Predicting Deepfake Virality on Social Networks

Seoyoon Jeong, JeeEun Kim, S. Shyam Sundar, Jinyoung Han*
Sungkyunkwan University

jsyoon0503@g.skku.edu, wldms15@g.skku.edu, sss12@psu.edu, *jinyounghan@skku.edu

Abstract

The proliferation of deepfake content on social media threatens information integrity. The role of deepfake media's unique visual characteristics in propagation has been largely overlooked by prior research. This paper proposes a proactive predictive model that forecasts a deepfake image's total shares over a 15-day period using its visual features and early propagation data. We constructed a unique deepfake propagation dataset and analyzed how image attributes correlate with virality. Our model, which integrates GraphSAGE with an LSTM network and CNN-extracted visual features, achieves superior predictive accuracy. The results show our approach can effectively identify deepfake content with high viral potential, contributing to the early prediction and mitigation of deepfake dissemination.

Ⅰ. Introduction

With the rapid advancement of generative models, deepfake technology has become widely accessible, enabling the creation of highly realistic synthetic media [1]. Malicious actors are exploiting this technology to spread misinformation and manipulate public opinion on social media [2]. The viral nature of deepfakes on social networks is dangerous, as their influence is difficult to control once they gain traction, posing a significant threat to public trust and safety. This underscores the critical need for early intervention strategies [3].

While deepfake detection technologies have made progress, most are reactive, identifying synthetic media only after it has been widely disseminated. This approach is often too slow to counter the rapid spread of deepfakes. Although research has explored content virality, the unique characteristics of deepfake media have been largely overlooked [4]. Prior studies have mainly focused on text-based content or benign images, ignoring the fact that deepfakes are a highly visual form of stimulus. As such, they are primarily shared based on their visual content, and user attributes or textual metadata are often limited or unreliable [5].

To address these challenges, we propose a novel predictive model that forecasts the total number of shares a deepfake image post will accumulate over a 15-day period, based on the deepfake image and its initial propagation data. Our model identifies posts with high virality potential shortly after they are published, offering a proactive solution to complement existing detection efforts and serving as a crucial step toward building a safer digital environment.

Ⅱ. Methods

Ⅱ-1. Data Collection and Processing

To train and evaluate our model, we created a novel dataset to track the spread of deepfake content, as no public dataset for this purpose existed. From November 2023 to March 2025, we tracked 6,833 deepfake image posts on Tumblr, chosen for its "reblog" function, which is ideal for studying content propagation. For each post, we collected the deepfake image, its full reblog cascade, and all relevant timestamps.

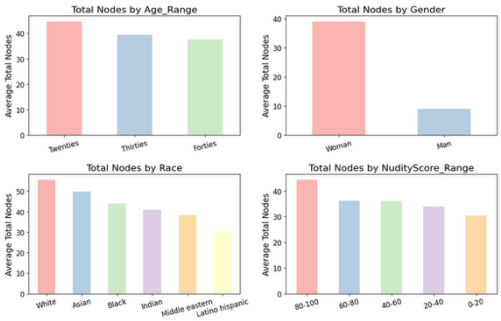

Figure 1. Influence of Demographics and Nudity on Shares

To better analyze the characteristics of deepfake images that tend to go viral, we used automated tools to extract visual and demographic attributes. For example, we used the DeepFace model to infer labels for age, gender, and race. We also used Nudity-Detection to calculate a nudity score for each image, which ranges from 0 to 100, with higher values indicating greater visual exposure.

Ⅱ-2. Data Analysis

To better understand how specific visual and demographic attributes of deepfake images correlate with their virality, we measured the size of their propagation cascades by the average total number of reshares. Deepfakes of individuals categorized as White and Asian also tend to have a higher average number of reshares. We found a positive correlation between visual exposure and virality, as deepfakes with a higher nudity score had the largest average propagation cascades.

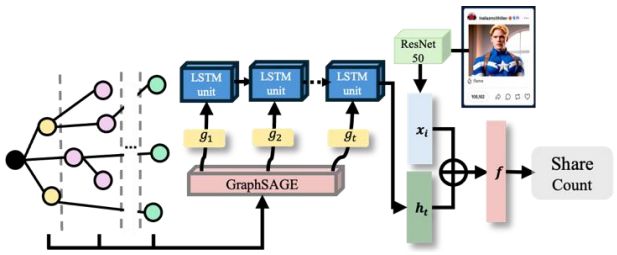

Figure 2. Overall architecture of the proposed model pipeline

This analysis confirms that the visual and demographic attributes of deepfake images are key factors in their propagation, providing crucial insights into the characteristics that drive content virality.

Ⅱ-3. Propagation Prediction Model

To predict the total number of shares, we developed a two-stage model that combines a GraphSAGE-based graph encoder with a Long Short-Term Memory (LSTM) network. This approach captures both the structural and temporal dynamics of content propagation.

In the first stage, a GraphSAGE encoder processes snapshots of the propagation graph over time. This encoder effectively captures the structural information of the sharing cascade, even with limited node-level data. The resulting sequence of graph embeddings is then fed into an LSTM network, which models the temporal evolution and learns how early propagation patterns influence the final share count.

Finally, the LSTM's final hidden state is concatenated with a visual feature vector extracted from the deepfake image using a pre-trained ResNet-50 model. This combined vector is then fed into a fully connected layer to predict the total share count, allowing our model to make a more accurate forecast by integrating both propagation and visual features.

Ⅱ-4. Result

To evaluate our model's performance, we used three metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Normalized Discounted Cumulative Gain (nDCG). While MAE and RMSE measure the accuracy of our numerical predictions for total shares, nDCG assesses the model's ability to rank content by its propagation potential, which is crucial for identifying high-risk, high-virality content for early intervention.

For benchmarking, we compared our model against two established temporal graph-based models, TGN (Temporal Graph Networks) and TGAT (Temporal Graph Attention Network), and a baseline model combining a GCN with an LSTM. These models are well-suited for capturing the evolving dynamics of social networks. To ensure a fair comparison, all models were provided with the same initial propagation graph structure constructed from the content sharing events.

Our model achieved the lowest MAE and RMSE values and the highest nDCG score, consistently outperforming the TGN and TGAT baselines in predicting the total number of shares. This demonstrates our model's superior predictive accuracy and its strong ability to identify content with high virality potential. By combining the structural and temporal dynamics of the sharing cascade with semantic features from the deepfake image, our approach proves to be a more effective way to model content propagation.

Methods	MAE	Share RMSE	nDCG
GCN + LSTM	5.339	8.431	0.653
TGN	6.207	9.005	0.602
TGAT	5.614	8.292	0.713
Ours	**4.821**	**7.901**	**0.781**

Table 1. Comparison of our method with baseline models

Ⅲ. Conclusion

In this paper, we proposed a novel predictive model to forecast the total number of shares of deepfake content using both the image's visual features and its early propagation data. We constructed a dedicated deepfake propagation dataset from Tumblr and analyzed how deepfake image features correlate with their virality. We employed a graph encoder combined with an LSTM network to capture spatiotemporal diffusion patterns, which were integrated with CNN-extracted visual features to predict the total number of shares. The results show that our model effectively identifies content with high viral potential. These findings suggest that our approach can contribute to the early prediction, thereby supporting the development of safer social networking environments.

ACKNOWLEDGMENT

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2023-00230337, Advanced and Proactive AI Platform Research and Development Against Malicious Deepfakes) and under the Global Scholars Invitation Program (RS-2024-00459638).

References

[1] Amerini, Irene, et al. "Deepfake media forensics: Status and future challenges." Journal of Imaging 11.3 (2025): 73.

[2] Xu, Zhaoxiang, et al. "Public perception towards deepfake through topic modelling and sentiment analysis of social media data." Social Network Analysis and Mining 15.1 (2025): 16.

[3] Abbas, Fakhar, and Araz Taeihagh. "Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence." Expert Systems with Applications 252 (2024): 124260.

[4] Wu, Cheng-Lin, et al. "MUFFLE: Multi-modal fake news influence estimator on Twitter." Applied Sciences 12.1 (2022): 453.

[5] Sundar, S. Shyam, Maria D. Molina, and Eugene Cho. "Seeing is believing: Is video modality more powerful in spreading fake news via online messaging apps?." Journal of Computer-Mediated Communication 26.6 (2021): 301-319.