From Laughter to Concern: Exploring Conversations about Deepfakes on Reddit - Trends and Sentiments

Harshitha Benakanahalli Nagaraj
Rochester Institute of Technology

Rahul Gowda Kengeri Kiran
Rochester Institute of Technology

Abstract

The escalating prevalence of deepfake content on online platforms raises concerns about its potential threats to individual privacy, national security, and democracy. [2] This phenomenon is closely tied to rapid advancements in deep learning technologies, enabling highly realistic manipulation and generation of synthetic content. With rapidly evolving tools for deepfake creation and changing public perceptions, there is a pressing need to keep pace with these developments to enable social media platforms, law enforcement agencies, and researchers to develop better deepfake detection capabilities. To contribute to this evolving field, we compiled and analyzed a dataset of deepfake-related SFW discussions on Reddit. Our systematic analysis revealed several key findings: the emergence of new creation tools each year, spikes in negative sentiment towards deepfakes following high-profile misuse incidents, and a diverse range of discussions including topics such as deepfake creation involving famous personalities, challenges in regulation, detection techniques, and instances of deepfake-related scams. These insights provide valuable information for understanding the evolving landscape of deepfake technology and its societal impact, potentially informing future strategies for detection, regulation, and public awareness campaigns.

1 Introduction

With the rapid advancement of Deep Learning technologies, deepfakes have blurred the line between reality and fabrication. [1] While deepfakes can be employed for humorous and entertaining purposes, they have always harbored the potential for disruption. Unfortunately, some bad actors have exploited this potential and wreaked havoc. Therefore, identifying the critical trends within the deepfake creation community and developing a comprehensive understanding of the current deepfake landscape will provide valuable insights into effectively and efficiently addressing the challenges posed by deepfakes.

In order to achieve this, we compiled and analyzed a deepfake-related SFW dataset of discussions from 2018 to February 2024, which includes information on creation and detection tools and techniques, news and events, and other relevant topics from Reddit conversations. We seek to answer the following research questions by analyzing the gathered data:

- *RQ1:* What are the prominent latent topics and themes present within the deepfake-related discussions on Reddit during this study period?
- *RQ2:* What individuals, groups, or entities have been the primary targets of deepfakes?
- *RQ3:* How does public sentiment towards deepfake content vary across different personalities, and to what extent does public perception of them impact the sentiment?

Through our analysis, we found diverse topics and themes, including the creation of deepfakes featuring famous personalities, concerns about deepfakes' implications, legislative challenges, audio deepfakes, deepfake detection techniques, and more. Famous personalities like Nicolas Cage, Donald Trump, Joe Biden, and Tom Cruise emerged as primary targets, with their prominence fluctuating across different years, potentially influenced by events or evolving public perception. Sentiment analysis reveals that there has always been negative sentiment towards deepfake with spikes in between where big controversies have occurred by the use of deepfakes.The findings underscore the complexities of online discourse surrounding deepfakes, where celebrity status, public perception, and contextual factors affect the reception of such content.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2024.
August 11–13, 2024, Philadelphia, PA, United States.

2 Ethical Considerations

We obtained Institutional Review Board (IRB) approval from our university for this study. The IRB committee declared that since we are using publicly available text from Reddit, which is accessible to anyone, this does not qualify as human subjects research under the federal regulations. Human subject research regulations apply only to private data, and posting on Reddit would not be considered private. During our data collection process, we solely obtained data from publicly accessible Reddit forums and did not attempt to access any private or restricted subreddit groups. Furthermore, we intentionally did not collect Reddit usernames, as they were irrelevant to the nature of our analysis.

Consequently, our study did not focus on specific user perspectives, individual subreddits, or specific moderators within these groups. Instead, our analysis concentrated on vast amount of data, examining overarching discussion topics, collective user perspectives, and drawing insights from the data as a whole.

The Institutional Review Board (IRB) at our university deemed our work exempt as it does not quality as human subjects research under federal regulations. All the data was collected from public Reddit conversations and anonymized.

3 Methodology

Our study analyzed 4,015 Reddit posts with 93,775 comments from 2018 to February 2024, collected using PRAW from deepfake-focused subreddits and keyword searches. Data was stored in Redis for efficient incremental crawling, with measures taken to comply with Reddit's rate limits and maintain a SFW dataset. Text preprocessing involved lowercase conversion, removal of irrelevant elements, tokenization, stop word removal, and lemmatization. The data was then separated into year-wise datasets for temporal analysis.

We conducted sentiment analysis using the VADER model, which is well-suited for analyzing informal language common on social media. This allowed us to categorize comments as positive, neutral, or negative, providing insights into public perceptions of deepfakes. We used spaCy's Named Entity Recognition (NER) to identify and count mentions of individuals in posts, revealing the most frequently deepfaked personalities. To understand the evolution of deepfake technology, we analyzed posts discussing creation tools, identifying newly emerging tools each year and tracking their frequency of mention.

We employed Google's gemini-pro model through the PaLM API to classify posts into predefined categories, using a carefully engineered prompt. For topic modeling, we utilized BERTopic, applying it to both year-wise and complete datasets. The resulting topics were refined through parameter tuning and qualitative coding methods to ensure accurate representation of discussion themes. Related topics were grouped into broader categories to provide a comprehensive overview of prominent themes in deepfake discussions on Reddit.

4 Initial Results

To answer RQ1, the analysis revealed a diverse range of deepfake-related topics evolving from 2018 to 2024. Discussions consistently centered on creating deepfakes of famous personalities, while early conversations focused on sharing tools and expressing concerns about the technology. Over time, the discourse shifted towards regulatory challenges, potential misuse, and the emergence of audio deepfakes. Later years saw increased emphasis on detection techniques, deepfakes in news and media, scam incidents, and hardware requirements for creation tools. Fig. 2 illustrates the extracted topics across the study period.

To answer RQ2, the analysis revealed that celebrities and public figures featured in deepfakes increased and varied each year. While Donald Trump was not the top deepfaked personality every single year, he emerged as the most frequently deepfaked individual overall across the entire study period from January 2018 to February 2024. The tools used for deepfake creation evolved annually, with new options emerging every year. Notably, DeepFaceLab stood out as a prominent tool that maintained its popularity throughout the study period.

To answer RQ3, the analysis showed that negative sentiment towards deepfakes tended to spike following incidents of misuse. A notable example occurred in January 2024, when sexually explicit deepfake images of Taylor Swift circulated widely on the internet. This event triggered a significant increase in negative sentiment among Redditors, reflecting public concern and disapproval of deepfake technology's harmful applications. This pattern of sentiment fluctuation in response to high-profile incidents demonstrates the public's sensitivity to the potential dangers and ethical implications of deepfake technology.

References

[1] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M Nguyen. Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding*, 223:103525, 2022.

[2] U.S. Department of Homeland Security. Increasing Threats of Deepfake Identities.