딥러닝를 사용한 온라인 게임에서의 욕설 탐지

박성희[O], 김휘강[**], 우지영[*]
순천향대학교, 미래융합기술학과[O]
순천향대학교, 미래융합기술학과[*]
고려대학교 정보보호대학원[**]
e-mail: {sunghee, jywoo}@sch.ac.kr[O*], cenda@korea.ac.kr[**]

Abusive Sentence Detection using Deep Learning in Online Game

Sunghee Park[O], Huy Kang Kim[**], Jiyoung Woo[*]
Dept. of Future Convergence Technology, Soonchunhyang University[O]
Dept. of Future Convergence Technology, Soonchunhyang University[*]
Graduate School of Information Security, Korea University[**]

● 요 약 ●

욕설은 게임 내 가장 큰 불쾌 요소 중 하나이다. 지금까지 게임 사용자들의 욕설을 방지하기 위해서 금칙어를 기반으로 필터링 해왔으나, 한국어 특성상 단어를 변형하거나 중간에 숫자를 넣는 등 우회할 방법이 다양하기 때문에 효과적이지 않다. 따라서 본 논문에서는 실제 온라인 게임 'Archeage'에서 수집된 채팅 데이터를 기반으로 딥러닝 기법 중 하나인 콘볼루션 신경망을 사용하여 욕설을 탐지하는 모델을 구축하였다. 한글의 자음, 모음을 분리하여 실험하였을 때, 87%라는 정확도를 얻었다. 한 글자씩 분리한 경우, 조금 더 좋은 정확도를 얻었으나, 사전의 수가 자소를 분리한 경우보다 10배 이상 늘어난 것을 고려해보면 자소를 분리한 것이 더 효율적이다.

키워드: 욕설 탐지(abuse detection), 딥러닝(deep learning), 온라인 게임(online game), 콘볼루션 신경망(convolutional neural network)

I. Introduction

게임 산업의 발달과 함께 온라인 PC게임, 모바일 게임 채팅 등에서 욕설을 포함한 언어폭력이 지속적으로 발생하고 있는 것으로 나타났다. 욕설이 심한 경우에 게이머들은 게임에서 이탈하기도 하고, 극단적으로는 모욕죄로 이어져 법적인 문제가 될 때도 있다. 이에 욕설을 제재하는 것은 게임사에서 중요한 문제가 된다. 현재 필터링 시스템은 보수적으로 탐지하는 경향이 있어 변형된 욕설을 간주하지 못하거나, 정상 단어이지만 욕설로 간주하는 경향이 있다. 심지어 한국어는 영어나 다른 언어에 비해 훨씬 까다로운데 이는 한국어가 어순이 중요하지 않고, 교착어이며, 띄어쓰기가 제대로 지켜지지 않기 때문이다.

기존의 욕설 연구는 대부분 영어이며, 기계학습 기반이다. 유튜브에서 악성 댓글 탐지[1], Yahoo의 금융 및 뉴스 기사를 활용한 욕설 분류[2] 모두 Naive Bayes, SVM 및 회귀 모델을 사용하였다. 최근 CNN(Convolutional Neural Network) 또한 텍스트 분석에 효과적인 것으로 나타나 자연어 처리 분야에서 각광받고 있다.[3] 이를 바탕으로 본 연구에서는 CNN을 활용하여 욕설 문장을 분류하는 방법을 연구하고자 한다.

II. Experiment

1. Data

엑스엘게임즈에서 개발한 MMORPG 게임 'ArcheAge'의 정식 서비스를 시작 제공하기 전, 클로즈 베타 서비스 5차의 데이터이다. 실제 접속 인원은 최소한 15만 명 이상이며 2012년 8월 15일 00:33:32~21:13:18의 65,499개 채팅 데이터가 있다. 그 중 금칙어 사전을 기반으로 욕설이라고 간주 된 5,000개를 무작위로 뽑아, 직접 보고 욕설 유무를 태깅하였다. 이 중 1,395개가 욕설로 잘 못 간주된 정상 문장이었다.

2. Features

영어와 달리 한국어는 단어가 띄어쓰기 기준이 아니기 때문에 어절 단위로 토큰을 삼으면 어미와 조사의 변화 때문에 사전 구축에 문제가 생긴다. '나/나는/내가/나를', '먹다/먹는다/먹으니'가 전부 다른 토큰이 되기 때문이다. 또한 한국어는 일상에서 띄어쓰기를 제대로 하기가 어려워 데이터의 불균일로 나타나고, 한국어 자연어처

리의 자동화를 어렵게 하는 한 요인이 된다. 따라서 본 연구에서는 한 글자씩 분리하고, 한글의 자소(초성, 중성, 종성)를 분리하여 실험을 했다. 각각의 사전을 구축하고 각 입력값을 숫자로 벡터화 하여 모델의 인풋으로 사용하였다. 문장을 일정한 길이로 고정하기 위해 0으로 패딩하였다. 예를 들면 [‘시발’]의 경우 [‘ㅅ’,‘ㅣ’,‘ㅂ’,‘ㅏ’,‘ㄹ’] 또는 [‘시’,‘발’]로 분리되고 각각의 입력값은 [6,5,12,3,8,0,0,0.....] 또는 [16,28,0,0,0,0,0....]이 된다.

3. Method

이미지 픽셀 대신 텍스트 데이터 작업에 대한 입력값은 매트릭스로 표현된 문장이다. 매트릭스의 각 행은 하나의 토큰, 일반적인 단어에 해당된다. 이 단어를 적합한 벡터로 수치화하는 임베딩 과정을 거쳤다. 본 모델의 구조는 3개의 콘볼루션 층으로 이루어져 있다. 64차원으로 임베딩 한 후 첫 번째 컨볼루션 층을 통과하면서 16차원으로, 두 번째 컨볼루션 층을 통과하면서 8차원으로 줄어든다. 층의 필터사이즈는 3이다. 이미지 처리에서 필터는 이미지의 지역 필터를 따라 움직이지만 텍스트에서는 매트릭스의 전체 행을 슬라이딩 하는 필터를 사용한다. 첫 번째와 두 번째 콘볼루션 층에서는 필터가 임의의 간격으로 벡터 매트릭스를 움직이며 간단한 문맥적 정보를 수집하고 마지막 컨볼루션 층에서는 분류에 영향을 미치는 욕설단어 정보를 추출하는 역할을 한다. 이 연속된 세 층에서 출력된 매트릭스는 pooling 층으로 주어지는데 주변 값들을 대표하고 가장 큰 값을 선택하는 max-pooling을 사용하여 나온 특징 벡터들을 1/2 크기로 줄인 후, fully-connected 층으로 보내기 위해 1차원으로 변환하는 flatten 과정을 거친다. fully-connected 층의 활성화 함수로 softmax를 채택하였고 각 문장이 정상, 욕설 중 하나의 클래스일 확률을 출력한다. fully-connected를 제외한 모든 층의 활성화 함수는 relu를 사용하였다. Adam optimizer를 사용하였고, L2 정규화를 사용하였다.

III. Results

5000개 중 75%를 훈련 데이터, 25%를 테스트 데이터로 사용하였으며 자소를 분리한 경우 총 사용된 사전의 수는 126개, 한 글자씩 분리한 경우 1065개였다. 자소를 분리한 경우 가장 긴 문장 기준으로 469개의 벡터 길이가 필요하였고, 분리하지 않은 경우에는 241개의 벡터 길이가 필요했다. 제안된 모델로 epoch=20, batch size=32일 때 결과는 [표1]과 같다.

Table 1. 분류 모델의 성능 평가 결과

	욕설 / 정상	
	자소 분리	한 글자 분리
Precision	0.88/0.84	0.90/0.81
Recall	0.95/0.66	0.94/0.75
F1 measure	0.91/0.74	0.92/0.79
Accuracy	0.874	0.886

IV. Conclusions

본 논문에서는 한국어로 된 게임 채팅 데이터를 이용하여 욕설을 판단하는 방법에 대해 연구하였다. 자소를 분리하지 않는 것이 조금 더 정확도를 보였으나 자소를 분리하는 것이 사전의 개수를 10배 이상 줄일 수 있고, 정상 문장 분류 성능에서는 자소를 분리하는 것이 더 좋은 결과를 나타내었다. 향후 연구로는 띄어쓰기나 특수문자를 제거해본 결과를 비교해보고, 다른 딥러닝 기법을 사용하여 연구할 계획이다.

ACKNOWLEDGEMENT

본 연구는 한국연구재단의 이공분야기초연구사업(과제번호 NRF-2017R1D1A3B03036050)의 지원을 받아 수행되었음

REFERENCES

[1] Y. Chen, Y. Zhou, S. Zhu, and H. Xu "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety," In Privacy, Security, Risk and Trust, IEEE, p. 71-80, 2012

[2] Nobata, C.; Tetreault, J.; Thomas, A.; Mehdad, Y.; and Chang, Y. "Abusive Language Detection in Online User Content," In WWW, p. 145-153, 2016

[3] Y. Kim “*Convolutional Neural Networks for Sentence Classification*” In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), p. 1746-1751

[4] Nemanja Djuric at el, "Hate Speech Detection with Comment Embeddings." In WWW, p. 29-30, 2015.

[5] S. Seo and S. Cho, “Transfer Learning Method for Solving Imbalance Data of Abusive Sentence Classification” Journal of KIISE, Vol. 44, No. 12, p. 1275-1281, 2017