validsoft

Protecting Social Media from Deepfakes

The Role of Voice Biometrics in Content Moderation

© 2022 ValidSoft. All Rights Reserved.

www.validsoft.com

There is considerable discussion and debate concerning the management of harmful content on the internet. How much longer will we tolerate the lack of resolution and allow this to continue? The risks faced by internet users, and the challenges addressing those risks, has become more real than ever. It's time to focus on protecting internet users from such content, and helping content platforms meet these challenges, while also protecting vital freedom of speech.

Illegal or seriously harmful content is a major issue for Social Media platform providers and the primary focus of such providers is to enhance the effectiveness of online content moderation in order to offer users greater protection from potentially harmful material. Content must be analyzed to ensure that it complies with strict rules and guidelines before it can be published. The process is complex, complicated and still very dependent on painstaking manual supervision and control - a mammoth task when the social media platform is global, multilingual and subject to very different attitudes to freedom of expression. Allowing harmful or illegal content to be inadvertently published is costly as well as unethical. Indeed, allowing abusive, mentally damaging, content to be distributed to new and even naive minds that will shape humanity's future, for better or worse. This is no longer purely a cost issue. It speaks to fundamental questions about how humanity wishes to express itself and evolve. Moreover, the psychological impact on individual (human) content moderators, and internet users alike, can be profound when exposed to extreme content. Artificial Intelligence (AI) and Machine Learning (ML) tools have helped screen for extreme content but as we can prove, it's no longer sufficient. The arms race between 'good' and 'bad' AI

"Voice is now becoming the new human User Interface"

The Role of Content Moderation

Facebook alone, with over 2 billion active users, employs more than 15,000 humans to catch things that existing technology can't catch – and it's still not enough!

"Chatbots must understand not only the *what*, but the *who*"

AI was a good first stage. Yes "was". Mostly, it has reduced the volume of extreme content that humans need to review, and in that way reduced the aggregate volume of psychological impacts to a degree. Its main impact has been to detect spam, fake accounts, adult nudity and terror-related posts for instance. However, it is not adept at detecting things such as hate speech or interpreting context, such as parody or humor, which still requires human intervention – and judgment – in determining compliance.

Human moderators and AI also have great difficulty in detecting Deepfake video and audio. If a Deepfake is simply a video of a person speaking "fake news" or "fake content", it will not be detectable to AI, nor to a human moderator as the quality of Deepfake technology is good enough to fool the human eye/ear and the technology that automates such screening.

© 2022 ValidSoft. All Rights Reserved.

Such authentic looking and sounding Deepfakes results in the publishing of something damaging, including to the platform publishing it. It's a continuation of the 'whack-a-mole' arms race.

"Authentication is invisible, fast and continuous"

Facebook and other social media platforms must therefore detect Deepfakes in the first instance, and subsequently determine if they meet the criteria for removal. This should involve both static media uploads and live broadcasts. The automatic detection and categorization of Deepfakes for immediate/subsequent examination, through AI or human moderation is therefore the key. Existing Content Moderation techniques alone will not detect them.

What if a different method exists? One which has been applied in the real world, for as long as humans have evolved to speak to one another, but which we never yet thought to apply to this problem?

Deepfakes – A Social Media Threat

Deepfakes are nefarious computer-generated audio and video that are a by-product of the advancements in AI and ML. Designed to sound and look exactly like the human they mimic, the technology is now sufficiently advanced to fool any human eye/ear. With the availability of Deepfake-generating software tools, the technology has been brought within the reach of not only fraudsters and cybercriminals, but anyone with a particular agenda or desire to spread fake news, cause fear or confusion or simply to humiliate or cause angst for any target individual. This is not something new! It all starts with the ability to digitize media, a picture, a sound, a written story. It's too late now to reverse this trend, so the least we can do is leave our future generation with something we can be proud of, the application of technology to reinstate the basic bonds of local person-to-person trust that we have evolved over millennia!

"Deepfakes will render existing impersonation attempts redundant"

Whereas Deepfakes were initially focused on humorous applications, or transposing celebrity faces into pornographic videos, their usage and the implications thereof are much broader. Nefarious actors, whether individuals, organizations or even governments will see greater value in the technology than mere harassment, nuisance or comedy value.

And as the technology has become more readily available and increasingly sophisticated, detection has become a far harder task. For social media platforms such as Facebook the emergence of Deepfakes is a real problem. In January of this year Facebook announced it was removing Deepfake media if it met certain criteria around misleading audio and video but excluding parody and satire amongst others.

© 2022 ValidSoft. All Rights Reserved.

Removing any existing media and blocking future Deepfake media, including file uploads and real-time live streaming is the responsibility of content moderators, both human and machine. The problem, however, is how to accurately identify them.

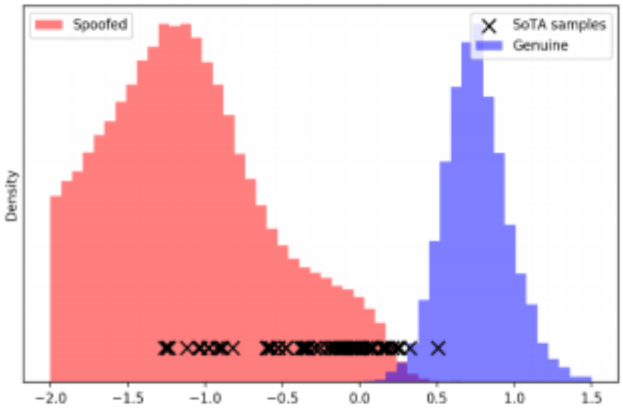
The Detection of Synthetic Voice: have you chosen the right Solution?

Only advanced voice biometric engines that can discriminate a human voice versus a synthetic, ML-generated voice can properly detect these Deepfakes.

Whilst Deepfakes, whether audio or audio combined with video, can trick any human into believing it is the genuine person, it cannot generate the unique characteristics, such as esophageal and physiological shape and vocal cord vibrations that are used in voice biometric analysis. These characteristics, their sequences, tones and aspects are as unique as our DNA!

Patented proprietary synthetic speech detection algorithms are multi-dimensional, measuring both the behavioral traits of an individual as well as the impact their physical body creates on sound. Linking the sound waves a human emits through the physiology of their body, we model the different frequencies and other characteristics in the sound spectrum created by the human voice, including those that the human ear cannot hear. The result is a unique and highly accurate scientific method of biometric identification. These perfected algorithms have inbuilt capabilities to detect differences between human voices and synthetically-created voices that neither a human, nor AI, can possibly do.

It is a data-driven classifier approach, an analysis of characteristics of audio that is able to identify unnatural features in a voice. Synthetic audio may sound identical to natural audio for a particular person, but once computer generated audio is digitized and analyzed by the ValidSoft synthetic voice detection patent, the differences are clear.



The distribution graph above demonstrates the ability of ValidSoft's synthetic detection capability to identify Deepfake audio, something the human ear could not.

© 2022 ValidSoft. All Rights Reserved.

Identification of Serial Offenders

ValidSoft's voice biometric technology can be applied to these challenges and more! For example, a solution could automatically create a "watchlist" of known and new synthetic audio characteristics, in the same way that it could also "watchlist" serial offenders, of human or synthetic voices linked to inappropriate and banned audio and video-audio posts.

Such a "watchlist" can then be used in real-time to compare and identify audio in files being uploaded or streamed to block repeat offenders - such applications can also be used in 'background mode', allowing previously published content to be identified and linked to a known offending synthetic or human 'voiceprint'. Another novel, but obvious, idea for some applications: Only allow (or promote) posts and content which have a valid voice biometric score, and are proven to be real humans. Whack-a-mole, whacked!

Literally only seconds of audio are required to obtain a result, with the service available on-premise or in the cloud. ValidSoft's biometric capabilities extend beyond authentication alone and can assist social media platforms in their ongoing content moderation processes.

This identification capability is not limited to Deepfake videos containing an author's own voice, but to any media uploaded or streamed by users. The ability to block offensive and artificial media, posted by repeat offenders, is another way of protecting content moderators and the value of content platforms themselves. It also shows the public and regulators around the world that platforms take their responsibilities to mankind seriously and are using all available technology solutions to address these real world, human, problems!

© 2022 ValidSoft. All Rights Reserved.

Learn More

ValidSoft is a leading voice biometrics software company with a long history of innovation in voice authentication and biometrics. Our technology is built using active, passive, and continuous voice-based authentication, guaranteeing that the speaker is who they are, always. Our solutions help to eliminate call fraud and identity theft. ValidSoft's EuroPriSe™ privacy seals ensure 100% compliance with EU GDPR and other leading Data Protection and Data Privacy laws like HIPAA, Digital Identity Guidelines, Vectors of Trust, Federal Identity Program Guidelines, etc. ValidSoft is consistently recognized by third-party analyst firms as a market leader. See how ValidSoft is powering the Future of Identity at www.validsoft.com

Contact ValidSoft

UK Office:

30 Moorgate
London EC2R 6JJ
United Kingdom

USA Office:

100 Pearl Street
Hartford, CT 06103
United States of America

Email

request@validsoft.com

Phone

+1(888) 392-0230

Confidentiality and Disclaimer

This document may contain references to information that has been obtained from sources believed to be reliable. ValidSoft does not guarantee the accuracy, completeness or adequacy of such information, and shall have no liability for errors, omissions or inadequacies. The recipient assumes sole responsibility for the interpretation and use of this material for its intended results. Predictions and forward-looking statements in this document reflect current expectations concerning future events and are subject to risks and uncertainties, many of which are beyond the control of ValidSoft. ValidSoft undertakes no obligations to update these statements as a result of new information. Opinions expressed in this document are subject to change without notice.

© 2022 ValidSoft. All Rights Reserved.