Misinformation and Disinformation on Social Media: An Updated Survey of Challenges and Current Trends

Fabrizio Lo Scudo(✉)

University of Calabria, Arcavacata, Italy
fabrizio.loscudo@unical.it

Abstract. Over the last decade, Social Media has been gradually shaping our world. From the Brexit to Ukraine war, passing through US election and COVID-19, there has been increasing attention on how social media affects our society. This attention has nowadays become an active research field in which researchers from different fields have proposed interdisciplinary solutions mainly aimed at fake news detection and prevention. Although this task is far to be solved.

Fake news detection is intrinsically hard since we have to cope with textual data; moreover the early detection requirement, to prevent wide diffusion, makes things even harder. If we now add a dynamic component to the problem definition we can easily understand why researchers have been keeping proposing new solutions to deal with new nuances of the problem. In this so fast-changing field, it is easy for newcomers to get lost. The scope of this work is not to provide a comprehensive review of the state-of-the-art approaches but instead a quick overview of the recent trends and how current technologies try to deal with the unresolved issues that characterize this task.

Keywords: Misinformation · Deep Learning · Social Media

1 Introduction

The advent of the Web2.0 [92] was introduced with a huge emphasis on collective culture and interoperability among end-users. The key change, compared to the previous generation, was the user-generated content which opened up endless opportunities for interacting and sharing information. But if this new feature has allowed the aggregation of people around common interests, facilitating the contamination of different cultures with healthy values and ethical principles, it also allows for the rapid dissemination of unsubstantiated rumors and incorrect interpretations which very often have often negatively impacted our society [37,91].

In general, the repercussions of bad information include opinion polarization, escalating fear and panic, weakening faith in scientific knowledge, historical negationism, or decreased access to health care. This was especially true during the

C. Comito and D. Talia (Eds.): PerSOM 2022, LNICST 494, pp. 17–40, 2023.
https://doi.org/10.1007/978-3-031-31469-8_2

COVID-19 pandemic since the fast spreading of misleading health information has increased vaccine hesitancy and delays in the provision of health care within population high-risk classes as shown in a recent WHO review [87]. The results of this work show the presence of much evidence that, during a crisis, the quality of the information tends to be low and that the development of adequate countermeasures, such as creating and promoting awareness campaigns, increase the amount of reliable content in mass media along with people's digital and health literacy, are needed. But since those policies require a huge amount of resources and time it is common practice to target the sources: the social media platforms.

Being social media platforms poorly regulated makes them nicely suitable for the task of spreading any kind of information. Of course not any kind of information is harmful to our society, and in this regard, it is useful to clarify the pieces of information we care about through the concept of *information disorder*. As described in [152], the notion of information disorder divides the alteration of information into three categories: mis-, dis-, and malinformation. With the term misinformation, we refer to false or inaccurate pieces of information, such as inaccurate dates, statistics, or translation errors whose degree of deliberately intended to deceive might be sometimes hard to assess. The same cannot be said for the idea of disinformation which is deliberately misleading or biased information, aimed to manipulate reality with narrative artifacts such as conspiracy theories, rumors, or simply propaganda. Finally, malinformation is explained as genuine (private) information about a person or corporate that is deliberately made public with the precise intent to cause harm: one famous example is the Russians hacked the Democrats' emails with the precise intent to unveil details to damage Clinton's reputation during her first presidential run.

This as many other definitions and classifications [54,132] of the possible nature of or way to analyze the information present on social media, and more generally on the web, are useful in the matter of enhancing our understanding; but those concepts are hard to formalize in languages useful for the artificial intelligence as described in [80].

In the following sections we thus report useful insights in the process of formalizing the problem (Sect. 2), we highlight the challenges we have to deal with (Sect. 3), and list recent works that face these issues with deep learning techniques (Sect. 4). Finally, in Sect. 5 we try to describe what, from our point of view, are the most evident shortcomings and possible future research trends.

2 A Socio-technological Problem

Following [107,108] the mis-/dis- information problem should be framed as a socio-technological one. This twofold view of the problem is something uncommon but it might be useful to design new operational features or indicators to be fed into algorithms.

In those works, authors propose a conceptual model, the disinformation and misinformation triangle, under which to capture key elements of harmful information and its spreading and propose interventions at a different level to detect and prevent that from happening. The model explains the spread of mis-/dis-information as the consequence of three causal factors which have to occur simultaneously to have a susceptible reader affected by harmful news which is propagating over social media. In this conceptual model, the factors of interest are the susceptible readers, the (un-)intentionally (false) information, and the medium by which the information reaches the readers.

Now, to prevent the diffusion and, as a consequence, the negative effect of the news on the readers, the authors propose three different kinds of interventions. The first concerns the automated identification of potentially harmful information which should support acts aimed to prevent its spreading. The second describes proactive educational campaigns to enhance a deeper critical judgment within the readers' minds. Third, a more structured legislative regulation of social media. But this last point should imply governments acts to push social media companies away from common marketing strategies [48] in favor of a more healthy society. Because of the complexity of discussing the acting at a legislative level, here we leave this aspect out in favor of a discussion about the first two components of the triangle: readers and information.

2.1 About Readers

In a recent work [74], authors try to highlight the importance of paying more attention to readers. The research questions posed in that study concern how the people, exposed to harmful information, would interpret it and which would be the right tools to intervene to prevent the negative effect. To answer those question authors extends a previous line of work on cognitive and ideologically motivated reasoning by introducing an aspect of information *familiarity*-vs-*novelty* to explain a major vulnerability when people are exposed to novel-vs-everyday news.

From a cognitive perspective, it seems that, in general, many individuals tend to rely on others' (possible famous ones') opinions to build their opinions[1] This form of *laziness* in the critical judgment process has been described in [94,95]. Those works suggest a certain inclination of such people towards believing fake news and such aspect is often exploited by mis-/dis- information makers to strengthen individuals' beliefs. In this regard, [31] highlights how people who experience a long exposition to fabricated information about a certain topic are more susceptible to strengthening their belief in that direction.

The ability to strengthen people's beliefs in a specific direction is the key to unlocking the real power behind mis-/dis- information. As it is shown in [60] stronger beliefs make easier the process of spreading the fake news, via *sharing* and *like*, as long as they match the beliefs. This in turn produces a process that amplifies the diffusion of the message allowing for a wider polarization [140]. At the basis of this phenomenon, there is the so-called *confirmation bias* [88], which is the condition in which people become more interested in the only news that is aligned with what they believe in. Overtime then people also become less

[1] Source: https://www.factcheck.org/2016/11/how-to-spot-fake-news/.

prone to challenge their beliefs with new information and only accept that that supports their views [82]. The analysis of this last point should however not be restricted to the mentioned conditions but should be also understood under the lens of ideologically motivated reasoning.

In [59] the author investigates the people's degree of acceptance of new information when they are exposed to a different political stimulus. In this study not only the acceptance but also the way, people process these new pieces of information is examined. The results show how ideological thinking lowered the people's acceptance level, restricting their interest to the only evidence that supports their own beliefs. Moreover, information processing in such contexts becomes lazier.

The discussion made so far might explain why certain people act irrationally while they are more inclined to misleading information. But it is worth noting that the majority of the cited works are based on exploratory studies due to the lack of theoretical guidance on this topic. Also, the described insights, being human-centered, do not find an easy spot within AI tools. For that reason, most of the research in the field only considers the information which is the topic of the next section.

2.2 About Information

In this section, we try to model the characteristics of mis-/dis-information that people may encounter online and how such a conceptual model can be used by AI systems. In this regard, we follow the conceptualization proposed in [80] which is used to facilitate the distinction among different types of information.

In [80] the authors use the term fake news as an umbrella term to start their analysis. This choice is motivated by the observation that over the years the term "fake news" has been used to refer to different types of content online regardless of whether it is intentional or not. This last distinction is important since the concept of fake news is very often tied to the idea of deceitful intent [5]. An example of that might be the results in [14] which show as reliable news outlets such as The New York Times, The Washington Post, and Associated Press were involved in disseminating false information. The authors of [80] thus propose a taxonomy of online content designed to identify signature features of fabricated news. With this taxonomy, they try to cover the nuances behind the definition of misinformation and also to extend its coverage to contents that are not intended for informational purposes, such as satirical expressions, commentary, or citizen journalism. The taxonomy is made of eight categories for the domain of fake news: real news, false news, polarized content, satire, misreporting, commentary, persuasive information, and citizen journalism. Each of these categories is characterized by unique features describing linguistic properties, sources, intentions, structural components, and network characteristics. Among these categories, we here focus on the difference between real and fake news and refer the readers to [80] for further details.

In general, recognizing fake news is a difficult task since it requires a consistent mental effort from readers who should use common-sense and background knowledge to assess the veracity [66]. However, although false news tries to imitate real information in its form, they often lack the news media's editorial style and references of reliable sources. So we could be tempt to use topic-specific characteristics, impartiality, and objectivity as indicators to understand the message's nature. For example, objectivity could be verified with tools for fact-checking and quote verification, described later, whereas impartiality might be verified by an analysis of sources and attributions [123]. Stylistic indicators, instead, are subtler to define since they are made by particular lexical and syntactical structures [7]. The real news should be written with a peculiar journalistic style [33] and moreover, it should lack any storytelling characteristics [123].

For example, the typical false news headlines have to catch the readers' attention straightforwardly and in a specific way, thus they are very often characterized by complete claims which makes them longer than real news ones [50]. This kind of engagement is similar to the technique called *click-bait* in which the user is tempted to follow/click on the link associated with the headline to read more about a specific event. Of course, the primary goal of this technique is not to spread misinformation but to advertise revenues. However news of that kind has also shown a low level of veracity [126].

Other than the mentioned features also moral-emotional words can be a suitable indicator since their presence could indicate low content veracity. As shown in [20] messages with moral-emotional language spread much faster.

Finally, besides the employed features one last distinction could be made on the amount of text considered in the analysis. The analysis with the least amount of information, that is claim-level methods [23,47,96], through medium size or article-level methods [51,98], to large amount of text that characterizes source-level methods [53,130].

The focus of the above-mentioned studies, regardless of the amount of the used information, is to build automated tools aimed to detect fake. We will discuss the fact-checking problem in the next section and later what are recent works on this topic.

3 Challenges

3.1 Fact-Checking

Without taking into account emotional and ideological aspects, we can say that assessing whether the news is true is a cognitively laborious process. In this process an individual, before accepting new evidence as facts, try to verify its reliability, truthfulness, and independence [17]. This becomes even more complicated in a highly dynamic environment in which new information is produced at an unprecedented rate under the need of engaging always larger audiences [77]. This has led to the launch of numerous fact-checking organizations, such as FactCheck[2], PolitiFact[3] and NewsGuard[4] and many others.

[2] https://www.factcheck.org/.
[3] https://www.politifact.com/.
[4] https://www.newsguardtech.com/.

The majority of these examples are based on laborious manual fact-checking which consist of a series of procedure, for example, identifying the claim, gathering evidence, check source credibility, which represents the cognitive effort required by the reader to assess the truthfulness of the news. However, manual validation only covers a small portion of the daily-produced new information. For this reason automatic fact-checking has been attracting attention in the context of computational journalism before [32,38] and within artificial intelligence community later [45,165]. In the AI field, especially, thanks to the advent of deep learning techniques the research on automated fact-checking has made important progress [40,158]. New insights in the fields of natural language processing (NLP) and information retrieval (IR) have allowed us to process large-scale textual information with increasing accuracy to assess the truthfulness of a claim. For example, in [141] authors design a pipeline to identify claims (to be checked), find appropriate evidence, and produce judgments. From there many datasets, systems, and simpler models for fact-checking were presented RumourEval [27], CLEF CheckThat [13], and ClaimBuster [47]. Those approaches share common components to verify web documents such as document retrieval, claim spotters, and claim validity checker. Other systems, such as FEVER2 [138] and SCIVER [145], are only designed to tackle claim validation under the assumption that claims are provided and worthy to be checked.

In general, once it is provided new information, automated fact-checking can be thought of as a four stages process, or sub-tasks:

1. **Claim detection and matching**: typically identified as the first step, this sub-task aims to identify claims that require verification [46] which is similar to the practices of journalistic fact-checking [18]. It also involves questions related to assessing the check-worthy of a claim [86] and how this worthiness varies over time [12]. Recently, [61] propose a model called Claim/not Claim, built on top of InferSent embeddings [24], with which pose attention to the question of whether or not a claim can be verifiable with the readily available evidence. Correlated with the claim detection there is the claim matching problem which is often framed as a ranking task and involves the retrieval of already checked facts w.r.t. the similarity with the fact to check [119] from some sort of database [93].
2. **Evidence retrieval**: its scope is to find sources supporting or refuting the claim. First attempts to solve the fact-checking task were based only on claims and pattern-recognition approaches without taking in account external knowledge [103,143,149]. Without supporting evidence, such attempts struggled to evaluate well-presented misinformation [116]. Nowadays, if we consider the quality of automatic text-generation tools, it is very difficult to distinguish between real news and fake news by only focusing on the style [157]. On the other side, the choice made by those works were dictated by the fundamental issue which is that not always possible to get access to trustful information. The methods, that try to include external knowledge sources, very often to assess the veracity of a claim postulate the access to trusted sources, such as encyclopedias, other media, or external knowledge bases [11,122,131,135].

This assumption were needed since, in general, assessing the trustfulness of a source and later verify a claim is a demanding task [68].

3. **Claim verification**: in this step based on the retrieved/available evidences researchers formulate the task as a classification problem. The outputs for this classification task ranges from a simple binary classification [84,96] to multi-class classification in which labels represent degrees of truthfulness [3,11,120]. By taking in account the well-known limitations, the multi-class setting is in general to prefer since the challenges of supporting strong position are very often hard to handle.
4. **Justification production**: this task concerns the production of human-interpretable explanations, or at least a set of evidence, supporting the classification decision. As discussed in [139] it is important, from a journalistic point of view, to convince readers of what the claim is saying. In the simplest case, we can start by presenting the evidence returned by a retrieval system. For example, in [70] authors build a justification employing an attention signal to highlight the salient parts of the retrieved information. However, more recent works have focused on the generation of textual justifications, as documented in [62], in which the system produces a summary as a proxy to explain its decision process [9]. However, although the created summary provides useful insights about how the model works, it misses to clarify the exact inference procedure; a possible solution to this issue might be relying on symbolic systems in which the justification is automatically produced as a result of the logical-inference process [1,34].

The description made so far allows us only to introduce a few concepts along with interesting works in the field. The methods present in the literature are much more and several works try to provide an exhaustive overview of the subject, such as [85,133], while [126,165] have more focus on social media.

3.2 Degrees of Truthfulness, Falsehood, and Subjectivity

Even with enough amount of information it could be not so easy to assess the truthfulness or the falsehood of a claim. In general, stories may be technically accurate but still misleading. In [8], for example, authors build a system for detecting cherry-picking to measure the amount of support a story has since it is not so rare to present well-chosen evidence to support misleading news. Since not all its information might be equally trustworthy, it is better to avoid considering a claim as a whole. Works that divide the veracity check among different sources [155] and that assess the agreement among those [161] are less prone to misclassify a claim although they still require improvements. Furthermore, new methods should however face a challenging problem which is subjective in the judgment process.

The degree of truthfulness or falsehood eventually has to do with a subjective interpretation of the reality. This interpretation is conditioned by the audience's social/cultural and religious system and education background. This last point allows us to introduce the next challenge which discusses the complexity of the annotation process while creating coherent datasets.

3.3 Datasets Building

State-of-the-art systems for the claim-related task and misinformation detection heavily rely on training large language models. Those models, although pre-trained on large-scale textual corpora, still require large and high-quality labeled datasets to be fine-tuned to the fake news task. Despite the recent research efforts, the available datasets are often synthetic, highly imbalanced in favor of fake news samples, and biased. For example, using crowd-sourcing based techniques datasets, as discussed for the more general task of reading comprehension in [49,153], easily conduct to biased models as documented for the related task of natural language inference NLI[5] in [43,76].

In the context of fact-checking, [117] highlighted the effect of claim-representative keywords on the predictions of models trained upon the dataset FEVER [136]. Adversarial training was proposed in the context of the FEVER 2 shared task [138] as an attempt to solve this issue. Other solutions to mitigate biases are based on making models less susceptible to catastrophic forgetting [73,134]. Finally, authors in [114] try to make models more sensitive to subtle differences in supporting evidence by building better contrastive samples.

The imbalance of datasets is another major source of issues since models trained on such datasets with a high chance tend to overfit. For example, [154] tries to alleviate this issue with a resampling procedure that involve only the samples of the minority class.

In the following of this section, we try to report a non-exhaustive list of the most commonly used datasets in the field of misinformation and disinformation. However, since each dataset has unique features and differences in the annotation process synthesizing all the datasets' nuances in a few lines would be misleading. We prefer to report the summary in the form of a simple table and provide the reference to the original paper to further details.

Claim-Related Dataset. For the claim-oriented datasets, we split the summary into two tables. In Table 1 on the top, we report datasets that were built to predict check-worthy claims in which the typical input is social media post with textual content. While in Table 1 on the bottom the datasets for claim validation.

Multimodal Dataset. In Table 2 we report a short list of most of the existing multi-modal datasets. Those datasets have recently become quite popular since the evolution of social media platforms which enhanced their text-based forums with multi-modal environments. This happened since visual modalities such as images and videos are more favorable and attractive to the users. As consequence misinformation producers have heavily relied on contextual correlations between modalities such as text and image. In Table 2, WS_O_TRN_TP stands for the ensemble of content providers: Wall Street, Onion, TheRealNews, and ThePoke.

[5] NLI is the task of determining whether a text h, the hypothesis, can (logically) be inferred from a given text p, called premise [19].

Table 1. In the top table, we report the claim detection datasets, where we split the datasets into two categories: Worthy Assessment and Checkable. Below is the table of claim validation datasets which are expressed in terms of factual verification.

Dataset for Worthy Assessment	Input Size	Num. Classes	Sources
CredBank [79]	1k	5	Twitter
Weibo [72]	5k	2	Twitter/Weibo
Suspicious [144]	131k	2/5	Twitter
CheckThat20-T1 [13]	8k	Ranking	Twitter
CheckThat21-T1A [86]	17k	2	Twitter
Debate [46]	1k	3	Transcript
ClaimRank [36]	5k	Ranking	Transcript
Dataset for Checkable	**Input Size**	**Num. Classes**	**Sources**
CitationReason [105]	4k	13	Wikipedia
PolitiTV [61]	6k	7	Transcript
SemEval19-TA[78]	2k	3	Forum

Dataset for Factual Verification	Input Size	Evidence	Num. Classes	Source
StatsProperties [142]	7k	KG[a]	Numeric	Internet
CreditAssess [97]	5k	Text	2	Fact Check/Wiki
PunditFact [104]	4k	-	2/6	Fact Check
Liar [150]	12k	Meta	6	Fact Check
Liar-Plus [4]	12k	Text/Meta	6	Fact Check
FEVER [136]	185k	Text	3	Wiki
NELA [52]	136k	-	2	News
BuzzfeedNews [99]	1k	Meta	4	Facebook
BuzzFace [111]	2k	Meta	4	Facebook
FakeNewsNet [125]	23,196	Meta	2	Fact Check
Snopes [44]	6k	Text	3	Fact Check
MultiFC [10]	36k	Text/Meta	2-27	Fact Check
Climate-FEVER [28]	1k	Text	4	Climate
SciFact [146]	1k	Text	3	Science
PUBHEALTH [62]	11k	Text	4	Fact Check
COVID-Fact [109]	4k	Text	2	Forum
TabFact [22]	92k	Table	2	Wiki
InfoTabs [42]	23k	Table	3	Wiki
HOVER [56]	26k	Text	2	Wiki
WikiFactCheck [112]	124k	Text	2	Wiki
FakeCovid [120]	5k	-	2	Fact Check
X-Fact [41]	31k	Text	7	Fact Check
AnswerFact [160]	60k	Text	5	Amazon
VitaminC [115]	488k	Text	3 Classes	Wiki
Sem-Tab-Fact [148]	5k	Table	3	Wiki
FEVEROUS [6]	87k	Text/Table	3	Wiki

[a] Stands for Knowledge Graph

Table 2. In this table we report the fake news datasets characterized by multi-modal input.

Dataset for Factual Verification	Input Size	Num. Classes	Modalities	Source
image-verification-corpus [16]	17k	2	image,text	Twitter
Fakeddit [83]	1M	2,3,6	image,text	Reddit
NewsBag [57]	215k	2	image, text	WS_O_TRN_TP
NewsBag++ [57]	589k	2	image,text	WS_O_TRN_TP
MM-COVID [69]	11,173	2	image,text,social context	Twitter
ReCOVery [164]	2,029	2	text,image	Twitter
CoAID [25]	5,216	2	image,text	Twitter
MMCoVaR [21]	2k articles+24k tweets	2	image,text,social context	Twitter
N24News [151]	60k	24	image,text	New York Times
MuMiN [90]	10k	3	image,text	Twitter

Although over recent years there has been an increasing interest in such kinds of multi-modal datasets there are still data-related challenges. The first, and perhaps most important, is the lack of comprehensive datasets since many datasets are small in size and often imbalanced in favor of fake examples. Other current flaws are the mono-lingual nature of most of them and the limited heterogeneity of their content (w.r.t. images and text of the articles). This last point becomes more apparent when we consider that many datasets are built to only cover a specific event, such as COVID-19 or elections. In this regard, only the recent Mumin Dataset [90] tries to address some of the issues just mentioned.

The Large-Scale Multilingual Multi-modal Fact-Checked Misinformation Social Network Dataset (MuMin) is quite large since it comprises 26 thousand Twitter threads (roughly 20M tweets). These threads have been aligned to 13 thousand fact-checked claims which, besides the labels, provide further information about the context than that contained in the tweets. Finally, the authors have chosen a conservative approach for the annotation strategy: if the claim is *mostly true* then it is labeled as factual, whereas when it is *half true* or *half false* it is labeled as misinformation. In this way, they collapse the claims' multi-class labeling into a binary choice under the assumption that the presence of a significant part of false information within a claim should expose the readers to misleading content.

4 Current Research Trends

Recent trends in the field of misinformation and disinformation detection largely rely on deep learning techniques. The common strategies can be divided into two major categories. The first is to use a pipeline whose components could be pre-trained large models or not. The pipeline's components are usually trained independently and evaluate each input separately. The second option is a joint distribution-based approach in which the output distribution is a function of multiple components. In the following, we discuss some solutions for the claim-related task and the misinformation detection with multi-modal inputs.

4.1 Claim-Related Tasks Solutions

Claim detection is an essential part of automated fact-checking systems as all other components need to rely on the output of this stage. Its goal is to select claims that need to be checked later in the pipeline. The task of claim detection, like many other tasks, has however the intrinsic issue related to the volume of data produced on a daily base. In this scenario, researchers have been trying not to use external evidence and frame the problem as a classification task.A binary decision is made on whether each input sentence constitutes a claim or not. Typically, a set of sentences is given as input.

The early systems were characterized by hand-crafted platform-dependent features such as Reddit karma and up-votes [2] or Twitter metadata [29]. Others approach relied on linguistic features or entities recognized in the text [167], and syntactic ones [163]. More recently, deep learning-based methods have taken hand-crafted features over. Recurrent and Graph neural networks have over time proved their value in this context. Especially the possibility of introducing user's activity context information [166] has allowed them to build more accurate models [39]. Graph Neural Networks has also provided a solid framework to model propagation behavior of (potentially harmful) claims [81,156].

Collecting evidence supporting or undermining a claim is a task that was typically carried out using consolidated indexing technologies, such as Lucene[6], and entity linking based on some knowledge bases [121]. For example in [137] authors use a pipeline, made of an evidence retrieval module and a verification module, in which a combination of TF-IDF for document retrieval and string matching using named entities and capitalized expressions was used. Advance in the field of embedding representations for textual input has later opened up the possibility of employing vectors as the element on which to compute similarity [58] and indexing [67]. Also, better methods for text generation have allowed to [30]'s authors to use an approach based on question-generated answering to provide information, in the form of natural language briefs about the claim before performing the check. In [65] authors propose to use language models as fact-checkers, but later works have shown as this approach might be prone to propagate the biases of the language models into the new task [64].

Something missing in all the above-mentioned methods it the lack of reasoning over multiple pieces of evidence. Of course, introducing a reasoning component into a differentiable system is not an easy task. The first attempt, for example, was based on the simple concatenation of different piece of evidence [71,89]. But more recent ones try to aggregate information from different evidence in a more elaborated way. [113] uses a joint reranking-and-verification model to fuses evidence documents, [162] uses semantic role labeling and graph structure to re-define the relative distances of words that, along with graph convolutional network and graph attention network, propagate and aggregate information from neighboring nodes on the graph.

Approaches for justification production could be based on attention to highlighting the span within the evidence [70,124]. However, later works [55,100,118]

[6] https://lucene.apache.org/.

have shown as removing high-score tokens may sometimes leave unaltered the final justification while low-score ones could heavily affect the results. In the opposite direction the research in [1,34] rely on logical languages to provide more robust methods. Those methods are essentially rule-based approaches with the constraint of representation power of the formalism. They employ a triplet-based format for the knowledge to guarantee scalability but, at the same time, limit the kind of information that can be stored in the knowledge base. Finally, following a recent trend, authors in [62] use a generative method, based on an abstractive approach, to provide a textual justification. However, as shown in [75], there is a chance that such an approach could generate misleading explanations due to hallucination phenomena.

4.2 Multi-Modal Misinformation Detection

Combinations of features e.g., text and image have been recently used to enhance the performance of misinformation detection systems. Different fusion mechanisms can be implemented, but most of them can be classified into early and late fusion. In early fusion, all the different kinds of features are fed into one model in their original form. The result will be later passed to the classifier as shown in [35]. Later fusion, on the other hand, performs the fusion on the extracted features provided by different components. Often features, such as text, images, and social networks are concatenated into a single vector that feeds the classifier [102,106,127]. However, it seems that simple concatenation is not very effective to build meaningful representations. In the attempt to generate better representation attention mechanism was used.

Different variants of attention have been proposed. For example, the Hierarchical Multi-modal Contextual Attention Networks [101] uses a hierarchical structural bias for the attention modules to extract more meaningful information. [110] propose a shared cross attention transformer encoder which, thanks to the shared layers, tries to learn correlations among modalities. Another cross-modal attention Residual system is presented in [128] aims to selectively extract the relevant information for a target modality from other modalities while preserving its distinctive features. Other examples of attention mechanism for misinformation detection are [63,70,147]. Besides the attention mechanism, the other most common types of neural architecture used for fake news detection are Graph Neural Networks (GNNs).

GNNs have gained huge success in recent years. [129] introduces a temporal propagation-based fake news detection framework in which structure, content semantics, and temporal information are used to recognize temporal evolution patterns of real-world news. By incorporating information from the medical knowledge graph DETERRENT [26] uses a GNN and an attention mechanism to build knowledge-guided article embeddings which are used for misinformation detection. Finally, [159] builds a deep diffusive network model to learn the representations of news articles, creators, and subjects simultaneously. These representations should incorporate the network structure information thanks to the connections among news articles, creators, and news subjects.

The last work we discuss is [15], which uses a continual learning approach for engagement prediction of a user in spreading misinformation. The authors propose an ego-graphs replay strategy in continual learning which is a different perspective compared to the work mentioned before. Ego-graphs are simple graphs composed of a single central node (an user) and its neighbors. Based on this kind of representation and using graph neural networks authors can predict whether users will engage in misinformation and conspiracy theories spreading. Also, the catastrophic forgetting issue related to the dynamic nature of online social networks is addressed with a continual learning approach.

5 Conclusion

In this study, we tried to give an updated not-exhaustive review of the state of the mis- and disinformation research field. We framed the problem as a socio-technological one and provided references to important works in the fields of psychology, journalism, and cognitive science. We paid particular attention to these aspects because any proposed solutions should take into account the way we, as humans, process information and how that information can be affected by deceptive intentions of other individuals.

We strongly believe that future high-quality datasets will continue to help progress the field if they succeed to have less biased content. This can be achieved with a multi-disciplinary approach and, of course, with some technological assistance. AI tools from natural language processing (NLP) and machine learning (ML) are advancing very quickly and can help, but the adoption of any tool should be carefully evaluated. Also corporate, such as Twitter, Facebook, YouTube, and Instagram, plays a critical role in this context since they are very often the medium through which potentially-dangerous information is spread. More regulated principles should guide those platforms.

The last point, which opens up a different discussion, regards how the challenges of this automation process concerning governance, accountability, and censorship would eventually impact our right to free speech.

References

1. Ahmadi, N., Lee, J., Papotti, P., Saeed, M.: Explainable fact checking with probabilistic answer set programming. arXiv preprint arXiv:1906.09198 (2019)
2. Aker, A., Derczynski, L., Bontcheva, K.: Simple open stance classification for rumour analysis. arXiv preprint arXiv:1708.05286 (2017)
3. Alhindi, T., Petridis, S., Muresan, S.: Where is your evidence: improving fact-checking by justification modeling. In: Proceedings of the first workshop on fact extraction and verification (FEVER), pp. 85–90 (2018)
4. Alhindi, T., Petridis, S., Muresan, S.: Where is your evidence: improving fact-checking by justification modeling. In: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), pp. 85–90. Association for Computational Linguistics, Brussels, Belgium (2018). https://doi.org/10.18653/v1/W18-5513, https://www.aclweb.org/anthology/W18-5513

5. Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. J. Econ. Perspect. **31**(2), 211–36 (2017)
6. Aly, R., et al.: FEVEROUS: fact extraction and verification over unstructured and structured information. In: 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks (2021)
7. Argamon-Engelson, S., Koppel, M., Avneri, G.: Style-based text categorization: what newspaper am i reading. In: Proceedings of the AAAI Workshop on Text Categorization, pp. 1–4 (1998)
8. Asudeh, A., Jagadish, H.V., Wu, Y., Yu, C.: On detecting cherry-picked trendlines. Proc. VLDB Endow. **13**(6), 939–952 (2020)
9. Atanasova, P., Simonsen, J.G., Lioma, C., Augenstein, I.: Generating fact checking explanations. arXiv preprint arXiv:2004.05773 (2020)
10. Augenstein, I., et al.: MultiFC: a real-world multi-domain dataset for evidence-based fact checking of claims. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 4685–4697. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1475, https://www.aclweb.org/anthology/D19-1475
11. Augenstein, I., et al.: MultiFC: a real-world multi-domain dataset for evidence-based fact checking of claims. arXiv preprint arXiv:1909.03242 (2019)
12. Barnoy, A., Reich, Z.: The when, why, how and so-what of verifications. Journal. Stud. **20**(16), 2312–2330 (2019)
13. Barrón-Cedeño, A., et al.: Overview of CheckThat! 2020: automatic identification and verification of claims in social media. In: Arampatzis, A., et al. (eds.) CLEF 2020. LNCS, vol. 12260, pp. 215–236. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58219-7_17
14. Benkler, Y., Faris, R., Roberts, H.: Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics. Oxford University Press, Oxford (2018)
15. Bo, H., McConville, R., Hong, J., Liu, W.: Ego-graph replay based continual learning for misinformation engagement prediction. arXiv preprint arXiv:2207.12105 (2022)
16. Boididou, C., Papadopoulos, S., Zampoglou, M., Apostolidis, L., Papadopoulou, O., Kompatsiaris, Y.: Detection and visualization of misleading content on twitter. Int. J. Multimed. Inf. Retr. **7**(1), 71–86 (2018)
17. Borden, S.L., Tew, C.: The role of journalist and the performance of journalism: ethical lessons from "fake" news (seriously). J. Mass Media Ethics **22**(4), 300–314 (2007)
18. Borel, B.: The Chicago Guide to Fact-Checking. University of Chicago Press, Chicago (2016)
19. Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326 (2015)
20. Brady, W.J., Wills, J.A., Jost, J.T., Tucker, J.A., Van Bavel, J.J.: Emotion shapes the diffusion of moralized content in social networks. Proc. Natl. Acad. Sci. **114**(28), 7313–7318 (2017)
21. Chen, M., Chu, X., Subbalakshmi, K.: MMCoVaR: multimodal COVID-19 vaccine focused data repository for fake news detection and a baseline architecture for classification. In: Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 31–38 (2021)

22. Chen, W., et al.: TabFact: a large-scale dataset for table-based fact verification. In: 8th International Conference on Learning Representations, ICLR 2020. Addis Ababa, Ethiopia (2020). https://openreview.net/forum?id=rkeJRhNYDH
23. Ciampaglia, G.L., Shiralkar, P., Rocha, L.M., Bollen, J., Menczer, F., Flammini, A.: Computational fact checking from knowledge networks. PLoS ONE **10**(6), e0128193 (2015)
24. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364 (2017)
25. Cui, L., Lee, D.: CoAID: COVID-19 healthcare misinformation dataset. arXiv preprint arXiv:2006.00885 (2020)
26. Cui, L., Seo, H., Tabar, M., Ma, F., Wang, S., Lee, D.: DETERRENT: knowledge guided graph attention network for detecting healthcare misinformation. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 492–502 (2020)
27. Derczynski, L., et al.: SemEVAL-2017 task 8: RumourEVAL: determining rumour veracity and support for rumours. arXiv preprint arXiv:1704.05972 (2017)
28. Diggelmann, T., Boyd-Graber, J.L., Bulian, J., Ciaramita, M., Leippold, M.: CLIMATE-FEVER: a dataset for verification of real-world climate claims. CoRR **abs/2012.00614** (2020). https://arxiv.org/abs/2012.00614
29. Enayet, O., El-Beltagy, S.R.: NileTMRG at SemEVAL-2017 task 8: determining rumour and veracity support for rumours on Twitter. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 470–474 (2017)
30. Fan, A., et al.: Generating fact checking briefs. arXiv preprint arXiv:2011.05448 (2020)
31. Fazio, L.: Pausing to consider why a headline is true or false can help reduce the sharing of false news. Harvard Kennedy School Misinformation Review 1(2) (2020)
32. Flew, T., Spurgeon, C., Daniel, A., Swift, A.: The promise of computational journalism. Journal. Pract. **6**(2), 157–171 (2012)
33. Frank, R.: Caveat lector: fake news as folklore. J. Am. Folk. **128**(509), 315–332 (2015)
34. Gad-Elrab, M.H., Stepanova, D., Urbani, J., Weikum, G.: ExFaKT: a framework for explaining facts over knowledge graphs and text. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, pp. 87–95 (2019)
35. Gallo, I., Ria, G., Landro, N., La Grassa, R.: Image and text fusion for UPMC food-101 using BERT and CNNs. In: 2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ), pp. 1–6. IEEE (2020)
36. Gencheva, P., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Koychev, I.: A context-aware approach for detecting worth-checking claims in political debates. In: 2017 Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP, pp. 267–276 (2017)
37. George, J.F., Gupta, M., Giordano, G., Mills, A.M., Tennant, V.M., Lewis, C.C.: The effects of communication media and culture on deception detection accuracy. MIS Q. **42**(2), 551–575 (2018)
38. Graves, D.: Understanding the promise and limits of automated fact-checking (2018)

39. Guo, H., Cao, J., Zhang, Y., Guo, J., Li, J.: Rumor detection with hierarchical social attention network. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pp. 943–951 (2018)
40. Guo, Z., Schlichtkrull, M., Vlachos, A.: A survey on automated fact-checking. Trans. Assoc. Comput. Linguist. **10**, 178–206 (2022)
41. Gupta, A., Srikumar, V.: X-fact: A new benchmark dataset for multilingual fact checking. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 675–682 (2021)
42. Gupta, V., Mehta, M., Nokhiz, P., Srikumar, V.: INFOTABS: inference on tables as semi-structured data. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2309–2324. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.210, https://www.aclweb.org/anthology/2020.acl-main.210
43. Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S.R., Smith, N.A.: Annotation artifacts in natural language inference data. arXiv preprint arXiv:1803.02324 (2018)
44. Hanselowski, A., Stab, C., Schulz, C., Li, Z., Gurevych, I.: A richly annotated corpus for different tasks in automated fact-checking. In: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), pp. 493–503. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/K19-1046, https://www.aclweb.org/anthology/K19-1046
45. Hassan, N., et al.: The quest to automate fact-checking. In: Proceedings of the 2015 Computation+ Journalism Symposium (2015)
46. Hassan, N., Li, C., Tremayne, M.: Detecting check-worthy factual claims in presidential debates. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp. 1835–1838 (2015)
47. Hassan, N., et al.: ClaimBuster: the first-ever end-to-end fact-checking system. Proc. VLDB Endow. **10**(12), 1945–1948 (2017)
48. He, S., Hollenbeck, B., Proserpio, D.: The market for fake reviews. Mark. Sci. **41**, 896–921 (2022)
49. Hermann, K.M., et al.: Teaching machines to read and comprehend. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
50. Horne, B.D., Adali, S., Sikdar, S.: Identifying the social signals that drive online discussions: a case study of reddit communities. In: 2017 26th International Conference on Computer Communication and Networks (ICCCN), pp. 1–9. IEEE (2017)
51. Horne, B.D., Dron, W., Khedr, S., Adali, S.: Assessing the news landscape: a multi-module toolkit for evaluating the credibility of news. In: 2018 Companion Proceedings of the The Web Conference, pp. 235–238 (2018)
52. Horne, B.D., Khedr, S., Adali, S.: Sampling the news producers: a large news and feature data set for the study of the complex media landscape. In: Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, 25-28 June 2018, pp. 518–527. AAAI Press (2018). https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17796
53. Horne, B.D., Nevo, D., O'Donovan, J., Cho, J.H., Adalı, S.: Rating reliability and bias in news articles: does AI assistance help everyone?. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 13, pp. 247–256 (2019)
54. Jack, C.: Lexicon of lies: terms for problematic information. Data Soc. **3**(22), 1094–1096 (2017)

55. Jain, S., Wallace, B.C.: Attention is not explanation. arXiv preprint arXiv:1902.10186 (2019)
56. Jiang, Y., Bordia, S., Zhong, Z., Dognin, C., Singh, M., Bansal, M.: HoVer: a dataset for many-hop fact extraction and claim verification. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 3441–3460. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.309, https://www.aclweb.org/anthology/2020.findings-emnlp.309
57. Jindal, S., Sood, R., Singh, R., Vatsa, M., Chakraborty, T.: NewsBag: a multimodal benchmark dataset for fake news detection. In: CEUR Workshop Proceedings, vol. 2560, pp. 138–145 (2020)
58. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Trans. Big Data **7**(3), 535–547 (2019)
59. Kahan, D.M.: Ideology, motivated reasoning, and cognitive reflection: an experimental study. Judgm. Decis. Mak. **8**, 407–24 (2012)
60. Kim, J., Tabibian, B., Oh, A., Schölkopf, B., Gomez-Rodriguez, M.: Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, pp. 324–332 (2018)
61. Konstantinovskiy, L., Price, O., Babakar, M., Zubiaga, A.: Toward automated factchecking: developing an annotation schema and benchmark for consistent automated claim detection. Digit. threats: Res. Pract. **2**(2), 1–16 (2021)
62. Kotonya, N., Toni, F.: Explainable automated fact-checking: a survey. arXiv preprint arXiv:2011.03870 (2020)
63. Kumari, R., Ekbal, A.: AMFB: attention based multimodal factorized bilinear pooling for multimodal fake news detection. Expert Syst. Appl. **184**, 115412 (2021)
64. Lee, N., Bang, Y., Madotto, A., Khabsa, M., Fung, P.: Towards few-shot fact-checking via perplexity. arXiv preprint arXiv:2103.09535 (2021)
65. Lee, N., Li, B.Z., Wang, S., Yih, W.t., Ma, H., Khabsa, M.: Language models as fact checkers? arXiv preprint arXiv:2006.04102 (2020)
66. Lewandowsky, S., Ecker, U.K., Cook, J.: Beyond misinformation: understanding and coping with the "post-truth" era. J. Appl. Res. Mem. Cogn. **6**(4), 353–369 (2017)
67. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advanced in Neural Information Processing System, vol. 33, pp. 9459–9474 (2020)
68. Li, Y., et al.: A survey on truth discovery. ACM SIGKDD Explor. Newsl. **17**(2), 1–16 (2016)
69. Li, Y., Jiang, B., Shu, K., Liu, H.: MM-COVID: a multilingual and multimodal data repository for combating COVID-19 disinformation. arXiv preprint arXiv:2011.04088 (2020)
70. Lu, Y.J., Li, C.T.: GCAN: graph-aware co-attention networks for explainable fake news detection on social media. arXiv preprint arXiv:2004.11648 (2020)
71. Luken, J., Jiang, N., de Marneffe, M.C.: QED: a fact verification system for the fever shared task. In: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), pp. 156–160 (2018)
72. Ma, J., et al.: Detecting rumors from microblogs with recurrent neural networks. In: Kambhampati, S. (ed.) Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9–15 July 2016, pp. 3818–3824. IJCAI/AAAI Press (2016). http://www.ijcai.org/Abstract/16/537

73. Mahabadi, R.K., Belinkov, Y., Henderson, J.: End-to-end bias mitigation by modelling biases in corpora. arXiv preprint arXiv:1909.06321 (2019)
74. Manikonda, L., Nevo, D., Horne, B.D., Arrington, C., Adali, S.: The reasoning behind fake news assessments: a linguistic analysis. AIS Trans. Human-Comput. Interact. **14**(2), 230–253 (2022)
75. Maynez, J., Narayan, S., Bohnet, B., McDonald, R.: On faithfulness and factuality in abstractive summarization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1906–1919. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.173, https://aclanthology.org/2020.acl-main.173
76. McCoy, R.T., Pavlick, E., Linzen, T.: Right for the wrong reasons: diagnosing syntactic heuristics in natural language inference. arXiv preprint arXiv:1902.01007 (2019)
77. Mihailidis, P., Viotty, S.: Spreadable spectacle in digital culture: civic expression, fake news, and the role of media literacies in "post-fact" society. Am. Behav. Sci. **61**(4), 441–454 (2017)
78. Mihaylova, T., Karadzhov, G., Atanasova, P., Baly, R., Mohtarami, M., Nakov, P.: SemEval-2019 task 8: Fact checking in community question answering forums. In: Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 860–869. Association for Computational Linguistics, Minneapolis, Minnesota, USA (2019). https://doi.org/10.18653/v1/S19-2149, https://www.aclweb.org/anthology/S19-2149
79. Mitra, T., Gilbert, E.: CREDBANK: A large-scale social media corpus with associated credibility annotations. In: Cha, M., Mascolo, C., Sandvig, C. (eds.) Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, 26–29 May 2015, pp. 258–267. AAAI Press (2015). http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10582
80. Molina, M.D., Sundar, S.S., Le, T., Lee, D.: "fake news" is not simply false information: a concept explication and taxonomy of online content. Am. Behav. Sci. **65**(2), 180–212 (2021)
81. Monti, F., Frasca, F., Eynard, D., Mannion, D., Bronstein, M.M.: Fake news detection on social media using geometric deep learning. arXiv preprint arXiv:1902.06673 (2019)
82. Moravec, P., Minas, R., Dennis, A.R.: Fake news on social media: people believe what they want to believe when it makes no sense at all. Kelley School of Business research paper (18–87) (2018)
83. Nakamura, K., Levy, S., Wang, W.Y.: r/Fakeddit: a new multimodal benchmark dataset for fine-grained fake news detection. arXiv preprint arXiv:1911.03854 (2019)
84. Nakashole, N., Mitchell, T.: Language-aware truth assessment of fact candidates. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1009–1019 (2014)
85. Nakov, P., et al.: Automated fact-checking for assisting human fact-checkers. arXiv preprint arXiv:2103.07769 (2021)
86. Nakov, P., et al.: The CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In: Hiemstra, D., Moens, M.-F., Mothe, J., Perego, R., Potthast, M., Sebastiani, F. (eds.) ECIR 2021. LNCS, vol. 12657, pp. 639–649. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-72240-1_75

87. Borges do Nascimento, I.J., et al.: Infodemics and health misinformation: a systematic review of reviews. Bull. World Health Org. **100**(9), 544–561 (2022)
88. Nickerson, R.S.: Confirmation bias: a ubiquitous phenomenon in many guises. Rev. Gen. Psychol. **2**(2), 175–220 (1998)
89. Nie, Y., Chen, H., Bansal, M.: Combining fact extraction and verification with neural semantic matching networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 6859–6866 (2019)
90. Nielsen, D.S., McConville, R.: MuMiN: a large-scale multilingual multimodal fact-checked misinformation social network dataset. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 3141–3153 (2022)
91. Olan, F., Jayawickrama, U., Arakpogun, E.O., Suklan, J., Liu, S.: Fake news on social media: the impact on society. Inf. Syst. Front., 1–16 (2022). https://doi.org/10.1007/s10796-022-10242-z
92. O'reilly, T.: What is Web 2.0. "O'Reilly Media Inc", Sebastopol (2009)
93. Passaro, L.C., Bondielli, A., Lenci, A., Marcelloni, F.: UNIPI-NLE at CheckThat! 2020: approaching fact checking from a sentence similarity perspective through the lens of transformers. In: CLEF (Working Notes) (2020)
94. Pennycook, G., Rand, D.G.: Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. Cognition **188**(-), 39–50 (2019)
95. Pennycook, G., Rand, D.G.: The psychology of fake news. Trends Cogn. Sci. **25**(5), 388–402 (2021)
96. Popat, K., Mukherjee, S., Strötgen, J., Weikum, G.: Credibility assessment of textual claims on the web. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pp. 2173–2178 (2016)
97. Popat, K., Mukherjee, S., Strötgen, J., Weikum, G.: Credibility assessment of textual claims on the web. In: Mukhopadhyay, S., et al. (eds.) Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, 24–28 Oct 2016, pp. 2173–2178. ACM (2016). https://doi.org/10.1145/2983323.2983661
98. Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., Stein, B.: A stylometric inquiry into hyperpartisan and fake news. arXiv preprint arXiv:1702.05638 (2017)
99. Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., Stein, B.: A stylometric inquiry into hyperpartisan and fake news. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 231–240. Association for Computational Linguistics, Melbourne, Australia (2018). https://doi.org/10.18653/v1/P18-1022, https://www.aclweb.org/anthology/P18-1022
100. Pruthi, D., Gupta, M., Dhingra, B., Neubig, G., Lipton, Z.C.: Learning to deceive with attention-based explanations. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4782–4793 (2020)
101. Qian, S., Wang, J., Hu, J., Fang, Q., Xu, C.: Hierarchical multi-modal contextual attention network for fake news detection. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 153–162 (2021)
102. Raj, C., Meel, P.: ARCNN framework for multimodal infodemic detection. Neural Netw. **146**, 36–68 (2022)
103. Rashkin, H., Choi, E., Jang, J.Y., Volkova, S., Choi, Y.: Truth of varying shades: analyzing language in fake news and political fact-checking. In: Proceedings of

the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2931–2937 (2017)
104. Rashkin, H., Choi, E., Jang, J.Y., Volkova, S., Choi, Y.: Truth of varying shades: analyzing language in fake news and political fact-checking. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2931–2937. Association for Computational Linguistics, Copenhagen, Denmark (2017). https://doi.org/10.18653/v1/D17-1317, https://www.aclweb.org/anthology/D17-1317
105. Redi, M., Fetahu, B., Morgan, J.T., Taraborelli, D.: Citation needed: a taxonomy and algorithmic assessment of Wikipedia's verifiability. In: Liu, L., et al. (eds.) The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, 13–17 May 2019, pp. 1567–1578. ACM (2019). https://doi.org/10.1145/3308558.3313618
106. Rezayi, S., Soleymani, S., Arabnia, H.R., Li, S.: Socially aware multimodal deep neural networks for fake news classification. In: 2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 253–259. IEEE (2021)
107. Rubin, V.L.: Disinformation and misinformation triangle: a conceptual model for "fake news" epidemic, causal factors and interventions. J. Documentation **75**, 1013–1034 (2019)
108. Rubin, V.L.: Misinformation and Disinformation: Detecting Fakes with the Eye and AI. Springer Nature, Berlin (2022)
109. Saakyan, A., Chakrabarty, T., Muresan, S.: COVID-Fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, 1–6 Aug 2021, pp. 2116–2129. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.acl-long.165, https://doi.org/10.18653/v1/2021.acl-long.165
110. Sachan, T., Pinnaparaju, N., Gupta, M., Varma, V.: SCATE: shared cross attention transformer encoders for multimodal fake news detection. In: Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 399–406 (2021)
111. Santia, G.C., Williams, J.R.: BuzzFace: a news veracity dataset with Facebook user commentary and egos. In: Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, 25–28 June 2018, pp. 531–540. AAAI Press (2018). https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17825
112. Sathe, A., Ather, S., Le, T.M., Perry, N., Park, J.: Automated fact-checking of claims from wikipedia. In: Calzolari, N., et al. (eds.) Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, 11–16 May 2020, pp. 6874–6882. European Language Resources Association (2020). https://aclanthology.org/2020.lrec-1.849/
113. Schlichtkrull, M., Karpukhin, V., Oğuz, B., Lewis, M., Yih, W.t., Riedel, S.: Joint verification and reranking for open fact checking over tables. arXiv preprint arXiv:2012.15115 (2020)
114. Schuster, T., Fisch, A., Barzilay, R.: Get your vitamin C! robust fact verification with contrastive evidence. arXiv preprint arXiv:2103.08541 (2021)
115. Schuster, T., Fisch, A., Barzilay, R.: Get your Vitamin C! robust fact verification with contrastive evidence. In: Proceedings of the 2021 Conference of the North

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 624–643. Association for Computational Linguistics (2021). https://www.aclweb.org/anthology/2021.naacl-main.52
116. Schuster, T., Schuster, R., Shah, D.J., Barzilay, R.: The limitations of stylometry for detecting machine-generated fake news. Comput. Linguist. **46**(2), 499–510 (2020)
117. Schuster, T., Shah, D.J., Yeo, Y.J.S., Filizzola, D., Santus, E., Barzilay, R.: Towards debiasing fact verification models. arXiv preprint arXiv:1908.05267 (2019)
118. Serrano, S., Smith, N.A.: Is attention interpretable? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2931–2951 (2019)
119. Shaar, S., Martino, G.D.S., Babulkov, N., Nakov, P.: That is a known lie: detecting previously fact-checked claims. arXiv preprint arXiv:2005.06058 (2020)
120. Shahi, G.K., Nandini, D.: FakeCovid–a multilingual cross-domain fact check news dataset for COVID-19 (2020)
121. Shen, W., Wang, J., Han, J.: Entity linking with a knowledge base: issues, techniques, and solutions. IEEE Trans. Knowl. Data Eng. **27**(2), 443–460 (2014)
122. Shi, B., Weninger, T.: Discriminative predicate path mining for fact checking in knowledge graphs. Knowl.-Based Syst. **104**, 123–133 (2016)
123. Shoemaker, P.J.: News values: reciprocal effects on journalists and journalism. Int. Encycl. Media Effects, 1–9 (2017)
124. Shu, K., Cui, L., Wang, S., Lee, D., Liu, H.: Defend: explainable fake news detection. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 395–405 (2019)
125. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: FakeNewsNet: a data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big Data **8**(3), 171–188 (2020). https://doi.org/10.1089/big.2020.0062
126. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor. Newsl. **19**(1), 22–36 (2017)
127. Singhal, S., Shah, R.R., Chakraborty, T., Kumaraguru, P., Satoh, S.: SpotFake: a multi-modal framework for fake news detection. In: 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), pp. 39–47. IEEE (2019)
128. Song, C., Ning, N., Zhang, Y., Wu, B.: A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. Inf. Process. Manage. **58**(1), 102437 (2021)
129. Song, C., Shu, K., Wu, B.: Temporally evolving graph neural network for fake news detection. Inf. Process. Manage. **58**(6), 102712 (2021)
130. Starbird, K., Arif, A., Wilson, T., Van Koevering, K., Yefimova, K., Scarnecchia, D.: Ecosystem or echo-system? Exploring content sharing across alternative media domains. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 12 (2018)
131. Syed, Z.H., Röder, M., Ngomo, A.-C.N.: Unsupervised discovery of corroborative paths for fact validation. In: Ghidini, C., et al. (eds.) ISWC 2019. LNCS, vol. 11778, pp. 630–646. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30793-6_36
132. Tandoc Jr., E.C., Lim, Z.W., Ling, R.: Defining "fake news" a typology of scholarly definitions. Digit. Journal. **6**(2), 137–153 (2018)

133. Thorne, J., Vlachos, A.: Automated fact checking: task formulations, methods and future directions. arXiv preprint arXiv:1806.07687 (2018)
134. Thorne, J., Vlachos, A.: Elastic weight consolidation for better bias inoculation. arXiv preprint arXiv:2004.14366 (2020)
135. Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: Fever: a large-scale dataset for fact extraction and verification. arXiv preprint arXiv:1803.05355 (2018)
136. Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: FEVER: a large-scale dataset for fact extraction and verification. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 809–819. Association for Computational Linguistics, New Orleans, Louisiana (2018). https://doi.org/10.18653/v1/N18-1074, https://www.aclweb.org/anthology/N18-1074
137. Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., Mittal, A.: The fact extraction and verification (fever) shared task. arXiv preprint arXiv:1811.10971 (2018)
138. Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., Mittal, A.: The fever2. 0 shared task. In: Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER), pp. 1–6 (2019)
139. Uscinski, J.E., Butler, R.W.: The epistemology of fact checking. Crit. Rev. **25**(2), 162–180 (2013)
140. Vicario, M.D., Quattrociocchi, W., Scala, A., Zollo, F.: Polarization and fake news: early warning of potential misinformation targets. ACM Trans. Web (TWEB) **13**(2), 1–22 (2019)
141. Vlachos, A., Riedel, S.: Fact checking: task definition and dataset construction. In: Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, pp. 18–22 (2014)
142. Vlachos, A., Riedel, S.: Identification and verification of simple claims about statistical properties. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2596–2601. Association for Computational Linguistics, Lisbon, Portugal (2015). https://doi.org/10.18653/v1/D15-1312, https://www.aclweb.org/anthology/D15-1312
143. Volkova, S., Shaffer, K., Jang, J.Y., Hodas, N.: Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers), pp. 647–653 (2017)
144. Volkova, S., Shaffer, K., Jang, J.Y., Hodas, N.: Separating facts from fiction: linguistic models to classify suspicious and trusted news posts on Twitter. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 647–653. Association for Computational Linguistics, Vancouver, Canada (2017). https://doi.org/10.18653/v1/P17-2102, https://www.aclweb.org/anthology/P17-2102
145. Wadden, D., et al.: Fact or fiction: verifying scientific claims. arXiv preprint arXiv:2004.14974 (2020)
146. Wadden, D., et al.: Fact or fiction: verifying scientific claims. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 7534–7550. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.emnlp-main.609, https://www.aclweb.org/anthology/2020.emnlp-main.609
147. Wang, J., Mao, H., Li, H.: FMFN: fine-grained multimodal fusion networks for fake news detection. Appl. Sci. **12**(3), 1093 (2022)

148. Wang, N.X.R., Mahajan, D., Danilevsky, M., Rosenthal, S.: SemEval-2021 task 9: fact verification and evidence finding for tabular data in scientific documents (SEM-TAB-FACTS). In: Palmer, A., Schneider, N., Schluter, N., Emerson, G., Herbelot, A., Zhu, X. (eds.) Proceedings of the 15th International Workshop on Semantic Evaluation, SemEval@ACL/IJCNLP 2021, Virtual Event / Bangkok, Thailand, 5–6 Aug. 2021, pp. 317–326. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.semeval-1.39
149. Wang, W.Y.: "Liar, liar pants on fire": a new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648 (2017)
150. Wang, W.Y.: "Liar, Liar Pants on Fire": A new benchmark dataset for fake news detection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 422–426. Association for Computational Linguistics, Vancouver, Canada (2017). https://doi.org/10.18653/v1/P17-2067, https://www.aclweb.org/anthology/P17-2067
151. Wang, Z., Shan, X., Yang, J.: N15news: a new dataset for multimodal news classification. arXiv preprint arXiv:2108.13327 (2021)
152. Wardle, C., Derakhshan, H.: Information disorder: toward an interdisciplinary framework for research and policymaking (2017)
153. Williams, A., Nangia, N., Bowman, S.R.: A broad-coverage challenge corpus for sentence understanding through inference. arXiv preprint arXiv:1704.05426 (2017)
154. Williams, E., Rodrigues, P., Novak, V.: Accenture at CheckThat! 2020: if you say so: post-hoc fact-checking of claims using transformer-based models. arXiv preprint arXiv:2009.02431 (2020)
155. Wu, L., Rao, Y., Yang, X., Wang, W., Nazir, A.: Evidence-aware hierarchical interactive attention networks for explainable claim verification. In: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, pp. 1388–1394 (2021)
156. Yang, X., Lyu, Y., Tian, T., Liu, Y., Liu, Y., Zhang, X.: Rumor detection on social media with graph structured adversarial learning. In: Proceedings of the Twenty-ninth International Conference on International Joint Conferences on Artificial Intelligence, pp. 1417–1423 (2021)
157. Zellers, R., et al.: Defending against neural fake news. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
158. Zeng, X., Abumansour, A.S., Zubiaga, A.: Automated fact-checking: a survey. Lang. Linguist. Compass **15**(10), e12438 (2021)
159. Zhang, J., Dong, B., Philip, S.Y.: FakeDetector: effective fake news detection with deep diffusive neural network. In: 2020 IEEE 36th International Conference on Data Engineering (ICDE), pp. 1826–1829. IEEE (2020)
160. Zhang, W., Deng, Y., Ma, J., Lam, W.: AnswerFact: fact checking in product question answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2407–2417. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.emnlp-main.188, https://www.aclweb.org/anthology/2020.emnlp-main.188
161. Zhang, Y., Ives, Z., Roth, D.: Evidence-based trustworthiness. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 413–423 (2019)
162. Zhong, W., et al.: Reasoning over semantic-level graph for fact checking. arXiv preprint arXiv:1909.03745 (2019)
163. Zhou, X., Jain, A., Phoha, V.V., Zafarani, R.: Fake news early detection: a theory-driven model. Digit. Threats: Res. Pract. **1**(2), 1–25 (2020)

164. Zhou, X., Mulay, A., Ferrara, E., Zafarani, R.: Recovery: a multimodal repository for COVID-19 news credibility research. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 3205–3212 (2020)
165. Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., Procter, R.: Detection and resolution of rumours in social media: a survey. ACM Comput. Surv. (CSUR) **51**(2), 1–36 (2018)
166. Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G., Tolmie, P.: Analysing how people orient to and spread rumours in social media by looking at conversational threads. PloS one **11**(3), e0150989 (2016)
167. Zuo, C., Karakas, A., Banerjee, R.: A hybrid recognition system for check-worthy claims using heuristics and supervised learning. In: CEUR Workshop Proceedings, vol. 2125 (2018)