# 머신러닝을 이용한 유튜브 악성 댓글 탐지 시스템

김나경*, 김정민*, 이혜원*, 국중진*
*상명대학교 정보보안공학과
kng8861@naver.com, 8065789@naver.com, hyenim0304@naver.com, kook@smu.ac.kr

# YouTube Malicious Comment Detection System

Na-Gyeong Kim*, Jeong-Min Kim*, Hye-Won Lee*, Joong-Jin Kook*
*Dept. of Information security engineering, Sang-Myung University

요 약

악성 댓글은 언어폭력이며 사이버 범죄의 일종으로 인터넷상에서 상대방이 올린 글에 비방이나 험담을 하는 악의적인 댓글을 말한다. 악성 댓글을 단순히 차단하는 다른 프로그램들과는 달리 해당 영상의 악성 댓글의 비율을 알려주고 악플러들의 닉네임과 그 빈도를 나타내주는 것으로 차별화를 두었다. 따라서 많은 유튜버들이 겪는 악성 댓글 문제들을 탐지하여 유튜브에 달리는 악성 댓글들을 탐지하고 시각화하여 제공한다.

## 1. 서론

인터넷의 익명성을 악의적으로 이용하여 명예훼손, 인신공격, 사생활 침해 등과 사회적으로 심각한 문제들이 계속해서 양산되고 있다.[1] '연예 뉴스 댓글 서비스'를 폐지하면서 악성 댓글의 수가 줄어드는 듯했으나, SNS와 유튜브로 사람들의 활동 반경이 넓어짐에 따라 인스타그램 DM, 유튜브 댓글 등 새로운 범위에서 악성 댓글을 사용한 범죄가 증가하고 있다.

본 연구에서는 악성 댓글이 온라인상에서 발생하는 불법적인 언사와 행위에 대해 가장 대두되는 문제라고 인식하여,[2][3] 많은 사람들이 고통받는 악성 댓글을 자동 탐지하여 피해를 줄이고자 이에 대한 필요성을 느끼고 프로젝트를 기획하게 되었다.

우리는 그중 유튜버들의 악성 댓글 문제의 주목하였다. 유튜브는 1인당 1달에 30시간 34분, 하루에 59분 이상 이용한다. 유튜브 앱 사용자 4천41만 명 중 10대가 13.4%, 20대가 17.2%, 30대가 19.4%, 40대가 21.3%, 50대 이상이 28.7%로[4] 다양한 의견과 댓글들이 달리는 경우가 많아 악성 댓글에 취약할 것이라고 예상되어 대상 플랫폼을 유튜브로 선택하였다.

악성 댓글은 내용에 따라 상대에 대한 비방, 협박 및 저주, 게시판 도배 및 광고, 기타 등으로 나눌 수 있고, 피해 대상에 따라 게시글에 나타난 대상, 게시글 작성자, 댓글 작성자, 기타 등으로 나눌 수 있다. 또한, 악성 댓글의 비율은 59%로 일반 댓글보다 악성 댓글의 비율이 더 높게 인터넷상에 작성되고 있다. [5][6]

<표 1> 사이버 명예훼손 및 모욕 고소·고발 건수[7]

| | |
|---|---|
| 2009 | 4,752 |
| 2012 | 5,684 |
| 2015 | 15,043 |
| 2019 | 16,633 |

## 2. 본론

1. 연구 과정

1.1 댓글 크롤링

여러 유형의 댓글이 필요하다 판단하여 다양한 어휘들을 최대한 수집하고 학습시키기 위해 먹방/게임/뷰티 등 카테고리를 세분화하여 동영상을 선정하여 댓글을 추출하였다. 또한, 동영상의 일반 댓글과 악성 댓글의 분포가 거의 같은 동영상뿐만 아니라, 분포가 2배 차이나는 동영상에서도 댓글을 추출하였다. 댓글 추출 방식은 Google API console에 접속하여 API 프로젝트를 생성하고 YouTube Data API v3를 활성화해 api key를 발급받아 파이썬에 수집한 댓글을 엑셀의 형태로 저장하기 위한 라이브러리인 pandas 라이

브러리를 사용하여 유튜브 영상의 댓글뿐만 아니라 작성자 닉네임을 포함하는 코드를 작성하여 엑셀 파일로 댓글을 수집하였다. [8]

1.2 데이터 정제

(1) 댓글 라벨링

수집된 댓글들은 학습을 위하여 임의로 일반 댓글과 악성 댓글로 분류하였다. 일반 댓글은 1, 악성 댓글은 0으로 라벨링 하는 작업을 수행하였다. 악성 댓글은 자기와 생각이 맞지 않거나 싫어하는 사람에게 단순히 재미로 욕설을 하는 댓글, 상대방의 약점을 들춰내고 헐뜯는 댓글, 같은 내용의 욕설이나 의미없는 글들을 연속해서 게시한 댓글, 성에 대한 노골적인 욕설을 하여 상대방에게 불쾌감과 수치심을 주는 댓글, 사실이 아닌 거짓 소문을 퍼트려 상대방에게 피해를 입히는 댓글 등을 기준으로 삼았다. 욕설이 포함되어 있지 않더라도 비방적인 댓글, 비꼬는 댓글 등 상대방이 불쾌감을 느낄 언사는 악성 댓글로 분류하였다. 라벨링을 마친 댓글 데이터들은 test 데이터와 train 데이터에 넣어 학습에 사용하였다. 본 연구에서 학습에 사용한 총 데이터는 51447개이다.

(2) 결측값 제거

(1)의 과정으로 분류된 데이터들의 956개의 중복값과 181개의 결측값을 확인하고 제거 작업을 마친 후, 일반 댓글과 악성 댓글의 비율을 맞추어 적절히 조절하였다. 댓글 중에 Null 값을 가진 샘플이 있는지 확인하여 Null 값 샘플을 제거한다.

(3) 정규 표현식 수행

수집한 데이터들에서 온점(.)이나 (물음표)?와 같은 각종 특수문자와 영어가 사용된 문장이 전체 댓글 수의 약 94%로 다수 존재하였다. 이러한 문장들은 정규 표현식을 사용하여 한글과 공백을 제외한 특수문자들을 모두 제거해 준다. 우선 자음과 모음에 대한 범위를 자음의 범위는 ㄱ ~ ㅎ, 모음의 범위는 ㅏ ~ ㅣ와 같이 지정할 수 있다. 또한, 완성형 한글의 범위는 가 ~ 힣과 같이 사용한다. 따라서 정규 표현식의 형태는 [^ㄱ-ㅎㅏ-ㅣ가-힣 ]가 된다. 위의 범위 지정을 모두 반영하여 데이터에 한글과 공백을 제외하고 모두 제거하는 정규 표현식을 수행한다.

(4) 토큰화

(3)의 과정 이후 불용어를 지정하여 KoNLPy의 Okt 토큰화를 이용해 제거해 준다. 불용어란 분석에 큰 의미가 없는 단어를 지칭하며 '의', '가', '으로', '를' 등과 같은 단어를 포함한다.

불용어는 한국어의 조사, 접속사인 ['의','가','이','은','들','는','좀','잘','강','과','도','를','으로','자','에','와','한','하다']를 정의해 댓글 데이터에서 이 단어들을 제거함으로써 형태소 분석의 효율을 높여주었다.

(5) 정수 인코딩

기계가 텍스트로 된 데이터를 숫자로 인식하여 처리할 수 있도록 test 데이터와 train 데이터에 정수 인코딩을 진행하였다. train 데이터에 대하여 단어 집합을 생성해 준 후 각 단어마다 고유한 정수를 부여해 준다. 만약 등장 회수가 2회 이하라면 train 데이터에서 차지하는 비중은 상대적으로 매우 적으므로 정수 인코딩 과정에서 제외시킨다.

(6) 패딩

(5) 과정 이후 길이가 기계가 전부 동일한 문서들에 대해서는 하나의 행렬로 보고, 한꺼번에 묶어서 처리할 수 있기 때문에 패딩 작업을 통해 데이터 샘플들의 길이를 일정하게 맞춰주었다. 약 97% 이상의 댓글 길이가 50이하인 것을 확인하였고, 이에 맞추어 댓글 길이를 50으로 맞추어 진행하였다.

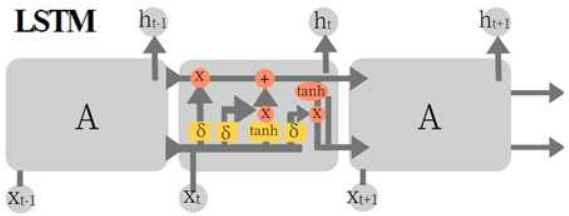
1.3 댓글 감성 분류

(1) LSTM

LSTM은 이전 스텝의 출력값이 다시 입력으로 연결되는 신경망인 RNN에서 잘 해결할 수 없는 장기 메모리가 필요한 경우에 문제를 해결하기 위해 개발된 신경망이다. LSTM은 기존 RNN에 cell state를 추가하여 먼 과거의 데이터를 얼마나 반영할 것인지 제어하며, 장기간 메모리를 수행하는 cell state와 연결의 강도를 조절하는 3개의 게이트로 구성된다. cell state에서 게이트를 조정하여 이전 state 정보가 현재 state로 끼치는 영향 조절한다.[9]

본 연구에서 LSTM을 사용하는 이유는 다른 방법인 RNN을 사용하면  관련 정보와 그 정보를 사용하는 지점 사이 거리가 멀 경우 역전파시 그래디언트가 점차 줄어 학습능력이 크게 저하되는 vanishing gradient problem이 존재하기 때문이다.[10] 임베딩 벡터의 차원을 100으로 정하고, 검증 데이터의 손실이 증가하면 과적합 징후로 인지하여 징후가 4번 발

생 시 조기 종료되도록 한다. 또한, 검증 데이터의 정확도(val_acc)가 이전보다 좋아질 경우에만 모델이 저장되도록 하였다. 에포크 15번 수행으로 훈련 데이터의 20%를 검증 데이터로 사용하여 정확도를 확인하려 하였으나, 조기 종료의 조건에 따라 10번의 실행에서 멈추어 약 92%의 정확도를 보였다. 훈련이 끝나면 test 데이터를 사용하여 학습 모델을 검증한다.




(그림 1) LSTM의 구조

1.4 시각화


(그림 2) GUI 실행화면


(그림 3) 크롤링 된 영상의 악성/일반 댓글 비율 그래프


(그림 4) 크롤링 된 영상의 악성 댓글 다수 작성자 수치 그래프

(1) PYQT5

PYQT란 QT의 레이아웃에 파이썬의 코드를 연결하여 GUI 프로그램을 만들 수 있게 해주는 프레임워크를 뜻한다.

완성된 GUI에서는 자신이 올린 영상의 주소를 넣어 [크롤링 시작] 버튼을 누르면 해당 영상의 댓글이 자동으로 크롤링 되고, 지정 폴더에 영상의 댓글과 작성자 등의 정보가 담긴 엑셀파일이 자동 저장된다. [악성 댓글 / 일반 댓글 비율 확인] 버튼을 누르면 해당 영상의 일반 댓글과 악성 댓글의 전체적인 비율을 나타내주는 그래프가 시각화되어 제공된다. 또한 [악성 댓글 작성자 확인] 버튼을 누르면 악성 댓글러들의 닉네임과 악성 댓글을 단 빈도를 나타낸 그래프를 출력해준다.

1.5 실험 결과

약 5만 개의 댓글로 학습된 데이터를 통해 악성 댓글 및 일반 댓글의 내용을 검색해보았을 때, 약 92%의 정확도를 보였다. 이 정확도를 근거로 하여 해당 유튜브의 주소를 넣어 댓글을 크롤링하고 악성 댓글 비율 및 악플러들을 보았을 때, 그에 맞는 결과가 잘 도출된 것을 확인할 수 있었다.

<표 2> 수집한 데이터 양에 따른 정확도

| 수집한 데이터의 양 | 정확도(약) |
|---|---|
| 10,000 | 89% |
| 20,000 | 94% |
| 50,000 | 92% |

## 3. 결론

인터넷에서 가장 큰 범죄인 악성 댓글은 익명성을 이용하여 모욕적인 어휘를 사용하며 상대방을 비방하고 공격한다. 이러한 행동에 대해 반성이나 잘못의 인식이 요구되고 있지만 해를 거듭할수록 악성 댓글로 인한 피해는 오히려 기하급수적으로 증가하고 있다. 이에 본 논문에서는 이러한 사이버상의 범죄를 예방하기 위해 악성 댓글 탐지시스템을 제시하였다.

최근 댓글들은 사람들의 인위적인 조작으로 특수문자를 이용한 한글을 사용하는 점이 많기 때문에 한글 정규화 과정을 통해 특수문자들을 모두 제거해주었으며 본 연구는 기계학습을 시키기 위해 필요한 데이터를 정제하는 과정에서 댓글에 특화된 감성사전을 구축하여 이를 감성분석에 이용할 수 있음을 확인했다. 또한 기존의 연구 방식에 비해 시각화를

통해 사용자의 사용면에서 고려했다는 점이 악성 댓글 탐지에 대해 효과가 더욱 높을 것이며 이를 좀 더 발전시킨다면 악성 댓글 방지 및 사용자가 사용할 수 있는 서비스가 될 수 있을 것이라고 생각된다.

"유튜브 악성 댓글 탐지시스템"은 단순히 악성 댓글 탐지가 아닌, 유튜버가 직접 악성 댓글의 빈도와 악플러들의 닉네임을 한눈에 파악할 수 있어 다양한 조치를 취할 수 있도록 도와준다. 본 연구는 악성 댓글을 방지하기 위한 직접적인 해결책은 제시할 수 없지만 근본적인 대책을 제시함으로써 기대효과를 높이고자 하였다.

악성 댓글 탐지 시스템의 기대효과는 자율적인 악성 댓글 방지와 개선이다. 시스템에서 제공하는 그래프를 통하여 유튜버 당사자는 댓글 관리에 적극적으로 이용이 가능하다. 시각화된 자료를 직접 공개하여 사람들이 경각심을 가져 악성 댓글의 비율을 줄이도록 유도하거나 또한 그 정도가 심하다고 판단되면 따로 저장된 댓글 파일과 함께 법적 증거로 활용이 가능하다. 이는 다른 말로 악플러들의 차단과 신고가 수월해진다는 점이다. 한눈에 파악이 가능하기 때문에 사용자가 어떻게 활용하느냐에 따라 많은 방향으로 사용될 것으로 판단된다. 본 연구가 건강한 댓글 문화를 형성하는데 도움이 될 수 있음을 기대한다.

이 논문은 2016년도 정부(교육부)/한국연구재단의 산업연계교육활성화선도대학사업(PRIME사업)의 사후관리 프로그램 지원을 받아 수행된 연구임.

## 참고문헌

[1] 홍진주, "인터넷 악성댓글 탐지 기법 : A Malicious Comments Detection Technique on the Internet," *숭실대학교 소프트웨어특성화대학원*, 11쪽, 2015년 12월

[2] 박현주, "소리없는 흉기 '악플' 공세…연예인 일반인 안가린다", 중앙일보, 2020.11.03
채지선&강보인, ""악성 댓글 참다가 죽을 것 같아 고소" 악플과 전쟁 분투기", 한국일보, 2020.04.04
소중한, "유튜브 잠깐 출연했다 봉변... 일상 덮친 이름 모를 '악플'", 오마이뉴스(스타), 2020.10.27

[3] 이동우, "어느 유튜버의 죽음…유튜브 댓글이 위험하다", 머니투데이, 2021.02.05

[4] 백봉삼, "유튜브 가장 많이 보는 세대는 "50대 이상", ZDNet Korea, 2021.02.23

[5] 안태형, "악성 댓글의 범위와 유형",KCI, vol(2013)., no.32, pp. 109-131 (23 pages)

[6] 정관철, "[기획] 악성댓글, 이대로 괜찮은가", 한국리서치 여론 속의 여론, 2019.11.29

[7] 박주희, "상습 악플러 이력 공개… 댓글수익 챙긴 포털에도 책임 물어야", 2020.04.03

[8] CHML, 2021. 1. 17,https://untitledtblog.tistory.com/16

[9] [Part Ⅶ. Semantic Segmentation] 7. RNN, LSTM, GRU(2017),
https://m.blog.naver.com/PostView.nhn?blogId=laonple&logNo=221027194402&proxyReferer=https:%2F%2Fwww.google.com%2F,(2021.07.04.)

[10] Gichang Lee, 2017.03.09, https://ratsgo.github.io/