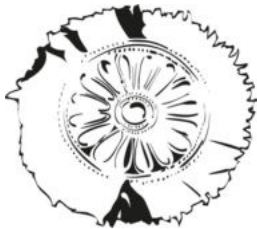


Content, Context and Propagation Approaches for Fighting Disinformation

Yiannis Kompatsiaris, CERTH-ITI, Director



Centre for Research and Technology Hellas



CERTH
CENTRE FOR
RESEARCH & TECHNOLOGY
HELLAS



Founded in 2000,
One of the leading
research institutions in
Greece

>1200 projects
1100 international
collaborations
5 institutes
>1000 employees

In TOP-15 E.U.
research institutions
in competitive
research grants

Centre for Research and Technology Hellas - Information Technologies Institute



The **largest** among
the five institutes of
CERTH

>500 employees

> 175 EU projects

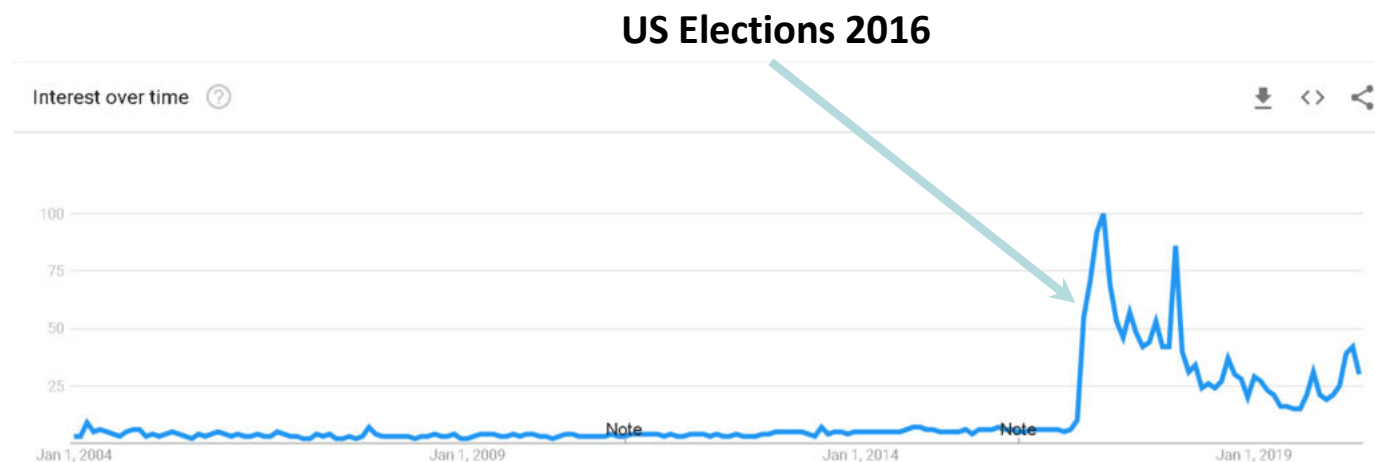
>100 national and
industrial projects

> 40M € via R&D
activity -last 5 years

Infrastructure
HPC, Smart Home,
Robots,
Autonomous
Vehicle, Drones,
EEG, 3D Printing,
VR

The Rise of Fake News

Volume for query “fake news” over time: A key milestone has been the US Elections in 2016, which marked the beginning of large-scale coordinated disinformation campaigns.



<https://trends.google.com/trends/explore?date=all&geo=US&q=fake%20news>

A challenge...
...as old as time

{disinformation}

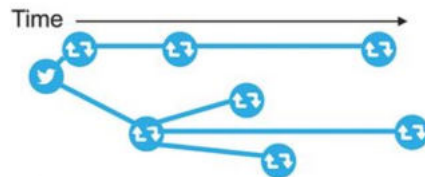
{misinformation}

high cost in both cases

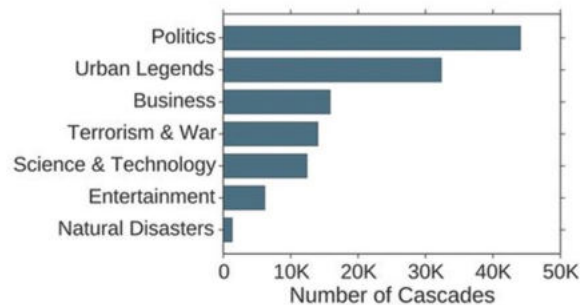


The Diffusion of Fake News

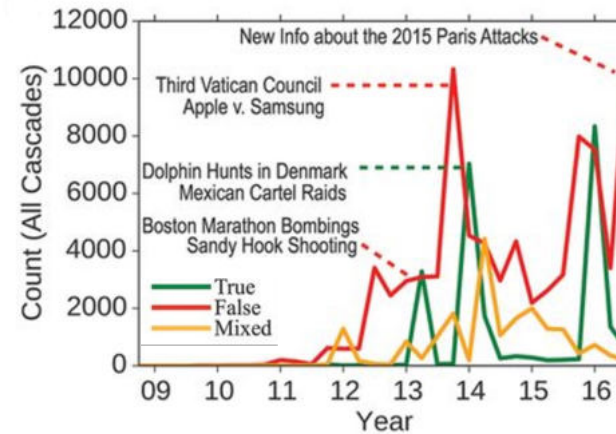
Example cascade



Topic frequency



Number of cascades



Misleading posts tend to spread faster and wider compared to accurate ones.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146-1151.

Visual disinformation is dangerous

- More persuasive than text
- Attracts more attention
- More tempting to share
- Can easily cross borders

Hameleers, M., Powell, T. E., Van Der Meer, T. G., & Bos, L. (2020). **A picture paints a thousand lies? The effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media.** *Political Communication*, 37(2), 281-301.

Dan, V., Paris, B., Donovan, J., Hameleers, M., Roozenbeek, J., van der Linden, S., & von Sikorski, C. (2021). **Visual mis-and disinformation, social media, and democracy.** *Journalism & Mass Communication Quarterly*, 98(3), 641-664.

Thomson, T. J., Angus, D., Dootson, P., Hurcombe, E., & Smith, A. (2020). **Visual mis/disinformation in journalism and public communications: current verification practices, challenges, and future opportunities.** *Journalism Practice*, 1-25.

...seeing is believing

The Famous Shark

2005



<https://www.snopes.com/photos/animals/puertorico.asp>



Maury Page
@mopage19

Follow

A shark photographed on I-75 just outside of Naples, FL

This is insane. [#HurricaneIrma](#)

8:12 PM - Sep 10, 2017

1,506 5,538 12,545



Jason Michael
@Jeggit

Follow

Believe it or not, this is a shark on the freeway in Houston, Texas. [#HurricaneHarvy](#)

9:00 AM - Aug 28, 2017

7,168 88,805 149,387



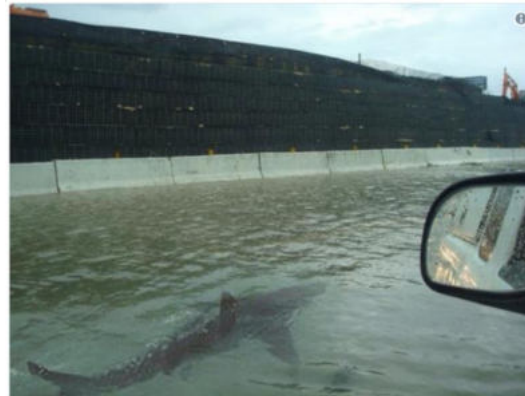
SoaR Gilli **Logan**
@Mcgilligan

Follow

A shark was pushed inland from [#HurricaneMatthew](#)

7:27 PM - Oct 7, 2016

22 263 506



Austin.
@austinelement

Follow

Shark in road [#sandy](#)

2:53 AM - Oct 30, 2012

6 86 13

DeepFakes

- Content, generated by deep neural networks, that seems authentic to human eye
- Most common form: generation and manipulation of human face



Source: [Media Forensics and DeepFakes: an overview](#)



Source:

<https://www.youtube.com/watch?v=iHv6Q9ychnA>



Source: <https://en.wikipedia.org/wiki/Deepfake>

DeepFakes Generation

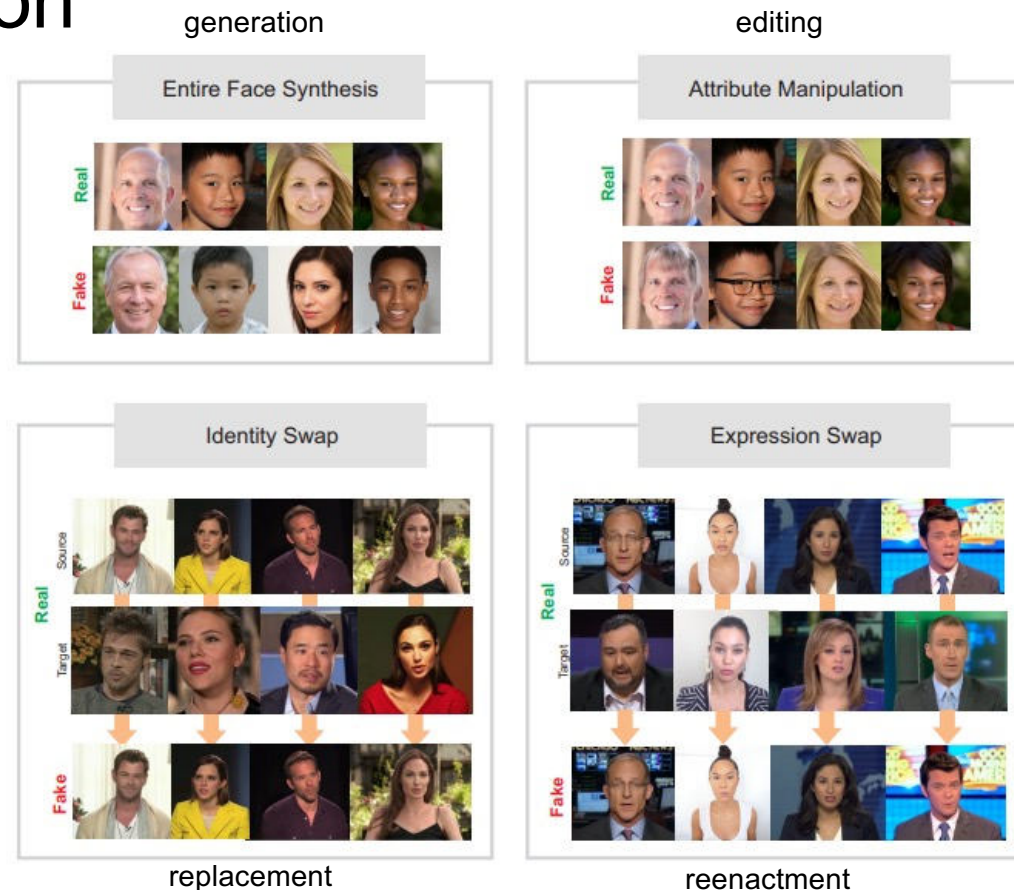
Four main types of face DeepFakes:
a) *Entire face synthesis*, b) *Attribute manipulation*, c) *Identity swap*,
d) *Expression swap*.

Lip syncing and *voice generation* are also common types in video and audio content.

Tolosana, R., et al. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64, 131-148.

Verdoliva, L. (2020). Media forensics and deepfakes: an overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5), 910-932.

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)*, 54(1), 1-41.



Source: *DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection* (Tolosana et al., 2021)

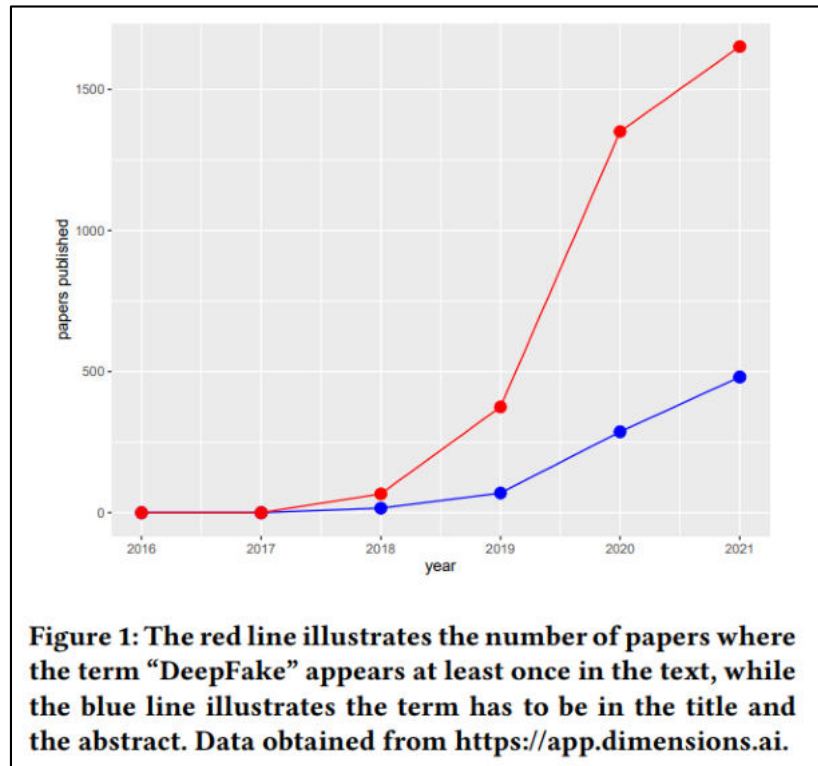
A New Level of Realism

- Created by Chris Ume, a VFX specialist
- Not detected by any of the commercial detection services
- Not discernible by human inspection
- Potential for misleading but to date barriers are still high
- a lot of expertise, skill and time
- an impersonator who looks like the target (Miles Fisher)

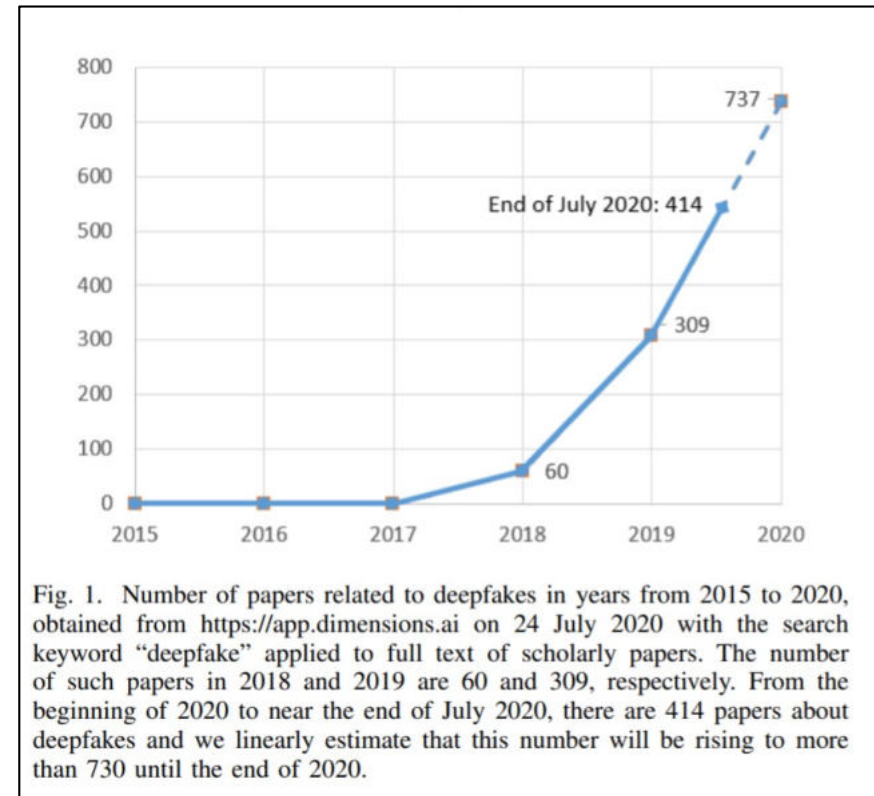
<https://www.theverge.com/2021/3/5/22314980/tom-cruise-deepfake-tiktok-videos-ai-impersonator-chris-ume-miles-fisher>



Gaining popularity



Baxevanakis, S., et al. (2022). The MeVer DeepFake Detection Service: Lessons Learnt from Developing and Deploying in the Wild. Submitted to ICMR MAD 2022



Nguyen, T. T., et al. (2019). Deep learning for deepfakes creation and detection. *arXiv preprint arXiv:1909.11573*, 1.

Potential Risks and Harms

| Psychological harm | Financial harm | Societal harm |
|---|---|---|
| <ul style="list-style-type: none">• (S)extortion• Defamation• Intimidation• Bullying• Undermining trust | <ul style="list-style-type: none">• Extortion• Identity theft• Fraud (e.g. insurance/payment)• Stock-price manipulation• Brand damage• Reputational damage | <ul style="list-style-type: none">• News media manipulation• Damage to economic stability• Damage to the justice system• Damage to the scientific system• Erosion of trust• Damage to democracy• Manipulation of elections• Damage to international relations• Damage to national security <p>Damage to Public Health</p> |

[Tackling deepfakes in European policy](#), Panel for the Future of Science and Technology, Scientific Foresight Unit (STOA), July 2021

Wired

BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY

SIGN IN

SUBSCRIBE

A new podcast from the most trusted voice in music

THE Pitchfork REVIEW


SUBSCRIBE

MATT BURGESS, WIRED UK

SECURITY 11.18.2020 09:00 AM

Telegram Still Hasn't Removed an AI Bot That's Abusing Women

A deepfake bot has been generating explicit, non-consensual images on the platform. The researchers who found it say their warnings have been ignored.



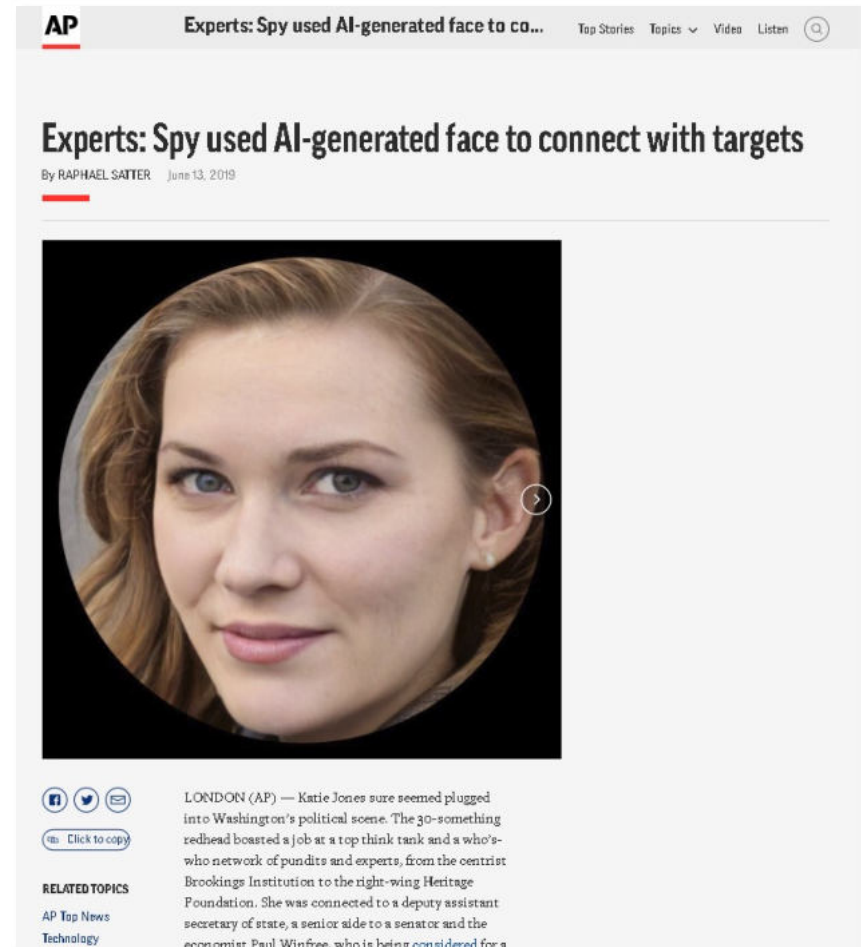
"The bot uses a version of the DeepNude AI tool, which was originally [created in 2019](#), to remove clothes from photos of women and generate their body parts. Anyone can easily use the bot to generate images. More than 100,000 such images have been publicly shared by the bot in several Telegram chat channels associated with it. "

<https://www.wired.com/story/telegram-still-hasnt-removed-an-ai-bot-thats-abusing-women/>

Fake Identities

But Katie Jones doesn't exist, The Associated Press has determined. Instead, the persona was part of a vast army of phantom profiles lurking on the professional networking site LinkedIn. And several experts contacted by the AP said Jones' profile picture appeared to have been created by a computer program....

<https://apnews.com/article/bc2f19097a4c4fffaa00de6770b8a60d>

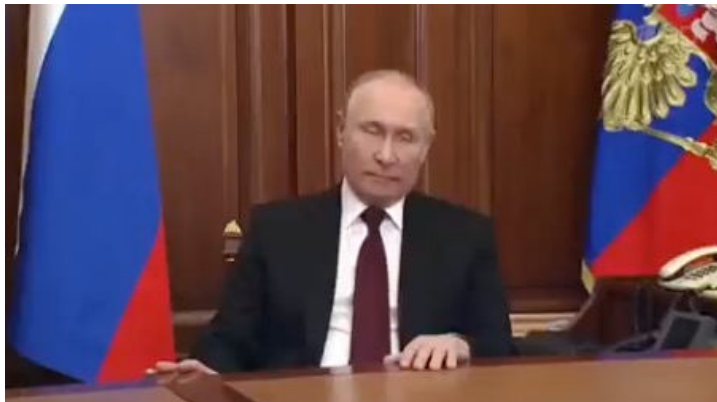


DeepFakes and National (cyber)Security



A 68-second deepfake video appeared in March 2022 during the third week of Russia's invasion in Ukraine, depicting Ukrainian President Volodymyr Zelenskyy calling for the surrender of arms. The video appeared on a compromised Ukrainian news web site and was then widely circulated on social media.

<https://nypost.com/2022/03/17/deepfake-video-shows-volodymyr-zelensky-telling-ukrainians-to-surrender/>



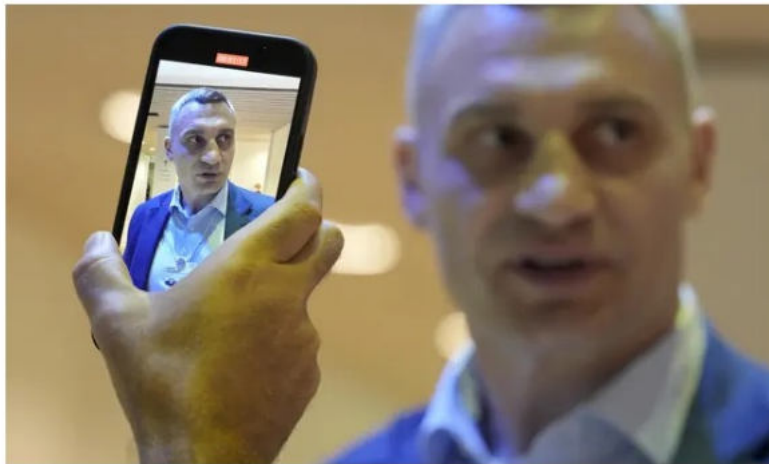
A few days later a video was circulated on social media that supposedly showed Russian President Vladimir Putin announcing that the Russian military was surrendering. A tweet sharing the video with a caption prompted Russian soldiers to lay down their weapons and go home.

<https://www.snopes.com/fact-check/putin-deepfake-russian-surrender>

DeepFakes and National (cyber)Security

European politicians duped into deepfake video calls with mayor of Kyiv

Person who sounds and looks like Vitali Klitschko has spoken with mayors of Berlin, Madrid and Vienna



Someone has been impersonating the mayor of Kyiv, Vitali Klitschko – the real one seen here.
Photograph: Markus Schreiber/AP

The mayors of several European capitals have been duped into holding video calls with a deepfake of their counterpart in Kyiv, **Vitali Klitschko**.

The mayor of Berlin, Franziska Giffey, took part in a scheduled call on the Webex video conferencing platform on Friday with a person she said looked and sounded like Klitschko.

... The mayor of Berlin, Franziska Giffey, took part in a scheduled call on the Webex video conferencing platform on Friday with a person she said looked and sounded like Klitschko.

“There were no signs that the video conference call wasn’t being held with a real person,” her office said in a statement.

<https://www.theguardian.com/world/2022/jun/25/european-leaders-deepfake-video-calls-mayor-of-kyiv-vitali-klitschko>

Face Swapping apps: DeepFakes going Mainstream

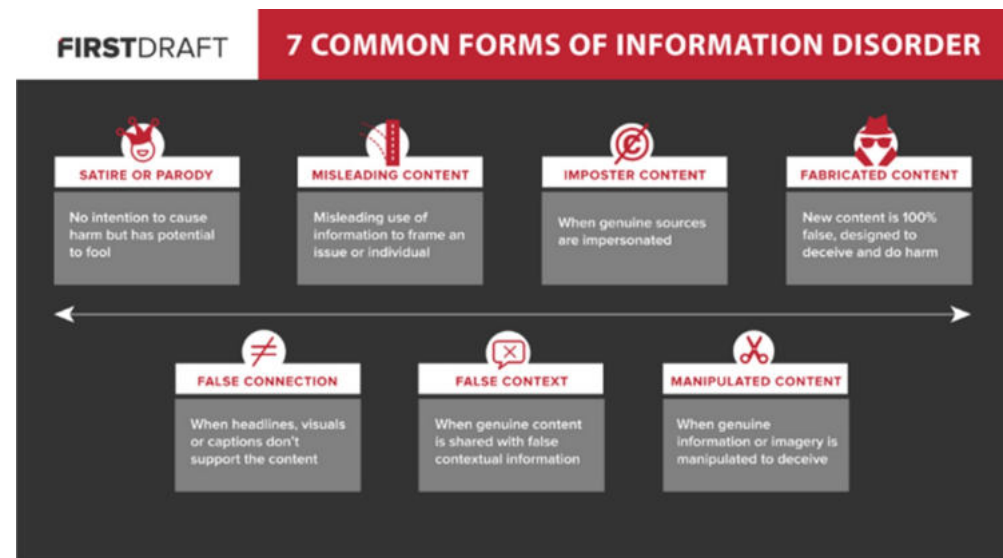
“The app [Reface] normalises deepfakes, and not everyone understands the concerns arising from them because not everyone has the digital know-how to differentiate what is real and what isn’t,” Apurva Singh, a privacy expert and volunteer legal counsel at Software Freedom Law Center, India....



<https://www.vice.com/en/article/wxqkbn/viral-reface-app-going-to-make-deepfake-problem-worse>

Visual disinformation comes in many forms

- Manipulated photos/video
- Deepfakes
- Visuals out of context
- False connections
- Visual memes



<https://medium.com/1st-draft/information-disorder-part-3-useful-graphics-2446c7dbb485>

...many different methods and tools are needed

MeVer tools deal with multiple information aspects

Content

- Image Verification Assistant
- DeepFake Detection
- Visual Location Estimation

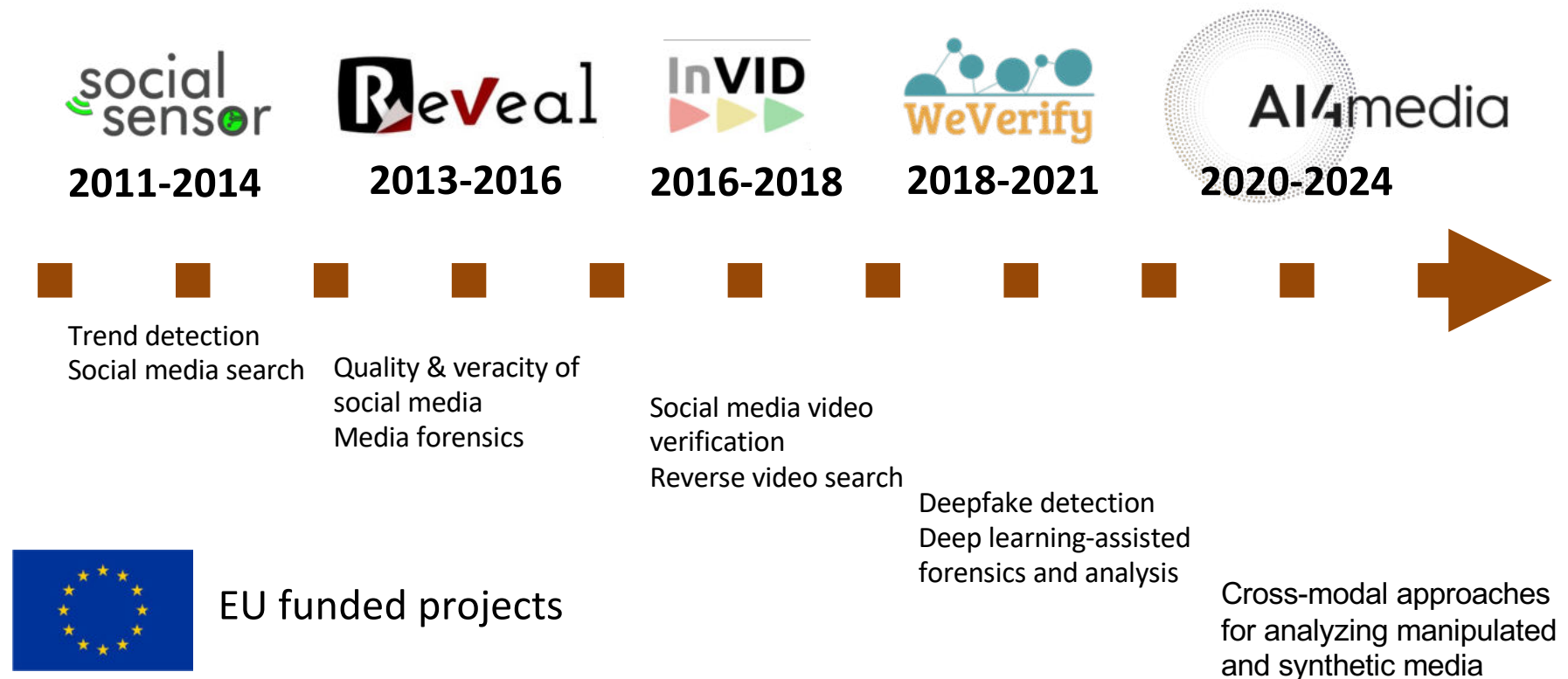
Context

- Near-duplicate Detection
- Context Aggregation and Analysis

Propagation

- Network Analysis and Visualization

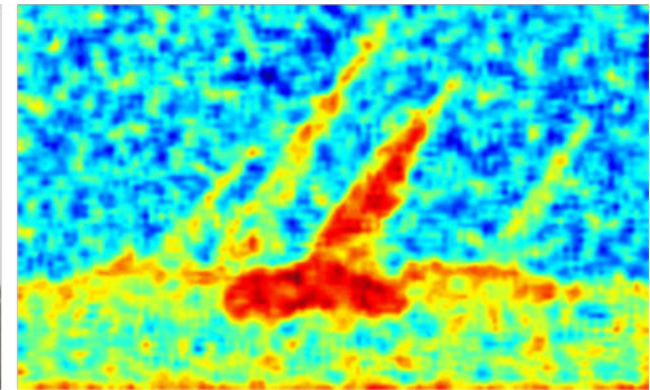
MKLab “historical” background



Content Verification

Digital image manipulation

- Image tampering localisation: highlight areas in an image that have been digitally manipulated
- Types of tampering:
 - **Splicing**
 - **Inpainting**
 - **Copy-move**
 - Cropping
 - Enhancement



<https://www.npr.org/templates/story/story.php?storyId=92442928>

Types of tampering localisation algorithms

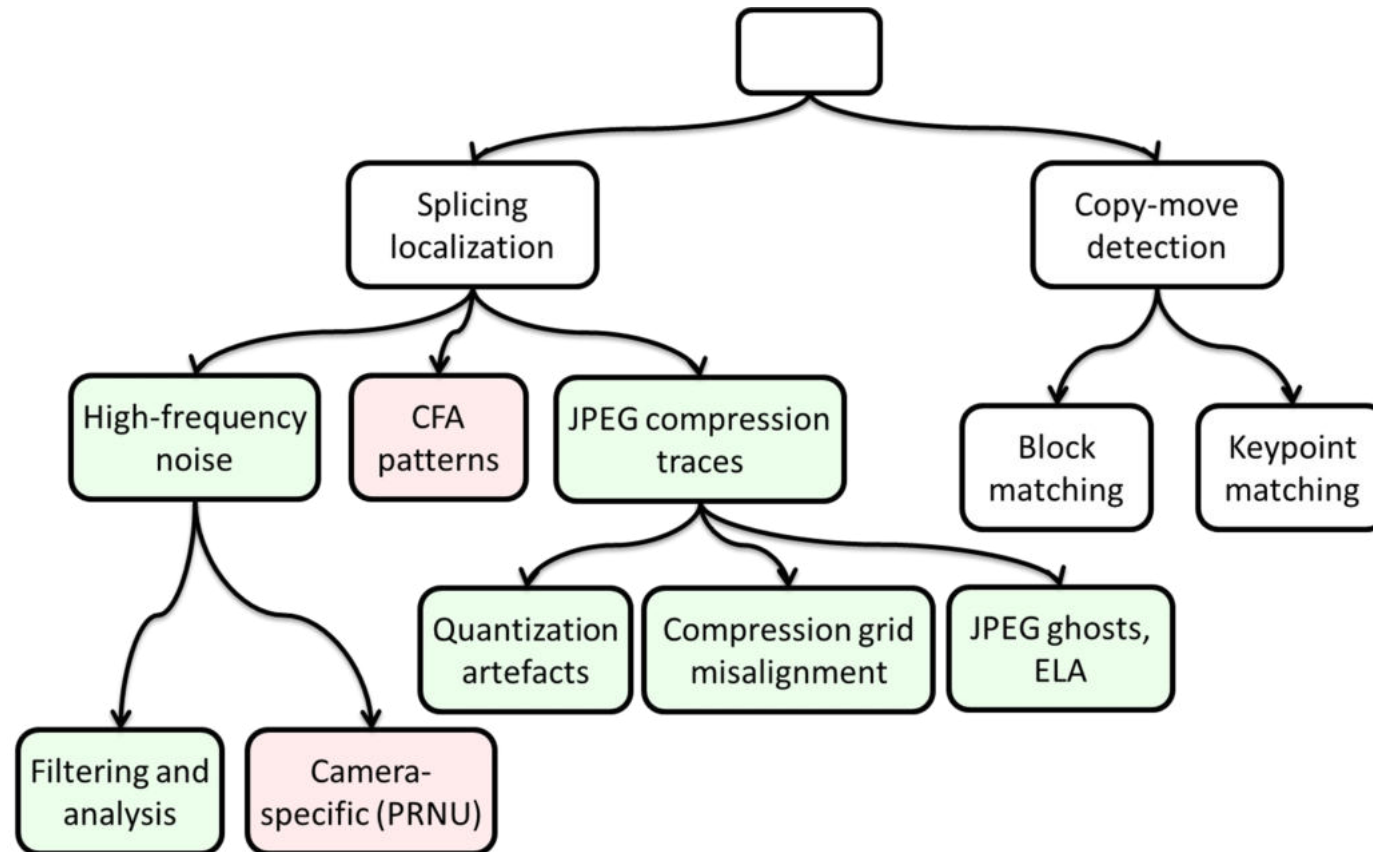
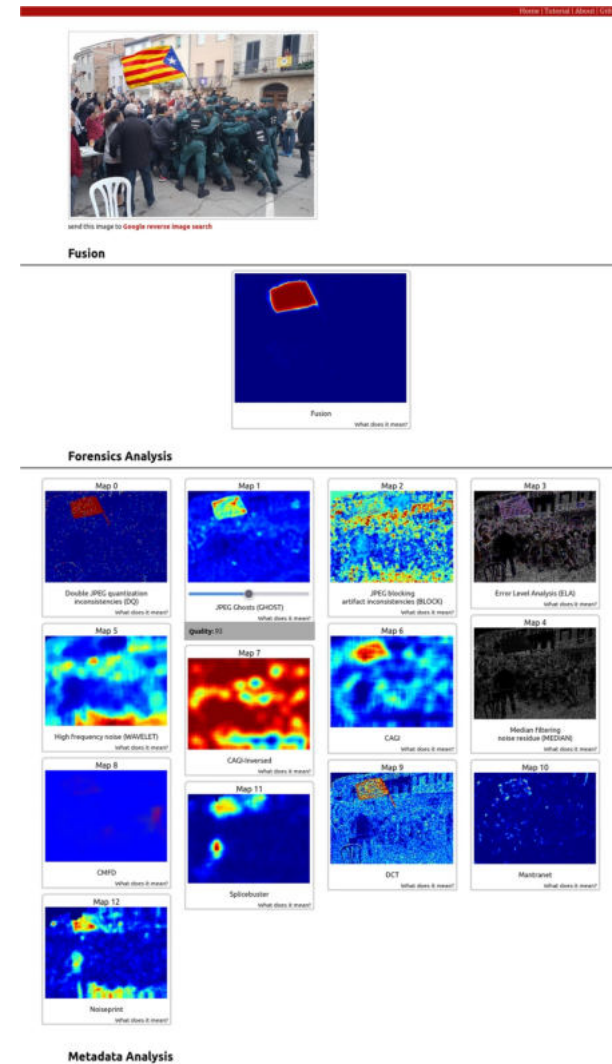


Image Verification Assistant

- Integrates 12 image forensics algorithms and a fusion algorithm
- Allows metadata inspection (EXIF, IPTC, Photoshop and more)
- Estimates the software/hardware origin of JPEG images
- Supports quick reverse image search on Google

<https://mever.itl.gr/forensics/>

Zampoglou, M., Papadopoulos, S., Kompatsiaris, Y., Bouwmeester, R., & Spangenberg, J. (2016, April). Web and Social Media Image Forensics for News Professionals. In *SMN@ ICWSM*.

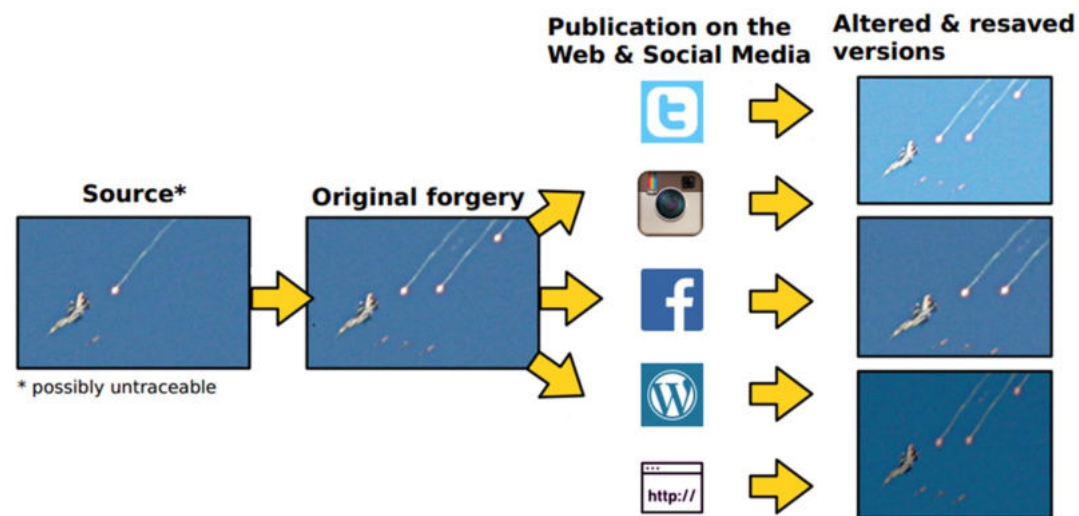


Integrated algorithms

- Error Level Analysis (ELA) (Krawetz, 2007)
- Inconsistencies of JPEG Blocking Artifact (DCT) (Ye et al., 2007)
- JPEG ghosts (Farid, 2009)
- Double JPEG quantization inconsistencies (Lin et al., 2009)
- Median filtering noise residue
- Inconsistencies of JPEG Blocking Artifact (BLK) (Li et al., 2009)
- High-frequency noise analysis (Wavelet) (Mahdian & Saic, 2009)
- SpliceBuster (Cozzolino et al., 2015)
- ➡ • **CAGI** (Iakovidou et al., 2018)
- ➡ • **MantraNet** (Wu et al., 2019)
- ➡ • **Copy move forgery detection** (Wu et al., 2018)
- Noiseprint (Cozzolino et al., 2018)
- ➡ • **Fusion** (Charitidis et al., 2021)

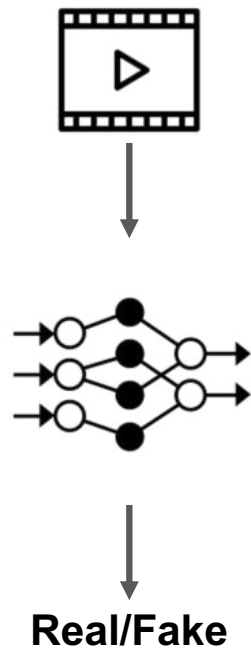
Caveat: Effectiveness of image forensics algorithms

- Certain image forensics algorithms work under very specific conditions
- Internet and social media images are particularly challenging due to multiple recompressions - often applied by the sharing platforms



Zampoglou, M., Papadopoulos, S., & Kompatsiaris, Y. (2015). Detecting image splicing in the wild (web). In *International Conference on Multimedia & Expo Workshops (ICMEW), 2015* (pp. 1-6). IEEE

DeepFake Detection



Common architectures

- CNN
- Visual Transformers (ViT)
- Capsule Networks

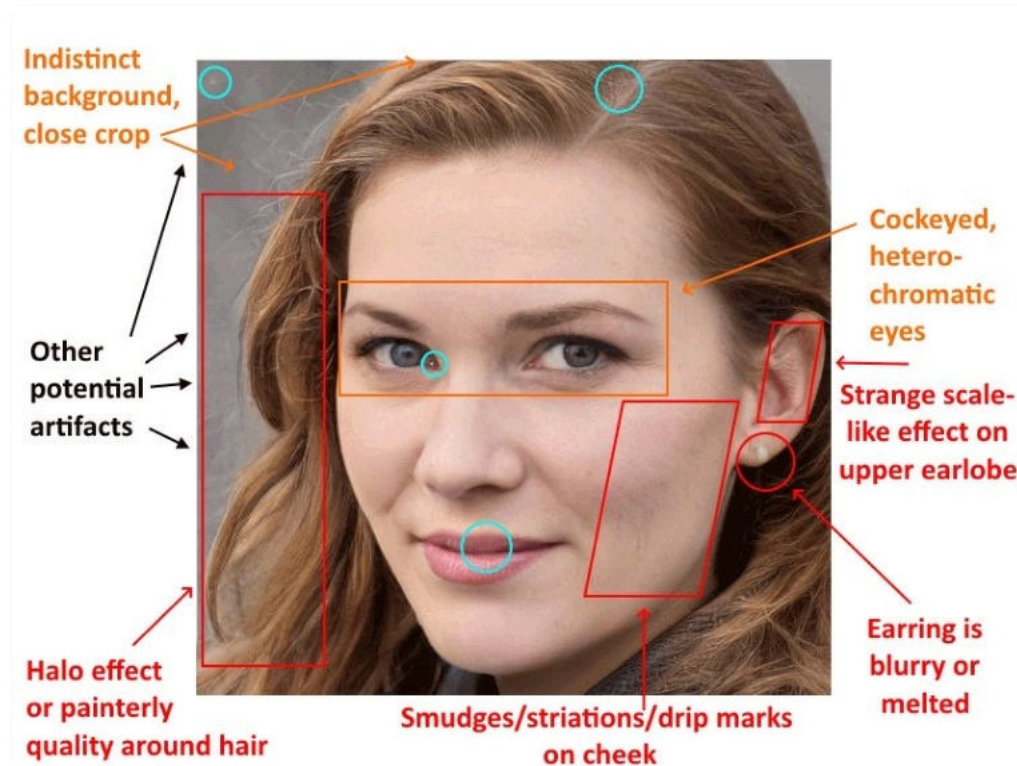
Key Ideas

- Detect abnormal changes in physiological signals e.g. **head poses**, **eye blinking**.
- Exploit left-over manipulation artifacts

Main Problem

Poor generalization to new manipulation techniques.

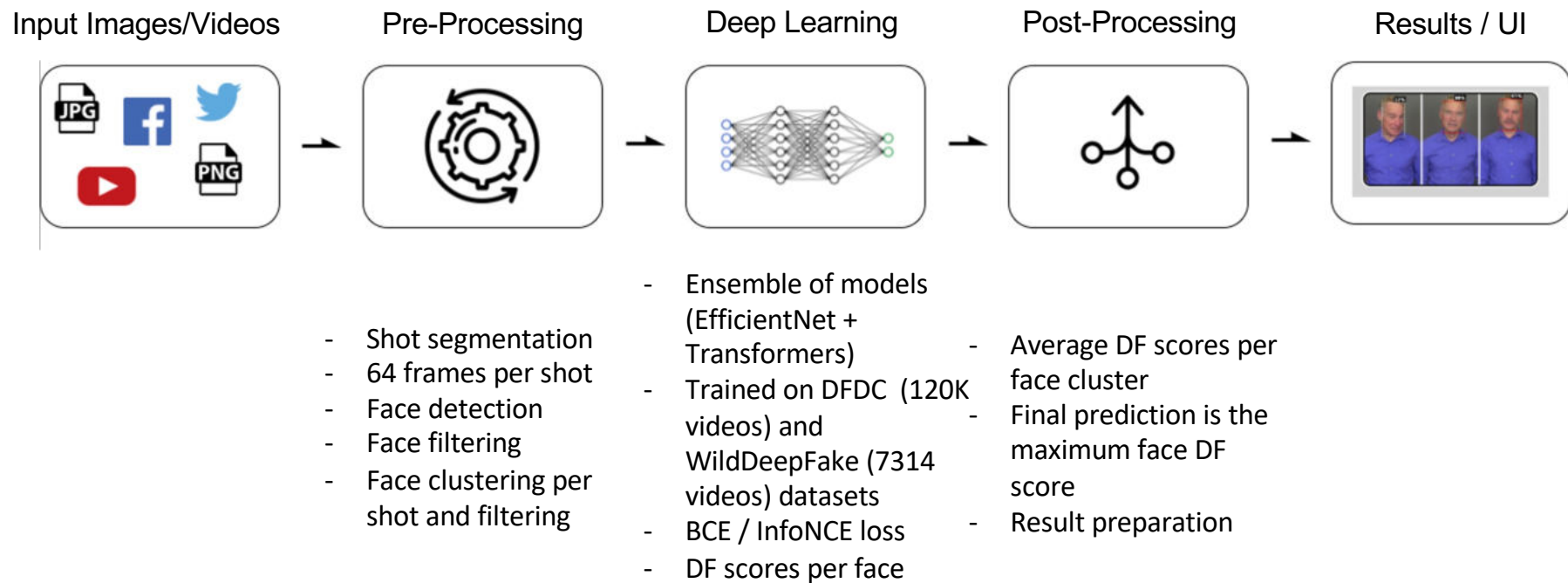
Signs of a DeepFake (in 2020)



<https://apnews.com/article/bc2f19097a4c4fffaa00de6770b8a60d>

- Different kinds of artifacts
- Blurry areas around lips, hair, earlobes
- Lack of symmetry
- Lighting inconsistencies
- Fuzzy background
- Flickering (in video)

Overview of Approach



P. Charitidis, G. Kordopatis-Zilos, S. Papadopoulos and I. Kompatsiaris. "Investigating the Impact of Pre-processing and Prediction Aggregation on the DeepFake Detection Task". In Proceedings of the Truth and Trust Online, 2020.

DeepFake Detection Services

- DeepWare: online DeepFake scanner and Android application
- DuckDuckGoose: DeepFake detection system and chrome plugin (only for images)
- DeepFake-o-meter: accepts video link or file and results are sent to user's email

<https://deepware.ai/>

<https://duckduckgoose.ai/>

<http://zinc.cse.buffalo.edu/ubmdfl/deep-o-meter/>

The MeVer DeepFake Detection Service

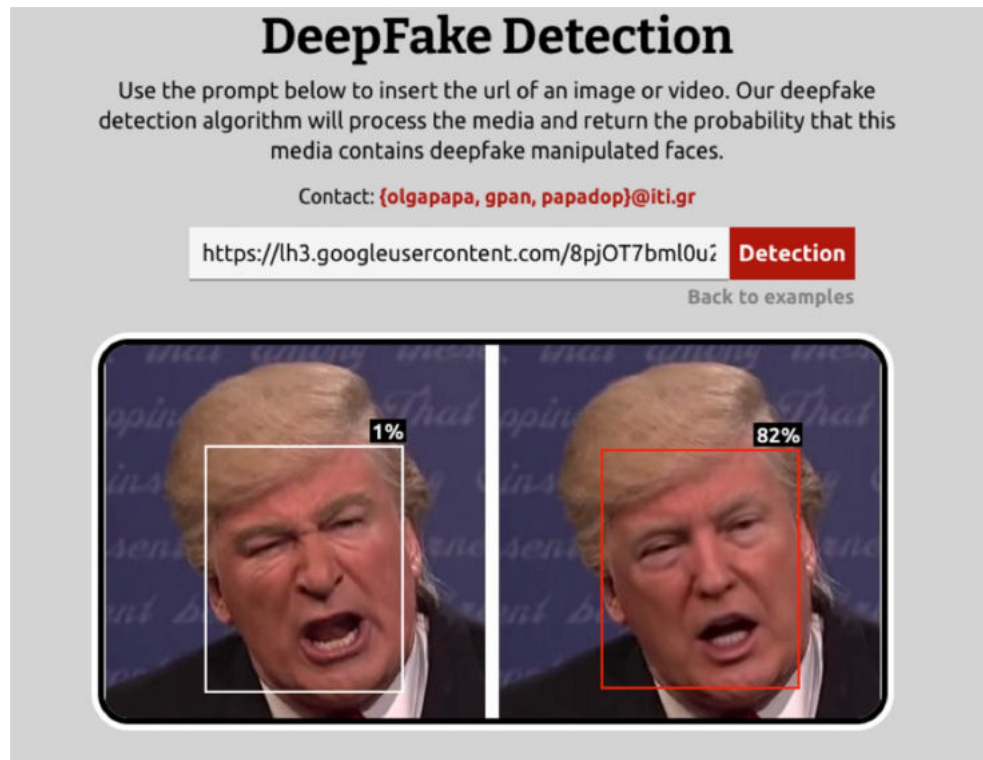
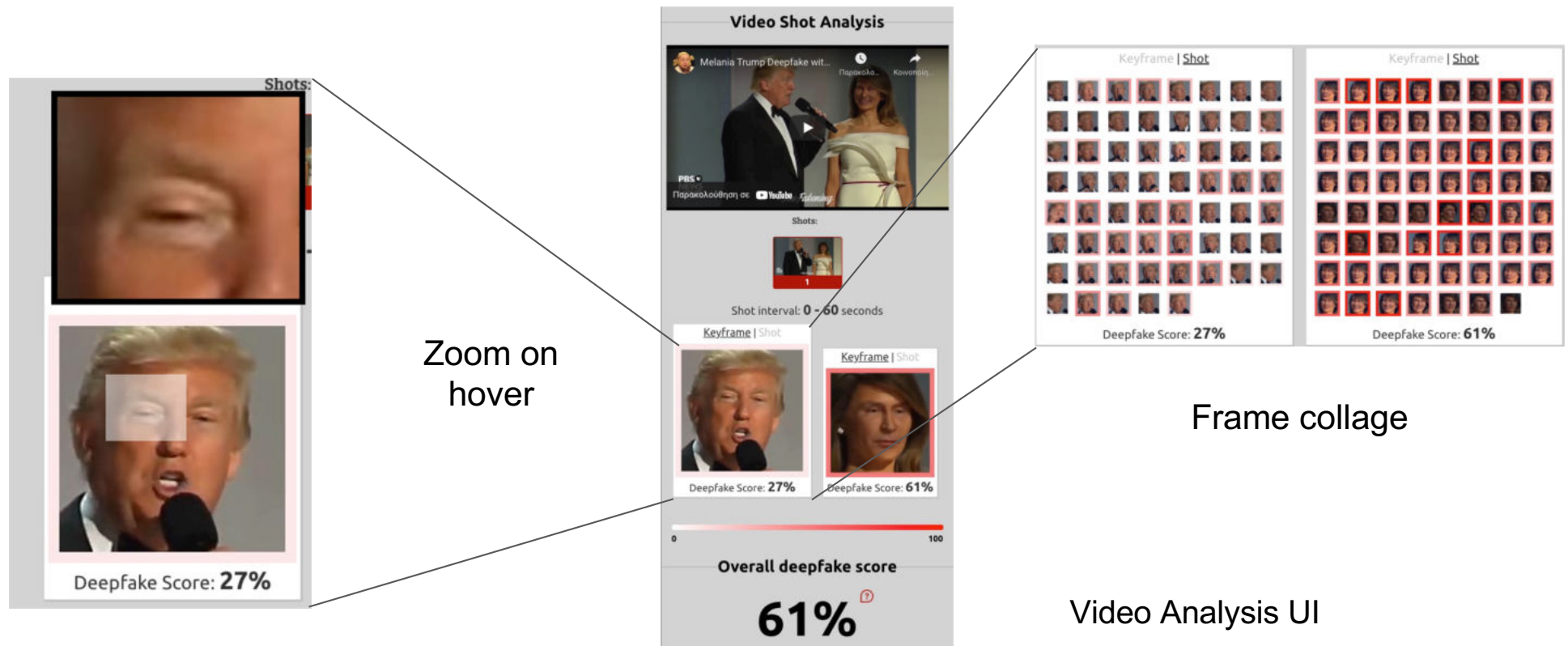


Image Analysis UI

<https://mever.iti.gr/deepfake>

The MeVer DeepFake Detection Service



Evaluation

Adversarial Robustness

Evasion attack: perform targeted alterations to an image

Projected Gradient Descent:
white-box evasion attack

Make use of the [Adversarial Robustness Toolbox \(ART\)](https://adversarial-robustness-toolbox.readthedocs.io/en/latest/) by IBM

| Dataset | norm-1 | norm-2 | norm-inf |
|-----------------|--------|--------|----------|
| FaceForensics++ | 70.31% | 64.04% | 50.53% |
| CelebDF | 82.75% | 76.01% | 50.00% |
| WildDeepFake | 84.94% | 63.04% | 50.00% |



<https://adversarial-robustness-toolbox.readthedocs.io/en/latest/>

Model Card

- inform & guide new users
- contains:
 - model architecture details
 - datasets used
 - evaluation results
 - versioning scheme
 - caveats and recommendations
 - factors that affect performance

Model Card - DeepFake Detection Service

Model Details

- Developed by: CREDIT-ITI Media Verification Team
- Model date: 02/02/2022
- Model version: 1.0. In this version, an ensemble of five models is deployed. Compared to previous versions, one model has been added, and several functionalities have been refactored to improve robustness.

Processing pipeline:

- Download the image/video from the input URL.
- In the case of images:
 - Use a Face Detector to detect all faces in the image.
 - Feed each face to the model ensemble to get a DeepFake probability score in range [0, 1].
- In the case of video:
 - Segment the input video into shots.
 - For each shot, use a Face Detector to detect faces in the shot's frames.
 - Perform a Face Clustering scheme to discard wrongly detected faces from the detector and organize the remaining faces into groups.
 - Feed each face to the model ensemble to get a DeepFake probability score in range [0, 1].
 - Use an Aggregation Strategy to derive a video-level DeepFake probability for the entire video.
 - The face predictions of each face cluster are averaged to generate a cluster prediction.
 - Segment predictions derive based on the maximum prediction of their clusters.
 - The final video-level prediction is the maximum segment prediction.

Model input: Video or image url.

Model output: The video-level DeepFake probability, and the probability for each detected person in each video shot. Probabilities closer to 0 means real and closer to 1 means fake.

Model type:

- DeepFake prediction: a five model ensemble is used:
 - a vanilla EfficientNet-b4.
 - a Transformer head based on DETR with fixed positional embeddings on top of an EfficientNet-b4.
 - a Transformer head based on DETR with learned positional embeddings on top of an EfficientNet-b4.
 - a Multi-head Transformer based on DETR on top of an EfficientNet-b4.
 - a vanilla EfficientNet-V2s.
- Face Detector: we use the faceconv-pytorch library.
- Face Clustering: we employ the method described in this paper, where we extract face features using the pretrained InceptionResNetV1 provided in faceconv-pytorch library, and used DISCAN for clustering.
- Shot segmentation: the feature extraction and similarity calculation described in this paper are used to extract peaks in the graph of distances of the consecutive frames.

Citation details: (CREDIT-ITI Media Verification Team, 2022) Media DeepFake Detection service.

Feedback & Contact: @pineluxx, georgioskardas, papadogi@iti.gr

Intended Use

- Primary intended use: Detect whether the faces present in the image or video from the provided URL have been manipulated using Deep Learning methods (DeepFake).
- Primary intended users: Journalists and media verification companies/organizations/groups.

Out-of-scope uses:

- The service cannot detect audio manipulations.
- The service cannot detect if the images/videos have been tampered with using non-facial manipulations or other forgeries (e.g. splicing, copy-move, inpainting).
- The service does not provide localized predictions on the extracted faces.
- The service does not process videos longer than 12 minutes containing more than 50 shots due to reliability issues. Refer to the *Caveats and Recommendations* section.

Relevant Factors

- Factors for which service performance may vary are:
 - Manipulations: whether the networks have been trained with the generated DeepFake manipulation method or not. Refer to the *Training Data* section for more information.
 - Background faces: if there are many background low-resolution faces in the input image/video, it may affect the service's final prediction because it treats all detected faces equally.
 - Image/Video quality: blurry or low quality faces can affect the predictions.
 - Adversarial Attacks: alterations in the images/video to evade detection can affect service performance.

Metrics

- Model performance measures:
 - Balanced Accuracy: defined as the mean of the recall computed on each class.
 - AUC Area Under the Receiver Operating Characteristic
- Metrics decision: since the evaluation datasets are unbalanced, we want to avoid skewed metrics that might favor one class or alter the datasets (e.g. see sampling to balance the dataset).
- Decision threshold: a face prediction greater than 0.5 is considered Fake whereas a prediction lower or equal than 0.5 is considered Real.

Relevant Datasets

- FaceForensics++ (FF++): The dataset is organized in two manipulation categories, Identity Swap and Expression Swap, each of which have two manipulation methods, FaceSwap and DeepFakes, and NeuralTextures and FaceFusion, respectively. It contains 1k videos for each manipulation derived from the combination of 1k real videos. Evaluation on FF++ provides a performance indicator on different manipulation categories and methods. It should also be mentioned that compared to more recent datasets (e.g. CddiDF, DFDC), the DeepFake quality in FF++ is really poor.
- CddiDF V2 (CddiDFV2): Compiled of videos from celebrity interviews that have been manipulated using an improved version of the DeepFake manipulation method used in the FF+++. It consists of 390 real and 5039 fake videos.

DeepFake Detection Challenge (DFDC): Published by Facebook in the context of a DeepFake Detection Challenge, it contains 20K videos from hundreds of paid actors that have been used to generate 100K manipulated videos using improved DeepFake, FaceSwap methods, and three GAN-based manipulations. Due to its size and quality, it is often used both in research and production.

WildDeepFake (WDF): This is one of the most recent datasets (2021) and in contrast to the previously mentioned datasets where the manipulations were applied automatically, it contains real-world DeepFakes scraped from various video-sharing websites as well as their corresponding real versions. It consists of 3.8k real and 3.8k fake videos. Due to its real-world nature, it is considered a challenging dataset.

Evaluation Data

- Datasets: FaceForensics++, CddiDF V2, WildDeepFake
- Preprocessing: The WildDeepFake dataset is already preprocessed via the procedure described on the original paper. For each video in the FF++ and CddiDF datasets, we follow the same preprocessing scheme used in the service. All face images are resized to 300x300 and normalized by the ImageNet mean and standard deviation.
- Postprocessing: we use the *Aggregation Strategy* described in the *Model Details* for all of the evaluation datasets.

Training Data

- Models 1 - 4 were trained on the DFDC dataset while model number 5 was trained on the WildDeepFake dataset.
- We expect that the models will demonstrate good performance on facial manipulations included in the DFDC and WildDeepFake datasets, i.e. Identity Swap manipulations based on DeepFake, FaceSwap, and GAN-based algorithms, and various real-world DeepFake manipulations. Thus, we expect the service to be accurate when detecting DFDC manipulations and more sensitive to real-world manipulations.

Ethical Considerations

- Risks and harms: The service presented should be used only as an auxiliary decision-making tool for anyone monitoring the validity of an image/video, thus, the results should not be seen as the absolute truth.

Caveats and Recommendations

- Manipulation methods: the performance of DeepFake detectors highly depends on the manipulations they have seen during training. For example, if a detector is trained using only one kind of DeepFake manipulation, it would perform very poorly in real-world DeepFakes since there are numerous manipulations. The generalization to novel manipulations is an open issue in the research community that almost all approaches suffer from, including our service. Our training data contain various manipulations, yet we cannot guarantee good performance on unseen manipulations due to this generalization issue.
- Multiple faces: It is recommended that the multi-faces inputs (images or videos) to the service contain only the face(s) in question and not any background faces that may distract the detection process and affect the final result.
- Video quality: It is also recommended that the input media be of the best quality possible since factors like quality and compression significantly affect the service performance.
- Video length: to ensure high-quality predictions and avoid computational overload, it is not recommended to submit very long videos and with many shots (i.e. Out-of-scope uses).

Quantitative Analysis

| Manipulation | Balanced Accuracy | AUC |
|----------------|-------------------|--------|
| FaceSwap | 78.40% | 86.74% |
| DeepFakes | 86.20% | 94.08% |
| NeuralTextures | 57.40% | 62.76% |
| FaceFusion | 59.02% | 64.02% |

Table 1: Balanced Accuracy and AUC for each manipulation in the FF++ dataset.

| Dataset | Balanced Accuracy | AUC |
|-----------------|-------------------|--------|
| FaceForensics++ | 70.31% | 77.05% |
| CddiDF | 82.73% | 92.59% |
| WildDeepFake | 84.94% | 93.73% |

Table 2: Balanced Accuracy and AUC for the service evaluated on three datasets.

| Dataset | norm-1 | norm-2 | norm-inf |
|-----------------|--------|--------|----------|
| FaceForensics++ | 70.31% | 64.04% | 70.50% |
| CddiDF | 82.73% | 78.03% | 90.00% |
| WildDeepFake | 84.94% | 63.04% | 90.00% |

Table 3: Balanced Accuracy scores on three datasets adversarially manipulated with the PGF attack (hyperparameters: $\alpha = 0.2$).

Performance Intuition

- Balanced Accuracy is the average of the accuracy in each class. Since our datasets are unbalanced, it would be misleading to just report the overall accuracy of the system. For example, in a dataset where 90% of the data are DeepFakes, a naive classifier that outputs only ones regardless of the input would get 90% accuracy, which is misleading. Owing to its intuitive nature, we consider Balanced Accuracy to be our primary metric to gauge our ensemble's performance.
- Area Under the Curve (AUC) takes into account the Miss Rate or, in other words, how often the model wrongly thinks a DeepFake is Real, as well as the True Positive Rate, meaning how often the model correctly classifies DeepFakes. Thus the AUC is an overall metric describing these two rates, and is a classification system, such as ours, higher is better. However, it does not consider the 0.5 probability threshold, which is an essential parameter in our setting, therefore, we consider it as an auxiliary metric.
- From the tables 1,2 it is evident that our system performs much better on the CddiDF and WildDeepFake datasets rather than the FF+++. It can be argued that this is due to our training data lacking Expression Swap examples which is one of the two manipulation categories of that dataset (i.e. *Relevant Datasets* and *Training Data* sections for more information).
- In table 1 the FaceSwap and DeepFakes manipulations belong in the *Identity Swap* category while the rest is the *Expression Swapping* category. Since we observe worse performance

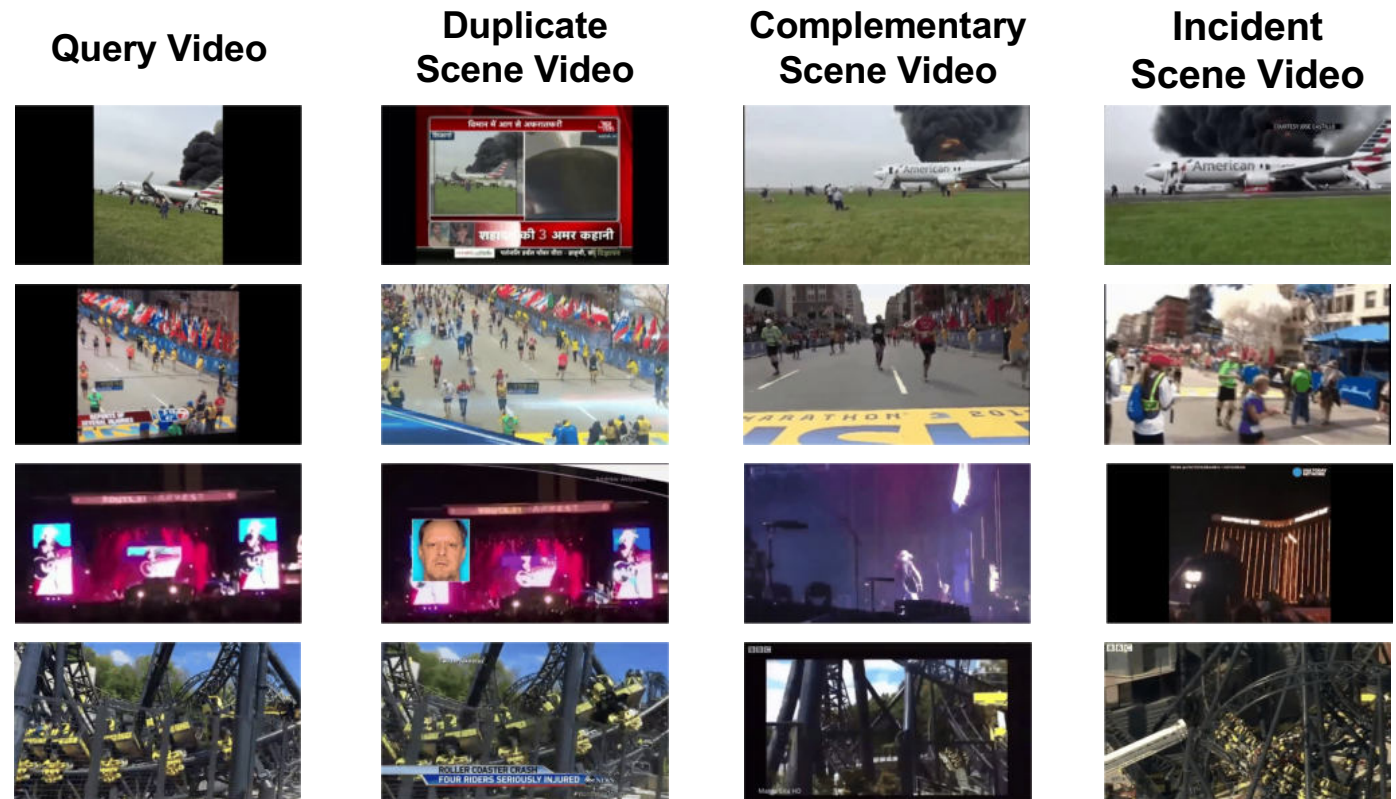
Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229).

Context Verification

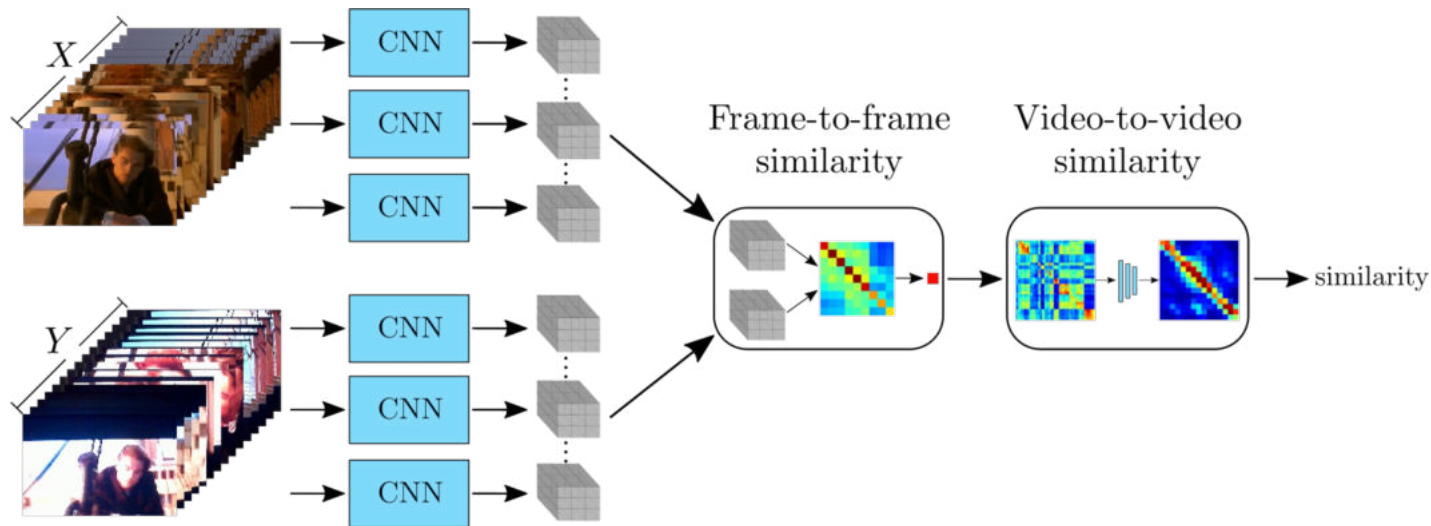
Near-Duplicate Detection

FIVR-200K dataset

- 225,960 videos
- 100 queries
- 4,687 news events
- 4 annotation labels



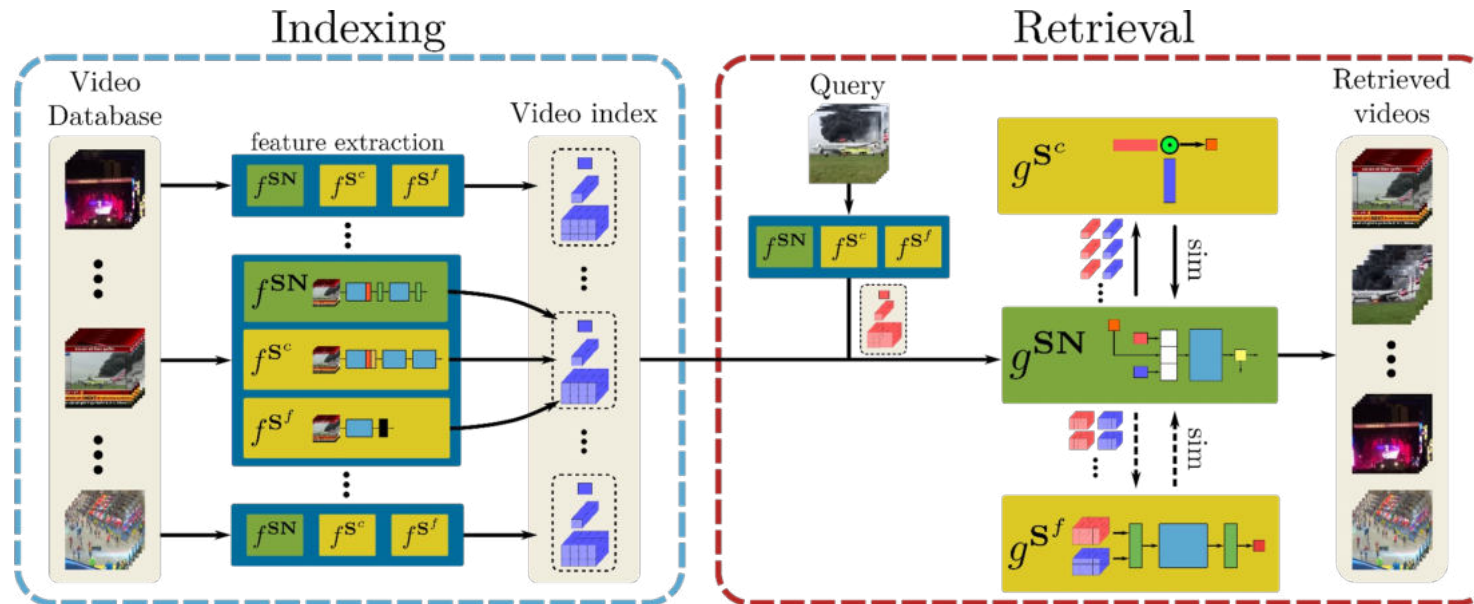
Near-Duplicate Detection



Video Similarity Learning (ViSiL)

- *Fine-grained similarity calculation*
- Learn a video similarity function that considers:
 - *Spatial structure* of video frames (intra-frame relations)
 - *Temporal structure* of videos (inter-frame relations)

Near-Duplicate Detection



DnS: Distill-and-Select for Video Indexing and Retrieval

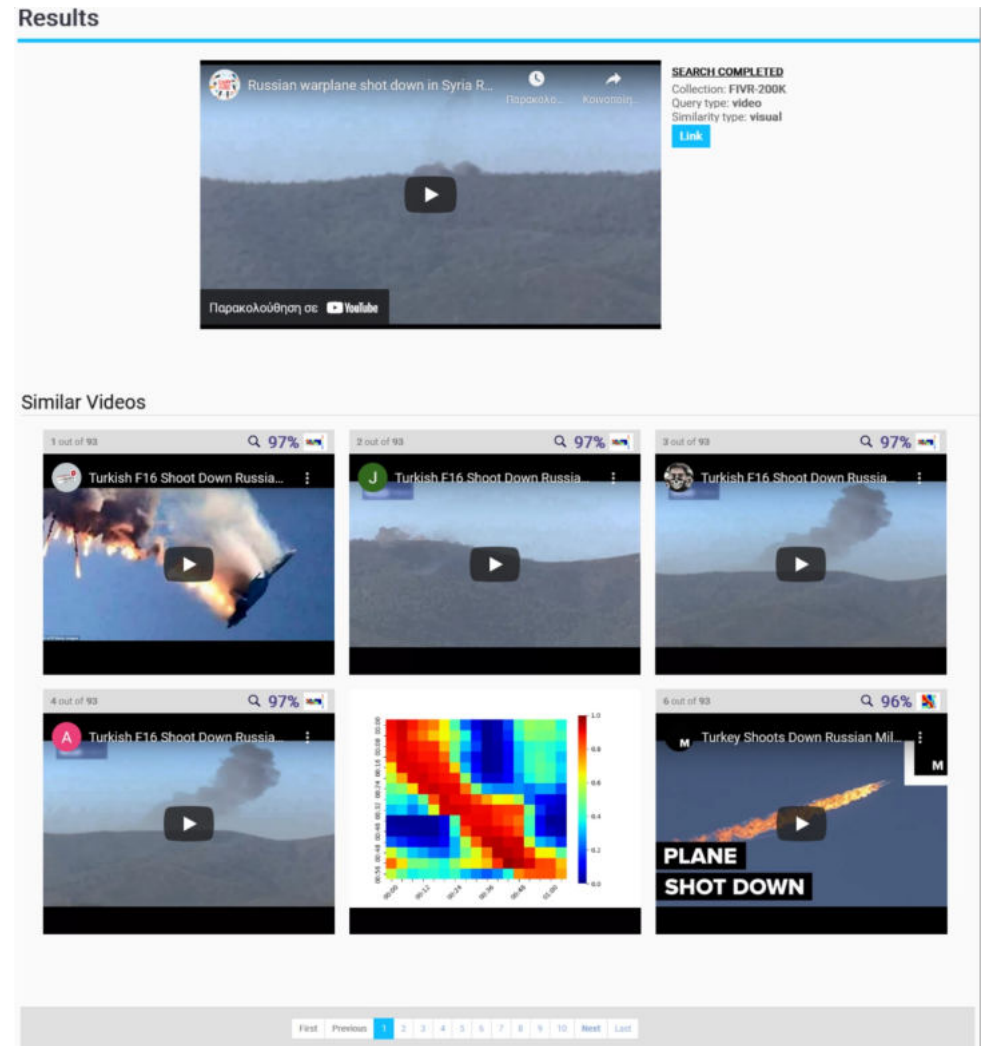
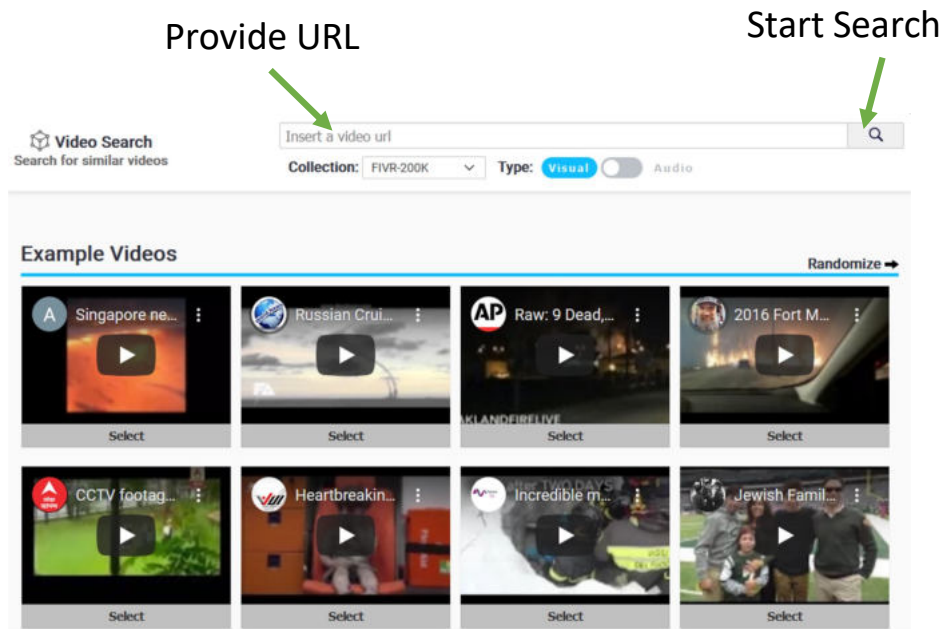
- *Knowledge Distillation* from a teacher network to multiple students
 - Different trade-off between accuracy/efficiency
- *Selection Mechanism* between two student networks
 - Select slow but accurate or fast but less accurate student

Kordopatis-Zilos et al. "DnS: Distill-and-Select for Efficient and Accurate Video Indexing and Retrieval." IJCV, 2021.

Near-Duplicate Detection

Online demo:

https://mever.itι.gr/video_search/



Context Aggregation and Analysis

In contrast to other approaches, which focus on the media items themselves for traces of forgery, this tool analyzes the media context

<https://mever.iti.gr/caa/>

Context Aggregation and Analysis

This is a demo platform aimed to facilitate the verification of UGC image and video content posted on YouTube, Twitter and Facebook. In contrast to other approaches, which attempt to analyze the media items themselves for traces of forgery, this platform analyzes the media context: The characteristics of the poster, any relevant user comments, the local weather reports at the time of the event, and other contextual pieces of information are aggregated and presented to the user for analysis. To test the service, simply copy and paste a YouTube, Facebook* or Twitter URL for videos into the box or a Facebook or Twitter URL for images, then click "Verify"

*Right click on the Facebook video and copy the video URL

Contact: [{olgapapa,papadop}@iti.gr](mailto:olgapapa,papadop@iti.gr)

☒ {Facebook, Youtube, Twitter} Video
☐ {Facebook, Twitter} Image

☐ Force Reprocess

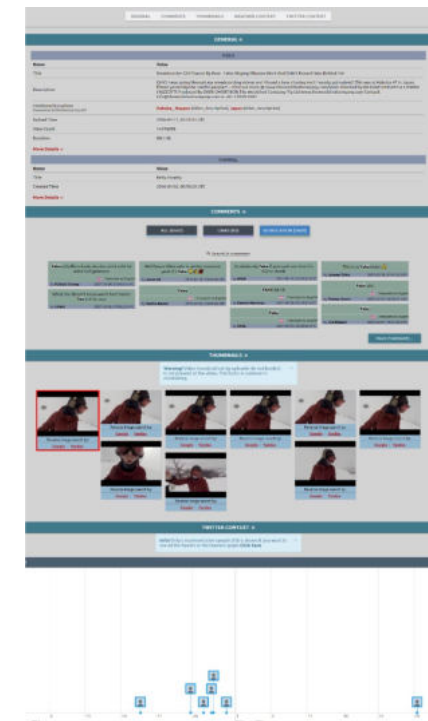
Verify

or have a look at some explanatory examples below

Facebook | Youtube | Twitter

Papadopoulou, O., Giomelakis, D., Apostolidis, L., Papadopoulos, S., & Kompatsiaris, Y. (2019). Context Aggregation and Analysis: A Tool for User-Generated Video Verification. In SIGIR 2019 Workshop on Reducing Online Misinformation Exposure (ROME 2019)

Papadopoulou, O., Zampoglou, M., Papadopoulos, S., & Kompatsiaris, I. (2019). Verification of Web Videos Through Analysis of Their Online Context. In Video Verification in the Fake News Era (pp. 191-221). Springer, Cham.



Context Aggregation and Analysis

Video and channel/user metadata

| GENERAL | |
|------------------------------|--|
| VIDEO | |
| Name | Value |
| Title | Muslims push a car down a subway stairs at the Metro in Brussels and burn the Christmas tree |
| Description | Happy New Year 2016 Translate to English |
| Upload Time | 2016-01-03, 09:43:50 UTC |
| View Count | 6819 |
| Duration | 00:00:53 |
| Mentioned Locations | Not Available |
| More Details | |
| CHANNEL | |
| Name | Value |
| Title | soim romania |
| Created Time | 2014-11-12, 05:01:59 UTC |
| More Details | |

Comments:

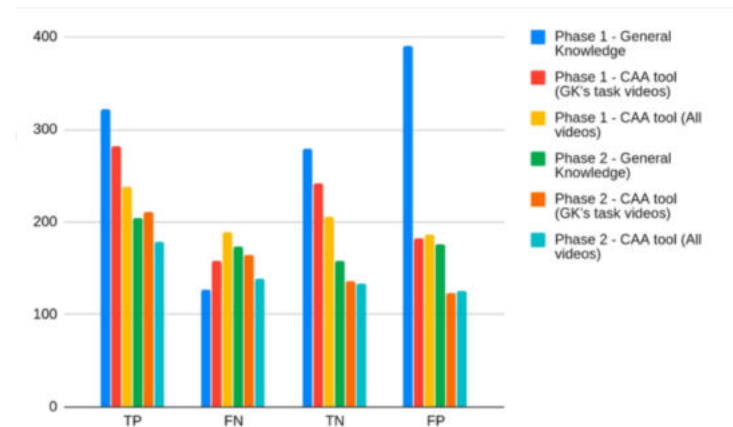
- All comments left below the video
- Links: comments containing links
- Verification: comments containing a verification-related keyword (pre-defined set of keywords)

| COMMENTS | | | |
|--|--|--|--|
| ALL (11) | | LINKS (1) | VERIFICATION (0) |
| <input type="text" value="Search in comments"/> | | | |
| Merkel's legacy by Top Kak 2020-04-01, 19:25:05 UTC Translate to English | One of the "likers" was the pope, other was Juncker... by Neza 2018-05-10, 07:06:26 UTC | In the first clip, did they kill someone in the stairs? by T-s Sosa 2016-07-12, 20:11:15 UTC | PEACEEEEEEE by The MrRick100 2016-06-05, 19:10:43 UTC Translate to English |
| "Liked/faved" for others to view. GET THEM. OUT. by offwiththefairies77 2016-05-09, 22:33:55 UTC | Religion of peace of shit. by Ty Ger 2016-05-19, 23:22:30 UTC | OR if you like reality, and value truth over sensationalist nonsense, here is the original video of the car, without the "Allahu akhbar": http://www.dailymail.co.uk/news/article-3381525/The-shocking-moment-gang-teenagers-pushed-CAR-stairs-packed-metro-platform-New-Year-s-Eve.html by Paul Smith 2016-05-12, 17:15:42 UTC | Good job Merkel. by Press X To Doubt 2016-04-06, 16:11:26 UTC Translate to English |
| Nice edit bro, the original didn't have all the 'allah arkbars' in it by Tim 2016-01-05, 22:52:01 UTC | Imao fml by Dopamine Cloud 2016-01-21, 18:33:33 UTC Translate to English | More Comments... | |

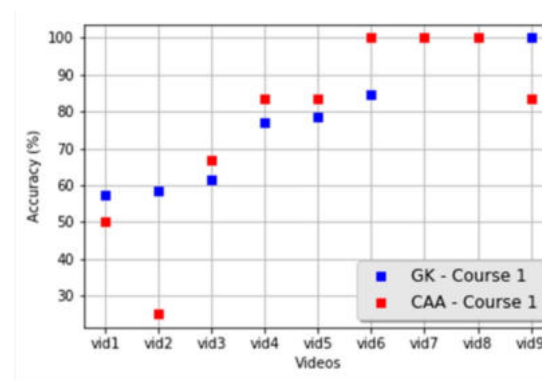
Context Aggregation and Analysis

Fake Verification Corpus: Corpus of debunked and verified user-generated videos.

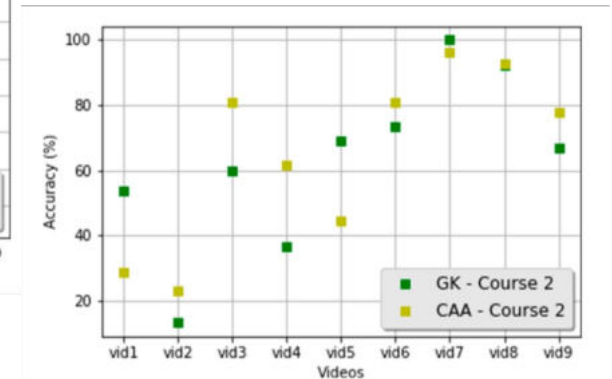
Study on the verification of news video content derived from social media platforms, using the CAA tool and a semi-automated verification practice.



Average time needed for the verification process



Average accuracy per video in GK (without CAA tool) and CAA tool tasks for the same set of fake videos.



Papadopoulou, O., Zampoglou, M., Papadopoulos, S., & Kompatsiaris, I. (2018). A corpus of debunked and verified user-generated videos. Online information review, 43(1), 72-88.

Giomelakis, D., Papadopoulou, O., Papadopoulos, S., & Veglis, A. (2021). Verification of News Video Content: Findings from a Study of Journalism Students. Journalism Practice, 1-30.

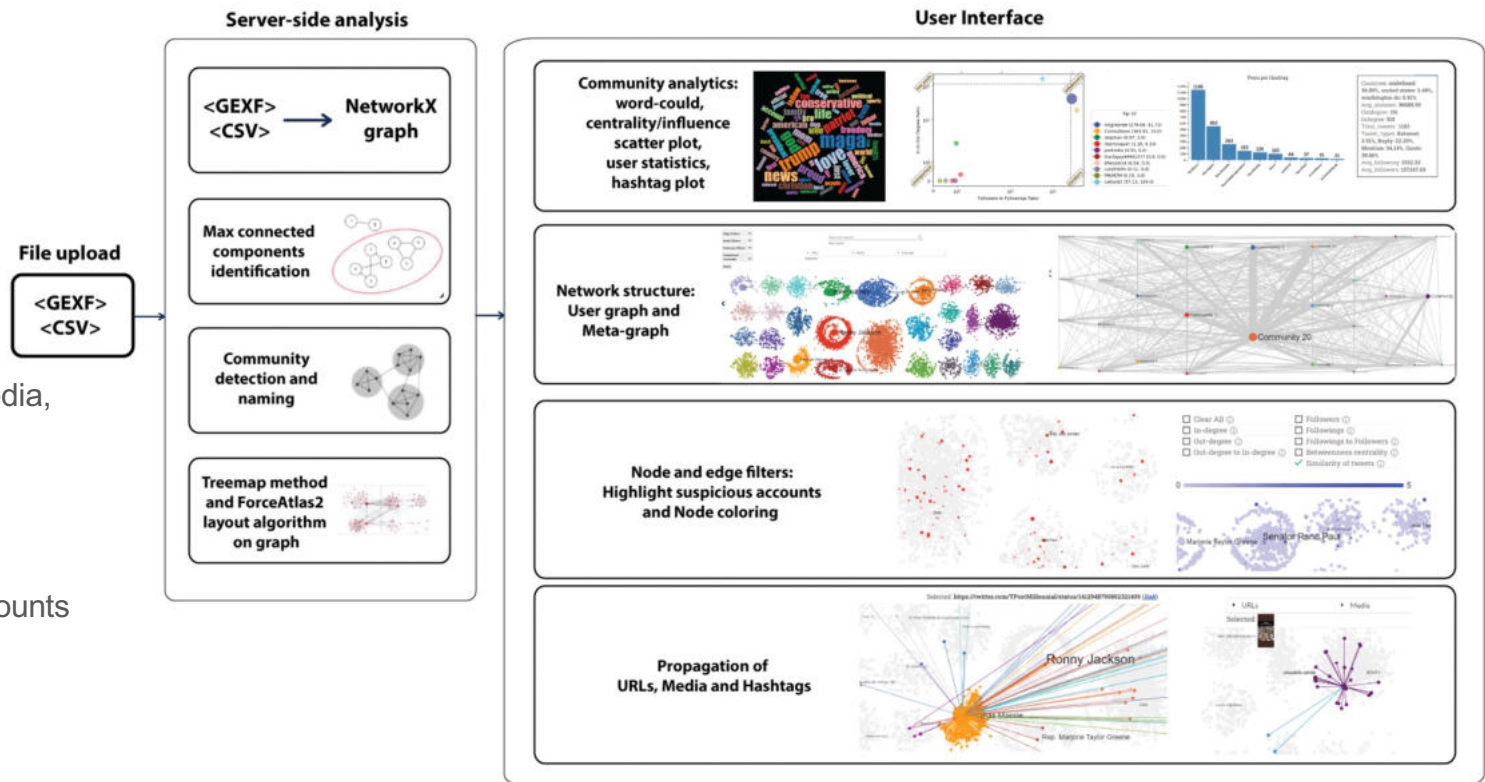
Information Propagation

Network Analysis and Visualization

MeVer NetworkX analysis and visualization tool helps users delve into **social media conversations**, helps users gain insights about how information propagates, and provides intuition about communities formed via **interactions**.

Features:

- Individual Account and Post Inspections
- Community Detection
- Community Analytics
- Metagraph
- Propagation of URLs, Media, and Hashtags
- Node and Edge Filters
- Node Coloring
- Highlight Suspicious Accounts



Network Analysis – Suspicious Accounts

- **Following rate** is the ratio of the number of followings to the number of days since an account was first created.
- **Status rate** is the ratio of the number of posts to the number of days since an account was created.
- **Average mentions per post** shows the average number of mentions in an account's tweets. A common strategy for spreading disinformation is mentioning many accounts in tweets.
- **Average mentions per word** shows the average number of mentions in a tweet's text. The tactic of posting tweets with many mentions and a single hashtag is often regarded as spam-like or suspicious. This feature is normalized to the total number of posts.
- **Average hashtags per word** calculates the average number of hashtags in a tweet's text.
- **Average URLs per word** calculates the average number of URLs in a tweet's text.

Support
Twitter,
Facebook
and Telegram

Network Analysis and Visualization - Analysis

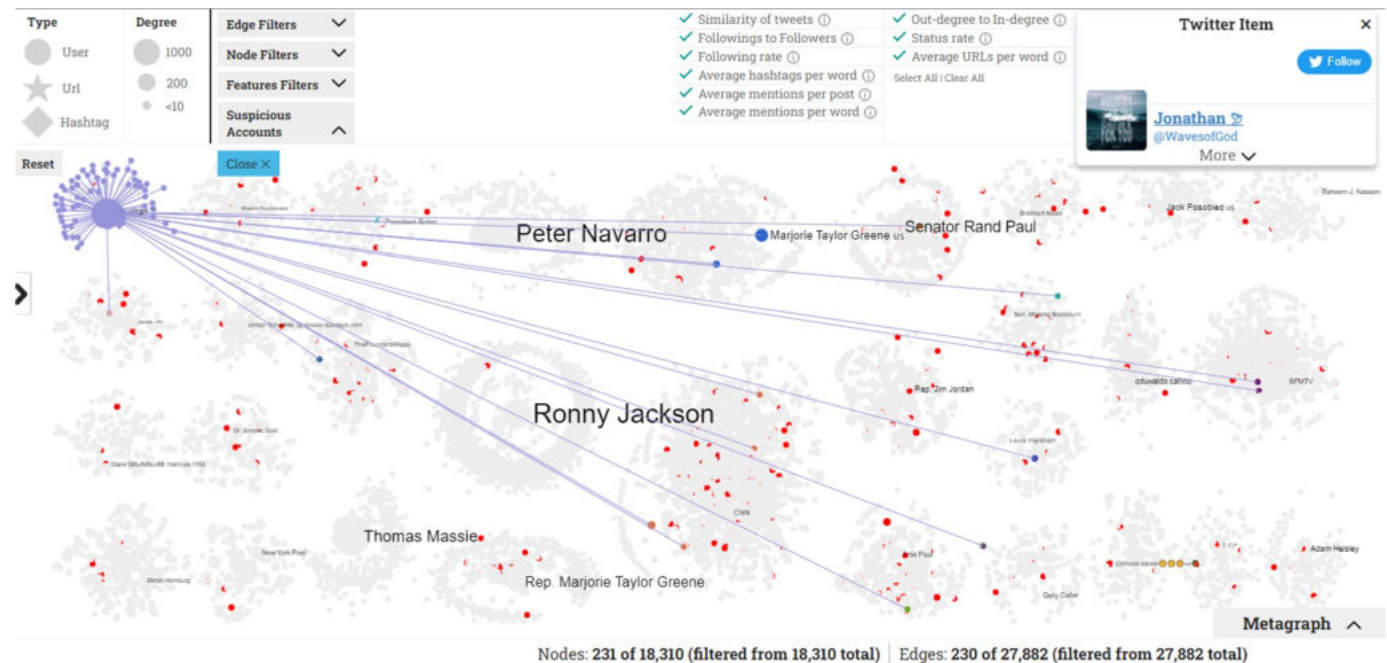
Four use cases to simulate a scenario in which, through the tool, an end user tries to identify and inspect suspicious accounts within a given dataset graph.

Fauci use case

425 accounts out of 18,310 highlighted as suspicious.

@WavesofGod (Community 14)

- Mentioned 228 other accounts in its tweets
- Mentioned popular accounts: President's Biden (POTUS) and Marjorie Taylor Greene ("mtgreenee").
- It is a strong supporter of Christianity and against vaccination.



This account has been suspended from Twitter

<https://mever.iti.gr/networkx/>

Papadopoulou, O., Makedas, T., Apostolidis, L., Poldi, F., Papadopoulos, S., & Kompatsiaris, I. (2022). MeVer NetworkX: Network Analysis and Visualization for Tracing Disinformation. Future Internet, 14(5), 147.

Key Challenges in AI against Disinformation

- Arms race nature of problem
 - Constantly improved AI models for synthetic media
 - New disinformation tactics
 - Gap between research and end users
 - Journalists/fact-checkers need intuitive tools
 - Tools need to be robust and trustworthy
 - Risks of discrimination
 - AI models might be biased against certain demographics
 - Use of AI models might be done in a biased manner
 - Sustainability
 - AI costs a lot in terms of research, maintenance and deployment
 - Fighting disinformation is not a profitable business
 - Required contribution from various disciplines
 - Content Analytics - NLP, Machine Learning, Network Analysis, Big Data Architectures,
 - Psychology – Social Sciences (patterns of presentation, sharing), Visualization
- **Continual learning**
 - **Out-of-domain generalization**
 - **Explainability**
 - **Human agency**
 - **Robustness**
 - **AI fairness**
 - **Green AI**
 - **Model compression**
 - **Knowledge distillation**

Emerging Challenges: Metaverse

*“...we might experience a **hyper-realistic VR environment that might make it much more difficult to assess media authenticity.** This is due to the fact that media assets like images and videos are blended in a more natural way in the VR environment, which could arguably increase the cognitive load for the human brain (more cognitive-neuroscience studies would be needed to study this)....”*

vera.ai EC project



<https://www.nytimes.com/2022/10/09/technology/meta-zuckerberg-metaverse.html>

Human Behaviour related challenges

Cyberpsychology, Behavior, and Social Networking, Vol. 22, No. 6 | Rapid Communications

 Free Access

What Debunking of Misinformation Does and Doesn't

Jeong-woo Jang  Eun-Ju Lee, and Soo Yun Shin

Published Online: 7 Jun 2019 | <https://doi.org/10.1089/cyber.2018.0608>

... participants' agreement with the news position was not attenuated by the explicit post hoc correction. Considering that information is judged as truth when it meets intuitive evaluation criteria (e.g., familiarity, compatibility with existing knowledge)... **debunking its falsehood may not be sufficient to undo it.**

How Health-Related Misinformation Spreads Across the Internet: Evidence for the “Typhoon Eye” Effect

Lei Zheng , Jincheng Cai, Fang Wang, Chenhan Ruan, Mingxing Xu, and Miao Miao 

Published Online: 11 Oct 2022 | <https://doi.org/10.1089/cyber.2022.0047>

 Sections PDF/EPUB

 Permissions & Citations  Share

... Our results highlight the importance of psychological approaches to understanding the propagation patterns of health-related misinformation. The present findings provide a new perspective for development of prevention and control strategies to reduce the spread of health-related misinformation during pandemics ...

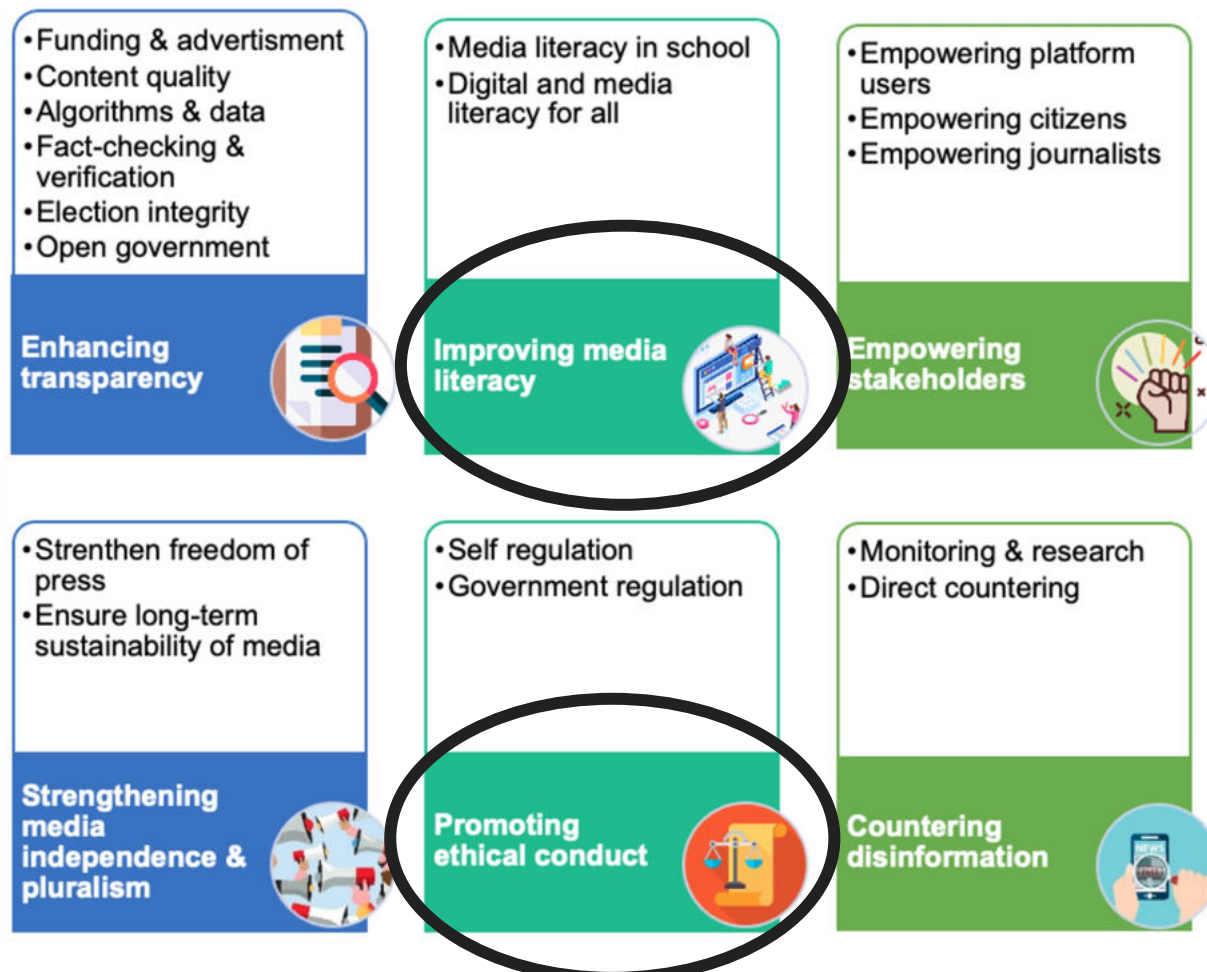
Industry ethicist: Social media companies amplifying Americans' anger for profit



NOVEMBER 6, 2022 / 7:32 PM / CBS NEWS

... The more moral outrageous language you use, the more inflammatory language, contemptuous language, the more indignation you use, the more it will get shared. So we are being rewarded for being division entrepreneurs. The better you are at innovating a new way to be divisive, we will pay you in more likes, followers and retweets....

Counter-disinformation policies classification



PRESS RELEASE | Publication 11 October 2022

Commission steps up action to tackle disinformation and promote digital literacy among young people

The Commission has published Guidelines for teachers and educators in primary and secondary schools, on how to address disinformation and promote digital literacy in their classrooms.

The guidelines provide practical support for teachers and educators and include definitions of technical concepts, class-exercises and how to encourage healthy online habits. This toolkit covers three main topics: building digital literacy, tackling disinformation, and assessing and evaluating digital literacy.

[Full press release](#)



Counter-disinformation policies



The EU is working in close cooperation with online platforms to encourage them to promote authoritative sources, demote content that is fact-checked as false or misleading, and take down illegal content or content that could cause physical harm.

Positive Applications of DeepFakes

Protecting the Identity of Interviewees

- Welcome to Chechnya is a 2020 documentary film by David France
- The film centres on the anti-gay purges in Chechnya of the late 2010s, filming LGBT Chechen refugees using hidden cameras as they made their way out of Russia
- This was the first film to use DeepFake technologies to protect the identities of speakers

https://en.wikipedia.org/wiki/Welcome_to_Chechnya



Animating Faces from the Past

- DeepNostalgia by my Heritage makes it possible to create animations from still photos, e.g. of historical figures or beloved ones



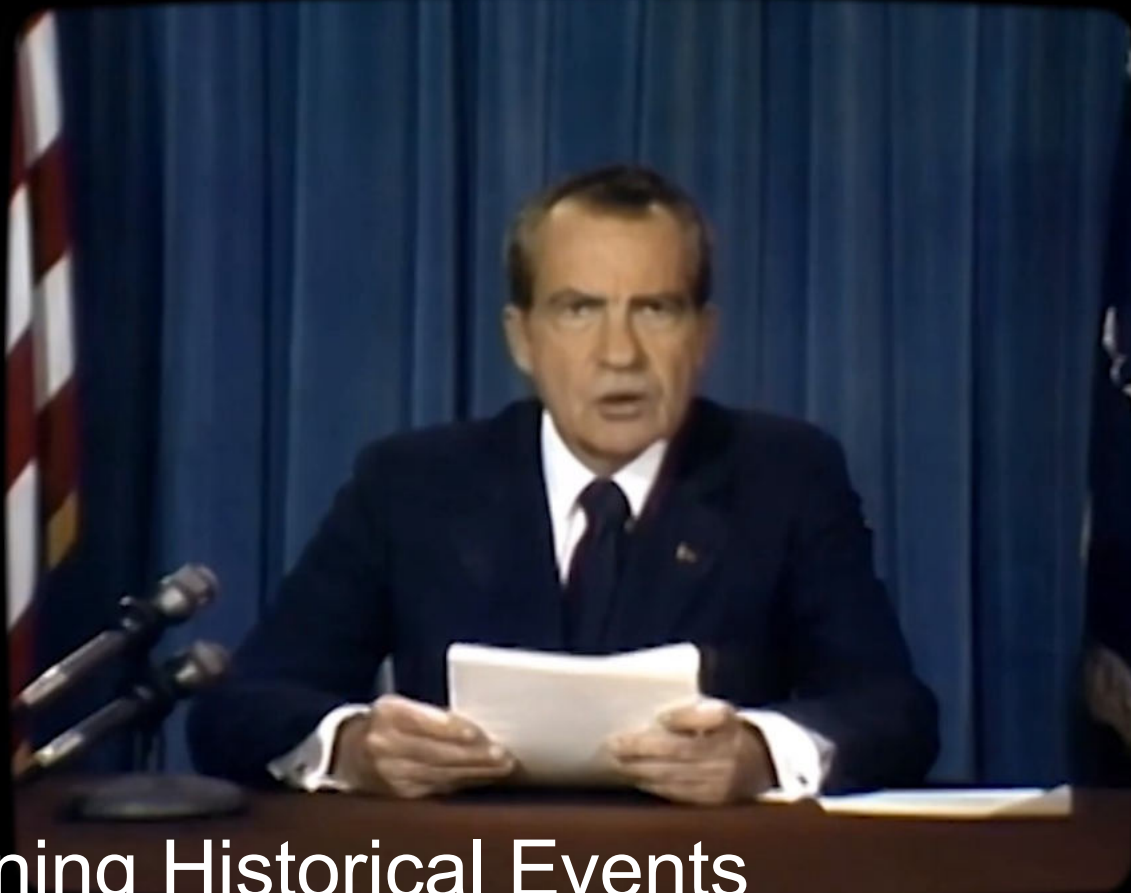
Dalí Lives



<https://www.theverge.com/2019/5/10/18540953/salvador-dali-lives-deepfake-museum>

IN EVENT OF
MOON DISASTER

THE FILM RESOURCES ABOUT

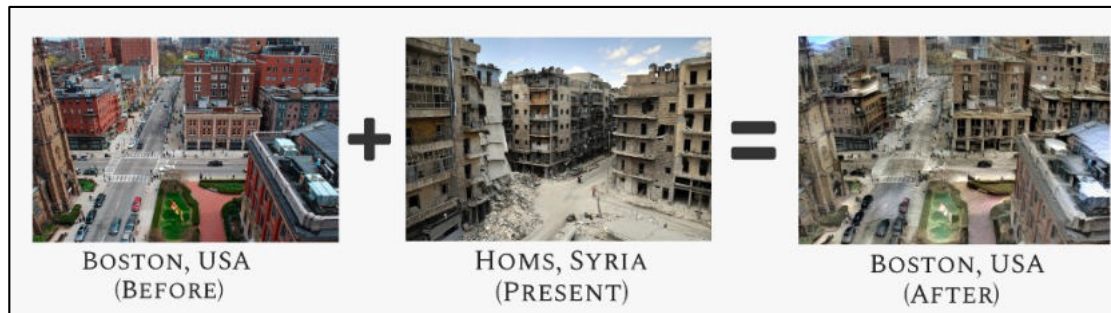


Reimagining Historical Events

<https://moondisaster.org/film>

Sensitizing about Global Issues

- DeepEmpathy (<https://deepempathy.mit.edu/>):
increase empathy by making our homes appear similar to the homes of victims of disasters/wars



- Using DF to Visualize Impacts of Climate Change

Luccioni, A., Schmidt, V., Vardanyan, V., & Bengio, Y. (2021). Using Artificial Intelligence to Visualize the Impacts of Climate Change. IEEE Computer Graphics and Applications, 41(1), 8-14.



Deliver the next generation of AI Research and Training
at the service of Media, Society and Democracy,
ensuring the embedment of European values of
ethical and **trustworthy** AI in the future deployments.

AI4media

ARTIFICIAL INTELLIGENCE FOR
THE MEDIA AND SOCIETY

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 951911



www.ai4media.eu



**A European Media AI
research agenda and
observatory**

#1



**Research and
Innovation activities**

#2



**AI real-world
applications through
use-cases**

#3



**Equity-free funding for
Innovative research and
applications in the AI sector**

#4



**AI Doctoral Academy -
AIDA**

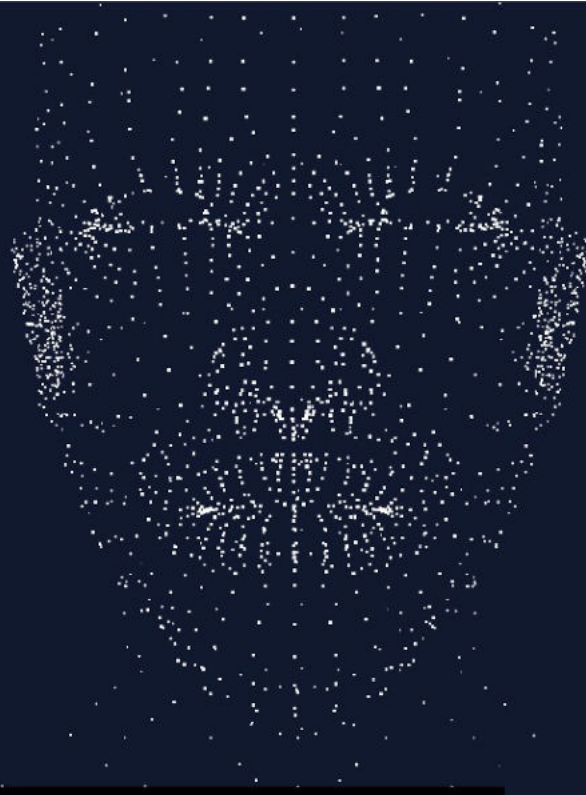
#5



**A Virtual Centre of
Excellence and
ecosystem**

#6

What AI4Media has to offer



Would you like
to become an
Associate Member?

Visit
www.ai4media.eu

AI4Media Associate Members



Open Call #2

10 projects will be
funded with
€50.000 each

Apply until 30th November
2022



www.ai4media.eu



Related projects



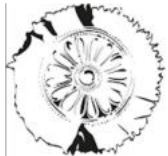
Thank you for your attention!



<https://mever.gr>

<https://twitter.com/meverteam>

Dr. Symeon Papadopoulos, Senior Researcher, Group Leader



CERTH
CENTRE FOR
RESEARCH & TECHNOLOGY
HELLAS

