Researching Disinformation Methodology and Annotation Guidelines

Magdalena Wilczyńska

Katarzyna Lipka
Mateusz Zadroga
Michał Pawela

September 25, 2024

Contents

1 Abstract

We present our methodology and annotation guidelines designed to standardize the assessment of articles for disinformation content, aiming to reduce subjectivity and enable comprehensive analysis. Utilizing these annotation guidelines, we analyzed a substantial volume of articles to identify trends in disinformation dissemination and mitigate its impact. The methodology outlines assessment stages, including overall categorization and detailed content analysis, focusing on identifying manipulation techniques, author motivations, narratives, and intents. By facilitating rigorous content evaluation, the methodology seeks to enhance understanding of disinformation patterns and promote information integrity.

As part of the project, we created the crowdsourcing platform, around which the community of Polish fact-checkers and debunkers involved in the fight against disinformation in various areas of social life will develop.

2 Introduction

The purpose of this paper is to describe the work methodology and annotation guidelines of disinformation analysts working. The methodology was developed by analysts (debunking experts) employed in the project, based on their experience as experts, scientific knowledge available in the subject, as well as the experience of other institutions and organisations involved in research and detection of disinformation. All authors of this methodology have at least three years of experience working for fact-checking or debunking organizations accredited by the International Fact-Checking Network [1].

The analysis carried out by InfoTester project experts is at all times based on a transparent methodology. In addition, the entirety of the operation of experts is aimed at working in the spirit of the principles set out by the International Fact-Checking Network [9], i.e.:

- Impartiality and fairness — organisations test each policy option equally. They follow the same rules, regardless of whose statement they verify. Importantly, they do not take a position on the issues they check.

- Transparency of sources — organizations disclose data and documents on which they base their analyses. Readers are also to be able to form their own opinion and follow the verification process. Except for the situations in which information source safety may be compromised.

[1] The International Fact Checking Network gives accreditation to debunking organizations that sign its code of principles. See https://www.poynter.org/ifcn/

- Transparency of funding and organizational structure — organisations describe the professional background of all key individuals, structure, and legal status. They also disclose sources of their financing, while potential sponsors cannot influence the results of the analyses. It is also important to clearly state how readers can contact the organisation.

- Transparency of methodology — organizations clarify on what basis they select materials for analysis. They describe the methodology they use to conduct research, write texts and make any corrections. They also encourage readers to report information that is worth being verified.

- Commitment to open and honest revisions — organisations publish revision policies They need to improve their work in a transparent and clear manner, as well as ensure that readers see the most up-to-date version of the analysis.

3 Main Assumptions of the Methodology

The aim of creating a uniform methodology is to guarantee the identity of the assessments made by analysts and to minimise their subjectivity. The adoption of clear criteria for the analysis and the partial nature (gradability) of the assessment made also allows for a detailed description of the collected data, which in turn will allow for their in-depth analysis. The methodology is to allow an analyst not only to correctly indicate whether a given content is disinformative, but also to describe this content in such a way as to enable the widest possible gathering of knowledge about disseminated disinformation, i.a. by identifying common features of disinformative articles. Thus, the methodology is to enable content analysis in the qualitative study regarding a set of articles [7].

The analysis of articles is carried out mainly via debunking technique, with the auxiliary use of the fact-checking technique. These terms, for the purposes of this methodology, are defined in a manner analogous to the methodology developed for the purposes of the NATO Strategic Communication Centre of Excellence [10]. Fact-checking is the long-standing process of checking that all facts in a piece of writing, news article, speech, etc. are correct. Debunking refers to the process of exposing falseness or manipulating in a systematic and strategic manner (based on a chosen topic, classifications of selected techniques, narrative, intention, or the like). The overall goal is to minimize the impact of potentially harmful content by classifying it and identifying existing trends. The main goals of debunking are not only investigating the truth, but also cataloguing evidence of false information, exposing false information and conspiracies, and finding sources of disinformation.

The methodology will be used to evaluate an unknown set of articles coming from both credible as well as mixed/biased and potentially unreliable sources. The basic assumption is, therefore, its adaptability to the content that will be identified as essential and the possibility of excluding irrelevant content. The goal of the project is to analyse a minimum of 15,000 articles (5,000 containing disinformation and 10,000 containing reliable information). Selected articles will be then classified into 10 topics. Only articles classified under one of the selected topics will be analysed for disinformation. The analysis consists of three phases of content assessment: overall assessment (topic definition, source analysis, author analysis, initial content analysis); detailed content analysis (with distinguishing the type of disinformation and eristic techniques used or fact-checking description); and the assessment of the author's intentions and motivations.

4 Overall Annotation Guidelines

4.1 Thematic category

Task Type: exhaustive list, single choice

The assessment of the article begins with an initial content analysis during which the topic is determined, and it allows it to be classified (single choice) into one of 10 exhaustive thematic categories. Thematic categories were pre-defined a priori. The selection of topics resulted from a prior analysis of the work of fact-checking and debunking organisations, such as Snopes, "Counteracting Disinformation" Foundation, Demagog Association, and Debunk EU. The topics in our dataset greatly overlap with the most significant disinformation topics published in the recent EU DisinfoLab report [11]: only 3 out of 10 topics (5G, Paranormal Activities and Political News) of our dataset are not mentioned in the report.

If the content does not correspond to any of the defined categories, the option "Not related to the topic" is selected. The selection of this category ends the assessment of the article, and its content is not included in the study results.

- COVID-19
- Migrations
- LGBT+
- Climate crisis
- 5G

- Paranormal Activities
- War in the Ukraine and situation of Ukrainian refugees in Poland
- Pseudomedicine
- Women's rights (sex education, contraception, IVF, abortion)
- News or Other
- Not related to any of the topics

4.2 Source Analysis

Task Type: exhaustive list, single choice

The subsequent element of the initial assessment is the analysis of the source, i.e. the domain in which the content has been posted. Assessing the credibility of a website requires an in-depth analysis of the content posted on it regularly, as well as checking it in reliable sources, including via the Media Bias/Fact Check search engine. The analysis consists in selecting the category that best suits a given domain:

1. **Reliable** — sources that are reliable/publishing reliable content on a specific topic, in particular traditional news portals.
2. **Unreliable** — sources publishing unreliable content, typically disinformation, e.g. all domains financed by the Kremlin, sites containing conspiracy theories, etc.
3. **Mixed/biased** — partially or potentially biased websites that may present false information on specific issues, e.g. typically far-right websites, and blog collections.

4.3 Author Analysis

Task Type: exhaustive list, single choice

Verification of the message sender is crucial in determining its credibility, as disinformation is largely disseminated by anonymous authors or people who use fictitious profiles. The goal of the next step is to analyse whether the content has a verifiable author, and it consists in choosing one of the three options:

1. **Anonymous** — there is no information enabling the identification of the author (e.g. no signature or content written under a pseudonym); editorials.

2. **Unknown author** — personal data provided does not allow for a broader identification of the person or the data may be potentially fabricated (e.g. people who cannot be found, such as Jan Nowak publishing on web portal.

3. **Known author** — content signed with personal data enabling a wider identification of the author

After selecting the known/unknown author option, the analyst enters the available data, in particular name and surname, in a separate field.

4.4 Content Analysis

Task Type: exhaustive list, single choice

The next step requires analysing the entire content of the article and recognizing whether the information is true or of disinformative nature. If the article provides only factual information, it is marked as "reliable information". Selecting this category ends the assessment of the article.

In a situation where information contained in the article is unreliable and misleads the recipients, the analyst determines whether the given content was disseminated with the intention of misleading or causing harm. The unintentional dissemination of false information is known as misinformation. The intention is critical to identify whether we are dealing with disinformation or misinformation. When a person provides information that is untrue with no intention of causing harm or even not knowing that it is false, we refer to that person's activity as misinformation. It should be emphasized, however, that even unintentional dissemination of false information without the goal of manipulating recipients can fuel disinformation. Disinformation is particularly difficult to detect as the author's intention is usually not specified and in most cases, it can only be presumed. In the case of articles published by reliable sources of information and credible authors (recognizable journalists), the analyst assumes that given content is of misinformative nature and there is no intention to cause harm/spread falseness.

Overall, most definitions of disinformation combine four elements [1]:

- type of information
- falseness of information
- intention of the author
- andconsequences of disseminating information

- including the personal effects, e.g. recipient views
- and social effects, e.g. disruption of democratic processes

For the purposes of this report, the definition of disinformation provided by the European Commission High-Level Group of Experts on False News and Disinformation on the Internet (HELG) will be used, as it covers all four aspects and does not exclude potentially harmful content presented in the form of political advertising or satire, as presented in the EU Code of Practice. The definition is as follows [5]:

> *" All forms of false, inaccurate, or misleading information designed, presented, and promoted to intentionally cause public harm or for profit."*

However, a necessary supplement to this definition is taking into account the European Union Code of Practice on Disinformation, according to which disinformation is defined as: "verifiable false or misleading information which, cumulatively, (a) is created, presented and disseminated for economic gain or to intentionally deceive the public; and (b) may cause public harm, intended as threats to democratic political and policy-making processes as well as public goods such as the protection of EU citizens' health, the environment or security." [3]. The detected information must be verifiable, which means that it can be proved that it is untrue, and, therefore, it cannot be, for example, a yet unproven theory or opinion, as long as it is not intended to mislead the recipients. In a situation where a given content is not verifiable (reliable/disinformative/misinformative), it is marked as the "Hard to say" category. Indicating this category ends the assessment. Below we present main categories:

1. Credible information
2. Misinformation
3. Disinformation
4. Hard to say

Disinformation is intentional misleading by providing misleading or false information [2]. Unlike disinformation, misinformation is *misleading information shared by people who do not recognize it as such* [5].

5 Detailed Content Annotation Guidelines

Only articles flagged as misleading (categories of disinformation and misinformation) are subject to further analysis. Detailed content analysis is based on the study of the article text, both in terms of manipulation techniques used and in terms of straightforward false statements (fabrication). Various types of verification are required depending on the type of disinformation identified.

5.1 Classification within a group disinformation/misinformation

Task Type: exhaustive list, multiple choice

The nature of the disinformation disseminated can vary significantly and requires different analysis. Therefore, the first step in a detailed content analysis is to classify content into one of four categories. These categories are not disjoint as it is possible that both manipulation and falseness/fabrication may be found in a single disinformation article. The analyst can choose appropriate categories multiple times.

1. Falseness/fabrication - all explicitly false content and articles containing fabricated documents/photos/quotes and the like.

Articles classified in this group require a fact-checking comment. The analyst's task is to identify explicitly false (such as lies) or fabricated content (so-called false evidence of disinformation), and then to find arguments/sources refuting the false statement. Selecting this option allows for entering comments in the appropriate window, which must each time include indicating the part of the text containing falseness/fabrication and attaching a fact-checking comment and/or source (in particular scientific sources and articles from reliable organisations dealing with disinformation study, which carried out truthfulness analysis) that refute a given content. When citing articles from independent NGOs, the analyst must pay special attention to the organisation's affiliation with the International Fact-Checking Network (IFCN).

Moreover, each individual analysis in terms of the credibility of the content should be based on an attempt to find the source of given information; if the information is not found directly, analogous or similar content should be searched for. Moreover, potentially false information should be confronted with available information coming from reliable sources, in particular from public institutions and scientific sources.

2. Manipulation - bending or distorting facts in order to prove one's point or to influence someone else's views and behaviour.

Manipulative articles are often not explicitly false but they contain incomplete statements, used in an inappropriate context, selected information, etc. In the further part of the analysis, it is necessary to indicate the appropriate manipulation technique, which will be described below.

3. Trolling/satire - posting controversial, often untrue content, showing certain phenomena in a caricatured, playful manner.

Disinformation articles including trolling/satire can be based on both falseness and manipulation. After selecting this option, an analysis adequate to the type of disinformation should be carried out.

4. Conspiracy theory - proposed explanation of an event assuming significant participation of a group of conspirators trying to hide the truth from the public. To exist, a conspiracy theory needs an official version of a given event, which the conspiracy theory refutes by proposing a true version in its place [12]. Any evidence that disproves a conspiracy theory can be interpreted as further evidence of the conspiracy's existence (self-sealing).

Understanding the purpose and motives behind creation and dissemination of conspiracy theories is of key importance for their identification. To this end, road signs for the analyst are the features of conspiratorial thinking identified by S. Lewandowsky and J. Cook [8]:

- something must be wrong — secret plan, conspiracy, fraud;
- immune to evidence — evidence against the conspiracy theory comes from a conspiracy;
- overriding suspicion — extreme degree of suspicion makes it impossible to believe anything that does not fit the conspiracy theory;
- contradictory — belief of the followers of conspiracy theories is absolute, it does not matter to them that their belief system is inconsistent;
- persecuted victim — conspiracy theorists see and present themselves as victims of organised persecution. At the same time, they see themselves as brave antagonists taking on the villainous conspirators. Conspiratorial thinking involves a self-perception of simultaneously being a victim and a hero;

- nefarious intent — the motivations behind any presumed conspiracy are invariably assumed to be nefarious. Conspiracy theories never propose that the presumed conspirators have benign motivations;
- re-interpreting randomness — overriding suspicion found in conspiratorial thinking frequently results in the belief that nothing occurs by accident Small random events are reinterpreted as caused by a conspiracy and are woven into a wider, interrelated pattern.

Preliminary analysis of the motivations of the potential conspiracy theories creators also allows for their identification. At this point, the analysis should be carried out based on the assumption that the conspiracy constitutes an explanatory function. The purpose of such a mental construction is, therefore, inter alia: easy explanation of complicated processes and events, providing a simple answer, regaining a sense of control and security, mythologising the causes and background of events, political change — strategic and program concept.

Choosing this category exempts the evaluation of the eristic techniques used. This is due to the fact that conspiracy theories use a wide range of eristic techniques, including an appeal to emotion, exaggeration, cherry picking, oversimplification, post hoc ergo propter hoc, anecdote, quote mining, ambiguity, strawman, argumentum ad populum, false analogy, false dilemma, false experts [4]. It would be excessively time-consuming to identify them all.

5.2 Eristic techniques used for disinformation

Task Type: exhaustive list, multiple choice

Each time, during the analysis of manipulative content, it is necessary to indicate what types of manipulation (eristic techniques) were used by the author. In most cases, several techniques are used in one disinformative article. The role of the analyst is to identify them all. For the purposes of this methodology, a set of the eleven most common eristic techniques used in the manipulation was developed [4]. In the case of identifying a technique that is not on the list, it should be indicated it in the "Comments" field at the end of the analysis performed. The list of selected techniques includes:

1. Cherry Picking
2. Quote Mining
3. Anecdote

4. Whataboutism
5. Strawman
6. Leading Questions
7. Appeal to Emotion
8. False Cause
9. Exaggeration
10. Reference Error
11. Misleading Clickbait

5.2.1 Cherry Picking

Presenting information using only data that supports a given thesis while ignoring the broader context. It may include the slothful induction (rejecting inconvenient evidence that challenges our beliefs) or the Texas sharpshooter error (ignoring differences and emphasizing similarities, using from among an extensive dataset a small slice that supports our thesis).
Example: *Data show that last winter was the coldest in 10 years, indicating that the climate is not warming at all.*
Explanation: Using a single case of data taken out of context, ignoring the trend seen over a longer time frame. Even an exceptionally cold winter is not evidence of a change in trend.

5.2.2 Quote Mining

Using a short excerpt from someone's longer speech/text in a way that significantly distorts its actual, original meaning.
Example: *It is as though fossils were just planted there, without any evolutionary history.*
Explanation: The quote comes from R. Dawkins' book, "Blind Watchmaker". It is used by proponents of creationism as if it were evidence to support their beliefs. In fact, as the full statement shows, this short quote completely distorts the meaning of Dawkins' statement, which criticizes creationists and disagrees with their hypothesis. The quote in full reads as follows: *It is as though fossils were planted there, without any evolutionary history. Needless to say this appearance of sudden planting has delighted creationists. Both schools of thought (Punctuationists and Gradualists) despise so-called scientific creationists equally,*

and both agree that the major gaps are real, that they are true imperfections in the fossil record. The only alternative explanation of the sudden appearance of so many complex animal types in the Cambrian era is divine creation and (we) both reject this alternative.

5.2.3 Anecdote

The use of evidence in the form of personal experience or an isolated case, possibly rumor or hearsay, most often to discredit statistics.
Example: *In Bavaria, Germany, 25 cm of snow has just fallen. Global warming is the biggest lie in human history.*
Explanation: Authors use a singular example they experienced personally to discredit the statistic. Individual experiences cannot be translated into an entire population or multi-year studies.

5.2.4 Whataboutism

Responding to a substantive argument not by addressing the heart of the matter but by raising a new point unrelated to the topic. Often referred to as dropping a false lead to divert attention from the topic.
Example: *You talk about LGBT+ people being persecuted. What about hungry children? No one thinks about them!*
Explanation: In this case, the author is trying to change the object of discussion by redirecting attention to something else. Often, as in this case, this is to hide the author's prejudice by expressing concern for another, usually worldview-neutral thing.

5.2.5 Strawman

Misrepresenting someone's argument in a way that makes it easier to refute. It usually boils down to attributing to an opponent a position the opponent does not share.
Example: *Since you criticize Russia's actions, that means you are Russophobic.*
Explanation: The author of this statement assumes that criticism implies prejudice. It makes it easy to portray oneself or someone as a victim.

5.2.6 Leading Questions

Flooding the target audience with consecutive questions or false arguments/studies that are suggestive. Guiding the recipient to a preconceived thesis. A statement consisting of a plethora of poorly related information, half-truths, and misinterpretations designed to overwhelm by their sheer volume.

Example: *I do not think we even know the actual mortality rate for monkeypox. Has a Westerner ever died from it? Could this possibly be the same money pox that occurs in Africa? If so, how did it suddenly appear in so many countries at once? (. . .) Perhaps it is here just to nudge us to get another shot?*
Explanation: The questions are intended to lead the recipient to a specific conclusion. The author wants to avoid accusations of spreading conspiracy theories, so instead, he will try to shape the viewer's opinion through suggestive questions. He can always defend himself by saying he is "just asking questions."

5.2.7 Appeal to Emotion

The use of words and phrases arouses in the recipient a strong emotion and attitude toward the presented matter. The person using this technique tries to resonate with the recipient's prejudices (Appeal to Fear/Prejudice) or their values and traditions (Appeal to Values). They may also use short, vital phrases, including stereotyping or labeling (Slogans) and offensive and hateful language (Loaded Language). It can also use group affiliation (Flag Waving) or suggest a time for action (Appeal to Time) to mobilize the recipient to take specific actions.
Example: *They are MURDERING our children with vaccines. They are DEVILS who are just trying to harm the innocent! All doctors who administer vaccines will face cruel punishment.*
Explanation: The above statements are highly emotionally charged. Keywords are emphasized to evoke negative emotions in the recipient. A specific group of culprits to be held accountable is also indicated.

5.2.8 False Cause

Assuming a cause-and-effect relationship solely based on the observed correlation. Among the manipulative statements used are those relating to time, such as those assuming that two events happening at the same time must be related (Cum Hoc Ergo Propter Hoc) or one following the other must be cause and effect (Post Hoc Ergo Propter Hoc).
Example: *Russia's military intervention in Ukraine is the culmination of U.S. and NATO aggression, dating back to their blitzkrieg against Serbia 23 years ago.*
Explanation: The above statement simplifies the whole issue, leaving out many events. Authors choose a convenient cause for them and omit all other aspects.

5.2.9 Exaggeration

The simplification and misrepresentation of a phenomenon or issue. For example, an author manipulating an audience may present a vision in which one decision can lead to unwanted negative consequences (Slippery Slope). Another way is to exaggerate minor or irrelevant aspects of an issue or the attitudes of individuals to denigrate an entire group or issue (Blowfish). One can also be used to manipulatively exaggerate the importance of a small group of people with different opinions than the rest of their community (Magnified Minority).
Example: *If we allow same-sex marriages, the next step will be to legalize pedophilia.*
Explanation: The author uses Slippery Slope. It is intended to falsely show the consequences of a decision or prove that one decision leads to another negative one. This technique is primarily intended to cause fear or reluctance in the recipient.

5.2.10 Reference Error

It is a reference to unreliable sources or people. It can involve passing on knowledge from anonymous individuals, such as from social media, citing propaganda claims by politicians or media, primarily from authoritarian countries. It can also involve using untrue quotes circulating online to prove a point. This technique often cites fake experts or others to pretend to be a supposed authority (Appeal to Authority).
Example: *Vladmir Putin said Western sanctions are not working against Russia, and the economy is only growing. Once again, Putin proves how good a president he is, and the West has shown its weakness.*
Explanation: This is an example of invoking the propaganda claims of politicians and considering them a reliable source of information. However, it serves to manipulate public opinion.

5.2.11 Misleading Clickbait

Giving the text a title that does not reflect the information presented in the article, often even contradicting it.
Example: *Covidian Church. Once again, the "Carol visit" will not take place because of the "pandemic"*
Explanation: The title uses the phrase "Covidian Church" and puts the word "pandemic" in quotes, implying skepticism or mocking the situation, which suggests people are overreacting or they are using the pandemic as an excuse to cancel the carol visit. However, the article itself explains in the first sentence *Due to the coronavirus pandemic, there will be no "carol visit" in the traditional form in the Archdiocese of Poznan this year either.*, which

shows that "Carol visit" will take place, but not in a traditional form. The title is misleading, creating a sense of controversy or conspiracy not reflected in the content, presenting a straightforward explanation. This technique is used to attract readers by distorting or exaggerating the facts.

6 Annotation and Interpretation of Motivations, Intentions and Narratives

The study of the motivation and intentions of the disinformation content authors is potentially the most subjective element of the analysis, and therefore it is particularly important to develop precise components of the assessment. This will allow for maintaining uniformity of the analysis carried out by different experts.

6.1 Motivation

Task Type: exhaustive list, multiple choice

The purpose of marking the probable intention of the author during the analysis is to allow for examination of the disinformation phenomenon in a broader manner, and in particular, it is an attempt to identify what the authors of unreliable content are guided by. Motivation is also a determinant that allows for determining what kind of disinformation we are dealing with.

6.1.1 Internal political

Categories: left/right/liberal

Internal political motivations are characterized by a strong marking of content against one political option. They can be directed against both a political minority (opposition) and the majority (currently in power). Typical of this type of motivation is a clear definition of one's political group (an affiliation of the author/author's views), which allows for indicating their motivation. In addition, it should be noted that disinformative content is more often promoted on websites promoting the so-called far-right views [6].

6.1.2 External political

External political - acting in favour of external political forces, e.g. Russia

External political motivations are usually identified by the nature of the website/domain that publishes them. It is typical in the Polish information space to promote this type of content through websites financed directly or indirectly by Russia or Russian entities (platforms such as sputniknews.com, dziennk-polityczny.pl). An additional premise indicating the above-mentioned motivation is the promotion of content that supports the demands of foreign countries, which at the same time may harm Polish state interests. Motivated in an external political manner is usually any kind of propaganda that weakens international alliances (such as the EU or NATO), as well as ridicules or weakens the defense capabilities of a given country. Typical of this type of disinformation is its promotion by authors who have connections with foreign countries or people who were accused of collaborating with them in the past (e.g. Konrad Rękas or Mateusz Piskorski). Moreover, this disinformation is spread under false pseudonyms by unknown authors (e.g. Jan Nowak publishing at dziennik-polityczny.pl, or Antony Ivanovitz publishing on blogging portals).

6.1.3 Economic / Financial

Economic/Financial motivation is identified through contextual analysis. In the case of this type of disinformation, it is characteristic to promote among recipients specific products (e.g. encouraging the purchase of a specific supplement/drug under an article that incorrectly informs about its effectiveness) or by encouraging contributions to the activities of an organization/author disseminating disinformative content (such a practice can be found not only in social media but also on websites, e.g. legaartis.pl).

6.1.4 Gaining popularity

Several authors disseminate false content in order to gain popularity. This motivation can occur both in the case of misinformation (e.g. a popular portal that provides false information by accident, wanting to be the first to inform about an event without first checking its truthfulness), and disinformation (usually through social media accounts of "influencers"). This motivation is often related to another, e.g. economic/financial, motivation. This motivation can also be recognized by the fact of publishing false content under one's own name (e.g. websites of doctors who promote alternative treatment methods). It also happens that false information is disseminated for this reason on Internet forums/blogospheres as well, in order to gain an appropriate position within a given environment. This motivation usually does not occur when an article is published anonymously.

6.1.5 Social

Social - activity for the benefit of specific social movements

Social motivation is extremely important in the case of disinformation because it is based on the formation of groups and opposition to them, creating a division of “us/them” (e.g. “us” who do not believe in vaccinations / “they” vaccinating themselves and their children). This motivation often comes down to antagonizing the society, confirming or reinforcing the erroneous views of a given group or falsely refuting the arguments of opponents. Its aim is also to strengthen faith in the false beliefs of one social group.

6.1.6 Chosen one / Saviour complex

Chosen one/saviour complex - ego building, willingness to make people aware of the ‘truth’

Note disinformation cannot be based solely on this motivation, its main application is misinformation

Motivation based on the so-called the chosen one complex is typical of authors of conspiracy theories. However, it also occurs when a given author promoting false content has a recognisable group of recipients who are convinced of the truthfulness of the statements promoted by that person. Usually, it involves recognizing the existence of a (false) conspiracy consisting in accusing public institutions or scientists of concealing the truth.

6.1.7 Fear

Note disinformation cannot be based solely on this motivation, its main use is misinformation)

False information that is disseminated out of fear should usually be categorised as misinformation. The author believes in false information they transmit and disseminate it to warn other recipients. Some conspiracy theories are also spread out of fear (authors’ concerns about the existence of global conspiracies)

6.1.8 Need for simple explanations

Note disinformation cannot be based solely on this motivation, its main use is misinformation)

As in the case of fear, false information being disseminated due to the author’s internal need for the existence of simple explanations suggests that the content is of misinformative nature. This motivation is also typical of conspiracy theorists (the existence of a global conspiracy explains several social turbulence and complicated problems). However, it can also be found in disinformation in a situation where a given author/domain, promoting extreme views, simplifies the message so as to lead to generalisation.

6.2 Intention

Task Type: exhaustive list with the possibility of modification, multiple choice

In this methodology, understanding the intention behind disinformation is crucial for effectively analyzing it. Disinformation, according to our definition, is always spread intentionally, emphasizing the significance of comprehending the motives driving its dissemination. The intention goes beyond just the narrative being promoted; it encapsulates the broader goal of the author, which guides the specific narratives they employ to achieve that goal.The purpose of false messages is not only to persuade recipients to believe in one specific message (e.g. that COVID-19 vaccines contain dangerous graphene), but to evoke a broader emotional response (e.g. denying safety of vaccinations). The classification of fake content in terms of the author's purpose/intention allows for classification of groups of articles in such a manner as to recognise trends in disinformation and, as a result, understand what content, how and by whom is promoted in specific disinformation trends.

The process of identifying intentions is informed by analyzing the narratives present in disinformation campaigns. Analysts identify common themes and patterns in the false messages being disseminated. These narratives are then grouped together under broader categories of intention, based on the overarching goals they serve. Analysts label both intention and narratives (in the form of multiple-choice drop-down lists) that aim at refining the identified intention/message.

To illustrate the connection between intention and narrative, let's consider a few examples:

Example 1 Analyst approached an article with a specific false claim: "Reports reveal shocking evidence of health officials deliberately inflating COVID-19 case numbers to justify draconian lockdown measures, raising serious doubts about the accuracy and integrity of government statistics." Such a claim would be considered under the narrative 'Alleging manipulation of disease statistics', which overall intention is 'Undermining the credibility of public institutions'. By spreading claims of manipulated disease statistics, this narrative aims to undermine the credibility of public health institutions, eroding trust in their ability to effectively manage health crises like the COVID-19 pandemic.

Example 2 Analyst approached an article with a specific false claim: "Leaked documents expose mainstream media collusion with political elites, revealing a concerted effort to suppress dissenting voices and manipulate public opinion in favor of establishment candidates, undermining the democratic process." Such a claim would be considered under the narrative ‘Allegations of mainstream media bias’, which overall intention is ‘Changing electoral beliefs’. By alleging mainstream media bias, this narrative seeks to change electoral beliefs by casting doubt on the integrity of media outlets, influencing public opinion and electoral outcomes in favor of certain political parties or candidates.

Example 3 Analyst approached an article with a specific claim: "A Russian diplomat accuses Baltic states of being puppets of the West due to agreeing to share information on the military with the US." Such a claim would be considered under the narrative ‘Alleging state subservience to NATO/USA/EU’, which overall intention is ‘Undermining the international position of the state’. By alleging state subservience to international entities, this narrative undermines the international position of the state, portraying it as lacking sovereignty and independence in its decision-making processes.

Example 4 Analyst approached an article with a specific claim: "Insiders from international climate summits expose the hypocrisy and incompetence of global leaders, revealing how so-called ’climate agreements’ serve the interests of wealthy nations while burdening developing countries with unfair regulations, perpetuating environmental injustice." Such a claim would be considered under the narrative ‘Discrediting international climate agreements’, which overall intention is ‘Undermining the international organization/its decisions’. By discrediting international climate agreements, this narrative aims to weaken the legitimacy and effectiveness of these agreements, undermining international efforts to address climate change and portraying them as ineffective or harmful to national interests.

Example 5 Analyst approached an article with a specific claim: "Muslim refugees from Syria are attacking and raping women in Sweden and planning terrorist attack." Such a claim would be considered under the narrative ‘Reinforcement of xenophobia in connection with security situation, which overall intention is ‘Promoting social stereotypes/antagonisms’. By reinforcing xenophobic attitudes towards immigrants, this narrative fosters division and hostility within society, exacerbating tensions and undermining efforts towards inclusivity and acceptance.

Example 6 Analyst approached an article with a specific claim: "The Earth is flat, we have been lied for centuries, there is no scientific evidence to prove that Earth is a globe." Such a claim would be considered under the narrative ‘Negating the sphericity/sphericity of the earth’, which overall intention is ‘Negating scientific facts’. By promoting the idea of a flat Earth, this narrative aims to undermine widely accepted scientific principles and discredit the credibility of scientific institutions, leading to confusion and distrust in established scientific knowledge.

Further we present the list of all categorized intentions and narratives which falls under each category.

6.2.1 Undermining the Credibility of Public Institutions

Narratives present for this intention:

- Undermining pandemic countermeasures undertaken (quarantine, lockdown, etc.)
- Allegations of human rights violations
- Alleging manipulation of disease statistics
- Alleging of manipulation of death statistics
- Negating the ability to defend oneself
- Allegation of the spread of Nazism by Ukraine
- Undermining measures taken to combat the migration crisis
- Undermining measures taken to combat climate change
- Alleging favouritism towards minorities
- Promoting stricter abortion laws
- Negating the achievements of space agencies
- Questioning the credibility of airlines

6.2.2 Changing Electoral Convictions

Narratives present for this intention:

- Blaming opposition parties for internal problems in the country
- Promoting the government's accomplishments or the effects of the actions taken
- Placing responsibility on external actors for the government's failures

- Allegations of inaction on issues of national importance
- Alleging mainstream media bias
- Incorrect/ incomplete reasons for the economic crisis
- Hiding/diminishing internal problems in the Catholic Church

6.2.3 Undermining the International Position of the State

Narratives present for this intention:

- Alleging state subservience to NATO/USA/EU

6.2.4 Undermining an International Organisation / its Decisions (EU, WHO, UN, NATO)

Narratives present for this intention:

- Blaming NATO/UN/EU for conflicts in the region
- Accusing of aggression against Russia and its allies
- Undermining defense capabilities
- Alleging violations of international treaties
- Discrediting international climate agreements
- Alleging favoritism toward LGBT+ persons
- Alleging favoring minorities
- Accusing the media of covering up global conspiracies
- Allegations of violations of the sovereignty of member states/accusations of interference in internal affairs

6.2.5 Weakening of International Alliances

Narratives present for this intention:

- Concerning Poland and NATO,
- Concerning Poland and Ukraine

6.2.6 Promoting Social Stereotypes / Antagonisms

Narratives present for this intention:

- Reinforcing homophobia
- Reinforcing transphobia
- Reinforcing xenophobia
- Reinforcing xenophobia in connection with the economic situation
- Reinforcing xenophobia in connection with security
- Reinforcing xenophobia in view of the health situation
- Reinforcing religious conflicts
- Strengthening anti-Semitism
- Reinforcing of antagonism over pandemic
- Reinforcing hatred directed against medics
- Accusation of Russophobia
- Fostering antagonisms in connection with the climate crisis
- Incitement of transphobia in connection with safety of women/children
- LGBT+ as a threat to the ’family’
- Negating discrimination against LG8T+ persons
- Negating the legitimacy of feminivatives
- Condemning persons after abortion/in vitro procedure
- Negating gender discrimination
- Perpetuating stereotypical social/occupational roles
- Undermining the credibility of feminist movements
- Reinforcing antagonism over abortion beliefs

- Undermining the competence/education of medical professionals
- Strengthening of antagonism in connection with beliefs about the sphericity of the earth/exploration of space
- Strengthening antagonism in connection with political beliefs

6.2.7 Denying Scientific Facts

Narratives present for this intention:

- Undermining the existence of COVID-19
- Downplaying COVID-19
- Undermining the safety and efficacy of vaccination
- Undermining the reliability of diagnostic metadata
- Undermining the safety/validity of wearing masks
- Challenging the validity of lockdowns
- Seeking alternative treatments for Covid-19
- Challenging the safety of 5G networks
- Negating climate change
- Downplaying human impact on climate change
- Negating existence of transgender persons
- Negating existence of non-binary persons
- Equating homosexuality with pedophilia
- Promoting conversion therapy
- Recognizing non-heteronormative orientation as a disease or fashion
- Equating abortion with murder
- Misrepresentation of the abortion procedure

- Misrepresentation of the process of embryological development of the human being
- False mental or physical complications after abortion/contraception
- False complications of a child from IVF
- Promotion of naprotechnology
- The in vitro procedure requires the death of children/abortions
- Equating the ’morning-after’ pill with abortion
- Undermining the methods used in sex education
- Questioning the safety of 5G/PEM technology in the context of human health
- Challenging the safety of 5G/PEM technology in the context of the environment
- Proposing methods to protect against 5G/PEM technology (e.g., foil caps)
- 5G as a tool for control and surveillance of humanity
- Blaming 5G networks for the development of pandemics
- Claiming the superiority of alternative medicine over conventional medicine
- Undermining the efficacy and safety of conventional medicine
- Undermining the efficacy and safety of conventional medications
- Promotion of homeopathy
- Promoting over-supplementation (orthomolecular medicine)
- Promoting potentially harmful nutrition/diet methods or therapies
- False causes of diseases (total biology)
- False assumption that everything natural is good
- Questioning the existence, meaning or etiology of mental illnesses
- Negating the sphericity/sphericity of the earth
- Negating achievements in space exploration

- Negation of the existence of gravity
- Condensation trails a threat to human life/health (chemtrails)
- Misleading evidence of extraterrestrial life
- Predicting the cosmic doom of civilization

6.2.8 Boosting the Morale of One Side of the Conflict.

Narratives present for this intention:

- Undermining the opponent's defensive capabilities
- Justifying aggression against the country/citizens
- Exaggerating military/diplomatic successes

6.3 Evoked Emotions

Task Type: exhaustive list, multiple choice

The indirect purpose of spreading disinformation is to arouse extreme emotional reactions in the recipient. The use of such techniques and content that evokes strong emotions also allows for faster and more effective promotion of fake content. Identifying what kind of emotional reaction the author of disinformation counts on may allow, in the future, for identifying what kind of content is disseminated using what kind of emotions (e.g. linking propaganda directed against the European Union with feelings of national pride). The analyst can choose from a catalogue of 6 typical disinformation emotions, which can be expanded after consultation with the team.

1. Fear/sense of threat
2. Anger
3. Opposition/rebellion
4. Uncertainty/sense of confusion
5. Hope (false hope)
6. Pride

7 Annotation Comments

The last element of the assessment is the possibility for the analyst to enter a general comment on the content in the comments field. The field also enables the collection of other information relevant to the analysis, in particular: new fake experts cited by the authors of the disinformation; new keywords (usually neologisms or offensive expressions) that indicate disinformative nature of a given content; information about the structure of the text, such as punctuation and grammar errors or automatic translation of the text using, for example, Google Translate; suggestions for a new topic or a new intention and others.

8 Double Evaluation and Consensus Establishment

According to this methodology, all content must undergo a double evaluation. Articles are evaluated two times by two analysts, working independently of each other. The second analyst does not read the first performed assessment, but only evaluates the content according to the methodology, independently of the results of the first evaluation. The analyst then compares the two performed assessments and makes the final decision on the choices made in the analysis process. Discrepancies spotted by the double-verification analyst are discussed by the team. Then, a common, consistent approach to content classification is established. The final registered assessment is therefore a consensus based on the first and second assessment, and can include elements of both independent evaluations. In addition, comments entered in the comments field may be discussed in the team, e.g. the need to add a new intention or narrative. The purpose of double verification is therefore not only to avoid the so-called human errors but also to the standardisation of the methodology's application.

References

[1] BAYER, J., BITIUKOVA, N., BARD, P., SZAKÁCS, J., ALEMANNO, A., AND USZKIEWICZ, E. Disinformation and propaganda–impact on the functioning of the rule of law in the eu and its member states. *European Parliament, LIBE Committee, Policy Department for Citizens' Rights and Constitutional Affairs* (2019).

[2] COMMISSION, E. Communication from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions on the european democracy action plan.

[3] COMMISSION, E. The strengthened code of practice on disinformation 2022.

[4] COOK, J. A history of flicc: the 5 techniques of science denial. Skeptical Science. Available online: https://skepticalscience.com/history-FLICC-5-techniques-science-denial.html, 2020.

[5] DE COCK BUNING, M. *A multi-dimensional approach to disinformation: Report of the independent high level group on fake news and online disinformation*. Publications Office of the European Union, 2018.

[6] EDELSON, L., NGUYEN, M.-K., GOLDSTEIN, I., GOGA, O., LAUINGER, T., MCCOY, D., ET AL. Far-right news sources on facebook more engaging. *Cybersecurity for Democracy* (2021).

[7] FLICK, U. *The SAGE handbook of qualitative data analysis*. Sage, 2013.

[8] LEWANDOWSKY, S., AND COOK, J. The conspiracy theory handbook.

[9] NETWORK, I. F.-C. 'commit to transparency — sign up for the international fact-checking network's code of principles'. Available at https://ifcncodeofprinciples.poynter.org/ (2024/03/20).

[10] PAMMENT, J., AND KIMBER, A. L. *Fact-checking and debunking: a best practice guide to dealing with disinformation*. NATO Strategic Communication Centre of Excellence, 2021.

[11] SESSA, M. Connecting the disinformation dots: insights, lessons, and guidance from 20 eu member states. https://www.disinfo.eu/publications/connecting-the-disinformation-dots/, December 2023.

[12] SZYMANEK, K. O teoriach spiskowych. *Folia Philosophica*, 30 (2012), 259–281.