악플 자동 수집기

팀명 : 선플 수호자
팀원 : 주일규 (팀장), 최유정
소속 : 위스콘신 대학교 (주일규), 고려대학교 (최유정)

1. 들어가며 : 사회 문제로서의 '악플'

현대 사회인에게 익숙한 이름의 권리, '표현의 자유'의 역사적 흐름은 인터넷의 등장과 함께 격변의 시기를 마주한다. 시공간을 초월하여 소통이 이루어지는 인터넷에서 개개인은 익명성과 비대면성 뒤에 숨어 광범위한 언어 폭력을 행사할 수 있게 되었다. 인터넷 댓글은 2000년 이후 한국의 담론 문화와 여론 형성을 이끌어왔다고 해도 과언이 아니다. 문제는 "인터넷의 역사가 댓글의 역사라면 다른 한편으로는 '악플의 역사'라고 표현해도 될 만큼" (황용석 2007) '악플' 은 그동안 우리 사회에 심각한 문제를 야기해왔다는 점이다.[1] (김민기 2008) '악플'이란 '악성 리플(Reply)' 또는 '악성 댓글'의 줄임말로, 인터넷상에서 다른 사람이 올린 글 또는 콘텐츠에 대해 악의적인 댓글을 달거나 험담을 하는 것을 지칭한다.[2] (나은영 2015) 악플은 논리성이 없으며 대부분 감정에 치우친 욕설이나 비방 등 인신공격적 내용을 담고 있다.[3] (김상겸 2011)

특히나 다수가 특정 대상을 집중적으로 공격할 시, 악플러들이 휘두르는 언어의 칼은 당사자에게 견딜 수 없는 폭력으로 다가오기도 한다. 악성 댓글은 인격권을 침해할 수 있으며 심리적·윤리적으로도 문제를 일으키기 때문에 악성 댓글을 표현의 자유 범주 내에서 관대히 보아 넘기기는 어렵다.

나아가 2010년경부터 사회문제로 대두된 혐오 표현이 악플 문화의 폐혜를 가속화시켰다. 혐오 표현은 기존의 악플 문화를 더욱 악화시켰는데, 이는 특정 집단에 대한 선입견과 차별 강화에 기여하는 혐오 표현의 특성 때문이다. 소수 집단에 대한 혐오에서 근거하며, 대상이 소수자와 일반 청중들이고, 소수자에 대한 차별을 공공연하게 드러낸다는 것은 혐오표현의 세 가지 구성요소이다.[4] (홍성수 2015) 위의 특성들과 무조건적인 비방이 합쳐져 이제 혐오 표현과 악플을 서로 뗄 수 없는 관계가 되었다.

2. 악플 자동 수집기의 필요성

악플로 인해 최근 몇 년간 상당수의 사람이 생명을 포기하였고 다수의 연예인이 우울과 괴로움을 호소했다. 악플에 대한 체계적인 제재가 부재하기에 악플로 인한 피해는 우리 사회에 계속해서 반복된다. 악플에 대한 법적 조치는 대체적으로 정보통신망 이용촉진 및 정보보호 등에 관한 법률 제70조의 명예훼손 혹은 형법 311조 모욕죄를 토대로 이루어진다. 하지만 이렇게 법적 절차 밟는 것을 시작하기도 전에, 당사자들은 자신들의 악플을 모으는 것에서부터 시작해야 한다. 자신을 무차별적으로 비방하는 악플들을 모두 하나하나 모으는 것에서부터 악플과의 전쟁이 시작되기 때문에 많은 이들은 악플과 맞서 싸울 의지를 잃기도 한다.

---

[1] 김민기, 이진로 (2008) 인터넷의 공공성과 '악플'에 대한 대처 방안에 관한 연구, 정치커뮤니케이션 연구, 9, 5-50

[2] 나은영 (2015) 인터넷상의 야누스, 악플의 사회심리학 , 언론중재, 13, 16-27

[3] 김상겸 (2011) 인터넷상 표현의 자유의 제한에 관한 연구 - 악성 댓글의 제재를 중 심으로, 비교법연구, 11:3, 47-72

[4] 홍성수 (2015) 혐오표현의 규제 : 표현의 자유와 소수자 보호를 위한 규제대안의 모색, 법과사회 50호, 2015. 12. pp. 287-336

악플에 대한 우리 사회의 대응은 중장기적 관점에서 미디어 교육부터 출발할 수 있다. 다만 당장 악플의 피해자들을 위한 법적 장치 마련과 법적 절차를 수월하게 도와줄 수 있는 기술이 필요하다. '악플 자동 수집기' 는 당장에라도 실용적이고 효과적인 도구로 활용될 수 있다.

악플 자동 수집기는 특정 키워드에 대한 악플을 자동적으로 수집하여 정리해주는 프로그램이다. 해당 프로젝트는 '특정 키워드에 대한 악플을 어떻게 더 쉽게 모을 수 있을까'하는 고민에서부터 시작되었다. 현재는 명예 훼손 피해를 본 당사자 혹은 그의 지인이나 법률 사무소에서 법적 절차를 밟기 위해서 악플을 하나하나 수집해야만 한다. 수동으로 수집하는 위의 방법은 크게 두 가지의 한계가 있다.

첫째, 악플을 수집하는 과정은 피해자들에게 심리적 고통을 준다. 자신에 대한 무조건적인 욕설과 비방을 직접 수집하는 과정은 피해자에게 더욱더 커다란 심리적인 부담을 주기 때문이다. 악플의 대상이 되어 여럿에 의해 그것을 직간접적으로 접했을 때 당사자는 엄청난 트라우마를 겪기도 한다. 자동으로 악플을 수집할 수 있다면 피해자들의 심리적 고통을 줄일 수 있을 것이다.

둘째, 악플은 실시간으로 업데이트된다. 매일, 매시간 마녀사냥의 피해자 혹은 그의 지인이 실시간으로 축적되는 악성댓글을 모으는 일은 엄청난 시간과 노력을 필요로한다. 또한 실시간으로 업데이트되는 댓글들을 수동으로 검수하고 수집하며 필연적으로 빠뜨리는 데이터가 생길 수 있다. 악플 자동 수집기는 수집의 신속성과 악플 데이터의 정확도를 보장할 수 있다.

3. 자연어처리를 활용한 '악플 자동' 수집기의 의의

자연어처리와 딥러닝 모델을 활용하여 우리는 악플을 효과적으로 자동 분류할 수 있다. 우선 악플을 일반 댓글과 분류하기 위해서는 키워드적 접근이 아닌 문맥적이고 확률적인 접근을 해야 한다. 일반적인 자연어의 화자와 청자는 발화 속의 공격성이나 인신 모독성을 존재 여부뿐만 아니라 정도(程度)로도 인식할 수 있다. 따라서 자연어처리 모델을 사용하여 악플을 분석할 때, 악플인지 아닌지의 여부를 가리기에 앞서 특정 발화가 악플일 확률을 계산하는 것이 선행되어야 한다.

실제로 대다수의 선행연구에서는 혐오 표현 (Hate speech)과 먼지 차별[5](Mircroaggression) 을 분석하기 위해 문맥을 분석할 수 있는 모델을 활용한다. 먼지 차별의 경우 혐오 표현보다 어휘적 특성이 적고 문맥적 특성에 집중해야 하기 때문에 단순 통계 혹은 검출로는 '악플'을 탐지해내지 못할 것이라는 점을 알 수 있다.[6] (Breitfeller 2019) 실제로 위 논문에서는 먼지 차별(Mircroaggression)이 포함된 데이터를 11가지의 기준에 맞춰 태깅을 거친 후 학습 데이터로 활용한다. 이를 참고했을 때, 먼지 차별과 혐오 표현을 아우르고 단순 비방이나 욕설까지 포함되는 개념인 '악플'을 제대로 분류하기 위해서는 문맥적, 확률적 접근이 필요하다.

따라서 이 프로젝트에서는 악플을 검출하기 위해 기존에 존재하는 'Korean HateSpeech Dataset'[7] 을 활용하여 사전 훈련 (pre-train)을 시켰다. 위 모델을 통해 특정 데이터 속의 '악플 확률'을 점수로 환산하였다. 이후 점수에 관한 기준점을 0.6 이하 '확률 낮음', 0.6-0.7 '확률 중간', 0.7 이상 '확률 높음'으로 설정하였다. 이후 웹사이

---

5 먼지차별이란 눈에 잘 띄지 않지만 도처에 깔린 미세먼지만큼 해로운 작은 차별을 뜻한다.
출처 : 한겨레, 미세먼지처럼 해롭고 만연한 '먼지차별' 당신은?, 2018.04.11.

6 Luke M. Breitfeller, Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts, 2019 Association for Computational Linguistics

7 Jihyung Moon, Won Ik Cho, Junbum Lee, BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection, 2020 Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media

트에서 “악플만 보기” 버튼을 누르면 점수가 0.5 이상인 댓글들만 등장하며, 버튼을 해제하면 키워드와 관련된 해당 데이터 베이스의 모든 댓글을 열람할 수 있다. 댓글들은 한번에 엑셀 파일(.xls)로 내보내어 사용자들이 손쉽게 내려받을 수 있도록 시스템을 설계했다. 사용자의 필요에 따라 엑셀에서 점수/ 닉네임/ 작성 일시 등으로 필터링하여 볼 수 있는 것 또한 시스템의 중요한 장점이다.

4. 악플 자동 수집기란?

가. 학습 데이터 셋 : Korean HateSpeech Dataset

본 프로젝트에서는 한국의 혐오 표현 데이터가 모여있는 ‘Korean HateSpeech Dataset’을 사용했다. 편견 (Bias), 혐오 (hate), 성별 고정관념 (contain_gender_bias)이라는 세가지 기준으로 데이터셋은 미리 태깅되어 있다. 위 데이터는 총 9381개의 댓글이며, 직접 어노테이팅(annotating) 되었다.

위 프로젝트에서 해당 데이터셋을 학습데이터로 선택한 이유는 데이터셋의 명확한 어노테이팅 기준과 충분한 학습 데이터 양 때문이었다. 데이터셋의 어노테이션 가이드라인에 따르면, 혐오(hate) 레이블은 “부정적인 정서를 함유한 댓글 중에서 대상에 대해 근거없이 비난하거나 깎아내리는 경우, 대상이 모욕감 혹은 수치감을 느낄 수 있는 발언”을 태깅한다. 특히, 가이드라인에서는 “최수종 유이 연기 왜케 어색하지?ㅠ” 와 같이 합리적인 수준의 비판이나, 부정적인 감정이 없는 뉴스 댓글에 대해서는 혐오로 태깅하지 않았다. 이 프로젝트에서 검출하고자 하는 악플은 감정에 치우친 욕설이나 비방 등 인신공격적 내용에 한정된다는 점에서, 정확한 태깅 기준을 가진 위와 같은 데이터셋을 학습데이터로 선정했다.

나. 사용 모델 : DistilKoBERT

본 프로젝트에서는 위 데이터셋을 활용하여 혐오 표현 여부를 예측하도록 DistilKoBERT 모델을 (KoBERT 의 경량화 버전, Github: @monologg/DistilKoBERT) 구성하였다.
(768-hidden, Dropout rate: 0.5, 92,186,880 params)

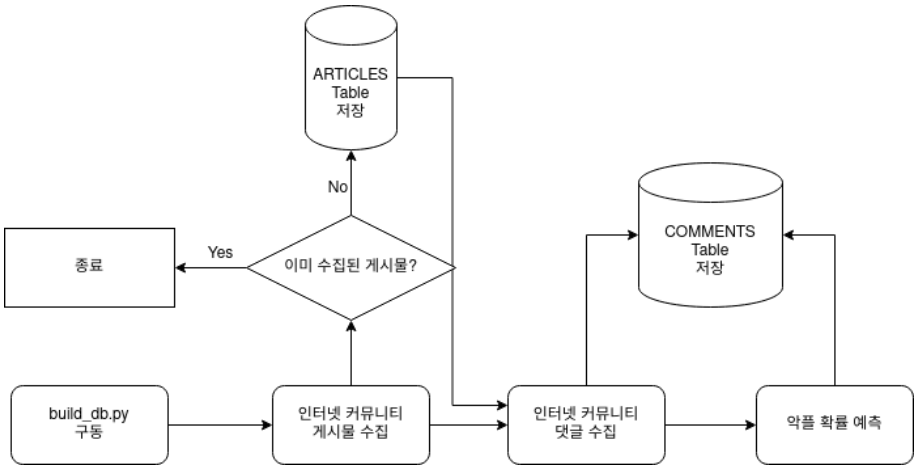
모델을 학습시킬 때는 각각의 데이터 속 ‘악플’을 먼저 찾아낸 후, 해당 예측이 정답일 확률을 수치화 하였다. 각각의 댓글 데이터 안에 악플이 없다면 ‘악플 없음’ (none), 비방의 표현이 있는 등 악플이 포함된 경우에는 ‘악플 존재’ (hate, offensive) 로 분류하도록 학습시켰다. 이모티콘과 반복된 문자의 사용 등은 사전에 제거하는 전처리 작업을 거쳤다. 훈련을 완료한 모델이 Korean HateSpeech Dataset에서 제공한 테스트 셋을 예측한 결과의 정확도는 0.7176이었다.

다. 백앤드 (Back end)

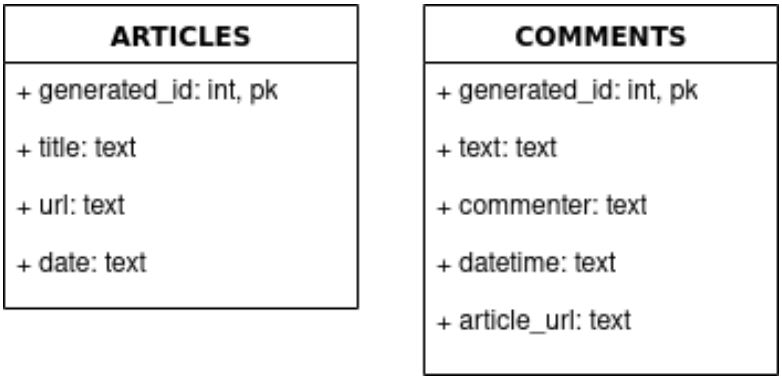
댓글 데이터베이스가 핵심인 백 엔드는 인터넷 커뮤니티에서 댓글을 자동 수집하고, 예측 모델의 평가 결과를 DB에 저장한다. 이렇게 구축된 DB와 악플 예측 모델은 FastAPI 기반의 API로 서비스되어 프론트 엔드에서 이용된다.

먼저, 인터넷 커뮤니티 댓글 자동 수집의 대상은 일베저장소(www.ilbe.com)의 〈일간 베스트 게시판〉과 인벤(www.inven.co.kr) 의 〈오픈 이슈 갤러리 3추글 게시판〉이었다. 선정한 두 게시판은 각 커뮤니티의 인기글이 올라오는 곳으로, 게시물 당 올라오는 댓글의 수가 일반 게시물에 비해 현저히 많다는 특징을 갖는다. 댓글 수집은 2020년 9월 7일부터 이뤄졌으며, 제출 시점(2020년 9월 15일)에 수집이 완료된 댓글은 총 494,331건이며, 제출 이후에도 꾸준히 업데이트 중이다.

댓글 데이터 베이스는 크론탭 (crontab) 을 통해 확보했다. 매일 실행되는 데이터베이스 구축 스크립트 (backend/build_db.py) 는 우선적으로 인터넷 커뮤니티의 인기 게시물을 수집한다. 만약 새로운 게시물이 있다면 해당 게시물에 대해 댓글을 수집하게 되며, 모델이 예측한 악플 확률을 데이터베이스에 저장한다. 댓글을 저장하는 데이터베이스로는 SQLite를 사용했으며, 데이터베이스의 구조 및 수집 결과는 아래와 같다.



[그림 1] 데이터베이스의 구조 및 수집 결과



[그림 2] 데이터베이스 수집 결과

ARTICLES: 게시물의 제목, URL, 게시 일자를 포함한다.
COMMENTS: 댓글의 내용, 작성자 닉네임, 작성 일시, 게시글 URL을 포함한다.

- 총 게시물 수: 22,582건 (일베:12,304건, 인벤: 10,278건)
  - 일베: 2020-08-28 ~ 2020-09-15 사이의 일간 베스트 게시글
  - 인벤: 2020-06-07 ~ 2020-09-15 사이의 오픈 이슈 갤러리 3추글
- 총 댓글 수: 494,331건 (일베:325,890건, 인벤: 198,636건)

라. 프론트앤드 (Front end)

본 프로젝트는 사용자들이 다양한 기기에서 프로젝트의 결과물을 보고 체험할 수 있도록 React 기반의 Ionic 을 개발 프레임워크로 사용했다. Ionic 은 하나의 코드베이스로 PC/안드로이드/iOS 에 호환하는 앱을 만들 수 있다는 장점이 있다. 사용자는 프로젝트의 데모 웹사이트 (http://akpl.xyz) 를 방문해 악플 자동 수집기가 모은 악성 댓글을 검색하고, 특정 키워드에 대한 댓글을 엑셀 파일(.xls) 형식으로 내려받을 수 있다. 키워드 검색 시, 사

용자는 로그데이터로 게시물의 경우 게시글의 제목, URL, 게시 일자를 볼 수 있으며, 댓글의 경우 댓글의 내용, 작성자 닉네임, 작성 일시, 게시글 URL을 볼 수 있다.

나아가, 데모 웹사이트는 사이드 기능으로 사용자가 지정한 문장에 악플에 준하는 표현이 존재하는지 예측해주는 서비스도 제공하고 있다. 사용자는 이 탭을 통해 여러 문장들을 직접 실험해보며 악플 검출 과정에 신뢰를 쌓을 수 있다.

5. 한계점 및 보완사항

이 프로젝트는 악플 자동 수집기의 의의, 사용 예시를 보여 주기 위한 개념 증명(Proof of Concept)의 일환으로, 한계점이 존재한다. 예를 들어, 컴퓨팅 자원의 한계와 개발 자원의 제한으로 인해 예측 모델의 경우 KoBERT 의 경량화 버전인 DistilKoBERT, 크롤링은 Python 의 BeautifulSoup 라이브러리, 데이터베이스는 SQLite 를 사용했다. 추후 (1) 모델 정확도를 높이기 위한 노력, (2) 크롤링 체계화 및 대상 다양화, (3) 방대한 댓글 데이터를 저장하는 더 나은 성능의 DBMS로 교체 등이 이루어진다면 향상된 정확도, 속도, 및 커버리지(다양한 인터넷 커뮤니티에 대한 댓글 수집)를 기대해 볼 수 있다.

악플 수집기의 이론적 한계는 악플의 기준이 본질적으로 모호하다는 점이다. 악플은 일종의 감정 표현이기 때문에 그 범위가 주관적일 수 밖에 없다는 점에서 완성도 있는 분류를 만들기에는 아직 어려움이 있다. 추후 데이터 셋을 구체적으로 구축할 때에는, 악플 탐지 어노테이션을 0부터 10까지의 스펙트럼으로 태깅하는 방법도 고려하고 있다.

6. 사회적·학문적 의의 및 활용 가능성

악플 자동 수집기는 다음과 같은 사회적 활용가치를 가지고 있다. 첫째, 악플로 인해 피해를 본 당사자들에게 법적 절차를 밟기 위한 기초 자료를 제공하고 심리적 부담을 덜어준다는 점에서 큰 의의가 있다. 악플로 인해 이미 괴로워하는 피해자의 추가적인 정신적 피해를 예방할 수 있으며, 법적인 절차를 진행하기 위한 과정에서의 인적, 시간적 자원을 아낄 수 있을 것이다. 둘째, 커뮤니티 별로 댓글 중 악플 비율의 수치화가 가능하다. 특정 커뮤니티의 전체 댓글들 중 악플의 비율을 수치화한다면 해당 사이트의 공격성을 평가할 수 있는 구체적인 근거가 될 것이다.

또한 악플 자동 수집기는 이후 연구의 훌륭한 학문적 도구로 사용될 것이다. 첫째, 커뮤니티 별 성향에 대한 양적 연구 진행을 수월하게 한다. 인터넷 커뮤니티 각각의 정치적, 사회적 성향을 측정할 수 있는 자료를 샘플링하는 것에 활용이 가능하다. 특정 도메인에서 공격하는 대상에 대한 데이터를 수치화하면 이는 해당 커뮤니티와 그 안의 문화에 대해 많은 것들을 설명해줄 것이다. 또한 둘째, 특정 집단을 향한 악플 연구의 토대가 되어줄 수 있다. 인종, 성별, 성적 지향성 및 성 정체성, 장애와 같은 다양한 소수자 집단을 지칭하는 말을 키워드로 검색한다면, 그들을 향한 해당 집단의 악플 수집이 가능하다. 이를 통해 새로 등장하는 비속어를 발견할 수도 있으며 그 추이를 기록하는 사회학적 연구에도 활용될 수 있다.

“악플은 안 질린다. 봐도봐도 새롭다. 무뎌지거나 질린다거나 그런 것이 없다.” 웹툰 작가 이말년씨의 지적이다. 여럿이 뭉쳐 특정 개인 혹은 집단을 망칠 수 있는 악플에 대한 고소와 신고의 과정이 피해자에게 또다른 부담이 되면 안된다. 악플의 피해자들의 기본적인 권리를 지키기 위한 험난한 과정 속에서 선플수호자는 그들의 든든한 일행이 되어줄 것이다.