



Artificial Intelligence in the Service of Verification

Published in January 2025



Authors:

Nikos Sarris - Centre for Research and Technology Hellas - Information Technologies Institute, Thessaloniki, Greece

Danae Tsabouraki - Athens Technology Center, Athens, Greece

Zoi Palla - Centre for Research and Technology Hellas - Information Technologies Institute, Thessaloniki, Greece

Spyros Papafragkos - Athens Technology Center, Athens, Greece

Stefanos-Iordanis Papadopoulos - Centre for Research and Technology Hellas - Information Technologies Institute, Thessaloniki, Greece

Efstratios Tzoannos - Athens Technology Center, Athens, Greece

Giorgos Giotis - Athens Technology Center, Athens, Greece

Ioannis Kompatsiaris - Centre for Research and Technology Hellas - Information Technologies Institute, Thessaloniki, Greece

Symeon Papadopoulos - Centre for Research and Technology Hellas - Information Technologies Institute, Thessaloniki, Greece

This report aims to provide an in-depth analysis of how Artificial Intelligence (AI) technologies are being used to address the pervasive issue of disinformation. It outlines AI's role in automating media verification, explores its applications in textual and visual content analysis, and examines its potential in multimodal disinformation detection. By addressing technical challenges, ethical considerations, and future directions, this report seeks to serve as a comprehensive guide for researchers, policymakers, and practitioners in the field of AI-based media verification.

The cover image of this report was generated using AI on the Canva platform with the prompt: "Create an image for a report on Artificial Intelligence in verification and disinformation detection, symbolizing representations of truth through a magnifying glass and a checkmark"

Introduction

In the age of digital connectivity and information dissemination, the spread of disinformation has become a remarkable challenge for societies worldwide. The development of Artificial Intelligence (AI) has added a layer of complexity to this issue, as it is both a facilitator and a potential solution. This report highlights the area of media verification based on AI methods, discussing the scope, key steps, challenges, and potential of using AI verification to combat the spread of disinformation.

The motivation behind exploring AI-based media verification lies in the urgent need to protect the reliability of information ecosystems. The digital landscape generated by social media and online platforms has become a breeding place for disinformation, increasing its impact on public opinion and undermining trust in information sources. As disinformation evolves, so must strategies to combat it. AI, which can process large amounts of data and detect patterns, is emerging as a powerful tool to optimise verification processes and mitigate the impact of false narratives.

This report extends beyond highlighting the existence of disinformation to examining the role AI plays in verifying the authenticity of information. From identifying helpful cues in textual content to recognising deepfakes and synthetically generated content, the application of AI in verification processes is broad and multifaceted.

Establishing a common understanding of key terms and concepts is essential to navigate the discourse on AI verification methods. AI, in this context, refers to the use of machine learning algorithms and computational models to simulate human intelligence in data analysis and verification.

Such methods face many challenges, both technical and social. Technically, the rapid evolution of disinformation techniques requires quickly adaptive AI models. The constant refinement of deepfake generation technologies, for example, and the potential for exploiting algorithmic vulnerabilities in digital platforms pose intimidating challenges to effective disinformation countering efforts. In social contexts, the acceptance and ethical use of AI in media verification is a delicate issue. Privacy concerns, algorithmic biases, and the possibility of censorship require careful consideration. Striking a balance between leveraging AI for media verification and ensuring transparency and accountability is crucial to encourage public trust in these technologies.

The report is organised in a way to provide a short overview of AI-based media verification methods. It lays the foundation by defining key terms and concepts, presenting a number of state-of-the-art approaches and their potential for supporting fully automatic fact-checking, but also explaining the ways in which textual, image, video and multimodal disinformation can be detected. As effective as such methods can be, they can still not substitute the role of human investigators but can be a useful aid to the work of fact-checkers and be instrumental in cases where visual inspection is not adequate.

AI for Automated Fact Checking and Textual Analysis

In this section, we explore how AI can assist both in automated fact-checking and textual content analysis for more efficient and reliable information processing. Fact-checking, often time-consuming, benefits significantly from AI automation, addressing the challenges posed by the rapid growth of online disinformation. Our exploration covers various domains and starts by discussing current research and development methods following the three components of automated fact-checking: identifying claims, gathering evidence, and reaching a verdict. Additionally, we delve into practical applications of advanced AI methods, from analysing sentiments to addressing biases and identifying synthetic text. This section provides guidance on diverse techniques, including claim detection, evidence matching, and thorough examination of synthetic text, showcasing AI's contributions to improving both automated fact-checking and textual content analysis.

Claim detection and classification is the first step in automated fact-checking, where AI algorithms identify potential claims in text that need verification. Advanced NLP techniques are used to distinguish between factual statements and opinions or non-verifiable statements. This area also includes identifying which claims are worth fact-checking based on their potential impact or likelihood of being false or misleading. Classical supervised learning methods such as Support Vector Machines, Naïve Bayes, and Random Forests are used for classifying text as “claim” or “non-claim” [1]. These are trained on hand-crafted features extracted from the text, such as word frequencies, part-of-speech tags, and syntactic structures. More recently, deep neural network architectures, including Recurrent Neural Networks, Graph Neural Networks, and Long Short-Term Memory networks, are also used for claim detection [2]. In particular, transformer-based models like BERT and its variants (e.g., RoBERTa, ALBERT) have been successfully used due to their superior ability in understanding context and semantics [3].

Evidence Retrieval is the next step in the fact-checking process. Once a claim is detected, AI-based systems search for evidence to verify the claim. This involves Information Retrieval (IR) algorithms, leveraging concepts such as TF-IDF (Term Frequency-Inverse Document Frequency) for traditional keyword matching and document ranking [4]. Recent research emphasises improving the relevance and reliability of retrieved evidence [5]. To achieve semantic relevance, text embeddings like Word2Vec, GloVe, and transformer-based models (e.g., BERT) are employed [6-7]. This step aligns textual content with pertinent evidence, utilising various AI techniques, including NLP, Named Entity Recognition (NER), and semantic matching algorithms [4]. By correlating claims with supporting evidence, this comprehensive approach enhances the credibility and reliability of fact-checking. However, challenges arise concerning the availability and quality of evidence, necessitating ongoing advancements in AI methodologies for a thorough and evidence-based verification process.

Claim verification is at the core of automated fact-checking and is often commonly addressed in tandem with the evidence retrieval task. AI systems use various models to categorise claims into different veracity levels based on the retrieved evidence. Research is ongoing to make these assessments more accurate and nuanced, considering the often-complex nature of information veracity [2]. Textual Entailment Models, often based on deep learning, assess whether a piece of evidence supports, contradicts, or is neutral towards a claim. Techniques like decomposable attention models or transformer-based approaches are used for this purpose. Convolutional Neural Networks are also used to determine the stance of the retrieved evidence relative to the claim (supporting, denying, querying, or commenting), while more recent work deals with the verification of more complex claims, where many pieces of evidence are retrieved (with irrelevant information being discarded) and combined for reaching a verdict [2].

After briefly discussing how AI technologies can be used for automated fact-checking, we move forward to present how AI is also instrumental in analysing textual content for various other purposes, such as sentiment analysis, bias detection, and hate speech identification, which can either be part of automated fact-checking or serve as standalone processes.

Recognising and categorising topics in textual content is imperative for effective organisation and comprehension of extensive datasets. Topic detection employs various AI methods, including traditional approaches like Latent Dirichlet Allocation (LDA) and advanced techniques such as neural network-based models [8]. By classifying text into predefined topics, this process plays a crucial role in structuring and categorising information for streamlined verification, thereby enhancing information retrieval and organisation. Employing AI methods for topic detection faces potential challenges, such as sensitivity to shifts in language usage and the possibility of overlap between topics. While proficient in organising large datasets and identifying predominant themes, this approach may encounter difficulties with emerging or evolving topics.

Sentiment extraction involves determining the emotional tone expressed in a piece of text, leveraging Natural Language Processing (NLP) techniques [9]. Machine Learning (ML) models, including recurrent neural networks (RNNs) and transformers, classify text into positive, negative, or neutral sentiments. It aids in understanding public opinion, assessing sentiment around specific claims, and offers quick analysis of societal response. Sensitive to context and cultural nuances, sentiment extraction may struggle with sarcasm or subtle expressions. Nevertheless, it can be effective for gauging public sentiment at scale, though limited by language complexity and context.

Detecting hate speech in textual content is crucial for maintaining online civility and preventing the spread of harmful content. Hate speech detection often involves ML models trained on labelled datasets containing instances of hate speech [10]. Identification and flagging of content containing hate speech contribute to a safer online environment. This approach allows for rapid identification of harmful content, fostering a healthier online discourse. However, potential biases in training data and the evolving nature of language usage pose risks. While scalable and effective in identifying explicit instances of hate speech, this method may struggle with subtle forms or context-dependent language.

Understanding and **mitigating bias in textual content** is essential for creating fair and unbiased AI models. Bias estimation involves scrutinising textual data to uncover potential biases. This is achieved through the application of machine learning models trained to discern and highlight patterns indicative of bias [11]. These models, developed through extensive training on diverse datasets, are designed to analyse linguistic nuances and contextual intricacies, enabling a comprehensive assessment of potential biases in the given text. It helps in identifying and rectifying biases in information sources, raising awareness for informed decision making. However, subjectivity in defining bias and reliance on labelled datasets pose risks. While contributing to building fair and unbiased AI systems, this method may struggle to capture subtle biases and faces challenges in defining a universal standard for bias.

Identifying **synthetic or generated text** is crucial for maintaining the authenticity of information, and recent advancements in language models, particularly large language models (LLMs), play a pivotal role in this process [12]. LLMs, such as ChatGPT, have demonstrated remarkable capabilities in discerning patterns and characteristics associated with synthetic text, contributing significantly to the identification and flagging of potentially fabricated information [12]. While effective in detecting common patterns, this approach encounters challenges when dealing with highly sophisticated synthetic text, exposing vulnerabilities to adversarial attacks and the evolving sophistication of text

generation techniques [13]. Nevertheless, the adoption of LLMs in synthetic text detection remains advantageous, as it enhances the reliability of information sources by employing a vigilant approach to identify and mitigate potential risks associated with synthetic text [14].

Fact-checking organisations are already experimenting with these AI technologies. Full Fact¹ and Chequeado's² collaboration led to a multi-component AI tool that includes claim detection, a matching tool for checking online repetitions of already-checked claims, and a stats-retrieval tool for finding relevant fact-checking statistics [15]. The Duke Reporters' Lab³ has developed "Squash" for live fact-checking, using a large database of claims to match statements in videos with previously published fact-checks, while Newtral⁴ has developed Claim Hunter, an AI tool that transcribes audio and detects statements for fact-checking, while it also monitors politicians' Twitter accounts, sending alerts when a factual statement is shared [15].

In summary, AI applications for text analysis are essential tools for thorough fact-checking and information verification. Each application has strengths and weaknesses, emphasising the need for a careful and adaptable approach. Continuous improvement of methods is crucial to match technological capabilities and text details. Overall, the focus is on accuracy and reliability in organising information within a complex informational environment. When it comes to AI-assisted automated fact-checking, despite the continuous advancement of AI technologies, several challenges and constraints remain, limiting the potential of automated fact-checking at scale and without human supervision. The complexity of language, which makes it difficult for AI to fully grasp context, nuances, and subtleties in human communication. Differentiating between facts, opinions, and sarcasm is particularly challenging. Then, the reliability and completeness of data sources used for verification are critical. Disinformation or incomplete data can lead to incorrect fact-checking outcomes. Another significant challenge is the rapid evolution of disinformation tactics, which requires constant updates in the detection algorithms. Additionally, maintaining impartiality and avoiding biases in AI algorithms is crucial to ensure fair and unbiased fact-checking. Lastly, the scalability of automated systems to handle the vast volume of information on the Internet and in media, while maintaining accuracy and speed, presents a considerable technical challenge.

AI in Verification of Visual Content

'A picture is worth a thousand words' - visual content has always been instrumental in conveying information more easily and in a much more persuasive way than mere verbal descriptions. Taking advantage of this fact, conveyors of disinformation have always been keen to use visuals in various deceitful ways to maximise the effect of their messages. Visuals can be deceitful either by being visually manipulated, by manipulating their context, or, lately, by being entirely artificially generated.

Visual manipulation is often implemented by means of image forgery or tampering, and, in particular, it has traditionally been carried out by photo editing tools, but lately, it is supported in much more advanced ways through AI technologies. There are generally three main kinds of image tampering: splicing, which is copying a region from a different image into a host image; copy-move, which is taking one object and replicating it within the same image; and inpainting, which is removing a piece of an image and filling in the missing parts through interpolation or through AI-generated content. Recent AI methods also enable "outpainting", which is filling in parts of image outside the original image frame based on advanced generative methods. There are also other kinds of image manipulations, typically innocuous, like content enhancement (e.g., changing of colours, contrast,

¹ <https://fullfact.org/>

² <https://chequeado.com/>

³ <https://reporterslab.org/>

⁴ <https://www.newtral.es/>

and brightness), which can often be used in combination with other tampering techniques (see Figure 1).



Figure 1: On top: An example of inpainting generated with Nvidia interactive demo. The inpainting masks (i.e., the deleted region that must be filled by the inpainting algorithm) is shown in yellow. On the right bottom (persons): A high profile, in-the-wild example of a splicing forgery that widely circulated in 2004. The figure of Jane Fonda captured in an unrelated 1972 photo, has been spliced into a pre-existing John Kerry picture taken in 1970. [16]

Recent research [16] has demonstrated that humans are not good in detecting image tampering. It is fortunate that the field of image forensics has emerged to equip human analysts with powerful tools to assess the authenticity and integrity of images. Image forensics is based on the fact that any photo generation process involves specific device hardware (e.g., lens, colour filter array, sensors) and (embedded) software, and depending on the model and the manufacturer, these leave different traces in an image. Any editing carried out outside the camera alters the patterns of these traces, and should, therefore, be detectable by specific forensics algorithms [17].

These methods have recently been significantly improved by leveraging the power of deep learning. Such an example is a popular deep learning-based approach called NoisePrint [18], where the key idea is that the convolutional neural network is trained with patches of many different images that are from the same camera and the same position within the photo. The network tries to focus on the noise patterns instead of the actual content of the image and after being trained on a sufficiently large sample of such patches, it can produce “noiseprints” which can reveal the areas of an image that have been tampered.

A recent trend [19] is to combine different streams of an image, e.g., the RGB stream together with the noise stream produced through some digital filtering of the image or even a frequency stream like the discrete cosine transform of the image. These streams train a deep learning architecture and then there are many ways to combine them through fusion layers.

However, visual manipulation has rapidly advanced in recent years in terms of level of realism and ease of production. This has been largely due to the emergence of AI technologies that are used to generate synthetic media, widely referred to as ‘deepfakes’, i.e. content generated, at least partly, by deep neural networks. Even though deepfake content can refer to any kind of synthetic content, the most common form is focusing on human faces and human subjects distinguishing four very well-known types: entirely synthetic faces; attribute manipulation (e.g., changing hairstyle); identity (face) swap; and reenactment (expression swap) [20-21].

Deepfake generation technologies have rapidly advanced since the first versions of Generative Adversarial Networks (GANs) [22] to the more recent emergence of Diffusion Models [23]. While GANs introduced the minimax game between the generator and the discriminator, which continuously lead to improved performance of the generator, the diffusion models are based on a noise diffusion process where noise is added progressively generating more noisy versions of an image in a Markovian way, followed by the inverse process, which tries from random noise to learn how to generate a synthetic denoised image. It turned out that this process led to very powerful generative models with very versatile capabilities, becoming really popular through OpenAI’s DALL-E and Google’s Imagen, and also enabled text prompting to generate corresponding images (see Figure 2). Since then, we have seen rapid advances with capabilities that are really astonishing, also

extending to text-to-video diffusion models. Furthermore, there is another family of generative models, named Neural Radiance Fields (NeRFs), which are inspired by more classic computer graphics approaches, trying to learn from very few images of different viewpoints of the same scene, in order to generate additional viewpoints and to provide a sense of a full 3D reconstruction [24].

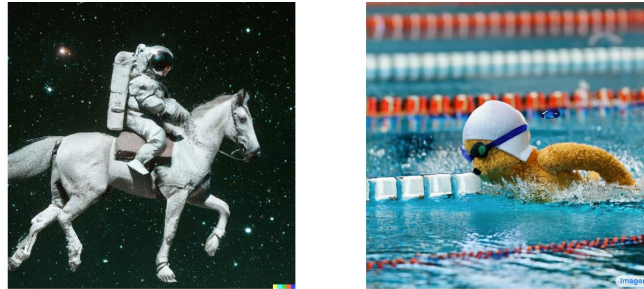


Figure 2: Two examples of text-to-image diffusion models by OpenAI and Google: (left) 'an astronaut riding a horse in photorealistic style' by DALL-E-2 www.openai.com/dall-e-2 (right) 'teddy bear swimming at the Olympics 400m Butterfly event' by Imagen <https://imagen.research.google>

Detecting deepfake content is consequently becoming increasingly challenging as there is a demand for detectors to continuously adapt to the fast-pacing advances of generators. Deepfakes generated by early models such as StyleGAN are still possible to detect even by human inspection as there are some imperfections like strange artefacts in the ears, or inconsistencies in the eyes allowing trained individuals to be able to spot them. Examples of such imperfections are nicely illustrated by various projects⁵ human observation and inspection is not as effective as an AI-based system.

For this reason, many models and approaches have been proposed and have been classified under various families and categories. One such categorization classifies approaches under a) those that try to capture and analyse physiological signals, b) those that try to detect artefacts, c) deep learning architectures, d) those that are based on the analysis of the frequency spectrum and e) multimodal approaches. Approaches based on physiological features examine features such as eye blinking, because it has been shown that there are specific natural characteristics associated with blinking, e.g. its frequency and speed, and most generative models cannot fully reproduce them. As a result, it is in principle possible to train deep learning models that can manage to distinguish between natural and synthetic blinking [25].

Other useful physiological features, for visual content with sufficient resolution, can be extracted by the analysis of corneal specular highlights, which refers to the way that light is reflected on the surface of the human eye [26].

Methods that try to detect generative artefacts typically focus in areas like the lips, teeth, or eyes as shown in Figure 3. They achieve this by specialised training and architectures, but also use specific data augmentation schemes, as artificially blurring parts of images to emulate the effect of a deepfake image and train more robust detection models [27-28]. Another more specific kind of artefact that is very common in face swapping methods is the so-called face X-ray, where there is a halo effect in the borders of swapped faces, as this is where the faces blend into the background [28].

⁵ such as www.whichfaceisreal.com or detectfakes.media.mit.edu

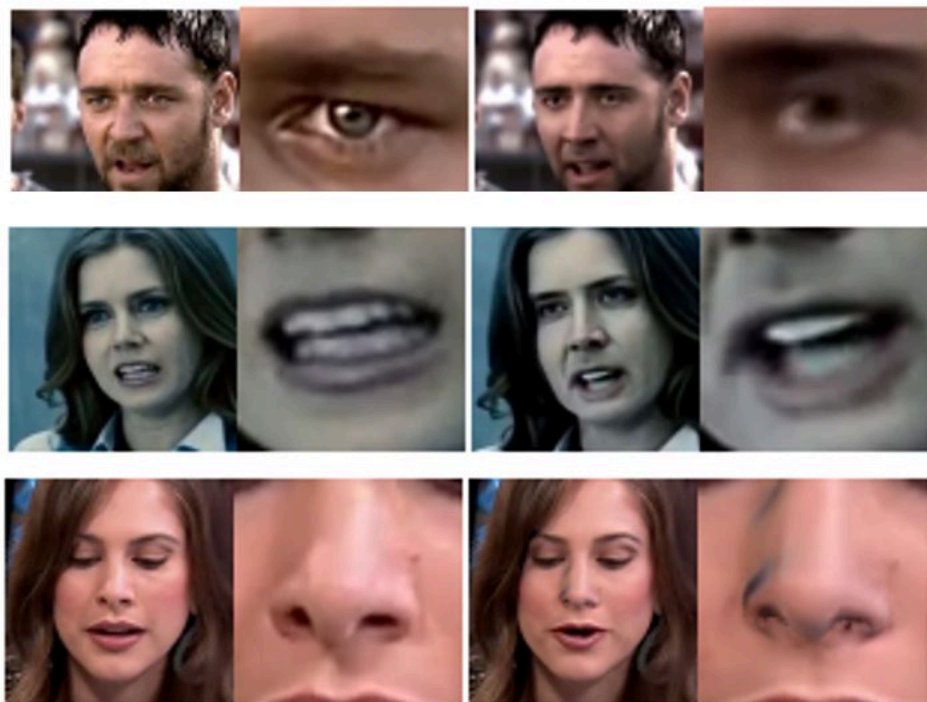


Figure 3: Exploiting visual artefacts to expose deepfakes and face manipulations [27]

The most commonly used methods, however, are those that train a deep neural network to detect whether an input image or video is deepfake or not. The earliest and most simple architecture is Mesonet [29], evolving to more effective approaches like XceptionNet [30], Capsule Networks [31] and Attention Heads [32]. The recent trend, however, is the use of Vision Transformer models to enhance the accuracy of deepfake detection as they proved to be quite effective especially in capturing the spatial and temporal relations of image patches [33].

Another way of detecting GAN generated images is based on analysis in the frequency domain. In [34] researchers have shown that it may be possible to create a universal detector for telling apart real images from those generated by a GAN, regardless of architecture or dataset used.

Finally, successful methods have recently been presented in attempting the cross examination of different modalities, such as speech audio and lips video in order to spot inconsistencies in their synchronization. The 'real-forensics' method [35] tries to learn in a self-supervised way how modalities interact with each other in a large set of authentic videos and then uses this representation in a supervised learning scheme to achieve a better distinction between fake and real videos.

The field of deepfake generation and detection offer a naturally adversarial setting with new methods constantly appearing in both sides. Already back in February 2021 a survey identified 70 generation and 108 detection methods indicating the fast evolution of the field [36]. Deepfake datasets are also evolving fast and already include three generations, starting from very small ones and moving now to datasets that contain tens of thousands of video snippets. Still however, there is a significant gap between benchmark datasets and actual deepfakes encountered in the wild. Although there have been attempts to source deepfakes from the Internet, the largest datasets to date are still generated in the lab, so they are often not really reflecting realistic conditions [37].

Despite intensive efforts by the relevant scientific communities, the challenges in deepfake detection remain unsolved. Robustness to adversarial attacks is extremely hard as most scientific results are openly published and become exposed to malicious actors [38]. Generalisation to

unseen generative models is only possible through continuous learning, although recently, approaches have been proposed that focus on learning the characteristics of authentic (instead of fake) content [18]. Efficiency and scalability aspects become a real problem as models become larger while transparency and explainability must also be constantly sought in order to be able to support cases with real data [39].

AI in countering multimodal disinformation

In the previous section we explained how multimodal features can help in the context of deepfake detection, but multimodal disinformation, as a problem, refers to false or misleading information that is spread using multiple modes of communication, such as text, images, audio and video. One common way is images and texts where legitimate/unaltered images are accompanied by texts that misrepresent some aspect of the image, e.g. its context, origin, location, depicted entities etc, as shown in Figure 4 below.

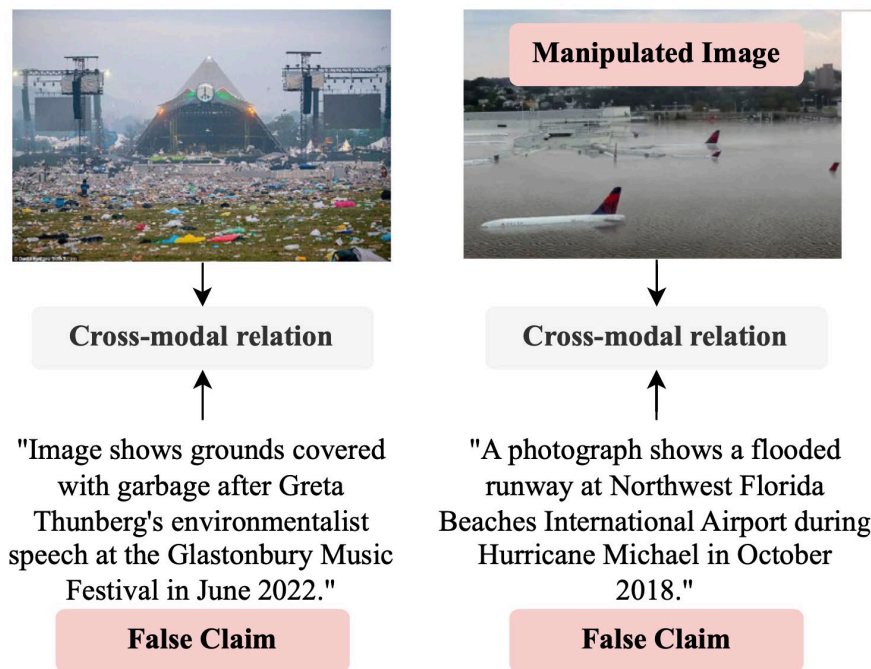


Figure 4: Examples of multimodal disinformation [40]

Detection methods require a training dataset and some (deep learning) model to extract and fuse features from multiple modalities. One of the first, very simple, approaches that used the VMU Twitter dataset [41], employed pre-trained text and visual encoders as feature extractors, namely BERT for text and VGG for images. The output features were concatenated and used to train a very simple layer of a network to distinguish between real and fake information [42]. Many more sophisticated architectures were proposed improving this approach, as the one proposed by [43] where such an initial step of separate image and text feature extraction, is followed by a fully connected layer (i.e. a bi-modal representation) that is directed both to a decoder and also to a 'fake news' detector classifier, in an effort to optimise the reconstruction loss and the binary classification loss.

Another version of this approach is using more sophisticated embeddings, as an improved vision transformer, along with a more sophisticated fusion mechanism, with even better results [44].

Previous datasets (VMU Twitter) had relatively limited size. Deep learning models benefit from larger datasets, but it is labour-intensive and costly to create a large, annotated dataset. For this reason, researchers have been trying to create "algorithmically generated" disinformation. Numerous works have focused on "out-of-context" multimodal disinformation, which is very often

used for generating and propagating misleading narratives, often to amplify pre-existing biases or beliefs (see Figure 5). This usually involves an image or video accompanied by a message which is misleading on purpose. It is typically sensational and is often used for clickbait purposes, often also referred to as “cheapfake”, because of the low effort and cost required to create such content.



Figure 5: Examples of out-of-context cases [45]

For out-of-context cases of multimodal disinformation, a variation has been proposed affecting the way the model is trained, maintaining the building components in principle the same. Universal sentence encoder embeddings are used together with a region convolutional neural network for the vision part, but the training is carried out with algorithmically created contrastive examples (i.e. an image with the real caption in contrast to the image with a false caption) utilising random sampling and guiding the model to distinguish these out of context relations between the two modalities [45].

This was followed by another approach for semi-synthetically generating a multimodal disinformation dataset, called NewsCLIPings [46]. Instead of solely relying on random sampling (as in COSMOS-training) which typically creates easy to detect disinformation, NewsCLIPings generates hard examples of out of context disinformation by using multimodal similarity using the CLIP model, along with Person matching and Scene matching computer vision models, in order to generate misleading pairs of image text that are at the same time also challenging to detect.

A recent trend for detecting such out-of-context disinformation is to leverage external knowledge sources in addition to pre-trained models [47-48]; for example, use the top-K results from a search engine. As a way to emulate how fact-checkers typically collect additional external information in order to verify a claim, this external knowledge is used together with the pair of image and text. Extending this approach [49] attempt to identify which of the collected external information is relevant to support or refute a claim. Leveraging external information showed significant improvements in detection accuracy.

Challenges

A key difficulty in multimodal disinformation is creating a mechanism to systematically integrate various modalities so that one adds value to the others. The current leading approaches mainly utilise early and late fusion, which are restricted methods and do not consistently produce strong outcomes. Recently, multimodal transformer models that are trained together (such as ViLBERT, Visual BERT, and Multimodal Bitransformers (MMBT))

have demonstrated the potential for significant enhancements. However, these models are developed with just two modalities in mind (text and visual), whereas fact-checking or disinformation-related materials typically involve a broader range of modalities (like text, speech, video, network, etc.).

Additionally, current approaches to detecting disinformation tend to lack contextualization, meaning they do not account for the wider context of a news piece, including readers' reactions and perceptions. In addition to the news and its context, factors such as the authenticity of the information, the reliability of the authors, and the factual accuracy of the content are also significant elements in identifying disinformation.

Bias, regional factors, and cultural understanding are also crucial. The effectiveness of many existing systems is constrained by the data they are based on, especially concerning demographic and cultural elements. For example, a model trained using an Indian political dataset may struggle to perform well on a US health-related dataset. Detection models must be structured in a manner that ensures their results are free from bias. Models designed for disinformation detection should convey results clearly so that users can interpret them, comprehend the reasons a piece of information is marked as disinformation, identify the corresponding factual news that informed the judgment, and pinpoint which aspects of the information are problematic.

Another challenge is that most current disinformation detection systems function as binary classifiers: they simply determine whether a piece of news is disinformation or not. While such binary signals may be adequate in certain situations, many others require more detailed labels and a deeper level of analysis. For instance, identifying whether a social media post is misleading can alert fact-checkers, but providing more specific categories such as true, satire/parody, misleading, manipulated, false connection, or imposter content could be even more beneficial.

Finally, the changing landscape of disinformation issues continues to present difficulties. Often, assertions or damaging information are spread in relation to ongoing events; information regarding COVID-19 and vaccines serves as examples of this phenomenon. Current models may struggle with these scenarios, making zero-shot or few-shot learning a significant path for future investigation.

Conclusion, challenges and future directions

This report described the impact of AI in automating verification processes, addressing the growing challenges posed by the increasing prevalence and complexity of online disinformation.

Three phases for automated fact-checking were examined: claim detection and classification, evidence retrieval, and claim verification. In all three, application of NLP

methods, machine learning classifiers, transformers, AI algorithms, and other sophisticated models have shown promising results, although not mature enough to establish them as something more than experimental assistive tools in the workflows of early adopter fact-checking organisations such as Full-Fact and Chequedo. We examined AI methods for textual analysis that can aid in the identification of underlying sentiments, hate speech, bias, topics, claims, as well as synthetically generated text. Progress of AI was examined in the automatic analysis of each modality used to convey (mis)information (i.e. text, image and video) separately, as well as in a multimodal way.

Visual analysis methods were examined for identifying image manipulations that range from very simple forgeries to deepfakes that require detection of complex signals, such as physiological imperfections and nearly invisible artefacts in the spatial or frequency domain.

Detection of multimodal disinformation is also explored, outlining methods that utilise algorithmically generated examples, external knowledge sources, and cross-modal inquiry. We stress the necessity for nuanced techniques to regulate the emergence of false information, highlighting issues with prejudice, cultural sensitivity, and the applicability of research findings in interpretation.

The current state of AI verification reflects significant advances in fact-checking and automated content analysis, as transformer-based architectures have proven effective and collaboration between organisations in the development of AI tools result to practical applications. Nonetheless, there are still issues with linguistic complexity, mitigating bias, and developing disinformation tactics.

Future directions may include refining techniques by addressing the limitations of current detection systems. Advancements in mitigating adversarial attacks, enhancing efficacy and extensibility, and promoting openness and explainability are imperative for enhancing AI-based media verification. Ongoing model training is essential to keep up with evolving disinformation techniques, while robust multimodal fusion methods and contextualised detection models need to be developed. It is also important to address bias, raise cultural understanding, and improve interpretability and transparency while optimising detection models for regional and cultural contexts ensures global applicability.

The complexity of language presents a challenge to understand nuances, sarcasm, and subtleties. Rapid evolution of disinformation techniques requires constant updating of detection algorithms. All the while, mitigating bias in AI algorithms and handling large amounts of information on the Internet, ensuring accuracy and speed for automated systems present increasing challenges.

References

- [1] Hassan N, Arslan F, Li C, Tremayne M. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining 2017 Aug 13 (pp. 1803-1812).
- [2] Guo Z, Schlichtkrull M, Vlachos A. A survey on automated fact-checking. Transactions of the Association for Computational Linguistics. 2022 Feb 9;10:178-206.
- [3] Nakov P, Da San Martino G, Elsayed T, Barrón-Cedeno A, Míguez R, Shaar S, Alam F, Haouari F, Hasanain M, Babulkov N, Nikolov A. The CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43 2021 (pp. 639-649). Springer International Publishing.
- [4] Thorne J, Vlachos A, Cocarascu O, Christodoulopoulos C, Mittal A. The fact extraction and VERification (FEVER) shared task. arXiv preprint arXiv:1811.10971. 2018 Nov 27.
- [5] Thorne J, Vlachos A. Automated fact checking: Task formulations, methods and future directions. arXiv preprint arXiv:1806.07687. 2018 Jun 20.
- [6] Lazarski E, Al-Khassaweneh M, Howard C. Using nlp for fact checking: A survey. Designs. 2021 Jul 14;5(3):42.
- [7] Soleimani A, Monz C, Worring M. Bert for evidence retrieval and claim verification. In Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42 2020 (pp. 359-366). Springer International Publishing.
- [8] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of machine Learning research. 2003;3(Jan):993-1022.
- [9] Pang B, Lee L. Opinion mining and sentiment analysis. Foundations and Trends® in information retrieval. 2008 Jul 6;2(1–2):1-35.
- [10] Burnap P, Williams ML. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. Policy & Internet. 2015 Jun;7(2):223-42.
- [11] Camerlo R, Marcone A, Ros LM. Polish metric spaces with fixed distance set. Annals of Pure and Applied Logic. 2020 Dec 1;171(10):102832.
- [12] Brown TB. Language models are few-shot learners. arXiv preprint arXiv:2005.14165. 2020.
- [13] Dai Z, Yang Z, Yang Y, Carbonell J, Le QV, Salakhutdinov R. Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860. 2019 Jan 9.
- [14] Zellers R, Holtzman A, Rashkin H, Bisk Y, Farhadi A, Roesner F, Choi Y. Defending against neural fake news. Advances in neural information processing systems. 2019;32.
- [15] Abels, G. (2022, June 28). What is the future of automated fact-checking? Fact-checkers discuss. Poynter. <https://www.poynter.org/fact-checking/2022/how-will-automated-fact-checking-work/>
- [16] Zheng L, Zhang Y, Thing VL. A survey on image tampering and its detection in real-world photos. Journal of Visual Communication and Image Representation. 2019 Jan 1;58:380-99.

- [17] Verdoliva L. Media forensics and deepfakes: an overview. *IEEE journal of selected topics in signal processing*. 2020 Jun 12;14(5):910-32.
- [18] Cozzolino D, Pianese A, Nießner M, Verdoliva L. Audio-visual person-of-interest deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2023* (pp. 943-952).
- [19] Zhou P, Han X, Morariu VI, Davis LS. Learning rich features for image manipulation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition 2018* (pp. 1053-1061).
- [20] Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*. 2020 Dec 1;64:131-48.
- [21] Mirsky Y, Lee W. The creation and detection of deepfakes: A survey. *ACM computing surveys (CSUR)*. 2021 Jan 2;54(1):1-41.
- [22] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. *Advances in neural information processing systems*. 2014;27.
- [23] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*. 2020;33:6840-51.
- [24] Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*. 2021 Dec 17;65(1):99-106.
- [25] Li Y, Lyu S. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*. 2018 Nov 1.
- [26] Hu S, Li Y, Lyu S. Exposing GAN-generated faces using inconsistent corneal specular highlights. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2021 Jun 6* (pp. 2500-2504). IEEE.
- [27] Matern F, Riess C, Stamminger M. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW) 2019 Jan 7* (pp. 83-92). IEEE.
- [28] Li L, Bao J, Zhang T, Yang H, Chen D, Wen F, Guo B. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2020* (pp. 5001-5010).
- [29] Afchar D, Nozick V, Yamagishi J, Echizen I. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS) 2018 Dec 11* (pp. 1-7). IEEE.
- [30] Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision 2019* (pp. 1-11).
- [31] Nguyen HH, Yamagishi J, Echizen I. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP) 2019 May 12* (pp. 2307-2311). IEEE.

- [32] Zhao H, Zhou W, Chen D, Wei T, Zhang W, Yu N. Multi-attentional deepfake detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2021 (pp. 2185-2194).
- [33] Heo YJ, Choi YJ, Lee YW, Kim BG. Deepfake detection scheme based on vision transformer and distillation. arXiv preprint arXiv:2104.01353. 2021 Apr 3.
- [34] Wang SY, Wang O, Zhang R, Owens A, Efros AA. CNN-generated images are surprisingly easy to spot... for now. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2020 (pp. 8695-8704).
- [35] Haliassos A, Mira R, Petridis S, Pantic M. Leveraging real talking faces via self-supervision for robust forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022 (pp. 14950-14962).
- [36] Juefei-Xu F, Wang R, Huang Y, Guo Q, Ma L, Liu Y. Countering malicious deepfakes: Survey, battleground, and horizon. International journal of computer vision. 2022 Jul;130(7):1678-734.
- [37] Wang J, Li Z, Zhang C, Chen J, Wu Z, Davis LS, Jiang YG. Fighting malicious media data: A survey on tampering detection and deepfake detection. arXiv preprint arXiv:2212.05667. 2022 Dec 12.
- [38] Ivanovska M, Struc V. On the vulnerability of deepfake detectors to attacks generated by denoising diffusion models. In Proceedings of the IEEE/CVF winter conference on applications of computer vision 2024 (pp. 1051-1060).
- [39] Tolosana R, Rathgeb C, Vera-Rodriguez R, Busch C, Verdoliva L, Lyu S, Nguyen HH, Yamagishi J, Echizen I, Rot P, Grm K. Future trends in digital face manipulation and detection. In Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks 2022 Jan 31 (pp. 463-482). Cham: Springer International Publishing.
- [40] Papadopoulos SI, Koutlis C, Papadopoulos S, Petrantonakis PC. VERITE: a Robust benchmark for multimodal misinformation detection accounting for unimodal bias. International Journal of Multimedia Information Retrieval. 2024 Mar;13(1):4.
- [41] Boididou C, Andreadou K, Papadopoulos S, Dang Nguyen DT, Boato G, Riegler M, Kompatsiaris Y. Verifying multimedia use at mediaeval 2015. In MediaEval 2015 2015 (Vol. 1436). CEUR-WS.
- [42] Singhal S, Shah RR, Chakraborty T, Kumaraguru P, Satoh SI. Spotfake: A multimodal framework for fake news detection. In 2019 IEEE fifth international conference on multimedia big data (BigMM) 2019 Sep 11 (pp. 39-47). IEEE.
- [43] Khattar D, Goud JS, Gupta M, Varma V. Mvae: Multimodal variational autoencoder for fake news detection. In The world wide web conference 2019 May 13 (pp. 2915-2921).
- [44] Yu C, Ma Y, An L, Li G. BCMF: A bidirectional cross-modal fusion model for fake news detection. Information Processing & Management. 2022 Sep 1;59(5):103063.
- [45] Aneja, S., Bregler, C., & Nießner, M. (2023, June). COSMOS: catching out-of-context image misuse using self-supervised learning. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, No. 12, pp. 14084-14092).

[46] Luo, G., Darrell, T., & Rohrbach, A. (2021, November). NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 6801-6817).

[47] Zhang F, Liu J, Zhang Q, Sun E, Xie J, Zha ZJ. ECENet: Explainable and Context-Enhanced Network for Multi-modal Fact verification. In Proceedings of the 31st ACM International Conference on Multimedia 2023 Oct 26 (pp. 1231-1240).

[48] Abdelnabi S, Hasan R, Fritz M. Open-domain, content-based, multimodal fact-checking of out-of-context images via online resources. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2022 (pp. 14940-14949).

[49] Papadopoulos SI, Koutlis C, Papadopoulos S, Petrantonakis PC. RED-DOT: Multimodal Fact-checking via Relevant Evidence Detection. arXiv preprint arXiv:2311.09939. 2023 Nov 16.