CNN-LSTM 기반 유해 텍스트 필터링 크롬 플러그인

황현빈*, 김한겸*, 정진우*, 정혁순*, 서충원**, 이수원*
*숭실대학교 소프트웨어학부
**홍익대학교 컴퓨터과학부
e-mail: hhb9817@naver.com, kgimhg@naver.com, jinwoo0527@gmail.com,
jhs9810299@gmail.com, royalcircle97@naver.com, swlee@ssu.ac.kr

A Chrome Plug-in for Harmful Text Filtering based on CNN-LSTM

Hyun-Bin Hwang*, Han-Kyum Kim* , Jinwoo Chung*, Hyuk-Soon Chung*,
Choong-Won Seo**, Soowon Lee*
*School of Software, Soongsil University
**School of Computer Science, Hongik University

요 약

최근 온라인 매체에서 무분별한 비속어나 욕설 사용이 늘어남에 따라 유해한 텍스트를 자동으로 필터링하는 시스템의 필요성이 증가하고 있다. 유해 텍스트 필터링 관련 기존의 접근방법은 채팅 프로그램 등 특정 프로그램에 한하여 적용이 되거나 특정 포탈의 웹페이지에 국한되어 적용이 되는 한계가 있다. 따라서 본 연구에서는 AI를 활용하여 모든 웹 페이지의 유해 텍스트를 필터링할 수 있는 Chrome Extension을 구현하고 그 유효성을 검증한다.

1. 서론

스마트폰의 보급과 인터넷의 발달에 따라 비속어나 욕설 등 올바르지 못한 표현을 사용한 글들이 쉽게 노출되어, 정신적으로 미성숙한 청소년들이 언어체계를 확립하는데 부정적인 영향을 끼치는 경우가 많아지고 있다. 또한, 악성 댓글 등의 자극적이고 공격적인 표현은 보는 이의 감정을 상하게 하거나 분란을 조장하고 집단 간의 갈등을 심화시키는 등 사용자의 서비스 이용 경험을 질적으로 저해한다. 온라인 소통의 영향력이 그 어느 때 보다 커진 현대사회에서 이러한 유해 텍스트 문제는 중요하게 다루어져야 하며 다양한 분야에서 활용되고 있는 AI 기술이 반드시 투입되어야 할 영역이다.

본 연구에서는 이러한 문제를 해결하기 위하여 머신러닝 기술을 활용하여 웹 상의 각종 유해 텍스트를 필터링하는 기술을 제시한다. 유해 텍스트 필터링 관련 기존 연구로 [1]과 [2]에서는 필터링 기술을 해당 연구에서 제작한 채팅 프로그램에 한하여 적용하였다. [3]은 네이버에서 자체적으로 개발한 유해한 댓글 및 이미지 등을 필터링 하는 프로그램으로, 네이버 웹 페이지에 한하여 필터링을 적용한다. 본 연구는 개발할 프로그램의 형태를 Google Chrome 브라우저의 확장 플러그인으로 하여, 사용자가 볼 수 있는 모든 웹 페이지를 실시간으로 필터링하는 기능을 설계하고 구현한다.

2. 관련 연구

유해 텍스트 필터링 관련 기존 연구로 [1]에서는 변형된 금칙어를 자동으로 검출해낼 수 있는 개선된 금칙어 필터링 기법과 이를 이용하는 실시간 채팅 검열 시스템을 제안하였다. 또한 변형된 금칙어를 자동으로 검출해낼 수 있는 금칙어 및 변형 금칙어 자동 필터링을 이용한 실시간 자동 채팅 제재 수위 판별 시스템을 제시하였다. [2]에서는 Text-CNN 모델을 바탕으로 비속어를 식별하여 필터링을 하는 형식의 채팅 프로그램을 구현하였다.

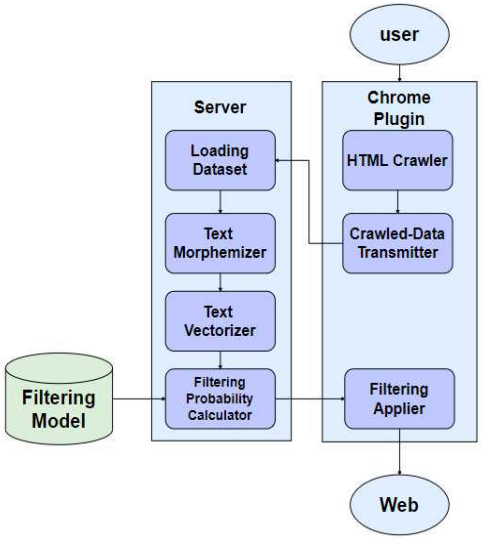
한국어 임베딩 관련 연구로 [4]에서는 Embedding layer에서 사용한 라이브러리에 따른 성능을 비교하였다. [4]에서 제시된 한국어의 특성을 고려한 단어 임베딩 기법과 그에 적합한 전처리 과정 및 파라미터에 의한 결과에 따르면 Fasttext 알고리즘을 사용하는 경우 그 성능이 가장 높게 도출됨을 제시하였다.

3. 유해 텍스트 필터링 크롬 플러그인

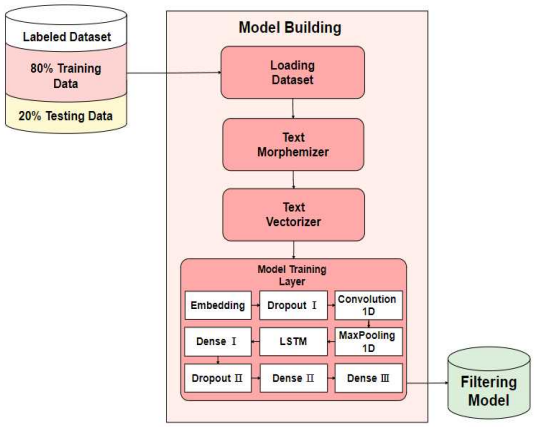
본 프로그램의 목적은 사용자들이 유해한 텍스트를 접하지 않게 하기 위해 필터링을 하는 것이다. 사용자가 Chrome을 통해 웹 페이지에 접근하면, 해당 웹 페이지에 있는 욕설을 탐지하여 유해하다고 판단된 경우 해당 욕설을 미리 설정한 텍스트로 대체하여 필터링한다. 제안 프로그램은 특정 웹 페이지에 국한되지 않고 모든 웹 페이지에 대해 필터링을 적용시킨다.

3.1. 프로그램 구조도

[그림 1]과 [그림 2]는 본 프로그램의 구조도 이다. 본 프로그램을 구성하는 모듈은 크게 Chrome Extension, Server, Model Building PC 3가지로 구성된다. 범용적인 적용을 위해 Chrome Extension을 사용하였고, 효과적인 필터링을 위한 AI 기술을 적용하였다. Model Building PC는 유해 텍스트를 식별하기 위한 CNN-LSTM 모델을 생성하고 학습하며, 학습된 모델을 Server에서 사용한다. Chrome Extension는 웹 페이지의 텍스트를 추출하여 Server로 전송한다. Server는 전송받은 텍스트에 대해 유해 여부를 식별하고 그 결과를 Chrome Extension으로 전송하며, Chrome Extension은 이를 토대로 웹 페이지에 필터링을 적용한다.



[그림 1] 확장 프로그램 구조도



[그림 2] 모델 학습 구조도

3.2. 프로그램 모듈

[표 1] 확장 프로그램 모듈 명세

구분	모듈명	입력 data	출력 data
Chrome Extension	Html Crawler	*.html	문장 별 텍스트
	Crawled-Data Transmitter	문장 별 텍스트	문장 별 텍스트
	Filtering Applier	문장 별 욕설 가능성	필터링이 적용된 문장 별 텍스트
		문장 별 텍스트	
Server	Loading Dataset	문장 별 텍스트	문장 별 텍스트
	Text Morphemizer	문장 별 텍스트	형태소 단위 배열
	Text Vectorizer	형태소 단위 배열	벡터화 된 배열
	Filtering Probability Calculator	모델 파일	문장 별 욕설 가능성
		벡터화 된 파일	

[표 1]은 확장 프로그램 모듈에 대한 명세이다. 사용자가 웹 페이지에 접근하면 웹 서버에서 Html code가 전송되며, Chrome의 확장 프로그램이 동작하여 HTML Crawler 모듈이 웹페이지 HTML Code를 Crawling하여 텍스트를 문장 단위로 가져온다. 그 후, Crawled-Data Transmitter 모듈이 서버의 Loading Dataset 모듈로 텍스트를 전달한다. Loading Dataset 모듈이 Text Morphemizer로 데이터를 전달한 후 형태소 단위로 텍스트를 분할한다. 그 후 분할된 텍스트를 모델에서 사용하기 위해 Text Vectorizer 모듈이 벡터화를 진행한다. 벡터화 된 텍스트에 대하여 Filtering Probability Calculator 모듈이 모델을 통해 각 문장이 욕설일 가능성을 계산하

여 Chrome Extension의 Filtering Applier 모듈로 보낸다. Filtering Applier 모듈은 전달 받은 각 문장의 욕설일 가능성을 통해 필터링 대상인 문장을 대체 텍스트로 치환한다.

[표 2] 모델 학습 모듈 명세

구분	모듈명	입력 data	출력 data
Model Building PC	Loading Dataset	*.txt	텍스트 배열
			분류 값 배열
	Text Morphemizer	텍스트 배열	형태소 단위 배열
	Text Vectorizer	형태소 단위 배열	벡터화 된 배열
	Training Model Layer	벡터화 된 배열	모델 파일
		분류 값 배열	

[표 2]는 모델 학습의 구성 모듈 명세이다. 모델을 학습할 때 사용할 데이터 셋을 불러와서 전처리를 해주어야 한다. 따라서 Loading Dataset 모듈이 Labeled Dataset을 불러온다. 불러온 데이터셋은 텍스트와 분류 값으로 구분되는데, 이 텍스트 데이터를 Text Morphemizer 모듈이 형태소 단위로 구분지어 준다. 이후 형태소 단위로 구분된 텍스트 데이터를 설계한 모델에서 사용하기 위해 Text Vectorizer 모듈이 벡터화를 진행한다. 이 후 Training Model Layer 모듈은 벡터화 된 텍스트 데이터와 분류 값 데이터를 사용하여 설계한 모델을 학습시킨 후 학습된 모델을 저장하여 후에 테스트 및 필터링 모듈에 삽입할 수 있게 한다.

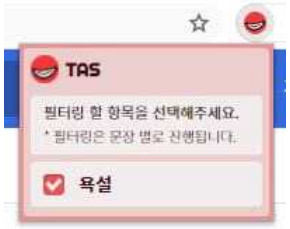
4. 구현

4.1. 데이터

본 연구에서는 네이버 뉴스 기사에 등록된 댓글들과 한국어 트위터 사용자의 트윗들, 비속/비윤리적 표현의 빈도수가 많은 특정 온라인 커뮤니티의 댓글들을 수집하여 Text 형식으로 구축된 'AI Hub - 인공지능 윤리 연구를 위한 비정형 텍스트 데이터셋 : Keti' 자료를 활용하였다. 이 중 부정적 정서로 인한 흥분상태일 때 혼자 감탄조로 사용하는 말, 상대방의 인격을 무시하고 모욕하거나 저주하는 공격적인 말을 욕설로 분류하였다. 이러한 방식으로 분류된 데이터셋의 80%를 훈련용, 20%를 테스트용 데이터셋으로 구분하였다.

4.2. 구현 결과

[그림 3]은 본 연구에서 개발한 유해 텍스트 필터링 플러그인 등록 화면이다. 플러그인을 가동하면 서버와의 통신을 통해 웹 페이지에 존재하는 유해 텍스트들이 필터링된다. [그림 4]와 [그림 5]는 각각 동일한 웹페이지에 대한 플러그인 가동 전과 플러그인 가동 후의 화면이다. [그림 5]에서 욕설 포함된 문장이 특정 문장으로 대체된 화면으로 사용자에게 출력된다.



[그림 3] 플러그인 등록 화면

[그림 4] 플러그인 실행 전 웹 페이지

[그림 5] 플러그인 실행 후 웹 페이지

4.3. 정확도 평가

학습된 모델에 테스트용 데이터셋을 적용하여 모델의 정확도를 검증하였다. 평가 기준은 F1-score를 사용하였으며, 실험 결과 약 84.4%로 산출되었다.

5. 결론 및 향후 연구

본 연구에서는 CNN-LSTM을 사용한 욕설 식별 모델 및 이를 적용한 웹 페이지 유해 텍스트 필터링 크롬 플러그인을 구현하였다. 본 프로그램를 통해 사용자는 웹 페이지에 존재하는 모든 텍스트에 대해 유해한 텍스트가 필터링 된 웹페이지를 제공받을 수 있다. 이를 통해, 인터넷 뉴스 혹은 SNS의 악성 댓글에 대한 문제를 해결할 수 있으며 깨끗한 인터넷 환경을 조성할 수 있는 효과가 있다.

향후 연구로는, 욕설뿐만 아니라 비속어, 성적 표현 등의 다양한 유해 텍스트를 필터링 할 수 있고, 사용자가 유해 텍스트의 범위를 선택하여 필터링할 수 있는 기능 등에 대한 개선이 필요하다. 또한, 유해 텍스트 식별의 정확도를 개선하기 위한 AI 모델의 파라미터 튜닝과 속도를 개선하기 위한 확장 프로그램의 알고리즘 구조의 개선 등이 필요하다.

감사의 글

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음 (2018-0-00209)

참고문헌

[1] ChanWoo. Kim†, Mee Young Sung, “Realtime Word Filtering System against Variations of Censored Words in Korean”, Journal of Korea Multimedia Society, Vol. 22, No. 6, pp. 695-705, June 2019.

[2] 이건환, 박주찬, 최동원, 이연경, 최호빈, 한연희, “딥러닝 기반 비속어 필터링 채팅 프로그램 설계 및 구현”, 한국정보처리학회 학술대회논문집, Vol. 26, No. 2, pp. 998-1001, November 2019.

[3] https://d2.naver.com/helloworld/7753273.

[4] 조현수, 이상구, “FastText를 적용한 한국어 단어 임베딩”, 한국정보과학회 학술발표논문집, pp. 705-707, December 2017.