A Mixed-method Approach to Analyze Deepfake-related discussions on Reddit and Twitter

Keywords: Deepfakes, BERTopic modeling, Discourse analysis, Reddit, Twitter

Extended Abstract

Introduction: With the advancement of artificial intelligence (AI), deepfakes—AI-generated synthetic content—are becoming extremely popular on social media platforms. Although some applications of deepfakes can be used for good, in many cases, social media shares deepfakes of unconsented synthetic nude images or synthetic pornography, or even political misinformation. Regardless of any implications, the use of deepfakes can be normalized and shaped by the communication in the social media communities that use them. Although platforms regulate or moderate harmful content shared on their platforms, it is not clear on deepfakes—how some content and communication may be harmful or impact society. To understand the gravity of the societal impact of deepfakes, it is essential to understand how online communities use this in their conversations. One way to explore this is to focus on multiple social media platforms where users exchange information and opinions about deepfakes, which provide us with a more nuanced view of what and how communities in each platform communicate and perceive this phenomenon.

Problem statements and Research Questions: Previous research demonstrates that the microblogging platform Twitter and social news aggregation platform Reddit have been playing instrumental roles in sharing and spreading opinions in many aspects—news, media, conspiracies, and propaganda including ideologies and perceptions of deepfakes. Especially, the subreddit community "r/Deepfakes" brought the concept of deepfakes into society and Twitter has been active in spreading deepfakes. Although these platforms have policies against deepfakes being shared [1], it does not capture the potential social harm instigated by the conversations of deepfakes. Deepfakes are not necessarily harmful, but solely dependent on the purpose it has been used and the conversation centered on it [2]. In the past, communities on Reddit and Twitter have been showing toxic, and harmful conversations under various phenomena. Due to the nature and the advancement of Deepfakes being a new phenomenon, it is yet unknown of the directions of discourse, the intentions of their use, or how the public is engaging with deepfakes across these platforms. Thus, we explored public discourse about deepfakes on Reddit and Twitter. We specifically asked:

R1: What are the major topics discussed in these platforms which are centered around deepfakes?

R2: How similar or different are these discussions on both platforms?

Methods: We employed mixed methods to answer these questions. We extracted discussions based on the search term "deepfakes" from Twitter–Twitter API V2 and Reddit- Pushshift API between 2018-2022 and collected 101,869 Reddit posts and 510,739 tweets. To answer the RQ1, a BERTTopic model analysis was conducted for both sets of data. Topic modeling was performed not only on the entire dataset but also on chunked data based on each year which reflects ideologies in each year. To answer the RQ2, we examined the topics in each platform and quantified similarities and differences in these topics using the cosine and generated a heatmap based on the cosine similarity matrix between topic embeddings. The matrix was constructed by having the Top 20 Twitter topics in the X axis and the Top 20 Reddit Topics in the Y axis. We considered 0.80 as a threshold to show the similarity based

on the ideologies in the discussed topics. Topics distributed across years were qualitatively examined and clustered based on similarities and differences.

Results: We found both similar and different topics discussed on Reddit and Twitter. Using BERTopic, we clustered the discussed 20 major topics for both platforms (Fig. 1a). The heatmap shows the most similar topics discussed in both platforms. Out of 40 topics (with 400 possible topic comparisons), only 29 were found to be within the similarity threshold. Some of those similar topics were mostly related to many deepfake videos shared by people (highest similarity of 0.949), discussions about viral deepfake videos such as Queen Elizabeth deepfake Christmas messages (0.903), and deepfake related news such as the Lucasfilm hiring of the Youtuber who created a de-aging deepfake (0.879) or news videos about a mother creating a deepfake pornographic video of cheerleader (0.855) are commonly discussed in both platforms. Also, discussions about users asking about the legitimacy of some videos being deepfaked or not (0.838), and discussions of nude or pornographic videos of celebrities (0.855), deepfake voice (0.929) were found to be similar in both platforms. On the other hand, the differences in both platforms were visible when the Reddit platform discuss specific deepfakes videos which have not been extensively discussed on Twitter such as the Terminator-3 film actress Scarlett Johansson on deepfaked sex videos (0.614), the House Full drama series actor Nick Offerman's deepfake series (0.590) which has received Platinum, Gold, and Silver Reddit coin awards from Reddit community, and deepfake video of Game of Thrones' actor Jon apologizing for season 8 (0.628) went viral and criticized in Reddit platform. These are all Reddit user-created deepfakes. Interestingly, most Twitter topics related to deepfakes have been discussed on Reddit but only two topics were found uniquely–Business deep learning Machine learning course (0.628) and discussions related to the truth about deepfakes (0.614).

To understand the in-depth topical changes and nuanced views between the two platforms, we quantified yearly topic distributions. The yearly distribution of the top 10 topics from each platform brought much detailed idea of how the discussed topics vary between years. Based on the qualitative analysis, we found the majority of topics discussed on these platforms were epistemically different in their ideologies, interactions, and interests. Twitter users more discussing or share news and event-related tweets. However, on Reddit, users were more discussing aspects of deepfakes, where some were about asking the community about the video's legitimacy of deepfakes, pornography-related videos, and discussions relating to creating deepfake videos. Examining the key topics discuss every year, we found that Reddit platform discussed "morality of deepfakes" in 2018, which discusses the immoral actions of deepfakes. However, such discussions never occurred in subsequent years, but discussed more on "deepfake creations", "technology use and training" or just shared many deepfakes seen on social media. In contrast to Reddit, Twitter acted more as a news platform where users mostly shared what's currently trending, and created deepfakes videos of popular actors and political leaders.

Implications: This study revealed that Twitter is a platform for sharing deepfake videos and news, whereas Reddit is deeply engaged in interactions on creating deepfakes for either ethical or unethical purposes (Fig.1b). Not just banning, platforms need different levels of moderation and regulations against deepfake conversations on Twitter and Reddit.

References

[1] J. D. Cochran and S. A. Napshin, "Deepfakes: Awareness, Concerns, and Platform Accountability," *https://home.liebertpub.com/cyber*, vol. 24, no. 3, pp. 164–172, Mar. 2021, doi: 10.1089/CYBER.2020.0100.

[2] A. De Ruiter and A. D. Nl, "The Distinct Wrong of Deepfakes porn · Social identity," 123AD, doi: 10.1007/s13347-021-00459-2.

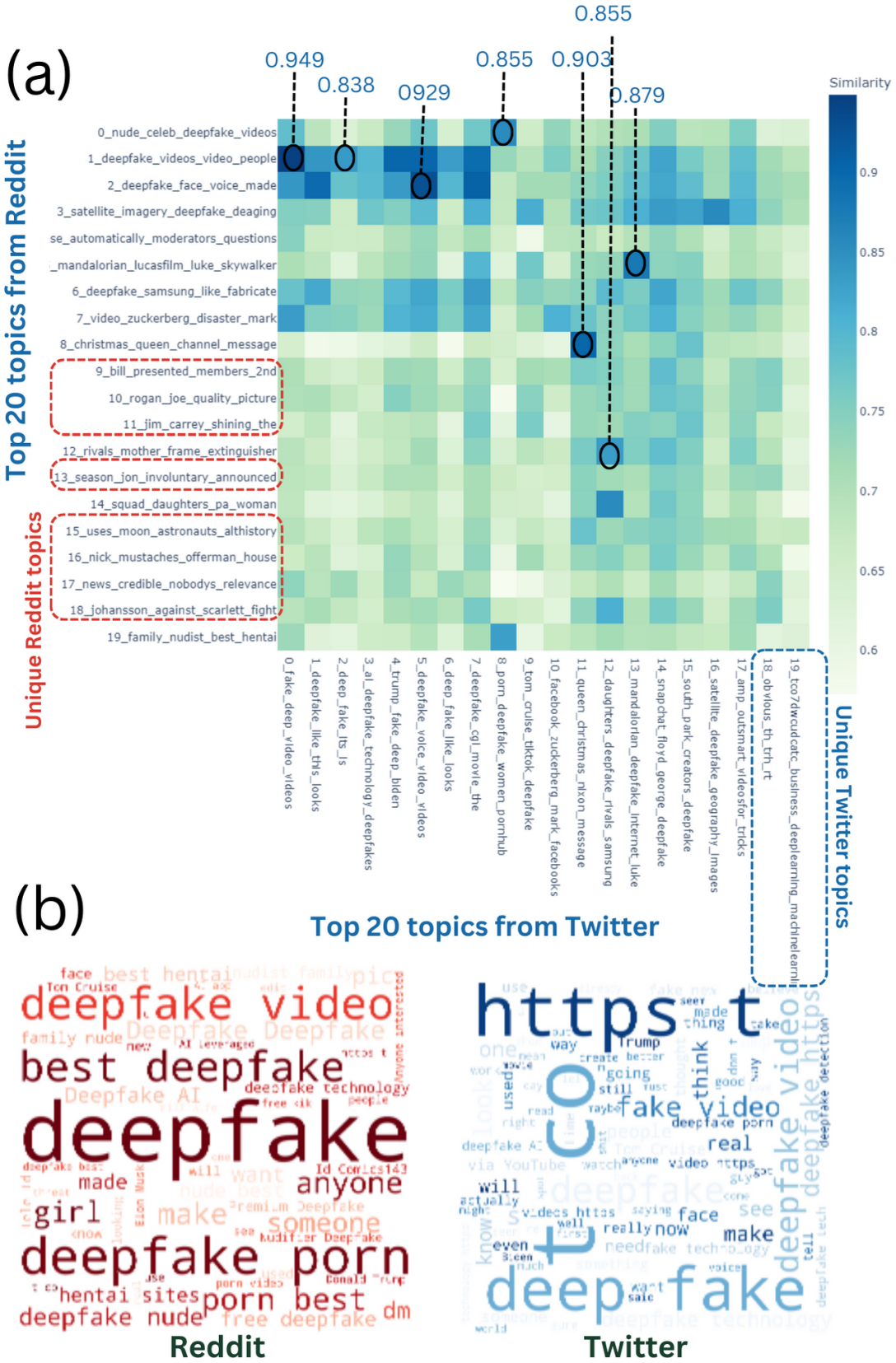

Figure **1a**. The heat map was generated based on cosine similarity index on the top 20 topics obtained from BERTopic Model. X-axis indicates the topics from Twitter and Y-axis indicates topics from Reddit. **1b.** Wordcloud generated from two platforms indicates Twitter is more sharing culture (http) whereas Reddit is more creating deepfakes videos (nude, porn or other)