

BAG: "BERT-Attention GNN" HTML 하이브리드 모델을 활용한 불법사이트 구조 탐지*

이세찬⁰¹, 허민석², 민무홍³
성균관대학교 컴퓨터교육과^{1,3}, 성균관대학교 실감미디어공학과²
{chan1031, alexhur3535, iceo}@g.skku.edu

BAG: A hybrid "BERT-Attention GNN" model for Illegal Site Detection from HTML

Sechan Lee⁰¹, Minseok Hur², Moohong Min³
SKKU Computer Education^{1,3}, SKKU Immersive Multimedia Engineering²

요 약

본 연구에서는 HTML 태그의 텍스트와 구조 정보를 결합하여 불법 도박 사이트를 효과적으로 탐지할 수 있는 하이브리드 모델 BAG (BERT-Attention GNN)을 제안한다. 제안된 모델은 HTML 각 태그의 텍스트를 BERT로 분석해 불법성 점수를 산출하고, 이를 GAT의 attention 연산에 반영하여 중요 노드의 정보를 강화한다. 또한, Weighted Pooling을 도입해 불법 키워드를 담은 노드가 많을수록 판정결과에 기여하도록 설정하였다. 불법 도박 사이트와 정상 사이트의 HTML 데이터를 활용한 실험 결과, 훈련 정확도 89%와 추론 성능 정확도 90%의 우수한 성능을 확인하였으며, 추가 제거 연구를 통해 BERT-Attention 연산과 텍스트 감정 분석이 HTML 분류 성능 향상에 기여하고 있음을 입증하였다.

1. 서 론

Graph Neural Network (GNN)는 그래프 자료구조 학습에 효과적인 모델로 활용되고 있다. HTML 데이터 셋의 분류 학습 시, 텍스트와 DOM Tree 그래프 구조를 동시에 학습하는 것이 가장 효과적이거나 GNN 모델로는 텍스트 정보를 학습 할 수가 없기에 정확한 분류에 한계가 존재한다. 특히, 불법 사이트 분류처럼 텍스트가 분류 결과에 중요한 정보로 사용되는 경우, '텍스트 감정 분석' 기능이 없어 GNN을 활용하기에 효율적이지 못하다는 단점이 있다 [1]. 이에 본 연구에서는 BERT를 조합하여 HTML 태그의 텍스트를 분류한 다음, 텍스트 분류 score를 Graph Attention Network (GAT)의 attention 점수에 반영하여, 효과적으로 HTML을 분류할 수 있는 특수한 하이브리드 모델인 'BERT-Attention GNN (BAG)'를 소개하고, 위 모델이 불법 도박사이트 탐지 도메인에 적용하여 효율적으로 활용될 수 있는지를 실험한다.

2. 이론적 배경

2.1 BERT

BERT는 트랜스포머의 attention 기반 인코더 구조를 활용하여 입력 텍스트의 양방향 문맥 정보를 동시에 학습하는 언어 모델로, Masked Language Modeling (MLM)과 Next Sentence Prediction (NSP) 태스크를 수행하여 대규모 말뭉치로부터 단어 간 의미 관계와 문장 간 연결성을 파악하는 데 특화된 모델이다 [2].

본 연구에서는 BERT 모델을 활용하여, 불법 도박사이트에서 사용되는

문장을 학습하고, 분류 결과 점수를 GAT의 self-attention에 반영하여 GNN이 효과적으로 HTML을 분류하게 한다.

2.2 GAT

GAT는 그래프 데이터에서 각 노드가 이웃 노드들의 정보를 집계할 때, 자신과 연결된 노드들의 상대적 중요도를 self-attention 메커니즘을 통해 동적으로 중요도를 계산하는 GNN 모델이다 [3]. self-attention은 자연어처리에서 주로 활용되는 기법이나, GAT는 각 이웃 노드의 중요도를 파악하여 그래프의 핵심구조를 파악하는 데 활용된다 [4]. 본 논문의 BAG에서는 불법 사이트의 HTML 구조를 효과적으로 분류하기 위해 각 태그의 BERT 텍스트 분류 점수를 산출하여 GAT의 self-attention 계산에 포함한다.

2.3 불법 사이트의 HTML 구조 특징



그림 1 불법사이트 유사도의 예시

불법 사이트는 지속적인 차단과 단속을 피하고자, 그림 1에서 확인할 수 있듯이, 서로 다른 사이트이지만 비슷한 HTML 구조를 공유한다는 특징을 보유하고 있다 [5]. BAG는 이러한 유사한 HTML 구조를 파악하기 위해 특화된 모델이며, 본 연구에서는 가장 많이 악용되고 있는 '불법 도박 사이트'를 모델에 학습한다.

* 이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구 결과 (No.RS-2024-00439139, 사이버 위기 대응 능력 및 복원력 시험평가 도구 개발이며, 메타버스융합대학원의 연구 결과로 수행되었음 (IITP-2025-RS-2023-00254129))

3. BERT-Attention GNN 모델 소개

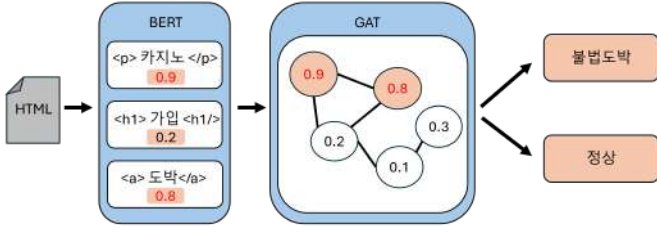


그림 2 BAG의 전체 구조도

BAG는 2-Stage 모델로 전체 구조도는 그림 2와 같다. HTML 태그의 텍스트를 파인튜닝된 BERT로 텍스트 감정 분석하여, 각태그를 0~1 사이의 점수로 측정한다. 이후 측정된 점수를 GAT의 self-attention에 반영하여 텍스트 분류에 기반한 HTML 구조를 학습한다. 기존의 Graph-BERT처럼 GNN과 BERT를 융합한 하이브리드 모델이 존재하지만 [6], 불법사이트 분류를 위한 키워드 텍스트에 집중한다는 점에서 그 차이가 존재한다.

3.1 Bert-Attention 연산

HTML 파일이 입력되면 DOM Tree의 그래프 구조로 변환한 다음, 불법 도박 키워드로 학습된 BERT가 각 태그 텍스트의 불법 도박 키워드 여부를 0~1 사이의 점수로 표현한다. 이후 GAT에서 효과적인 학습을 위해 해당 점수를 활용한 self-attention 연산을 한다.

$$\alpha_{i,j}^{final} = \lambda \gamma_j \frac{\exp(\text{LeakyReLU}(a_{src}^T z_i + a_{dst}^T z_i))}{\sum_{k \in N(i)} \exp(\text{LeakyReLU}(a_{src}^T z_i + a_{dst}^T z_i))}$$

수식 1 Bert-Attention 계산 수식

Bert-Attention의 수학적 표현은 수식 1과 같다. 수식 1은 노드 i 에서 소스 노드 j 로부터의 어텐션 계수이고 $z_i = Wx_i$ 는 노드 i 의 선형 변환된 임베딩이며, a_{src} 와 a_{dst} 는 각각 소스와 타겟 노드에 대응하는 attention 파라미터를 의미한다. 여기서 BERT에서 계산된 태그별 텍스트 분류 점수인 r_j 를 곱하여 self-attention 계산에 포함한다. 여기서 λ 는 r_j 의 영향도를 조정할 수 있는 하이퍼 파라미터를 의미한다.

3.2 Weighted Pooling Layer

하지만, 이 경우 불법 도박사이트와 동일한 사이트 구조를 가지는 정상 사이트를 잘못 분류하는 경우가 존재할 수 있다. 이를 방지하기 위해 모델의 마지막 Layer에 Weighted Pooling Layer를 추가하여 불법 도박 점수가 높은 노드가 최종 결과에 더 기여하도록 설정하였다. 이를 표현한 수식 2를 살펴보면, 최종 계산된 각 그래프의 노드 벡터들 (h_{graph})에 softmax 연산을 활용하여 각 노드벡터에 불법도박키워드가 있을 경우 해당 노드의 값을 증가시킨다 즉, 이 방법을 통해 웹 사이트의 구조가 동일하더라도 불법 도박으로 판정된 노드의 텍스트가 많을수록 불법 도박사이트로 분류하도록 유도할 수 있다.

$$h_{graph} = \sum_{i=1}^N \frac{\exp(T \cdot s_i)}{\sum_{j=1}^N \exp(T \cdot s_j)} x_i$$

수식 2 Pooling 연산 수식화

이후 Classifier layer를 거치면 최종적으로 이진분류를 통해 점수가 0.5 이상이면, 불법 도박사이트로 분류하도록 모델을 제작하였다.

3.3 커스텀 GAT 모델 구조

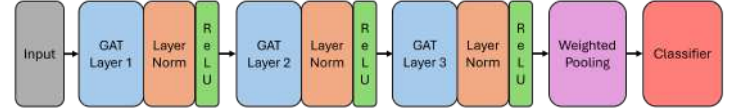


그림 3 BAG에 활용된 커스텀 GAT 모델 구조도

최종적으로 본 모델에 사용한 커스텀 GAT 모델 구조는 그림 3과 같다. 모델은 3개의 GAT Layer와 각 Layer는 Layer Norm, ReLU 활성화 함수를 거치도록 구성하였으며, 3.2에서 언급한 Pooling Layer를 거치고 난 후 이진 분류하도록 모델을 설계하였다.

4. 실험 진행 및 결과

4.1 실험 환경 및 모델 설정

실험에는 RTX 3090 GPU 2개를 사용하였으며, PyTorch 프레임워크를 활용하여 BAG 모델을 제작하였다. BERT 모델은 한국어 처리에 특화되어있는 'Kc-BERT [7]'를 파인튜닝하여 0.991의 정확도를 가지도록 훈련하였다. 이후 GAT 모델은 사전 학습 없이 데이터 셋을 수집하여 학습을 진행하였으며 기본적으로 3000 epoch를 설정하고 30 epoch 동안 모델 성능이 개선되지 않을 경우, 조기 종료되도록 설정하였다.

4.2 데이터 셋

BERT 훈련을 위한 텍스트 데이터 셋은 불법 사이트의 특성상 데이터 셋이 부족하다는 한계로 인해 GPT-4o API를 활용하여 데이터를 수집하였다. 우선, 불법 도박 사이트에 자주 쓰이는 단어들 ('바카라', '카지노', '콤프' 등)를 260개 준비한 다음 GPT-4o API를 활용하여 불법 도박사이트에서 사용 할 만한 문장들을 생성하였다. 이후 대조군인 정상 사이트 텍스트 데이터는 [8]의 한국어 데이터 셋에서 일부를 추출하여 사용하였다. 최종적으로 불법 도박, 정상 사이트를 합쳐 총 10만 개의 데이터를 준비하였다. GAT 훈련을 위한 HTML 데이터 셋은 크롤링을 활용하여 200개의 불법 도박사이트를 수집하고, 대조를 위한 정상 사이트의 경우 '정부 기관', '대학', '공공기관' 등으로 범주를 나누어 총 200개의 HTML 데이터 셋을 수집하였으며, 이후 HTML을 DOM Tree로 구조로 변환하여 GNN 이 학습할 수 있도록 하였다.

4.3 훈련 결과

성능 지표	값
Accuracy	0.89
Precision	0.82
F1 Score	0.90
AUC	0.95

표 1 BAG의 훈련 성능 지표

Kc-BERT와 커스텀 GAT 모델을 조합한 BAG 모델의 학습 결과는 표 1과 같다. 주어진 지표를 살펴보면, 모델은 0.89의 Accuracy로 우수한 성능을 보이고 있다. 특히, AUC 지표가 높으며, 이는 모델이 불법 도박사이트 클래스와 정상 사이트 클래스를 구분하는 능력이 우수함을 입증한다.

4.4 추론 성능 실험

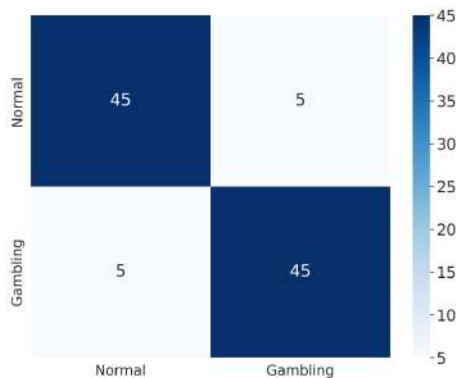


그림 4 실험 결과 혼동행렬

훈련된 모델이 실제 불법 도박사이트를 잘 분류할 수 있는지 실험하기 위해 다음과 같이 데이터 셋을 수집하였다. 구글 검색, SNS 홍보 게시물, KISA SMS 스팸 메시지에서 불법도박사이트 50개를 추출하고, 이후 50개의 정상 사이트를 수집하여 추론 성능을 테스트 하였다. 그림 4에서 확인할 수 있듯이, 추론성능은 90%의 높은 정확도를 보임을 확인할 수 있었다. 잘못 분류한 정상사이트는 특징을 추출할 수 없이 메인 페이지가 로그인 페이지만 존재하는 경우들이었다. 이를 분류하기 위해서는 메인 페이지의 이미지를 분류하는 모델을 개발하여 BAG와 앙상블 모델로 활용한다면 높은 정확도를 보일 것이라 생각한다.

5. 제거 연구

성능 지표	값
Accuracy	0.75
Precision	0.73
F1 Score	0.75
AUC	0.75

표 2 BERT를 제거한 모델의 훈련결과

BAG 모델에서 BERT 모델 추론 부분을 제외했을 때의 모델의 성능을 측정하여 BAG의 BERT-Attention과 텍스트 감정분석이 사이트 분류에 얼마나 중요한 역할을 하는지를 실험하였다. 실험은 이전과 동일한 환경에서 BERT만 제외하여 GAT의 기존 Attention연산을 활용하여 HTML의 그래프 구조에만 집중하도록 구성하였다. 표 2의 실험 결과를 살펴보면, 기존 BAG 모델보다 Accuracy가 0.14 감소한 것을 확인할 수 있었다. 그렇지만 성능이 크게 하락하지는 않았는데, 이는 2.3에서 언급한 불법 도박사이트의 유사한 HTML 그래프 구조만으로도 분류가 가능하기 때문으로 보인다. 하지만 불법 사이트의 텍스트 감정 분석과 BERT-Attention 연산이 모델의 성능 향상에 큰 기여를 하고 있는 점 또한 확인할 수 있었다.

6. 결론 및 한계

본 연구는 BAG라는 텍스트 분류에 특화된 BERT와 HTML DOM Tree 분석에 특화된 GAT를 융합한 하이브리드 모델을 소개하고, 본 모델이 효과적으로 HTML을 분석하여 불법 도박사이트 분류라는 특수한 도메인에 효과적으로 활용할 수 있음을 증명하였다. 이에 더 나아가면 불법 도박사이트 이외의 사이트에서도 효과적으로 적용할 수 있을 것이라 판단된다. 불법 사이트 탐지라는 특수한 상황으로 인해 수집할 수 있는 데이터 셋 수집에 한계가 있었으나, 자동화된 크롤링 서비스와 연계하여 데이터 셋을 더 수집할 수 있다면, 지금보다 더 높은 성능을 가질 수 있을 것이라 기대한다.

참고 문헌

- [1] Abedi Rad et al, "A survey of sentiment analysis methods based on graph neural network. *International Journal of Data Science and Analytics*", 19(3), 389-417, 2025.
- [2] Jacob Devlin et al, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. of NAAACL-HLT*, 2019.
- [3] Petar Veličković et al, "Graph Attention Networks," in *Proc. of ICLR*, 2018.
- [4] Ashish Vaswani et al, "Attention Is All You Need," in *Proc. of NeurIPS*, 2017.
- [5] Moohong Min et al, "Illegal Online Gambling Site Detection using Multiple Resource-Oriented Machine Learning," *Journal of Gambling Studies*, vol. 40, pp. 2237-2255, 2024.
- [6] J zhang et al, "Graph-Bert: Only Attention is Needed for Learning Graph Representations", arXiv:2001.05140, 2020.
- [7] 이준범, "KcBERT: 한국어 댓글로 학습한 BERT", 한글 및 한국어 정보처리 학술대회, 2020.
- [8] 한국지능정보사회진흥원(NIA), "대규모 웹 데이터 기반 한국어 말뭉치," AI-Hub, <https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=624>.

부록

본 논문에서 활용된 코드는 오픈소스로 제작되었으며, <https://github.com/chan1031/BAG> 에서 확인 가능하다.