자연어를 이용한 유해 영상 탐지

이정훈*
*경기대학교 응용통계학과
e-mail:vhrehfdl7tp@naver.com

Inappropriate Video Detect Using Natural Language Process

Jung-Hoon Lee*
*Dept. of Application Statistics, Kyonggi University

요 약

최근 청소년들은 욕설, 폭력적, 선정적, 비하적 표현을 일상생활에서 자연스럽게 사용하고 있다. 현재 청소년들은 자극적이고 폭력적인 개인 방송을 시청하며 유해 표현을 학습한다. 그래서 여러 기업에서는 모니터링 요원을 배치하거나 사용자들의 신고를 통해 유해 영상을 제재하는 중이다. 하지만 방대한 규모의 동영상 때문에 사람이 직접 모든 영상을 확인하는 것은 물리적으로 불가능하다. 따라서 본 논문에서는 자연어 처리 기술을 활용하여 자동으로 유해 영상을 탐지하는 시스템을 제안하고자 한다. 본 시스템은 데이터 수집, 텍스트 변환, 형태소 분석, 유해 사전 구성, 유해 판단 5가지 과정으로 이루어진다.

1. 서론

최근 청소년들은 일상생활에서 비속어, 욕설, 폭력적, 선정적, 비하적 표현을 거리낌 없이 사용하고 있다.[1] 이러한 유해 표현은 청소년 심리건강에 부정적 영향을 끼친다. 구체적으로 비속어를 자주 사용하는 그룹은 비속어를 사용하지 않는 그룹보다 어휘력이 떨어지고 공격성과 충동성이 증가한다.[2]

청소년들은 주위 친구들과 대중매체를 통해 유해 표현을 학습한다.[3] 왜냐하면 청소년들은 자신이 좋아하거나 동경하는 대상을 모방하기 때문이다.[4-9]

청소년들은 다양한 대중 매체 중 특히 인터넷 개인 방송을 선호한다. 개인 방송은 다양한 콘텐츠를 활용하고 시청자와 쌍방향 소통을 통해 청소년들에게 큰 인기를 얻고 있다.[10] 하지만 일부 개인 방송에서는 시청자 수를 늘리기 위해 경쟁적으로 자극적인 콘텐츠를 방송하고 있다. 이로 인해 청소년들은 자극적인 콘텐츠에 담긴 유해 표현과 행동을 학습하고 자연스럽게 사용한다.[11-13]

현재 아프리카 TV에서는 모니터링 요원을 배치해 유해 영상을 제재하고 유튜브에서는 사용자들의 신고를 받아 관리자가 영상을 제재하는 방법을 사용하고 있다. 하지만 사람이 수작업으로 유해 영상을 탐지하는 것은 물리적인 한계가 존재한다. 왜냐하면, 아프리카 TV에서는 수백 명의 BJ가 동시에 방송하고 Youtube에서는 1분당 400시간 이상의 영상이 업로드되기 때문이다. 따라서 모든 영상을 사람이 일일이 확인하고 탐지하는 것은 현실적으로 불가능하다.

선정적 영상이나 이미지를 찾는 음란물 필터링 시스템은 이미 개발되고 현실에서 적극적으로 활용되고 있다. 하지만 욕설, 폭력적, 선정적, 비하적 표현과 같이 유해 표현을 찾는 시스템은 현실에서 활용되지 않고 있다.

따라서 본 논문에서는 자연어 처리 기술을 이용하여 자동으로 유해 표현이 포함된 영상을 탐지하는 시스템을 제안하고자 한다. 이를 통해 영상을 탐지하는 시간과 비용을 줄이고 유해 영상을 제재하여 청소년들에게 건전한 영상을 제공하려 한다.

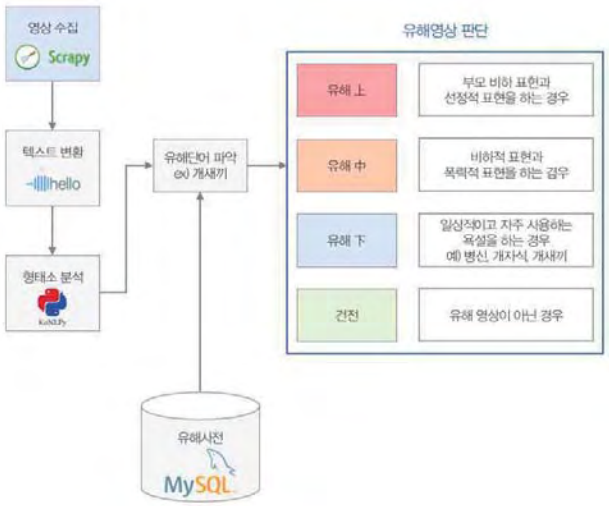
2. 관련 연구

국내에서 유해 표현이 포함된 영상을 탐지하는 연구는 진행된 적이 없다. 그렇지만 비속어와 욕설이 포함된 텍스트를 탐지하는 시스템은 과거부터 활용됐다.[14] 비속어 필터링 시스템은 채팅, 검색엔진, 댓글 등 다양한 곳에서 사용되는 욕설들을 필터링하기 위해 사용되었다.

비속어 필터링 시스템은 크게 사전기반 방식과 기계학습기반 방식으로 나누어진다. 사전기반 방식은 욕설 사전을 구성하고 해당 단어가 욕설 사전과 매칭되면 필터링하는 시스템이다. 사전 기반 방식은 욕설 사전이 얼마나 정교하고 방대하게 구성되어 있는지가 성능에 큰 영향을 미친다.

기계학습기반 방식은 유해 문장과 단어들을 학습 데이터로 구성하고 학습시켜 검출 모델을 만든다. 기계학습기반 방식은 알고리즘과 학습데이터가 성능에 큰 영향을 미친다.[15-16]

3. 유해 영상 검출 모형



(그림 1) 유해 영상 검출 모형 흐름도

3.1 데이터 수집

데이터 수집은 Scrapy를 사용해서 영상의 URL, 제목, 썸네일 이미지를 수집한다. 그리고 수집한 영상의 URL을 Pytube에 입력하여 고화질 mp4 파일을 다운받는다.

텍스트 변환을 하기 위해서는 영상 파일이 아닌 음성 파일이 필요하다. 따라서 음성 파일로 변환하기 위해 FFMPEG를 사용해 mp4 파일을 wav 파일로 변환한다.

3.2 텍스트 변환

영상에서 사용되는 유해 표현을 검출하기 위해 영상 속 음성을 텍스트로 변환하는 Speech To Text API를 사용한다. 본 논문에서는 1분에서 180분 이내의 한국어 영상을 대상으로 유해 영상 탐지를 진행하였다. 따라서 STT API 중 한국어를 지원하고 1분 이상의 음성 파일을 텍스트로 변환해주는 Google STT API를 사용한다.

3.3 형태소 분석

텍스트 변환이 완료되면 형태소 분석을 한다. 형태소 분석은 어미와 조사가 붙어 변형된 단어의 원형을 복원하기 위해 사용한다. 예를 들어 '뒤지다'라는 유해 단어는 어미와 조사들이 붙어 '뒤져', '뒤져버려라', '뒤진다' 같은 다양한 형태로 변형된다. 단어의 원형을 복원하지 않으면 유해 사전 규모가 불필요하게 방대해지므로 형태소 분석기를 사용해 원형복원을 한다.

형태소 분석기는 Komoran 형태소 분석기를 사용한다. 왜냐하면 Komoran 형태소 분석기는 다른 형태소 분석기보다 원형복원 능력이 뛰어나기 때문이다. 그리고 본 논문의 수집 대상인 유튜버는 문법적으로 올바르지 않은 문장을 자주 사용한다. Komoran 형태소 분석기는 문법적으로 올바르지 않은 문장에 대해서도 비교적 정확한 분석을 해주기 때문에 선택하였다.

3.4 유해 사전 구성

유해 사전은 유해 표현을 찾는 데 필요하며 검출 정확도에 영향을 미치는 핵심 요소이다. 유해 사전은 유해 단어, 유해 카테고리, 세부 카테고리로 구성한다. 유해 카테고리는 크게 욕설, 폭력적 표현, 선정적 표현, 비하적 표현으로 구분했다. 그리고 각 세부 카테고리는 <표1>과 같이 분류했다.

<표 1> 유해 카테고리

카테고리	욕설		
세부 카테고리	가족 욕설	동물에 비유	일상적 욕설
카테고리	폭력적 표현		
세부 카테고리	구타	살인	죽음
카테고리	선정적 표현		
세부 카테고리	성기 언급	매춘 언급	성적 표현
카테고리	비하적 표현		
세부 카테고리	종교 비하	여성 비하	지역 비하
	정치인 비하	동성애 비하	외국인 비하
	특정인 비하	장애인 비하	노인 비하

유해 사전은 <표2>와 같이 구성되어 있으며 유해 단어는 총 1228개로 구성되어 있다.[17] 유해 사전은 기존에 비속어 필터링 시스템에서 만들었던 욕설 DB, 다른 논문에서 언급한 욕설, 최근 청소년들이 자주 사용하는 신종 욕설 등을 수집하여 구성했다.[18]

<표 2> 유해 사전

유해 단어	카테고리	세부 카테고리
새끼	욕설	일상적 욕설
니애미	욕설	가족 욕설
패버린다	폭력적 표현	구타
좆같다	선정적 표현	성기 언급
씨발	선정적 표현	매춘 언급
홍어놈	비하적 표현	지역 비하
틀딱충	비하적 표현	노인 비하
된장녀	비하적 표현	여성 비하
짱깨	비하적 표현	외국인 비하

3.5 유해 판단

유해 판단은 2가지 검출 모델을 사용한다. 1차 검출 모델은 영상의 음성을 텍스트로 변환하여 유해 표현을 탐지하는 모델이다. 2차 검출 모델은 영상의 제목을 대상으로 유해 표현을 탐지하는 모델이다.

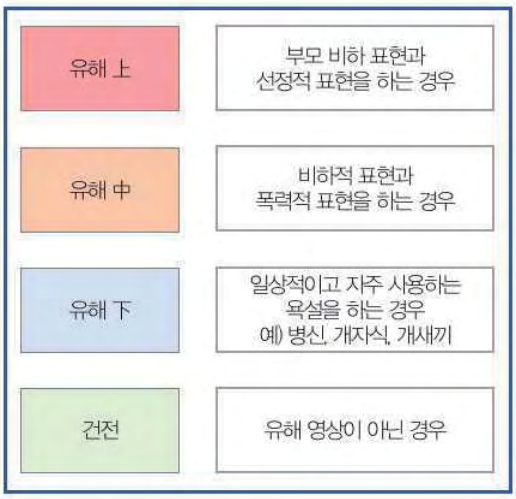
(그림 2) 유해 판단 모델

2가지 모델을 사용해 유해 영상을 검출한 후 영상의 유해 등급을 분류한다. 유해 등급 기준은 그 당시 사회 분위기와 유해 등급 심사위원의 주관에 따라 결정되며 절대적인 기준은 존재하지 않는다. 따라서 방송통신위원회와 영상물등급위원회의 기준을 바탕으로 (그림 3)과 같이 유해 등급 기준을 정의했다.

유해 등급은 유해 上, 유해 中, 유해 下, 건전 4가지로 분류한다. 유해 上은 영상에서 선정적 표현과 가족 욕설을 한 번이라도 한 경우에 해당한다. 유교문화권인 한국에서는 부모를 욕하는 표현과 성적인 표현에 대해 비교적 강한 분노와 거부감을 느낀다.[19-20]

유해 中은 비하적 표현과 폭력적 표현을 한 번이라도 한 경우에 해당한다. 유해 下는 일상적 욕설을 한 번이라도 한 경우에 해당한다. 일상적 욕설은 공영 방송에서도 사용될 만큼 일상생활에서 자주 사용되기 때문에 비교적 유해도가 낮다고 판단했다.

건전은 유해 上, 유해 中, 유해 下에 포함되지 않는 경우 건전으로 분류하였다.



(그림 3) 유해 등급 기준

4. 실험 및 결과

4.1 데이터 수집

영상은 10명의 유튜버를 선정하여 유튜버 채널에 업로드된 영상을 수집했다. 5명은 평소 방송에서 유해 표현을 자주 사용해 사회적으로 논란이 된 유튜버를 선정했다. 나머지 5명은 게임방송 유튜버 중 구독자 수가 가장 많은 5명을 선정했다. 유튜버 한 명당 30개 영상을 수집했고 총 300개의 영상을 수집했다.

4.2 텍스트 변환

Google STT API를 사용해 영상의 음성을 텍스트로 변환하면 <표3>과 같은 결과가 나온다. <표3>을 보면 현재 Speech To Text 기술로는 영상 속 음성을 완벽하게 텍스트로 변환하지 못하는 것을 알 수 있다. 하지만 자주 사용되는 유해 표현은 텍스트로 변환되기 때문에 유해 판단은 가능하다.

<표 3> 음성 텍스트 변환 예시

실제 음성 파일	1. 뭐 어 그거 계속 하라고요 2. 아 그거 듣지 말고 그냥 돌아와 듣지 말고 돌아와 3. 이제 돌아와 개 쓰레기 새끼야 4. 그리고 마우스에서 손 때 구경만 하고 있어
텍스트 변환 파일	1. 어 어 그거 계속 하라고요 2. 아 그거 말고 그냥 보라고 돌아와 3. 이제 또라이 개 쓰레기 새끼야 4. 발 마사지 존댓말 하고 있어

4.3 유해 판단

텍스트 변환과 형태소 분석을 완료하면 유해 사전과 일치하는 단어가 존재하는지 확인한다. 그다음 유해 등급을 분류하고 영상에서 사용된 유해표현을 보여준다.

(그림 4) 유해 영상 화면

4.5 결과

유해 표현 탐지 시스템의 정확도 측정을 위해 검증 데이터를 직접 구성하여 테스트했다. 검증 데이터는 유해 영상 15개와 건전영상 15개 총 30개를 활용하였다. 검증 데이터를 유해 표현 탐지 시스템에 적용한 결과 15개의 유해 영상 중 9개를 유해 영상으로 탐지하였다. 해당 시스템의 정분류율은 86%이고 오분류율은 14%이다.

<표 4> 오분류표

예측 \ 실제	T	F
T	11	0
F	4	15

5. 결론

본 논문에서는 자연어를 이용해 유해 영상을 탐지하는 시스템을 제안하였다. 본 시스템 성능이 개선되고 실제로 활용된다면 유해 영상을 찾는 시간과 비용이 크게 절감되고 많은 유해 영상을 제재할 수 있을 것이다.

본 연구는 유해 표현을 정확히 탐지하지 못 한다는 한계점을 가지고 있다. 첫 번째 이유는 한국어 STT API가 음성을 텍스트로 정확하게 변환해주지 못하기 때문이다. 현재 한국어 STT API는 문법적으로 오류가 적고 발음이 명확한 음성에 대해서만 정확히 변환해주고 있다. 따라서 영상에 유해 표현이 존재하더라도 텍스트로 정확히 변환되지 않아 유해 영상을 탐지하지 못하는 문제가 많이 발생했다. 첫 번째 문제는 추후 별도로 한국어 STT API 성능을 높이는 연구를 진행하여 해결하려 한다.

두 번째 이유는 새로 만들어진 유해 표현을 탐지하지 못했기 때문이다. 본 논문에서 구성한 유해 사전은 청소년들이 자주 사용하는 유해 표현을 포함하지 못하고 있다. 왜냐하면, 청소년들의 신조어와 은어 생산 속도는 점차 빨라지고 있으므로 모든 유해 표현을 사전에 입력하는 것은 많은 시간과 비용이 필요하기 때문이다. 따라서 두 번째 문제를 해결하기 위해서는 시간과 비용을 투자해 유해 사전의 규모를 키워야 한다. 그리고 기계 학습을 이용하여 다양한 형태의 유해 표현을 찾아낼 수 있는 탐지 모델을 만든다면 기존 시스템보다 높은 정확도를 얻을 수 있을 것이다.

참고문헌

[1] 한국교육개발원 “학교생활에서의 욕설 사용 실태 및 순화 대책”

[2] 변은숙 “중학생 욕설사용과 공격성과의 관계에서 충동성의 매개효과” 강원대학교 교육대학원 학교상담전공 석사학위 논문 2016. 8

[3] 보건복지부 “방송의 청소년 유해환경 실태조사 및 개선방안” 2004 국무총리 청소년보호위원회

[4] 양은령 “여중생의 화장품 소비행동과 아이돌 연예인 모방행태” 이화여자대학교 교육대학원 석사학위 청구논문 2011

[5] 김우준 “폭력적 영상물에의 노출이 청소년의 비행행동에 미치는 영향” 한국치안행정논집 제8권 제1호 : 305~326

[6] 이미숙 “TV미디어가 청소년의 신체이미지와 의복행동 및 연예인 모방행동에 미치는 영향” 충남대학교 대학원 의류학과 의류학전공 박사학위 논문 2000. 8

[7] 김재숙 “연예인 모방행동이 청소년의 의복행동에 미치는 영향” 2002년 04월 Family and Environment Research 40권 4호 201-210

[8] 유상미 “폭력 영상매체 노출에 따른 청소년 비행연구 - 사회학습 이론과 둔감화 이론을 중심으로” 청소년문화포럼 제23권, 2010.04, 41-71

[9] 임동찬 “청소년의 은어와 비속어 사용 실태 의식 연구” 아주대학교 국어교육 석사학위 논문 2015. 8

[10] 이영주, 송진 “개인방송 콘텐츠 수용에 대한 탐색적 연구” 방송통신연구 2016년 가을호 (통권 제96호)

[11] 김기정 “폭력적 인터넷 1인 방송에 대한 청소년의 인식과 모방행동에 관한 연구” 서강대학교 언론대학원 디지털미디어 전공 석사학위 논문 2017. 8

[12] 김재화 “코미디 프로그램의 폭력성과 선정성이 청소년 가치관 형성에 미치는 영향에 관한 연구” 중앙대학교 신문방송 대학원 석사학위 논문

[13] 정윤이 “TV연예·오락프로그램의 가학성이 청소년에게 미치는 영향” 중앙대학교 신문방송 대학원 석사학위 논문.

[14] 조아영 “웹 게시판 비속어 처리 프로그램의 설계 및 구현” 컴퓨터산업학회논문지 2001. 10, vol. 2, No. 10, October

[15] 이호석, 이홍래, 한요섭 “반자동 학습 기반의 비속어 및 욕설 탐지 시스템” 2017년 한국컴퓨터종합학술대회 논문집

[16] 박교현, 이지형 “SVM을 이용한 온라인게임 비속어 필터링 시스템” 한국정보과학회 2006 가을 학술발표논문집 제33권 제2호(B), 2006.10, 260-263

[17]https://github.com/vhrehfdl/Inappropriate_Video_Filter 유해 사전 github 주소

[18] 권선미 “통신 언어 성 욕설의 실태 분석 -10대와 20대 누리꾼을 중심으로-” 단국대학교 교육대학원 국어교육과 국어교육 전공 석사학위 논문 2008. 2

[19] 정진구, 임동호 “한국인의 전통적 효사상과 부모부양에 관한 연구” Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology Vol.7 No.5 [2017]

[20] 강혜경 “중학생 욕설사용과 공격성과의 관계에서 충동성의 매개효과” 유교문화연구 제15집 , 83~108쪽