H2020-ICT-2018-2 /ICT-28-2018-CSA

SOMA: Social Observatory for Disinformation and Social Media Analysis

Detecting mis- and disinformation in digital media

D2.3: Outlier (disinformation) detection solution

Project Reference No	SOMA [825469]
Deliverable	D2.3: Outlier (disinformation) detection solution
Work package	WP2: Methods and Analysis for disinformation modeling
Type	Report
Dissemination Level	Public
Date	09/11/2020
Status	Final
Authors	Jessica Gabriele Walter, DATALAB, Aarhus University Mathias Holm Sørensen, DATALAB, Aarhus University Anja Bechmann, DATALAB, Aarhus University
Contributor(s)	
Reviewers	Luca Tacchetti, Luiss Data Lab Emanuele Camarda, Luiss Data Lab
Document description	This report details the tools and work done on outlier detection and evaluate the transferability to disinformation and the potential actions needed in order to carry out effective detection

Document Revision History

Version	Date	Modifications Introduced	
		Modification Reason	Modified by
v0.1	10/30/2020	Consolidation of first draft	DATALAB, Aarhus University
v0.2	05/11/2020	Review	LUISS Datalab, LUISS University
v0.3	05/11/2020	Proofread	DATALAB, Aarhus University
v1.0	06/11/2020	Final version	DATALAB, Aarhus University

Executive Summary

The spread of disinformation poses a threat to society and disinformation detection still is a complicated and challenging endeavor. In this report, we summarize practice-oriented and research-oriented approaches to disinformation detection and discuss the challenges they face. Based on this discussion we propose several steps in order to improve disinformation detection. Practice-oriented approaches refer to services provided by fact-checking organizations and private companies in order to detect or publish disinformation. We focus on examples of the services from non-profit organizations namely FactCheck.org and PolitiFact as well as on the fact-checking tools provided by Google namely the Fact Check Explorer and the Fact Check Markup Tool in order to describe procedures and methods used by practice-oriented approaches. With regard to research-oriented approaches to disinformation detection, we describe methods within three subcategories related to the component-based category of disinformation: creator and user analysis, news content analysis and social context analysis. Furthermore, we address the data-mining based category of disinformation by describing supervised, semi-supervised and unsupervised methods. We also address the implement-based category of disinformation by distinguishing real-time and offline detection. In addition, we describe outlier detection methods, which can potentially be used for disinformation detection in more detail.

With regard to practice-oriented approaches, we discuss the issues of media coverage, used methods, independence, audience, reach and influence, transparency, comparability and velocity. Identified challenges concern the documentation of debunked stories, the dependence on experts, the transparency regarding procedures and set up of databases, the facilitation of research for example by increasing comparability and access and the missing information about usage.

With regard to research-oriented approaches, we discuss generalizability, the issue of time, multi-modality information and components of disinformation. In addition, we address methodological and conceptual challenges and discuss the purposes of disinformation detection. Identified challenges concern the general applicability of methods and development of methods allowing for an analysis of event-invariant features of disinformation or early on disinformation detection. Furthermore, challenges are related to the formats of disinformation with for example an existing neglect of visual features and to a combination of methods focusing on different components of disinformation such as social context and textual features. Additional challenges apply to disinformation detection in high-dimensional data, to the comparability of methods, to interdisciplinary approaches to disinformation detection and methodological challenges associated with the conceptualization of disinformation.

We propose an improvement of disinformation detection by a promotion of interdisciplinary approaches, by an extension of automated disinformation detection, by finding a consensus for the conceptualization and definition of disinformation, by facilitating access to information and data and by exploiting new research areas.

Table of Contents

List of Figures

List of Tables

List of Terms and Abbreviations

Abbreviation	Definition
ABOD	Angle-based outlier detection
AI	Artificial intelligence
CNN	Convolutional Neural Networks
COF	Connectivity-based Outlier Factor
GNNs	Graph Neural Networks
GRU	Gated Recurrent Unit
HiCS	High Contrast Subspaces
IFCN	International Fact-Checking Network
kNN	k nearest neighbors
LOCI	Local Correlation Integral
LOF	Local Outlier Factor
LSTM	Long short-term memory
NGO	Non-Governmental Organisation
PFCG	Probabilistic context free grammars
PINN	Projection-indexed Nearest-neighbours
RNN	Recurrent Neural Network
RvNN	Recursive Neural Network
SOMA	Social Observatory for Disinformation and Social Media Analysis

SVM	Support Vector Machine
URL	Uniform Resource Locator
U.S.	United States

1 Introduction

During the last 15 years, the way people interconnect and socialize changed dramatically with social media being a new form of socialization. Social networks, media and platforms serve the purpose of communication, business, information exchange, learning and information gathering, thus affecting many aspects of our everyday lives. The extreme growth of social media and the high velocity in which news are created as well as the limited control over the content that is shared thereby leads to a discussion about the veracity of the information. The spread of disinformation online poses a threat to society in various ways as the creators' reasons behind the spreading of disinformation, as well as the issues affected, are manifold. The intentions can vary for example from an interest in disrupting or influencing (foreign) societies to having profit-oriented ideological, normative or financial aims that potentially profit from the spread of disinformation (see e.g. EU DisinfoLab, 2020; Shu et al., 2020; Tandoc et al., 2018). However, not all information with a lower degree of veracity (be it false or half-true) is intentionally spread. We define disinformation or respectively fake news in this deliverable according to the Oxford English Dictionary as "the dissemination of deliberately false information, esp. when supplied by a government or its agent to a foreign power or to the media, with the intention of influencing the policies or opinions of those who receive it; false information so supplied"[1]. However, we also consider misinformation, as not all disinformation or fake news e.g. regarding the pressing topic Covid-19 is disinformation in this sense. Misinformation differs from disinformation, as the former is not necessarily created with the intention of causing harm or exerting influence (e.g. Carmi et al., 2020). However, we characterize disinformation as an information outlier, that is dis- or misinformation still is an abnormality and not the norm within the information universe and an important goal is to detect and prevent dis- or misinformation in order to prevent harm to societies. The concept of "fake news" also relates to the concept of disinformation and the two terms are often used as synonyms (e.g. Shu et al., 2020), however, they are not equivalent since disinformation focuses on the information and fake news on the news aspect (e.g. Gelfert, 2018; Vargo et al., 2018). "Fake news" is also not a new concept, as misinformation was always present in the media. But it is now more "used to describe false stories spreading on social media" (Tandoc et al., 2018, p.138) and thus changed meaning over time as it is now more used referring to social media instead of traditional media. Based on a literature review, Tandoc et al. (2018) distinguish within the concept of fake news between news satire, news parody, news fabrication (which comes close to the definition of disinformation), photo manipulation and propaganda. They conclude that two domains of fake news can be distinguished namely facticity "the degree to which fake news relies on facts" (p. 147) and intention "the degree to which the creator of fake news intends to mislead" (p.147). This also applies to disinformation, as disinformation stories as well can be false in various degrees. False information can be mixed with true elements. Also for disinformation, the degree of harmful intention can vary. These characteristics of disinformation and misinformation makes disinformation detection an even more

[1] https://www.oed.com/view/Entry/54579?redirectedFrom=disinformation#eid; last access October 30th, 2020

complicated endeavor. Besides the categories disinformation, misinformation and fake news, the terms “rumor” and “hoax” are also used as specific types of false information. Habib et al. (2019) describe rumors as “a piece of information whose truthfulness is in doubt and source is unreliable and probably produces under the emergency situation which creates panic in public, diminishing the government credibility, disturbs the social order and even threatens the national security” (p. 3). They describe hoaxes as “electronic messages with evil intention to misguide recipients consist[ing] of audio, text and multimedia content” (p. 4). Shu et al. (2020) mention that hoaxes aim at manipulating or persuading receivers in order to provoke or prevent a specific action, often using a threat or deception, and are usually spread to a large number of receivers. Even though we focus on dis- and misinformation, rumors and hoaxes are considered regarding the identification of disinformation detection methods, as some focus explicitly at hoaxes or rumors. Within the last decades we have seen an increased effort in disinformation detection also as a response to its grown impact on politics be it for example on the US presidential election in 2016 or the Brexit referendum (e.g. Gelfert, 2018).

1.1 Purpose and Scope

In this deliverable, we will lay out existing solutions for outlier detection with disinformation being the specific type of outlier we are looking at. Here we will distinguish two main approaches. One approach is practice-oriented and is predicated on existing fact-checking tools or services provided by different NGOs or platforms. Based on an identification of existing tools within this approach, we will address several aspects of these existing solutions, namely:

a. applied detection method
b. topic coverage
c. media coverage
d. user information

A second approach is research-oriented and here we will outline different disinformation detection methods and discuss them. By critically assessing both approaches our investigation will explore two main research questions: *1. Which methods and tools for disinformation detection currently exist? 2. In order to optimally detect disinformation, how can existing detection methods be improved?*

1.2 Structure of the report

The report is divided in two main theoretical and methodological sections that address the two main disinformation detection approaches. In the second chapter, we define the practice-oriented and the research-oriented approaches in more detail and therefore address the first question of which disinformation detection methods and tools currently exist. In the third chapter, we discuss the disadvantages and advantages of the outlined methods and tools within the two approaches. Based on this discussion, the fourth chapter outlines steps, which can be undertaken in order to improve disinformation detection. Thus, the third and fourth chapter address the questions of how we can optimally detect disinformation and which improvements of current methods and tools are needed to do so. The conclusion summarizes the main points.

2 Review of existing solutions for disinformation detection

As disinformation detection within social media has become more important over the years, the number of tools and methods to do it have improved and developed as well. Especially in journalism, fact-checking has become a new force (Graves, 2018). However, the tools have not been developed very systematically. Since disinformation itself is a complex concept, the solutions and methods to detect it are manifold. With the refocus of the concept of fake news to social media, online fact-checking tools gained importance for journalists in order to distinguish fake news from facts. Fact-checking tools can be used by journalists, and others who might be interested, to check information. The objectives of fact-checking organizations can be described as "informing the public, improving political rhetoric, and influencing other journalists" (Vargo et al., 2018, p.2033). Beyond this, some fact-checking organizations also aim at holding publishers of disinformation accountable for their actions in order to proactively discourage a repetition of this behavior and the spread of dis- or misinformation and to obtain corrections (Dias & Sippitt, 2020). Usually tools provided by fact-checking organizations offer the opportunity to look for specific topics or actors or to define a time span within the search function. In the first section of this chapter, we review these fact-checking tools. In the second section, we will provide an overview of disinformation detection methods that are also partly used by fact-checking tools in order to debunk disinformation stories but can also be applied by researchers on specific datasets. We refer to checking fake news via fact-checking tools as the practice-oriented approach of disinformation detection. The more general applicable use of disinformation detection methods is called the research-oriented approach in this report.

2.1 Practice-oriented approach to disinformation detection

There is a variety of fact-checking tools or services available used to distinguish fake news or disinformation from verified facts and information (see for the U.S. e.g. Lowrey, 2017). However, fact-checking was mainly developed to verify public political claims and therefore does often not address all kinds of dis- or misinformation. Fact-checking, in addition, can also be distinguished from internal fact-checking of media which aims at eliminating errors before a study or news is published (Graves, 2018). Graves (2018) reports the following definition for fact-checking: "fact checkers and fact-checking organizations aim to increase knowledge by re-reporting and researching the purported facts in published/ recorded statements made by politicians and anyone whose words impact others' lives and livelihoods. Fact checkers investigate verifiable facts, and their work is free of partisanship, advocacy and rhetoric." (p. 615). Typically, five phases of a fact-check can be distinguished: "1. choosing claims to check, 2. contacting the target, 3. tracing false claims, 4. consulting experts and sources, 5. publishing the check as transparently as possible" (Nieminen & Rapeli, 2019, p. 303). However, as we discuss in this report, fact-checking tools differ in their approaches regarding these five phases.

Many fact-checking tools have a strong link to journalism (Graves, 2018) especially in the U.S., but outside the U.S. this can differ, e.g. Nieminen & Rapeli (2019) mention a stronger influence of NGOs in Eastern Europe. Even though there is probably no exhaustive list of fact-checking tools available from which to pick the appropriate one, some lists are maintained, which provide an overview of the most important ones. Vargo et al. (2018) for example base their study on a list maintained by the Duke Reporter's Lab at Duke University and on the Poynter Institute's International Fact-Checking Network (IFCN) (also see Amazeen, 2020). However, it is in itself a difficult task for fact-checkers and journalists or researchers to identify trustworthy tools and find the appropriate tool to check a specific information or fact related to a specific topic, person or organization, since fact-checking tools apply different methods and evaluate different statements (Lim, 2018). Furthermore, new fact-checking sites continue to emerge (e.g. Nieminen & Rapeli, 2019), which also impedes their evaluation and effective usage. Lowrey (2017) also observes that at least in the U.S. fact-checking tools tend to diversify again. For instance, fact-checking approaches differ regarding the way claims are checked, where for example some services integrate the creator of a checked claim in the process of fact-checking while others don't. The services also differ in the way debunked stories are presented or whether and how a rating system is used (Graves, 2018). Thus, the proceedings of different fact-checking service providers differ and so does a statement's phrasing which is evaluated by different platforms (Lim, 2018). These differences make it more difficult to validate fact-checking efforts. Even though fact-checking sites first started to rise in the U.S., their numbers increase rapidly across the world and an international fact-checking movement has emerged. Research about fact-checking services, however, still has a strong focus on the U.S. (Dias & Sippitt, 2020; Nieminen & Rapeli, 2019). In general, three different fact-checking approaches can be distinguished:

1) expert-oriented
2) crowdsource-oriented
3) computational-oriented

Expert-oriented fact-checking relies on experts who do fact-checking. Crowdsourcing-oriented fact-checking relies on normal people to comment on news in order to detect disinformation. Computational-oriented fact-checking is based on automatic techniques to classify information as true or false whereby two tasks can be conducted automatically ¬ namely the identification of check-worth news and the discrimination of the veracity of the claims (Shu et al., 2017). There are also image-based detection tools, which help to examine whether target images have been modified or to assess the authenticity of images (Guo et al., 2020). However, in this report we focus on text-based detection tools.

Based on the diversity of fact-checking services, we will focus in this section on the two fact-checking services that belong to the most important ones for journalists respectively researchers. For this purpose, we take the provider of the service into consideration as we assume that services provided by profit-oriented providers such as Facebook, Twitter or Google differ from those provided by non-profit oriented providers for example regarding procedures, transparency or spread. Furthermore, journalists and researchers have different needs regarding fact-checking tools and services. Thus, we identify the tools and services, which are influential in journalism and research separately. We base the identification of the fact-checking tools, which play an important role in journalism and research on several indicators.

First, we conducted a short written interview with an expert within the fact-checking community asking for information about the most widely used fact-checking tools. The director of Pagella Politica ¬ an Italian fact-checking organization and partner in SOMA ¬ pointed out three services that play an important role. Besides FactCheck.org ¬ as the oldest, still active fact-checking project in the U.S. ¬ he refers to PolitiFact ¬ the winner of the 2009 Pulitzer Prize ¬ and, finally, the fact-checking tools provided by Google given that they are widely used by fact-checkers. Google's Markup Tool for example is used by Pagella Politica as well as by many additional fact-checking projects.

Second, google scholar searches were conducted for all fact-checking services that contribute as members of the International Fact Checking Network (IFCN) to the #CoronaVirusFacts Alliance (a list is provided in the appendix). The IFCN is a unit at the Poynter Institute established in 2015, which "monitors trends, formats and policy-making about fact-checking worldwide" [2] and aims at unsheathing common positions of fact-checkers, promoting standards by applying the fact-checkers' code of principles and promoting fact-checking activities for example by convening conferences or providing training. The #CoronaVirusFacts Alliance, which is led by the IFCN, was launched in January 2020 and it unites around 100 fact-checkers worldwide in order to publish, share and translate facts surrounding the new coronavirus. Thus, all contributors to this database are still actively debunking misinformation and therefore relevant services.[3]

The Google Scholar searches included the name of the respective fact-checking organization and the term "fact check" (to increase the comparability of results) in order to assess the organization's influence on academic research. Even though this can only be an indicator, since the number of citations or similar indicators are not taken into consideration and search results are not analyzed in detail, these searches indicate significant differences between the fact-checking organizations. The number of search results since the year 2016 and without citations and patents also indicate that PolitiFact and FactCheck.org (with approximately 1320 respectively 920 hits) are the most used fact-checking websites in research as well (most search results revealed no more than 10 hits, the organization with the next frequent hits was Full Fact with approximately 190 hits).[4][5]

Regarding fact-checking services provided by profit-oriented companies, an assessment based on Google Scholar searches is less constructive as search terms are difficult to identify. Instead, we applied an exclusion approach looking at social media and platform services with a high penetration in society ¬ namely Facebook, Twitter and Google[6]. Facebook approaches disinformation in several

[2] https://www.poynter.org/ifcn/, last access October 3rd, 2020

[3] https://www.poynter.org/coronavirusfactsalliance/, last access October 3rd, 2020

[4] search conducted on October 3rd, 2020

[5] a similar search was conducted on ProQuest, leading to the same conclusion

[6] https://gs.statcounter.com/social-media-stats/all/europe; https://www.smartinsights.com/search-engine-marketing/search-engine-statistics/; last access October 16th, 2020

ways. On the one hand, they rely on cooperation with third-party fact-checkers, with both FactCheck.org and PolitiFact being partners. On the other hand, Facebook also relies on the users to identify and report disinformation by giving feedback, which is then checked by Facebook[7] (cp. Shu et al., 2020). Furthermore, Facebook's service CrowdTangle can be used to get information about how information spreads on social media. This, however, not with a focus on disinformation and with restricted access[8]. Social Science One has administered access to a limited number of researchers to the so-called URL dataset from Facebook which also includes debunked stories according to the dataset documentation available.[9] Thus, disinformation detection is not bundled into one service and therefore requires a higher effort by researchers since datasets have to be constructed with different services, which differ in the way they log information. Twitter has so far been reluctant to introduce fact-checking in its services. However, Twitter recently labeled tweets by the Americas President Donald Trump, as potentially misleading or glorifying violence, and this labeling led to a discussion about the procedure. Many fact-checking organizations criticized the intransparency of the labeling, e.g. asking for more transparency about why and how and by whom the labeling was conducted (Mantas, Harrison, 2020). Twitter does not refer to this procedure as fact-checking but as providing context (Pham, Sherisse, 2020). However, the information about whether and how a tweet was labeled cannot be accessed by scraping Twitter data.[10] Besides the labeling of tweets and the provision of context of tweets, Twitter does not provide additional tools to detect disinformation. In contrast to Facebook or Twitter, Google recently introduced fact-checking tools, which are at least partly freely available, and thus provides a bundled service for disinformation detection. These characteristics ¬ facilitated access and concentrated services ¬ are presenting a promising service for research, which is why this report focuses on the Google services. From a journalistic and a research perspective, thus, the same practice-oriented services of disinformation detection, provided by non-profit and profit-oriented providers, are important.
Table 1 lists the practice-oriented approaches of disinformation detection from a journalistic and a research perspective distinguished by profit-oriented and non-profit oriented providers. In the following section, we describe these practice-oriented approaches in more detail.

[7] https://www.facebook.com/help/572838089565953; https://www.facebook.com/help/1952307158131536, last access October 6th, 2020

[8] https://www.facebook.com/formedia/solutions/crowdtangle, last access October 6th, 2020

[9] https://socialscience.one/blog/social-science-one-announces-access-facebook-dataset-publicly-shared-urls, last access October 28th, 2020

[10] https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/overview/tweet-object, last access October 30th, 2020

Table 1 Examples of practice-oriented approaches of disinformation detection

	Journalistic perspective	Research perspective
non-profit oriented provider	PolitiFact/ FactCheck.org	PolitiFact/ FactCheck.org
profit-oriented provider	Google Fact Checking Tools	Google Fact Checking Tools

2.1.1 Detailed description of fact-checking tools and services: non-profit oriented providers

The two examples for fact-checking services provided by non-profit oriented providers are FactCheck.org and PolitiFact.

FactCheck.org ¬ FactCheck.org was the first fact-checking organization in the U.S. based on the work of professional journalists. It was founded in 2003 and covers almost all national print and broadcast newsrooms within the U.S. on the national, state and local level (Graves, 2018). Since 2016, it is one of several organizations working with Facebook in order to debunk misinformation shared on Facebook (as part of the Facebook Initiative).[11] FactCheck.org itself is a non-partisan, non-profit service from the Annenberg Public Policy Center at the University of Pennsylvania and aims at reducing the level of deception and confusion in U.S. politics. Hence, FactCheck.org has a strong focus on political disinformation within the U.S.. On the website, several services are provided for information about mis- and disinformation. Besides the political fact-checking, services related to specific areas such as science or the Covid-19 pandemic are also provided. The search function allows the user to search for any topic, actor or other search terms and provides a list of results that can be sorted by relevance or date. A debunked story includes, among others, information about the date of publishing, the context and sources of the claim, the author of the debunked story, the sources used to debunk the story and a rating. The search results are not provided in a file that can be downloaded but have to be collected by the user. The website FactCheck.org provides information about the topics, which are selected for fact-checking and their sources. Sources thereby vary in their format ¬ from TV ads and shows to social media posts of politicians. If a claim is identified by FactCheck.org as potentially being misinformation, FactCheck.org contacts the publisher and requests material to check the initial doubt. Stories or claims, which cannot be verified by the publisher's material, are checked by FactCheck.org's employees using diverse research collaborations and methods. A written story then enters an editing and review process within FactCheck.org.[12] Thus, FactCheck.org relies on an expert-oriented approach towards disinformation detection. Regarding the use of the service, the website does not provide information for example about how often searches are conducted or debunked stories cited. Figure 1A in the appendix provides an example for how search results are presented by FactCheck.org.

[11] https://www.factcheck.org/fake-news/, last access 29th of September 2020

[12] https://www.factcheck.org/our-process/, last access October 6th, 2020

PolitiFact ¬ PolitiFact was founded in 2007 by the Tampa Bay Times (Singer, 2019). In 2018, the Poynter Institute (a nonprofit school for journalists) acquired PolitiFact and now runs the service as "a nonpartisan fact-checking website to sort out the truth in American politics"[13]. Hence, PolitiFact also has a strong focus on misinformation within American politics. Facebook finances PolitiFact to some extent, whereas this does not mean that Facebook has an influence on the issues of Politifact or that Politifact supports products, services or opinions of Facebook or other donors. However, PolitiFact collaborates with Facebook and TikTok to slow the spread of online misinformation. Hereby, Facebook and TikTok flag doubtful posts that then are checked by PolitiFact's fact-checkers. Besides these collaborations, PolitiFact's journalists choose claims that they fact-check. In addition, PolitiFact allows everyone to suggest a topic or claim for fact-checking on the website[14]. PolitiFact covers stories from diverse media ¬ e.g. from TV as well as from social media. The debunked stories are made available using a search function, which allows for the search of any search term. Besides the general search function, categories (issues, people, state editions and media) are provided to guide the search. The search results provide, for example, information about the authors who rated the claim, the publisher and publishing date of the claim, its context, the publishing medium, the sources for fact-checking and the rating. PolitiFact uses the so-called "truth-o-meter" to rate claims and distinguishes in six categories the level of truth ¬ true, mostly true, half true, mostly false, false and pants on fire. PolitiFact does not provide a function to export the search results and, hence, no data file can be downloaded. In addition, information about users of the provided services is not made public. The disinformation detection approach used by PolitiFact can also be described as expert-oriented. Figure 2A in the appendix provides an example of how search results are presented by PolitiFact.

Both described services ¬ FactCheck.org and PolitiFact ¬ therefore apply a. an expert-oriented approach of disinformation detection, b. focus on dis- and misinformation related to American politics, c. take diverse media formats into consideration and d. provide not much information about the usage of the provided services.

2.1.2 Detailed description of fact-checking tools and services: profit-oriented providers

The Google Fact Checking Tools consists of two tools: the Fact Check Explorer and the Fact Check Markup Tool. Both tools "aim to facilitate the work of fact checkers, journalists and researchers", whereas Google "does not endorse or create any of these fact checks"[15]. This is an immense difference to the described non-profit services from FactCheck.org and PolitiFact. Google does not hire any experts or applies any automated methods in order to detect disinformation, but relies on third parties' fact-checking only. The Fact Check Explorer allows users to browse and search for fact checks by using keywords. The results can be restricted to a specific publisher, language or the most

[13] https://www.politifact.com/who-pays-for-politifact/, last access October 3rd 2020

[14] https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/, last access October 6th 2020

[15] https://toolbox.google.com/factcheck/about, last access October 7th, 2020

recent ones and consists of a list of matching claims and fact checks related to the used keywords. Fact check articles are listed if they meet specific Google guidelines. The Fact Check Markup Tool facilitates the process of providing a structured markup by allowing users to submit a markup using a simple web form. Authorized publishers who write a fact check article, can add a “schema.org/ClaimReview” markup which contains information about the fact check, e.g. about the dates of a rating or the rating itself[16]. This information is, among others, also provided for the search results of the Fact Check Explorer. The search allows for searches regarding any topic and diverse media channels are covered including social media. Thus, the Google fact-checking tools are a. expert-oriented in the extent to which they rely on services with expert-oriented approaches and otherwise provide an automated search and technical supported markup, b. cover all topics whereas the search results also depend on the topics covered by the authorized publishers of the fact checks, c. cover a large spectrum of media and d. are limited in the provision of information about users. Figure 3A in the appendix provides an example of how search results are presented by the Google Fact Check Explorer.

2.1.3 Collaborative approaches to fact-checking

Even though the different providers of fact-checking tools operate more or less independently, international networks have been established in order to facilitate collaboration and define common standards. The IFCN is one of the examples of collaboration between fact-checkers. Also within the SOMA project, the effort was made to support fact-checking via TrulyMedia.

TrulyMedia is a collaborative platform for content verification and the D3.1 report “Social Media Observatory Guide” (Tsabouraki, Danae et al., 2018) provides a detailed description and analysis of the operations and functionalities. TrulyMedia was developed mainly as a tool for journalists and fact-checkers to facilitate the handling of online disinformation by an optimization and facilitation of collaborations and workflow, an “integration of multiple fact-checking tools, the use of analytics and big data” (D3.1, p.3) and therefore also by a more efficient and effective use of working time. TrulyMedia is a tool, which can be used by members of the SOMA network. Its main features are:

1. Content aggregation ¬ the finding and collection of large volumes of information from various social media platforms,
2. Content curation ¬ the organization of content into collections,
3. A robust search and filtering function ¬ enabling the search across and within collections by applying various filters,
4. Automated translation ¬ the translation from and to diverse languages,
5. Content analysis ¬ the analysis of the aggregated content with TrulyMedia and third-party tools,
6. Editing ¬ the creation and publishing of content,
7. Collaboration ¬ the connection of team members and other organizations in order to facilitate the analysis and verification of content.

[16] https://toolbox.google.com/factcheck/about, last access October 7th, 2020

2.2 Research-oriented approach to disinformation detection

Fact-checking efforts, as they have been described in 2.1, to some extent also rely on academia and nonprofit worlds in order to validate information and approaches or to gain practical insights (Graves, 2018). There are also research articles, which use the services of practice-oriented approaches to disinformation detection in order to detect disinformation in social media. A research article, which gained a lot of attention and was cited in many other studies, is the one from Vosoughi et al. (2018). In their study, they analyze how true and false news spread online using Twitter as an example. Based on rumor-investigations published on websites of six different fact-checking organizations, they automatically collect the corresponding cascades to those rumors on Twitter. Their conclusion is that false news ¬ especially false political news ¬ diffuse farther, faster, more broadly and deeper than true news. However, as their approach relies on fact-checking services, it cannot surmount limitations of these services. Shao et al. (2018), for example, point out that the study of Vosoughi et al. (2018) only relies on a limited set of fact-checked information and therefore potentially takes information from low-credibility sources not into consideration, since these are often not fact-checked. Furthermore, it does not take resharing of information into account. It also focuses on Twitter data, and it might be that the results cannot be generalized to other social media (Kumar & Shah, 2018). A research approach, which is based on fact-checking services, is also always limited with regard to the amount and velocity of disinformation spreading. Especially due to the high velocity of the creation and spreading of dis- and misinformation, the improvement of automated detection methods and their application increased the attention from researchers. In this section, we will summarize these methods. The starting point for this summary was a literature review with the search terms "disinformation detection", "misinformation detection", "false information detection" and "fake news detection" in combination with "social media" and "online". The first ten search results for each search term combination is listed in Table 1A in the appendix.

2.2.1 Conceptualization of dis- or misinformation

For a better understanding of how research-oriented approaches of disinformation detection address disinformation, a more detailed conceptualization is necessary.

Disinformation or misinformation consists of several aspects, which can be analyzed using automated techniques. First, there is the creator of the dis- or misinformation. Second, there is its content and the context in which the dis-or misinformation is published. Finally, there are the receivers or victims of dis- or misinformation. This distinction is also made by Zhang & Ghorbani, (2020) for Fake News. In their review of disinformation detection, they provide the graph presented as Figure 1 in order to visually conceptualize the term "Fake News".

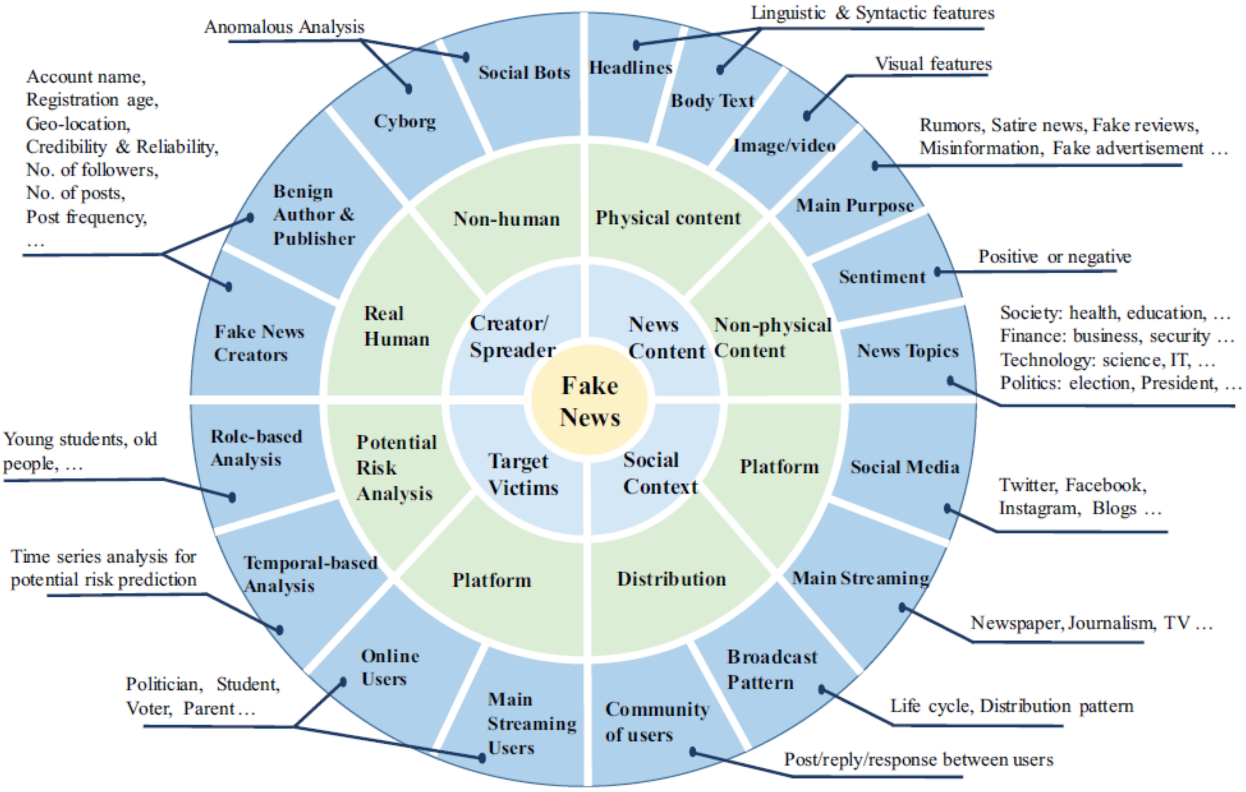

Fake News
Creator/ Spreader
News Content
Target Victims
Social Context
Non-human
Physical content
Real Human
Non-physical Content
Potential Risk Analysis
Platform
Platform
Distribution
Social Bots
Cyborg
Benign Author & Publisher
Fake News Creators
Headlines
Body Text
Image/video
Main Purpose
Sentiment
News Topics
Social Media
Main Streaming
Broadcast Pattern
Community of users
Main Streaming Users
Online Users
Temporal-based Analysis
Role-based Analysis
Anomalous Analysis
Account name, Registration age, Geo-location, Credibility & Reliability, No. of followers, No. of posts, Post frequency, …
Young students, old people, …
Time series analysis for potential risk prediction
Politician, Student, Voter, Parent…
Linguistic & Syntactic features
Visual features
Rumors, Satire news, Fake reviews, Misinformation, Fake advertisement …
Positive or negative
Society: health, education, …
Finance: business, security …
Technology: science, IT, …
Politics: election, President, …
Twitter, Facebook, Instagram, Blogs …
Newspaper, Journalism, TV …
Life cycle, Distribution pattern
Post/reply/response between users

Figure 1 Visual conceptualization of Fake News by Zhang & Ghorbani (2020, p.4)

Zhang & Ghorbani (2020) distinguish between human and non-human creators of dis- or misinformation. Non-human creators can be social bots or cyborgs and anomalous analysis can be used to detect them. Cyborgs are thereby bot-assisted humans or human-assisted bots. Social bots are defined as “computer algorithms that are designed to exhibit human-like behaviors, and automatically produce content and interact with humans on social media” (Zhang & Ghorbani, 2020, p.5). Human creators can be authors and publishers, who do not intend to harm, or authors and publishers, who intentionally create disinformation. In both cases, several indicators and information about these authors and publishers can be used to detect disinformation. For example, the account name, geo-location, credibility, posting behavior or number of followers.

With regard to the news content, physical content and non-physical content can be distinguished. The physical content consists of characteristics of the dis- or misinformation itself that is linguistic, syntactic and visual features. Non-physical content addresses characteristics such as purpose, the sentiment or topic of the dis- or misinformation. That is, information and disinformation consists of explicitly accessible features and implicitly accessible features.

The social context of dis- or misinformation addresses the distribution and the platform on which dis- or misinformation is published, which can be a main streaming platform for example from a newspaper or TV channel or a social media platform such as Twitter or Facebook. The aspect distribution includes the user community and patterns of distribution ¬ for example, how the dis- or misinformation spreads or users reply to it.

The aspect receiver and victims of dis- or misinformation includes the users of the platforms on which the information is published and also addresses analyses of potential risks of experiencing or being exposed to dis- or misinformation.

The practice-oriented approaches described in section 2.1 also address some of the aspects of this conceptualization, such as the creator of dis-or misinformation, aspects of the social context or the content. In this section, however, we focus on disinformation detection methods that go beyond the efforts of fact-checking and also address automated approaches.

2.2.2 Research-oriented methods to address different aspects of dis- or misinformation

Research-oriented approaches to disinformation detection usually address different aspects of dis- or misinformation. Whereas some approaches are more useful for specific kinds of dis- or misinformation than others, for example with traditional linguistic processing and embedding techniques being not as useful for the detection of fake news ¬ for which deep learning techniques gained importance ¬ as it is for the detection of reviews or rumors (Zhang & Ghorbani 2020). Also regarding different kinds of media, methods can vary. Disinformation detection on traditional news media is often based on news content, while social context information can often be used for disinformation detection on social media (Shu et al. 2017).

Zhang & Ghorbani (2020) distinguish three types of models for research-oriented approaches with several subcategories each: component-based, data mining-based and implement-based. Table 2 provides an overview of the component-based category of research-oriented approaches, which are described in detail next.

Table 2 Component-based category of research-oriented approaches to disinformation detection

	Description	Methodological examples
Creator and user analysis		
user profiling analysis	characteristics of user are taken into account, description of activity and evaluation of suspicion	used language, geographic information, creation respectively registration time, account verification
temporal and posting behavior analysis	takes temporal aspects of posting behavior and posting behavior itself into account	signal similarity to Poisson process, average time between posts, reply/ share or mention frequency
credibility related analysis	assesses the credibility of a disinformation creator	use number of friends or followers/ ratio between friends and followers as indicator
sentiment related analysis	Focuses on sentiments related to a specific dis- or misinformation	illustration of emotions, attitudes and opinions; psychological keyword analysis; combined analysis of various sentiment variables
News content analysis		
linguistic and semantic-based analysis	analysis of linguistic patterns and writing styles; characterization of the syntactic structures	methods which represent raw news texts such as “bag-of-words” “n-grams or word2vec; natural language processing, deep syntax analysis
knowledge-based analysis	direct assessment of truthfulness of claims stated in the news	fact-checking as it is described in 2.1; artificial intelligence (AI)-based learning models
style-based analysis	consists of physical and non-physical style analysis	extracting influential physical features such as writing style, text syntax; identifying suspicious tokens (e.g. use of URLs or hashtags); analysis of complexity or readability
Social context analysis		
user network analysis	analyses of networks of news creators and interaction between online users	analyzing interactivity between users, network size, credibility of the network
distribution pattern analysis	analyses of characteristics of information spreading	anomalous pattern detection

Note: own presentation based on review from Zhang & Ghorbani (2020)

Component-based ¬ According to Zhang & Ghorbani (2020) the component-based category of approaches can be divided into creator and user analysis, news content analysis and social context analysis. That is, this category addresses different aspects and components of dis- or misinformation. The first category *"creator and user analysis"* consists of four different approaches. First, user-profiling analysis takes the characteristics of users into account by looking for example at the used language, account creation or registration time or available geographic information. Furthermore, the activity of the creator or user is analyzed and suspicious behavior detected. Here creation time of news or the account verification can be analyzed. Second, a temporal and posting behavior analysis takes temporal aspects and characteristics of the posting behavior into account. For example, the signal similarity to a Poisson process can be assessed, the average time between posts can be analyzed or the frequency in which posts or information is shared, replied to or mentioned. Third, with regard to a credibility related analysis the credibility of a dis- or misinformation creator is assessed by looking for example at the number of friends and followers. Finally, sentiment related analysis focuses on sentiments related to specific dis- or misinformation. The aim is to illustrate emotions, attitudes or opinions. For example, a potential method is a psychological keyword analysis. Often several sentiment variables are analyzed in combination. Shu et al. (2020), in addition, summarize methods to detect social bots. According to the classification of Zhang & Ghorbani (2020) presented in Figure 1, the question of whether a human or non-human spread the disinformation also belongs to the creator and user analysis category. Shu et al. (2020) distinguish graph-based methods, crowdsourcing methods and feature-based methods to detect social bots. Graph-based methods "lie on the assumption that the connectivity of bots is different from human users on social media" (p. 10). Crowdsourcing approaches rely on humans for the identification of bots. Feature-based methods are based on the assumption that social bots can be distinguished from humans by analyzing their features in order to detect characteristics that are specific for respectively bots or humans. Here, the content, activity patterns or network connections are taken into consideration.

The subcategory "*news content analysis*" consists of three different approaches according to Zhang & Ghorbani (2020) ¬ a linguistic and semantic-based analysis, knowledge-based analysis and a style-based analysis. With regard to a linguistic and semantic-based analysis, linguistic patterns and writing styles can be analyzed or the syntactic structures can be characterized. Here methods are applied which represent raw news texts, for example "bag-of-words" models, "n-grams" models or "word2vec". "Bag-of-words" and "n-grams" are text representation models, with "bag-of-words" relying on allocating all words of a text document or a set of text documents to a container that allows ignoring the order of words or the grammar. "N-grams" models "either count [...] word frequencies or weight of a term in a document to characterize the input text" (Al Asaad & Erascu, 2018, p. 382). "Word2vec" is described by Ioannis, Konstantinidis (2018) as "an unsupervised learning technique that learns word embeddings from a collection of documents using contextual information integrated in a shallow neural network, i.e. a neural network that contains only one hidden layer" (p. 35). Regarding linguistic-based analysis the fact that fake news are often phrased in an "opinionated and inflammatory language, crafted as 'clickbait' (i.e., to entice users to click on the link to read the full article) or to incite confusion" is often used (Shu et al., 2017, p.26). Also deep syntax analysis can be used to analyze dis- or misinformation. Deep syntax models use "probabilistic context free grammars (PCFG), with which sentences can be transformed into rules that describe the syntax structure" (Shu et al., 2017, p.28). Furthermore, knowledge-based analysis consists of a direct assessment of the truthfulness of claims. Beyond the practice-oriented approaches described in chapter 2.1, artificial intelligence (AI)-based learning models are also applied in this regard. The third approach is a style-based analysis and it consists of physical and non-physical styles analysis. The aim is to extract influential physical features such as writing style or text syntax or the identification of suspicious tokens, for example related to the use of URLs or hashtags, in order to find indicators for the objectivity of news content. A hyperpartisan style or yellow-journalism is an example for lacking objectivity (cp. Shu et al., 2017). With regard to style analyses, in addition the analyses of the complexity or readability of news texts takes place (Zhang & Ghorbani, 2020). Belonging to this third approach, visual features can also be analyzed, as visual cues can be manipulated for fake news propaganda. Visual features that are analyzed are "clarity score, coherence score, similarity distribution histogram, diversity score, and clustering score" (Shu et al., 2017, p. 26). In addition, statistical features can be used such as "image ratio, multi-image ratio, hot image ratio [or] long image ratio" (Shu et al., 2017, p.26).

Regarding the category "*social context analysis*", a user network analysis as well as a distribution pattern analysis can be distinguished. The former means the analysis of networks of news creators and the interaction between online users. This is done by looking, for example, at the interactivity between users, the network size of a dis- or misinformation creator or the credibility of the creator's network. The latter describes the analysis of characteristics of information spreading using anomalous pattern detection (Zhang & Ghorbani, 2020). In addition, the stance-based approaches can be used to detect disinformation by looking at "users' viewpoints from relevant post contents to infer the veracity of original news articles" (Shu et al., 2017, p. 28). Besides analyzing stances based on explicitly expressed emotions or opinions, topic-modelling methods can also be used to learn latent stance from topics in order to identify implicit representations of stances (Shu et al., 2017). Hence, news veracity in this case is assessed on the basis of stance values of relevant posts.

Component-based approaches to disinformation detection therefore focus on different aspects of dis- or misinformation and aim at detecting patterns or characteristics, which are specific for dis- or misinformation in contrast to truthful information in order to detect dis- or misinformation.

Zhang & Ghorbani (2020) also provide literature examples for different component-based types of models regarding dis- or misinformation detection approaches (see also Guo et al., 2020).

Data mining-based ¬ Zhang & Ghorbani (2020) summarize supervised, semi-supervised and unsupervised models regarding the data mining-based category of disinformation detection approaches with a focus on supervised and unsupervised machine-learning models. Machine-learning refers to an artificial intelligence discipline that allows systems automatic learning and improvement based on experience (Habib et al., 2019). Habib et al. (2019) describe data mining-based disinformation detection using machine-learning techniques in four steps:

1. data preprocessing; in this step noise within the (mainly text) data is reduced in order to improve the performance of classifier;

2. feature extraction; in this step the extraction of relevant information for content classification takes place with the extracted information used as input for the next step;

3. train-test split; in this step the dataset is divided into a training and a testing part with the former being used for learning the algorithm and the latter for an evaluation of the model's performance;

4. applying machine-learning classifier; the chosen model provides an estimation for whether the features belong to a specific class, for example to false or credible information.

Supervised methods produce predictions based on labeled examples (Habib et al., 2019). Supervised learning methods, which are used to detect online hoaxes, frauds or to classify deceptive information, are the application of machine-learning algorithms such as "Decision Tree, Random Forest, Support Vector Machine (SVM), Logistic Regression, K-nearest Neighbour" (cp. Habib et al., 2019; Zhang & Ghorbani, 2020, p. 15). Effective supervised learning models depend on high quality datasets for training the model. Furthermore, deep learning algorithms have been used for visual object recognition and speech recognition. Deep learning is a subfield of machine-learning and is also referred to as deep machine-learning, hierarchical learning or deep structured learning. It means a "set of algorithms originated by the function and structure of the brain" (Habib et al., 2019, p. 8). That is, "deep learning-based methods learn the latent depth representation of information through neural networks" (Guo et al., 2020, p.3) and many studies model related posts as time-series data (Guo et al., 2020). Deep learning approaches can handle large-sized data (Islam et al., 2020). Furthermore, deep learning algorithms can process raw data and automatically detect representations. Zhang & Ghorbani (2020) highlight the use of deep learning algorithms such as Recurrent Neural Network (RNN) for the revelation of sequence structures in high-dimensional data which can be applied regarding natural language processing and used for topic classification, sentiment analysis, question answering or language translation. Deep learning-based methods are also good for disinformation detection, since deep-learning algorithms such as LSTM (long short-term memory), bidirectional LSTM or Gated Recurrent Unit (GRU) do not rely on hand-crafted textual features of disinformation and can help to analyze information about creators and context (cp. Islam et al., 2020; Zhang & Ghorbani, 2020).

Semi-supervised learning algorithms use labeled and unlabeled data for training (Habib et al., 2019). Unsupervised learning methods allow applying models, which do not depend on a labeled dataset and therefore overcome obstacles such as the large size and incompleteness of most online datasets, which are also often unstructured, uncleaned and unlabeled, and the high velocity of spreading and the diversity of dis- or misinformation. However, they have not been used often so far to detect disinformation and if so often focus on sentiment analysis or semantic similarity analysis (Zhang & Ghorbani, 2020). Unsupervised methods, which are promising for disinformation detection, are cluster analysis, semantic similarity analysis, outlier analysis or unsupervised news embedding. Regarding disinformation detection, cluster analysis can be used to identify clusters of news or authors in order to examine for example homogeneity.

Semantic similarity analysis can be used to identify similarities between texts and therefore detect duplicates. Since dis- and misinformation often relies on already published content, disinformation detection is a potential field of application for such analysis. Unsupervised news embedding highlights the importance of embedding as part of natural language processing and "refers to a process of extracting distributed representations of raw textual data" (Zhang & Ghorbani, 2020, p. 21). The chosen embedding methods determines the success of disinformation detection by affecting the way the underlying nature of news or information is measured. Habib et al. (2019) also reviewed some supervised, semi-supervised and unsupervised methods for the detection of rumors, fake news or misinformation and also reported which machine-learning techniques are used most often. Also Guo et al. (2020) review disinformation detection methods and additionally describe Convolutional Neural Networks (CNN), RNN, Recursive Neural Network (RvNN), Auto-Encoder, Generative Adversarial Network and attention mechanisms as deep-learning approaches to disinformation detection (see for details p. 8-9). Shu et al. (2020), in addition, highlight the potential of graph mining-based disinformation detection methods. These methods use graph neural networks (GNNs), which are neural models that learn latent node representations in graphs. Islam et al. (2020) distinguish discriminative models, generative models and hybrid deep learning models to detect disinformation and review models in each category with CNN or RNN being examples.

Implement-based category ¬ Dis- or misinformation detection can take place real-time or offline. Especially a classification of dis- or misinformation with categories such as fake review, satire or hoaxes often takes place offline. Offline approaches to dis- or misinformation detection, however, are associated with the limitations of a dependence on specific datasets, which may not reveal the underlying characteristics of dis- or misinformation and an application of specific learning models, which makes applying it to different datasets more difficult. Real-time detection methods can be used to identify dis- or misinformation in real-time while information is produced and spread and has the potential to contribute to an improvement of applied offline methods or to a prediction of dis- or misinformation. However, only few studies use real-time detection approaches (Zhang & Ghorbani, 2020).

Besides Zhang & Ghorbani (2020), further studies review disinformation detection methods and present studies that follow different disinformation detection approaches, for example Kumar & Shah (2018), Guo et al. (2020), Islam et al. (2020) or Habib et al. (2019).

2.2.3 Outlier detection methods for disinformation detection

In this section, we focus explicitly on outlier detection methods as potential approaches to detect disinformation. Outlier analysis aims at detecting abnormal characteristics of objects and therefore can potentially be applied to disinformation as well. That is, disinformation is not only defined theoretically as an outlier but also methodological. However, only a few studies use outlier detection methods to detect disinformation. Outlier detection focuses on the provision of statistical measures and applies distance or density-based methods in order to identify outliers (Zhang &

Ghorbani, 2020). Boukerche et al. (2020) provide a good overview of outlier detection methods. They mention diverse areas for application, for example fraud detection or medical anomaly diagnosis. Most approaches are unsupervised since labeled datasets are lacking, whereas there is no method that suits all datasets or scenarios. However, also supervised and semi-supervised outlier detection methods are used. Outliers are thereby "different from the norm with respect to their features" and "rare in a dataset compared to normal instances" (Boukerche et al., 2020, p. 2) and can refer to individual data instances or a collection of data instances. Furthermore, vector outliers can be distinguished from graph outliers. The former "are mentioned with vectorlike multi-dimensional data, while [the latter…] exist in graph data" (Boukerche et al., 2020, p. 3). Boukerche et al. (2020) focus on vector outliers. In Figure 2 their classification of outlier detection techniques is provided.

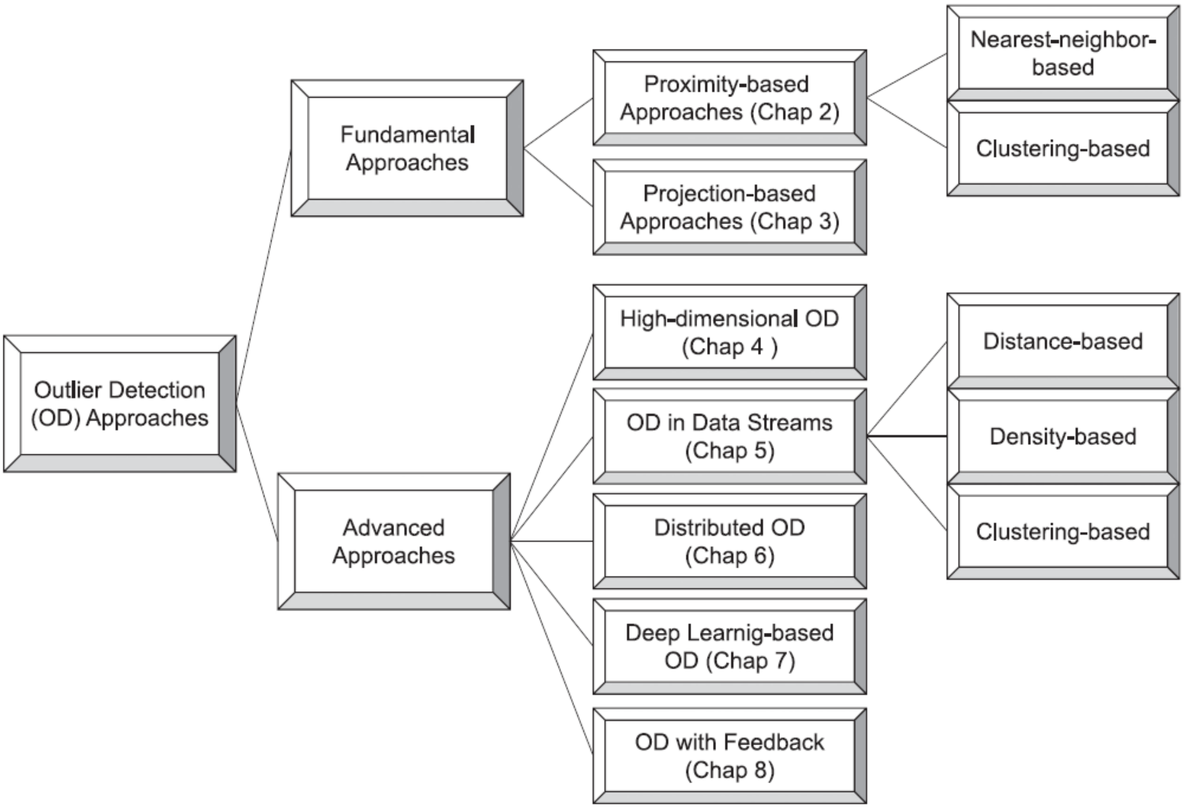

Outlier Detection (OD) Approaches
Fundamental Approaches
Advanced Approaches
Proximity-based Approaches (Chap 2)
Projection-based Approaches (Chap 3)
Nearest-neighbor-based
Clustering-based
High-dimensional OD (Chap 4)
OD in Data Streams (Chap 5)
Distributed OD (Chap 6)
Deep Learnig-based OD (Chap 7)
OD with Feedback (Chap 8)
Distance-based
Density-based
Clustering-based

Fig. 2 Classification of outlier detection techniques by Boukerche et al. (2020, p.4)

Advanced approaches built on the fundamental approaches to address new challenges such as high-dimensional data. A high number of features in the dataset characterizes high-dimensional data. That is, the number of attributes can exceed the number of observations, as it is the case for example for genome data. To detect outliers in high-dimensional data is especially challenging as many dimensions may be noisy as "irrelevant attributes have a dilution effect on the accuracy of distance computations and therefore the resulting outlier scores might also be inaccurate" (Aggarwal, 2017, p. 17). If this is the case, often lower-dimensional local subspaces of relevant attributes can be used and these methods are called subspace outlier detection methods. "The assumption in subspace outlier detection is that outliers are often hidden in the unusual local behavior of low-dimensional subspaces, and this deviant behavior is masked by full-dimensional analysis" (Aggarwal, 2017, p. 19). Some of the methods described by Boukerche et al. (2020) are such subspace methods. They group fundamental approaches into two groups based on the techniques they use ¬ namely proximity-based and projection-based methods. Proximity-based approaches again are grouped into nearest-neighbor-based and clustering-based methods.

Nearest-neighbor-based outlier detection approaches use the relation of a data point to its nearest neighbors to measure the degree of abnormality. "Neighborhood" can be defined as *k nearest neighbors (kNN)* or as *neighborhoods within a pre-specified radius with a data point as center*. K nearest-neighbors-based methods are already used for disinformation detection (Zhang & Ghorbani, 2020). These methods rely on the assumption that "normal data instances are closer to their neighbors, thus forming a dense neighborhood, whereas outliers are far from their neighbors, thus sparsely populated." (Boukerche et al., 2020, p. 5). Local Outlier Factor (LOF) is one of the basic approaches for local outliers ¬ that is an outlier that is defined by differences to its nearest neighbors and not the entire dataset (Global outlier) ¬ and is based on the idea that local outliers differ significantly from their closeby data points. For a specific data instance the LOF score is based on "the average ratio of the instance's neighbor's density to that instance's density" (Boukerche et al., 2020, p. 6). There are enhancements of this approach using different definitions for neighborhood. For example, the Connectivity-based Outlier Factor (COF) "uses the notion of 'isolativity,' which is the degree that a data point is connected with others" (Boukerche et al., 2020, p. 7) with 'isolativity' relying on the concept of 'chaining distance', that is the shortest path to connect k neighbors and a data instance. The Local Correlation Integral (LOCI) uses the definition of local density, where neighbors within a radius r around a data point are counted (Boukerche et al., 2020). Nearest-neighbor-based methods enable us to differentiate between strong and weak outliers and this provides an advantage over clustering-based methods. But they are often associated with a high computation complexity and depend strongly on the choice of k. That is, "an overly large k results in a weak distinction between outliers and normal points. An overly small k results in an unreliable estimation of the proximity density" (Boukerche et al., 2020, p. 9). Clustering-based methods for outlier detection usually group data with clustering algorithms and then assess the degree of deviation based on the results of the clustering for example by analysing the distance of a data point to the cluster center or describing differences between clusters with outliers and clusters without outliers. Projection-based methods are often efficient methods and can also be applied to high-dimensional data. Examples are the Projection-indexed Nearest-neighbours (PINN) method, which "is based on a random projection scheme to reduce the data dimensionality and thus decrease the computation cost of determining the k-NN relations" (Boukerche et al., 2020, p. 12). Also tree-based approaches belong to these methods and they are based on the idea of mapping the original data points to specific tree nodes which contain proximity information. An example is the Isolation Forest method meaning that multiple Isolation Trees represent an Isolation Forest and "can be viewed as the unsupervised counterpart of decision trees [...and] the intuition behind is that outliers have a higher chance of being isolated on an earlier stage than normal data instances [...and therefore] have a shorter height in the isolation trees" (Boukerche et al., 2020, p. 14). Boukerche et al. (2020) also describe several other methods that can be used for outlier-detection in high-dimensional data. Examples for such methods are an angle-based outlier detection method (ABOD) where the "outlier score for a data points relies on the variance of the angles having that data point as the vertex, weighted by the distances to the pair of other data points” (Boukerche et al., 2020, p. 16). Another approach is “based on the assumption that outliers in high-dimensional data are hidden in multiple subspaces that exhibit non-uniformity and high contrast“ (Boukerche et al., 2020, p. 17). The method is called High Contrast Subspaces (HiCS) and measures the contrast of subspaces (see for details Boukerche et al., 2020, p. 17). However, they still see challenges regarding outlier detection in high-dimensional data especially with regard to the use of subspace methods. Furthermore, Boukerche et al. (2020) describe several outlier detection methods that can be used in data streams and also describe approaches that can handle big data. In addition, they also describe methods based on deep learning. Even though these outlier detection methods have not often been applied to detect disinformation, there is a high potential in doing so and further research should address these methods with a focus on methods used for high-dimensional data and based on unsupervised methods (cp. Zhang & Ghorbani, 2020).

3 A discussion of disinformation detection methods

In order to assess how disinformation detection can be improved, the described detection methods need to be discussed. Also in the discussion, we focus on practice-oriented approaches and research-oriented approaches separately.

3.1 Practice-oriented approaches

Even though fact-checking tools/ services provided by non-profit and profit organizations help to detect mis- and disinformation, several aspects need to be addressed in order to improve these services or to discuss their limitations. We will discuss the aspects media coverage, methods, independence, audience, reach and influence, transparency, comparability and velocity in the following section. Table 3 provides an overview and description of these aspects and characteristics for the practice-oriented tools described in 2.1.

Table 3 Characteristics of practice-oriented approaches of disinformation detection

	non-profit oriented (examples: FactCheck.org/ PolitiFact)	**profit-oriented** (example: Google Fact Checking Tools)
media coverage	broad media coverage	broad media coverage
methods	expert-oriented approach to fact-checking	outsourcing of fact-checking
independence	non-partisan orientation	profit-orientation
audience	social influencer, journalists, educators	fact-checkers, journalists and researchers
reach/ influence	limited transparency	limited transparency
transparency	decision processes intransparent, sources and debunked stories transparent	database construction partly intransparent, sources and debunked stories transparent
comparability	within platform comparability given, across platform comparability limited	within platform comparability limited
velocity	medium velocity of publishing	higher velocity of publishing

Media Coverage ¬ Some fact-checking tools or services not only take social media and online sources into consideration, but also news media or television ¬ as it is the case for FactCheck.org or PolitiFact. The Google Fact Checking Explorer also provides debunked stories about claims from diverse media and in diverse formats. Depending on a researcher's approach and research question, it could be important to distinguish between different sources and enable filtering based on the publishing channel of the mis- or disinformation. That is, to increase comparability and facilitate analyses of debunked stories, the information, which media channel and which format was checked and the debunked story published on, is essential. In this regard, some improvements are necessary in the documentation of the debunked stories.

Methods ¬ The methods to debunk dis-or misinformation differ by fact-checking organization. Even though some fact-checking organizations and service providers rely on automated disinformation detection methods ¬ for example Classify.news or Factmata.com ¬ (Zhang & Ghorbani, 2020), FactCheck.org and PolitiFact both have an expert-oriented approach to disinformation detection. That is, journalists or experts write the debunked stories. The Google Fact-Checking Explorer also allows for searches of debunked stories that are based on this approach. Google itself does not fact-check claims but outsources this task to third parties. Exhaustive information is missing about the constitution of the database of the Fact Checking Explorer. For example, information about the number of eligible organizations and the selection of debunked stories by eligible organizations is missing. The time lack between the publishing of a claim and its debunking as dis- or misinformation is thereby a disadvantage of the expert-oriented approach. The limitation regarding the number of facts or claims that can be checked within a given timespan, relying on the expert-oriented approach, poses another disadvantage of the expert-oriented approach.

Independence ¬ Independence is one of the main goals of fact-checking service providers (Singer, 2019). All organizations belonging to the IFCN network, follow for example the developed code of principles. Some of these principles directly address the issue of independence. One of the commitments is to nonpartisanship and fairness, meaning that fact-checking follows a standard procedure and does not focus on any one side. Conclusions are drawn based on the evidence and no policy positions on a specific fact-checked issue are taken. Furthermore, funding is made transparent and if funding is accepted from another organization, this organization has no influence on drawn conclusions by the fact-checking organization. Both ¬ FactCheck.org and PolitiFact ¬ are members of the IFCN.[17] The Google Fact Checking Tools are run independently by Google ¬ a profit-oriented company ¬ and Google sets its standards for reporting on fact-checking stories. The Google Fact-Check Explorer includes debunked stories from organizations, which committed to standards such as the IFCN code of principles, however, transparency regarding the eligibility of organizations for contributing to the database is limited and should be improved.

Audience ¬ A study conducted by Singer (2019) shows that the fact-checking service providers she interviewed (among them PolitiFact and FactCheck.org) see social influencers as their audience. That is, politicians and policymakers get information on whether and how they have been checked. The second audience they identify are journalists. Furthermore, they mention educators as an audience. In Singer's study, the interviewees also show some interest in expanding the audience of the fact-checking services or tools. Thus, a discussion of whom to address and how is still in progress. According to a study of Mena (2019), journalists see the purpose of fact-checking mainly in the evaluation of the accuracy of statements made by public figures but also as a means to uphold journalistic ideals and to debunk disinformation spread on social media. Researchers are not seen explicitly as an audience so far and an open question is how researchers perceive the potential and purposes of fact-checking. The audience for the Google fact-checking tools is defined broadly, but mainly fact-checkers, journalists and researchers are addressed[18]. Even though researchers are mentioned as an audience, research could be facilitated for example with a provision of files that can be analyzed. This also applies to the non-profit oriented providers of fact-checking services. So far, data about, or of, the debunked stories has to be scraped by the researchers or users themselves. The issues of transparency and comparability also play a role for the usage of the provided services for researchers and journalists. In general, a stronger dialogue between researchers and fact-checker organizations is needed in order to improve fact-checking and

[17] https://www.poynter.org/ifcn-fact-checkers-code-of-principles/, last access October 9th, 2020

[18] https://toolbox.google.com/factcheck/about#fcmt-creators, last access October 9th, 2020

disinformation detection (Dias & Sippitt, 2020). Hereby, services by fact-checking organizations can profit from research in order to improve methods but can also serve as study subjects themselves.

Reach/ Influence ¬ Another aspect that needs to be addressed regarding practical fact-checking tools is their reach. Data is needed to assess to what extent and by whom these websites are used and with which consequences. To some extent, studies already provide examinations of the impact of false news (e.g. Fletcher, Richard et al., 2018; Vargo et al., 2018), but fact-checking tools themselves have not been explored often. First, the transparency of the tools regarding usage is limited. More information is needed about who and how often and how users use the services. However, in order to assess how often they are used, actual user numbers are needed in addition to studies, which base their findings on survey data. Robertson et al. (2020) examine how fact-checking services in the U.S. are perceived based on survey data and show that they are perceived as more positive by liberals than by conservatives and that fact-checking sites despite proclaiming non-partisanship are perceived as being political by the audience. In addition, more studies should address whether the perception of debunked stories or the use of fact-checking services and tools leads to a change in behavior or attitudes. Nyhan & Reifler (2015), for example, show that fact-checking can potentially have an influence on the behavior of politicians.

Transparency ¬ Transparency is also a main goal of fact-checking service providers and addresses methods as well as data (Singer, 2019). Transparency is also one means to establish trust in the services (Shawcross, Alistair, 2016). However, the level of transparency can be increased. The services provided by FactCheck.org, PolitiFact and other organizations need a higher transparency of why claims are fact-checked. Even though the procedures are often made transparent, decision processes regarding the selection of fact-checking stories have been criticized for lacking transparency (Nieminen & Rapeli, 2019). A numbering of how many claims have been suggested for fact-checking, how many of them were finally checked and published would be one option to improve transparency regarding decision processes. FactCheck.org also lacks transparency regarding the rating of claims (cp. Zhang & Ghorbani, 2020). In general, information is missing about how valid and reliable ratings are, especially if the rating is provided in a non-binary way, as is the case for PolitiFact. The context of the debunked stories ¬ that is for example information about the author of the debunked story, the source and publication of the claim and the sources for an evaluation ¬ are provided. Regarding the Google Fact Checking Tools, transparency is given for the context and content of the debunked stories that are in the database, however, less information is available about the condition and actual constitution of the database.

Comparability ¬ Fact-checking tools vary concerning the content they provide and their scope. Lim (2018) compares for example two policy related fact-checking tools and finds that they only to some extent overlap a. regarding the examined topic and b. regarding their rating of a statement. Also Lowrey (2017) finds that fact-checking sites within the U.S. that are developed in between 2011 and 2013 show a higher diversity than those developed before. However, the question of how consistent fact-checking organizations are needs additional research (Nieminen & Rapeli, 2019). In order to facilitate and improve fact-checking via the fact-checking tools provided by diverse organizations, it is necessary to provide a transparent description of how and why statements are checked, to establish means of validation and to harmonize outputs to facilitate comparisons across providers.

Regarding validation, information about the sources, which were used to check a claim, are essential ¬ and this information is often already available. Furthermore, services need to be controlled by externs to verify that checks were conducted correctly and on the basis of the right material. Information about to what extent this is already done is not made public so far. Regarding the harmonization of debunked stories and documentation, improvements are necessary. FactCheck.org and PolitiFact use the same procedures for all the stories, which are checked within their organization, however, the outcome across these two organizations is comparable only to some extent. The schema.org-ClaimReview-Markup used by Google, provides standards for the markup of claims[19]. However, it is unclear which information is mandatory for the documentation of debunked stories. Thus, the search results of the Google Fact Check Explorer provide some overlapping information, but often information is missing and not comparable since the contributors to the database work with different procedures and details about the debunked stories are only provided on the website of the third-parties that published the story. This hinders cross-platform and cross-national analysis. Services such as TrulyMedia or the collective database #CoronaVirusFacts Alliance of the IFCN are good examples of how to increase comparability across fact-checking organizations.

Velocity ¬ Furthermore, many fact-checking websites operate on the basis of expertise provided by established journalists or researchers to rate information and detect disinformation. This expert-oriented approach is very time-consuming (Shu et al., 2019) and leads to a time lack between the publishing or declaring of facts or claims and their checking by fact-checking providers. This time lack is even more detrimental in times in which a large amount of information is spread in high velocity via social media and due to technology and new forms of journalism (Chen et al., 2015; Hassan, Naeemul et al., 2015). Hassan, Naeemul et al. (2015) point out several disadvantages associated with human-based fact-checking. Not only is the fact-checking time consuming but also requires a higher level of skills regarding research and writing both amplifying the time lack. These disadvantages lead to a high need of automated fact-checking which ideally would make fact-checking more effective by reducing the time lack and increasing the amount of information that can be checked.

3.2 Research-oriented approaches

In this section, we will address some challenges for disinformation detection associated with the described research-oriented approaches. Research-oriented approaches have the potential to overcome some of the outlined disadvantages associated with practice-oriented approaches. Automated detection enables us to handle the high velocity and extent to which disinformation spreads and to cover many different topics, datasets and formats. However, also regarding research-oriented approaches we still face challenges. We will address several challenges here ¬ namely generalizability, multi-modality, component related challenges, time related challenges, methodological as well as conceptual challenges and the issue of purpose.

[19] https://schema.org/ClaimReview; last access October 9th, 2020

Generalizability ¬ Most detection mechanisms are based on an assumption of how the disinformation was generated (Shu et al., 2020). However, it is not often assessed whether this assumption holds for disinformation in general and can be applied to other disinformation datasets (also see Guo et al., 2020). Furthermore, most detection strategies are based on learning the identification of event-specific features and therefore cannot be transferred to newly arising events. Thus, there is no method that can be used for all topics or kinds of datasets. Newly developed methods, however, allow for an extraction of event-invariant features and need to be taken into consideration in future research (Shu et al., 2020).

Dimension of time ¬ Even though automated approaches have the potential of detecting disinformation in near-real time, the challenge of early disinformation detection still remains, as the researcher does not know responses in an early stage. Only recently methods are developed that have the potential at detecting fake news early on (Guo et al., 2020; Shu et al., 2017, 2020) and, thus, more research is needed to address the issue of early detection. Especially since disinformation has the potential to harm from the very beginning of publication.

Multi-modality information ¬ Disinformation detection also depends on the feature of disinformation that is analyzed. Regarding the detection of fake images, Shu et al. (2020) describe the problem that often the methods used by the creator or spreader of a fake image is not known to the ones trying to detect the fake image and this hinders the choice of appropriate methods. Shu et al. (2020) also argue for more research in the area of fake video detection saying that visual components have been neglected (also cp. Guo et al., 2020).

Components of disinformation ¬ Related to the challenge of multi-modality information, we also have to address the issue of different components of disinformation in general. Disinformation detection would be most efficient if different components of disinformation are taken into consideration ¬ for example information about victims, content or distribution patterns. Most studies, however, focused on specific components and are therefore limited in disinformation detection (Shu et al., 2017, 2020). Even though the number of studies that use an integrated approach by combining different disinformation detection methods increased, there is still a need for further studies that combine different methods and look at different components. Shu et al. (2019) propose a Social Article Fusion model that examines linguistic aspects as well as social context of news to detect fake news. Guo et al. (2020) refer to this approach as a feature fusion-based approach. These approaches use content features and social context features comprehensively. Shu et al. (2017) also advocate that more studies are needed for the examination of social context information. Guo et al. (2020), in addition, argue that most common deep learning-based methods focus on decision results instead of the reasons that lead to this decision and that more studies should focus on explanations of decision results in order to increase trust in methods.

Methodological challenges ¬ Supervised, semi-supervised and unsupervised detection methods are also associated with different limitations. Supervised methods rely on labeled datasets, whereas the creation of these datasets is time consuming and requires human experts. Shu et al. (2017) conclude that there is still the need for the creation of a benchmark dataset which allows for a better extraction of relevant features in order to detect disinformation (cp. also Guo et al., 2020; Islam et al., 2020). Furthermore, the use of different methods makes it difficult to compare the results of existing studies and not many studies so far compare and assess different disinformation detection methods and techniques (Habib et al., 2019; Islam et al., 2020; Kumar & Shah, 2018). Furthermore, the issue of multidisciplinarity needs to be addressed. On the one hand, most automated methods require special skills and an expert knowledge and are therefore difficult to apply for most social scientists or psychologists. That is, they need the expertise of scientists who can apply disinformation detection methods. On the other hand, the improvement of disinformation detection methods would profit from an interdisciplinary approach. Several researchers have pointed out additional potential research areas for disinformation detection, which need further investigation. Shu et al. (2020) see potential in the use of threat modeling in order to detect disinformation with threat modeling being "a widely used technique in the field of computer security to identify and combat the threat" (p. 16). Furthermore, they see an examination of disinformation from a psychological point of view as being promising since approaches in psychology probably help to better understand why people show different responses to disinformation than to true information. Thus, an interdisciplinary approach to disinformation detection is needed (also see Guo et al., 2020). Guo et al. (2020) also mention methods used in neuroscience that could be promising to examine cognitive mechanisms of disinformation.

Conceptual challenges ¬ For the distinction between dis - and misinformation the intention behind the creation and publishing of information is essential. However, intention is difficult to detect since explicit indicators are often not available. Most methods so far focus on the assessment of authenticity of and not on intention behind information. Future research should address to what extent data mining methods can be used to examine intentions (Shu et al., 2017). Furthermore, most studies focus on fake news and rumors or hoaxes as specific forms of mis- and disinformation are neglected (Habib et al., 2019). Thus, besides the development of a common understanding of how disinformation or false information is defined, we also need a discussion of how to distinguish different kinds of false information and about which methods can be applied to which kind of false information.

Purposes of disinformation detection ¬ While many approaches focus on how to detect disinformation, less research is done regarding the implications of this detection. That is, how to proceed after disinformation is detected. A further neglected research area is the one of intervention. That is, research about how consequences of disinformation can proactively be prevented by intervening. Proactive intervention methods, for example, a. remove creators of disinformation in order to prevent further spreading or b. help users to be aware of and to identify false information by exposing them to true news (Shu et al., 2017). Furthermore, we need more research about the consequences of disinformation, that is a focus on the victims of disinformation.

4 Actions needed in order to improve disinformation detection

In the previous sections, we addressed several disadvantages and challenges for disinformation detection associated with practice-oriented as well as research-oriented approaches. Based on this discussion, we propose several steps in order to improve disinformation detection.

1. Promotion of interdisciplinary disinformation detection

As practice-oriented approaches often rely on experts and are based on a journalistic perspective, an interdisciplinary approach can facilitate disinformation detection by expediting the process and improving the methods. Organizations which provide practice-oriented disinformation detection services, need to address researchers to a stronger extent by advancing the documentation of the disinformation detection process, by harmonizing this process across organizations and by external validations of the debunked stories. Such improvements would facilitate, for example, an analysis of biases in different disinformation databases that could create down-stream challenges when policy is built on top of such (true) knowledge repositories. In addition, a facilitated access to the data of these organizations for researchers is necessary, for example, to promote the study of the impact of disinformation detection. This access needs to recognize the work processes of researchers involving taking out lists and comparing these dynamic (e.g. json from API access or .csv files) database content lists to large-scale external datasets. Furthermore, these organizations should increase transparency regarding the whole disinformation detection process and implement transparency regarding the usage of the services. Thus, a stronger communication between fact-checkers and academics is necessary in order to address needs and identify potential academic contributions. Compared to U.S. based organizations, European organizations providing fact-checking services especially have to increase their attraction for researchers.

Regarding research-oriented approaches, we also need a stronger exchange between research disciplines and sectors in order to find and improve methods for disinformation detection. We also need to take advantage of methodological expertise and diverse theoretical approaches for a better understanding of how and why disinformation spreads and to advance models for detection, prediction and prevention. Steps to increase interdisciplinarity can be the establishment of an interdisciplinary network for disinformation detection, or the establishment of a disinformation detection database in which research studies and databases can be accessed.

2. The extension of automated disinformation detection

On top of the previous point, organizations which provide practice-oriented disinformation detection services need to continue to implement automated disinformation detection in the disinformation detection process, for example regarding the selection of facts they check. This would increase the transparency of why facts are checked and therefore make replication and verification possible. Furthermore, the disinformation detection process could be organized more efficiently and conducted faster.

3. Consensus regarding definitions and conceptualization

In order to facilitate interdisciplinary research and disinformation detection, further discussions need to take place to find a common understanding of the conceptualization of false information and disinformation in particular and to discuss how different types of false information can be distinguished from each other theoretically as well as methodologically. Furthermore, we need an assessment about the extension of the application of detection methods for specific types of disinformation on other ones.

4. Facilitation of data access and access to information

The importance of access to data and information for interdisciplinary research and an improvement of disinformation detection methods needs to be stressed in more detail. Not only should organizations that provide practice-oriented services for disinformation detection enhance documentation and access to their data and provide additional information about usage and procedures, but discussions should also lead to a stronger focus on how to compare and assess research-oriented disinformation detection methods and to help researchers make a baseline for disinformation by potentially doing more comparison between various databases and repositories with extended documentation designed for research purposes. Standards are needed to validate and assess the success of different methods, in different settings and for different datasets.

5. Exploitation of new research areas

In addition to a better assessment and usage of current disinformation detection methods, we also need further research to improve and extend disinformation detection. It is still challenging to detect disinformation in high dimensional data and early on. For example, research should focus more on the potential of methods used in different fields such as neuroscience or psychology, on the application of different methods, as they are used for example for outlier detection and in a combination of different methods. Regarding the latter, methods that focus on specific components of disinformation, for example on social context or textual features, should be combined. Furthermore, neglected components such as visual features of information and disinformation should gain attention.

5 Conclusion

Regarding disinformation detection, we can distinguish two main approaches ¬ a practice-oriented and a research-oriented approach. Practice-oriented approaches refer to those services provided by fact-checking organizations, which often rely on experts in order to debunk disinformation and to publish debunked stories. Many of these services have a journalistic background. The services reflected most in research papers ¬ namely PolitiFact and FactCheck.org ¬ are located in the U.S. and focus on disinformation in the realm of politics. However, there was a considerable rise in numbers of these services within approximately the last fifteen years also in Europe and across the world. More recent developments are the establishment of an international fact-checking network (IFCN) and the launch of services provided by profit-oriented organisations such as Google which enable both a search for debunked stories and the mark up of disinformation. Practice-oriented approaches to disinformation detection face several challenges especially regarding the high velocity and vast spread of disinformation in times of a high usage of social and digital media. In order to improve disinformation detection practice-oriented approaches should increase transparency and improve documentation with regard to the source of fact-checked information and the detection and publishing process, with regard to the usage of the services or the constitution of the database. Furthermore, researchers need to be addressed as an audience explicitly and research should be facilitated for example by the provision of data files that can be analyzed. Also a harmonization of debunked stories would help to improve research. In addition, practice-oriented services should continue to implement automated disinformation detection in their procedures. Research-oriented approaches can be applied to a larger variety of datasets and topics and have the potential to overcome some of the disadvantages associated with practice-oriented disinformation detection. Research-oriented approaches usually use automated methods in order to detect abnormalities related to specific components of disinformation ¬ for example, natural language processing methods is used to detect abnormalities within the text body. Potential improvements of these research-oriented approaches relate to event-invariant and early disinformation detection as well as to methods, which use neglected features of disinformation, for example virtual ones, or use several components of disinformation, for example social context information as well as linguistic aspects, in order to detect it. Furthermore, more research is needed especially regarding methods that can be applied in high-dimensional data or/and are used for outlier detection. In addition, more studies are needed that compare and assess methods. There are also neglected aspects of disinformation, which need further attention by researchers such as the identification of intentions or the impact of disinformation. Research should also apply a more interdisciplinary approach across disciplines such as computer science, neuroscience, psychology or social sciences in order to improve disinformation detection methods. In general, we argue for a promotion of interdisciplinary disinformation detection, for an extension of automated disinformation detection, for further discussions regarding the conceptualization and definition of disinformation, for a facilitated access to information and data and for additional research in order to improve disinformation detection.

6 References

Aggarwal, C. C. (2017). *Outlier Analysis*. Springer International Publishing. https://doi.org/10.1007/978-3-319-47578-3

Al Asaad, B., & Erascu, M. (2018). A Tool for Fake News Detection. *2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, 379–386. https://doi.org/10.1109/SYNASC.2018.00064

Amazeen, M. A. (2020). Journalistic interventions: The structural factors affecting the global emergence of fact-checking. *Journalism*, *21*(1), 95–111. https://doi.org/10.1177/1464884917730217

Boukerche, A., Zheng, L., & Alfandi, O. (2020). Outlier Detection: Methods, Models, and Classification. *ACM Computing Surveys*, *53*(3), 1–37. https://doi.org/10.1145/3381028

Carmi, E., Yates, S. J., Lockley, E., & Pawluczuk, A. (2020). Data citizenship: Rethinking data literacy in the age of disinformation, misinformation, and malinformation. *Internet Policy Review*, *9*(2). https://doi.org/10.14763/2020.2.1481

Chen, Y., Conroy, N. K., & Rubin, V. L. (2015). News in an online world: The need for an "automatic crap detector." *Proceedings of the Association for Information Science and Technology*, *52*(1), 1–4. https://doi.org/10.1002/pra2.2015.145052010081

Dias, N., & Sippitt, A. (2020). Researching Fact Checking: Present Limitations and Future Opportunities. *The Political Quarterly*, 1467-923X.12892. https://doi.org/10.1111/1467-923X.12892

EU DisinfoLab. (2020). *The Few Faces of Disinformation* (p. 7). https://www.disinfo.eu/wp-content/uploads/2020/05/20200512_The-Few-Faces-of-Disinformation.pdf

Fletcher, Richard, Cornia, Alessio, Graves, Lucas, & Kleis Nielsen, Rasmus. (2018). *Measuring the reach of "fake news" and online disinformation in Europe* (FACTSHEET, p. 10). Reuters Institute for the Study of Journalism; University of Oxford. https://reutersinstitute.politics.ox.ac.uk/our-research/measuring-reach-fake-news-and-online-disinformation-europe

Gelfert, A. (2018). Fake News: A Definition. *Informal Logic*, *38*(1), 84–117. https://doi.org/10.22329/il.v38i1.5068

Graves, L. (2018). Boundaries Not Drawn: Mapping the institutional roots of the global fact-checking movement. *Journalism Studies*, *19*(5), 613–631. https://doi.org/10.1080/1461670X.2016.1196602

Guo, B., Ding, Y., Yao, L., Liang, Y., & Yu, Z. (2020). The Future of False Information Detection on Social Media: New Perspectives and Trends. *ACM Computing Surveys*, *53*(4), 1–36. https://doi.org/10.1145/3393880

Habib, A., Asghar, M. Z., Khan, A., Habib, A., & Khan, A. (2019). False information detection in online content and its role in decision making: A systematic literature review. *Social Network Analysis and Mining*, *9*(1), 50. https://doi.org/10.1007/s13278-019-0595-5

Hassan, Naeemul, Adair, Bill, Hamilton, James T., Li, Chengkai, Tremayne, Mark, Yang, Jun, & Yu, Cong. (2015). *The Quest to Automate Fact-Checking* (Proceedings of the Computation and Journalism Symposium). http://cj2015.brown.columbia.edu/papers/automate-fact-checking.pdf

Ioannis, Konstantinidis. (2018). *Disinformation Detection with Model Explanations* [Master of Science thesis in Data Science, International Hellenic University]. https://repository.ihu.edu.gr/xmlui/handle/11544/29281

Islam, M. R., Liu, S., Wang, X., & Xu, G. (2020). Deep learning for misinformation detection on online social networks: A survey and new perspectives. *Social Network Analysis and Mining*, *10*(1), 82. https://doi.org/10.1007/s13278-020-00696-x

Kumar, S., & Shah, N. (2018). False Information on Web and Social Media: A Survey. *ArXiv:1804.08559 [Cs]*. http://arxiv.org/abs/1804.08559

Lim, C. (2018). Checking how fact-checkers check. *Research & Politics*, *5*(3), 205316801878684. https://doi.org/10.1177/2053168018786848

Lowrey, W. (2017). The Emergence and Development of News Fact-checking Sites: Institutional logics and population ecology. *Journalism Studies*, *18*(3), 376–394. https://doi.org/10.1080/1461670X.2015.1052537

Mantas, Harrison. (2020, May 29). Fact-checkers support Twitter labels, but more than that, they want transparency. *Https://Www.Poynter.Org/Fact-Checking/2020/Fact-Checkers-Support-Twitter-Labels-but-More-than-That-They-Want-Transparency/*.

Mena, P. (2019). Principles and Boundaries of Fact-checking: Journalists' Perceptions. *Journalism Practice*, *13*(6), 657–672. https://doi.org/10.1080/17512786.2018.1547655

Nieminen, S., & Rapeli, L. (2019). Fighting Misperceptions and Doubting Journalists' Objectivity: A Review of Fact-checking Literature. *Political Studies Review*, *17*(3), 296–309. https://doi.org/10.1177/1478929918786852

Nyhan, B., & Reifler, J. (2015). The Effect of Fact-Checking on Elites: A Field Experiment on U.S. State Legislators: THE EFFECT OF FACT-CHECKING ON ELITES. *American Journal of Political Science*, *59*(3), 628–640. https://doi.org/10.1111/ajps.12162

Pham, Sherisse. (2020, June 3). Twitter says it labels tweets to provide ´context, not fact-checking´. *CNN Business*.

Robertson, C. T., Mourão, R. R., & Thorson, E. (2020). Who Uses Fact-Checking Sites? The Impact of Demographics, Political Antecedents, and Media Use on Fact-Checking Site Awareness, Attitudes, and Behavior. *The International Journal of Press/Politics*, *25*(2), 217–237. https://doi.org/10.1177/1940161219898055

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, *9*(1), 4787. https://doi.org/10.1038/s41467-018-06930-7

Shawcross, Alistair. (2016). *Facts We Can Believe In: How to make fact-checking better* (Beyond Propaganda Series, p. 44). Legatum Institute; Transitons Forum. https://www.lse.ac.uk/iga/assets/documents/arena/archives/facts-we-can-believe-in-how-to-make-fact-checking-better-web-pdf.pdf

Shu, K., Bhattacharjee, A., Alatawi, F., Nazer, T. H., Ding, K., Karami, M., & Liu, H. (2020). Combating disinformation in a social media age. *WIREs Data Mining and Knowledge Discovery*. https://doi.org/10.1002/widm.1385

Shu, K., Mahudeswaran, D., & Liu, H. (2019). FakeNewsTracker: A tool for fake news collection, detection, and visualization. *Computational and Mathematical Organization Theory*, *25*(1), 60–71. https://doi.org/10.1007/s10588-018-09280-3

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36. https://doi.org/10.1145/3137597.3137600

Singer, J. B. (2019). Fact-checkers as Entrepreneurs. *Journalism Practice*, *13*(8), 976–981. https://doi.org/10.1080/17512786.2019.1646613

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining "Fake News": A typology of scholarly definitions. *Digital Journalism*, *6*(2), 137–153. https://doi.org/10.1080/21670811.2017.1360143

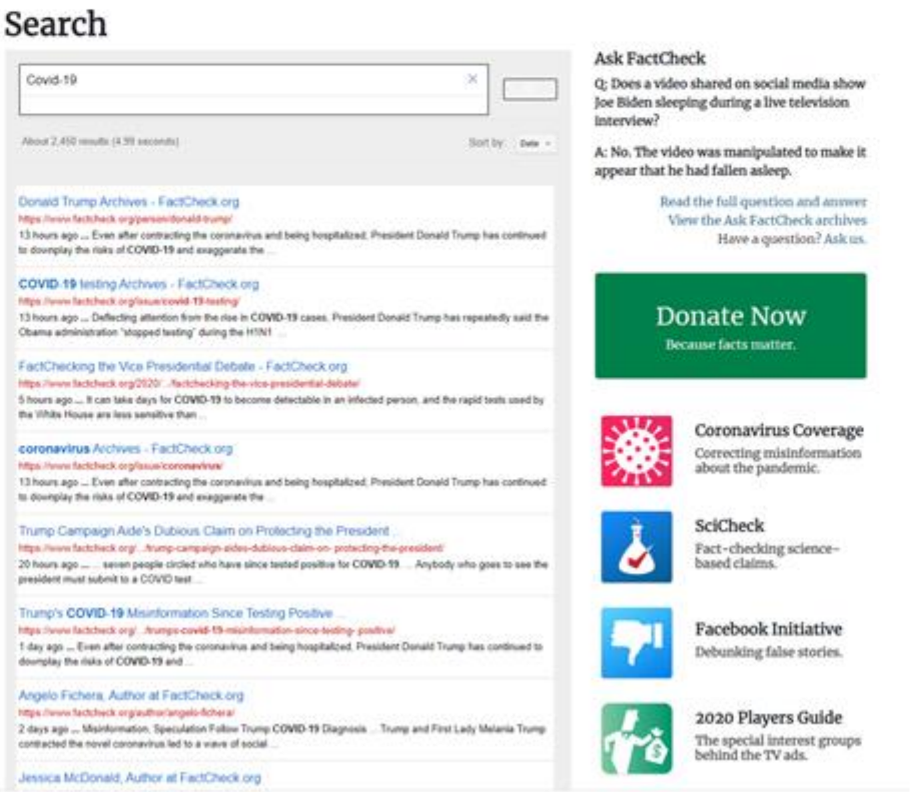
Tsabouraki, Danae, Klitsi, Marina, & Sarris, Nikos. (2018). *D3.1 Social Media Observatory Guide* (p. 40) [Report for SOMA (Social Observatory for Disinformation and Social Media Analysis)]. ATC. https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5c04fb352&appId=PPGMS

Vargo, C. J., Guo, L., & Amazeen, M. A. (2018). The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New Media & Society*, *20*(5), 2028–2049. https://doi.org/10.1177/1461444817712086

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

Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, *57*(2), 102025. https://doi.org/10.1016/j.ipm.2019.03.004

7 Appendix



Search

Covid-19

About 2,450 results (4.59 seconds)

Sort by: Date

Donald Trump Archives - FactCheck.org
https://www.factcheck.org/person/donald-trump/
13 hours ago ... Even after contracting the coronavirus and being hospitalized, President Donald Trump has continued to downplay the risks of COVID-19 and exaggerate the ...

COVID-19 testing Archives - FactCheck.org
https://www.factcheck.org/issue/covid-19-testing/
13 hours ago ... Deflecting attention from the rise in COVID-19 cases, President Donald Trump has repeatedly said the Obama administration "stopped testing" during the H1N1 ...

FactChecking the Vice Presidential Debate - FactCheck.org
https://www.factcheck.org/2020/.../factchecking-the-vice-presidential-debate/
5 hours ago ... It can take days for COVID-19 to become detectable in an infected person, and the rapid tests used by the White House are less sensitive than ...

coronavirus Archives - FactCheck.org
https://www.factcheck.org/issue/coronavirus/
13 hours ago ... Even after contracting the coronavirus and being hospitalized, President Donald Trump has continued to downplay the risks of COVID-19 and exaggerate the ...

Trump Campaign Aide's Dubious Claim on Protecting the President ...
https://www.factcheck.org/.../trump-campaign-aides-dubious-claim-on- protecting-the-president/
20 hours ago ... ... seven people circled who have since tested positive for COVID-19. ... Anybody who goes to see the president must submit to a COVID test ...

Trump's COVID-19 Misinformation Since Testing Positive ...
https://www.factcheck.org/.../trumps-covid-19-misinformation-since-testing- positive/
1 day ago ... Even after contracting the coronavirus and being hospitalized, President Donald Trump has continued to downplay the risks of COVID-19 and ...

Angelo Fichera, Author at FactCheck.org
https://www.factcheck.org/author/angelo-fichera/
2 days ago ... Misinformation, Speculation Follow Trump COVID-19 Diagnosis ... Trump and First Lady Melania Trump contracted the novel coronavirus led to a wave of social ...

Jessica McDonald, Author at FactCheck.org

Ask FactCheck

Q: Does a video shared on social media show Joe Biden sleeping during a live television interview?

A: No. The video was manipulated to make it appear that he had fallen asleep.

Read the full question and answer
View the Ask FactCheck archives
Have a question? Ask us.

Donate Now
Because facts matter.

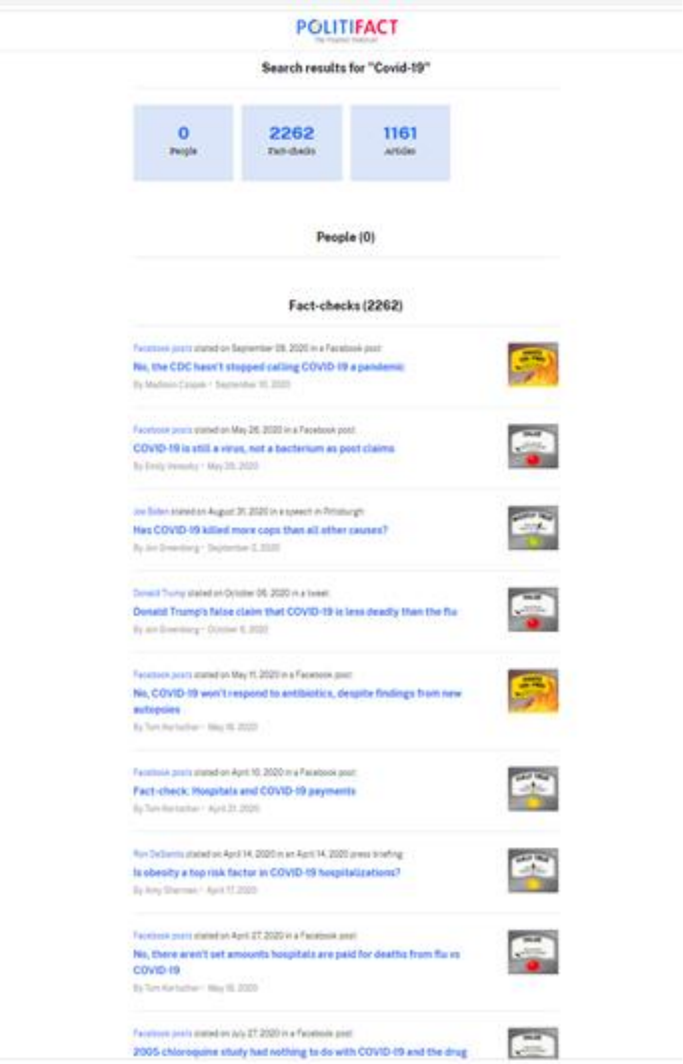
Coronavirus Coverage
Correcting misinformation about the pandemic.

SciCheck
Fact-checking science-based claims.

Facebook Initiative
Debunking false stories.

2020 Players Guide
The special interest groups behind the TV ads.

Figure 1A Example for presentation of search results at FactCheck.org; Note: Source https://www.factcheck.org/search/#gsc.tab=0&gsc.q=Covid-19&gsc.sort=date; screenshot taken October 8th, 2020



POLITIFACT

Search results for "Covid-19"

0 People	2262 Fact-checks	1161 Articles

People (0)

Fact-checks (2262)

Facebook posts stated on September 09, 2020 in a Facebook post:
No, the CDC hasn't stopped calling COVID-19 a pandemic
By Madison Czopek • September 10, 2020

Facebook posts stated on May 26, 2020 in a Facebook post:
COVID-19 is still a virus, not a bacterium as post claims
By Emily Venezky • May 28, 2020

Joe Biden stated on August 31, 2020 in a speech in Pittsburgh:
Has COVID-19 killed more cops than all other causes?
By Jon Greenberg • September 2, 2020

Donald Trump stated on October 06, 2020 in a tweet:
Donald Trump's false claim that COVID-19 is less deadly than the flu
By Jon Greenberg • October 6, 2020

Facebook posts stated on May 11, 2020 in a Facebook post:
No, COVID-19 won't respond to antibiotics, despite findings from new autopsies
By Tom Kertscher • May 18, 2020

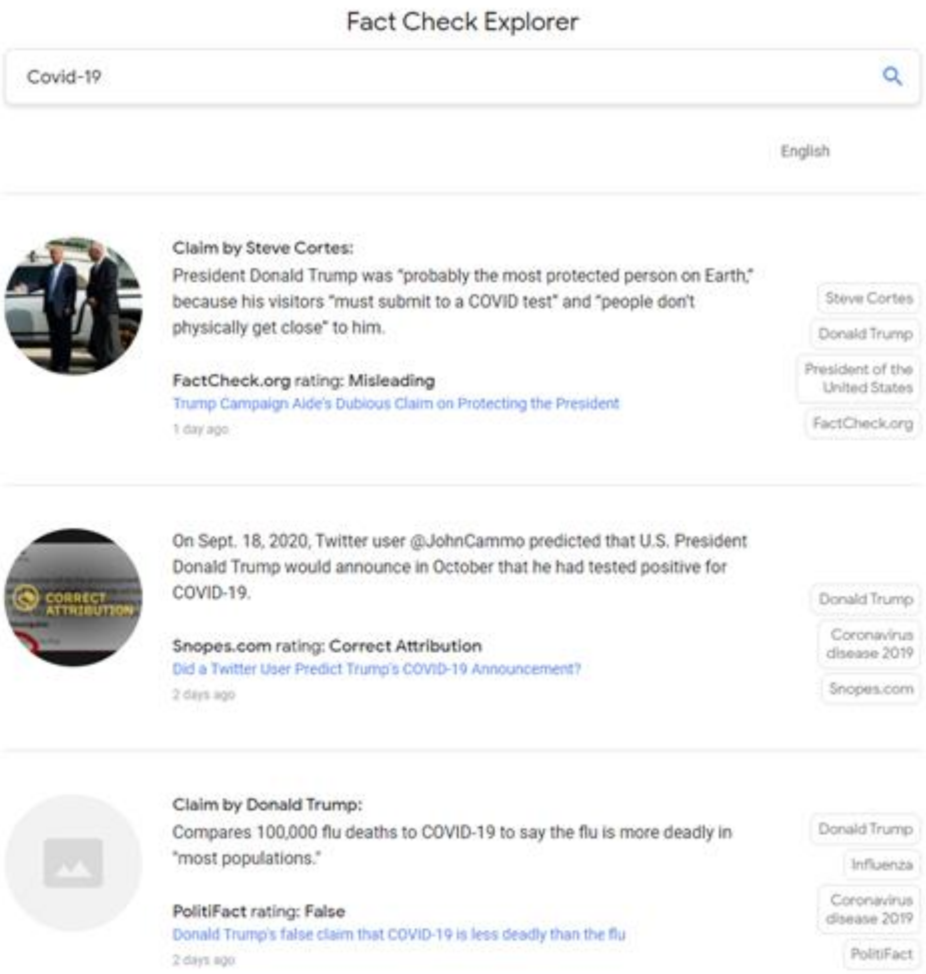
Facebook posts stated on April 10, 2020 in a Facebook post:
Fact-check: Hospitals and COVID-19 payments
By Tom Kertscher • April 21, 2020

Ron DeSantis stated on April 14, 2020 in an April 14, 2020 press briefing:
Is obesity a top risk factor in COVID-19 hospitalizations?
By Amy Sherman • April 17, 2020

Facebook posts stated on April 27, 2020 in a Facebook post:
No, there aren't set amounts hospitals are paid for deaths from flu vs COVID-19
By Tom Kertscher • May 18, 2020

Facebook posts stated on July 27, 2020 in a Facebook post:
2005 chloroquine study had nothing to do with COVID-19 and the drug

Figure 2A Example for presentation of search results at PolitiFact; Note: Source: https://www.politifact.com/search/?q=Covid-19, screenshot taken October 8th, 2020



Fact Check Explorer

Covid-19

English

Claim by Steve Cortes:
President Donald Trump was "probably the most protected person on Earth," because his visitors "must submit to a COVID test" and "people don't physically get close" to him.

FactCheck.org rating: Misleading
Trump Campaign Aide's Dubious Claim on Protecting the President
1 day ago

Steve Cortes
Donald Trump
President of the United States
FactCheck.org

On Sept. 18, 2020, Twitter user @JohnCammo predicted that U.S. President Donald Trump would announce in October that he had tested positive for COVID-19.

Snopes.com rating: Correct Attribution
Did a Twitter User Predict Trump's COVID-19 Announcement?
2 days ago

Donald Trump
Coronavirus disease 2019
Snopes.com

Claim by Donald Trump:
Compares 100,000 flu deaths to COVID-19 to say the flu is more deadly in "most populations."

PolitiFact rating: False
Donald Trump's false claim that COVID-19 is less deadly than the flu
2 days ago

Donald Trump
Influenza
Coronavirus disease 2019
PolitiFact

Figure 3A Example for presentation of debunked stories for the Google Fact Check Explorer; Note: Source: https://toolbox.google.com/factcheck/explorer/search/Covid-19;hl=en; screenshot taken on October 8th, 2020

Table 1A Search results of the literature review

Search terms	Author(s)	Title
“disinformation detection” AND “online” + ”disinformation detection” AND ”social media”	Alaphilippe et al. (2018)	Disinformation detection system: 2018 Italian elections
“disinformation detection” AND “online” + ”disinformation detection” AND ”social media”	Chorás et al. (2019)	SocialTruth Project Approach to Online Disinformation (Fake News) Detection and Mitigation
“disinformation detection” AND “online”	Collado et al. (2020)	Falling victims to online disinformation among young Filipino people: Is human mind to blame?
“disinformation detection” AND “online”	Fallis (2015)	What is Disinformation?
“disinformation detection” AND “online”	Hounsel et al. (2020)	Identifying Disinformation Websites Using Infrastructure Features
“disinformation detection” AND “online” + “disinformation detection” AND ”social media”	Pierri et al. (2020)	A multi-layer approach to disinformation detection on Twitter
“disinformation detection” AND “online”	Pierri et al. (2020)	HoaxItaly: a collection of Italian disinformation and fact-checking stories shared on Twitter in 2019
“disinformation detection” AND “online”	Shu et al. (2020)	Combating disinformation in a social media age
“disinformation detection” AND “online” + ”disinformation detection” AND ”social media”	Vargas et al. (2020)	On the Detection of Disinformation Campaign Activity with Network Analysis

Continuation Table 1A

“disinformation detection” AND “online” + ”disinformation detection” AND ”social media”	Yu & Lo (2020)	Disinformation Detection using Passive Aggressive Algorithms
”disinformation detection” AND ”social media”	Alam et al. (2020)	Fighting the COVID-19 Infodemic in Social Media: A Holistic Perspective and a Call to Arms
”disinformation detection” AND ”social media”	Shu et al. (2020)	Mining Disinformation and Fake News: Concepts, Methods, and Recent Advancements
”disinformation detection” AND ”social media”	Shu et al. (2020)	Combating disinformation in a social media age
”disinformation detection” AND ”social media”	Wolverton & Stevens (2020)	The impact of Personality in recognizing disinformation
“false information detection” AND “online”	Arshad et al. (2018)	A survey of local/cooperative-based malicious information detection techniques in VANETs
“false information detection” AND “online” + “false information detection” AND “social media”	Dong et al. (2019)	Dual-stream Self-Attentive Random Forest for False Information Detection
“false information detection” AND “online” + “false information detection” AND “social media”	Ghanem et al. (2020)	An Emotional Analysis of False Information in Social Media and News Articles

Continuation Table 1A

“false information detection” AND “online” + “false information detection” AND “social media”	Guo et al. (2020)	The Future of False Information Detection on Social Media: New Perspectives and Trends
“false information detection” AND “online” + “false information detection” AND “social media”	Habib et al. (2019)	False information detection in online content and its role in decision making: a systematic literature review
“false information detection” AND “online” + “misinformation detection” AND “online”	Islam et al. (2020)	Deep learning for misinformation detection on online social networks: a survey and new perspectives
“false information detection” AND “online” + “false information detection” AND “social media”	Kumar & Shah (2018)	False Information on Web and Social Media: A Survey
“false information detection” AND “online” + “false information detection” AND “social media”	Wu et al. (2018)	False Information Detection on Social Media via a Hybrid Deep Model
“false information detection” AND “online”	Zannettou et al. (2019)	The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans
“false information detection” AND “online” + “false information detection” AND “social media”	Zhang & Ghorbani (2020)	An overview of online fake news: Characterization, detection, and discussion

Continuation Table 1A

“false information detection” AND “social media”	Allcott & Gentzkow (2017)	Social Media and Fake News in the 2016 Election
“false information detection” AND “social media”	Giachanou et al. (2019)	Leveraging Emotional Signals for Credibility Detection
“false information detection” AND “social media”	Tian et al. (2020)	QSAN: A Quantum-probability based Signed Attention Network for Explainable False Information Detection
“fake news detection” AND “online”	Ahmed et al. (2017)	Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques
“fake news detection” AND “online”	Ajao et al. (2019)	Sentiment Aware Fake News Detection on Online Social Networks
“fake news detection” AND “online”	Conroy et al. (2016)	Automatic deception detection: Methods for finding fake news
“fake news detection” AND “online”	Della Vedova et al. (2018)	Automatic Online Fake News Detection Combining Content and Social Signals
“fake news detection” AND “online”	Karimi & Tang (2019)	Learning Hierarchical Discourse-level Structure for Fake News Detection
“fake news detection” AND “online”	Ozbay & Alatas (2020)	Fake news detection within online social media using supervised artificial intelligence algorithms
“fake news detection” AND “online”	Pérez-Rosas et al. (2017)	Automatic Detection of Fake News

Continuation Table 1A

“fake news detection” AND “online” + “fake news detection” AND “social media”	Shu et al. (2017)	Fake News Detection on Social Media: A Data Mining Perspective
“fake news detection” AND “online” + “fake news detection” AND “social media”	Tschiatschek et al. (2018)	Fake News Detection in Social Networks via Crowd Signals
“fake news detection” AND “online”	Zhang et al. (2018)	Fake News Detection with Deep Diffusive Network Model
“fake news detection” AND “social media”	Guo et al. (2019)	Exploiting Emotions for Fake News Detection on Social Media
“fake news detection” AND “social media”	Okoro et al. (2018)	A Hybrid Approach to Fake News Detection on Social Media
“fake news detection” AND “social media”	Monti et al. (2019)	Fake News Detection on Social Media using Geometric Deep Learning
“fake news detection” AND “social media”	Shu et al. (2019)	Beyond News Contents: The Role of Social Context for Fake News Detection
“fake news detection” AND “social media”	Tacchini et al. (2017)	Some Like it Hoax: Automated Fake News Detection in Social Networks
“fake news detection” AND “social media”	Shu et al. (2018)	Understanding User Profiles on Social Media for Fake News Detection
“fake news detection” AND “social media”	Wang et al. (2018)	EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection

Continuation Table 1A

“fake news detection” AND “social media”	Yang et al. (2019)	Unsupervised Fake News Detection on Social Media: A Generative Approach
“misinformation detection” AND “online”	Almaliki (2019)	Online Misinformation Spread: A Systematic Literature Map
“misinformation detection” AND “online”	Antoniadis et al. (2015)	A Model for Identifying Misinformation in Online Social Networks
“misinformation detection” AND “online”	Fernandez & Alani (2018)	Online Misinformation: Challenges and Future Directions
“misinformation detection” AND “online” + “misinformation detection” AND “social media”	Jain et al. (2016)	Towards automated real-time detection of misinformation on Twitter
“misinformation detection” AND “online” + “misinformation detection” AND “social media”	Sharma et al. (2020)	COVID-19 on Social Media: Analyzing Misinformation in Twitter Conversations
“misinformation detection” AND “online”	Wei et al. (2019)	QuickStop: A Markov Optimal Stopping Approach for Quickest Misinformation Detection
“misinformation detection” AND “online”	Zhang et al. (2016)	Detecting misinformation in online social networks before it is too late
“misinformation detection” AND “online”	Zhang et al. (2016)	Misinformation in Online Social Networks: Detect Them All with a Limited Budget
“misinformation detection” AND “online”	Zhang et al. (2015)	Monitor placement to timely detect misinformation in Online Social Networks

Continuation Table 1A

“misinformation detection” AND “social media”	Cardoso Durier da Silva (2019)	Can Machines Learn to Detect Fake News? A Survey Focused on Social Media
“misinformation detection” AND “social media”	Jiang & Wilson (2018)	Linguistic Signals under Misinformation and Fact-Checking: Evidence from User Comments on Social Media
“misinformation detection” AND “social media”	Radjev & Lee (2015)	Fake and Spam Messages: Detecting Misinformation During Natural Disasters on Social Media
“misinformation detection” AND “social media”	Shang et al. (2020)	FauxWard: a graph neural network approach to fauxtography detection using social media comments
“misinformation detection” AND “social media”	Torabi & Taboada (2019)	Big Data and quality data for fake news and misinformation detection
“misinformation detection” AND “social media”	Wu et al. (2019)	Misinformation in Social Media: Definition, Manipulation, and Detection
“misinformation detection” AND “social media”	Wu et al. (2016)	Mining Misinformation in Social Media
“misinformation detection” AND “social media”	Zhang et al. (2019)	Reply-Aided Detection of Misinformation via Bayesian Deep Learning

Note: literature review for disinformation detection with several different keywords; search conducted October 14th, 2020 on google scholar; first 10 hits; results since 2015 taken into consideration if publication medium provided; by search term and in alphabetical order of first author

List of contributors to the IFCN #CoronaVirusFacts Alliance database[20]:

15min.lt, 20 Minutes Fake off, AAP FactCheck, AFP, AfricaCheck, Agência Lupa, Agencia Ocote, Animal Político, Annie Lab, Aos Fatos, Bolivia Verifica, BOOM FactCheck, BuzzFeed Japan, Check Your Fact, CheckNews, Chequeado, Colombiacheck, Congo Check, Convoca.pe, Correctiv, Décrypteurs – Radio-Canada, Delfi Melo Detektorius (Lie Detector), Demagog, Détecteur de rumeurs, Deutsche Presse-Agentur, Digiteye India, Doğruluk Payı, Dubawa, Ecuador Chequea, EFE Verifica, Efecto Cocuyo, Effecinque – SkyTg24, El Surtidor, Ellinika Hoaxes, Estadão Verifica, FactCheck Georgia, Factcheck.kz, FactCheckNI, FactCheck.org, Factcheck.Vlaanderen, FactCrescendo, Factly, Factnameh, Faktabaari/FactBar, Faktograf, Fatabyyano, France 24 Observers, franceinfo, Full Fact, GhanaFact, India Today, INFACT, Istinomer, JTBC news, Källkritikbyrån, La Nación, La Voz de Guanacaste, La Silla Vacía, LeadStories, Les Décodeurs, Maldita.es, MediaWise, Misbar, Mygopen, Myth Detector, Newschecker, Newsmeter.in, NewsMobile, Newtral.es, Nieuwscheckers, Observador, OjoPúblico, Open, Pagella Politica, Periodismo de Barrio, PesaCheck, Poligrafo, PolitiFact, Rappler, Raskrinkavanje, Re:Check, Salud con lupa, Science Feedback, Spondeo Media, StopFake.org, Sure And Share Center MCOT, Taiwan FactCheck Center, TEMPO, Teyit, The Quint, TheJournal.ie, TjekDet.dk, VERA Files, Verificado, Verificador de La República, Vishvas News, Vistinomer, VoxCheck and Washington Post Fact-Checker

[20] https://www.poynter.org/coronavirusfactsalliance/; last access October 16th, 2020