WatersTechnology

Proliferation of Fake Information: Deepfakes & the Capital Markets

By Josephine Gallagher
January 17, 2020

Can you trust what you read or see online? Notice the irony as you read this story from—more than likely—a web browser on your laptop or mobile device. The truth is that this question is very relevant in the world we live today.

"I think people should be aware of the fact that what they see on the internet might not be true, since images and videos can be easily manipulated or even completely synthetically generated," says Luisa Verdoliva, a professor at the University of Federico II of Naples. "It is important because otherwise people just believe everything they see."

The warning signs of false information online and its power to influence came to international attention around four years ago during the 2016 US presidential election. It later emerged that the Russian government interfered in the election in "sweeping and systematic fashion" by spreading disinformation through a social media campaign aimed at "disparaging candidate Hillary Clinton" and releasing stolen documents tied to the Clinton campaign, according to the Special Counsel investigation conducted by Robert Mueller.

There are already great concerns for the November 2020 elections as bots have already been used to spread false information to a fake website created to look like Democratic candidate and former Vice President Joe Biden's official site. Even US Supreme Court chief justice John Roberts warned in his annual report on behalf of the federal judiciary that the "spread [of] rumor and false information" is threatening public confidence in government institutions. Guy Harrison, general manager of risk and compliance at Dow Jones, says the stakes for error or misunderstandings are high for risk and compliance departments. Piping fake news reports or false information into systems can lead to misinformed decisions that can result in reputational damages, fines, and loss of revenue.

"The reason it's so dangerous for our customers is that they are doing more and more straight-through processing on their side," Harrison says. "So, they're plugging those data sources directly into their processes without human intervention. And again, if it's in the system, it can really start to wreak havoc." Dow Jones has several hundred researchers responsible for verifying sources that are used to feed its risk and compliance solutions, such as its recently launched adverse media and monitoring solution. The firm also uses Factiva, an international news database that combines information from over 32,000 sources.

An Old Problem Evolved

There is a line belonging to the 17th-century writer Jonathan Swift that helps to show the perniciousness of disinformation: "Falsehood flies, and the truth comes limping after it." Swift's words were written more than 300 years ago, long before the digital age and superfast broadband. Today it is cheaper and easier to doctor videos or images using simple editing software or even a smartphone. For example, in May 2019, a clip of Nancy Pelosi, speaker of the US House of Representatives, was edited to make it appear as if her speech was slurred, causing speculation over her health and condition. This video was viewed millions of times on Facebook and Twitter before it emerged that the audio was adjusted, and the speed of the footage was simply slowed down.

Presently, this is not an issue on the minds of many banks and asset managers. This article is not about how capital markets firms are fighting false information; rather, it's to inform on developments in the space because this will become a more concerning issue for investment firms in the (potentially not-too-distant) future.

The reason for this concern is technological advancements have further evolved to produce an uneasy threat known as deepfakes, or machine-manipulated media. Deepfakes use deep learning, a sophisticated subset of machine learning, to manipulate media, such as videos, pictures, and audio. The artificial intelligence (AI) is trained using large amounts of data—video, images, and audio—of an individual, learning their facial features, behavior, and tone of voice. (See Understanding Deepfakes, below.)

Deepfakes are truly horrifying as they can be used to create dangerous videos of highly influential people, from President Donald Trump threatening airstrikes on a foreign nation or a celebrity's face being superimposed into a pornographic movie. Other instances include examples of fraud where deepfake voices are used to imitate senior level executives in firms to approve large corporate payment transactions.

"There was an example of about $250,000 where they used a deep-voice—I think it was the CEO to the chief financial officer or the CFO to the treasury person—but the person believed it was their boss and undertook the instruction," says Robert Tharle, fraud and authentication professional at NICE Actimize, a provider of financial crime solutions.

While some deepfakes are relatively easy to root out, the tech and sophistication are improving exponentially and the barrier to entry is lowering. Professor Luisa Verdoliva, a researcher in the space, is presently working on an automated benchmark for detecting facial manipulation, in conjunction with the Technical University of Munich, which is co-sponsored by Google. She says almost anyone with a graphics processing unit (GPU) with deep learning and large amounts of data, which can be downloaded from the internet, can produce deepfakes. A GPU with deep learning can cost anywhere from $1,000 to tens of thousands of dollars. Verdoliva adds that using a GPU takes very little training and someone could quickly learn how to use one from YouTube tutorials.

"It is easy nowadays to manipulate a video using deep learning, and although this is great in terms of technology [advancements], it can also be very scary because you don't need a lot of competency to use the software—you just need the data and the GPU, and the software will do everything," she says.

BOX: Arms Race

Today trying to detect deepfakes or manipulated media has become increasingly difficult. A year or two ago, you could spot visual defects or unnatural-looking behavior of an individual in an image or video, but now technology advancements mean that we can no longer—for the most part—rely on the human eye to look for signs of visual manipulation. For this reason, academic institutes, human rights organizations, and big tech companies from around the world are channeling resources into combating the threat of disinformation online.

On January 7, Facebook finalized new rules that bans users from posting deepfakes on the platform. Facebook also announced in September 2019 that it was partnering with Microsoft, the Partnership with AI, WITNESS, and academics from several universities and institutes to develop a Deepfake Detection Challenge (DFDC).

Today it also includes other organizations such as Amazon, BBC, CBC, First Draft, the New York Times and XPrize. The goal of the challenge is to build a technology that will enable everyone to identify when AI is being used to manipulate media to mislead the public. One of the biggest limitations to creating highly sophisticated detection tools has been the lack of a large enough dataset. As a result, Facebook launched the start of the DFDC project on December 11, 2019, along with the release of the full version of its deepfakes dataset, which uses media taken of paid actors. Verdoliva is one of the academics involved in the challenge and says that creating and publishing the large dataset will enable the industry to develop more effective technologies for detection.

"This can be helpful for training the algorithm in order to discover the manipulation," Verdoliva says. "Every deep learning-based method needs lots of data for training and to make the network learn what is pristine and what is forged. Large datasets of pristine and manipulated data can help to train the network to distinguish the two."

Similarly, Google is collaborating with Jigsaw to release its own large dataset of visual deepfakes, which will be incorporated into the University of Federico II Naples and University of Munich's FaceForensics benchmark. The two universities also conducted a study that looks at four types of manipulation across images and videos, two of the samples were modified by computer graphics (e.g. someone manually changing an individual's expression) and two were deepfakes (e.g. an AI swapping one person's face onto another). The results showed that deepfakes could be detected using deep learning, but not in cases where the media underwent a high level of compression.

Visual detection methods aren't enough as everyday technology continues to progress and hackers find new ways to outsmart algorithms. Therefore, the two universities are also looking at less obvious methods of manipulation known as multimedia forensics. Forensic analysis of digital images and videos can be done by analyzing the internal data or digital history of the media to see if it has been modified. The research proposes training neural networks to be able to suppress other values in the image, such as high-level scene content or other disturbances to recover the necessary characteristics that are considered the "camera's fingerprint" or "noiseprint". For example, if deep learning is used to manipulate an image or video, it creates synthetic data, which is unlike the original characteristics of the camera's fingerprint—such detail about the camera's sensor, optical lense and filter.

"Inside a camera there are several internal operations that leave some specific traces in an image or a video," Verdoliva says. "If a manipulation is carried out then these traces will be modified and an anomaly will arise.

This is an indication that that area was manipulated."

Authentication Case Study

Currently, many of the biggest technologies firms in the world—including Facebook, Google Twitter, Amazon and Microsoft—are developing tools to fight the spread of deepfakes. There is also a growing field of specialists entering the market.

Truepic, for example, is one of the vendors investing in authentication technologies, while others include Amber Video and eWitness. Mounir Ibrahim, vice president of strategic initiatives at Truepic, says there is no reliable way for people to verify existing media online unless it has been previously authenticated. This is a particularly significant challenge for media companies as they are sent large volumes of images and footage from events around

the world and have difficulty proving their legitimacy. One example of this dates back to October 2019, when ABC news mistakenly used footage alongside a report on Turkish attacks on Northern Syria against Kurdish civilians, when in fact the video was believed to have been taken from a nighttime gun demonstration in Kentucky that was originally published on YouTube in 2017.

"There is no way that the media company can have faith that the user-generated content they are receiving is, in fact, real, especially with the rates of image manipulation increasing," Ibrahim says. "And then, of course, there's deepfakes and all this advanced technology to manipulate imagery, that you cannot trust the actual imagery, which leads into a very dangerous predicament for media companies."

He adds that the reason image detection is incredibly difficult is largely because of the format of most pictures, a JPEG. This is because JPEGs are stripped of their metadata or are altered immediately when sent from one digital location to another. One example is when uploading a picture to Facebook, the image is automatically compressed to a lower quality.

Truepic, founded in late 2014, has developed a solution that authenticates media at the point of creation, and then curates and crowdsources them into its Truepic Vision platform. Once an image or video is captured using its Controlled Capture technology, it undergoes over 20 verification tests including the integrity of the capture device, the date and time, pixilation, the location of capture, and whether an image is a recaptured image of an existing one. Only those media that pass the tests are deemed authentic. The platform also uses sophisticated algorithms to detect any spoofings and cross references available data sensors on the device, such the time on a phone, the server time, and the time zone using GPS.

All the data from the capturing device is encrypted and transmitted to its servers in the cloud without any breaks in transmission to prevent people from going offline and trying to alter the media at any point. Following the authentication process, a unique cryptographic signature—also known as a SHA256 hash—is then created and written onto a public blockchain for anyone to access.

"When all of these tests are completed, we have the ability to log this critical information onto the blockchain for immutability," Ibrahim says. "This process all happens in 10 seconds or less. So, the second something is taken, we prove it's real, put it in a safe place, and then let you then share that."

Anyone looking to then verify an image or video captured in the system can then cross-check it by running the same algorithms and ensure it produces the same code.

Although, authentication technologies and large databases of verified media sounds like a promising solution, there is one big challenge yet to overcome—scalability. For these types of products to work, they require mass adoption on a global scale because all the images and videos have to be captured using the vendor's technology. Truepic it seems, has already thought of this—they are partnering with major chip manufacturer Qualcomm Technologies.

"The idea is that we will move our technology into the actual hardware of cell phones or smartphones, so that at some point in the future, you will have the option to capture a verified image directly from the native camera in the phone," Ibrahim says. "That's how you go about addressing the issue of scale—you move from a software to a hardware."

Truepic hopes to have real-life prototypes in production within the next two years.

Closer to Home

Although many financial market firms might not be considering how false news and deepfakes will impact their workflows today, there is a definite need to recognize the risks they will pose in the future, particularly as the appetite for new and alternative sources of data increases and AI systems are trained in less supervised environments. Another risk to consider for the trading floor is the need to react quickly to market disruptions, meaning verifying a data source can sometimes become an afterthought.

Additionally, as noted earlier by Dow Jones' Harrison, middle-office teams are also under increasing pressure to ensure they do their due diligence with regards to risk and compliance, as they are responsible for ensuring that the firm is using reliable, credible, and verifiable sources of data for informing decisions, as guided under organizations such as the Financial Action Task Force and local regulators.

Other considerations for financial market firms include the bias in media reports or other forms of online information. Many media outlets have political orientations, which need to be considered when selecting data sources. Some might include opinion columns, satirical publications, or less credible news sites. Furthermore, just as what happened during the bring-your-own-device era, firms need to worry about workers bringing malicious apps into the office. Recently, popular apps ZAO and FaceApp, which use deepfake technology, made headlines because of concerns over user privacy and information leakage.

Abraham Thomas, co-founder and chief data officer at Quandl, an alternative data provider recently acquired by Nasdaq, says false information can also incorporate examples such as obfuscated data in company filings or manipulating ratings on performance websites. Firms, for example, might choose to reformat some of their data for competitive reasons, from a text into an image file to make it harder for computer platforms to ingest the data. Additionally, company review websites like Glassdoor are subject to manipulation if a competitor makes several negative reviews under its name to bring down their performance score.

More worrying, of course, is the damaging impact of fake images or videos to a company's reputation and valuation.

"It's entirely possible that somebody might create a fake video of a CEO of a major company doing something utterly reprehensible, and that could cause the company's stock price to fall—but we haven't seen that, yet," Thomas says.

The word "yet," is worth emphasizing.

BOX: Understanding Deepfakes

Deepfakes—a portmanteau of deep learning and fake—are machine-manipulated media, the most common example being one person's face superimposed onto another. They are produced using a model called generative adversarial networks (GANs), which uses deep-learning techniques and neural networks to alter media content.

Using large volumes of data—images, videos, or audio—to train from, the GAN will generate an artificial output. This is done by the first neural network, known as the generator. Then,

using the second neural network, known as the discriminator, the GAN will determine whether the outputs are real by comparing them with the training data samples.

This process is continued back and forth between the generator and the discriminator until the discriminator cannot distinguish the difference between the output created and the training samples, thus creating the final product—the deepfake.

While deepfake technology has been used widely in movies—and that is only going to increase—for some eerie examples, search Jordan Peele and Barack Obama deepfake or, even weirder, Bill Hader and Tom Cruise deepfake.