딥러닝 기반 욕설 탐지

김유민*,+, 강효빈*,+, 한수현*,+, 정희용*,**
*전남대학교 전자컴퓨터공학부
**전남대학교 소프트웨어공학과
culyumin33@gmail.com, 154206@jnu.ac.kr, 182765@jnu.ac.kr, h.jeong@jnu.ac.kr
+공동 1저자를 표시

Swear Word Detection through Convolutional Neural Network

Yumin Kim*,+, Hyobin Gang*,+, Suhyeun Han*,+, Hieyong Jeong*,**
*School of Electronics & Computer Engineering, Chonnam National University
**Dept. of Software Engineering, Chonnam National University
+indicates co-first authors.

요 약

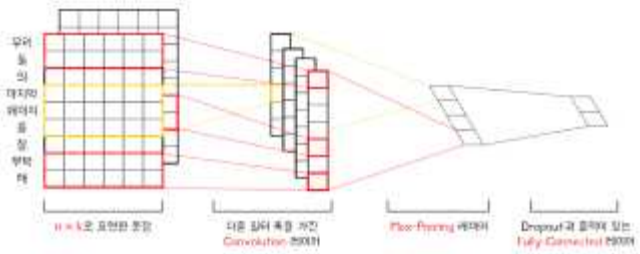
개인의 소셜미디어 활동이 활발해지면서 익명성을 악용하여 타인에게 욕설을 주저없이 해버리는 사용자가 늘고 있다. 본 연구는 욕설이 난무하는 채팅창에서 욕설 데이터를 크롤링하여 데이터셋을 구축하여 컨볼루션 네트워크로 학습시켰을 때 욕설을 탐지하고, 전체 문장에서 그 탐지한 욕설의 위치를 파악하여 블러링 처리를 할 수 있는지를 확인하는 것을 목적으로 한다. 전처리 작업으로 한글과 공백을 제외하고 형태소 단위로 토큰화한 후 불용어를 제거해서 패딩처리를 하였다. 학습 모델로는 1차원 컨볼루션을 사용하여 수집한 데이터의 80%를 훈련에 사용하고 나머지 20%를 테스트에 사용하였다. 키워드를 이용한 단순 분류 모델과 비교하였을 때, 본 연구에서 이용한 모델이 약 14% 정확도가 향상된 것을 확인할 수 있었다. 테스트에서 전체 문장에서 욕설이 포함되었을 때 욕설과 그 위치 정보를 잘 획득하는 것도 확인할 수 있었다.

1. 서론

소셜미디어와 온라인 게임 산업의 발달과 더불어 욕설을 포함한 사이버폭력이 지속적으로 발생하고 있다. 심한 욕설피해를 받은 피해자는 소셜미디어나 게임에서 이탈하거나 법적 소송을 해서 가해자가 법적 처벌을 받게 하기도 한다. 이에 욕설을 제재하는 것은 소셜미디어나 게임사에서 중요한 문제이다 [1]. 소셜미디어 활동은 사용자간 공감과 커뮤리케이션을 돕는 순기능도 있기는 하지만, 혐오표현이나 욕설 등 악용사례도 적지 않다. 이를 방지하기 위해 게임회사는 욕설 필터링이나 신고 제도 등 다양한 방지책을 도입했지만 완벽한 욕설 근절로 이어지지는 못하는 것이 현실이다. 이를 방지하기 위해 게임회사는 욕설 필터링이나 신고 제도 등 다양한 방지책을 도입했지만 완벽한 욕설 근절로 이어지지는 못하는 것이 현실이다. 현재 필터링 시스템은 보수적으로 탐지하는 경향이 높아 변형된 욕설을 놓치거나, 바른 표현이지만 욕설로 간주하는 경우가 자주 발생한다. 특히 한국어는 영어에 비해 어순히 중요하지 않고 교착어이며 띄어쓰기가 제대로 지켜지지 않기 때문에 탐지하기가 훨씬 까다롭다.

욕설 탐지 연구는 대부분 영어이며, 기계학습 기반으로 학습되었다. 유투브의 악설 댓글 탐지, 야후의 금융 및 뉴스 기사를 활용한 욕설 분류가 대표적인 예로 둘 수 있다. 한국어에서는 금칙어를 정해놓고, 해당 단어를 블러링하는 기존의 욕설 방지책은 한계가 분명하다. 모음 대신에 숫자를 사용한 변형된 욕설이나, 욕설을 사용하지 않았지만 그 의미가 공격적인 표현은 탐지하기 어렵다. 금칙어 기준을 강화하자니 평범한 말도 욕설로 제재되는 현상이 발생하여 사용자에게 오히려 불편함을 안겨주게 된다. 결국 현재는 운영자가 신고가 들어온 내역을 일일이 수작업으로 확인하여 처리해야 하는 방식을 취할 수 밖에 없다.

따라서 본 연구에서는 소셜미디어에서 욕설을 크롤링하여 데이터셋을 구축하여 단어를 형태소 단위로 토큰화한 다음 컨볼루션으로 학습을 시켰을 때 욕설을 잘 분류할 수 있는지 확인해 보는 것이 목적이다. 또한 욕설이 포함되어 있다면 문장 가운데 어느 위치



(그림 1) 컨볼루션에 의한 욕설탐지 모델.

Model: "model"

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 20)]	0
embedding (Embedding)	(None, 20, 256)	2294272
conv1d (Conv1D)	(None, 18, 317)	243773
max_pooling1d (MaxPooling1D)	(None, 9, 317)	0
conv1d_1 (Conv1D)	(None, 9, 317)	100806
conv1d_2 (Conv1D)	(None, 8, 317)	201296
global_max_pooling1d (Global	(None, 317)	0
dense (Dense)	(None, 128)	40704
preds (Dense)	(None, 1)	129

Total params: 2,880,979
Trainable params: 2,880,979
Non-trainable params: 0

(그림 2) 실제 학습에 사용된 컨볼루션 모델.

에 포함되어 있는지 표시하여 욕설로 분류된 단어만 블러링 처리할 수 있는지도 살펴보도록 하겠다.

2. 컨볼루션

그림 1은 컨볼루션에 의한 욕설탐지 모델을 보여주고 있다. 우선 전처리 과정으로 한글과 공백을 제외한 후 형태소 단위로 토큰화 작업을 수행한다. 이때 불용어는 제거하고 행렬의 크기를 동일하게 하기 위하여 패딩처리를 수행한다. 다음으로 1차원 컨볼루션을 사용하여 토큰화된 단어로 (행 x 열 = k x n) 크기의 행렬을 만들어 문장을 행렬로 표현한다. 다양한 필터 크기로 훈련을 시키면서 특징을 추추한 레이어에서 최대값을 추출한 후 flatten 레이어로 변환을 하여 학습을 시키게 된다. 학습의 결과로 욕설이면 1을 욕설이 아니면 0으로 분류한 결과값이 출력이 된다.

그림 2는 실제 학습에 사용된 컨볼루션 모델을 보여주고 있다. 입력값을 넣은 후 임베딩을 거쳐 (k x n = 20 x 256) 크기로 행렬화 작업을 마친 후 다양한 필터 크기로 컨볼루션으로 특징을 추출한 후 맥스풀링에서 최대값을 추출하여 평탄화 작업을 마치면 최대 2,880,970개의 파라메터를 생성하게 되고 이 파라메터가 학습에 사용되게 된다.

```
final_model = load_model('best_model2.h5')
print("\n 테스트 정확도: %.4f" % (final_model.evaluate(X_test, y_test)[1]))
```

테스트 정확도: 0.8565

테스트 정확도: 0.8565

(그림 3) 컨볼루션으로 학습된 결과.

욕설일 때

이 댓글은 욕설이 포함되어 있습니다.

(그림 4) 욕설 탐지 및 위치 정보 파악 결과.

3. 실험 결과

그림 3은 본 연구에서 구축한 컨볼루션 모델을 이용하여 학습시킨 결과를 보여주고 있다. 데이터의 80%를 학습에 20%를 테스트에 사용하였는데 테스트의 정확도가 86%인 것을 알 수 있다. 그림 4는 주어진 문장에서 욕설을 탐지할 수 있는지, 그리고 탐지된 욕설이 전체 문장에서 어디에 위치하고 있는지를 검출하고 있는 결과를 보여주고 있다.

실험 결과를 바탕으로 욕설탐지 뿐만 아니라 욕설의 위치 정보까지 잘 파악하고 있는 것을 확인하였다.

4. 결론

본 연구에서는 1차원 컨볼루션을 이용하여 욕설탐지 뿐만아니라 문장에서 욕설의 위치정보도 표시해 줄 수 있다는 것을 확인하였다.

사사

이 논문은 2021년도 전남대학교 SW중심대학 산학협력프로젝트와 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(2021R1I1A305521011)의 연구비 지원을 받아 수행되었음.

참고문헌

[1] 박성희, 김휘강, 우지영, “딥러닝를 사용한 온라인 게임에서의 욕설 탐지”, 한국컴퓨터정보학회 하계학술대회 논문집, 27(2), 2019.