

소셜 네트워크에서의 딥페이크 콘텐츠 전파정도 예측

정서윤, 강채원, 우사이먼 성일, 한진영*
성균관대학교

jsyoon0503@g.skku.edu, codnjs3@g.skku.edu, swoo@g.skku.edu, *jinyoungghan@skku.edu

Predicting the Virality of Deepfake Content on Social Networks

Seoyoon Jeong, Chaewon Kang, Simon S. Woo, Jinyoung Han*
SungKyunKwan University

요약

딥페이크 기술의 발전으로 인한 사회적 문제, 특히 딥페이크 포르노 콘텐츠의 확산이 심각한 이슈로 부각되고 있다. 기존의 콘텐츠 확산 예측 연구는 주로 텍스트 기반의 가짜 뉴스에 초점을 맞추고 있어, 시청각적 특성을 가진 딥페이크 콘텐츠의 확산 예측에는 한계가 있다. 본 연구는 딥페이크 콘텐츠의 전파 양상을 분석하고 예측하기 위해, 딥페이크 포르노 콘텐츠를 수집하고 이를 바탕으로 전파 트리를 모델링하였다. 네트워크 분석을 위해 위너 인덱스를 사용하여 딥페이크 콘텐츠의 확산 구조를 정량화하고, CNN 기반의 예측 모델을 통해 확산 가능성이 높은 콘텐츠를 식별하였다. 실험 결과, 위너 인덱스 상위 20%에 해당하는 딥페이크 콘텐츠를 효과적으로 예측할 수 있었으며, 이는 딥페이크 콘텐츠 확산의 조기 대응을 위한 중요한 기준을 제공한다.

1. 서론

딥페이크(Deepfake)는 인공지능 기술을 활용하여 영상 또는 사진 속 인물의 얼굴을 조작하는 기술로, 최근 이미지 처리 및 생성 기술의 발전은 딥페이크 콘텐츠 생성의 가능성을 크게 확장시켰다. 그러나 딥페이크를 활용한 정치인 및 유명인 대상 허위 정보 생성과 같이 디지털 범죄 사례가 증가하고 있으며 [1], 딥페이크 포르노(Deepfake pornography)로 인한 피해가 가장 큰 비중을 차지한다 [2]. 딥페이크 기술의 발전 및 접근성 향상으로, 방대한 양의 딥페이크 콘텐츠는 소셜 네트워크 내에서 다발적으로 공유되고 급속히 확산되어 이를 통제하거나 억제하는 것은 상당히 까다로운 문제이다 [3]. 이로 인한 사회적 피해를 예방하기 위해서는 딥페이크 전파양상 분석 및 예측을 통한 대응이 필수적이다.

소셜 네트워크 내에서 콘텐츠 확산을 예측하기 위한 노력은 다수의 연구자들에 의해 꾸준히 이루어져왔다. Islam et al. [4]는 콘텐츠 전파정도(virality), 다음 전파대상 및 전파시점을 예측하는 연구를 수행했다. Murayama et al. [5]은 가짜뉴스의 초기 확산 양상은 일반적인 뉴스와 유사하나, 이후 해당 내용의 진위여부가 드러나며 기존과는 상이한 양상을 보임을 입증했다. 그러나 딥페이크 기술로 생성된 콘텐츠의 전파 양상에 초점을 맞춘 연구는 전무하며, 일반 콘텐츠와 딥페이크 콘텐츠의 특성 차이로 인해 기존 연구 결과는 딥페이크 콘텐츠 전파에 적용이 어렵다.

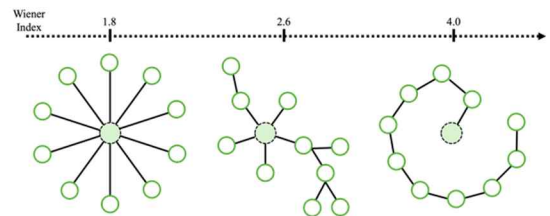
본 연구에서는 딥페이크 전파 양상 분석을 위한 양질의 데이터를 수집 및 전처리하고, 딥페이크 콘텐츠의 전파 정도를 파악하기 위하여 ‘위너 인덱스 (Wiener Index)’ [6]를 통해 전파의 구조적 특성을 정량화 한다. 위너 인덱스는 네트워크 내 노드 (Node) 쌍의 평균 홉 (Hop) 수로 정의되며, 이는 노드 간의 연결성을 평가하는 지표로 사용된다. 위너 인덱스 ($W I$)는 식 1과 같이 표현되며, 여기서 U 는 전체 노드 집합을 나타내고, d_{ij} 는 노드 i 와 노드 j 사이의 최단 경로 거리를 의미한다. 위너 인덱스 값이 낮을수록 콘텐츠 전파 양상이 방사형에 가까움을 의미한다 (그림 1). 실험을 통해 위너 인덱스 상위 20%에 해당하는 딥페이크 게시물을 식별하는 것이 확산 가능성이

높은 핵심 콘텐츠를 가장 효과적으로 식별하고, 이를 통해 악의적 딥페이크 콘텐츠의 조기 통제 가능성을 제시한다.

식 1. 위너 인덱스(Wiener Index) 계산식

$$W I = \frac{1}{|U|(|U|-1)} \sum_{i,j \in U, i \neq j} d_{ij}$$

그림 2. 위너 인덱스(Wiener Index)에 따른 전파 양상



II. 본론

2.1 Data Collection and Processing

딥페이크 전파 데이터셋이 전무함에 따라, 본 연구에서는 소셜 네트워크 내에서 이루어진 딥페이크 전파 데이터를 크롤링한 뒤 분석 및 실험한다. 딥페이크 기반 디지털 범죄 사례의 대부분은 딥페이크 포르노에 의한 것이므로 [2], 딥페이크 포르노 콘텐츠를 데이터 수집 대상으로 한다. 대부분의 기존 전파 연구는 게시글의 공유 정보를 얻을 수 있는 트위터 데이터를 사용하였으나 [7], 2023년 7월부터 Twitter API 사용이 제한됨¹에 따라 트위터와 유사한 방식으로 게시글 업로드 및 공유가 이루어지는 소셜 네트워크인 텀블러(Tumblr)²를 데이터 수집 플랫폼으로 선택하였다.

데이터는 2023년 11월 8일부터 12월 21일까지 약 43일에 걸쳐 수집되었으며, 최종적으로 5,261 개의 딥페이크 포르노 콘텐츠를 확보하였다. 게시글 또는 게시글의 해시태그에 ‘딥페이크’ 관련 어휘가 포함된 콘텐츠를 선별한 후, 해당 게시글의 딥페이크 이미지, 최초 게시자, 좋아요 수, 댓글 수, 공유 수, 사용자 간 공유 관계(Source→Sharer), 그리고 해시태그 정보를 수집하였다. 추가로, 사용자 정보를 얻기 위해

¹ <https://x.com/en/tos>

² <https://www.tumblr.com>

프로필 소개, 팔로잉 수, 빈번히 사용된 해시태그, 텀블러 배지(Badge) 유무 정보를 수집하였다. 사용자 정보를 기반으로 사용자 유형을 파악하기 위해 유저 클러스터링을 진행했으며, 이 때 Elbow method [8] 통해 총 7 개의 클러스터로 사용자를 구분하였다.

2.2 Method

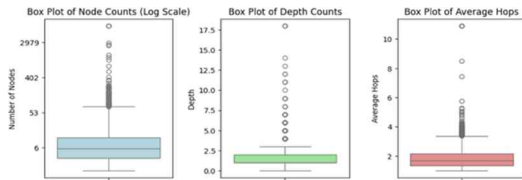
위너 인덱스는 네트워크 내에서 콘텐츠가 장거리로 확산되는 정도를 나타내며, 이는 콘텐츠가 네트워크 내에서 얼마나 넓은 범위로 확산되었는지를 나타낸다. 따라서, 상위 위너 인덱스를 가진 딥페이크 콘텐츠를 식별하는 것은 해당 콘텐츠가 네트워크 내에서 광범위하게 확산될 가능성을 초기에 인식하고 대응 전략을 마련하는 데 중요하다. 본 연구에서는 딥페이크 전파 양상을 분석하기 위해 딥페이크 콘텐츠를 전파한 개별 사용자는 노드로, 사용자가 콘텐츠를 공유한 네트워크 내의 연결을 엣지로 가지는 딥페이크 전파 트리를 모델링한다.

딥페이크 콘텐츠의 상위 위너 인덱스 해당 여부 예측을 위해 CNN 기반 모델을 설계하였다. 먼저, 64x64 크기의 RGB 이미지를 입력으로 받아 두 개의 Conv2D 레이어를 통해 특징을 추출한 뒤, 핵심 정보만을 전달할 수 있도록 window 크기가 2x2 인 MaxPooling2D 레이어로 공간적 차원을 축소한다. 이후 Flatten 레이어를 통해 1 차원 벡터로 변환하고 128 개의 유닛(unit)으로 구성된 Dense 레이어와 0.5 비율의 Dropout 레이어를 적용하여 모델의 일반화 성능을 높인다. 마지막으로 sigmoid 함수를 사용하여 최종 예측을 진행한다.

2.3 Experiment

딥페이크 콘텐츠의 상위 위너 인덱스 해당 여부 예측을 위하여 딥페이크 이미지, 최초 게시자의 사용자 유형, 게시물의 해시태그 정보를 입력으로 활용하였다. 실험 진행 시 배치 크기는 32 로 설정하고, 학습률은 0.001 로 조정하였다. 드롭아웃, 옵티마이저, 그리고 에포크는 각각 0.5, Adam, 3 으로 설정하였다. 모든 실험은 NVIDIA GeForce 631 RTX 3090 GPU 에서 수행되었다.

그림 3. 딥페이크 전파 트리 통계



수집된 5,261 개의 딥페이크 게시글로 생성된 전파 트리를 분석한 결과는 그림 2 와 같으며, 노드 개수, 트리 깊이, 위너 인덱스의 평균이 각각 6.0, 1.77, 1.81 인 것으로 나타났다. 이는 딥페이크 포르노에 대한 전파 양상이 평균적으로 얇고 한정된 네트워크 내에서 집중적으로 퍼지는 경향이 있음을 보여준다. 노드 수에 따른 위너 인덱스의 평균을 세분화하여 분석한 결과, 노드 수가 30 개 이하인 경우 1.67, 30에서 100 개 사이인 경우 2.71, 100 개 이상인 경우는 3.19 로 확인되었다. 따라서 전파 트리의 노드 수가 증가할수록 위너 인덱스도 증가하는 경향이 있으나 노드 수가 100 개를 초과하더라도 위너 인덱스의 증가는 상대적으로 완만하다는 것을 알 수 있다. 이는 딥페이크 콘텐츠가 전파되는 네트워크에서 사용자 수가 많아지더라도, 전파 거리는 짧은 수준으로 유지된다는 것을 시사한다.

표 1. 위너 인덱스 상위 범주에 따른 성능 (%) 비교

상위 범주	Precision	Recall	F1 score	Accuracy
10%	78.23	74.35	76.26	75.97
20%	82.34	79.12	80.70	80.30
30%	80.17	77.49	78.81	78.63

확산 가능성이 높은 딥페이크 콘텐츠 식별을 위한 최적의 기준을 도출하기 위해 위너 인덱스의 상위 범주별로 예측을 수행하였다. 표 1 은 범주에 따른 예측 정확도를 나타내며, 위너 인덱스 상위 20%에 해당하는 딥페이크 포르노 콘텐츠를 예측한 경우에 가장 높은 정확도를 보였다. 이는 상위 10%에 집중할 경우 예측 대상이 지나치게 제한되어 일부 중요한 전파 경로를 놓칠 가능성이 있으며, 반대로 상위 30%를 포함할 경우에는 확산 가능성이 낮은 콘텐츠까지 예측에 포함되어 정확도가 떨어지는 것으로 분석된다. 따라서 상위 20%에 해당하는 범위가 전파정도와 크기의 균형을 가장 잘 반영하는 적절한 지점으로, 딥페이크 콘텐츠의 효과적인 확산 예측과 대응을 위한 최적의 기준으로 판단된다.

III. 결론

본 논문에서는 딥페이크 콘텐츠의 확산 양상을 분석하고, 전파 가능성이 높은 핵심 콘텐츠를 효과적으로 식별하기 위한 최적의 기준을 도출하고자 한다. 딥페이크 포르노 콘텐츠를 수집하여 전파 트리를 모델링하고, 네트워크 내 확산 양상을 정량화하기 위해 위너 인덱스를 활용하였다. 실험 결과, 딥페이크 콘텐츠의 확산은 주로 제한된 범위에서 이루어지며 상위 위너 인덱스를 가진 콘텐츠가 더 넓은 네트워크로 확산될 가능성이 높음을 확인하였다. 특히, 위너 인덱스 상위 20%에 해당하는 딥페이크 콘텐츠를 가장 정확하게 예측할 수 있었으며, 이는 딥페이크 전파 예측 및 대응 전략 수립에 중요한 시사점을 제공한다. 연구 결과는 딥페이크 포르노와 같은 조작된 콘텐츠의 조기 통제 및 피해 예방을 위한 실질적인 기반을 제공하며, 향후 관련 분야의 연구 및 실무적 대응 전략 수립에 기여할 수 있을 것으로 기대된다.

Acknowledgement

본 논문은 2023 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행되었으며 (No.RS-2023-00230337, 악의적 변조 콘텐츠 대응을 위한 딥페이크 탐지 고도화, 생성억제, 유포 방지 플랫폼 개발), 과학기술정보통신부 및 정보통신기획평가원의 디지털분야 해외석학유치지원 연구결과로 수행되었음 (RS-2024-00459638).

참 고 문 헌

- [1] Kohli A, Gupta A (2021) Detecting DeepFake, FaceSwap and Face2Face facial forgeries using frequency CNN. *Multimed Tools Appl* 80(12):18461-18478
- [2] S. Minaee, Y. Y. Boykov, F. Porikli, A. Plaza, N. K. Arnavaz, and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.44, No.7, 2021.
- [3] Yazdinejad A, Parizi RM, Srivastava G, Dehghantaha A (2020b) Making sense of blockchain for AI deepfakes technology. In: 2020 IEEE globecom workshops (GC Wkshps). Taipei, Taiwan. IEEE, pp 1-6
- [4] Islam, Mohammad Raihanul, et al. "Deepdiffuse: Predicting the'who'and'when'in cascades." 2018 IEEE international conference on data mining (ICDM). IEEE, 2018.
- [5] Murayama, Taichi, et al. "Modeling the spread of fake news on Twitter." *Plos one* 16.4 (2021): e0250419..
- [6] Han, Jinyoung, et al. "Predicting popular and viral image cascades in pinterest." *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 11. No. 1. 2017.

- [7] Yang, Cheng, et al. "Full-scale information diffusion prediction with reinforced recurrent networks." *IEEE Transactions on Neural Networks and Learning Systems* 34.5 (2021): 2271-2283.
- [8] Bholowalia, Purnima, and Arvind Kumar. "EBK-means: A clustering technique based on elbow method and k-means in WSN." *International Journal of Computer Applications* 105.9 (2014).