobic Hate Speech on Social Media,” *Journal of Information Technology & Politics*, Vol. 17, No. 1, pp. 66-78, January 2020. https://doi.org/10.1080/19331681.2019.1702607

이신행(Shin Haeng Lee)

2008년 : 고려대학교 사회학과 (학사)
2010년 : Indiana University, School of Journalism (MA)
2016년 : University of Washington, Dept. of Communication (Ph.D)

2020년～현 재: 중앙대학교 사회과학대학 미디어커뮤니케이션학부 조교수
※관심분야 : 미디어 빅데이터(Media Big Data), 알고리즘 편향(Algorithmic Bias), 텍스트 마이닝(Text Mining), 혐오 댓글(Malicious Comments) 등

이주연(Ju Yeon Lee)

2022년 : 중앙대학교 미디어커뮤니케이션학과 (문학석사)

※관심분야 : 빅데이터(Big Data), 데이터분석(Data Analytics), 기계학습(Machine Learning)

조민정(Min-Jeong Jo)

2022년 : 중앙대학교 미디어커뮤니케이션학과 (문학석사)

※관심분야 : 데이터 마이닝(Data Mining), 빅데이터(Big Data), 기계학습(Machine Learning)

박태강(Tae-Kang Park)

2022년 : 한양대학교 정책학과 (학사)

2021년～현 재: 네덱스 대표
※관심분야 : 전자민주주의(Electronic Democracy), 디지털 리터러시(Digital Literacy), 자연어처리(Natural Language Processing), 블록체인(Blockchain)