NORDIS – NORdic observatory for digital media and information DISorders

State of the art in fact-checking technology

Date: 14-03-2022

Final version

Action No:	2020-EU-IA-0189
Project Acronym:	NORDIS
Project title:	NORdic observatory for digital media and information DISorders
Start date of the project:	01/09/2021
Duration of the project:	24
Project website address:	https://datalab.au.dk/nordis
Authors of the report:	Carl-Gustav Lindén, University of Bergen Duc-Tien Dang-Nguyen, University of Bergen Camilla Salas-Gulliksen, University of Bergen Sohail Ahmed Khan, University of Bergen Maria Amelie, Factiverse Laurence Dierickx, University of Bergen
Activity number	Activity 4
Task	4.2

Reviewers: Morten Langfeldt Dahlback (Faktisk.no, Norway), Matteo Magnani (Uppsala University, Sweden), Jessica Gabriele Walter (Aarhus University, Denmark), Marina Charquero Ballester (Aarhus University, Denmark), Mathias Holm Tveen (Aarhus University, Denmark)

Funding:
The research was funded by EU CEF grant number 2394203.

Table of content

Executive summary

The purpose of this report is to summarise the state of art in fact-checking technology in Europe and the United States. We aim to build a knowledge foundation to inform and guide coming work in the NORDIS project. The report contains an exploration of how fact-checking practices are augmented with different technical tools and an overview of available or emerging technology. The report ends with a brief analysis of gaps in available technology in relation to working processes and opportunities for development.

1. Introduction

The purpose of this report is to summarise the state of art in fact-checking technology in Europe and the United States and to build a knowledge foundation to inform and guide future work in the NORDIS project. In Activity 4 of the Nordis project, “Innovation & technology”, researchers at the University of Bergen are collaborating with fact-checkers, technology companies, and other hubs in the European Digital Media Observatory (EDMO) to analyse user needs and develop new tools for verifying content. To be able to do that we start by mapping and evaluating the state of art in fact-checking technology. We provide an overview of the existing technical solutions (AI systems/tools/assistants) and methods for verification of facts and content. We review work processes and processes that could be of interest to the project, mainly from the perspective of relevant technologies and their applications. The overview is based on systematic document analysis of fact-checking: gathering, following and analysing industry news reporting and analysis in specialised and quality media, research centres and projects, media reports and technology reviews. This includes analysing the impact of platforms on fact-checking practices and tools. To gain practical insights to guide our mapping, the team had a group discussion with Andy Dudfield, who is Head of Product at UK based fact-checking organisation Full Fact. In the end of the report, we draw conclusions with regard to gaps in available technology and where there is potential for augmenting fact-checking practices with new tools and applications.

2. Disinformation, misinformation and fake news

First, we give a short overview of the problem that fact-checkers are struggling with, a kind of tsunami of information that in combination with rumours, conspiracy theories, fabrication and falsehood which related to the COVID-19 epidemic was deemed an infodemic (Zarocostas, 2020). The terms “misinformation,” “disinformation,” and “propaganda” are sometimes used interchangeably, with shifting and overlapping definitions (Guess & Lyons, 2020). All three concern false or misleading messages spread under the guise of informative content, whether in the form of elite communication, online messages, advertising, or published articles (Guess & Lyons, 2020). Disinformation is understood as ”verifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm” (Action Plan against Disinformation, EU Commission, 2018)[1]. This is conceptually distinct from rumours or conspiracy theories, whose definitions do not hinge on the truth value of the claims being made (Guess & Lyons, 2020).

We can also note that disinformation is a subset of misinformation: Disinformation is meant to deceive, while misinformation may be inadvertent or unintentional. There are different reasons why disinformation has become such a problem in digital societies and in general for democracies. For instance, political motives are often behind the deliberate spreading of

[1]https://ec.europa.eu/info/publications/action-plan-disinformation-commission-contribution-european-council-13-14-december-2018_en

factually dubious content for creating mistrust, confusion and disorder. One important actor has been Russia's state-funded Internet Research Agency (IRA) which participated in Russian meddling in the 2016 US elections. The extent of the harm is disputed with some researchers claiming that only a small group of society was affected (Bail et al., 2020).

Other motives might be economic. In the Macedonian town Veles around 100 young people created websites with names such as USADailyPolitics.com, WorldPoliticus.com, and DonaldTrumpNews.com. They filled these with made up news and could earn as much as $8,000 a month, twenty times the average wage in Veles. "We ended up finding a small cluster of news websites all registered in the same town in Macedonia called Veles," Craig Silverman, media editor, recalls (Wendling, 2018). He and a colleague started to investigate the phenomena, and shortly before the US election in 2016 they identified at least 140 fake news websites which were pulling in huge numbers on Facebook.

Disinformation can also be used for propaganda and brand management purposes as was the case during the Olympic games in Peking in 2022 where Chinese authorities created a massive amount of bots and fake Twitter accounts to whitewash the country's flaws and human rights abuses. The controversy around the disappearance of Peng Shuai, the professional tennis player, ignited a similar propaganda campaign on social media (Myers et al, 2022).

A specific type of disinformation is the false content known as "fake news," or deliberately misleading articles designed to mimic the look of actual articles from established news organisations (Waisboard, 2018). The concept is contested and as an example, the British government has actually decided that fake news should not be mentioned at all in its public communication. Few public persons have probably done more to popularise the term "fake news" than the former US president Donald Trump who regularly on Twitter and on television began to describe any unfavourable press coverage of him as "fake news" (Ott & Dickinson, 2019). In fact, it is likely that the widespread perception of polarisation on social media is due to a minority of highly active and visible partisan individuals (Barberá & Rivero, 2015). These we can call influencers and their impact is defined by the number of subscribers, followers and friends.

Judging from these events, social media algorithms, along with social platforms' intensive emphasis on generating revenue from user data, has eroded the mutual trust of networked publics and opened the way for dis-/misinformation campaigns. Empowering companies, such as Facebook, Twitter and Youtube, say they do not want to take the role as "arbiters of truth" by regulating speech on their platforms but they are increasingly pressed to regulate what can be posted (Kreiss & McGregor, 2019).

There is a great desire to design interventions to reduce disinformation such as the spread of fake news (Persily & Tucker, 2020) even if there is little evidence that the phenomena has any major impact on society (Tsfati, 2020). Yet, we know little of who and why people share fake

news. Is it because they think it is true or, instead, because they agree with it ideologically and do not care if it is fake? There are even signs that people share fake news because they find it interesting, even though they know it is not true (Guess & Lyons, 2020). How can we design appropriate policies and methods to reduce its spread? To what extent will fact-checking suffice in fighting fake news? Moreover, if we do not know the effects of exposure to disinformation such as fake news, then we cannot know the “benefits" of reducing exposure to it (Parsely & Tucker, 2020).

In this report we are not going deep into the epistemology of fact-checking but there are a number of limitations to the extent it is possible for a human to state what is true and what is false. However, while we might have some presumption that facts represent a taken-for-granted sense of external reality, it is neglecting that facts are built upon the institutional structure that underlies public claims (Graves, 2016). Facts we are aware of are often incomplete, conditional, or otherwise uncertain and often mixed up with arguments. According to Waisboard (2018, 1872) what the truth about some state of affairs is, is “forever unstable, disputed, challenged” and subject to “social conditions in which arguments are developed, shared, and discussed”. Sources of uncertainty are plenty such as ignorance or inattention. Thus, separating legitimate views from disinformation is “difficult and often controversial work” (Graves, 2018, 1). This is a limitation that carries over to attempts to automate fact-checking.

3. Fact-checking as a practice

Checking facts is actually the “essence” of journalism, which basically is “a discipline of verification” (Kovach & Rosenstiel, 2007, p. 79; Graves & Amazeen, 2019). Fact-checking has a long tradition in traditional media, which have had experts going through articles before publication (ante hoc) to find factual mistakes made by journalists (Harrison Smith, 2004) but also for checking and challenging statements by established figures in public life (Dobbs, 2012). The emergence of digital media and specifically social media meant that public figures now had direct channels to a broader part of the population without gatekeepers. This inspired the birth of online organisations performing and publishing fact-checks (post hoc) such as the foundation of Snopes in 1994 (Full Fact, 2020).

If we look at their mission there is no universal understanding about who fact-checkers are and what they do: for some it is about politics or campaigns, for others it is journalism. For instance, Full Fact focuses on people in power in the UK such as politicians, journalists or other people who are producing information that is consumed by many people. Other fact-checkers are interested in viral misinformation or will be particularly focused on health misinformation. Some are mostly active during elections. Accordingly, there are also different approaches to selecting what claims should be fact-checked. Should they include media and journalists and risk putting their peers under pressure?

Essentially, from a computational perspective there are four core tasks in the work process that would be key to the development of automated fact-checking pipelines (Hanselowski et al., 2019):

- Document retrieval is applied to identify documents that contain important information for the validation of the claim.
- Evidence extraction aims at retrieving text snippets or sentences from the identified documents that are related to the claim.
- This evidence can be further processed via stance detection to infer whether it supports or refutes the claim.
- Claim validation assesses the validity of the claim given the evidence.

However, from a broader societal perspective there are other important tasks beyond identifying claims, choosing which claims to fact-check, doing the actual fact-check and writing a claim report. In addition, fact-checkers need to make sure that fact-checking has an impact, that the results published reach the general public. And there is another outreach task, to follow up and to revisit claims already checked to find out if they have been corrected. If one politician misquotes some statistics, or leads people into a wrong direction, and if that is repeated by multiple other people, fact-checkers want to be able to identify that and ask them to correct the record (A Dudfield, personal communication, January 26, 2022).
Arnold (2020) has defined three major challenges for fact-checkers.

1) Within the monitoring and selection process, fact checkers are grappling with large pools of potential claims to check, questions over how to define virality, the opaque nature of some platforms, and worries about whether cyber armies are gaming reporting.
2) When it comes to researching the accuracy of a claim, fact checkers experience challenges including accessibility of information and the transparency of authorities, highly repetitive claims and research tasks, and changes to or discontinuation of online investigation tools.
3) When publishing and distributing their work, challenges include grappling with the demand for fact checks from internet companies and the impact of these partnerships, internet shutdowns, the large effort required to set up new social media channels, and the difficulty of sustaining media partnerships.

Adding to the challenges is the fact that there are very few people focusing on verification of claims. The International Fact-Checking Network (IFCN), which was established in 2015 and serves as a professional trade organisation for fact-checkers, has around 100 certified members. According to another estimate by Duke Reporters´ Lab in 2021,[2] there were 353 global fact-checking sites operating in more than 100 countries. In any case, it is a small closed community with an invitation-only Slack channel managed by the IFCN as an important communication channel, but it attracts increasing interest from many parties in society. The Russian military and information attack on Ukraine in February 2022 illustrated how important it is for the whole of the EU to see through all these disinformation attempts. Global Fact, a yearly conference for fact-checkers and interested parties such as researchers and technologists, attracted around 1,000 participants in October 2021.

[2] https://reporterslab.org/fact-checking/

There are also important questions regarding impact. Relatively few people ever visit professional fact-checking sites; the public appreciates fact-checking in theory but shows little interest in practice (Nyhan & Reifler 2015). However, one can argue, as Dias and Sippit (2020) do, that research has focussed on the short-term impact of individual fact-checks, thus overlooking the cultural and systemic changes that fact-checkers might achieve. As an example, the authors mention seeking improvements to press complaints systems so that unsubstantiated reports can be corrected. Even if individuals do take the initiative to visit fact-checking sites, these sites frequently choose to cover markedly different topics. In fact, even when their coverage does overlap, fact-checking organisations often reach diametrically opposed conclusions about the factual basis for a given piece of information (Marietta et al., 2015).

Nissen et al. (2022), in their study of digital infrastructures in fact-checking, note that there is no real unifying global infrastructure of fact-checked stories. Basically, there are two main systems for fact-checked claims. One is ClaimReview, a tagging system to identify fact-check articles for search engines and apps. In 2016, a project at the Duke Reporters´ Lab started to help fact-checking organisations increase the speed and impact of their work and they designed ClaimReview as a new data standard, which was later adopted by Google to feature in search results (Graves & Anderson, 2019). ClaimReview entries contain summaries of fact-checkers' findings in a machine-readable format.

The other infrastructure is Facebook's Third-Party Fact-Checking Programme (3PFC)[3], launched in the aftermath of Donald Trump's election in 2016. The company had failed to stem the spread of "fake news" on its platform during the political campaign and needed to take action. 3PFC is now one of the main funders of fact-checking activities in the world. A necessary precondition for participating in 3PFC is that the organisation is a member of IFCN. Fact-checks are embedded into Facebook where visitors that try to share problematic content will face a sign on the screen saying that fact-checkers have reviewed this and that they might want to consider this before sharing it. Instagram has a similar function and so does Youtube in its search results.

For some organisations, the revenues from the platform can become rather large. For instance, between July 2019 and November 2020, UK based Full Fact received £312,507 for its work. The amount of money that Full Fact is entitled to depends on the amount of fact checking done under the programme (Full Fact, 2020).

Fact-checkers who otherwise believe in transparency and integrity find that with platform companies, the use of non-disclosure agreements is common. There are also other sources of controversy. While Facebook and Google are funding fact-checking, they are at the same time paying millions of dollars to the operators of clickbait pages, bankrolling the deterioration

[3] https://www.facebook.com/journalismproject/programs/third-party-fact-checking

of information ecosystems around the world (Hao, 2021). Karen Hao illustrates how Facebook changed the whole information ecosystem in Myanmar, away from verified sources. She writes: “In 2015, six of the 10 websites in Myanmar getting the most engagement on Facebook were from legitimate media, according to data from CrowdTangle, a Facebook-run tool. A year later, Facebook [...] offered global access to Instant Articles, a program publishers could use to monetize their content. One year after that rollout, legitimate publishers accounted for only two of the top 10 publishers on Facebook in Myanmar. By 2018, they accounted for zero. All the engagement had instead gone to fake news and clickbait websites. In a country where Facebook is synonymous with the internet, the low-grade content overwhelmed other information sources.”

Furthermore, a study by Nissen et al. (2022) indicates that fact-checking infrastructures have inherent biases that affect what is being fact-checked. For instance, the claims that come in the feed to fact-checkers in Facebook’s system is the viral mean end of mis- and disinformation but the company does not fund fact-checking of political statements or advertisements. This means that Facebook’s preferences guide fact-checkers to tasks that secure funding while neglecting other, potentially more important claims.

Google does not fund fact-checkers in the same way as Facebook does but their priority in providing high quality neutral information for search results creates other problems, for instance an overreliance on information from Wikipedia. This has its own inherent challenges such as task conflict regarding controversial and “edit war” articles, issues of social status in knowledge production and a lack of diversity in both participation and content (Ruprechter et al., 2020; Lerner & Lomi, 2020; Menking & Rosenberg, 2021).

Techno-optimists tend to believe that they will be able to solve wicked societal problems with the help of advanced systems in artificial intelligence. However, today’s technology is severely limited in many senses. Full Fact, a UK based fact-checking organisation uses Bidirectional Encoder Representations from Transformers (BERT), one of the large open-source language models made available by Google and trained with annotations from fact-checkers to identify claims. Full Fact also uses spaCy, a free open-source library for Natural Language Processing in Python and used for semantic similarity. The organisation also uses “some really old-fashioned things”, says Dudfield, such as BM25 which is a document processing technique from the 1980’s.

“I am not precious about what tools we use. It does not have to be the most cutting edge technology. I just want to use things that are going to give the most value at any particular time.”

Full Fact also scrapes information from websites using the Python library Beautiful Soup. The data team also applies speech to text technology on radio and television broadcasts. Full Fact then uses quotation detection to find out who said what and what is the characteristic of the claim.

“Once we have got everything in text we can break stuff down into sentences [so] we can identify claims [and] start classifying those [and] because we have scraped all of that information we can bring more structure.” (A. Dudfield)

The organisation mines around 100,000 mainly text-based claims every day and shares these with its network of fact-checkers in the UK, Kenya, Nigeria and South Africa that can log in and retrieve them from the system. The aim is to find around 30-40 claims made by prominent politicians during the past 24 hours that can be analysed. For fact-checks of claims that are based on numerical data, Full Fact often uses the UK office of national statistics through their API. However, in most cases there is no good solid statistical foundation to use to check claims.

However, a quick reality check indicates that technology is not enough for solving social and political challenges such as information disorder or the so-called infodemic. Besides fundamental issues, such as lack of deep understanding of the problems we want to solve, there are also severe technical limitations due to, for instance, data quality issues. Full Fact, for instance, have used supporters to manually annotate hundreds of thousands of claims for training purposes but the data is of low quality and very noisy: when it takes one or two fact-checkers to forensically review a statement to decide if it is right or wrong, the corresponding number of people from the public is around 20. Machine learning methods require a large corpus with reliable annotations for the different tasks in the fact-checking process. Existing fact-checking corpora are either too small in size, do not provide detailed annotations, or are limited to a single domain (Hanselowski, 2019).

In the next section we present the results from our mapping project.

4. Practical applications of technology in fact-checking

In this part we provide a brief overview of AI systems, tools and assistants and methods for verification of facts and content. In total we found 134 fact-checking tools and services tools and a brief analysis shows a broad spectrum of technologies and methods of finding, verifying and classifying facts. Their approaches vary, but they all seem to share a common objective of providing true and trustworthy information to a user – be that a journalist, decision maker, businessperson, politician or private person.

A smaller number of the resources such as Fakey or Interland are purely for educational purposes, aiming to provide the public with the tools to check their own facts. Other tools focus on the detection of plagiarism, or detection of malicious bot activity such as Botcheck.me or Botsentinel.

Most of the tools found are mining and analysing content as texts, for instance Fakerfact. By analysing entire articles, opinions, or social media posts and feeds, many of the providers claim to use artificial intelligence, natural language processing and machine learning to identify

and classify claims. Some, such as Exorde, store this insight in a blockchain to allow users to constantly verify and debunk claims.

Few of the tools are aimed towards verifying images, and videos, and most of the tools that do have this feature, only have the ability to detect tampered footage – not necessarily advanced synthetic footage. This leaves those services vulnerable in the detection of i.e. a deep fake video. One of the exceptions is Defudger that detects manipulations in images and videos including detecting deepfake video or the toolkit Forensically.

A minority of the tools seem to be aimed to be used as an integral part of a journalistic work process, or embedded into editorial publishing platforms. A large part of the tools are web based applications or plugins, which do not necessarily fit into a journalist's workflow. Many seem to aim towards only detecting mis- and disinformation, and not stop it at the source.

During the mapping it became clear that there is no easy way to verify the tools source code or technology without some degree of technical proficiency. The technical description of the services varies, and thus they will also do so in the mapping.

None of the services investigated provides truly extensive, fully automated reports. Several provide labelling such as “true”, “false”, or “suspicious”, and some of the more in-depth services provide “nutritional labels” summarising the fact check results. More extensive reports are mainly found in the manually reviewed fact checks and are written by human beings. Besides ClaimReview there are also some search engine markup tools, which use fact checks and verifications to mark content in search engines or social media feeds if they have been fact checked.

The overview is based on systematic document analysis of fact-checking: gathering, following and analysing industry news reporting and analysis in specialised and quality media, research centres and projects, media reports and technology reviews. The link to the table of tools is here:
https://docs.google.com/spreadsheets/d/1oIFvwR8b_7v9osdJwueNo4KSBb2p7PuqOQD4XKhO7k8/edit?usp=sharing

4.1 Visual Content Verification

In the analysis of existing tools and solutions we recognised a lack of solutions for verification of visual content. However, fake news and mis-/ disinformation shared online on social media platforms accompanying visual content (images, videos) is more popular than the text-based mis-/disinformation (Guy, H. 2017). Thus it is important to check the truthfulness of the visual content during fact checking of the textual content. That is why we decided to include this section with a more detailed look at tools and solutions available.

To verify if the shared visual content is manipulated or genuine, a number of tools are available online that can help verify manipulated visual content, for example, FotoForensics[4], Forensically[5], InVID[6], and WeVerify[7] etc. These tools, if utilised properly can be helpful in detecting a number of different image/video manipulations, for example, (1) copy/move forgeries, (2) image splicing forgeries, (3) face swapping etc.

However, at present, most of the visual mis/disinformation shared online does not always contain manipulated visual content, rather the genuine visual content is presented out-of-context, i.e., when visuals appear along with two (or sometimes even more) different and contradictory text captions online (Aneja et al., 2021). This type of visual mis/disinformation is not difficult to produce and can have devastating consequences. We briefly describe a number of different image/video forgeries below.

4.1.1 Copy/Move Forgeries

In Copy-Move forgery a small block of an image is copied and pasted at some other location within the same image. Copy-Move forgeries are typically employed to conceal objects or other information within an image, or sometimes employed to increase the number of objects present inside an image. Since the copy-move forgery uses regions from within the same image, it can be detected by doing an exhaustive search. The exhaustive search strategy analyses the image under consideration for any identical blocks present within the image (Shivakumar & Baboo, 2010). However, the exhaustive search strategy is computationally expensive and thus not suitable for real-world applications. After the advent of deep learning, more and more researchers employed deep neural network architectures to detect copy/move forgeries. The deep learning models do not require handcrafted features, rather, these algorithms are capable of extracting underlying patterns themselves during the training phase, and thus are more effective in detecting forgeries (Rao & Ni, 2016; Cozzolino et al., 2017; Zheng et al., 2019).

4.1.2 Image Splicing Forgeries

Image splicing forgeries are carried out by replacing a small segment from within the (target) image with a segment copied from another (source) image. Carefully manipulated images using splicing forgeries are difficult to detect and cannot even be perceived by the human eye in most cases. This detection becomes even more difficult after using post-processing operations on the spliced image, for example, flipping, rotation, skewing, stretching, scaling, and others (Kaur & Jindal, 2020).

[4] http://fotoforensics.com/
[5] https://29a.ch/photo-forensics/
[6] https://www.invid-project.eu/tools-and-services/invid-verification-plugin/
[7] https://weverify.eu/

Several detection techniques were proposed in the past, which were based on traditional machine learning based models as for example, support vector machines. The machine learning classifiers were trained on handcrafted features, (Kaur and Jindal, 2020; Ng et al., 2004; Popescu & Farid, 2005; Hsu &Chang, 2010). Dimensionality reduction techniques were also employed in the proposed techniques to make the detection models efficient.

However, after the widespread employment of deep convolutional neural network (CNN) architectures for vision related tasks, most of the new studies employed deep CNN architectures to detect splicing forgeries, achieving better results as compared to the previously proposed classical machine learning techniques (Pomari et al., 2018; Jaiswal & Srivastava, 2020; Rao et al., 2020).

4.1.3 Image Cropping

Image cropping forgeries are carried out by cropping out the image segments around the corners. Cropping operation is not considered to be malicious most of the time, however, it can be employed to spread mis/disinformation by cropping out certain objects from within an image to conceal or misrepresent the information present in the image.

Image cropping detection techniques detect this kind of manipulation by analysing the image for traces of resampling i.e., compression inconsistencies, etc. The resampling traces appear whenever a frame from a video or an image is cropped and enlarged. In order to make the resolution of all the frames across a video consistent, the process of resampling (precisely up-sampling) is carried out (Singh & Aggarwal, 2018). In the case of images, the cropped image is saved again. The saving operation introduces traces of resampling which can be detected.

Researchers employed a number of different feature types to detect upscale crop forgeries in images. For example, Hyun et al. (2013) employed sensor pattern noise (SPN) features to detect statistical correlations introduced by the resampling operation. Fanfani et al. (2020) proposed a fully automated asymmetrical image cropping detector on Manhattan-World scenes. The proposed technique employed the camera principal point features to detect image processing operations as for example, resizing and compression. The authors demonstrated the effectiveness of the proposed approach through extensive experimentation on a number of cropping scenarios.

Image cropping forgeries are not as widespread as other kinds of image forgeries e.g., copy/move, image splicing etc, but these forgeries can be used to spread misinformation (Singh & Aggarwal, 2018).

4.1.4 Deepfake Media

Nowadays with the disposal of low-cost graphics processing units (GPUs), and with the swift progress in the field of deep learning, more and more advanced classes of visual content forgeries are becoming widespread. Deepfake media is an example of contemporary visual content forgeries where the manipulation or generation of synthetic visual content is done using advanced deep learning models known as Generative Adversarial Networks (GANS), initially proposed by Goodfellow et al. (2014).

Some of the widely used open source deep fake media generation models are, (1) StyleGAN (Karras et al., 2019), (2) StyleGAN 2 (Karras et al, 2020), (3) FSGAN (Nirkin et al., 2019), (4) DCGAN (Radford et al., 2015), (5) FaceShifter (Li et al., 2019), (6) CycleGAN (Zhu et al., 2017). These deep generative models are capable of producing credible fake media, including real world scenes, faces, videos and audios.

There are numerous useful applications of these deep generative models, for example, (1) text-to-image generation and vice versa, (2) image-to-image translation (3) generating high quality cartoon/anime characters, (4) applications in the film industry to produce vfx effects etc.

However, at present these models are employed for malicious purposes, for example, to carry out face swapping, facial re-enactment etc. In face swapping operation, a generative model is employed to generate fake visual content by warping the face of one person (source) onto the face of another person (target) (Mirsky & Lee, 2021). In facial re-enactment operation, the model transfers the source person's face to the target person while keeping the identity of the target person intact.

Deep fake media detection is a difficult task and a lot of solutions have been proposed in the past to detect deep fake media. Nevertheless, nearly all the proposed solutions have one problem in common i.e., poor generalisation capability. Poor generalisation capability indicates that the deepfake media detection systems work brilliantly on the training data, but when these systems are evaluated on the real-world data or the data from other deepfake detection datasets, the detection systems perform poorly (Ferrer et al., 2020).

In addition to this, most of the studies focus on detecting the facial deep fakes and not the other types of deep fake content, for example, real world scenes etc.

4.1.5 Cheap Fake Media (Image Re-contextualisation)

Cheap fake media includes images and their accompanying textual descriptions. In most of the cases, the textual description or the associated image is not tampered, but instead presented out of context. In some rare cases the image or the accompanying text or both are manipulated. This makes the automatic detection of such content an extremely challenging

task, and demands human interference while validating this kind of manipulated media (Aneja et al., 2021).

Most of the recently proposed detection strategies employ deep learning models to detect cheap fake media. Some approaches tend to train two separate models, i.e., one model for textual feature extraction and another model for image feature extraction. The extracted features are fused to get a prediction of whether the evidence expressed by the image text pair is true or false (Aneja et al., 2021).

5. Conclusions

In our report**,** we are keen on the idea that technology is an enabler rather than a complete solution that can be of great help in terms of monitoring, claim matching, distribution and managing communities (Arnold, 2020). As done by Full Fact, the path to increased automation is to break down the constituent parts of fact-checking and testing whether any of these parts can be performed accurately by machines (Arnold, 2020). Automated fact-checking has received significant attention in the NLP community in the past years (Hanselowski et al., 2019). Technology can offer limited functionalities at the moment, so fully automating concepts around fact-checking, such as reviewing a piece of information and deciding whether it is right or wrong, is difficult. Many of the challenges experienced by fact-checkers are dependent on the political situation within a country; issues such as obtaining information from certain governments or lack of transparency and access to information (Arnold, 2020). One of the limitations of attempts to apply AI solutions on fact-checking is the lack of training data since the number of fact-checks around the world is quite small.

There are also general problems with introducing AI technology in organisations, such as lack of human and financial resources, misinterpretations and unrealistic expectations of technology (Kapoor & Klueter, 2017; Davenport & Ronanki, 2018)

In the report we detected gaps in technology for fact-checking and focused on verification tools for visual content, which also fits the research agenda at University of Bergen. However, we will develop a deeper understanding in our next deliverable that focuses on the user needs of fact-checkers. For that we are working closely with other hubs in the European Digital Media Observatory. Issues we will address are, for instance, how accurate are the tools in detecting disinformation and to what extent can they be used by different user groups? How much technical knowledge does one need to have in order to use these tools? How widespread are these tools we identified? How does language and context affect how sensitive they are?

There are also limitations when it comes to transparency and open access in fact-checking technology. There is a need to keep methods and tools hidden from the entities that spread false claims to reduce the risk of reengineering by bad actors.

Fact-checking can be seen as a form of investigative journalism and when it comes to automation of work tasks there are limited opportunities in unique cases (Stray 2019): the biggest near-term potential lies in data preparation tasks, such as data extraction from diverse documents and probabilistic cross-database record linkage. Shortly, automation is most valuable taking care of routine tasks, such as mining claims.

The perceived promise of automated fact-checking might easily lead to misinterpretations of the technology and exemplified as leading to unrealistic expectations – similar issues as other industries struggle with (Davenport & Ronanki, 2018). As a note of caution, we underline that disinformation is largely a social and political problem that needs a broader approach than technical solutions alone.

6. References

Aneja, S., Bregler, C. and Nießner, M., 2021. COSMOS: Catching out-of-context misinformation with self-supervised learning. *arXiv preprint arXiv:2101.06278*.

Arnold, P. 2020, The challenges of online fact checking: how technology can (and can't) help.. London: Full Fact.

Bail, C. A., Guay, B., Maloney, E., Combs, A., Hillygus, D. S., Merhout, F., ... & Volfovsky, A. (2020). Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017. Proceedings of the national academy of sciences, 117(1), 243-250.

Barberá, P., & Rivero, G. 2015. Understanding the political representativeness of Twitter users. Social Science Computer Review, 33(6), 712–729.

Cozzolino, D., Poggi, G. and Verdoliva, L., 2017, June. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security* (pp. 159-164).

Davenport, T. H., & Ronanki, R. 2018. Artificial intelligence for the real world. Harvard business review, 96(1), 108-116.

Dias, N., & Sippitt, A. 2020. Researching fact checking: present limitations and future opportunities. The Political Quarterly, 91(3), 605-613.

Dobbs, M. 2012. The Rise of Political Fact-checking How Reagan Inspired a Journalistic Movement: A Reporter's Eye View. New America Foundation.

Fanfani, M., Iuliani, M., Bellavia, F., Colombo, C. and Piva, A., 2020. A vision-based fully automated approach to robust image cropping detection. *Signal Processing: Image Communication*, *80*, p.115629.

Ferrer, C.C., Dolhansky, B., Pflaum, B., Bitton, J., Pan, J. and Lu, J., 2020. Deepfake detection challenge results: an open initiative to advance AI. *Facebook AI,[online], https://ai. facebook. com/blog/deepfake-detection-challenge-resultsan-open-initiative-to-advance-ai*.

Full Fact. 2020. Report on the Facebook Third-Party Fact-Checking programme. https://fullfact.org/media/uploads/tpfc-2020.pdf

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. Generative adversarial nets. *Advances in neural information processing systems*, *27*.

Graves, L. 2016. Deciding what is true: The rise of political fact-checking in American journalism. Columbia University Press: New York.

Graves, L. 2018. Understanding the promise and limits of automated fact-checking. Oxford: Reuters Institute for the Study of Journalism.

https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-02/graves_factsheet_180226%20FINAL.pdf

Graves, L., & Amazeen, M. A. 2019. Fact-checking as idea and practice in journalism. Oxford Research Encyclopedia of Communication. Oxford, UK: Oxford University Press. https://doi.org/10.1093/acrefore/9780190228613.013.808

Graves, L., & Anderson, C. W. 2020. Discipline and promote: Building infrastructure and managing algorithms in a "structured journalism" project by professional fact-checking groups. New media & society, 22(2), 342-360.

Guess, A. M., & Lyons, B. A. 2020. Misinformation, disinformation, and online propaganda. In: Social media and democracy: The state of the field, prospects for reform, 10-33.Eds Parsely, N. & Tucker J. Cambridge, UK: Cambridge University Press.

Guy, H., 2017. Why we need to understand misinformation through visuals. FirstDraft, [online] *https://firstdraftnews.org/articles/understanding-visual-misinfo* (accessed 22.11.2021).

Hanselowski, A., Stab, C., Schulz, C., Li, Z., & Gurevych, I. 2019. A richly annotated corpus for different tasks in automated fact-checking. arXiv preprint arXiv:1911.01214.

Hao, K. 2021, How Facebook and Google fund global misinformation. https://www.technologyreview.com/2021/11/20/1039076/facebook-google-disinformation-clickbait/ (accessed 22.11.2021).

Hsu, Y.F. and Chang, S.F., 2010. Camera response functions for image forensics: an automatic algorithm for splicing detection. *IEEE Transactions on Information Forensics and Security*, *5*(4), pp.816-825.

Hyun, D.K., Ryu, S.J., Lee, H.Y. and Lee, H.K., 2013. Detection of upscale-crop and partial manipulation in surveillance video based on sensor pattern noise. Sensors, 13(9), pp.12605-12631.

Jaiswal, A.K. and Srivastava, R., 2020. A technique for image splicing detection using hybrid feature set. *Multimedia Tools and Applications*, *79*(17), pp.11837-11860.

Kapoor, R. & Klueter, T. 2017, Organizing for new technologies, MIT Sloan Management Review, vol. 58, nr. 2, s. 85.

Karras, T., Laine, S. and Aila, T., 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4401-4410).

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J. and Aila, T., 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8110-8119).

Kaur, H. and Jindal, N., 2020. Image and video forensics: A critical survey. *Wireless Personal Communications*, *112*(2), pp.1281-1302.

Kreiss, D., & McGregor, S. C. (2019). The “arbiters of what our voters see”: Facebook and Google’s struggle with policy, process, and enforcement around political advertising. Political Communication, 36(4), 499-522.

Lerner, J., & Lomi, A. (2020). The free encyclopedia that anyone can dispute: An analysis of the micro-structural dynamics of positive and negative relations in the production of contentious Wikipedia articles. Social Networks, 60, 11-25.

Li, L., Bao, J., Yang, H., Chen, D. and Wen, F., 2019. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*.

Lind, D. 2018. President Donald Trump finally admits that “fake news” just means news he doesn’t like like. https://www.vox.com/policy-and-politics/2018/5/9/17335306/trump-tweet-twitter-latest-fake-news-credentials (accessed 2.3.2022)

Lindén, C-G. 2020, AI to the Rescue of Combating Fake News? It is largely a social and political problem. Media Motor Europe. https://mediamotoreurope.eu/ai-to-the-rescue-of-combating-fake-news-it-is-largely-a-social-and-political-problem/ (accessed 01.02.2022).

Marietta, M., Barker, D. C., & Bowser, T. 2015. Fact-checking polarized politics: Does the fact-check industry provide consistent guidance on disputed realities? The Forum, 13(4), 577–596.

Menking, A., & Rosenberg, J. 2021. WP: NOT, WP: NPOV, and other stories Wikipedia tells us: A feminist critique of Wikipedia’s epistemology. Science, Technology, & Human Values, 46(3), 455-479.

Mirsky, Y. and Lee, W., 2021. The creation and detection of deepfakes. *ACM Computing Surveys*, *54*(1), p.7.

Myers, S.L., Mozur, P. & Kao, J. 2022, How Bots and Fake Accounts Push China’s Vision of Winter Olympic Wonderland. https://www.propublica.org/article/how-bots-and-fake-accounts-push-chinas-vision-of-winter-olympic-wonderland?token=F1W4XoVHukeDvXTmrb8HqM4rO-mJyOjp (accessed 23.2.2022).

Ng, T.T., Chang, S.F. and Sun, Q., 2004, May. Blind detection of photomontage using higher order statistics. In 2004 IEEE international symposium on circuits and systems (IEEE Cat. No. 04CH37512) (Vol. 5, pp. V-V). IEEE.

Nirkin, Y., Keller, Y. and Hassner, T., 2019. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 7184-7193).

Nissen, I. A., Walter, J. G., Charquero-Ballester, M., & Bechmann, A. 2022. Digital Infrastructures of COVID-19 Misinformation: A New Conceptual and Analytical Perspective on Fact-Checking. Digital Journalism, 1-23.

Nyhan, B. & Reifler, J. 2015. Estimating fact-checking's effects: Evidence from a long-term experiment during campaign 2014. Working Paper, American Press Institute.

Ott, B. L., & Dickinson, G. 2019. The Twitter presidency: Donald J. Trump and the politics of White rage. Routledge.

Persily, N. & Tucker, J.A. (eds.) 2020. *Social Media and Democracy: The State of the Field, Prospects for Reform*. Cambridge: Cambridge University Press.

Popescu, A.C. and Farid, H. 2005. Exposing digital forgeries by detecting traces of resampling. *IEEE Transactions on signal processing*, *53*(2), pp.758-767.

Pomari, T., Ruppert, G., Rezende, E., Rocha, A. and Carvalho, T. 2018, October. Image splicing detection through illumination inconsistencies and deep learning. In *2018 25th IEEE International Conference on Image Processing (ICIP)* (pp. 3788-3792). IEEE.

Radford, A., Metz, L. and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Rao, Y. and Ni, J., 2016, December. A deep learning approach to detection of splicing and copy-move forgeries in images. In *2016 IEEE International Workshop on Information Forensics and Security (WIFS)* (pp. 1-6). IEEE.

Ruprechter, T., Santos, T., & Helic, D. 2020. Relating Wikipedia article quality to edit behavior and link structure. Applied Network Science, 5(1), 1-20.

Shivakumar, B.L. and Baboo, L.D.S.S., 2010. Detecting copy-move forgery in digital images: a survey and analysis of current methods. *Global Journal of Computer Science and Technology*.

Singh, R.D. and Aggarwal, N., 2018. Video content authentication techniques: a comprehensive survey. *Multimedia Systems*, *24*(2), pp.211-240.

Smith, S.H. 2004. The Fact Checker's Bible: A Guide to Getting it Right. New York: Anchor Books.

Stray, J. 2019. Making artificial intelligence work for investigative journalism. *Digital Journalism*, *7*(8), 1076-1097.

Tsfati, Y., Boomgaarden, H.G., Strömbäck, J., Vliegenthart, R., Damstra, A. & Lindgren, E. 2020. Causes and consequences of mainstream media dissemination of fake news: literature review and synthesis, Annals of the International Communication Association, vol. 44, nr. 2, s. 157-173.

Waisbord, S. 2018. Truth is What Happens to News, Journalism Studies, 19:13, 1866-1878, DOI: 10.1080/1461670X.2018.1492881

Wendling, M. 2018. The (almost) complete history of 'fake news' . https://www.bbc.com/news/blogs-trending-42724320.

Zarocostas, J. (2020). How to fight an infodemic. The Lancet, 395(10225), 676.

Zheng, L., Zhang, Y. and Thing, V.L. 2019. A survey on image tampering and its detection in real-world photos. *Journal of Visual Communication and Image Representation*, *58*, pp.380-399.

Zhu, J.Y., Park, T., Isola, P. and Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223-2232).