Free Speech and False news - Disinformation Detection

2024

Authors: *Dimitris Stathopoulos, f3352318*

Dionysios Voulgarakis, f3353307

Introduction

With the rapid rise of digital technologies, the spread of disinformation has become a significant threat to democratic processes and societies. Disinformation (MEDIA DEFENCE)[1] is information that is false, and the person who is disseminating it knows it is false. “It is a deliberate, intentional lie, and points to people being actively disinformed by malicious actors”. Misinformation is information that is false, but the person who is disseminating it believes that it is true. Mal-information is information that is based on reality but is used to inflict harm on a person, orgasnisation or country. We will refer to all of the as disinformation. We will refer to all of them as disinformation in this article.

This article examines the role of Artificial Intelligence (AI) in combating disinformation online. While AI technologies offer promising approaches to detect and address disinformation, they also raise concerns about their potential impact on freedom of expression and media pluralism. A comprehensive understanding of these implications is crucial for policymakers and stakeholders seeking to develop effective strategies to mitigate the spread of disinformation while safeguarding fundamental rights.

I. The Role of Technology in the Spread of Fake News and Disinformation

Social media has changed how we share and find information, offering new ways to communicate and express ourselves.With an estimated user base of 4.59 billion as of 2022 and projected to exceed 5.5 billion by 2027, its impact is substantial (STATISTA)[2].The problem with that arises with the fact that it has become a hub for spreading fake news and disinformation. The quick spread of disinformation on social media can lead to misleading content without proper fact-checking. Based on a study (SCIENCE)[3] “tweets containing falsehoods reach people on Twitter six times faster than truthful tweets”. Algorithms that boost user engagement may unintentionally promote false narratives,

contributing to the spread of deceitful information. In addition, the anonymity on social media allows for fake accounts and manipulation of public discussions, making it hard to distinguish between reliable sources and malicious users.

The spread of fake news and disinformation on the internet has been made even more widespread and advanced with the introduction of language models (LLMs) and deepfake technology. LLMs like OpenAI's GPT series can produce text that closely resembles human writing, making it easy to create large quantities of content quickly. When combined with deepfake technology, which can create realistic audio and video of people saying or doing things they never did, these tools become a powerful force for spreading false information.

“A growing number of websites, with generic names such as iBusiness Day or Ireland Top News, are delivering fake news made to look genuine, in dozens of languages from Arabic to Thai.” (Pranshu Verma, 2023)[4]. Large language models (LLMs) are often used to create content such as articles, social media posts, and comments that sound like they were written by a human. This makes it harder to tell what information is real and what is fake. These computer-generated texts can be customized to support certain storylines or take advantage of biases, blurring the line between truth and lies. Additionally, LLMs can create a lot of content very quickly, spreading false information even further. Adding to that the Hallucinations of LLMs, disinformation can spread even without a bad motive. Hallucinations “are the events in which ML models, particularly large language models (LLMs) like GPT-3 or GPT-4, produce outputs that are coherent and grammatically correct but factually incorrect or nonsensical” (Iguazio.com)[5]

Deepfake technology exacerbates the issue by enabling malicious individuals to produce extremely lifelike audio and video content that can be utilized to manipulate public opinion. This can lead to politicians, celebrities, and other prominent figures being depicted saying or doing things that never actually happened, resulting in widespread confusion and mistrust. The merging of LLM-generated text with deepfake audio/video forms a potent weapon for spreading false information, as these altered materials can be distributed on numerous online channels to specifically target certain groups and deepen societal divisions. As an example with have the case where “A finance worker at a multinational firm was tricked into paying out $25 million to fraudsters using deepfake technology to pose as the company’s chief financial officer in a video conference call, according to Hong Kong police.” (Kathleen Magramo, 2024)[6]

With the increasing use of automated bots and algorithms, LLMs and deepfakes are amplifying the dissemination of false information by introducing a higher level of complexity and convincing realism to the misleading content landscape. Bots utilize text generated by LLMs to inundate social media platforms with deceptive material, and deepfake technology allows the production of compelling audio and video content to bolster these fictitious

storylines. These tools collectively boost the reach and believability of fake news, thereby complicating the task of distinguishing truth from falsehood for users.

II. Threats to Free Speech Posed by Disinformation

In today's digital world, disinformation is a major challenge to free speech. Although free speech is crucial for democracy, the spread of false information can harm its core values. Disinformation, which includes false or inaccurate information, can cause confusion, distrust, and negative outcomes. Social media and online information sharing have made it easier for disinformation to spread rapidly and widely, creating a conducive environment for falsehoods to thrive.

Disinformation is a serious issue as it has the power to influence public discussion and viewpoints. Fake news can warp the truth, molding how people see things in a way that may not be accurate. This distortion undermines the exchange of ideas, a crucial part of free speech. If disinformation spreads without consequences, it can silence valid perspectives and hinder important discussions, which in turn weakens the basis of democracy.

Furthermore, spreading false information can be harmful and cause divisions in society. Whether it's spreading lies about health, politics, or social issues, disinformation can have serious effects. For example, during the COVID-19 pandemic, false information about the virus and treatments caused confusion, and risky behaviors that put public health at risk. Similarly, spreading disinformation to fuel societal tensions or incite violence can create serious problems for communities, making existing divides worse and escalating conflicts.

Additionally, the spread of false information damages the trust in reliable sources of information. With the rapid dissemination of disinformation, distinguishing between truth and lies becomes more challenging for individuals. This loss of faith in trustworthy sources can greatly impact democracy and the ability to make well-informed decisions. Without trusting the media, scientific organizations, or other credible sources, it becomes difficult to ensure accountability and make informed decisions as members of society. An example is the pizzagate conspiracy theory in which “A man fired a rifle on Sunday inside a Washington pizza restaurant that has been subjected to harassment based on false stories tying it to child abuse, the police said. No one was hurt, and the man was arrested.” (Eric Lipton, 2016)[7]

III. Approaches to Detecting Fake News and Disinformation

The fight against fake news and disinformation has led to the development of various strategies. As states in this study (Ryan Kraski, 2018)[8] “the Würzburg District Court made clear that social media networks, such as Facebook, are typically neither content creators nor collaborators in specific acts of defamation. Thus, functioning only as service providers, in the context of the German Teleservices Act, service providers cannot be obligated to proactively search through all of the content published on their platforms and to delete content that is defamatory, or otherwise affecting an individual’s rights”. Fact-checking efforts are key in addressing disinformation, with teams of journalists and researchers working diligently to verify facts and debunk false claims. By ensuring the public has access to accurate information, fact-checkers play a crucial role in preserving truth in a time filled with deceptive narratives. X (former twitter) which is amongst the platforms with the highest rates of fake news in 2017 (STATISTA)[9], implemented a tool, called community notes, which is a Twitter program that allows certain users to submit helpful context to tweets that might be misleading or missing important information. According to X, people are on average 30% less susceptible to agree to the contents of a post after having read accompanying community posts, and they are also less likely to repost it to their followers.(X (former twitter))[10]. But based on Alex this community notes fail to combat disinformation efficiently. (Madison Czopek, 2023)[11]

In addition, and contrary to the previous discussion about the role of AI in spreading disinformation, AI and machine learning have, also, transformed the battle against disinformation. AI technology allows for the quick and accurate analysis of large data sets, helping to identify patterns and anomalies in content that may be misleading. Machine learning algorithms can also improve their ability to detect false information by constantly learning from new data, making them better at distinguishing between trustworthy and deceptive content. Dr Demartini claims that “Our technology not only identifies fake news, but also explains and substantiates why that is the case.(University of Queensland)[12]

Yet, the best way to tackle fake news may be through partnerships between technology firms, governments, and civil society. With a deep understanding of the complex issue at hand, stakeholders from different fields are coming together to put in place effective plans. Tech companies are key players in this collaboration as they create and enforce strong content moderation rules, increase transparency in algorithms, and promote digital literacy among their users. Governments, too, must step up by passing laws and regulations that can hold those spreading disinformation responsible while still protecting freedom of speech. Also, journalist Peter Pomerantsev states that the most effective way to combat misleading information in media is through political and scientific transparency from governments and companies. Through this, he suggests, we can start to build back the trust that has been eroding for decades and combat the spread of misleading information around Covid-19, as fake news is often an outcome of ignorance. (Londons Kings College, 2022)[13]

IV. Balancing Free Speech and Disinformation Regulation

Exploring the ethical aspects of content moderation poses a complex challenge in today's digital era. The main aim is to protect people from the negative impacts of disinformation, but doing so without limiting freedom of speech inadvertently requires a careful equilibrium. Ethical guidelines for content moderation need to be clear, flexible, and responsible to guarantee unbiased and fair actions (Chris MARSDEN et al., 2019)[14]. This includes not only detecting and deleting harmful content but also considering the wider societal consequences of censorship. Platforms must weigh the potential damage of hate speech or false facts against the significance of amplifying marginalized voices (Richard BARRET, Herdis KJERULF THO., 2019)[15]. Moreover, there's a growing recognition of the need to incorporate diverse perspectives and community feedback into content moderation decisions to mitigate biases and ensure inclusivity as implied on (Chris MARSDEN et al., 2019)[14].

Legal frameworks and regulations play a pivotal role in shaping the boundaries of free speech and combating disinformation in the digital era. Governments around the world have enacted various laws and regulations aimed at addressing the spread of fake news and disinformation online as state in (Richard BARRET, Herdis KJERULF THO., 2019)[15]. For example, countries like Germany have implemented legislation to address the spread of fake news, although such measures have sparked debates over their potential impact on freedom of expression (Ryan Kraski, 2017)[16]. Finding a balance between upholding the law and safeguarding individual freedoms is crucial. This involves weighing the importance of basic rights like free speech and press freedom, alongside recognizing the negative impact of inaccurate information on society.

Finding the right balance between censoring harmful content and protecting free speech is a challenging task that goes beyond just relying on technology. Content moderation algorithms have difficulty in accurately identifying real news from fake news, leading to the accidental removal of valid content or the oversight of harmful content. Such errors are technically referred to as Type I-II errors as explained in (Chris MARSDEN et al., 2019)[14]. Furthermore, the widespread reach of the internet makes it challenging to regulate effectively, with rules implemented in one place potentially causing unexpected issues in others. Moreover, there is a concern about platforms or governments going too far in their attempts to address disinformation by utilizing bad practices, which could result in censorship and the silencing of opposing viewpoints as (Mariya Gabriel, Madeleine de Cock Buning 2018)[17] and (Richard BARRET, Herdis KJERULF THO., 2019)[15] explain such utilization. To tackle these issues, we need to take a careful approach that focuses on being transparent, holding ourselves accountable, and working closely with all parties involved. This means working with knowledgeable individuals, non-governmental organizations, and communities directly impacted to create solutions tailored to the circumstances that respect both the right to free speech and the common good.

V. Case Studies and Examples

The impact of false information on public perception has been shown in many well-known cases, highlighting the pressing importance of implementing successful solutions. One of the most notable examples of the impact of false narratives and disinformation campaigns proliferating on social media platforms is evident in the 2016 U.S. presidential election. Research has shown that fake news spread within platforms, with a significant role played by automated accounts (Nir Grinberg et al., 2019)[18]. Surveys and web browsing data among U.S. voters indicate that a considerable portion of the population encountered and remembered fake news stories about the election, with visits to fake news sources more prevalent among certain demographic groups, particularly conservatives (Nir Grinberg et al., 2019)[18]. This highlights the pervasive influence of false information on voter perceptions and the exacerbation of political polarization (Nir Grinberg et al., 2019)[18]. The discussion about fake news has led to actions by both citizens and government to address disinformation online. Fact-checking groups work to verify information and claims on social media around the world. Certain nations have put in place rules, and Facebook has adjusted its data access policies due to worries about its impact on election results (Thorsten Quandt et al.)[19]. The extensive sharing of fake news online shows how damaging disinformation can be to society and emphasizes the need for strong measures to tackle the issue.

In recent years, collaboration across different sectors has led to substantial progress in tackling disinformation. A standout example of this is the International Fact-Checking Network (IFCN), which has played a key role in setting high standards for ethical fact-checking methods as more carefully explained here (Mariya Gabriel, Madeleine de Cock Buning 2018)[17]. Similarly, the High Level Expert Group on Fake News and Online Disinformation (HLEG) from the European Commission has put forth several suggestions to improve transparency, promote media literacy, and strengthen backing for fact-checking efforts (Chris MARSDEN et al., 2019)[14].The various methods mentioned highlight the significance of working together to fight against the spread of false or misleading information on online platforms. Furthermore, there has been a rise in new initiatives aimed at debunking false statements and confirming the truthfulness of online content in the battle against disinformation(Mariya Gabriel, Madeleine de Cock Buning 2018)[17]. Through the expertise of both humans and technology, these efforts are essential for protecting the accuracy of information in today's digital world.

Key takeaways from the fight against disinformation highlight the significance of collaboration, transparency, and media literacy. Successful strategies often involve forming partnerships with various stakeholders, such as those endorsed by the International Fact-Checking Network (IFCN), as supported by organizations like the ACLU Foundation of

Northern California and the Electronic Frontier Foundation. Additionally, the European Commission's recommendations for improving transparency and media literacy, in alignment with suggestions from the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, underscore the importance of these efforts. By leveraging partnerships and promoting transparency, organizations can effectively counter the spread of false information online.

VI. The Future of Free Speech in the Age of Disinformation

Emerging technologies such as artificial intelligence (AI) and deep learning hold both promise and peril in the fight against disinformation as it has implied by the previous. AI-powered algorithms can help quickly identify and remove fake news, but they can also lead to worries about bias and unintended results. The growth of deepfake technology is a major risk to the honesty of visual media, making it harder to tell what's true and what's not. It's important to take action to lessen the negative effects of these technologies while using their potential for good.

The battle against disinformation is far from over, with new challenges and evolving threats constantly emerging in the digital information landscape. Bad actors continue to exploit vulnerabilities in online platforms and exploit societal divisions for their gain, posing significant challenges to efforts aimed at combating disinformation. Moreover, the rise of decentralized communication channels such as encrypted messaging apps presents new challenges for content moderation and detection efforts, as traditional approaches to identifying and mitigating the spread of fake news may prove less effective in these contexts. To tackle these challenges, we need a united effort from governments, tech companies, NGOs, and universities. It's crucial to facilitate global collaboration and sharing of information to combat the widespread distribution of false information and encourage a stronger and more knowledgeable public conversation.

Preserving free speech while combating disinformation requires a delicate balancing act that prioritizes transparency, accountability, and inclusivity (Richard BARRET, Herdis KJERULF THO., 2019)[15]. By encouraging open discussions and respectful arguments, we can create a culture that values critical thinking and the ability to withstand disinformation. Teaching people to properly evaluate media and be responsible online is crucial in helping each individual differentiate between what's true and what's not in today's digital world. Improving transparency and accountability on the internet can also play a key role in restoring trust in the information we receive and stopping the spread of false facts. In addition, policymakers need to use proven methods when creating rules that find the correct mix of safeguarding essential rights and lessening the negative impacts of false information. By studying past incidents and advocating for united solutions, communities can more effectively handle the problems brought about by disinformation in the online era and encourage a more educated and strong public conversation.

COCLUSION

Around the world, governments are scrambling to tackle the spread of disinformation. But as they do, many people worry about what this means for our freedom to speak our minds. Finding the right balance is tough. It calls for us to be open about what we're doing and to work together. Groups like the International Fact-Checking Network show how important it is for us to team up. They bring people together to check the facts and make sure we're getting the truth. And as we explore new technologies, like AI and social media, we need to be careful. They can do a lot of good, but they also come with risks. To really make a difference, we've got to work together across the globe. It's going to take a mix of different approaches to protect both our right to speak and the well-being of our societies.

[1] Misinformation, Disinformation and Mal-Information, **MEDIA DEFENCE**
https://www.mediadefence.org/ereader/publications/introductory-modules-on-digital-rights

-and-freedom-of-expression-online/module-8-false-news-misinformation-and-propaganda/misinformation-disinformation-and-mal-information/ (last visited Apr. 5, 2024).

[2] Number of Social Media Users Worldwide from 2010 to 2021, **STATISTA**, https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ (last visited Apr. 4, 2024).

[3] Fake news spreads faster than true news on Twitter—thanks to people, not bots, **SCIENCE,** https://www.science.org/content/article/fake-news-spreads-faster-true-news-twitter-thanks-people-not-bots (last visited Apr. 4, 2024).

[4] The rise of AI fake news is creating a 'misinformation superspreader', **WASHINGTON POST** https://www.washingtonpost.com/technology/2023/12/17/ai-fake-news-misinformation/ (last visited Apr. 3, 2024).

[5] What are LLM Hallucinations? , **IGUAZIO** https://www.iguazio.com/glossary/llm-hallucination/ (last visited Apr. 6, 2024).

[6] Finance worker pays out $25 million after video call with deepfake 'chief financial officer', **CNN,** https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html (last visited Apr. 4 2024).

[7] Man Motivated by 'Pizzagate' Conspiracy Theory Arrested in Washington Gunfire, **NEW YORK TIMES,** https://www.nytimes.com/2016/12/05/us/pizzagate-comet-ping-pong-edgar-maddison-welch.html (last visited Apr. 3 2024).

[8] Combating Fake News In Social Media: U.S. and German Legal Approaches, **ST. JOHN UNIVERSITY** https://scholarship.law.stjohns.edu/lawreview/vol91/iss4/5/ (last visited Apr. 6 2024).

[9] Most likely sources of fake news stories in the United States as of January 2017, **STATISTA** https://www.statista.com/statistics/697774/fake-news-sources/ (last visited Apr. 6 2024).

[10] Community Notes Introduction, **X**
https://communitynotes.twitter.com/guide/en/about/introduction
(last visited Apr. 6 2024).

[11] Why Twitter's Community Notes feature mostly fails to combat misinformation, **POYNTER**
https://www.poynter.org/fact-checking/2023/why-twitters-community-notes-feature-mostly-fails-to-combat-misinformation/ (last visited Apr. 6 2024).

[12] How AI is being used to fight fake news **THE CHRONICLE OF HIGHER EDUCATION**
https://sponsored.chronicle.com/how-ai-is-being-used-to-fight-fake-news/index.html (last visited Apr. 6 2024).

[13] The fine line between fake news and freedom of speech **KINGS COLLEGE LONDON**
https://www.kcl.ac.uk/the-fine-line-between-fake-news-and-freedom-of-speech
(last visited Apr. 6 2024).

[14] European Parliamentary Research Service-2019, page 7 **EPRS**
(last visited Apr. 6 2024).

[15] CoEAI-DIGITAL-TECHNOLOGIES-ELECTIONS-2019, page 37

[16] Combating Fake News In Social Media: U.S. and German Legal Approaches, p. 33
https://scholarship.law.stjohns.edu/lawreview/vol91/iss4/5/?utm_source=scholarship.law.stjohns.edu%2Flawreview%2Fvol91%2Fiss4%2F5&utm_medium=PDF&utm_campaign=PDFCoverPages

[17] A multi-dimensional approach to disinformation, page 16

[18] FAKE-NEWS-AND-US-ELECTIONS, page 1
Direct dowload: https://www.science.org/doi/10.1126/science.aau2706
Journal: https://www.science.org/journal/science

[19] FakeNewsQuandt2019, page 5