딥러닝 기반 비속어 필터링 채팅 프로그램 설계 및 구현

이건환*, 박주찬*, 최동원*, 이연경*, 최호빈**, 한연희***†
*한국기술교육대학교 컴퓨터공학부
**한국기술교육대학교 대학원 컴퓨터공학과
***한국기술교육대학교 첨단기술연구소
e-mail: {hwanism, green669, lipoij, yeonk, chb3350, yhhan}@koreatech.ac.kr

Design and Implementation of Profanity Filtering Chat Program Based on Deep Learning

Geon-Hwan Lee,* Joo-Chan Park*, Dong-won Choi*, Yeon-Gyeong Lee*, Ho-Bin Choi**, Youn-Hee Han***†
*Dept of Computer Engineering, KoreaTech University
**Dept of Computer Engineering, Graduate School of KoreaTech University
***Advanced Technology Research Center, KoreaTech University

요 약

최근에 게임이나 채팅 프로그램 내에서의 비속어 필터링은 금칙어 기반으로 운영되고 있다. 하지만 금칙어 기반의 프로그램은 여러 한계점을 보이며, 따라서, 본 논문에서는 'Text-CNN'을 활용한 딥러닝 기법에 기반하여 비속어 필터링 프로그램을 제안한다. 데이터의 자질을 '자모' 단위로 전처리하여 학습시키고 어느 부분이 비속어인지 검출하여 마스킹 처리하는 'LIME 알고리즘'을 사용하여 우리의 프로그램을 이용하는 사용자들에게 바른 언어습관을 지향하며 더 나아가 올바른 인터넷 문화를 조성할 수 있도록 필터링 채팅 프로그램을 제안한다.

1. 서론

1.1 배경

작품도출 과정에서 현 사회에서 문제시되고 있는 언어폭력과 사이버 폭력, 이로 인해 야기되는 명예훼손, 모욕죄 등을 해결할 방법들이 연구되어왔다. 더 큰 문제점은 해가 지날수록 앞서 언급한 사례들이 점차 급증하는 추세를 보인다는 점이다. 이런 문제가 앞으로도 지속된다면 추후에 인터넷에서 마주하는 상대방을 향한 개개인의 언행에서 무분별한 모독과 욕설 행위를 양심의 가책 없이 더 쉽게 행할 것이며, 도리어 또 다른 문제를 초래할 수 있다. 따라서, 본 논문에서는 이런 문제점들을 예방하고자 딥러닝 기반 비속어 필터링 채팅 프로그램을 제안한다.

1.2 필요성

지난 2015년, 경찰청이 조사한'사이버 명예훼손'발생 건수[1] 를 살펴보면 2012년에는 5,684건이 발생하였으며, 6,320건(2013년), 8,820건(2014년)으로 매년 크게 증가하고 있다.'모욕죄'발생 건수[1] 또한 4,258건(2007년)에서 36,931(2015년)으로 약 8.7배 증가해 이에 대한 심각성이 대두되고 있다. 그러므로 올바른 언어 사용을 지향하며 상대방의 기분을 해칠만한 인격적인 모독, 무분별한 모욕, 비방 등의 문제를 해결해야 한다. 현시대에 대두되는 사회적인 문제를 완화하기 위해 인간의 개입 없이 비속어가 필터링할 수 있는 딥러닝 기반의 채팅 프로그램이 필요하다.

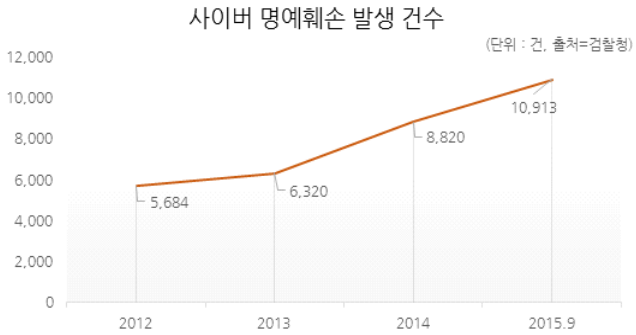

그림 1. 사이버 명예훼손

† 교신저자: 한연희(한국기술교육대학교)
이 논문은 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임
(No.2018R1A6A03025526 및 No.2016R1D1A3B03933355)

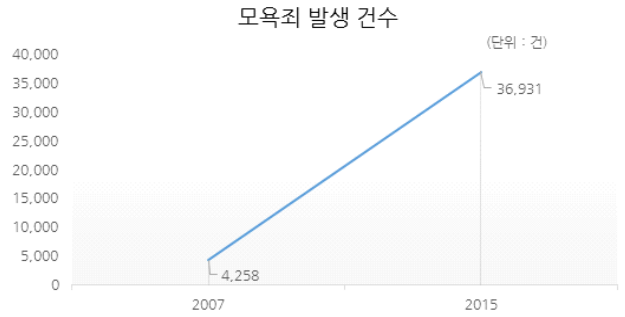

그림 2. 모욕죄 발생 건수

2. 제안 방법

2.1 데이터 전처리

딥러닝을 이용하기 위해서는 방대한 데이터가 필요하다. 제공되는 한국어 욕설 데이터가 없어, 직접 특정 악성 커뮤니티 사이트들에서 댓글들을 크롤링(Crawling)하여 수집 후, 각각의 댓글마다 분류 기준을 두어 레이블링(Labeling) 처리를 한다. 총 수집한 데이터 수는 약 400,000개이다. 데이터를 형태소 단위와 자모 단위 두 가지 방법으로 전처리를 한다. 형태소 단위의 단점으로는 방대한 단어사전을 구축해야 하며 사전에 없는 실제 데이터가 오게 된다면 성능이 저하된다. 또한, 형태소 분석기에 따른 한계점이 발생한다. 반면에 자모 단위는 (표 1)과 같이 초성, 중성, 종성, 숫자, 알파벳, 특수문자, 단일 문자로 총 127개의 한정된 단어사전 [2]만 있으면 모든 데이터를 벡터(Vector)로 전처리할 수 있다. 자모 단위의 장점으로는 채팅 데이터와 같은 비정형 데이터에서 형태소 단위 분해보다 성능이 높다.

표 1. 자모 단위 사전

카테고리	심볼
초성	ㄱ, ㄲ, ㄴ, ㄷ, ㄸ, ㄹ, ㅁ, ㅂ, ㅃ, ㅅ, ㅆ, ㅇ, ㅈ, ㅉ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ
중성	ㅏ, ㅐ, ㅑ, ㅒ, ㅓ, ㅔ, ㅕ, ㅖ, ㅗ, ㅘ, ㅙ, ㅚ, ㅛ, ㅜ, ㅝ, ㅞ, ㅟ, ㅠ, ㅡ, ㅢ, ㅣ
종성	ㄱ, ㄲ, ㄳ, ㄴ, ㄵ, ㄶ, ㄷ, ㄹ, ㄺ, ㄻ, ㄼ, ㄽ, ㄾ, ㄿ, ㅀ, ㅁ, ㅂ, ㅄ, ㅅ, ㅆ, ㅇ, ㅈ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ, (종성이 없는 경우 포함)
숫자	1, 2, 3, 4, 5, 6, 7, 8, 9, 0
알파벳	a, b, c, d, e, f, g, h, I, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z
특수문자	!, ", #, $, %, &, ', ?, @, *, +, , , -, ., /, ~, :, ^, (공백 포함)
단일 문자	ㄳ, ㄵ, ㅄ, ㄺ

자모 단위로 분해한 각각의 심볼(Symbol)들은 크기가 127인 One-hot Vector로 표현되며 입력의 길이의 최대 크기는 200으로 두었다. 200보다 작은 길이의 데이터는 Zero-padding을 붙여 크기가 200이 되게 (표 2) [2]와 같이 맞췄다.

2.2 모델 구조

데이터들을 학습시킬 딥러닝 모델로는 CNN 모델

표 2. 자모 단위 벡터화 도식도 [2]

로서, CV task에 이용된다. 최근 CNN을 NLP(Natural Language Processing, 자연어 처리) 문제에 활용하기 위한 많은 시도가 있었고, CNN 모델이 NLP 문제에 적합하다는 것이 여러 실험을 통해 밝혀졌다. CNN의 가장 큰 장점은 입력 데이터의 특징을 추출하는 데 적합하다는 것이다. 합성 곱은 컴퓨터 그래픽의 핵심적인 부분이며, GPU 레벨로 구현되어 있다. 본 논문에서 사용하는 CNN 모델의

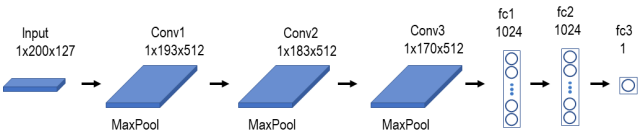

그림 3. CNN 모델 구조

구조는 (그림 3)와 같다.

총 3개의 Convolution Layer와 3개의 FC Layer를 사용하였다. Convolution의 kernel size는 순서대로 6, 9, 12이다. 모든 Convolution Layer의 뒤에는 Max Pooling을 사용하였으며 Kernel size는 3으로 동일하다. 각 FC Layer의 마지막 단에는 Overfitting을 방지하기 위해 Drop out을 두었다. Loss function은 Binary Cross Entropy Loss를 사용하였고, Optimizer는 Adam을 사용했다. Learning rate는 0.0001로 두고 20 epoch마다 Learning rate를 0.1씩 감소시켰다.

2.3 마스킹 처리

마스킹 처리는 LIME 알고리즘을 사용하였다. LIME 알고리즘은 학습모델이 측정한 결과에 대해 해석하는 알고리즘이다. LIME 알고리즘의 처리 과정은 (표 3)과 같다. Input data를 어절 단위로 나누

어 욕으로 판단된 어절을 '*'로 마스킹한다.

표 3. 마스킹 처리 과정

Input	존나 배고프다 시~~발	
Analysis	모든 경우의 수	결과(욕일 확률)
	배고프다 시~~발	99.89
	존나 시~~발	99.98
	존나 배고프다	97.68
	시~~발	99.99
	배고프다	00.02
	존나	97.89
Output	** 배고프다 ****	

3. 채팅 프로그램

채팅 프로그램은 Flask에서 WebSocket을 기반으로 개발하였다. Flask는 Python Framework이기 때문에 다른 환경보다 딥러닝 모델과의 연동이 쉽다는 장점이 있다 [3].

채팅 메시지가 마스킹 되어 출력되는 과정은 Input Data로 채팅 메시지를 받은 후, Emit 이벤트를 통해 Input Data를 전달한다. 전달받은 Input Data를 딥러닝 모델을 통해 욕인지 아닌지 예측하고 마스킹 처리해준 결과인 Output Data와 교체시킨다. 마스킹 처리된 Output Data는 다시 emit 이벤트를 통해 채팅창으로 출력된다.

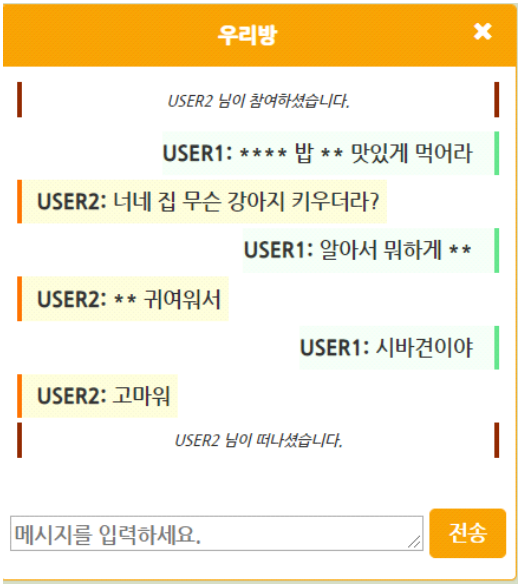

그림 4. 채팅 프로그램 화면

4. 실험 및 결과 분석

4.1 실험 데이터

실험을 위해 욕설이 포함된 특정 악성 커뮤니티 사이트들에서 댓글들을 Crawling 한 뒤, 분류로 평서문은 0, 욕설은 1로 'Labeling'을 해주었다. 총 데이터의 수는 400,000개이며, 욕과 평서문의 비율을 각각 50:50의 비율로 맞췄다. 70%를 Training data로 두고, 15%를 Validation data, 나머지 15%를 Test data로 학습을 진행했다.

4.2 모델 성능

자모 단위의 CNN과 형태소 단위의 LSTM 총 두 가지에 대해서 실험을 해보았고, 결과는 (표 4)과 같다.

표 4. 성능 측정

자질 단위	성능
자모단위 CNN	94.13%
형태소 단위 LSTM	85.82%

형태소 단위의 LSTM보다 자모 단위 CNN의 성능이 약 8% 정도 높게 측정되었다. 웹 사이트의 댓글이라는 비정형 데이터에 대해서 자모 단위로 했을 때 성능이 뛰어난 것을 볼 수 있다.

4.3 마스킹 처리 결과

(표 5)는 마스킹 처리 결과이다.

표 5. 마스킹 처리 결과

Input	욕일 확률	Output
몇학년 몇반이니?	00.94005	몇학년 몇반이니?
시바견	01.95044	시바견
시바	73.17643	**
ㅈ1랄 하고있네	96.48115	*** 하고있네
뭐하냐 ㅅㅐㄱ1야	92.81957	뭐하냐 *****

(표 5)의 '시바견'을 보면 기존 금칙어 기반의 욕설 필터링이었다면, '시바'를 욕으로 보고 마스킹 처리를 했지만, 딥러닝 기반의 모델은 욕일 확률을 약 2%로 예측하고 욕으로 판별하지 않는다. 또한 'ㅈ1랄', 'ㅅㅐㄱ1'같은 자음과 숫자를 결합하거나, 자모음을 서로 띄어서 교묘하게 우회된 욕설도 90%가 넘는 수치로 예측하며 우회된 욕설에 대해서도 좋은 결과를 보였다.

5. 결론

기존의 비속어를 필터링할 수 있는 기술력은 사람들이 수동으로 금칙어를 지정해야 하는'금칙어 기반'에 그쳤다. 본 논문에서는 최근 많은 곳에서 주목받고 응용되고 있는 인공지능의 기술 중 하나인 '딥러닝'을 활용한 필터링 채팅 프로그램이다. 딥러닝을 통하여 모델에 데이터를 입력하면 스스로 학습하여 데이터들을 분류할 수 있는 프로그램이 되기 때문에 사람이 직접 금칙어들을 제작할 필요가 없어지게 되고, 우회된 데이터와 비정형 데이터들에 대한 필터링 정확도도 증가한다. 우리의 아이디어가 활성화된다면 최근 사회적인 문제 중 급증하는 추세를 보이는 '사이버 명예훼손' 및 '모욕죄' 등의 심각한 문제들을 해결하는 데 도움이 된다. 또, 새롭게 인터넷 문화를 접하는 유아, 청소년, 노인들에게도 올바른 인터넷 문화를 조성해줄 수 있으며, 올바른 언어습관을 갖게 해주므로 많은 문제를 예방한다.

참고문헌

[1] 박지연, "모욕하는 대한민국… 최근 9년간 모욕죄 고소・고발 급증," 한국일보, 2016.

[2] 신해빈, 서민관, 변형진, "한국어 자모 단위 기반의 Convolution Neural Network를 이용한 텍스트 분류," 한국정보과학회 2017년 한국컴퓨터종합학술대회, pp. 587-589, 2017.

[3] https://flask-socketio.readthedocs.io/en/latest.

[4] 허지웅, 김영랑, 천현강, 이재환, "Text-CNN을 활용한 다중 감정 분류 모델 개발 및 챗봇 학습 데이터 연동 방안 연구,"한국정보과학회 2017 한국소프트웨어종합학술대회, pp. 25-27, 2017.

[5] https://ratsgo.github.io.

[6] 김건영, 이창기, "Convolutional Neural Network를 이용한 한국어 영화평 감성 분석," 한국정보과학회 2016년 한국컴퓨터종합학술대회, pp. 747-749, 2016.