Slovakia as the Precursor to Deepfake-Enabled Election Interference: Lessons Learned and Pathways Forward

Matyas Bohacek

Stanford University, USA
maty@stanford.edu

On September 28, 2023, just two days before Slovakia's 2023 elections, a hot mic of Michal Simecka, the liberal party leader, went viral on Slovak social media[1]. In the recording, Simecka confessed to conspiring to rig the upcoming election using votes from the Roma minority to a prominent journalist, who was seemingly covering the story up. Simecka's political opponents rushed to share the message on their profiles, further igniting public outrage. The only catch? The whole recording was fake[2]. What's more, because of a moratorium on election coverage two days before the election[3] – as mandated by Slovakia's law – the media was not able to set the record straight. Whether this event had a tangible effect on the outcome of the election is still debated[4]. However, it is already clear that the video tricked thousands of people into believing it, received mainstream attention, and left the media ecosystem partially paralyzed– all of that despite its poor quality. In fact, the video did not even include a visual of Simecka or Todova talking; it only contained a static image serving as the background for an audio voice-over with captions, as shown in Figure 1.

This video was made possible by the recent advancements in image (Zhang et al. 2023; Bie et al. 2023), audio (Khanjani, Watson, and Janeja 2023), and video (Mirsky and Lee 2020) deepfake technology and the democratization of this technology through online platforms. It now takes just a single photo of a person to generate a new image of them in any desired setting using text-to-image models, half a minute of a person's speech to clone their voice, and one underlying video to render a lip-sync deepfake[5]. Importantly, these can be created through online services that, for very little or no fee, make this technology accessible to many users with basic technical skills[6]. And, despite the creative promise, as Slovakia's case clearly shows, this technology can be used for fraud, disinformation, and interfering with democratic processes.

In a critical year for democracy, in which almost half the world will vote[7], Slovakia's case begs the question of what preventive measures and direct interventions should be implemented to minimize the impact of future deepfake-enabled election interference. What lessons can be learned from this case? And is there any way forward for safe and independent elections?

For an intervention to be undertaken against a deepfake, the deepfake must be first identified and localized. While deepfake media is getting increasingly hard to distinguish from authentic media for individuals (Lu et al. 2023; Milewski, Zaporowski, and Czyżewski 2023), the computer vision and media forensics literature has been studying the automatic detection of such media quite extensively. The literature has introduced methods that can robustly detect image (Xu, Zhang, and Shi 2024), audio (Barrington et al. 2023), and video (Boháček and Farid 2022) deepfakes, yielding in-the-wild accuracies of over 95%. To that end, it is critical to recognize that the speed of development of deepfake technology is high, and so it is only a matter of time before contemporary detection methods become obsolete. Still, it appears that, at least for the time being, the forensics community has managed to keep up with the deepfake models and engines to robustly detect them. However, there are some gaps in the existing literature. For example, almost all of the existing work focuses on English and American-centric deepfakes (Yi et al. 2023; Yu et al. 2021).

A bigger challenge than classifying a single image, video, or audio as deepfake or not is the identification of deepfakes in the mass of content uploaded on the internet every day. Deepfakes usually get on the radar when they become viral[8] – only then can an intervention, such as a response from the social media platform that hosts the content, be drafted and executed[9]. However, some argue that this is too late (Khan,

[1] https://www.wired.com/story/slovakias-election-deepfakes-show-ai-is-a-danger-to-democracy/

[2] https://fakty.afp.com/doc.afp.com.33WY9LF

[3] https://www.brookings.edu/articles/the-impact-of-generative-ai-in-a-global-election-year/

[4] https://www.thetimes.co.uk/article/was-slovakia-election-the-first-swung-by-deepfakes-7t8dbfl9b

[5] https://www.theverge.com/2019/6/20/18692671/deepfake-technology-singing-talking-video-portrait-from-a-single-image-imperial-college-samsung

[6] https://www.deepswap.ai, https://deepfakesweb.com, https://elevenlabs.io

[7] https://www.theglobeandmail.com/world/article-world-elections-2024-list-of-countries/

[8] https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf

[9] https://reutersinstitute.politics.ox.ac.uk/news/spotting-deepfakes-year-elections-how-ai-detection-tools-work-and-where-they-fail

Michalas, and Akhunzada 2021; De Beer and Matthee 2021; Pérez-Rosas et al. 2017), because the deepfake has at that point realized most of its potential. For it to become viral, it must have already been presented to a large body of users, and the lifetime of the post after its identification as deepfake likely will not be very long.

Due to the shortcomings of reactive interventions, the literature has examined other means of identifying emerging disinformation and deepfake threats that would allow for the responses to be more timely. In many cases, especially in Eastern Europe, hoax websites are a reliable proxy for the upcoming hoax stories that will erupt on social media in the coming days[10]. However, this did not appear to be the case for Slovakia's deepfake case. The most prominent hoax-spreading websites in Slovak online space, identified by Konspiratori.sk, have not used the deepfake in building their narrative. We analyzed 40 most prominent websites from Kinspiratori.sk's analysis[11], finding no references to the deepfake.

Instead, the deepfake mostly circulated on social media (notably, Facebook and Telegram) and in disinformation email campaigns[12]. These channels are closed to public monitoring, making the detection difficult. Social networks prevent developers and researchers from scraping their content, and emails are, by nature, private, so they can't be analyzed without their recipient's direct, proactive interest. In summary, timely identification of deepfake content as it is emerging on the internet, particularly social media, is challenging.

Nonetheless, once the deepfake content is identified, the focus of the intervention turns to platform policy and regulations. In Slovakia's case, the deepfake fell through the cracks – both in terms of the platform policy and law. Facebook's policies restrict video deepfakes but not audio deepfakes[13], so the fake audio spliced into a static video theoretically did not violate this policy. In terms of the law, there are no regulations concerning deepfakes in Slovakia as of yet (Unit 2021), similar to most other European jurisdictions (Unit 2021). Therefore, the video still exists on Facebook[14] as of the time of writing (April 2024). The most severe measure taken against the deepfake to this day is that the police informed the public about it on their social media profile.

According to many analysts, Slovakia's 2023 election was the first one that saw major deepfake interference[15]. It will not be the last. Moving into the future elections, we can expect the number and sophistication of such high-profile, election-targeted deepfakes to rise. As illustrated in this extended abstract, we believe Slovakia offers multiple lessons

Figure 1: Screenshot from the Simecka and Todova deepfake shared on Facebook.

that can be learned and applied in future research, community action, and policy-making in elections-targeted deepfake interventions:

1. Deepfake detection technology is not sufficiently developed and evaluated for non-English (especially low-resource) languages, leading to lesser reliability and coverage for world elections;
2. Election-targeted deepfakes seem to spread on social media and hoax email chains without the possibility for automatic third-party analysis, leading to slower response;
3. Despite their usefulness in other analyses, conventional websites and channels for spreading disinformation may not serve as a reliable observatory for deepfakes that are gaining traction;
4. Platform policies of social networks and laws have not kept up with deepfakes, and so they can often linger online even after their identification, leading to more potential deception and damage.

Possible pathways to address these identified challenges, in the respective order, are:

1. The literature should focus on developing language- and context-diverse deepfake detection methods;
2. Social media platforms should allow for researchers and independent organizations to access public social media data for purposes relating to election monitoring;
3. Conventional websites and channels for spreading disinformation should not be treated as a reflection of the deepfake content spreading elsewhere;
4. Platforms should adopt policies – and countries should adopt laws – that clearly define what use of deepfakes is allowed and how these rules should be enforced.

[10]https://libguides.lib.cwu.edu/c.php?g=625394&p=4391900

[11]https://konspiratori.sk/zoznam-stranok

[12]https://www.icjk.sk/275/Analyza-Pred-volbami-na-novinarov-a-media-najviac-utocil-Robert-Fico-a-strana-Smer-platili-si-aj-reklamu

[13]https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/

[14]https://www.facebook.com/100012139860134/videos/285237664331210/

[15]https://www.thetimes.co.uk/article/was-slovakia-election-the-first-swung-by-deepfakes-7t8dbfl9b

References

Barrington, S.; Barua, R.; Koorma, G.; and Farid, H. 2023. Single and multi-speaker cloned voice detection: From perceptual to learned features. In *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1–6. IEEE.

Bie, F.; Yang, Y.; Zhou, Z.; Ghanem, A.; Zhang, M.; Yao, Z.; Wu, X.; Holmes, C.; Golnari, P.; Clifton, D. A.; et al. 2023. RenAIssance: A Survey into AI Text-to-Image Generation in the Era of Large Model. *arXiv preprint arXiv:2309.00810*.

Boháček, M.; and Farid, H. 2022. Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms. *Proceedings of the National Academy of Sciences*, 119(48): e2216035119.

De Beer, D.; and Matthee, M. 2021. Approaches to identify fake news: a systematic literature review. *Integrated Science in Digital Age 2020*, 13–22.

Khan, T.; Michalas, A.; and Akhunzada, A. 2021. Fake news outbreak 2021: Can we stop the viral spread? *Journal of Network and Computer Applications*, 190: 103112.

Khanjani, Z.; Watson, G.; and Janeja, V. P. 2023. Audio deepfakes: A survey. *Frontiers in Big Data*, 5: 1001063.

Lu, Z.; Huang, D.; Bai, L.; Liu, X.; Qu, J.; and Ouyang, W. 2023. Seeing is not always believing: A quantitative study on human perception of ai-generated images. *arXiv preprint arXiv:2304.13023*.

Milewski, K.; Zaporowski, S.; and Czyżewski, A. 2023. Comparison of the Ability of Neural Network Model and Humans to Detect a Cloned Voice. *Electronics*, 12(21): 4458.

Mirsky, Y.; and Lee, W. 2020. The Creation and Detection of Deepfakes. *ACM Computing Surveys (CSUR)*, 54: 1 – 41.

Pérez-Rosas, V.; Kleinberg, B.; Lefevre, A.; and Mihalcea, R. 2017. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*.

Unit, S. F. 2021. Tackling deepfakes in European policy. Technical report, Tech. rep., European Parliamentary Research Service. https://www. europarl ....

Xu, K.; Zhang, L.; and Shi, J. 2024. Detecting Image Attribution for Text-to-Image Diffusion Models in RGB and Beyond. *arXiv preprint arXiv:2403.19653*.

Yi, J.; Wang, C.; Tao, J.; Zhang, X.; Zhang, C. Y.; and Zhao, Y. 2023. Audio deepfake detection: A survey. *arXiv preprint arXiv:2308.14970*.

Yu, P.; Xia, Z.; Fei, J.; and Lu, Y. 2021. A survey on deepfake video detection. *Iet Biometrics*, 10(6): 607–624.

Zhang, C.; Zhang, C.; Zhang, M.; and Kweon, I. S. 2023. Text-to-image diffusion model in generative ai: A survey. *arXiv preprint arXiv:2303.07909*.