ACIG
APPLIED CYBERSECURITY & INTERNET GOVERNANCE

NASK

AI in disinformation detection

Julia Puczyńska | IDEAS NCBR Sp. z o.o., 69 Chmielna Street, 00-801 Warsaw, Poland, and IPPT PAN, Pawińskiego St. 5B; 02-106 Warsaw, Poland | ORCID: 0009-0009-5304-7092

Youcef Djenouri | IDEAS NCBR Sp. z o.o., 69 Chmielna Street, 00-801 Warsaw, Poland, and University of South-Eastern Norway (USN), Post Office Box 4, 3199 Borre, Norway | ORCID: 0000-0003-0135-7450

Abstract

The Russian Doppelganger campaign was a flop. It tried targeting European governments and institutions with fake news and cloned websites, but their measurable impact on real users — views, likes, or shares — was near zero [1]. However, as a part of continuous efforts to influence Western media, this campaign does contribute to changing the online discourse and normalising hate speech. The potential for harm from such attacks has proven to be even more extreme. Such threats require international efforts to identify and counter such campaigns effectively.

In this article, we consider the use of artificial intelligence (AI) in disinformation detection. The recent explosion of AI performance and popularity is a double-edged sword. On the one hand, AI makes generating fake news faster. On the other hand, it helps fight back; in fact, nowadays leveraging AI-driven techniques – such as Natural Language Processing (NLP), multimedia analysis, and network analysis – is crucial in the fight against fake news.

Our discussion is based on the DISARM Framework, a disinformation-focused counterpart to the MITRE ATT&CK® framework, designed to standardise disinformation-related terminology and analytical methods [2]. We focus particularly on a key tactic of disinformation that relies on overwhelming the target, apparent in many social engineering plots. Be it news or messages, the 21st century is overfilled

Received: 17.11.2024

Accepted: 20.12.2024

Published: 30.12.2024

Cite this article as:
Julia Puczyńska, Youcef Djenouri, "AI in disinformation detection," ACIG, vol. 3, no. 2, 2024, pp. 211–232. DOI: 10.60097/ ACIG/200200

Corresponding author:
Julia Puczyńska, 69 Chmielna Street, 00-801 Warsaw, Poland, julia.puczynska@ideas-ncbr.pl
0009-0009-5304-7092

with content, forcing people into constant stress, weakening their decision-making, and increasing their susceptibility to manipulation. We discuss the practical overview of disinformation detection. In this discussion, we include uncertainty quantification (UQ) as a groundbreaking tool to counteract this challenge (a solution introduced by Julia Puczyńska, Youcef Djenouri, Tomasz Pawel Michalak and Piotr Sankowski in 'Knowledge Base Monte Carlo for Uncertainty Quantification in Fake News Detection', mimeo, IDEAS NCBR, 2024). UQ enhances reliability, explainability, and adaptability in disinformation detection systems, as it enables estimation of model confidence.

Our framework demonstrates the potential of AI-driven systems to counteract disinformation through multimodal analysis and cross-platform collaboration while maintaining transparency and ethical integrity. We underscore the urgency of integrating UQ into fake news detection methodologies to address the rapid evolution of disinformation campaigns. The paper concludes by outlining future directions for developing scalable, transparent, and resilient systems to safeguard information integrity and societal trust in an increasingly digital age.

Keywords

disinformation, fake news, artificial intelligence, uncertainty quantification, social media

1. Introduction

Disinformation became a very popular topic after the 2016 US presidential election and again after Russia's 2022 invasion of Ukraine, and now artificial intelligence (AI)-powered technologies are raising the stakes even further. They're powering sophisticated disinformation campaigns, through, for example, Natural Language Processing (NLP) and generative AI models that help spread falsehoods at lightning speed [3].

Ironically, these same technologies provide innovative solutions to identify, analyse, and counteract disinformation. However, there's a gap between the tools researchers write about in their papers and the ones that actually get used. People on the frontlines of combating disinformation often do not know whether these solutions exist, cannot apply them, or cannot afford to integrate them into their work. This is why we believe it is time to bridge this gap. In this paper, we dig into how AI both enables and combats disinformation. We are using the 'Doppelganger' campaign as a base for a

case study, an example for the challenge, and a reminder of what's at stake.

We emphasise the need for a comprehensive approach to disinformation detection that combines technological innovation and ethical responsibility. This should include accessible and explainable AI-powered, uncertainty quantification (UQ)-based, robust fact-checking systems. We argue that the same technological advances fueling disinformation can and must be harnessed to safeguard the truth and rebuild societal trust.

1.1. Doppelganger Campaign

'Olaf Scholz has betrayed the German economy' says a bold headline, 'European Union will manage without Poland' – says another headline on the Polish Radio's site. Or do they? The Doppelganger campaign got its name for impersonating trusted media sources and spreading such disinformation. It is attributed to Russian influence operations and has been actively spreading propaganda in the United States, Germany, and Ukraine [1]. As of today (December 2024), the researchers from Recorded Future's Insikt Group are tracking over 2000 fake social media (SM) accounts associated with this campaign, which relies on impersonating news outlets and creating fake websites to disseminate false narratives. Key tactics include undermining Ukraine's political stability, military strength, and international alliances; promoting narratives of Germany's domestic decline; and exploiting the US political and social divisions ahead of the 2024 election. Notably, some content is likely generated using AI, reflecting an evolving approach to bypass detection and establish long-term influence networks.

The campaign has been linked to Russian companies Structura National Technologies and Social Design Agency, both sanctioned by the European Union (EU) and the United States for their involvement. These operations highlight the Kremlin's strategic use of disinformation in its broader information warfare, leveraging AI tools to scale propaganda efforts.

The campaign's attack flow, as illustrated in Figure 2, is focused on a singular goal, which is spreading content. These undertaken steps made it very persistent, despite the continuous efforts to mitigate its spread. However, as mentioned above, the campaign's reach is negligible compared to the resources it requires. It would seem that the sole purpose of these actions is the content's generation and not its appeal or its reach.

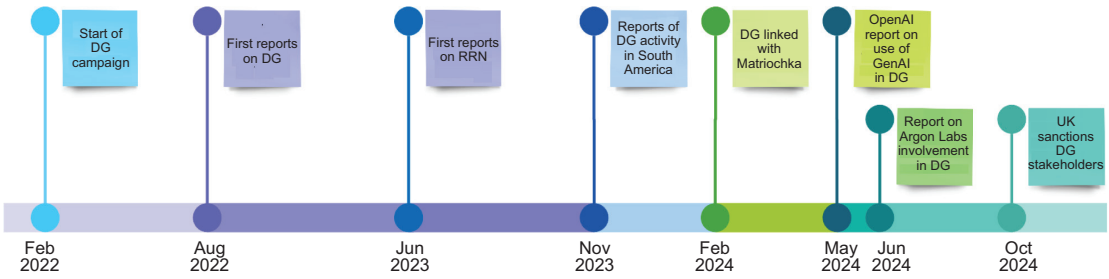

Figure 1. Timeline of reports regarding the Doppelganger (DG) campaign and the linked sub-campaigns.

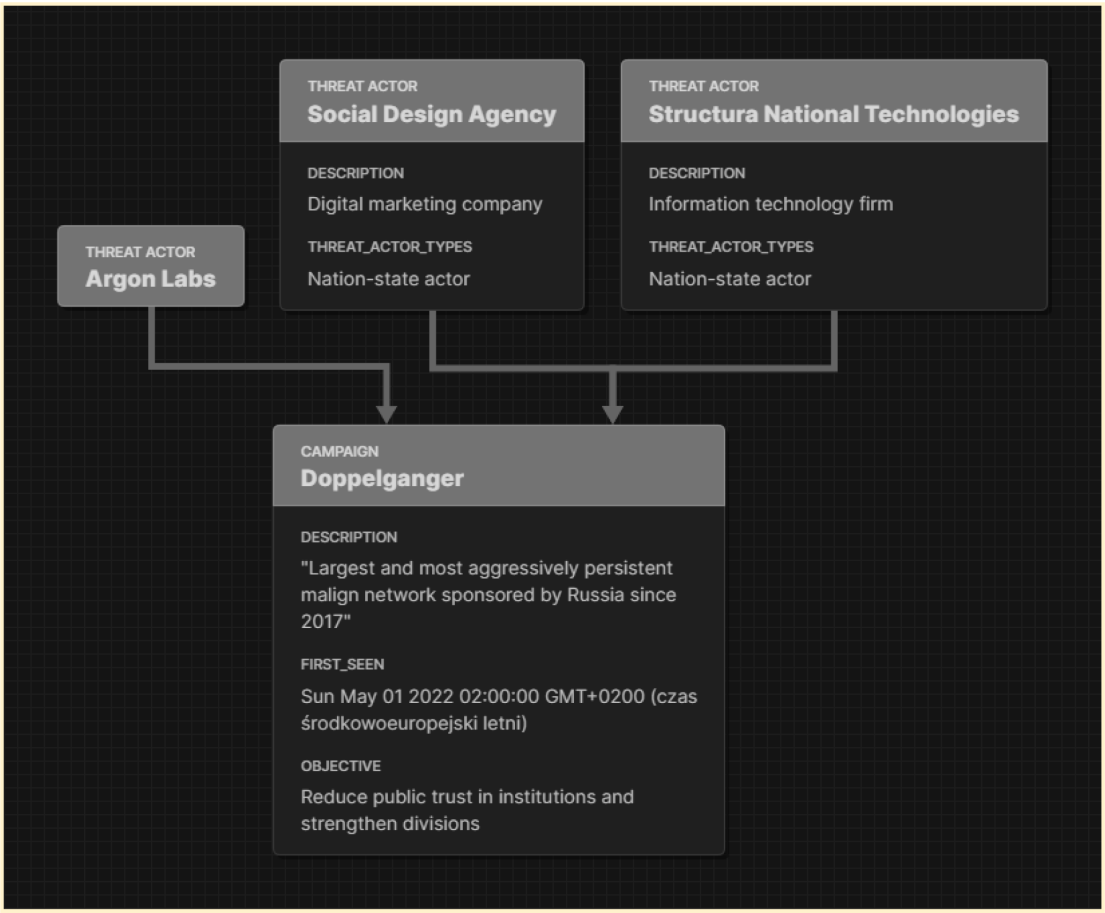

Figure 2. Doppelganger campaign-related threat actors.

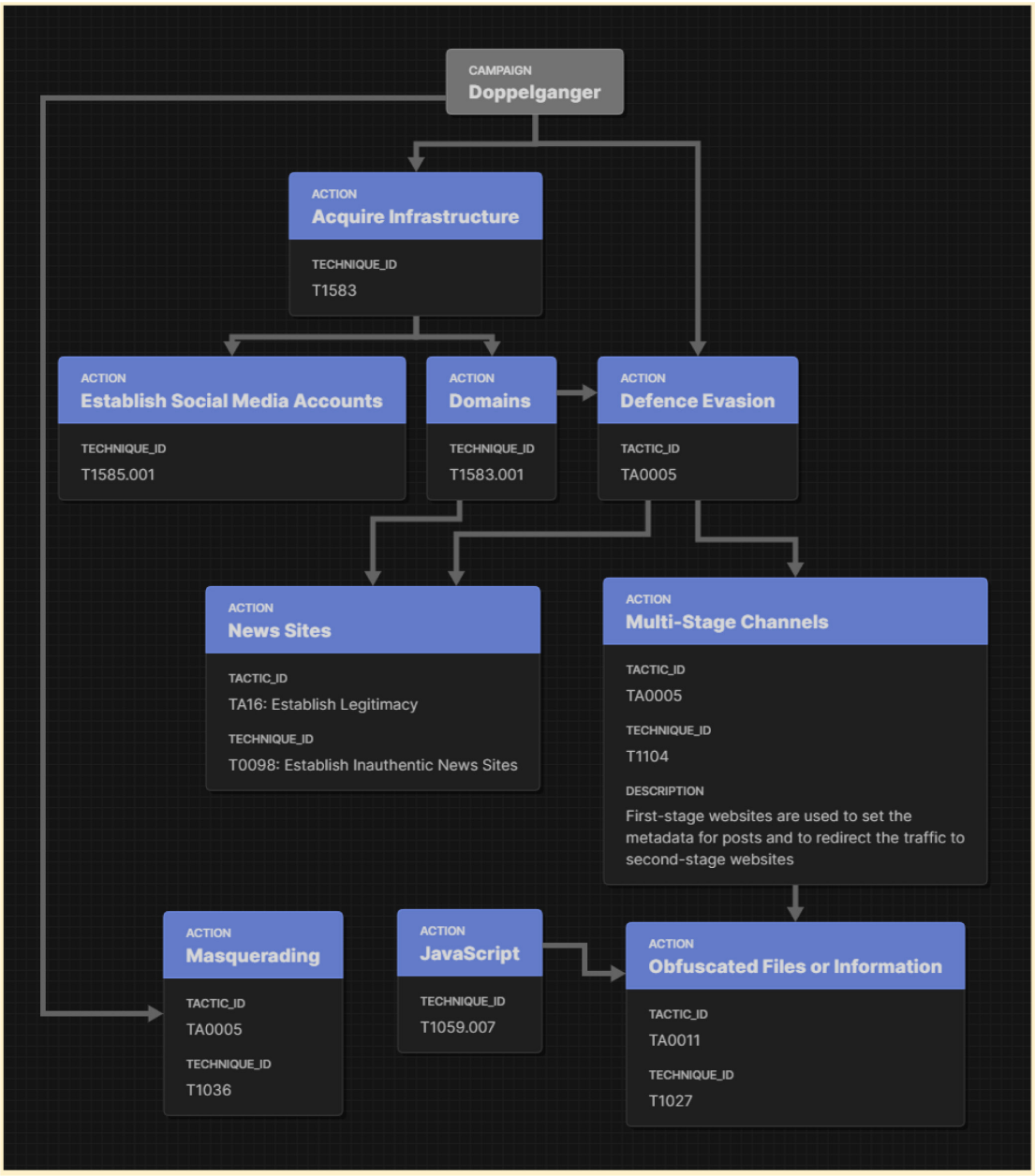

Figure 3. The campaign's attack flow – technical aspects.

Figure 4. The campaign's attack flow – narrative aspects. The campaign lacked measures to adjust and read just content to the target audiences. While it is difficult to assess the true extent of threat actors' efforts to analyse and construct their messages, the reality still is that they were not well fitted to their audience.

1.2. Contributions

The objective of this paper is to widen the perspective on disinformation. Our contributions are as follows:

- We apply the DISARM Framework to the Doppelganger campaign analysis. This helps with cementing the framework's role in disinformation-related research.
- We discuss a practical, AI-based approach to current problems in disinformation detection.
- We highlight the role of UQ in disinformation detection as a solution to the problem of strained content moderation and fact-checking apparatus.

2. Background

Disinformation refers to deliberately false or misleading information created and shared with the intent to deceive or manipulate public opinion [4]. Unlike misinformation, which involves spreading incorrect or misleading information without malicious intent, disinformation is intentionally crafted to cause harm, confusion, or disruption. Fake news, a term frequently used in the context of SM, is a specific type of disinformation. It involves fabricated stories or media designed to resemble legitimate news, intending to deceive readers [5]. While fake news is always based on a lie, it often serves as a vehicle for spreading either disinformation or misinformation. The primary distinction between these terms lies in the intent and factuality, with disinformation being intentional and fake news always rooted in fabrication.

Recognition of these differences is crucial for developing effective strategies to combat harmful content and introduce appropriate consequences for its spread. Specifically, in our understanding unaware users share disinformation without intent to deceive, that is not misinformation, because the content itself is being crafted and originally shared in order to manipulate the recipient. Therefore, while simply detecting harmful and misleading content usually does not include detection of intention (which is difficult to establish), we choose to keep this definition in order to retain the induced accountability for both its creation and spread.

2.1. Artificial Intelligence

Artificial Intelligence is a branch of computer science that aims to develop systems capable of performing tasks that typically require human intelligence, such as learning, reasoning,

problem-solving, and perception [6]. In the realm of disinformation, AI plays a dual role. On the one hand, it is used to create misleading content, such as deepfakes [7] and, on the other, it is utilised to combat disinformation through various detection and verification systems [8]. AI technologies, particularly NLP [9] and Large Language Models (LLMs) [10], are instrumental in both creation and identification of false narratives. NLP, a subfield of AI, focuses on enabling machines to understand, interpret, and generate human language. It plays a critical role in disinformation detection, as it can analyse patterns in online conversations, identify manipulated text, and track emerging trends. For instance, sentiment analysis techniques in NLP can identify manipulative language, often present in disinformation, by classifying text as having a positive, negative, or neutral sentiment [11].

The sentiment score can be as simple as a mean average of sentiment value associated with each word in a piece of text; that is, it can be calculated using a simple formula:

$$S = \frac{\sum_{i=1}^{N} score(w_i)}{N}, \tag{1}$$

where w_i represents individual words in the text, and score(w_i) is the sentiment score for each word, which is typically drawn from a predefined lexicon. The value of N is the number of words in the document.

2.2. Large Language Models

Large Language Models, such as Open AI's GPT models and Google's Bard, are trained on vast datasets to generate and understand text. These models contribute to the creation of sophisticated fake narratives by bots and are also used to counter disinformation by performing advanced text analysis, summarisation, and verification tasks. LLMs rely on neural networks that process vast amounts of textual data and learn the underlying patterns of language. For example, GPT models [12, 13] use transformer architecture, and the model's responses are based on both input and a set of parameters, which are defined during training. The transformer model uses self-attention mechanisms to weigh the importance of each word relative to others, enabling it to capture syntactic and semantic relationships in language and to generate a coherent text.

This process can be represented using the following transformation:

$$y = f(X, \theta), \tag{2}$$

where X is the input sequence (text data), θ represents the model's parameters, and y is the predicted output (e.g., the next word in the sequence).

2.3. Uncertainty Quantification

It is a mathematical framework designed to assess the uncertainty inherent in model predictions [14]. By identifying areas where predictions are uncertain, UQ provides confidence levels for specific outcomes, allowing for more informed decision-making. In the context of AI systems used for disinformation detection, UQ can help improve the robustness of models by quantifying their reliability. Simply put, where a model can classify content as disinformation or not, UQ returns the certainty of such classification, so how sure we are that this response is accurate.

Statistical inference based on a single data point, for example, an article, requires artificial multiplication of data. The article can then be assessed as false with a 70% confidence – because in 70% of these multiplied cases the article has been deemed false. One common approach in UQ is the use of Bayesian methods [15], which infer distributions over model parameters. This allows for a more probabilistic interpretation of model predictions, rather than providing deterministic outputs. For instance, if we have a model that predicts the likelihood $\hat{y}$ of a claim being false, the Bayesian approach provides a distribution over the prediction:

$$p(\hat{y} \mid x) = \int p(\hat{y} \mid \theta, X) p(\theta \mid X) d\theta, \tag{3}$$

where θ represents the model parameters, and $p(\theta \mid X)$ *represents the updated probability distribution of the model's parameters after*

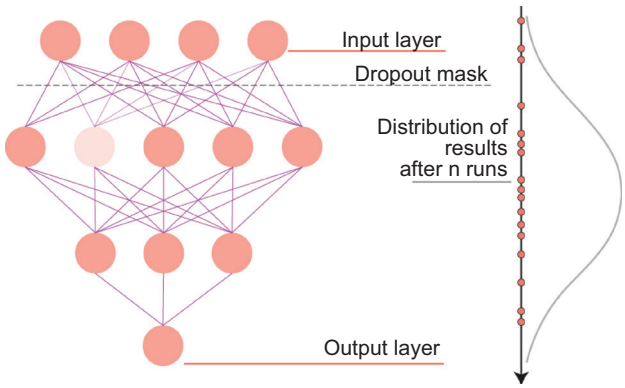

Figure 5. Monte Carlo Dropout.

incorporating the observed data X. This distribution represents the uncertainty in the model's predictions, allowing us to quantify how confident the model is about its conclusions.

A practical example of UQ in disinformation detection can be found in Monte Carlo Dropout [16], a method that estimates the uncertainty by applying dropout during inference. Dropout is a technique typically used during training to prevent over fitting, where certain neurons in the neural network are randomly 'dropped' or ignored during each forward pass. To quantify uncertainty, Monte Carlo Dropout keeps the dropout layers active during inference. The final prediction $\hat{y}$ is made by averaging multiple forward passes, each with different random neurons omitted, producing a distribution of predictions:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} f(X, \theta_i), \tag{4}$$

where θ_i are the parameters of the model after each forward pass with different dropout configurations, and N is the number of forward passes. The variance across these predictions provides an estimate of the model's uncertainty.

Another method used in UQ is Deep Ensembles [17], where multiple models are trained on the same dataset and their predictions are aggregated to estimate uncertainty. This approach captures the range of possible outcomes by training different models, each with slightly varied parameters, and combining their predictions. The uncertainty can then be calculated as the variance between the predictions of ensemble models:

$$U(\hat{y}) = \frac{1}{M}\sum_{j=1}^{M} \hat{y}_j - \hat{y}_{avg}, \tag{5}$$

where $\hat{y}_j$ is the prediction from the j-th model in the ensemble, $\hat{y}_{avg}$ is the average prediction across all models, and M is the number of models in the ensemble.

3. Practice

As mentioned, there is a disparity between the current state-of-the-art solutions in theoretical works and the solutions actually employed by some of those that take on the responsibility to fact check viral news. Therefore, we decided to describe in detail the reality of fact-checking.

3.1. Sources of Disinformation

Disinformation is similar to scams: both are based on influencing people and exploiting their vulnerability at the moment of contact with manipulative content. The endgoals may differ, but plenty of methods stay the same: overwhelming, inspiring fear, and impersonating trusted figures and sources. The spread of disinformation is incredibly complicated because it includes different social media platforms, TV, newspapers both online and offline, advertisements, and simply word-of-mouth [2]. Depending on the demographics, different sources matter more and less, but social media is increasingly significant.

One of the reasons for social media's popularity amongst security researchers is that regulations cannot keep up with the underlying technology. While the literature might already exist for plenty of possible threats related to social media, the public is often not informed and equipped well enough to recognise and appropriately react to them. It is worth noting that plenty of people base their knowledge about current events on social media. Polish IBIMS (Instytut Badań Internetu i Mediów Społecznościowych) and IBRIS report investigated the percentage of users who draw information from the Internet but differentiated between online news outlets (60% of respondents) and social media (38.8%) [18]. However, it is worth noting that users often access articles through links to news outlets on social media. Therefore, they are still subjected to, for example, biased selection of content by the social media algorithms. Interestingly, the Doppelganger campaign's fake news websites, which posed as trusted news outlets, could only be accessed through links in sponsored content posted on Facebook and X (formerly Twitter).

3.2. Fact-Checking

Efforts to combat disinformation involve a combination of strategies from governments, non-governmental organisations (NGOs), and social media companies. Each entity approaches the problem from different angles, leveraging its unique capabilities and areas of influence.

3.3. Government

Governments often focus on regulatory measures, public awareness campaigns, and collaboration with organisations to mitigate the spread of disinformation. For example, the EU's Digital Services Act (DSA) holds platforms accountable for harmful

content, including disinformation. Dedicated bodies to monitor and counteract disinformation include the US Cybersecurity and Infrastructure Security Agency (CISA, at https://www.cisa.gov/) or the European External Action Services East StratCom Task Force, which runs the EUvsDisinfo project – a database of articles and media considered to be disinformative (at https://euvsdisinfo.eu/). Such organisations create and implement their solutions, which have proven useful in France against the Doppelganger campaign. Their Service for Vigilance and Protection against Foreign Digital Interference (VIGINUM) agency, subject to Secretariat-General for National Defence and Security (fr. *Secrétariat général de la défense et de la sécurité nationale*, SGDSN), detected imitations of four French media outlets [19]. The organisation reported its findings on the RRN (*rrussianews*), an anonymous newsmedia organisation behind these fake websites. The RRN serves as a content repository for Doppelganger [1].

3.4. Non-Governmental Organisations

Non-governmental organisations focus on research, education, and advocacy to combat disinformation while supporting free speech and human rights. NGOs, like FactCheck.org, PolitiFact, and the International Fact-Checking Network (IFCN), identify and debunk disinformation and often partner with social media companies to label or flag false content (https://www.poynter.org/ifcn/) (Accessed: Nov. 18, 2024). This helps to create data sets for training disinformation detection systems [20]. Such organisations also develop programmes to improve critical thinking skills and media literacy among the public (https://geremek.pl/program/cyfrowa-akademia-walki-z-dezinformacja/) (Accessed: Nov. 18, 2024).

Input through NGOs is invaluable. Their impartial nature makes them the much-needed judges of the system's efficacy and equity. However, that also means they are highly dependent on donations, which often lead to underfunding and understaffing. Government and corporate funding helps solve this problem. In turn, it largely affects the impartiality of these organisations.

3.5. Social Media Companies

Social media platforms, as primary vectors for disinformation, focus on improving content moderation, transparency, and user awareness. Strategies include content moderation using AI and human moderators to detect and label or remove disinformation.

Platforms actively suspend fake or bot accounts spreading disinformation. In some cases, they remove coordinated inauthentic behaviour, as seen in campaigns linked to state-sponsored actors. Meta's Ad Transparency Tool is an example of providing access to information about political ads and their funding. It is worth noting that plenty of their efforts are legally required, for example, the Ad Transparency Tool or counteracting the spread of disinformation and reporting the results of their efforts. Without the regulation, these platforms wouldn't have the incentive to invest in countering the spread of disinformation; which became especially clear when Meta resigned from fact-checking programs in favor of an X-style 'community notes'.

4. Challenges for Disinformation Detection Framework

Fact-checking faces numerous challenges in the digital age. These obstacles can be broadly categorised into technical, operational, and societal domains, each presenting unique complexities that must be addressed for effective disinformation detection and mitigation.

4.1. Technical challenges

Volume and velocity: The digital ecosystem generates daily an overwhelming volume of content, ranging from social media posts and news articles to multimedia content. The rapid pace at which disinformation spreads often outpaces fact-checking efforts. Viral misinformation can reach millions within hours, while corrections, even when issued, struggle to achieve similar penetration. For example, during crises or high-profile events, false narratives dominate public discourse long before accurate information is disseminated. This imbalance underscores the need for scalable, automated tools capable of processing and verifying large quantities of data in real time.

Lack of datasets: Available datasets for disinformation detection include: FakeNewsNet, LIAR, ISOT FakeNews Dataset and WEIBO. However, more datasets are needed: firstly, there is a need for diverse datasets, including platform-specific and language-specific data. Nuances and contexts present in different cultures, platforms, and modalities are underrepresented. Existing datasets focus predominantly on text, leaving a gap in multimodal detection capabilities, and limiting the applicability and usefulness of fake news detection systems. Secondly, new topics and forms

of disinformation arise during global events (e.g., pandemics, elections, wars). Dynamic and up-to-date datasets are crucial to address evolving challenges. Since platforms like X, Instagram, and Facebook are not supported by any fact-checking programs, the access to data is limited even more.

Generative AI: Recent advancements in AI, particularly in generative technologies, have exacerbated the challenge. Tools like large language models and generative adversarial networks (GANs) are now capable of creating highly convincing fake content [21, 22, 23]. Deepfake videos can depict public figures engaging in fabricated acts, while AI-generated articles mimic credible news sources with alarming accuracy. The sophistication of such content makes it difficult for both humans and existing automated tools to discern authenticity, requiring the development of advanced detection algorithms tailored to generative outputs.

Multimodal disinformation: Disinformation campaigns increasingly utilise multimodal formats, blending text, images, and videos to enhance believability and engagement [24]. For instance, a false claim might be accompanied by a doctored image or a video with altered context, creating a layered narrative that appears credible. Detecting and analysing such multimodal disinformation demands cross-modal AI systems capable of correlating information across different formats – a complex and resource-intensive task.

4.2. Operational Challenges

Cross-platform propagation: Disinformation effortlessly migrates across platforms, exploiting the lack of coordinated detection mechanisms between social media, messaging apps, and traditional news outlets. A false narrative might originate on one platform, such as a tweet, and subsequently be amplified on others, including Facebook, Instagram, or WhatsApp. This fragmented ecosystem complicates detection efforts, as each platform employs varying policies, tools, and capabilities to address disinformation. Building interoperable solutions and fostering collaboration among platforms is critical but remains an unresolved challenge.

Language and cultural nuances: Disinformation often leverages specific linguistic and cultural contexts to increase its impact. A narrative tailored for one region may exploit local events, historical tensions, or societal biases, making it challenging to detect using generalised tools. Furthermore, many fact-checking systems and datasets are optimised for dominant languages like English, leaving significant

gaps in coverage for regional languages and dialects. Effective detection requires a nuanced understanding of cultural context and linguistic subtleties, necessitating localised datasets and AI models.

4.3. Societal Challenges

Polarisation and bias: In politically polarised environments, fact-checking is often perceived as an extension of one ideological viewpoint, undermining its credibility. Disinformation campaigns exploit these divisions, framing corrections as biased attempts to suppress dissenting opinions. This skepticism is further fueled by bad actors who discredit fact-checkers and promote narratives of censorship.

Overcoming this challenge requires transparent methodologies, diverse fact-checking teams, and the inclusion of multiple perspectives in verification processes to build public trust.

Trust deficits: A growing distrust in institutions, including media organisations and fact-checking bodies, significantly hampers efforts to combat disinformation. In many cases, people are more likely to trust information shared within their social or ideological circles than corrections issued by external entities. Addressing this trust deficit involves not only improving the accuracy and transparency of fact-checking efforts but also engaging communities directly to foster grassroots awareness and resilience against disinformation.

To overcome these challenges, a multi-pronged approach is required. Technical advancements must prioritise scalability and multimodal capabilities. Operational strategies should emphasise cross-platform collaboration and localised solutions. On the societal front, rebuilding trust through transparency, community engagement, and education is imperative. These measures, when integrated into a cohesive framework, can enhance the effectiveness of fact-checking efforts in the digital age.

5. Methodology and Tool Set

The dynamic and multifaceted nature of disinformation necessitates a diverse arsenal of tools that automate and enhance the detection process. These tools leverage cutting-edge AI, statistical techniques, and domain-specific expertise to identify, verify, and counter disinformation [25]. They can be broadly categorised into text analysis tools, multimedia analysis tools, and network analysis tools, each addressing specific challenges in the fact-checking landscape.

Text tools analysis: NLP techniques are pivotal in identifying and countering textual disinformation by analysing the tone, intent, and content of the text. Sentiment analysis helps flag emotionally charged or manipulative language often used in disinformation, such as fear-mongering or sensationalism designed to provoke rapid sharing without scrutiny. Entity recognition, another critical NLP capability, extracts and categorises names, organisations, and locations within a text, enabling cross-referencing with trusted databases to spot inconsistencies or fabrications. Claim matching, meanwhile, identifies recurring patterns or exact matches of previously debunked statements, aiding in the rapid recognition of recycled disinformation narratives. Beyond these, advanced language models like GPT, BERT, and T5 enhance the process by retrieving and cross-referencing documents from credible sources to verify claims [26]. Where simple sentiment analysis may fail, these models excel in understanding complex linguistic nuances, such as sarcasm or context-dependent meanings, which are often employed in sophisticated disinformation. Furthermore, integrating such models into automated fact-checking pipelines, supported by structured datasets, accelerates the generation of fact-checking reports for emerging claims, providing a scalable and efficient approach to combating textual disinformation [27].

Multimedia tools analysis: The rise of multimedia disinformation has necessitated the development of specialised tools for analysing and detecting manipulated visual content, from altered images to synthetic videos [28]. Image forensics plays a crucial role by examining metadata – such as timestamps, geolocation, and camera settings – to uncover inconsistencies indicative of tampering [29]. Algorithms also detect visual artifacts like irregular pixel patterns, lighting mismatches, or compression anomalies, which often result from editing processes. Additionally, reverse image search techniques allow cross-referencing of suspect visuals with existing databases to identify duplicates or modifications. Similarly, video analysis tools tackle the challenges posed by deepfakes and splicedfootage. Biometric inconsistencies, such as unnatural blinking or misaligned facial expressions, are flagged using deepfake detection algorithms, including those that analyse generative model fingerprints imperceptible to humans. Temporal analysis further aids detection by identifying irregularities in motion, lighting, or audio synchronisation, which often signal manipulation. Advanced scene reconstruction techniques complement these efforts by contextualising video content, enabling evaluators to determine whether depicted events genuinely align with the associated narrative. Together, these tools

form a comprehensive framework for combating multimedia-based disinformation.

Network tools analysis: Disinformation campaigns often leverage complex dissemination networks to amplify their reach and legitimacy, necessitating robust network analysis tools for effective detection and mitigation. Propagation mapping is a critical technique in this context, allowing researchers to track the evolution and spread of disinformation narratives across social platforms [30]. By identifying key actors, influential hashtags, and clusters responsible for amplifying false information, these tools enable targeted interventions. Algorithms that detect influential nodes within the network – individuals or groups with a disproportionate impact on disinformation dissemination – are particularly valuable in disrupting these campaigns. Temporal dynamic analysis further strengthens this approach by examining the timing and frequency of posts to identify patterns indicative of coordinated campaigns, such as those orchestrated by bot networks or state-sponsored entities.

Bot detection forms another essential component of network analysis, addressing the role of automated accounts in disseminating disinformation. The behavioural analysis identifies suspicious patterns, such as excessive posting frequency, identical content shared across multiple accounts, or activity during improbable hours, all of which suggest automation. Network-specific features, including low engagement rates, clustering within particular communities, or repeated interactions with known disinformation agents, further assist in distinguishing bots from human users. Machine learning models trained on diverse datasets enhance this process, classifying accounts based on multidimensional behavioural characteristics. Together, these tools provide a comprehensive approach to mapping, analysing, and ultimately disrupting the networks that propagate disinformation.

Other tools: We believe that there is a need for reliable and accessible fact-checking tools that can be used by both specialists and general public. These include web plug-ins, news apps, and dedicated SM profiles; all of these should focus on increasing the users' ability to determine what is trustworthy. The 'Ground News' app, which aims to provide informative news headlines and insight into the bias of reported news, serves as a great example of what is wildly needed. Today's users are overwhelmed with content. Anything that helps with limiting the quantity of content they receive, without

jeopardising its quantity and the users' choice over what they can get access to, is of value.

6. Vision for the Doppelganger

The Doppelganger campaign succeeds mainly in the sheer amount of content, created and/or generated by AI – what it lacks in likes, views, and shares, it makes up for in scale and persistence. While it may seem like a waste of resources, we think that the actual lesson that needs to be learned from it is that it would not take much for this campaign to be significantly more successful. Had these articles and their content caught on and spread among actual users, the mitigation couldn't have been limited to the continuous blocking of websites and fake accounts. Once real users would have been involved, their accounts would often not be blocked just because they shared disinformative content and the content itself might not be blocked nor marked as untrue. Such infrastructure as the one created for the sake of the Doppelganger campaign would keep providing new articles and links for these users, overwhelming even more our already strained system.

6.1. Uncertainty Quantification

Uncertainty quantification presents a transformative opportunity to enhance the robustness and reliability of fake news detection systems, particularly in the context of complex disinformation campaigns like Doppelganger. This campaign, which relied on cloned websites and targeted social media manipulation, demonstrates the challenges of distinguishing fabricated narratives from legitimate content. Traditional detection models often provide binary classifications, lacking the nuanced confidence metrics needed to guide critical decisions. Integrating UQ into these systems can address this limitation by estimating the reliability of predictions. For example, when analysing cloned content, UQ can pinpoint regions of high uncertainty, prompting additional human verification. Similarly, in cross-platform disinformation campaigns, where the context and format of narratives can vary, UQ can identify instances of low-confidence classifications. This capability ensures that questionable results are flagged for further scrutiny, reducing the risk of false positives or missed threats.

In addition to bolstering detection accuracy, UQ enhances the adaptability and transparency of fake news detection frameworks. Disinformation campaigns like Doppelganger evolve rapidly, with adversaries employing novel tactics to evade detection. UQ enables

systems to dynamically recalibrate their predictions by identifying areas where the model lacks sufficient training data or encounters new patterns. This adaptive capability ensures resilience against the evolving tactics of disinformation actors. Furthermore, the integration of UQ fosters greater transparency, particularly in politically sensitive contexts. By providing explanations alongside confidence metrics, UQ empowers stakeholders – such as fact-checkers, policymakers, and the public – to better understand and trust the decisions made by AI-driven detection systems. This transparency is critical in countering skepticism and ensuring that automated systems are perceived as reliable partners in combating disinformation.

Looking forward, UQ can play a pivotal role in improving the efficiency of resource allocation and the overall scalability of fake news detection efforts. Disinformation campaigns operate on a massive scale, often overwhelming human fact-checkers and investigative teams. UQ facilitates the prioritisation of high-risk cases by flagging predictions with elevated uncertainty for manual review. This targeted approach allows human resources to focus on the most critical and ambiguous cases, improving the efficiency of detection efforts. Furthermore, UQ strengthens defences against adversarial tactics, such as subtle content modifications that seek to exploit detection system vulnerabilities. By identifying instances of high uncertainty – often indicative of adversarial interference – UQ provides an early warning system for emerging threats. Finally, as cross-platform disinformation becomes more prevalent, the standardisation of UQ protocols enables seamless collaboration between platforms, fostering trust towards automated fact-checking and enabling coordinated responses to campaigns like Doppelganger. Together, these advancements position UQ as a cornerstone of future efforts to safeguard information integrity and societal trust.

7. Conclusions

In conclusion, the fight against disinformation, exemplified by campaigns, like Doppelganger, presents a growing challenge in the digital age. The dual role of AI in enabling and mitigating disinformation underscores the complexity of addressing this issue effectively. This paper has outlined a comprehensive framework for disinformation detection, emphasising the importance of integrating advanced AI techniques, such as NLP, multimedia analysis, and network analysis, into the detection process. Moreover, it has discussed UQ as a critical innovation, offering enhanced reliability

and interpretability for AI-driven detection systems. UQ not only improves confidence in predictions but also provides valuable insights that can guide human intervention, prioritise resources, and ensure the system remains adaptable to emerging disinformation tactics. As disinformation campaigns continue evolving in sophistication and scale, the need for adaptive, transparent, and collaborative detection mechanisms becomes increasingly urgent. This framework offers a promising direction for developing systems that can not only identify false narratives across multiple platforms but also respond to them in a way that is both efficient and ethically responsible. Moving forward, future research should focus on refining UQ techniques, improving cross-platform collaboration, and developing scalable solutions that can handle the ever-increasing volume and velocity of disinformation. By harnessing the full potential of AI and UQ, we can build a more resilient and trustworthy information ecosystem, safeguarding truth and societal trust in an increasingly complex digital world.

8. Acknowledgements

The authors used LLMs for editing and polishing the author-written version of the text, replacing about 15% of the original text.

References

[1] Insikt Group (2023). Obfuscation and AI content in the Russian influence network "doppelgänger" signals evolving tactics [Online]. Available: https://www.recordedfuture.com/research/russian-influence-network-doppelgangers-ai-content-tactics [Accessed: Nov. 18, 2024].

[2] DISARM Foundation (2022). DISARM disinformation TTP (tactics, techniques and procedures) framework [Online]. Available: https://github.com/DISARMFoundation/DISARMframeworks [Accessed: Nov. 18, 2024].

[3] J. Puczyńska, M. Podhajski, K. Wojtasik T. P. Michalak. *Large language models in jihadist terrorism and crimes*. Warsaw: Agencja Bezpieczeństwa Wewnętrznego (Internal Security Agency), 2024, 351 p.

[4] S. Abdali, S. Shaham, B. Krishnamachari. *Multi-modal misinformation detection: Approaches, challenges and opportunities*. New York, NY: ACM Computing Surveys, 2022.

[5] J. Alghamdi, S. Luo, Y. Lin, "A comprehensive survey on machine learning approaches for fake news detection," *Multimedia Tools and Applications*, vol. 83, no. 17, pp. 51009–51067, 2024, doi: 10.1007/s11042-023-17470-8.

[6] Y. Xu, X. Liu, X. Cao, C. Huang, E. Liu, S. Qian, ... J. Zhang, "Artificial intelligence: A powerful paradigm for scientific research," *The Innovation*, vol. 2, no. 4, p. 100179, 2021, doi: 10.1016/j.xinn.2021.100179.

[7] Á. F. Gambín, A. Yazidi, A. Vasilakos, H. Haugerud, Y. Djenouri, "Deepfakes: Current and future trends," *Artificial Intelligence Review*, vol. 57, no. 3, p. 64, 2024, doi: 10.1007/s10462-023-10679-x.

[8] V. U. Gongane, M. V. Munot, A. D. Anuse, "A survey of explainable ai techniques for detection of fake news and hate speech on social media platforms," *Journal of Computational Social Science*, vol. 7, pp. 1–37, 2024, doi: 10.1007/s42001-024-00248-9.

[9] G. G. Devarajan, S. M. Nagarajan, S. I. Amanullah, S. S. A. Mary, A. K. Bashir, "AI-assisted deep NLP-based approach for prediction of fake news from social media users," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 4, pp. 4975–4985, Aug. 2024, doi: 10.1109/TCSS.2023.3259480.

[10] B. Hu, Q. Sheng, J. Cao, Y. Shi, Y. Li, D. Wang, P. Qi, "Bad actor, good advisor: Exploring the role of large language models in fake news detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 20, pp. 22105–22113, 2024, doi: 10.48550/arXiv.2309.12247.

[11] J. Li, L. Xiao, "Multi-emotion recognition using multi-EmoBERT and emotion analysis in fake news," in *Proceedings of the 15th ACM web science conference 2023* (WebSci '23). Association for Computing Machinery, New York, NY, USA, 128–135. doi: 10.1145/3578503.3583595.

[12] V. Alto, *Modern generative AI with ChatGPT and OpenAI models: Leverage the capabilities of OpenAI's LLM for productivity and innovation with GPT3 and GPT4*. Birmingham: Packt Publishing, 2023.

[13] K. S. Kalyan, "A survey of GPT-3 family large language models including ChatGPT and GPT-4," *Natural Language Processing Journal*, vol. 6, p. 100048, 2023. doi: 10.1016/j.nlp.2023.100048.

[14] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, S. Nahavandi, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243–297, 2021.

[15] G. Franchi, A. Bursuc, E. Aldea, S. Dubuisson, I. Bloch, "Encoding the latent posterior of Bayesian neural networks for uncertainty quantification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 4, pp. 2027–2040, April 2024, doi: 10.1109/TPAMI.2023.3328829.

[16] D. Bethell, S. Gerasimou, R. Calinescu, "Robust uncertainty quantification using conformalised Monte Carlo prediction," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 19, pp. 20939–20948, 2024. doi: 10.1609/aaai.v38i19.30084.

[17] R. Rahaman, "Uncertainty quantification and deep ensembles," *Advances in Neural Information Processing Systems*, vol. 34, pp. 20063–20075, 2021.

[18] IBIMS & IBRIS (2022). Skąd polacy czerpią informacje? [Online]. Available: https://ibims.pl/wp-content/uploads/2021/01/Raport-IBIMS-IBRIS-Zrodla-informacji-Polakow.pdf [Accessed: Nov. 18, 2024].

[19] Service for Vigilance and Protection against Foreign Digital Interference (VIGINUM) (2023). RRN: A complex and persistent information manipulation campaign [Online]. Available: https://www.sgdsn.gouv.fr/files/files/Publications/20230719_NP_VIGINUM_RAPPORT-CAMPAGNE-RRN_EN.pdf [Accessed: Nov. 18, 2024].

[20] K. Shu, G. Zheng, Y. Li, S. Mukherjee, A. H. Awadallah, S. Ruston, H. Liu, "Leveraging multi-source weak social supervision for early detection of fake news," arXiv preprint arXiv:2004.01732, 2020, doi: 10.48550/arXiv.2004.01732.

[21] R. Raman, V. K. Nair, P. Nedungadi, A. K. Sahu, R. Kowalski, S. Ramanathan, K. Achuthan, "Fake news research trends, linkages to generative artificial intelligence and sustainable development goals," *Heliyon*, vol. 10, no. 3, p. e24727, 2024, doi: 10.1016/j.heliyon.2024.e24727.

[22] A. Bashardoust, S. Feuerriegel, Y. R. Shrestha, "Comparing the willingness to share for human-generated vs. AI-generated fake news," *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. CSCW2, pp. 1–21, 2024, doi: 10.1145/3687028.

[23] D. Xu, S. Fan, M. Kankanhalli, "Combating misinformation in the era of generative AI models," in *Proceedings of the 31st ACM international conference on multimedia*, Association for Computing Machinery, New York, NY, USA, 9291–9298, 2023, doi: 10.1145/3581783.3612704.

[24] S. Tufchi, A. Yadav, T. Ahmed, "A comprehensive survey of multimodal fake news detection techniques: Advances, challenges, and opportunities," *International Journal of Multimedia Information Retrieval*, vol. 12, no. 2, p. 28, 2023, doi: 10.1007/s13735-023-00296-3.

[25] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, F. Menczer, "Detecting and tracking political abuse in social media," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 5, no. 1, pp. 297–304, 2011, doi: 10.18653/v1/N19-1423.

[26] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the Association for computational linguistics: Human language technologies*, vol. 1, p. 2, pp. 4171–4186. Association for Computational Linguistics, Kerrville, TX 78028, USA, 2019, doi: 10.18653/v1/N19-1423.

[27] S. Bansal, N. S. Singh, S. S. Dar, N. Kumar, "MMCFND: Multimodal multilingual caption-aware fake news detection for low-resource indic languages," arXiv preprint, arXiv:2410.10407, 2024, doi: 10.48550/arXiv.2410.10407.

[28] F. Marra, D. Gragnaniello, D. Cozzolino, L. Verdoliva, "Detection of gan-generated fake images over social networks," in Proceedings of the *2018 IEEE conference on multimedia information processing and retrieval* (MIPR) pp. 384–389, 2018, New York, NY: IEEE. doi: 10.1109/MIPR.2018.00084.

[29] L. Verdoliva, "Media forensics and deepfakes: An overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020, doi: 10.1109/JSTSP.2020.3002101.

[30] L. Vargas, P. Emami, P. Traynor, "On the detection of disinformation campaign activity with network analysis," in *Proceedings of the 2020 ACM SIGSAC conference on cloud computing security workshop*, pp. 133–146, ACM. New York, NY, 2020. doi: 10.1145/3411495.3421360.