크라우드소싱 기반의 딥러닝 분류 알고리즘을 이용한 댓글 분류 시스템

박희지, 하지민, 박혜림, 강정호*
배화여자대학교 정보보호과
(pheeji, basejm07, hyae_lim, kangsamm)@naver.com

Comment Classification System using Deep Learning Classification Algorithm based on Crowdsourcing

Heeji Park, Jimin Ha, Hyaelim Park, Jungho Kang*
Dept. of Information Security, Baewha Women's University

요 약

뉴스, SNS 등의 인터넷 댓글은 익명으로 의견을 자유롭게 개진할 수 있는 반면 댓글의 익명성을 악용하여 비방이나 험담을 하는 악성 댓글이 여러 분야에서 사회적 문제가 되고 있다. 해당 문제를 해결하기 위해 AI를 활용한 댓글 분류 알고리즘을 개발하려는 많은 노력들이 이루어지고 있지만, 댓글 분류 모델에 사용되는 AI는 오버피팅의 문제로 인해 댓글 분류에 대한 정확도가 떨어지는 문제점을 가지고 있다. 이에 본 연구에서는 크라우드소싱을 활용하여 오버피팅으로 인한 악성 댓글 분류 및 판단 정확도 저하 문제를 개선한 크라우드소싱 기반 딥러닝 분류 알고리즘(Deep Learning Classification Algorithm Based on Crowdsourcing: DCAC)과 해당 알고리즘을 사용한 시스템을 제안한다. 또한, 실험을 통해 오버피팅으로 낮아진 판단 정확도를 증가시키는 데 제안된 방법이 도움이 되는 것을 확인하였다.

1. 서론

댓글은 인터넷 사용자가 올린 게시물에 대하여 다른 사용자들이 답하는 글이다.[1] 댓글은 익명으로 자유로운 의사 표현이 가능하지만 이를 악용하는 악성 댓글은 인신공격, 혐오, 조롱 및 허위사실 유포 등의 문제를 야기하고 있다. 실예로 유명 연예인들의 자살 등의 피해가 발생하였고 연예뿐만 아니라 스포츠, 정치와 같은 분야에서도 악성 댓글에 대한 피해가 확산되면서 사회문제로 대두되었다. 이를 해결하기 위해 AI를 활용한 악성 댓글 탐지 방법이 제안되고 있지만 AI를 사용함으로써 항상 오버피팅의 문제를 가지고 있다. 오버피팅이란 입력 데이터를 과도하게 학습하여 발생하는 문제로 새로운 데이터 학습 시 판단의 유연성을 저하시키고 정확도를 떨어뜨린다.[2]

이에 본 논문에서는 크라우드소싱을 활용하여 오버피팅으로 인한 악성 댓글 분류 및 판단 정확도 저하 문제를 개선하는 크라우드소싱 기반 딥러닝 분류 알고리즘(Deep Learning Classification Algorithm Based on Crowdsourcing: DCAC) 과 해당 알고리즘을 활용한 댓글 분류 시스템을 제안한다.

2. 관련 연구

2.1 KoBERT

BERT는 Natural Language Processing(NLP) 작업을 위해 구글에서 개발한 pre-training 모델이다.[3] BERT는 33억 개의 단어가 외국어를 중심으로 학습되어 있어 외국어 정확도는 높지만 한국어에 대한 정확도가 저하되는 문제가 있다. KoBERT[4]는 BERT의 문제점인 한국어 정확도 개선을 위해 SKTBrain에서 개발한 모델로 위키피디아, 뉴스 등에서 모은 수백만 개의 한국어 문장 corpus를 학습한다.

2.2 Crowdsourcing

Crowdsourcing은 Crowd와 Outsourcing의 합성

* 교신저자 : 배화여자대학교 정보보호과 교수

어로 200년 제프 하우가 아웃소싱의 대안으로 만들어낸 개념이다. 크라우드소싱은 아웃소싱처럼 해당 분야의 전문가에서 하청을 주는 것이 아니라, 대중의 참여를 통해 아웃소싱하는 프로세스를 말한다.[5] 크라우드소싱의 대표적인 예로는 위키피디아가 있다. 위키피디아는 사용자가 직접 참여하는 무료 인터넷 백과사전으로 일반 사용자에 의해 내용의 생성 및 수정이 된다는 특징이 있다.[6]

2.3 SVM 악성 댓글 탐지 방법

감성 분석과 SupportVector Machine(SVM)을 이용한 연구에서는 감정 사전으로 악성 지수를 계산하고 SVM으로 악성 댓글 여부를 판단할 계산식을 도출하여 악성 댓글을 탐지하는 기법을 제시하였다.[7]

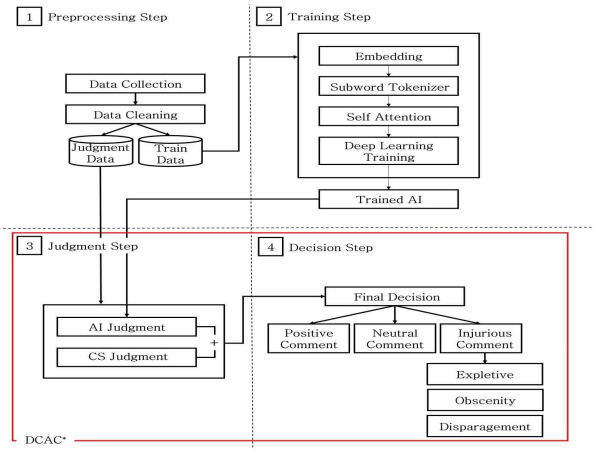
2.4 CNN 악성 댓글 탐지 방법

Highway Network 기반 Convolution Neural Network 모델링 연구에서는 CNN, Highway Network, Out of Vocabulary 사전학습 임베딩을 활용하여 댓글을 6가지 클래스로 분류하고 문장 맥락을 고려하여 학습시킨 모델을 제안하였다.[8]

3. 제안 시스템

본 논문에서는 Deep Learning Classification Algorithm Based on Crowdsourcing (DCAC)과 해당 알고리즘을 기반으로 한 댓글 분류 시스템을 제안한다.

3.1 DCAC 기반 댓글 분류 시스템 구조



(그림 1) DCAC를 기반으로 한 댓글 분류 시스템

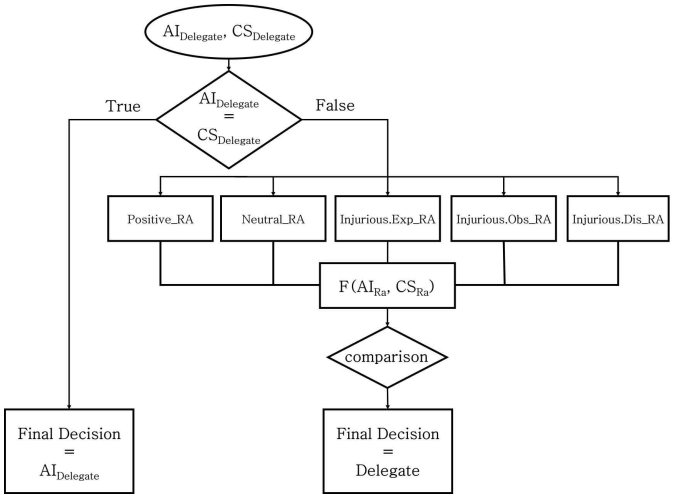
(그림 1)의 댓글 분류 시스템은 Preprocessing Step, Training Step, Judgment Step과 Decision Step으로 구성되어 있다. 첫 번째 단계에서는 데이터를 수집하고 Data Cleaning을 하여 Training Data와 Judgment Data로 나눈다. 두 번째 단계에서는 딥러닝 학습 정확도를 높이기 위해 Training Data를 가공하는 Embedding, Self Attention, Subword Tokenizer 과정을 거친 후 딥러닝 모델에 학습시킨다. 세 번째 단계에서는 Judgment Data와 Trained AI를 이용해 AI Judgment를 진행하고 사람들을 대상으로 CS Judgment 을 진행한다. 마지막으로, 네 번째 단계에서는 $AI_{Delegate}$과 $CS_{Delegate}$을 비교하여 둘 중 큰 값을 Final Decision으로 한다. 댓글 분류 시스템 중 Judgment Step과 Decision Step 합쳐서 본 논문에서 DCAC로 명명하여 제안하였다. DCAC에 대한 자세한 설명은 3.2에서 다루기로 한다.

다음의 <표 1>은 DCAC에서 사용하는 대표적인 기호를 정리한 표이다.

<표 1> DCAC 구조에서 사용하는 기호표

기호	설명
CA	댓글의 Classification을 나타내는 기호이다. Positive_CA: 긍정적인 댓글 Neutral_CA: 중립적인 댓글 Injurious.Exp_CA: 유해한 댓글 중 Expletive (비속) Injurious.Obs_CA: 유해한 댓글 중 Obscenity (음란) Injurious.Dis_CA: 유해한 댓글 중 Disparagement (비하)
RA	AI 또는 크라우드소싱으로 판단된 댓글의 분류 별 차지하는 Ratio(%)이다. Positive_RA: 긍정적인 댓글 RA Neutral_RA: 중립적인 댓글 RA Injurious.Exp_RA: 유해한 댓글 중 Expletive RA Injurious.Obs_RA: 유해한 댓글 중 Obscenity RA Injurious.Dis_RA: 유해한 댓글 중 Disparagement RA
Delegate	판단된 RA 중 가장 큰 값을 갖는 CA를 의미한다. $AI_{Delegate}$: AI에서 판단된 RA 값 중 가장 큰 값의 분류 $CS_{Delegate}$: CS에서 판단된 RA 값 중 가장 큰 값의 분류
F(AI_{RA}, CS_{RA})	최종 판단을 위한 CA 각각의 값을 구하는 공식으로 AI_{RA}와 CS_{RA}를 사용한다. - 식: $\frac{(AI_{RA}\times X)+(CS_{RA}\times Y)}{X+Y}$ (X, Y는 정수이고 각각은 AI_{RA}와 CS_{RA}의 비율을 의미한다.)
Z_c	정확도 판단을 위해 정의한 댓글 분류 정답들의 집합이다.

3.2 DCAC 세부 구조



(그림 2) DCAC 세부 구조

(그림 2)는 (그림 1)의 3, 4에 해당하는 DCAC 세부 구조를 나타낸 것이고 (그림 2)에 대한 설명은 다음과 같다.

True) $AI_{Delegate}$와 $CS_{Delegate}$이 같을 경우 $AI_{Delegate}$를 Final Decision으로 결정한다.
False) $AI_{Delegate}$와 $CS_{Delegate}$이 다를 경우 아래와 같이 진행한다.
a) AI_{RA}와 CS_{RA} 비가 X : Y라 할 때, $F(AI_{RA}, CS_{RA})$를 활용하여 모든 RA를 계산한다.
b) 계산된 RA들을 비교하고 Delegate를 Final Decision으로 한다.

크라우드소싱 판단 과정을 추가한 DCAC는 기존 알고리즘들에 비해 오버피팅의 문제점으로부터 자유로울 수 있다. DCAC의 이점을 활용하여 기존 AI를 활용한 댓글 분류 모델보다 악성 댓글을 좀 더 정교하게 판별하고, 이를 통하여 바른 댓글 환경 조성이 가능하다.

4. 실험

4.1 실험 환경

개발 환경은 Python 3.7 언어 기반 Colaboratory 개발 브라우저를 이용하고 딥러닝은 여러 모델 중 KoBERT를 사용한다.

실험의 데이터는 크롤링을 활용하여 한겨레, 중앙일보 정치면의 기사 댓글을 수집하고 그중 1,200개를 임의로 선정한다. 선정된 댓글은 Training Data 1080개와 Judgment Data 120개로 나눈다. 이때 비율은 오버피팅 에러율 비교 시 주로 사용되는 9 : 1이다.[5] 직접 1,200개의 데이터를 분석하여 데이터 각각의 Z_C을 정하고 CS Judgment에서 사용되는 Judgment Data는 약 26명에게 설문을 받아 수집했다. 수집된 크라우드소싱 값을 확인한 결과 제대로 작성하지 않거나 동일한 분류로 선택하는 경우가 존재했다. 이 경우 크라우드소싱 정확도가 제대로 나오지 않기 때문에 해당 데이터를 제거했다.

4.2 실험

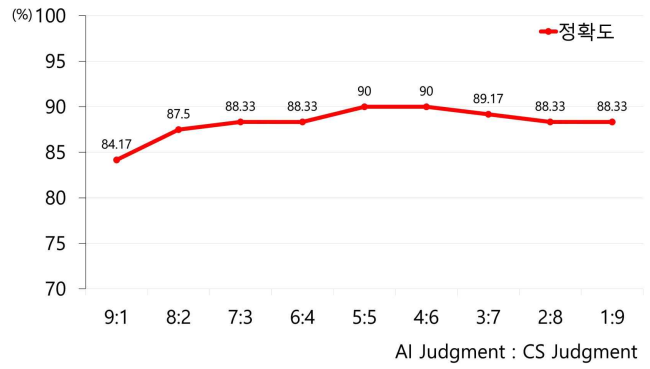
실험은 CS Judgment가 AI Judgment보다 우수하다는 것을 전제로 진행하였지만, 악의적으로 크라우드소싱을 잘못 판단하는 예외가 발생할 수 있다는 점을 감안하였다. 따라서 AI_{RA}와 CS_{RA} 비를 X : Y (정수 X, Y, X+Y=10)로 나누어 다양한 형태의 정확도를 비교하고 그중 정확도가 가장 높은 계수를 $F(AI_{RA}, CS_{RA})$의 비율로 결정하였다. 이때 악성 댓글 분류 및 판단 정확도는 $\frac{Z_C\text{와 일치하는 개수}}{\text{전체 댓글 개수}} \times 100$으로 계산하였다.

4.3 실험 결과

<표 2>는 AI · CS Judgment Decision과 Z_C를 비교한 결과이다. AI Judgment는 Z_C와 총 95개의 댓글 분류가 일치하여 79.17%의 정확도를 보여주었고, CS Judgment는 Z_C와 총 106개의 댓글 분류가 일치하여 88.33%의 정확도를 보여주었다.

<표 2> AI · CS Judgment Decision과 Zc 비교표

	Positive Comment	Neutral Comment	Injurious Comment			Total	Accuracy(%)
			Expletive	Obscenity	Disparagement		
Z_C의 개수	10	81	11	4	14	120	100
Z_C와 일치하는 AI Judgment 개수	5	70	8	4	8	95	79.17
Z_C와 일치하는 CS Judgment 개수	9	75	10	4	9	106	88.33



(그림 3) 판단 비율에 따른 정확도 비교 그래프

(그림 3)은 AI Judgment와 CS Judgment 비율에 따라 9 : 1부터 1 : 9까지의 DCAC를 활용한 정확도를 나타낸 그래프이다. 10 : 0과 0 : 10은 각각 AI Judgment 시 정확도와 CS Judgment 시 정확도를 의미하기 때문에 이를 제외하고 그래프에 나타내었다. 제안한 알고리즘을 활용한 정확도는 9 : 1에서 점차 증가하여 4 : 6과 5 : 5에서 가장 높았고 그 후로 점점 감소하는 것을 알 수 있다.

다음의 <표 3>은 DCAC를 사용한 댓글 분류 시스템의 정확도와 AI, 크라우드소싱 정확도를 비롯한 다른 알고리즘을 사용한 댓글 분류 시스템의 정확도를 비교한다. DCAC 악성 댓글 분류 및 판단 정확도 실험 중 가장 높은 정확도를 가지는 4 : 6과 5 : 5의 값은 90%로 CS Judgment 정확도보다 1.67% 증가하였고, AI Judgment 정확도보다 10.83%로 정확도가 눈에 띄게 향상된 것을 확인하였다. 동일한 환경에서 실험을 한 것은 아니기 때문에 절대적인 비교는 어렵지만 상대적인 비교를 위해 다른 논문에서 실험한 결과를 비교, 제시하였다. 실험을 통해 기존 댓글 분류 알고리즘보다 DCAC의 댓글 분류의 정확도가 향상된 것을 확인할 수 있다.

<표 3> 알고리즘별 댓글 분류에 대한 판단 정확도 비교표

실험에 사용한 알고리즘	알고리즘별 댓글 분류에 대한 정확도(%)
AI	79.17
SVM 알고리즘[7]	87.80
CNN 알고리즘[8]	67.49
크라우드소싱	88.33
DCAC	90.00

5. 결론 및 향후 연구

위 실험을 통해 DCAC를 활용한 댓글 분류 시스템은 다른 알고리즘보다 댓글 분류에 있어 정확도가 높다는 것을 증명하였다. DCAC는 한글 댓글에서도 정확한 판단을 수행하였다는 점, 댓글 분류에 사람들의 의견을 포함시켜 오버피팅으로 인해 파악하지 못한 악성 댓글을 추가적으로 잡아냈다는 점에서 의미가 있다. 이는 오버피팅 문제 해결의 새로운 방향성을 제시하고 선한 댓글 문화 형성에 도움이 될 것으로 판단한다.

본 연구는 악의적으로 크라우드소싱 할 경우 기존 알고리즘보다 정확도가 떨어질 수 있다는 한계점이 존재한다. 따라서 다양한 상황에 대한 실험 기준과 DCAC 알고리즘을 회피하고자 하는 악의적인 판단을 제거하기 위한 방안, 크라우드소싱 기준에 대한 연구가 필요하다.

참고문헌

[1] Park, Jihyun, et al. "Solving the Abusive Comments Problem through ML-based Visualization." Journal of Digital Contents Society 21.4, 771-779, 2020.

[2] Park, Sungsoo, et al. "Motion Monitoring using Mask R-CNN for Articulation Disease Management." Journal of the Korea Convergence Society 10.3, 1-6, 2019.

[3] Devlin, Jacob, et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding” NAACL, 2019, 4171-4186.

[4] https://github.com/SKTBrain/KoBERT

[5] Kim, Jaeil, et al. "A Case Study of Crowdsourcing Platform : Focused on Bros& Company." Korea Business Review 23.1, 29-56, 2019.

[6] Shim, Wonsik, et al. "Analysis of Wikipedia Citations in Peer-Reviewed Journal Articles." Journal of the Korean Society for Library and Information Science 47.2, 247-264, 2013.

[7] Hong, Jinju, et al. "A Malicious Comments Detection Technique on the Internet using Sentiment Analysis and SVM." Journal of the Korea Institute of Information and Communication Engineering 20.2, 260-267, 2016.

[8] Lee, Hyunsang, et al. "A Study on the Toxic Comments Classification Using CNN Modeling with Highway Network and OOV Process." The Journal of Information Systems 29.3, 103-117, 2020.