Co-inform

Context Matters,
Your Sources Too

Survey of Misinformation Detection Methods

D3.2

#ThinkCheckShare

Document Summary Information

Project Title:	Co-Inform: Co-Creating Misinformation-Resilient Societies		
Project Acronym:	Co-Inform	**Proposal Number:**	770302
Type of Action:	RIA (Research and Innovation action)		
Start Date:	01/04/2018	**Duration:**	36 months
Project URL:	http://coinform.eu/		
Deliverable:	D3.2 Survey of Misinformation Detection Methods		
Version:	1.1		
Work Package:	WP3		
Submission date:	March 31st 2019		
Nature:	Type of Document	**Dissemination Level:**	Public
Lead Beneficiary:	The Open University		
Author(s):	From The Knowledge Media Institute (KMi): Tracie Farrell - Research Associate Martino Mensio - Research Assistant Gregoire Burel - Research Associate Lara Piccolo - Research Fellow Harith Alani - Professor of Web Science		
Contributions from:	Ipek Baris, Orna Young, Allan Leonard, Anton Wicklund, Sarah Denigris, Ronald Denaux, Syed Iftikhar Shah, Eleni A. Kyza		

The Co-inform project is co-funded by Horizon 2020 – the Framework Programme for Research and Innovation (2014-2020) H2020-SC6-CO-CREATION-2016-2017 (CO-CREATION FOR GROWTH AND INCLUSION).

Revision History

Version	Date	Change editor	Description
1.1	5.3.2019	Tracie Farrell	First draft of outline
1. 2	5.3.2019	Martino Mensio	Adaptations of the structure of section 3 to the literature
1.3	8.3.2019	Tracie Farrell	Additions to the background sections on misinformation and information
1.4	15.03.2019	Martino Mensio	Section 3 full draft of sections and graphics
1.5	18.03.2019	Gregoire Burel	Section 3 Claim Review
1.6	19.03.2019	Ipek Baris	Section on Rumour detection and evaluation
1.7	19.03.2019	Tracie Farrell and Martino Mensio	Additions to sections 1 and 3
1.8	21.03.2019	Orna Young and Allan Leonard	Reviewing the early draft of the deliverable and helping to scope
1.9	26.03.2019	Lara Piccolo Tracie Farrell Martino Mensio Gregoire Burel	Section integrations, resolving outstanding comments, and conclusion
2.0	27.03.2019	Open for partners	With comments and additions from the above-named contributing authors

Disclaimer

The sole responsibility for the content of this publication lies with the authors. It does not necessarily reflect the opinion of the European Union. Neither the Co-Inform Consortium nor the European Commission are responsible for any use that may be made of the information contained herein.

Copyright Message

©Co-Inform Consortium, 2018-2021. This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both. Reproduction is authorised provided the source is acknowledged.

Executive Summary

This document is a Survey of Misinformation Detection Tools and Algorithms that we have examined within the Co-Inform project under WP3.

The introduction outlines how this work fits the objectives of the Co-Inform project and more specifically of WP3. In the body of the document, we discuss the properties and characteristics of such tools and algorithms, as well as how they fit into the larger "misinformation ecosystem", which includes also social, political and personal motivations. We briefly outline some of the specific problems that misinformation detection seeks to address and the types of misinformation that have been identified, including which properties of misinformation can be detected. We also explain some different models of how misinformation spreads that are relevant for detection. We discuss manual fact-checking and the opportunities and limitations of this skilled work that inform our detection approach. Then, we provide a detailed description of both the types of tools we were able to identify, such as style-based, knowledge-based, propagation-based or credibility-based tools and the granularity of the analysis they are able to provide (source, document or claim level). We provide an extensive list of tools with descriptive categories. This is intended to map tools to their position in the ecosystem, with regard to how, when and what kinds of misinformation they handle.

Table of Contents

Table of Figures

1. Introduction

In 2017, the Pew Research Centre, along with Elon University consulted a number of experts and theoreticians about the problem of misinformation and the promise of technology to help resolve some of the most pressing issues (Anderson & Rainie, 2017). Respondents were nearly split between confidence that technological solutions would advance significantly enough to provide adequate solutions to misinformation and those who believed that human nature is such that we will shape technology to our own “less than noble” interests, thus deepening the problem. However, the connection between technology and fears around misinformation has been seen many times before.

In the Co-Inform project, we are aware that people form opinions on complex issues of public concern based on what they see on social media, and that echo chambers and filter bubbles can result in limiting perspectives. However, we express an optimistic view on how technology can help fight misinformation, in particular misinformation online. We are integrating technology throughout the project to ensure that we are creating the best tools possible to assist human beings in the complex task of determining fact from fiction. Technology is not meant to replace humans in the jobs they do best, but to support them in reducing the scale of some of their tasks (through better detection methods, for example) or in amplifying their work (through the inclusion of claim standards, for example, as we will discuss in the sections below).

1.1. Objectives of WP3 and Task 3.2

In WP3, we are investigating the best and most appropriate methods for the accurate detection, monitoring and prediction of misinformation. We are cooperating with our partners from WP2 and WP4 to define the best sources of information, from social media, blogs, news sources and data from fact-checking websites to produce a language independent, integrated service on the Co-Inform platform. WP3 involves training and constructing semantically enriched classification models to help automatically detect misinformation content, in real time. Later in the project, we will extract and leverage flow patterns of misinformation and develop new models to predict these dynamics. We will also design and develop tools and models for understanding perceptions and behaviour of citizens on social media.

In Task 3.2, we are focused on the problem of volume and velocity of misinformation spread on social media. Effective algorithms and tools to automatically detect and rank such content, and to calculate its misinformation-probability will be essential for fact-checkers, journalists and citizens. In this task, we are working with partners to design, develop, and deploy intelligent algorithms for measuring misinformation-probability of text articles and their sources, and to rank them based on their relative importance using a mixture of supervised and, eventually, unsupervised models (e.g., linguistic and cascade patterns from T3.3). The detection algorithms will follow relevant policies from WP2, and integrate historical data from FCNI and other fact checking sites (collected in T3.1) to train and test these models. Models will be produced using content, temporal, semantic, network, and user features. Additionally, methods for detecting bots will also be developed

or adapted from existing tools, to capture nodes in the network that are deliberately set-up for the purpose of disseminating misinformation.

In this deliverable, D3.2, we provide a targeted survey of misinformation detection methods, ensuring that we have identified tools and algorithms that are appropriate for detecting various types of misinformation (discussed below). It is our aim to expand our survey from the time since the proposal, to incorporate the many new approaches and tools that have become available since the time of developing the project idea.

1.2. Relationship with other work packages

At this point in the Co-Inform project, we have had our first feedback from some of the co-creation workshops and technical deliverables that have given the project shape. From D1.1 and the subsequent co-creation workshops, we have outlined four stakeholder groups that are of particular importance: citizens, journalists, fact-checkers and policy makers. With this in mind, the technology we seek to develop should support each of these groups in developing misinformation resilience. From the project proposal, we aimed to have an impact on society through education, in particular around "detecting and handling misinformation". We also anticipated the interest and cooperation with fact-checking networks, to provide them with tools that meet real needs. Finally, we wished to contribute to evidence-based policy making, with regard to misinformation handling.

From pilot activities conducted under WP1, we have confirmed that social media is an important part of the news diet that many individuals consume, and that our tools will have to respond to new developments in how citizens access (and are exposed to) news content. In WP2, we adapted and developed ontologies for modelling misinformation, including indicators for "handling known misinformation", and identifying potential management and intervention policies that would be important for the platform. In WP3, we are now considering how best to implement this work within the technical work packages.

In D3.1 we outlined the ways in which the project will gather information from the web, through crawling various RSS feeds and keyword-based sources, as well as public social media content (through APIs). In D4.1 and 4.2, we delivered the generic Co-inform architecture, to describe the different components such a platform could have and how stakeholders might engage with it. Task 3.2 (and later Task 3.3 and 3.4) is situated between these two points to describe potential mechanisms for properly analysing and categorising information that is gathered through the means described in D3.1. In Task 3.2, we have been asked to "experiment with various novel supervised and unsupervised models for misinformation identification techniques from big data." In our following tasks, we will start to extract patterns and discover ways of predicting misinformation flow. In T3.4, we will be taking this further in exploring methods for sentiment analysis and behaviour categorisation.

1.3. The Misinformation Ecosystem and Detection

Even before misinformation enters the news life cycle, certain features of our online engagement already create an environment that allows for misinformation to thrive. The structure of social media platforms, the scale and scope of the problem, world events, and the characteristics of the information 'consumer' (Ciampaglia, Mantzarlis, Maus, & Menczer, 2018) all create the context in which information is received and interpreted. In this section, we intend just to illustrate the additional layer that is impacting all of the other layers beneath it, as we will discuss further in sections 2, 3, and 4. More specifically, the section addresses two challenges to misinformation detection that arise from the misinformation ecosystem: increased connectivity in the news cycle and how to encourage active agency in news consumers.

1.3.1 Extended News Cycles

Within a very complex, increasingly globalised network of people, news cycles are operating with the creation, distribution and ultimate consumption of news. However, social media has disrupted the traditional news cycle, such that creation, distribution and consumption are no longer controlled by the media. Ellyn Angelotti of the Poynter Institute for Media Studies has re-envisioned the news cycle, taking into account the new **interactive nature of news**, in which we are not only receiving content, but **content re-contextualised** through the thoughts and opinions of our own social networks and the public in general (Angelotti, 2012) (see Figure 1) Misinformation can be inserted into the public domain at the time of publishing, or at any other point in the cycle, not only by publishers, but also by consumers. This makes the problem of misinformation very difficult to contain or control.

Figure 1 "The Bew News Cycle". From Angelotti (Angelotti, 2012)

1.3.2 Limitations to the Current State-of-the-Art

In the course of the Co-Inform project, we have already published one paper in 2018 surveying the detection methods for misinformation that existed at that time. Fernandez and Alani had proposed some challenges and limitations to misinformation detection that continue to be relevant as we expand this work in D3.2. (Fernandez & Alani, 2018). These limitations include:

- Alerting users to misinformation without providing a **rationale**
- Treating users as passive consumers rather than **active agents**
- Ignoring the **dynamics** of misinformation and the influence of the **typology** of social networks (e.g. the kind of social network, it's coverage topics, it's architecture)
- Lacking **progression beyond the "generation of facts** and the early detection of malicious accounts"
- Focusing on the technological, rather than the **social and human factors** involved in misinformation

Many of the limitations that were documented at the beginning of the Co-Inform project, as is exemplified by this list, involve the human and social elements of misinformation. In addition, the authors cite general difficulties of manual methods that are unable to cope with the volume of misinformation generated online.

In the context of the Co-Inform project, we aim to mitigate these limitations, primarily through two means: a) **empowering human agents** to be more active and informed in dealing with misinformation online, and b) **going beyond detection** to consider how detection approaches can be leveraged to help **educate the public** and support awareness raising. Of course, we continue to seek misinformation detection methods that will improve the **accuracy** and **efficiency** in detecting potential sources of misinformation or misinforming content.

1.4. Contribution of this deliverable to the project

The scope of D3.2 has been determined through a) the connection with previous work packages and the overall direction of the Co-Inform project, and b) the extent to which the detection methods incorporate technology to help automate some of the processes of misinformation detection. As the project work has already scoped the work of D3.2, we present here the approaches to misinformation detection that we feel best describe the state-of-the-art, that represent the most novel approaches and that provide the best picture of what the Co-Inform platform could offer. We have largely focused on detection approaches for online settings. This is because most technological approaches are developed for digital media, and have been applied within the context of social media.

However, some of the tools and technologies described in this section would also be relevant for any kind of print media that could be digitally analysed.

The scope of this deliverable means that we will not go into great detail about issues of misinformation spread, as this will be covered by D3.3. However, as some detection methods described in section 3 relate to detection of spread patterns, we will cover the subject briefly under the heading of misinformation dynamics.

We also do not go into detail about intervention mechanisms, although detection and intervention are related. This task will be covered in more detail in D3.4. Still, we are experimenting with methods that we feel also have the potential to help educate the public, if the public can adequately understand how the tools work. This is discussed in the future work section.

1.5. Structure of the document

In section 2, we describe the misinformation ecosystem has it has been described through the literature and the importance of understanding how the ecosystem relates to misinformation detection. This includes a brief review of the different types of misinformation and their origins, who they target, the dynamics of misinformation spread and the politics of the misinformation ecosystem. In this section we outline the importance of considering this context when working with or interpreting the results of different misinformation detection tools. In section 3, we also review the manual fact-checking process, to demonstrate how this profession is working with technology to improve the quality and efficiency of their work. In section 4, we outline the misinformation detection ecosystem in technical terms, describing computational methods for identifying and dealing with misinformation (particularly, misinformation online). This includes a description of the various approaches, the type of data and processes on which they rely, and the level of granularity they can provide in their outputs. In section 5, we present more in-depth discussions of different tools to illustrate and explain these attributes. We also provide an extensive table of tools. We wrap up the deliverable in section 6, with a review of the contribution to the Co-Inform project and how our approach will both fill gaps and maximise synergies in combining social and computational approaches.

Glossary

Term	Definition
Source	A term that identifies the actors in the misinformation ecosystem. A source can be a news outlet (with an associated domain name), a public figure (as a person, website or as a social media profile), a citizen (considered on online social media as a profile).

Document	A general term to indicate what is being analysed. It can belong to different types: (news) article, social media post, public speech. Its properties can be grouped in *i)* content (text, headline, media attachments, references) and *ii)* context (what is the source, where it is shared, by whom, when).
Claim	An assertion containing information/facts that can potentially be checked (there are also claims that are not check-worthy or are not verifiable).
Stance	The position of a certain idea/somebody with respect to a claim.
Credibility	A measure, based on a set of evidences and motivations, of how a certain source is oriented towards good information (it's not the intent, it's based on behaviour and adherence to criteria).
Judgement	The output of the fact-checking on Misinformation

2. Mapping Misinformation for Detection

A recent paper by Derek Ruths has highlighted the importance of mapping misinformation types and sources to specific detection methods to better understand and contextualise sometimes conflicting accounts of what is happening in the misinformation ecosystem. Ruths points out inconsistencies within the research literature, for example, regarding the importance of bots in spreading misinformation, or the ways in which misinformation proliferates in more homogenous or heterogenous networks. He argues that understanding the definition of misinformation that is utilised by different researchers and the type of problem they are investigating, shapes how their findings are to be interpreted (and thus how tools should be used) (Ruths, 2019). Some examples of these inconsistencies are provided in section 2.5 on the politics of misinformation.

Before defining the entities and dynamics involved in `misinformation detection', however, it is worth briefly reviewing what is meant by *'information'* and what it is that is being distorted or misrepresented. We can utilise this discussion to better **contextualise detection methods** and what it is that they are able to highlight or leverage.

Israel and Perry defined information in terms of information *carriers*, informational *content* and *context* (Israel & Perry, 1991). In other words, each piece of information has provenance (we refer to this later as the **source**), which includes where it came from, when it was transmitted and who transmitted it. Each piece of information also has some **proposition** (we refer to this later in the document as a 'claim'), which may also be attached to other pieces of information that add complexity to the informational content. Finally, each piece of information has some sort of **temporal, spatial and social context** that impacts how that information should be perceived. Consider the example below, adapted from Israel and Perry:

> *"The x-ray indicates that Jackie has a broken leg"*

The presence of the word 'indicates' (or another action word, such as 'shows', 'proves' etc.), along with the noun of 'x-ray', describes the provenance of the information. The informational content itself is that Jackie has a broken leg. Perry and Israel describe this kind of information as 'factive' - if the provenance is to be trusted, then the information is true. If the x-ray is to be believed, then Jackie has a broken leg.

However, in the context of misinformation, there are many places where **provenance and content can be distorted**. Using the example from above, what if the x-ray is three years old? What if that x-ray is not of Jackie, but of Rita? What if the x-ray was read by a lawyer,

rather than a medical professional? What about the additional information that Jackie is a dog? Additional information shapes perceptions about information. The statement about Jackie's broken leg is just one simple factual statement that could be easily verified. Online, distortion can be amplified, as informational content can be much more complex (Hochschild & Einstein, 2015), **divorced from the source and context** (Cheney, Finkelstein, Ludaescher, & Vansummeren, 2012) and **manipulated with increasing precision** (Chesney & Citron, 2018)**.**

2.1 Types of Misinformation

Misinformation can be considered in terms of its related concepts, typologies and models. In this section, we describe the definitions and descriptions of misinformation that informed the development of this deliverable and to which we refer later.

With regard to concepts, the terms **misinformation** and **false news** are used interchangeably within the media. However, two related concepts, **disinformation** and **'fake news'**, are subsets of misinformation that have specific qualities, and should be differentiated. Disinformation is largely about *intent to deceive*, whereas misinformation may not involve deception at all (Lazer et al., 2018). Fake news generally refers to **fabricated content**, but the term has become convoluted through overuse and misapplication in the media, such that it is not only meaningless but potentially harmful (Wardle, 2017). These definitions and descriptions are important to the subject of misinformation detection because they help to identify features that could be analysed, such as deceptive language, or false context (Burgoon, Blair, Qin, & Nunamaker, 2003).

Columbia Journalism Review[1] identified 6 types of misinformation in the context of the 2016 US presidential election, that continue to persist online:

- **Authentic material used in the wrong context**
- **Imposter news sites designed to look like brands we already know**
- **Fake news sites**
- **Fake information**
- **Manipulated content**
- **Parody content**

[1] **https://www.cjr.org/tow_center/6_types_election_fake_news.php**

These basic types of misinformation have also been identified by other entities, such as the international network of FirstDraft, which researches and proposes standards for "monitoring, verification and reporting on disinformation." FirstDraft identified 7 types of mis- and disinformation (see Figure 2), dividing false context from false connection. We will continue to use these taxonomies in mapping the type of misinformation that certain detection approaches are intended to capture.

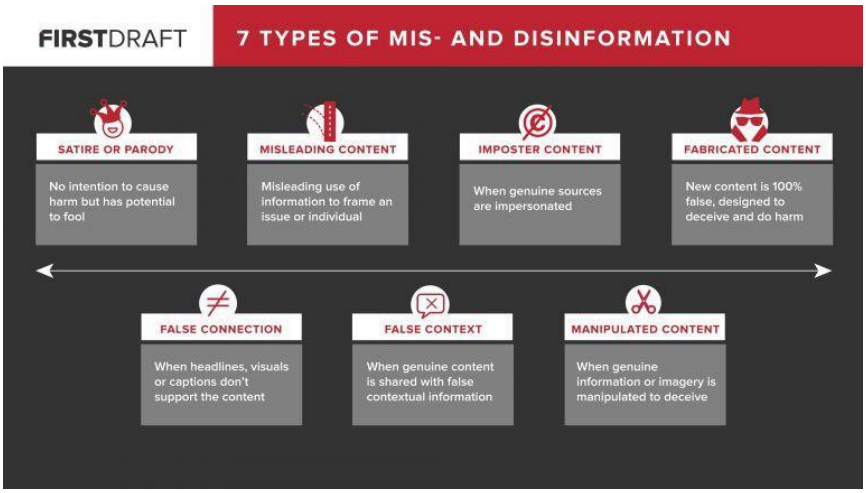

Figure 2 Key types of misinformation from FirstDraft, Claire Wardle (2017)

An example of **authentic material used in the wrong context** might be a photograph of someone taken at one event, that is then falsely used within the context of a different event or activity. One example from Germany involved a photograph from 2012, taken of refugees at a protest in Bonn. The photo shows some young men, with a flag that may or may not be associated with the so-called Islamic State, in apparent conflict with German police. This photograph, taken before the declaration of the caliphate, was used more recently to promote an image of refugees as terrorists and aggressors in the context of the mass migration as a result of the Syrian war[2]. FirstDraft distinguishes two types of authentic material used in the wrong context. **False connection** occurs when there is a mismatch of visuals and captions with the content. **False context** involves sharing potentially genuine content, but without the appropriate or necessary contextual information. Both can result in **misleading the public** about important and controversial issues.

Imposter news sites and imposter content can leverage the reputation of other institutions with sometimes very simple methods to drive consumers to their content. In Brazil, the fact-checking organisation Aos Fatos (which uses the web address .org) was

[2] **https://www.independent.co.uk/news/world/europe/isis-flag-picture-that-claims-to-show-refugees-attacking-police-goes-viral-and-is-a-lie-10501290.html**

compromised by an imposter website using a .com address of the same name.[3] Imposter news sites can also capitalise on popular typos or subtle changes in the domain name to trick users into engaging with misinforming content.[4]

Fake news sites typically generate fabricated, misleading or antagonistic stories (Wingfield, Isaac, & Benner, 2016). **Fake information** might include **rumours**, for example, which can create significant social disruption and emotional turmoil when they are false (Zubiaga, Aker, Bontcheva, Liakata, & Procter, 2018). Zubiaga *et al* (*ibid*) identified two main classes of rumours, those that emerge when news breaks about a given event and those that persist over time. For example, the authors identified several false rumours surrounding the fatal shooting of Michael Brown in Ferguson that were disseminated close to the time the event occurred, about his arrest record other factors in the investigation, as well as false rumours that persisted long after his death, particularly regarding the causes of resulting riots and unrest.[5][6] These divisions can help researchers to identify rumours on the basis of how and when they are circulated. Rumours may be about individual people, such as the idea that former US president Barack Obama is a muslim, or specific events (Dang, Smit, Moh'd, Minghim, & Milios, 2016), and can occur without intention to deceive in the absence of enough quality information. **Hoaxes** are another type of fake information, disguised to emulate the truth, sometimes successfully in mainstream media (S. Kumar, West, & Leskovec, 2016). This may include the presence of fabricated evidence or other information that is deliberately misrepresented as truth. One example of a recent hoax is the Momo Challenge, which recently made headlines for allegedly inciting young children to engage in self-harm or to commit suicide. The UK Safer Internet Centre and numerous UK charities working on mental health and youth have stated that they have seen no evidence of this challenge actually having the stated impact, but view the hoax itself as dangerous and "raising the risk of harm" for vulnerable young people.[7]

Parody or satire continues to be included in the discussion of misinformation for two reasons. First, satire and parody are covered under free speech laws in many cases, providing some protection for those who may intend to mislead or misinform if they label their content as satire (Klein & Wueller, 2017). An example is the website America's Last Line of Defense, which argues in its "About Us" page that they are producing satire, rather than hyperbolic and incendiary content.[8] Second, the characteristics of real satire and

[3] **https://www.poynter.org/fact-checking/2019/this-website-impersonated-a-fact-checking-outlet-to-publish-fake-news-stories/**
[4] **https://www.forbes.com/sites/christopherelliott/2019/02/21/these-are-the-real-fake-news-sites/#7930cac33c3e**
[5] **https://www.snopes.com/fact-check/blm-ferguson-russia/**
[6] **https://www.snopes.com/fact-check/riot-act/**
[7] **https://www.theguardian.com/technology/2019/feb/28/viral-momo-challenge-is-a-malicious-hoax-say-charities**
[8] **https://thelastlineofdefense.org/about-us/**

parody may assist in detection. Satire typically contains “cues revealing its own deceptiveness”, that may make it discernable from other types of fake news content disguised as satire (Rubin, Conroy, Chen, & Cornwell, 2016).

Types of misinformation can also be classified across different dimensions. Tandoc, Lim and Link proposed a typology for misinformation across the two dimensions of **`facticity’** and **`deceptiveness’** (Tandoc & Ling, 2018). Within those dimensions there are many different approaches to misinformation and misinforming. What might be called propaganda or `disinformation’, for example, would fall under the description of having both high deceptiveness and low facticity. McCright and Dunlap (2017) described two other dimensions of the **rhetorical style of the carrier** and **primary audience (recipient)** of misinformation (see Figure 3). On one axis, the authors have the realism/constructivism continuum, which refers to the information carrier’s “ontological position on truth and facts.” In other words, this axis refers to how important the truth appears to be to the person or entity that is the source of the (mis)information. On the other axis, the authors have the style of the information carrier and their use of language from very formal to very informal style. News sources such as Breitbart, according to the authors, are less attached to the facts and use a highly informal language, aimed toward regular citizens. Political pundits in the United States, such as Bill O’Reilly and Sean Hannity appear to have an interest in the truth (whatever their interpretation of that is), but still use a rather informal style of communication.

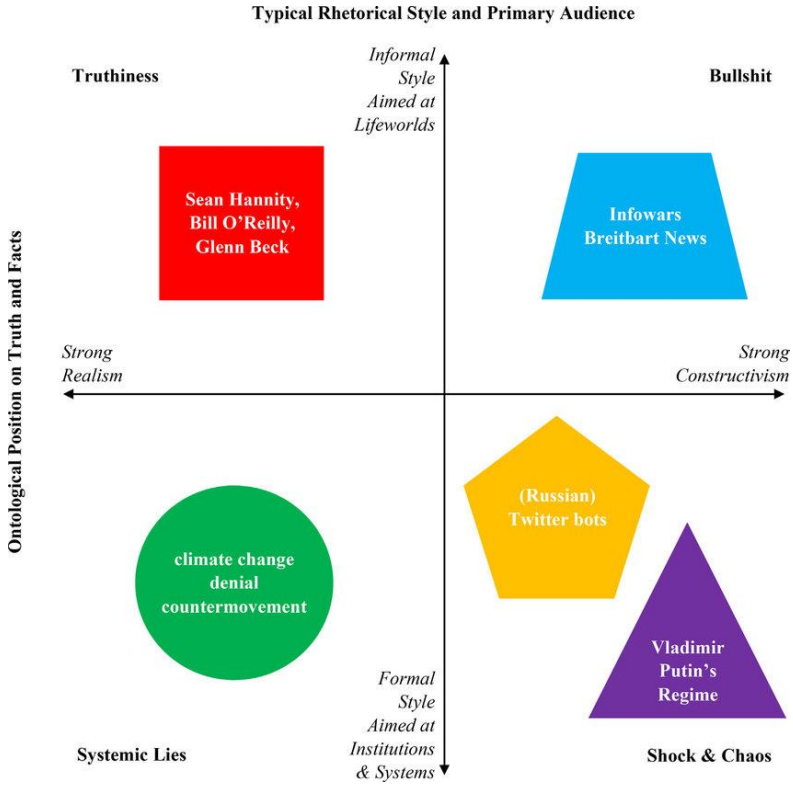

Figure 3 Key types of misinformation (with illustrative examples) From McCright & Dunlap, 2017)

2.2 Misinformation Carriers and Sources

We have mentioned several examples in section 2.1 that have implied who or which entities may see an advantage in *producing* misinformation. With regard to propaganda or political partisanship, for example, **state agents and politicians** may wish to mislead the public to influence their emotions, opinions, attitudes or actions through misinformation (Ellul, 1966). **Imposter news sites** and some **satire sites** also produce misinformation (intentionally or not). Washington Post reporter Eli Saslow has interviewed Christopher Blair, the person behind a popular website America's Last Line of Defense, about the impact of his self-proclaimed "satirical" news content on the attitudes and opinions of the American public[9]. America's Last Line of Defense has been under scrutiny several times in the past years for spreading misinformation about polarizing issues.[10] Blair argues that he wants to highlight the ignorance of the far-right and expose their bigotry, despite the fact that his end-consumers often miss this subtle rebuke. As Saslow points out in his article, Blair also makes a considerable profit from these deceptions. The financial gains of misinformation are therefore of general concern in considering how best to fight it.

For misinformation detection, however, the focus is less on producers and more on **sources** of misinformation. Fact-checking organisation Full Fact recognises several potential sources of misinformation that are necessary to monitor: **television and subtitles, radio, newspapers, debates, press releases, websites, blogs, emails, adverts (on- and offline), and social media**.

Two categories of misinformation sources are of particular interest in misinformation detection within the Co-Inform project are **media entities** (both real and imposter) and **bots**. As we have mentioned previously, the role of bots blends into work that we plan to conduct in the context of D3.3 on information cascades. For this reason, we do not go extensively into bots during this deliverable 3.2. However, we mention the existence of bots insofar as bot detection makes up a subset of detection activities that have been a controversial part of the discourse on misinformation. Media entities and understanding the credibility of media entities through automated or semi-automated means is one of the outputs of the Co-Inform project that have emerged across work packages.

2.3 Targets of misinformation

The **recipient or consumer** of the information is a critical aspect of misinformation in general (Karlova & Fisher, 2013). Who is being misinformed and what does it mean to be misinformed? First, looking at the contexts in which misinformation is prevalent, we can draw some conclusions. For example, misinformation spikes during **times of conflict and war**,

[9] **https://wapo.st/2Hq1amN**
[10] **https://www.factcheck.org/tag/americas-last-line-of-defense/**

when each party involved hopes to be able to control the message to the public (Lewandowsky, Stritzke, Freund, Oberauer, & Krueger, 2013). To a large extent this continues more generally within the field of politics, where information is used to promote ideas through **persuasion and issue framing**, especially in the context of elections (Kuklinski, Quirk, Jerit, Schwieder, & Rich, 2000). Misinformation often accompanies **breaking news developments**, when people are looking for more details (Starbird, Maddock, Orand, Achterman, & Mason, 2014), as well as during **disasters**, when they might desperately need information about where to go or what to do next (Lu & Yang, 2011). Misinformation may be targeted at certain groups of people, because they are more susceptible to certain types of misinformation that conforms to their beliefs (Jost, van der Linden, Panagopoulos, & Hardin, 2018), or because their behaviour is viewed as important to shape and control. The Computational Propaganda Research Project, for example, found that Russian interference in the US election was extremely targeted, encouraging African Americans to boycott elections, for example, or encouraging conservatives to become "more confrontational" (Howard, Ganesh, Liotsiou, Kelly, & François, 2018). In a very real sense, misinformation is able to **leverage and manipulate the emotions of people** (Huang, Starbird, Orand, Stanek, & Pedersen, 2015), intentionally or otherwise, influencing their reasoning behaviour and ultimately their choices.

In our view, misinformation detection will be more accurate and efficient if it involves an understanding of the psychological, social and technological factors that are involved in identifying and interrogating misinformation online. For our work on detecting misinformation cascades and spread patterns, we plan to put more emphasis in **understanding the people who are most targeted by misinformation** and their networks.

2.4 Dynamics of Misinformation

Within the dynamics of misinformation, there are different models for describing the entities involved and relationships that exist between them. These dynamics can be studied to understand more about how misinformation spreads and how to stop or minimize it.

In the **social diffusion model** from Karlova and Fisher (see Figure 4), the recipients' judgement about both the truth of the information content and the credibility of the information carrier are central to decision-making. The model acknowledges many unknowns, such as the intentions of both the information carrier and the recipient (Karlova & Fisher, 2013). Social diffusion models seek to understand the relationship between the receivers and producers of misinformation and how this extends to discussions on credibility and trust. These models are also useful for understanding why socially encountered misinformation can have longer lasting effects than misinformation that has not come across through social means (Gabbert, Memon, Allan, & Wright, 2004). In addition, cognitive models of misinformation can support social diffusion models by exploring the effects of consistency, coherence, credibility and acceptability of information in end-consumers (K. K. Kumar & Geethakumari, 2014).

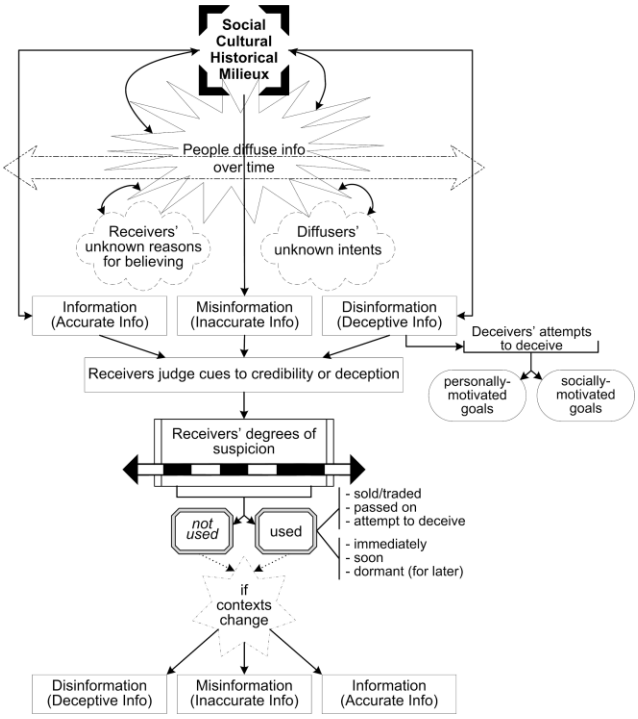

Figure 4 Social Diffusion Model of Misinformation. From Karlova & Fisher, 2013

Wu, Morstatter, Hu, & Liu, (2016) examined other more mathematical approaches to understanding **information cascades** and misinformation **containment**, some of which has been adapted from health research, to better control the spread of misinformation. The SIR model, for example, assesses the "health" of the community in terms of who is susceptible to misinformation, who is currently "infected" and who is recovered. This metaphor is carried over into other mathematical models, such as **linear threshold models** and **tipping point models**, that can help to identify hotspots in the network.

We plan to more fully explore the **dynamics of misinformation in D3.3**. However, providing an overview of these models will help to frame the discussion of detection approaches in the following sections.

2.5 The Politics of Misinformation Detection

The impacts of misinformation and expectations of the public in the aftermath of a large-scale misinformation-related incident shape how and when different types of misinformation detection tools begin to have traction in the research literature. For example, in 2007, AI-assisted fake pornography or "Deep fakes" (altered images that look especially realistic) began to appear on the internet. Within a period of only a few years, images were increasingly difficult to tell apart from a non-altered image[11]. This has even led to the creation

[11] https://motherboard.vice.com/en_us/article/gydydm/gal-gadot-fake-ai-porn

of popular apps, such as FaceSwap, which allows a user to switch faces with another user[12]. The AI technology behind deep fakes is developed by training the software on thousands of images. As celebrity images are quite easily available on the internet, they are quite often the subject of deepfake pornography. For misinformation detection, however, deepfakes represent a real problem, not only for those who are falsely represented. Deep fakes also **erode trust in what we see**, which can create an increasing **skepticism toward any kind of evidence** (Chesney & Citron, 2018). The tools to detect such fakes have had to evolve and be innovative in their design, such as looking for lighting irregularities (Pomari, Ruppert, Rezende, Rocha, & Carvalho, 2018) or blink patterns (Li, Chang, Farid, & Lyu, 2018), because these tools will have to provide robust evidence to counter both increased vulnerability and skepticism .

Bot-detection software, to provide another example, became more visible in the lead up to and just following the US presidential election in 2016. It was feared that bots are capable of pushing messages out into different communities and attempting to shift the public discourse (Ferrara, Varol, Davis, Menczer, & Flammini, 2016). Bessi and Ferrara provided evidence that bots represented 1/5th of the conversation (using hashtags) about the 2016 US election. They argue that the negativity that the bots added to this conversation can erode the public trust and **spoil the conversation between real users** (Bessi & Ferrara, 2016). Benigni, Joseph and Carley demonstrated in late 2018 that bots could account for up to 60% of network influence through **promoting certain accounts** (Benigni, Joseph, & Carley, 2019). Bots also exist in sub-communities that have specific coverage topics, from advertising to political campaigning. However, in the examples above, researchers have not demonstrated a) any impact on the decision-making of voters or b) that bots are more likely to spread misinformation than humans. Vosoughi, Roy and Aral argued that although bots do spread a lot of misinformation, **they appear to spread real information as well** (Vosoughi, Roy, & Aral, 2018a). The authors argued that, even if bots can put misinformation into the network, it is **humans who are most responsible for spreading** it. Could it be that one's position toward the 2016 US election impacts how invested one might be in understanding the influence of bot accounts? It is up to all of those involved in designing, using or interpreting these tools to reflect on the **powerful motivations and biases that might exist within the whole ecosystem.**

While politics has influenced researchers and companies to provide solutions to growing difficulties related to misinformation, it is also important to consider how tools are also expressions of political ideology. One recent example was provided to us from our partners most intimately involved in fact-checking (FactCheckNI). The NewsGuard fact checking site[13], which has the tagline of restoring trust and accountability, recently came under fire from the Daily Mail for having assigned the publication a poor rating. This rating claimed that the Daily Mail was often not conforming to journalistic standards, to which the Daily Mail

[12] **https://techwiser.com/best-face-swap-apps/**
[13] **https://www.newsguardtech.com/**

objected in writing. NewsGuard responded to the critiques from the Daily Mail and has since changed the rating from red to green[14]. This story highlights two difficulties facing fact-checking organisations that are relevant for detection. First, fact-checkers and fact-checking organisations have their own blind spots and ideologies. Understanding this, many fact-checkers, according to our experts, are rarely comfortable providing a simple “true” or “false” label on a news item. In addition, making judgements about the intentions of a creator of a piece of information, does not fit their professional mandate. Rather, they prefer to outline the reasons why the piece of information might or might not be true. Even these careful considerations come under scrutiny (in particular by conservatives) for being too fastidious in making their judgements.[15]

However, and this is the second challenge, as we attempt to give the public a better idea of what to look for in misinformation, we often speak to the idea of “credibility” or “trust”. **It is a difficult task to marry together the ideas of trust in reputable sources, along with critical thinking and information literacy**. We will have to apply both, even to our own potential biases and assumptions, across activities in the Co-Inform project and within the technologies we create.

3. Manual Fact-checking

"*From politicians to marketers, from advocacy groups to brands — everyone who seeks to convince others has an incentive to distort, exaggerate or obfuscate the facts." (UNESCO report pg 86).* Fact-checkers are equipped with skills and a methodology to detect fact-checkable claims and evaluate evidence critically, in line with ethical norms and standards (Posetti, 2017)

In the ecosystem of Misinformation (described in section 4 below), manual fact-checking takes a central role that is always needed. Misinformation is hardly ever so cut and dry as a fake image or fake caption as we outlined in section 2. Many times, the truth is much more complicated and the need for human fact-checkers continues. In this section, we describe briefly the elements of the manual fact-checking process, in particular, those that integrate with automated detection and procedure (such as the ClaimReview standard described below).

External post-hoc fact checking started being practised by independent organisations in the early 2000 (Graves, 2013). Since then, many different organisations (such as Politifact[16],

[14] **http://www.niemanlab.org/2019/01/newsguard-changed-its-mind-about-the-daily-mails-quality-its-green-now-not-red/**
[15] **https://thefederalist.com/2019/02/06/state-american-fact-checking-completely-useless/**
[16] **https://www.politifact.com/**

Snopes[17], FullFact[18]) have opened the way to many others (as can be seen in the lists reported by IFCN[19], Reporterslab,[20] FactCheckEU[21]).

In September 2015 the Poynter Institute created a unit dedicated to bringing together fact-checkers worldwide: the International Fact-Checking Network.[22] The idea is to support and coordinate (in terms of standards of practices and principles) a wide number of diverse initiatives.

The result is that nowadays this is the *de-facto* coordination network. The adherence of its signatories[23] to a set of principles makes it the regulator of real and trustworthy fact checkers.

The principles that are signed as commitment and evaluated periodically are the following[24]:

1. **Organisation**
    a. The applicant is a **legally registered** organisation set up exclusively for the purpose of fact-checking or the distinct fact-checking section of a legally registered media outlet or research institution.
    b. The applicant publishes reports that **evaluate distinct claims** exclusively on the basis of their accuracy. It does so on a regular basis (an average of at least one report a week over the previous three months).
2. **Nonpartisanship and fairness**
    a. The applicant demonstrates that fact checks cover a **variety of subjects** or speakers and do not unduly concentrate on one side of the topic/context they fact check.
    b. The organisation must not support a candidate in any election nor advocate or take policy positions on any issues not strictly related to fact-checking. The organisation should also explain its policies on preventing staff from direct involvement in political parties and advocacy organisations.
3. **Transparency of sources**
    a. The applicant links to the **sources of the claim** it is fact-checking and the e**vidence it uses** to fact-check it (or identifies them in detail where these can't be linked).

[17] **https://www.snopes.com/**
[18] **https://fullfact.org/**
[19] **https://www.poynter.org/ifcn/**
[20] **https://reporterslab.org/fact-checking/**
[21] **https://factcheckeu.info/**
[22] **https://www.poynter.org/ifcn/**
[23] **https://ifcncodeofprinciples.poynter.org/signatories**
[24] **https://ifcncodeofprinciples.poynter.org/know-more/what-it-takes-to-be-a-signatory**

4. **Transparency of funding and of organisation**

    a. All applicants that are standalone organisations must have a page on their site detailing each **source of funding** over the past calendar year accounting for 5% or more of total revenue. This page should also set out an overview of spending that year and indicate the form in which the organisation is registered (e.g. as a non-profit, as a company etc) and, if this would allow the public to verify certain financial information about it.

    b. All **authors and key actors** behind the fact-checking project must be clearly listed on the site and their biographies indicated.

    c. Signatories must indicate an easy way for the **public** to reach out directly to the fact-checking initiative with complaints or comments.

5. **Transparency of methodology**

    a. The applicant explains its **fact-checking methodology** publicly and clearly in an accessible place.

    b. The applicant indicates to readers how they can send claims to fact-check while making it clear what readers can legitimately expect will be fact-checked and **what isn't fact-checkable.**

6. **Open and honest corrections policy**

    a. The applicant has a **public corrections policy** and can point to the means in which its audience can **request a correction**. If the organisation *y* is the distinct fact-checking section of a legally registered media outlet or research institution, the parent organisation must have the same corrections policy.

    b. The applicant indicates to readers how they can **send claims to fact-check** while making it clear what readers can legitimately expect will be fact-checked and what isn't fact-checkable.

While on one side these principles assure the **validity of its signatories**, on the other side there are a wide set of critics to how the fact-checkers check and their political fairness. Some of these critics come from studies that do cross-comparison (Lim, 2018) of reviews, **finding discrepancies**. Once again, this is evidence of the political aspects of misinformation detection and fact-checking that was set out in section 2.5. It is also evidence for the Co-Inform project that relying on any one tool, approach or entity

3.1 Manual Fact-checking Process

In the recent scenario in which social media plays in disseminating news and misinformation, the work of fact-checkers go beyond fact-checking public claims, but may include debunking viral hoaxes. Debunking is a subset of fact-checking and requires a specific set of skills that are in common with verification, especially of user-generated content, as described in the diagram below (Ireton & Posetti, 2018).

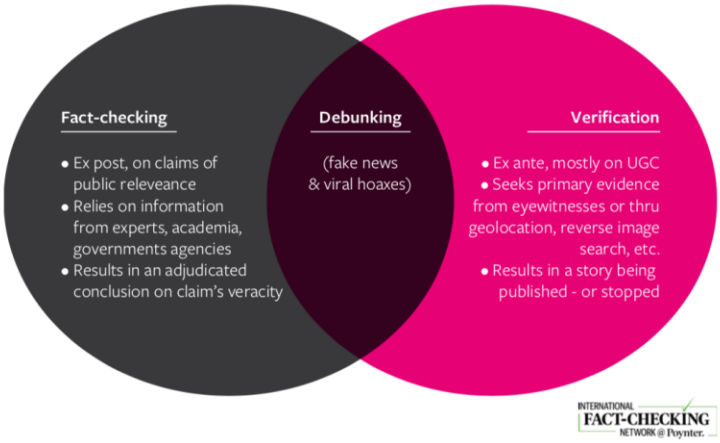

Figure 5 The difference between Fact-checking and verification from Ireton & Posetti. 2018

Generally speaking, fact-checking is composed of three phases (Ireton & Posetti, 2018):

1) Finding fact-checkable claims by scouring through legislative records, media outlets and social media. This process includes determining which major public claims (a) can be fact-checked and (b) ought to be fact-checked.
2) Finding the facts by looking for the best available evidence regarding the claim at hand.
3) Correcting the record by evaluating the claim in light of the evidence, usually on a scale of truthfulness.

3.2 Towards Standardisation

One of the most important decisions that has been made in order to converge both in the direction of coordination (e.g. data exchange) and ability to increase the reach of the fact-checking articles is **standardisation**.

This standardisation process was born from the needs of different players: search engines, researchers, fact-checking organisation and also social media platforms. It is needed to have a common format for describing who fact-checked what, to be then able to put together data coming from different fact-checkers and make it searchable and integratable into different platforms.

The standardisation process, lead by Google and Schema.org, brought the new **ClaimReview**[25] schema.

The schema standard was developed by the fact-checking community, but takes many of its properties from other digital standard tools for reviewing restaurants and movies, and for describing digital content (publisher, dates, creative licenses, authors, etc.).

As explained in the following table 1, this data format allows to describe the output of the fact-checking process, presenting features of the claim that is being reviewed as well as the classification and the author of fact-checking.

Property name	Type	Usage
Claim Reviewed	text	A short summary of the specific claims reviewed in a ClaimReview.
Item Reviewed	Thing or CreativeWork	The item that is being reviewed/rated. It also contains a property *author* that identifies the author of the claim.
	Claim (pending[26])	The new type allows to describe the *appearances* and *firstAppearance* of the claim, belonging to the type CreativeWork and recently chosen to hold the URLs containing the claims[27]
Review Rating	Rating	A rating of the claim. Can be expressed with free text in the subproperty *alternateName* or quantified by using a *ratingValue* between *WorstRating* and *BestRating*.
url	URL	The location where this claimReview is published, together with a full article published by the fact-checker
Author	Organisation or Person	The author of this review
Date Published	Date	The date when the review has been published

[25] https://schema.org/ClaimReview
[26] https://pending.schema.org/Claim
[27] https://github.com/schemaorg/schemaorg/issues/1828

Table 1 Main Properties of ClaimReview Standard

Being a quite recent standard, there are still parts of it under discussion (as for example the pending *Claim* type). And there are also inconsistencies of how the different fact-checking organisations use it. For example, to indicate an URL to the claim that is being reviewed, the most common solution before the proposition of the *Claim* type was to indicate it in the nested field *ItemReviewed.author.sameAs* that instead is better suited to contain informations about the author of a certain claim (e.g. a URL pointing to the twitter profile of a public figure, or a wikipedia page in order to be able to recognise and link the entity). Or another inconsistency when indicating the claim reviewed is to treat not consistently the several appearances of the same claim: an appearance can support the claim or already reviewing and debunking it. And these inconsistencies make it problematic to recognise URLs that have been already fact-checked, resulting in dirty data raising the need of manual re-annotation, as in (Monti, Frasca, Eynard, Mannion, & Bronstein, 2019) where the authors had to rely on Turk annotators to recognise the appearances of the claims.

As of today, the ClaimReview standard is starting to empower a wide variety of platforms. The first to use it was Google[28] in the search engine, that allows the users to face the misinformation right at the moment when they are potentially being subject to misinformation performing a search that is similar to a claim that has already been reviewed. This feature has been temporarily removed because of complaints from The Daily Caller and other conservative news outlets[29]. Once again, the politics of misinformation and fact-checking make it difficult to negotiate what evidence will be considered in making such determinations. Youtube has recently started experimenting with fact-checking features on the Indian region, focusing on coverage topics that are particularly vulnerable to misinformation in the area[30]. Another platform that is now making use of ClaimReview items is Facebook[31], that engages with the user in the publishing phase (by asking confirmation before sharing something that has been fact-checked) or in the consumption (presenting "related stories") next to the debunked stories.

And even for research and journalists there are growing datasets[32], feeds[33] and platforms that allow to retrieve and search the ClaimReviews.

[28] **https://www.blog.google/products/search/fact-check-now-available-google-search-and-news-around-world/**
[29] **https://www.poynter.org/fact-checking/2018/google-suspends-fact-checking-feature-over-quality-concerns/**
[30] **https://www.poynter.org/fact-checking/2019/youtube-is-now-surfacing-fact-checks-in-search-heres-how-it-works/**
[31] **https://www.facebook.com/help/1952307158131536**
[32] **https://datacommons.org/factcheck/download**
[33] **https://storage.googleapis.com/datacommons-feeds/claimreview/latest/data.json**

One notable example is Google FactCheck Explorer[34]. It can be used on one side to annotate and submit ClaimReviews to the platform, intended for journalists to submit their work and make it available. Every user of the platform can submit ClaimReviews but only the authors that are approved by Google (using the IFCN list) will be considered and retrievable from other users. The second feature of the platform is the explore: users can perform queries based on text or get the most recent fact-checks selecting also the language. And the third feature, that has been very recently added, is an API to retrieve the ClaimReviews.[35]

[34] https://toolbox.google.com/factcheck/about
[35] https://developers.google.com/fact-check/tools/api/

4 Misinformation Detection Ecosystem

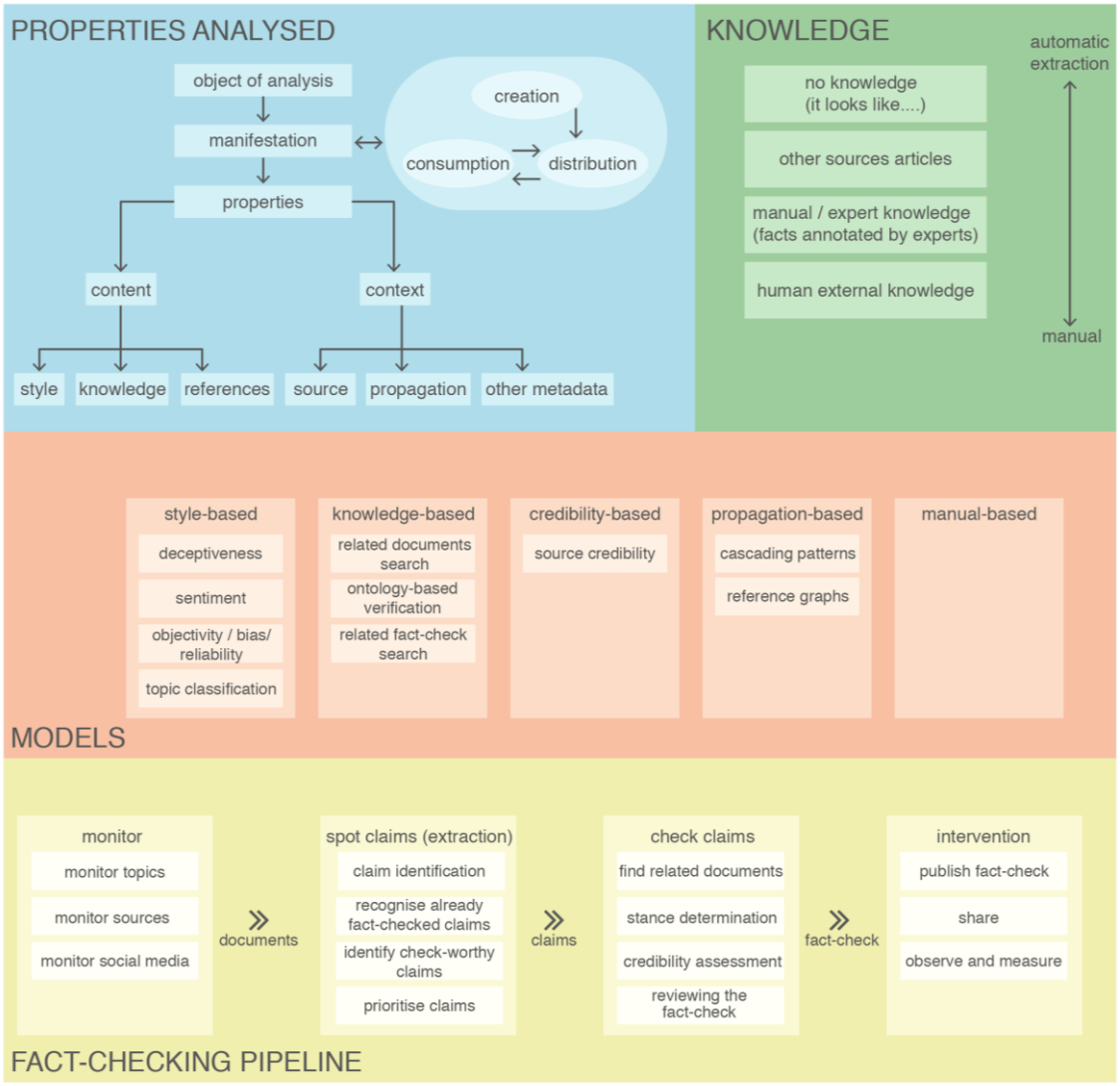

Figure 6 An overview showing different components of the misinformation detection ecosystem

The next part of this deliverable involves looking more deeply into the different detection approaches that we have begun to reference in sections 1 and 2, and demonstrating how they fit into the misinformation ecosystem. The ecosystem includes all of the contextual features described in section 2 (visible in the upper left, blue square of the diagram in Figure 5). However, misinformation detection tools can also be mapped to the different publishing, monitoring, fact-checking and analysis procedures, different models of analysis, and how the methods use knowledge (for example, from experts) We have tried to capture this complex environment in Figure 5, Figure 7 and Figure 8 (which is Figure 5, but including relationships and connections). The following subsections relate back to their various positions within this figure, to help the reader understand which processes and procedures are connected to them.

Considering the wide variety of detection methods, we have to consider many different perspectives and give some common descriptive attributes in order to see the general picture. Starting from the lifecycle of content propagation (independently from the genuineness of the item), we outline first the general steps of the processing pipeline for fact-checking (4.1). Then, we examine the different types of properties that are available in each stage (4.2), describe some general properties of the detection itself, like the granularity at which it is applied (4.3), and where it retrieves the external knowledge (4.4). Then we are seeing which are the most common types of analysis grouped by feature types (4.5). We will provide a table with concrete examples in section 5. This analysis refers generally to the automated techniques but can also be applied to the manual verification techniques and also contains some references to it.

4.1 The Fact-checking Pipeline

For fact-checking, responding to misinformation either in manual, automated or mixed ways is a process that Babakar & Moy (2016) described as a pipeline. This process is visible in the light-green-coloured block at the bottom of Figure 5. It starts from monitoring public spaces, and then moves to spotting, selecting and sorting suspect claims, checking them and then intervening. Misinformation detection is part of this process, concentrated in spotting and sorting claims.

Monitoring refers to defining a set of criteria, such as sources or topics, to collect documents (articles and posts). The collection can happen from media (TV, newspapers), social media, etc, by using specific platform APIs, web-scraping, news aggregators, etc. In section 1, we outlined the connections that monitoring has within the work of Work Package 3, and more specifically D1.1, D2.1, D3.1., D4.1 and D4.2. D1.1 described the topics and triggers of interest for the Co-Inform platform. D2.1 described the policies by which different sources of information (described in D3.1) would be handled on the platform. D4.1 and D4.2 help us to visualise the architecture of how these various components will come together in the system. Now, with D3.2, we share some of the different approaches we may use to determine how the data is processed.

Spotting claims, according to (Babakar & Moy, 2016), involves different tasks that can be automated. First of all it is necessary to distinguish claims that have already been fact-checked from the new claims made. And the new claims need to be identified with certain properties. A claim needs to be verifiable and for this reason different types of claims have been defined by (Konstantinovskiy, Price, Babakar, & Zubiaga, 2018): personal experience, quantity in the past or present, correlation or causation, current laws or rules of operation, prediction, other type of claim, not a claim. And the claim has to be check-worthiness (Zuo, Karakas, & Banerjee, n.d.), also by prioritising the needs of the citizens considering the current spread and importance of its topic.

All this has to be done by considering the complications that come from the use of natural language: the same concepts can be expressed in different ways and it is not very simple to recognise the claims along different linguistic variations.

Checking claims, the third stage of the pipeline, analyses the veracity of the claims.

The methods for that can be automated, manual or hybrid, relying on a wide variety of features and models as detailed in later sections. Given the level of accuracy of existing methods, the validation of automated results by an expert is still required. The automated tools can give indications to the human judges who, based on the explanations and evidence provided, performs the checking and review the claim.

Given the possibility of having unexpected predictions by the automated models and given the time required by purely manual fact-checking, the hybrid collaborations seems to be the strategy that promises best results (in terms of accuracy and throughput).

Interventions are the last step of the fact-checking pipeline. This is the most critical in terms of effectively promoting the fact-checked information. The way the information is published and the explanations regarding the process and criteria are essential to avoid backfiring and stimulate critical thinking. In D2.1, we described some of the possible interventions that the Co-Inform platform could consider. We will be researching those ideas within the context of D3.3 and D3.4, to understand more about what interventions have the most success in bringing about behavioural change.

4.2 Object of Analysis

The data available to be retrieved and analysed during the misinformation detection differs according to stage of propagation of a piece of news. The small diagram on the top left corner of Figure 5 illustrates this 3-stage propagation cycle, which is detailed in the Figure 6 below. **Creation, distribution and consumption of news** are the three stages of this cycle, with a possible looping from consumption to distribution and consumption to creation.

Creation takes place on news websites in the shape of articles, or on social media in the shape of posts or also on online forums. Distribution considers how the users reach the created content. It may be by manually visiting the source, or by having news feed that automatically aggregate from other sources, or private messages (e.g. a link on WhatsApp) or emails, or by being mentioned/tagged on social media. For example, Facebook's news feed used to prioritise content that produced a lot of "clicks" ("clickbait"). Clickbait, however, is designed to promote engagement, so it says absolutely nothing about the value of the content. Facebook has supposedly attempted to shift it's newsfeed to include less clickbait[36].

[36] https://newsroom.fb.com/news/2014/08/news-feed-fyi-click-baiting/

This stage is usually non observable by third parties but just the platform owners. The only exceptions are public posts like twitter mentions.

The consumption is only observable when users reply or share publicly. The consumption can be performed both by citizens (e.g. on social media or personal blogs) or by other news agencies (this is possible when a news agency cites another source or a post from social media). Consumption can also have different orientations such as support, deny, query, comment (Gorrell et al., 2018).

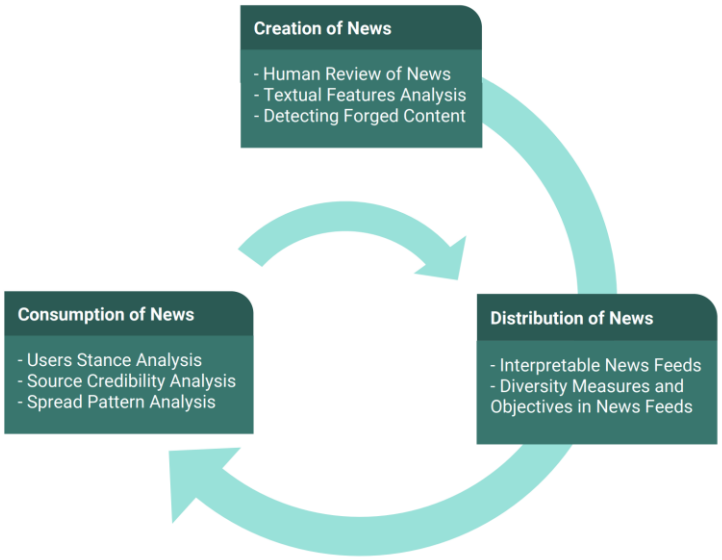

Figure 7 A summary of fake content and detection methods at different stages of the news life cycle (Mohseni & Ragan, 2018)

In the Table 2 below, the properties available at each stage of the new propagation are described.

	Properties available
Creation	- The text of the content - Creation date - Author - Context: which website, which group is addressed to (e.g. it is a post in a facebook group, or on a certain facebook page, or on a specific reddit)

	- Attachments: links, images, videos
Distribution	- Actors involved (source, destination) - Distribution date
Consumption	- Consumer - Original content - Added comment/content (if any) - Date of the action

Table 2 Properties that can be detected at different stages of the news life cycle

4.3 Detection levels

In general, misinformation detection occurs at different levels that are either **source-based**, **document-based** or **claim-based** all using varying detection methods (Figure 5). Source-based methods tend to be simple and focus on identifying if an information source is credible based on previous source behavior (i.e. if the previous source claims have been trustworthy). Document-based approaches check if a document contains misinformation. Depending on the complexity of the information judgement method, the approach can be shallow and only look at document source or can be much more complex and investigate internal document claims as well as related claims from other sources. Finally, claim-based methods start from a textual-claim and investigate individual claims rather than documents.

4.3.1 Source-level detection

As a general level of granularity, this approach consider the source associated to an information that needs to be checked. A source can be both a **news outlet** (corresponding to a specific domain name from its website) or a **social media profile** (owned by a citizen, a company, automated account) or other **websites/blogs/forums**. Analysis on this level consider **domain names**[37][38][39] or social media profiles[40] based both on manual (compliance with some rules) and automated assessments (predictive models). The assumption that lies beneath all these analysis is that a certain source is more likely to behave as in the past. Generally, the approach does not account for the differences between singular instances (e.g. a reliable website can potentially publish a misinforming content, or the opposite) and

[37] http://www.opensources.co/
[38] https://www.newsguardtech.com/
[39] https://www.mywot.com/
[40] https://botometer.iuni.iu.edu/

provide a generic measure (credibility, bot-likeness) of the source (see 4.6.4 **credibility-based** for a description of how this analysis is performed).

4.3.2 Document-level detection

A more detailed analysis level compared to source analysis. This approach analyse documents as a whole rather than individual sentences or arguments. Typical approaches rely on available annotated documents datasets[41][42][43] that present labelled document instances (or URLs). Even with long texts, it is possible to have an idea of the document veracity and credibility as a whole. The main limitation of such approach is that a document may contain multiple statements about different information and contain a mix of true and false information.

4.3.3 Claim-level detection

Sometimes though it is needed to break down the document into more fine-grained pieces as documents may contain information that are both trustworthy and misinformation. This is where claim-level analysis is necessary. A claim can be traced to the field of argument analysis and defined as an '**assertion that deserves attention**' (Daxenberger, Eger, Habernal, Stab, & Gurevych, 2017) it is therefore important to determine which claims are check-worthy (Zuo et al., n.d.). Claim-level classification requires multiple steps such as claim extraction, stance determination and credibility assessment (Figure 7). Claim representation has been loosely standardised in the ClaimReview standard, where the field ClaimReviewed is of type text and is the piece of information that is being reviewed. Fact checking agencies also describe in detail claims as the part of the reviewed sources/documents that are true or false, also by having multiple ClaimReviews embedded in the same page.

[41] https://www.kaggle.com/mrisdal/fake-news
[42] https://www.kaggle.com/mdepak/fakenewsnet
[43] https://www.kaggle.com/jruvika/fake-news-detection

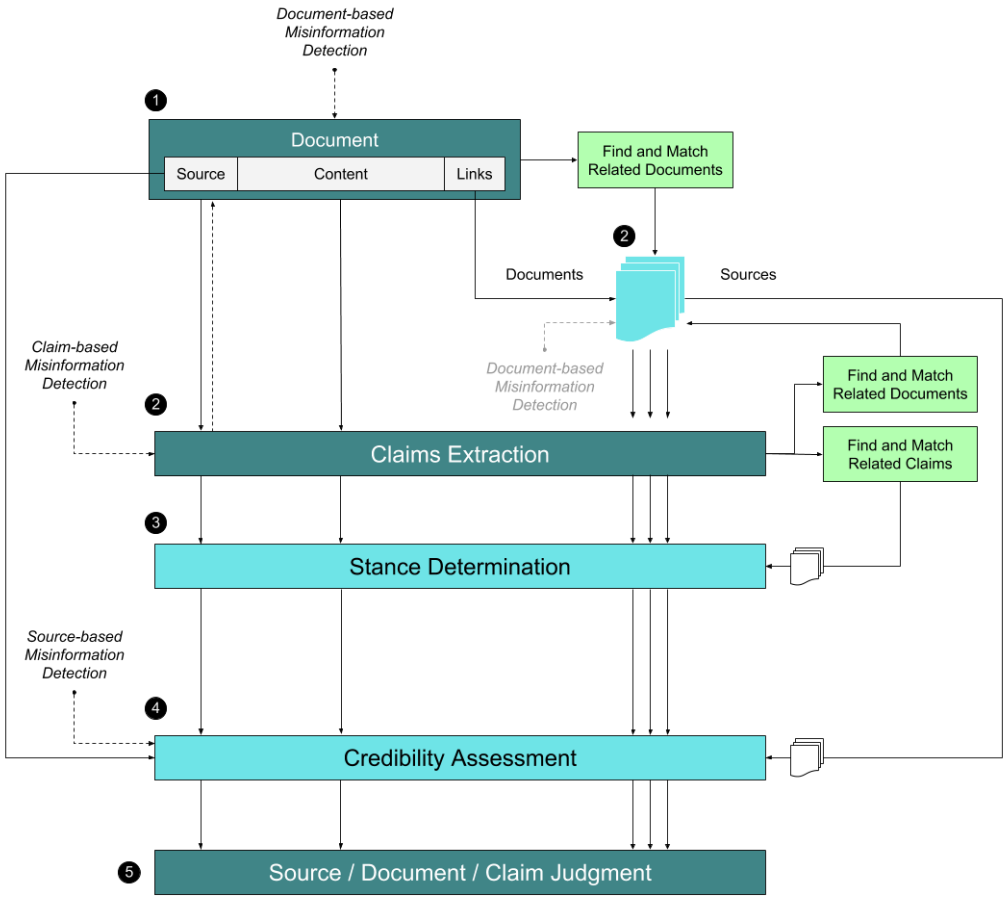

Figure 8 Misinformation Detection Process based on Popat, Mukherjee, Strotgen & Weikum, 2017

4.4 Detection workflow

Multiple steps are required for judging documents, information sources and claims (Figure N) with some methods more thorough than others. In a **document-oriented situation** the most thorough approach typically follows the following schema:

1. **Document analysis** First, a document is divided into three sub-components that are then analysed in more details in the following steps: a) The source or author of the document is extracted; b) The content of the document, and; c) Links or references to related documents (e.g. cited documents or documents citing the analysed article).

2. **Claim extraction** The second step that is sometimes omitted when documents as a whole are analysed rather than individual claims is the claim extraction phase. In this phase, claims are extracted or compiled from referred documents or the previous source claims. The importance of claims may be also considered by comparing their similarity with the currently analysed claim.

3. **Stance Detection** The third step that is also sometimes skipped is the identification of the stance associated with a claim. The aim of this part of the process is to find if

a claim is supported or unsupported and is critical for obtaining a better assessment of the considered document or claim.

4. **Credibility Assessment** After determining the stance of the document claims and the related claims, the credibility of the supported and unsupported claims is assessed. This may also include credibility information from the document or claims authors.

5. **(Mis)Information Judgment** Finally, a judgment is performed to assess if the document contains information by merging and weighting the different credibility assessments for each document and related claims.

Depending on the complexity of the information judgement method, the approach can be shallow and only look at the source, or can be much more complex and investigate on the document level or even breaking it down into more detailed claim-level analysis. In this subsection we are presenting the different levels of granularity that can be applied.

4.4.1 Document Analysis

Many approaches for identifying misinformation start from documents. In the context of online misinformation detection, document URLs may be the first part of the analysis where documents are first fetched using their URL and then divided in different components such as (1) the source where the document source come from (typically inferred from the URL domain), (2) the textual content of the document, and; (3) the links to and from external documents by using hyperlinks within the document and search engine results linked do the document.

The source of the document is then analysed using credibility assessments results from the document previously analysed from the same source. For the links extracted or linked to the document, document analysis may be also performed on them. When analysing documents, an existing labeled document database may be also used so that previously annotated documents are simply matched.

4.4.2 Claim Extraction

Claim extraction is the task that extracts statements from documents. This step may involve (1) the identification of content that is potentially misinformation or trustworthy, and; (2) the ranking of important statements that needs to be analysed. This phase can be linked with research in argument mining (Lippi & Torroni, 2016) where the aim is to identify relevant document section from documents that can be potentially misinforming.

In the area of misinformation detection, claim detection mainly focuses on the identification of check-worthy claims in documents. A common approach is to use labelled sentences and machine learning algorithms for identifying claims in documents followed by ranking

methods for identifying fact-worthy documents (Hassan, Arslan, Li, & Tremayne, 2017) (Zuo et al., n.d.) (Hassan, Nayak, et al., 2017). Other methods rely on more static approaches that simply rely on rules and keywords for identifying important sentences (Zuo et al., n.d.) (Hassan, Nayak, et al., 2017). FInally, approaches that rely on pre-existing claim databases have been also used, usually as part of the previous approaches. These methods are mostly based on similarity metrics between the claims that are already in the database with the investigated sentences (Popat et al., 2017) (Zhi, Sun, Liu, Zhang, & Han, 2017a) (Hassan, Nayak, et al., 2017). This last approach may be more precise but miss new claims.

4.4.3 Stance Detection

Stance detection can be defined as a reaction to a claim made by a primary actor. Therefore, this step aims at identifying if a particul statement is supported or unsupported in the analysed document. Stance detection is particularly important in reporting articles such as news documents where journalist may report claims that are not their own but then provide support or non supporting arguments. When stance detection is omitted, misinformation detection may report incorrect predictions due to a mismatch between the document claims and their stance.

Approaches for stance detection has been investigated in the Fake News Challenge[44] and the SemEval tasks and generally involve a mix of machine learning methods coupled with the detection of cue words (e.g., refuse, wonder, confirm, etc..) in order to determine if a claim is supported or unsupported (Bahuleyan & Vechtomova, 2017) (Ghanem, Rosso, & Rangel, 2018) (Popat et al., 2017).

4.4.4 Credibility Assessment

Credibility assessment aims at identifying if a claim can be trusted.This is usually done based on previous source credibility track record or based on propagation methods where credibility is assessed based on similar claims or referenced/referring graph (i.e., credibility of the linked and linking document sources).

Work on credibility assessment is mostly related to the identifying if the source of a claim is credible. Approaches varies from analysing the credibility of the claim previously posted by a given source or other source of knowledge (Section 4.5)

4.4.5 (Mis)Information Judgment (Verification)

[44] http://www.fakenewschallenge.org/

The last part of the process is to decide if a document is misinformation. This step depends on the type of method used for combining the different individual credibility assessment of individual claims, related claims and information sources. The approach are typically fitted to a particular type of misinformation (Section 2.1).

4.5 Sources of knowledge

When it comes the analysis of a piece of information, there are different levels of sources to assess its veracity/credibility/informativeness.

The knowledge to classify information could come in different ways. Here we present the analysis by considering increasing levels of manual human knowledge (ranging from completely automated without handcrafted knowledge, passing through automated assessment with handcrafted knowledge bases, arriving to the completely human expert knowledge).

Here we give a description about the collateral knowledge that is used instead of describing the models using it (see 4.5 for the models).

4.5.1 No knowledge

The first and most shallow source of knowledge is when there is no actual verification of the contents. This can happen in different cases:

1) style-based analysis (4.5.1): in this case the models rely on the correlation between the usage of linguistic patterns and the informativeness of the item under consideration.

2) context-based: credibility-based (4.5.4) and propagation-based (4.5.3) that, in their base implementation (can be combined with other sources of information too), predict the informativeness without even looking at the content. This will be explained in detail in the models subsection, but in general relies on the fact that misinforming content can be recognised by how it spreads (cascading pattern analysis), which sources it cites (reference graphs) or by the credibility of the actors involved (publishers and spreaders).

4.5.2 Cross-comparison

Moving to approaches that actually consider the knowledge contained in the contents, we can have different sources of knowledge. To the group of cross-comparison belong methods that assess the truthiness of claims by using different news articles, coming from different sources (with their relative credibility scores) (Bountouridis, Marrero, Tintarev, & Hauff, 2018; Nguyen, Kharosekar, Krishnan, et al., 2018; Zhi, Sun, Liu, Zhang, & Han, 2017b), covering the same topic. To have models that can assess the veracity of the analysed items, it is very important to understand whether the other sources are denying or supporting the current one. With this kind of information it is possible to predict the correctness of the current claim analysed and also to have evaluations of inter-rater agreement of the sources themselves.

Another category of approaches that is based on comparison of opinions is the one based on stance detection. As described in the RumourEval challenge (part of SemEval)[45], the task A is tracking how other subsequent posts orient to the accuracy of the rumoured story (Support, Deny, Query, Comment).

4.5.3 Manual/Expert Knowledge

Moving away from automated extraction of facts, there are methods that consider knowledge coming from manually curated sources of information. These sources can manifest under different shapes (ontologies that store facts, lists of manually verified articles and claims). With this kind of knowledge it is possible to create models that are able to find direct matches with stored facts and previous fact checks, but also to enable a process of automated reasoning based on derivation rules (Gad-Elrab, Stepanova, Urbani, & Weikum, 2019).

There are mainly two types of expert knowledge-bases that we are considering:

1. Ontologies of common knowledge: manually curated set of known facts
2. Fact-checking based: from the claims that have already been fact-checked by expert and certified organisations (e.g. valid signatories of IFCN), it is possible to recognise instances of claims that have already been debunked (thanks also to the standardisation of the ClaimReview schema)

4.5.4 Human external knowledge

Then the last step towards manual verification is the human knowledge. With this group we consider all the work that is done by journalists to retrieve additional information, as evidence or argumentations, when they need to fact-check a new claim. This is the most complex source of knowledge and it is the one that assures the highest level of reliability.

4.6 Types of models

Going deep in the details of the models used in the fact-checking pipeline, a first classification of the methods can be done by considering the perspective that is taken: different families of methods exist and they focus on different features and hypothesis. The coarse-grained classification that we apply here is the one also presented in (Potthast, Kiesel, Reinartz, Bevendorff, & Stein, n.d.) and (Zhou & Zafarani, 2018).

As highlighted in Figure 6, from the propagation lifecycle we can retrieve different types of features that can be grouped in:

[45] https://competitions.codalab.org/competitions/19938

- **Content:** textual content of articles (headline, full text) and of successive propagation steps (comments, retweets) as well as references to other entities (other news articles, users, websites). The content can also include attachments like media (images, videos) or generic files.
- **Context**: features that come from the surrounding information about the document. The most important are its source (who published this document) and the propagation (intended as the set of user posts/reactions and other documents that contain references to this item).

In the following paragraphs we are analyzing them specifically.

4.6.1 Style-based

Starting with approaches that focuses on the content, we have a first group that is basing the classification on the stylistical surface. The style can be defined as a set of linguistic features for text, or visual features for attached media. The theory that lies behind is that true stories/images and fabricated ones can be recognised by how they appear, because its intent to deceive or imitate.

For the analysis of text, usually the useful methods for misinformation are able to predict deception (W. Y. Wang, 2017), credibility indicators (reliability, bias, objectiveness, subjectivity, readability) (Horne, Dron, Khedr, & Adali, 2018), sentiment (Go, Bhayani, & Huang, n.d.) and topic analysis (S. Wang & Manning, 2012). These models are able to provide measures across several dimensions that can be used to predict the probability of misinformation[46].

To the style-based category belong also models that analyse media for detecting whether they are counterfeit or not. Beyond simple reverse search methods, there are studies (Li & Lyu, 2018; Ruchansky, Seo, & Liu, 2017) and tools that focus on finding forgery and deep fakes.

A limitation of these methods is that misinforming content can be really similar to genuine content from the stylistical point of view.

4.6.2 Knowledge-based

This other family of methods, instead of stopping at the stilistic surface, relies on the extraction of the semantic content from the document. The entities and their relations extracted from the claim that is analysed are compared against a Knowledge-Base that contains the ground truth (Ciampaglia et al., 2015; Shi & Weninger, 2016). The Knowledge-Base can be one of the many ontologies publicly available (such as DBPedia, Cyc,

[46] http://nelatoolkit.science/visualizationtoolkit

ConceptNet) or be appositely built by extracting knowledge from news articles or by comparing to other news articles on the same topic (Nguyen, Kharosekar, Lease, & Wallace, 2018) (more coverage with recent events, but less reliable).

The comparison is done by having two main processes: *entity linking/resolution* that maps the entities found in the claim with the entities in the KB, and the *reasoning* that tries to verify or deny the existence of relationships identified in the claim by using the known relationships in the KB.

This type of method can cover the problems of style-based approaches since they don't consider the style but are based on the semantic level. However, they are very sensitive to errors during entity linking (complications such as co-referencing) and reasoning. Furthermore they are limited on what is contained in the KB and unknown entities and relationships are very difficult to manage. On a socio-cultural level, these types of approaches are complicated. Our partners at Fact-check NI shared some experiences with us about how fact-checks that initially seem straightforward become "an analysis of a series of version of events" that make it difficult to determine what actually occurred. One example they shared with us involved the claim that "city deals" were part of the Democratic Unionist's Party deal with the Conservatives.[47] After publishing an initial verdict of false to this claim, the DUP produced additional documentation in support of their claim, which lead FCNI to eventually relabel the claim as unclear. Negotiating what is known, accessibility and acceptance of evidence remains a significant issue for such types of fact-checking methods.

4.6.3 Propagation-based

Moving instead to context-based methods, there is a wide variety of studies that focus on the identification of cascading patterns and the difference between those generated by genuine content and misinforming content (Del Vicario et al., 2016; Shao, Hui, et al., 2018; Vosoughi, Roy, & Aral, 2018b), also considering the role of specific nodes like bots (Shao, Ciampaglia, et al., 2018).

Although this is usually relative not to Misinformation Detection but to Misinformation Flow Analysis Prediction, there are some methods that exploit the diffusion features and discriminators of misinforming content to classify and distinguish misinformation (Monti et al., 2019).

The assumption behind these models is that the misinformation spreads following different patterns than news.

[47] **https://factcheckni.org/facts/were-city-deals-part-of-dups-confidence-and-supply-agreement/**

When considering the propagation, for example, there come into play methods that analyse the stance of the comments and retweets (Krejzl, Hourová, & Steinberger, 2017) (Baly et al., 2018). Stance detection is also used on article headlines to understand agreement between different sources (Bourgonje, Moreno Schneider, & Rehm, 2017). Another type of propagation is the one that considers the references belonging to an article in order to assess the credibility of it. This has been applied, for example, in assessing the credibility of articles on vaccination based on the references they provide (Shah et al., 2019).

4.6.4 Credibility-based

A last family of approaches tries to detect misinformation using indicators of credibility of the different actors involved, such as news outlets and online accounts, that are involved both in publishing, consuming and spreading (mis)information.

The credibility analysis is based on information that are both news-related and social-related. This method overlaps with the propagation-based, but the difference is that it does not consider the content and the social relationships. The current studies focus on the credibility of:

- News headlines: this task can be reduced to detecting clickbaits. The main purpose of clickbaits is to attract users by using authority and sensationalism on the target page that then contains or fake content or unrelated content. Current detection studies are using a wide variety of linguistic features (Biyani, Tsioutsiouliklis, & Blackmer, 2016; Bourgonje et al., 2017). As mentioned previously however, what is determined to be clickbait is not an absolute feature either. The example of NewsGuard and the Daily Mail, mentioned previous, is evidence of this. The editor of the Daily Mail was able to successfully argue in his favor that the language used in the Daily Mail, despite how it may be perceived, did not differ substantially from many other sources that NewsGuard had listed as “green” (good).
- News sources: this task is done by predicting the credibility of a certain source by analysing a set of content and link features (Esteves, Reddy, Chawla, & Lehmann, 2018)
- News comments: review spam detection (Dungs, Aker, Fuhr, & Bontcheva, 2018). This overlaps with propagation-based, with the difference that the comments are used to model the credibility of the subject of analysis (as in task A of RumourEval). This analysis includes also identifying the participants and their relationship with the spread (malicious and naive users, non participants).

As concrete examples of this group, we have tools that provide evaluation of:

- News outlets: assigning “nutrition labels”[48], labeling domains with credibility indexes[49][50] and looking at business registry for the companies that are mapped to that domain[51][52]
- Social profiles: analyse profiles to predict if they are bot or not (see 5.1) and if they spread misinformation or not, etc.

As it can be seen, there is an overlap with the other types of models. The main difference is the perspective: this set of models characterises the credibility of the different actors, that can be news outlets (with the corresponding domains), and online accounts (spreaders, agencies and bots). They are based on the assumption that a certain actor is more likely to behave similarly as in the past.

This assumption is biased and can be limiting because is is acting on an average over the shared posts, without exploiting the differences.

For example, a certain profile/outlet may be very expert in a particular topic and very naive in another one but the current models do not explore this feature (at the time of writing).

4.6.5 Summary

The figure below summarises the entities, processes and relationships that have been described in the subsections above, by illustrating with arrows the different connections between them. One can see from the diagram the different dependencies and focus areas.

[48] https://www.newsguardtech.com/
[49] http://opensources.co/
[50] https://www.mywot.com/
[51] https://bolagsverket.se/om/oss/nyheter/arkiv/2018/en-uppgift-en-gang-pa-ett-stalle-1.17178
[52] http://www.toop.eu/assets/custom/docs/TOOP_brief.pdf

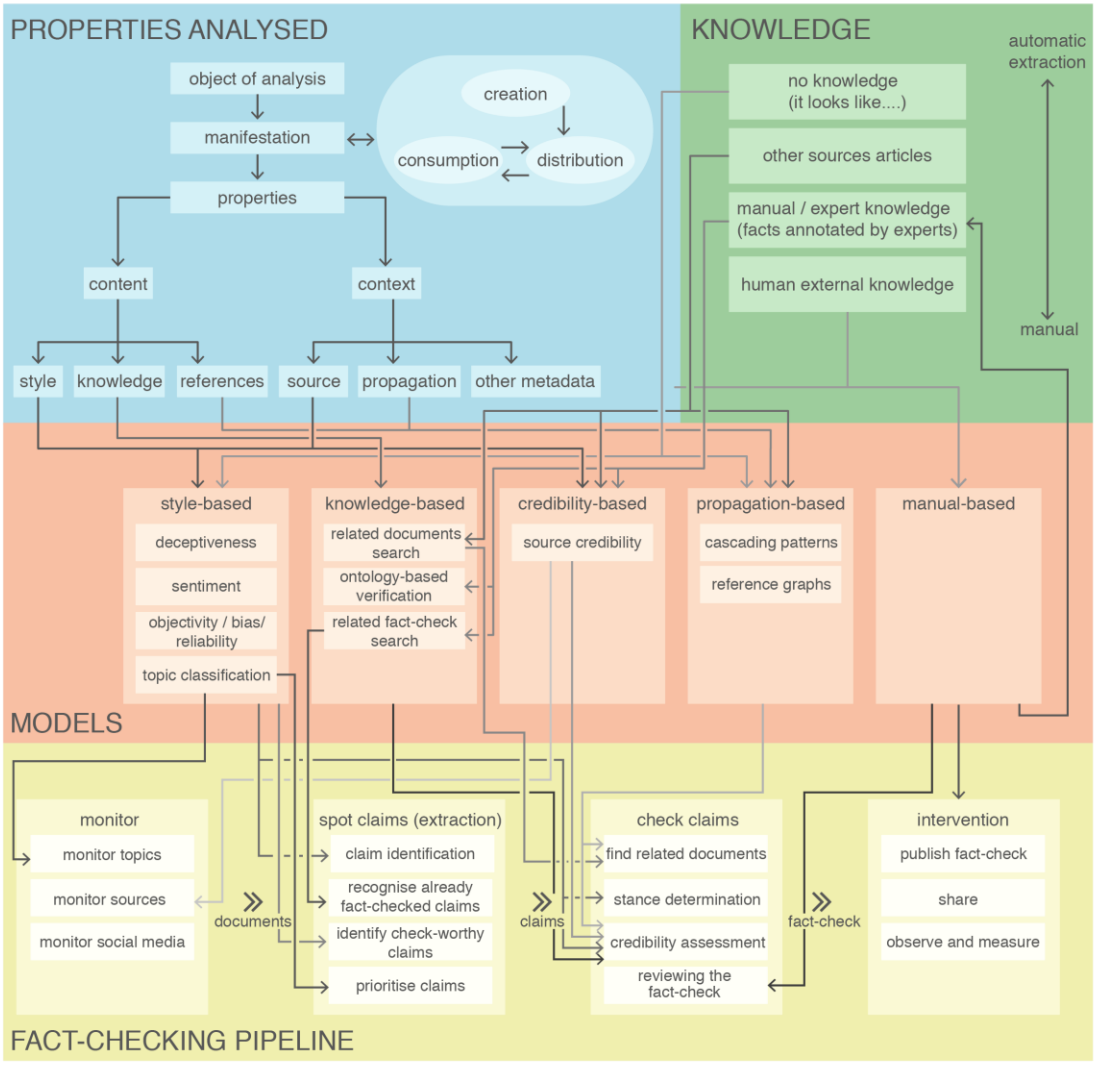

Figure 9 An overview showing the different relationships and processes involved in misinformation detection, including the fact-checking pipeline, the different models, the objects analysis and the external knowledge used

5 Tools and Algorithms

Reflecting the diversity of methods and approaches to tackle misinformation, tools are deployed in many different ways, some of which are accessible to the general public without much technical knowledge and others that require in-depth knowledge of computing. In addition, some tools have been tested in limited environments and may only be available as a **demonstration**. Complete **web platforms** offer the most accessibility to the general user and can include different **subservices for verification** of images, video, identity and the source, which also exist on-line, on their on domains. **Browser plugs-ins** or **applications**, though they require installation, are also suitable for regular users and professionals to deploy. At the other end of the spectrum are tools such as **APIs**, and core **machine learning** processes that are not typically stand-alone tools, but part of larger structures or research initiatives.

This subsection focuses on representative examples of tools according to the basic type of detection method on which they rely. We will also explain their relationship with the different perspectives that have been described before, such as when such tools are used in the fact-checking pipeline and how their approaches are dynamic and evolve.

5.1 Credibility-based: Assessment of profiles (Botometer)

The Botometer tool[53] analyses twitter profiles in order to evaluate their probability of being bot accounts.

It focuses on different features:

- English specific: sentiment analysis and content
- Language-independent: friend, network, temporal, user

It can be used also to analyse followers or friends on twitter.

This tool comes in our analysis as a source-level assessment (profile-based), with models that are based on patterns without using external knowledge. In the pipeline of fact checking it is not applied directly but its result can be used when analysing the propagation of the claims in order to enrich the study (as has been done for example in (Shao, Ciampaglia, et al., 2018).

This tool is mainly based on a credibility-based model, assessing the role of a certain actor in the ecosystem of misinformation. But it also integrates style-based features coming from the content created and shared.

5.2 Style-based: Rumour detection with ELMo embeddings

[53] **https://botometer.iuni.iu.edu/**

Determining veracity of social media content is crucial task if authorized evidence is absent. Stance analysis on discussion threads towards a rumour gives a hint for identifying its veracity in early stage. (Qazvinian, Rosengren, Radev, & Mei, 2011) introduced rumour stance classification of each tweet as supporting, querying or denying a rumour. (Procter, Vis, & Voss, 2013) expanded the annotation scheme with comment label that is used if a post is either unrelated to a rumour or does not contribute for identifying veracity of the rumour.

RumourEval (Gorrell et al., 2018) is SemEval shared task aiming determining veracity of a rumour and stance classification of posts towards the rumour. First RumourEval(Derczynski et al., 2017) was held in 2017 and attracted researchers in this field. RumourEval task consists of two subtask: stance detection and veracity. The aim of stance detection is to classify the type of interaction between a rumorous social media post and a reply post as support, query, deny, or comment. The aim of veracity prediction is to determine the veracity of rumour post that initiates the discussion. The winner system on stance detection, Turing (Kochkina, Liakata, & Augenstein, 2017), employed branch-LSTM sequential classifier where stance label of each tweet depends on the features and labels of the previous tweets. NileTMRG (Enayet & El-Beltagy, 2017), winner on second task, used predictions from the task of stance detection along with additional hand-crafted features as feature of linear support vector machine for determining veracity. Recent attempts focus on jointly learning stance classification and veracity prediction(Kochkina, Liakata, & Zubiaga, 2018; Ma, Gao, & Wong, 2018)

We have participated in both subtasks of RumourEval 2019 to evaluate initial version of our system, namely CLEARumor (Baris, Schmelzeisen, & Staab, n.d.). After pre-processing and embedding the posts with ELMo (Peters et al., 2018), CLEARumor runs CNN layer with different kernel sizes, averages all resulting vectors, combines those with auxiliary features (e.g whether post is from Twitter/Reddit, number of followers/friends, ratio of followers/friends), and then feeds resulted features into a MLP for stance classification. Average of these estimations are concatenated with further auxiliary features (e.g presence of media, upvote ratio of source post if it is Reddit) and fed into an MLP block for predicting veracity. Our official submission achieved second rank on task of determining veracity[54]. After competition, we continued to tune hyperparameters of our architecture and observed improvements on scores of both subtasks.

The reason because this approach is classified as *Style-based* is that this approach is based on ELMo representations, that contain syntactic and semantic information. The semantic is not explicit, with entity recognition and linking, but is a shallow implicit one, derived simply from distributional semantics. It does not qualify to be classified as Knowledge-based because there is no entity extraction and any ontology matching/reasoning. Due to the task

[54] **https://competitions.codalab.org/competitions/19938**, CLEARumor is “ukob-west” on the list.

A, that is relevant to stance detection, this approach also presents propagation-based characteristics.

5.3 Knowledge-based: ExFaKT

Belonging to the knowledge-base group of models, (Gad-Elrab et al., 2019) is an approach that provides validation for claims based on Knowledge Graphs and Text. Together with the verification results it also provides explanations, in the form of semantic traces, that come both from the automated reasoning applied on the considered sources. This tool is not yet deployed as public demo, but its source code and evaluation datasets can be found on the Max Plank Institute website,[55] as well as a poster illustrating the model.[56]

Starting from the input text to be verified (claim), it is converted into a query constituted of entities and relationships (a predicate with a subject and object). The predicate is compared with the Knowledge Graphs and other unstructured text resources (that are converted into sets of predicates following the same approach used for the claim.

The comparison and verification is based on a set of rule of derivation, that can be manually created by experts or automatically mined. An example of such rules is:

citizenOf (X, Y) ← mayorOf (X, Z), locatedIn(Z, Y)

That means that being *X* the mayor of a certain city *Z* and being the city *Z* located in the country *Y*, it is a consequence that *X* is a citizen of *Y*. Both predicates can be easily found into structured KB like dbpedia,[57] while the derived information is usually not represented. The process of expanding these rules recursively, applying them to the claim to be verified is called **semantic reasoning**. Its result can be the verification of the claim.

One thing that is particular of this approach is that it provides also explanations for the provided output: the steps of reasoning are used to justify the prediction of veracity of the claim.

5.4 Propagation-based: Hoaxy

A tool that is representative of the propagation-based models is *Hoaxy*[58]. This tool allows to visualise the spread of claims and fact-checking, using keywords and enabling the users to find relevant and/or recent articles. It is possible to select a wide number of articles and visualise them together, making the distinction between claims and fact-checking articles.

The visualisation provides a plot of the popularity of the articles over time, and a graph visualisation where it is possible to see (also with temporal animation) the spread of the

[55] **https://www.mpi-inf.mpg.de/impact/exfakt**
[56] **https://people.mpi-inf.mpg.de/~gadelrab/downloads/WSDM2019/poster.pdf**
[57] **https://wiki.dbpedia.org/**
[58] **https://hoaxy.iuni.iu.edu/**

articles over a topology constituted of nodes, which are characterised with their probability of being bots or human, and edges that represent the influence between users (as tweet mentions and retweets).

This tool is also linked to credibility-based features, such as the bot-probability of the nodes in the graph.

In the pipeline of fact-checking, this tool can be used to monitor and find a set of articles that is spreading over social media. However it does not provide verification for the articles provided: the information about their belonging to the fact-checking category is based on a domain list.

5.5 Tool Collection

As part of the Co-Inform project we have been analysing different tools according to our schema of the various characteristics and functions described in section 4. We have created a matrix of tools that will grow during the project. By expanding this table with additional features that the misinformation detection and fact-checking communities find relevant, it should become a resource. In this table, the top level headings are the following:

Tool/Algorithm: The name of the tool, or of the publication if it references a method, rather than a specific instance

Reference: Reference to the publication where the tool or algorithm is listed

Source Code: If there is a link within the paper to any source code, we list it here

Misinformation Type: The type of misinformation to which the tool was applied (as described in section 2.1

Stage of the pipeline: The stage of the fact-checking pipeline as described in section 4.1, which helps to define the type of data that will be collected

Coverage of Topics/Additional Context: When it is clear from the publication in which context the tool was used, we will provide details here. This information can help fact-checkers understand how to interpret the outputs of the tool

Method of Deployment: This describes whether or not the tool is a demo, or whether it is available as a web platform and API, a browser extension or some other self-deployable model

Documentation/Description: This describes whether or not the source code is available, as well as the publication

Detection Method: This lists which detection method(s) are most relevant for the tool as described in section 4.3

Model Outputs: This explains what kind of output the user can expect, such as a probability, or a “nutrition label”.

Target User: This label will explain the type of user that would be able to deploy the tool. This is a new category and will be updated as we explore propagation dynamics in D3.3

Visualisations: This describes what kinds of visual outputs are available to the user

Dynamicity: This category describes the extent to which a tool is ”hand-crafted” and requires rewriting features and models, whether it is static and requires new input by hand, or whether it automatically refreshes itself

Languages Supported: This category lists whether a tool or algorithm is language dependent or whether it can be deployed in many different language contexts

Link: a link for the tools that can be tried online

This table and the related category descriptions are available in the appendixes to this document.

6 Conclusion and Future Directions

In this deliverable we have discussed the misinformation ecosystem focusing on the role of detection tools and current methods and approaches applied towards recognising pieces of misinformation for stopping their spreading.

The number of tools and methods to automate or to support human activities on misinformation detection and fact-checking are fastly growing, as well as their capacity to deliver more accurate and effective results. However, we recognise that each method or approach contributes to specific nuances of the problem, and all the solutions are naturally limited. In line with the EU report that warns against technocentric optimism as a solution to mis/disinformation online (Marsden, Meyer, European Parliament, European Parliamentary Research Service, & Scientific Foresight Unit, 2019), at Co-inform we understand automate detection as one piece of broader solutions.

In this context, technical developments should evolve not only towards improving accuracy, but also integrating aspects of human complexity in judging and acting upon information, ensuring the transparency of the automated solutions, and transforming regular users into active agents more capable of distinguishing fact from fiction.

Towards this direction, Co-inform Work Package 3 developments are targeting to integrate aspects of human values in the new methods and algorithms; facilitate the access to pieces of information fact-checked by trustworthy sources and to support fact-checkers in their work; and contribute to raising awareness and information literacy of social media users. To this end, explainable artificial intelligence techniques will be investigated and applied to translate accountable signals of misinformation detected automatically into arguments that can be understood by humans, including all of our stakeholder groups.

References

Anderson, Janna, and Lee Rainie. “The Future of Truth and Misinformation Online,” n.d., 224.

Anderson, J., & Rainie, L. (n.d.). *The Future of Truth and Misinformation Online*. 224.

Angelotti, E. M. (2012). Twibel Law: What Defamation and Its Remedies Look like in the Age of Twiter. *J. High Tech. L.*, *13*, 430.

Babakar, M., & Moy, W. (2016, August). *The State of Automated Factchecking*. Retrieved from https://fullfact.org/media/uploads/full_fact-the_state_of_automated_factchecking_aug_2016.pdf

Bahuleyan, H., & Vechtomova, O. (2017). UWaterloo at SemEval-2017 Task 8: Detecting Stance towards Rumours with Topic Independent Features. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 461–464. Retrieved from http://www.aclweb.org/anthology/S17-2080

Baly, R., Mohtarami, M., Glass, J., Màrquez, L., Moschitti, A., & Nakov, P. (2018). Integrating stance detection and fact checking in a unified corpus. *ArXiv Preprint ArXiv:1804.08012*.

Baris, I., Schmelzeisen, L., & Staab, S. (n.d.). CLEARumor at SemEval-2019 Task 7: ConvoLving ELMo Against Rumors. *SemEval@ACL2019*, 5.

Benigni, M. C., Joseph, K., & Carley, K. M. (2019). Bot-ivistm: Assessing Information Manipulation in Social Media Using Network Analytics. In N. Agarwal, N. Dokoohaki, & S. Tokdemir (Eds.), *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining* (pp. 19–42). https://doi.org/10.1007/978-3-319-94105-9_2

Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday*, *21*(11). https://doi.org/10.5210/fm.v21i11.7090

Biyani, P., Tsioutsiouliklis, K., & Blackmer, J. (2016). “8 Amazing Secrets for Getting More Clicks”: Detecting Clickbaits in News Streams Using Article Informality. *Thirtieth AAAI Conference on Artificial Intelligence*. Presented at the Thirtieth AAAI Conference on Artificial Intelligence. Retrieved from https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11807

Bountouridis, D., Marrero, M., Tintarev, N., & Hauff, C. (2018). Explaining Credibility in News Articles using Cross-Referencing. *Proceedings of the 1st International Workshop on ExplainAble Recommendation and Search (EARS 2018). Ann Arbor, MI, USA*, 5.

Bourgonje, P., Moreno Schneider, J., & Rehm, G. (2017). From Clickbait to Fake News Detection: An Approach based on Detecting the Stance of Headlines to Articles. *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing Meets Journalism*, 84–89. https://doi.org/10.18653/v1/W17-4215

Burgoon, J. K., Blair, J. P., Qin, T., & Nunamaker, J. F. (2003). Detecting Deception through Linguistic Analysis. In H. Chen, R. Miranda, D. D. Zeng, C. Demchak, J. Schroeder, & T. Madhusudan (Eds.), *Intelligence and Security Informatics* (pp. 91–101). Springer Berlin Heidelberg.

Cheney, J., Finkelstein, A., Ludaescher, B., & Vansummeren, S. (2012). Principles of Provenance (Dagstuhl Seminar 12091). *Dagstuhl Reports*, *2*(2), 84–113. https://doi.org/10.4230/DagRep.2.2.84

Chesney, R., & Citron, D. K. (2018). *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*.

Ciampaglia, G. L., Mantzarlis, A., Maus, G., & Menczer, F. (2018). Research Challenges of Digital Misinformation: Toward a Trustworthy Web. *AI Magazine*, *39*(1).

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., & Flammini, A. (2015). Computational fact checking from knowledge networks. *PloS One*, *10*(6), e0128193.

Dang, A., Smit, M., Moh'd, A., Minghim, R., & Milios, E. (2016). Toward understanding how users respond to rumours in social media. *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 777–784. IEEE.

Daxenberger, J., Eger, S., Habernal, I., Stab, C., & Gurevych, I. (2017). What is the Essence of a Claim? Cross-Domain Claim Identification. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2055–2066. https://doi.org/10.18653/v1/D17-1218

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., … Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554–559.

Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Hoi, G. W. S., & Zubiaga, A. (2017). SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. *ArXiv:1704.05972 [Cs]*. Retrieved from http://arxiv.org/abs/1704.05972

Dungs, S., Aker, A., Fuhr, N., & Bontcheva, K. (2018). Can Rumour Stance Alone Predict Veracity? *Proceedings of the 27th International Conference on Computational Linguistics*, 3360–3370. Retrieved from http://www.aclweb.org/anthology/C18-1284

Ellul, J. (1966). *Propaganda*. Knopf New York, NY.

Enayet, O., & El-Beltagy, S. R. (2017). NileTMRG at SemEval-2017 Task 8: Determining Rumour and Veracity Support for Rumours on Twitter. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 470–474. Retrieved from http://www.aclweb.org/anthology/S17-2082

Esteves, D., Reddy, A. J., Chawla, P., & Lehmann, J. (2018). Belittling the Source: Trustworthiness Indicators to Obfuscate Fake News on the Web. *EMNLP 2018*, 50.

Fernandez, M., & Alani, H. (2018). *Online Misinformation: Challenges and Future Directions*. 595–602. https://doi.org/10.1145/3184558.3188730

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96–104.

Gabbert, F., Memon, A., Allan, K., & Wright, D. B. (2004). Say it to my face: Examining the effects of socially encountered misinformation. *Legal and Criminological Psychology*, *9*(2), 215–227.

Gad-Elrab, M. H., Stepanova, D., Urbani, J., & Weikum, G. (2019). ExFaKT: A Framework for Explaining Facts over Knowledge Graphs and Text. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 87–95. ACM.

Ghanem, B., Rosso, P., & Rangel, F. (2018). Stance Detection in Fake News A Combined Feature Representation. *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, 66–71. Retrieved from https://www.aclweb.org/anthology/W18-5510

Go, A., Bhayani, R., & Huang, L. (n.d.). *Twitter Sentiment Classification using Distant Supervision*. 6.

Gorrell, G., Bontcheva, K., Derczynski, L., Kochkina, E., Liakata, M., & Zubiaga, A. (2018). RumourEval 2019: Determining Rumour Veracity and Support for Rumours. *ArXiv:1809.06683 [Cs]*. Retrieved from http://arxiv.org/abs/1809.06683

Graves, L. (2013). *Deciding What's True: Fact-Checking Journalism and the New Ecology of News* (Columbia University). https://doi.org/10.7916/D8XG9Z7C

Hassan, N., Arslan, F., Li, C., & Tremayne, M. (2017). Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17*, 1803–1812. https://doi.org/10.1145/3097983.3098131

Hassan, N., Nayak, A. K., Sable, V., Li, C., Tremayne, M., Zhang, G., … Kulkarni, A. (2017). ClaimBuster: the first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, *10*(12), 1945–1948. https://doi.org/10.14778/3137765.3137815

Hochschild, J., & Einstein, K. L. (2015). *Do Facts Matter?: Information and Misinformation in American Politics*. Norman, OK: University of Oklahoma Press.

Horne, B. D., Dron, W., Khedr, S., & Adali, S. (2018). Assessing the News Landscape: A Multi-Module Toolkit for Evaluating the Credibility of News. *Companion Proceedings of the The Web Conference 2018*, 235–238. https://doi.org/10.1145/3184558.3186987

Howard, P. N., Ganesh, B., Liotsiou, D., Kelly, J., & François, C. (2018). *The IRA, Social Media and Political Polarization in the United States, 2012-2018*. University of Oxford.

Huang, Y. L., Starbird, K., Orand, M., Stanek, S. A., & Pedersen, H. T. (2015). Connected Through Crisis: Emotional Proximity and the Spread of Misinformation Online. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*, 969–980. https://doi.org/10.1145/2675133.2675202

Ireton, C., & Posetti, J. (2018). *Journalism,'Fake News'& Disinformation*. United Nations Educational, Scientific and Cultural Organization.

Israel, D. J., & Perry, J. (1991). *What is information?* CSLI.

Jost, J. T., van der Linden, S., Panagopoulos, C., & Hardin, C. D. (2018). Ideological asymmetries in conformity, desire for shared reality, and the spread of misinformation. *Current Opinion in Psychology*, *23*, 77–83.

Karlova, N. A., & Fisher, K. E. (2013). Plz RT": A social diffusion model of misinformation and disinformation for understanding human information behaviour. *Information Research*,

18(1), 1–17.

Klein, D., & Wueller, J. (2017). *Fake news: a legal perspective*.

Kochkina, E., Liakata, M., & Augenstein, I. (2017). Turing at SemEval-2017 Task 8: Sequential Approach to Rumour Stance Classification with Branch-LSTM. *ArXiv:1704.07221 [Cs]*. Retrieved from http://arxiv.org/abs/1704.07221

Kochkina, E., Liakata, M., & Zubiaga, A. (2018). All-in-one: Multi-task Learning for Rumour Verification. *ArXiv:1806.03713 [Cs]*. Retrieved from http://arxiv.org/abs/1806.03713

Konstantinovskiy, L., Price, O., Babakar, M., & Zubiaga, A. (2018). Towards Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection. *ArXiv:1809.08193 [Cs]*. Retrieved from http://arxiv.org/abs/1809.08193

Krejzl, P., Hourová, B., & Steinberger, J. (2017). Stance detection in online discussions. *ArXiv Preprint ArXiv:1701.00504*.

Kuklinski, J. H., Quirk, P. J., Jerit, J., Schwieder, D., & Rich, R. F. (2000). Misinformation and the currency of democratic citizenship. *Journal of Politics*, *62*(3), 790–816.

Kumar, K. K., & Geethakumari, G. (2014). Detecting misinformation in online social networks using cognitive psychology. *Human-Centric Computing and Information Sciences*, *4*(1), 14.

Kumar, S., West, R., & Leskovec, J. (2016). *Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes*. 591–602. https://doi.org/10.1145/2872427.2883085

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., … Rothschild, D. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096.

Lewandowsky, S., Stritzke, W. G. K., Freund, A. M., Oberauer, K., & Krueger, J. I. (2013). Misinformation, disinformation, and violent conflict: From Iraq and the “War on Terror” to future threats to peace. *American Psychologist*, *68*(7), 487–501. https://doi.org/10.1037/a0034515

Li, Y., Chang, M.-C., Farid, H., & Lyu, S. (2018). In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. *ArXiv Preprint ArXiv:1806.02877*.

Li, Y., & Lyu, S. (2018). Exposing DeepFake Videos By Detecting Face Warping Artifacts. *ArXiv:1811.00656 [Cs]*. Retrieved from http://arxiv.org/abs/1811.00656

Lim, C. (2018). Checking how fact-checkers check. *Research & Politics*, *5*(3), 2053168018786848.

Lippi, M., & Torroni, P. (2016). Argumentation Mining: State of the Art and Emerging Trends. *ACM Trans. Internet Technol.*, *16*(2), 10:1–10:25. https://doi.org/10.1145/2850417

Lu, Y., & Yang, D. (2011). Information exchange in virtual communities under extreme disaster conditions. *Decision Support Systems*, *50*(2), 529–538.

Ma, J., Gao, W., & Wong, K.-F. (2018). Detect Rumor and Stance Jointly by Neural Multi-task Learning. *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW ’18*, 585–593. https://doi.org/10.1145/3184558.3188729

Marsden, C., Meyer, T., European Parliament, European Parliamentary Research Service, & Scientific Foresight Unit. (2019). *Regulating disinformation with artificial intelligence: effects of disinformation initiatives on freedom of expression and media pluralism*. Retrieved from http://www.europarl.europa.eu/RegData/etudes/STUD/2019/624279/EPRS_STU(2019)624279_EN.pdf

McCright, A. M., & Dunlap, R. E. (2017). Combatting misinformation requires recognizing its types and the factors that facilitate its spread and resonance. *Journal of Applied Research in Memory and Cognition*, *6*(4), 389–396.

Mohseni, S., & Ragan, E. (2018). Combating Fake News with Interpretable News Feed Algorithms. *ArXiv:1811.12349 [Cs]*. Retrieved from http://arxiv.org/abs/1811.12349

Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake News Detection on Social Media using Geometric Deep Learning. *ArXiv Preprint ArXiv:1902.06673*.

Nguyen, A. T., Kharosekar, A., Krishnan, S., Krishnan, S., Tate, E., Wallace, B. C., & Lease, M. (2018). Believe it or not: Designing a Human-AI Partnership for Mixed-Initiative Fact-Checking. *The 31st Annual ACM Symposium on User Interface Software and Technology*, 189–199. ACM.

Nguyen, A. T., Kharosekar, A., Lease, M., & Wallace, B. (2018). An interpretable joint graphical model for fact-checking from crowds. *Thirty-Second AAAI Conference on Artificial Intelligence*.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, *1*, 2227–2237.

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

Popat, K., Mukherjee, S., Strötgen, J., & Weikum, G. (2017). Where the truth lies: Explaining the credibility of emerging claims on the web and social media. *Proceedings of the 26th International Conference on World Wide Web Companion*, 1003–1012. International World Wide Web Conferences Steering Committee.

Posetti, J. N. (2017). *UNESCO report: surveillance and data collection are putting journalists and sources at risk*.

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (n.d.). *A Stylometric Inquiry into Hyperpartisan and Fake News*. 10.

Procter, R., Vis, F., & Voss, A. (2013). Reading the riots on Twitter: methodological innovation for the analysis of big data. *International Journal of Social Research Methodology*, *16*(3), 197–214.

Qazvinian, V., Rosengren, E., Radev, D. R., & Mei, Q. (2011). Rumor Has It: Identifying

Misinformation in Microblogs. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1589–1599. Retrieved from http://dl.acm.org/citation.cfm?id=2145432.2145602

Rubin, V., Conroy, N., Chen, Y., & Cornwell, S. (2016). Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, 7–17. https://doi.org/10.18653/v1/W16-0802

Ruchansky, N., Seo, S., & Liu, Y. (2017). CSI: A Hybrid Deep Model for Fake News Detection. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 797–806. https://doi.org/10.1145/3132847.3132877

Ruths, D. (2019). The misinformation machine. *Science*, *363*(6425), 348–348. https://doi.org/10.1126/science.aaw1315

Shah, Z., Surian, D., Dyda, A., Coiera, E., Mandl, K. D., & Dunn, A. G. (2019). Automatically applying a credibility appraisal tool to track vaccination-related communications shared on social media. *ArXiv Preprint ArXiv:1903.07219*.

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-06930-7

Shao, C., Hui, P.-M., Wang, L., Jiang, X., Flammini, A., Menczer, F., & Ciampaglia, G. L. (2018). Anatomy of an online misinformation network. *PLOS ONE*, *13*(4), e0196087. https://doi.org/10.1371/journal.pone.0196087

Shi, B., & Weninger, T. (2016). Fact checking in heterogeneous information networks. *Proceedings of the 25th International Conference Companion on World Wide Web*, 101–102. International World Wide Web Conferences Steering Committee.

Starbird, K., Maddock, J., Orand, M., Achterman, P., & Mason, R. M. (2014). Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing. *IConference 2014 Proceedings*.

Tandoc Jr, E. C., Lim, Z. W., & Ling, R. (2018). Defining “fake news” A typology of scholarly definitions. *Digital Journalism*, *6*(2), 137–153.

Vosoughi, S., Roy, D., & Aral, S. (2018a). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Vosoughi, S., Roy, D., & Aral, S. (2018b). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Wang, S., & Manning, C. (2012). Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 90–94. Retrieved from http://www.aclweb.org/anthology/P12-2018

Wang, W. Y. (2017). “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection. *ArXiv:1705.00648 [Cs]*. Retrieved from http://arxiv.org/abs/1705.00648

Wardle, C. (2017). Fake news. It's complicated. *First Draft News*, *16*.

Wingfield, N., Isaac, M., & Benner, K. (2016). Google and Facebook take aim at fake news sites. *The New York Times*, *11*, 12.

Wu, L., Morstatter, F., Hu, X., & Liu, H. (2016). Mining misinformation in social media. *Big Data in Complex and Social Networks*, 123–152.

Zhi, S., Sun, Y., Liu, J., Zhang, C., & Han, J. (2017a). ClaimVerif: a real-time claim verification system using the web and fact databases. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2555–2558. ACM.

Zhi, S., Sun, Y., Liu, J., Zhang, C., & Han, J. (2017b). ClaimVerif: a real-time claim verification system using the web and fact databases. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2555–2558. ACM.

Zhou, X., & Zafarani, R. (2018). Fake News: A Survey of Research, Detection Methods, and Opportunities. *ArXiv Preprint ArXiv:1812.00315*.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2018). Detection and Resolution of Rumours in Social Media: A Survey. *ACM Comput. Surv.*, *51*(2), 32:1–32:36. https://doi.org/10.1145/3161603

Zuo, C., Karakas, A. I., & Banerjee, R. (n.d.). A Hybrid Recognition System for Check-worthy Claims Using Heuristics and Supervised Learning. *CLEF 2018 Working Notes. Working Notes of CLEF 2018-Conference and Labs of the Evaluation Forum*, *8995*, 14.