Information Processing and Management 60 (2023) 103219

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

The state of human-centered NLP technology for fact-checking

Anubrata Das *, Houjiang Liu, Venelin Kovatchev, Matthew Lease

School of Information, The University of Texas at Austin, Austin, TX, USA

ARTICLE INFO

Keywords:
Natural Language Processing
Misinformation
Disinformation
Explainability
Human-AI teaming

ABSTRACT

Misinformation threatens modern society by promoting distrust in science, changing narratives in public health, heightening social polarization, and disrupting democratic elections and financial markets, among a myriad of other societal harms. To address this, a growing cadre of professional fact-checkers and journalists provide high-quality investigations into purported facts. However, these largely manual efforts have struggled to match the enormous scale of the problem. In response, a growing body of Natural Language Processing (NLP) technologies have been proposed for more scalable fact-checking. Despite tremendous growth in such research, however, practical adoption of NLP technologies for fact-checking still remains in its infancy today.

In this work, we review the capabilities and limitations of the current NLP technologies for fact-checking. Our particular focus is to further chart the design space for how these technologies can be harnessed and refined in order to better meet the needs of human fact-checkers. To do so, we review key aspects of NLP-based fact-checking: task formulation, dataset construction, modeling, and human-centered strategies, such as explainable models and human-in-the-loop approaches. Next, we review the efficacy of applying NLP-based fact-checking tools to assist human fact-checkers. We recommend that future research include collaboration with fact-checker stakeholders early on in NLP research, as well as incorporation of human-centered design practices in model development, in order to further guide technology development for human use and practical adoption. Finally, we advocate for more research on benchmark development supporting extrinsic evaluation of human-centered fact-checking technologies.

1. Introduction

Misinformation and related issues (disinformation, deceptive news, clickbait, rumors, and information credibility) increasingly threaten society. While concerns of misinformation existed since the early days of written text (Marcus, 1992), with recent development of social media, the entry barrier for creating and spreading content has never been lower. Moreover, polarization online drives the spread of misinformation that in turn increases polarization (Cinelli et al., 2021a, 2021b; Vicario, Quattrociocchi, Scala, & Zollo, 2019). Braking such a vicious cycle would require addressing the problem of misinformation at its root.

Fields such as journalism (Graves, 2018b; Graves & Amazeen, 2019; Neely-Sardon & Tignor, 2018) and archival studies (LeBeau, 2017) have extensively studied misinformation, and recent years have seen a significant growth in fact-checking initiatives to address this problem. Various organizations now focus on fact-checks (e.g., PolitiFact, Snopes, FactCheck, First Draft, and Full Fact), and

* Corresponding author.
E-mail address: anubrata.das@utexas.edu (A. Das).

https://doi.org/10.1016/j.ipm.2022.103219
Received 15 June 2022; Received in revised form 29 November 2022; Accepted 30 November 2022
Available online 21 December 2022

organizations such as the International Fact-Checking Network (IFCN)[1] train and provide resources for independent fact-checkers and journalists to further scale expert fact-checking.

While professional fact-checkers and journalists provide high-quality investigations of purported facts to inform the public, human effort struggles to match the global Internet scale of the problem. To address this, a growing body of research has investigated Natural Language Processing (NLP) to fully or partially automate fact-checking (Graves, 2018a; Guo, Schlichtkrull, & Vlachos, 2022; Nakov, Corney et al., 2021; Zeng, Abumansour, & Zubiaga, 2021; Zhou & Zafarani, 2020). However, even state-of-the-art NLP technologies still cannot match human capabilities in many areas and remain insufficient to automate fact-checking in practice. Experts argue (Arnold, 2020; Nakov, Corney et al., 2021) that fact-checking is a complex process and requires subjective judgment and expertise. While current NLP systems are increasingly better at addressing simple fact-checking tasks, identifying false claims that are contextual and beyond simple declarative statements remains beyond the reach for fully automated systems (Chen, Sriram, Choi, & Durrett, 2022; Fan et al., 2020). For example, claims buried in conversational systems, comment threads in social media community, and claims in multimedia contents are particularly challenging for automated systems. Additionally, most fact-checking practitioners desire NLP tools that are integrated into the existing fact-checking workflow and reduce latency (Alam, Shaar et al., 2021; Graves, 2018b; Nakov, Corney et al., 2021).

In this literature review, we provide the reader with a comprehensive and holistic overview of the current state-of-the-art challenges and opportunities to more effectively leverage NLP technology in fact-checking. Our objectives in this work are twofold. First, we cover all aspects of the NLP pipeline for fact checking: task formulation, dataset construction, and modeling approaches. Second, we emphasize the human-centered approaches that seek to augment and accelerate human fact-checking, rather than supplant it. In contrast, prior literature reviews (Guo et al., 2022; Oshikawa, Qian, & Wang, 2020; Zeng et al., 2021) either provide an overview of the existing approaches or capture the details of only a specific part of the fact-checking pipeline (Demartini, Mizzaro, & Spina, 2020; Hanselowski et al., 2018; Hardalov, Arora, Nakov, & Augenstein, 2021; Kotonya & Toni, 2020a).

Furthermore, we argue that it is important to extend the review of NLP technologies for fact-checking from modeling development to the area of Human–Computer Interaction (HCI) because technology design should reflect user needs so that its development can be better integrated in the real-world use context (Graves, 2018a; Juneja & Mitra, 2022; Kovatchev et al., 2020; Lease, 2020; Micallef, Armacost, Memon, & Patil, 2022; Nakov, Corney et al., 2021). Specifically, we point the reader toward Section 7 where we propose concrete directions for future work.

Current challenges are largely due to the relatively early stage of development of the automated fact-checking technology. Specifically, current studies tend to adopt an intrinsic evaluation of components of the fact-checking pipeline rather than an end-to-end extrinsic evaluation of the entire fact-checking task. Moreover, component-wise accuracies may remain below the threshold required for practical adoption. Furthermore, while the research community's focus on prediction accuracy has yielded laudable improvements, human factors (e.g., usability, intelligibility, trust) have garnered far less attention or progress yet are crucial for practical adoption.

Such limitations have implications for future research. First, practical use of NLP technologies for fact-checking is likely to come from hybrid, human-in-the-loop approaches rather than full automation. Second, as the technology matures, end-to-end evaluation becomes increasingly important to ensure practical solutions are being developed to solve the real-world use-case. To this end, new benchmarks that facilitate the extrinsic evaluation of automated fact-checking applications in practical settings may help drive progress on solutions that can be adopted for use in the wild. Finally, to craft effective human-in-the-loop systems, more cross-cutting NLP and HCI integration could strengthen design of fact-checking tools, so that they are accurate, scalable, and usable in practice. Toward this end, it may be fruitful to collaborate more with stakeholders early on in NLP research and incorporate human-centered design practices in developing models.

We have written this article with different audiences in mind. For researchers and fact-checkers who are new to automated fact-checking, this article provides a comprehensive overview of the problem. We discuss the challenges, the state-of-the-art capabilities, and the opportunities in the field, and we emphasize how machine learning and natural language processing can be used to combat disinformation. We recommend researchers new to this topic read the article in its entirety, following the logical structure of sections. Other readers who have more experience in the field may already be familiar with some of the concepts that we discuss. For them, *this paper offers a novel human-centered perspective of automated fact checking* and a discussion on how that perspective can affect system design, implementation, and evaluation. To facilitate the use of the paper by more experienced readers, we provide a quick overview of the content covered in each section.

- Section 2 introduces the automated fact-checking pipeline. We provide an overview of the process for human fact-checkers and for automated solutions.
- Section 3 discusses the *task formulation*: the goals and formal definitions of different sub-tasks in fact checking.
- Section 4 describes the process of dataset construction, presents the most popular corpora for automated fact checking, and outlines some limitations of the data.
- Section 5 reviews approaches for automating fact checking. We discuss general NLP capabilities (Section 5.1), explainable approaches (Section 5.2), and human-in-the-loop (Section 5.3) approaches for fact-checking.
- Section 6 surveys existing tools that apply NLP for fact-checking in a practical, real-world context. We argue that the human-centered perspective is necessary for the practical adoption of automated solutions.

[1] https://www.politifact.com/, https://www.snopes.com/, https://www.factcheck.org/, https://firstdraftnews.org/, https://fullfact.org/, and https://www.poynter.org/ifcn/, respectively.

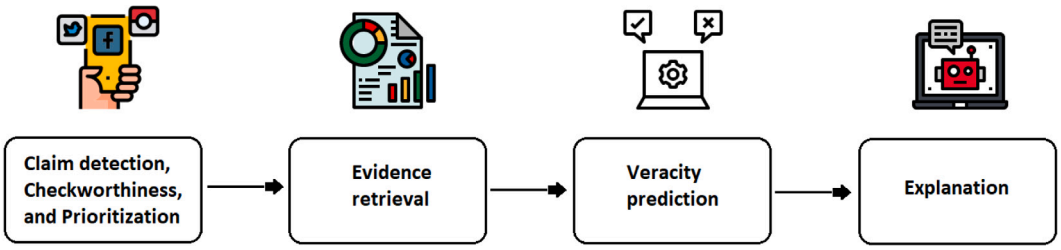

Fig. 1. Fact-checking pipeline.

- Section 7 provides future research directions in the context of human-centered fact checking. We discuss the work division between human and AI for mixed-initiative fact-checking in Section 7.1. In Section 7.2 we propose a novel concept for measuring trust and a novel human-centered evaluation of NLP to assist fact-checkers.
- We conclude our literature review with Section 8.

2. Fact-checking pipeline

The core idea behind automated fact-checking is enabling AI to reason over available information to determine the truthfulness of a claim. For successful automation, it is essential first to understand the complex process of journalistic fact-checking that involves human expertise along with skilled effort toward gathering evidence and synthesizing the evidence. Additional complexity comes from the need to process heterogeneous sources (e.g., information across various digital and non-digital sources). Data is also spread across different modalities such as images, videos, tables, graphs, among others. Moreover, there is a lack of tools that support effective and efficient fact-checking workflows (Arnold, 2020; Graves, 2018a; Nakov, Corney et al., 2021).

Graves (2017) breaks down the practical fact-checking mechanism for human fact-checkers into multiple steps such as (a) identifying the claims to check, (b) tracing false claims, (c) consulting experts, and (d) sharing the resulting fact-check. A growing body of AI literature – specifically in NLP – focuses on automating the fact-checking process. We synthesize several related surveys (Graves, 2018a; Guo et al., 2022; Micallef et al., 2022; Nakov, Corney et al., 2021; Zeng et al., 2021) and distinguish four typical stages that constitute the automated fact-checking technology pipeline (illustrated in Fig. 1). Note that the pipeline we describe below closely follows the structure of Guo et al. (2022), though the broader literature is also incorporated within these four stages:

- **Claim Detection, Checkworthiness, and Prioritization:** Claim detection involves monitoring news and/or online information for potentially false content to fact-check. One must identify claims that are potentially falsifiable (e.g., purported facts rather subjective opinions) (Guo et al., 2022; Zeng et al., 2021). Moreover, because it is impractical to fact-check everything online given limited fact-checking resources (human or automated), fact checkers must prioritize what to fact-check (Arnold, 2020). NLP researchers have sought to inform such prioritization by automatically predicting the "checkworthiness" of claims (Nakov, Corney et al., 2021). Additionally, to avoid repeated work, fact-checkers may consult existing fact-checking databases before judging the veracity of a new claim (*claim matching Zeng et al., 2021*). We see claim matching as a part of prioritizing claims, as fact-checkers would prioritize against checking such claims.
- **Evidence Retrieval:** Once it is clear which claims to fact-check, the next step is to gather relevant, trustworthy evidence for assessing the claim (Guo et al., 2022; Zeng et al., 2021).
- **Veracity Prediction:** Given the evidence, it is necessary to assess it to determine the veracity of the claim (Guo et al., 2022; Zeng et al., 2021).
- **Explanation:** Finally, for human use, one must explain the fact-checking outcome via human-understandable justification for the model's determination (Graves, 2018a; Guo et al., 2022; Kotonya & Toni, 2020a).

In the subsequent sections, we discuss each of the tasks above in the context of existing NLP research in automated fact-checking. Some other steps (for example, detecting propaganda in text, click-bait detection) are also pertinent to fact-checking but do not directly fit into the stages described above. They are briefly discussed in Section 3.5.

3. Task formulation for automated fact-checking

Modern Natural Language Processing is largely-data driven. In this article, we distinguish task formulation (conceptual) vs. dataset construction (implementation activity, given the task definition). That said, the availability of a suitable dataset or the feasibility of constructing a new dataset can also bear on how tasks are formulated.

3.1. Claim detection, checkworthiness, and prioritization

Fact-checkers and news organizations monitor information sources such as social media (Facebook, Twitter, Reddit, etc.), political campaigns and speeches, and public addresses from government officials on critical issues (Arnold, 2020; Nakov, Corney et al., 2021). Additional sources include tip-lines on end-to-end encrypted platforms (such as WhatsApp, Telegram, and Signal) (Kazemi, Garimella, Shahi, Gaffney and Hale, 2021). The volume of information on various platforms makes it challenging to efficiently monitor all sources for misinformation. Zeng et al. (2021) define the *claim detection* step as identifying, filtering, and prioritizing claims.

To identify claims, social media streams are often monitored for rumors (Guo et al., 2022). A rumor can be defined as a claim that is unverified and being circulated online (Zubiaga, Aker, Bontcheva, Liakata, & Procter, 2018). Rumors are characterized by the subjectivity of the language and the reach of the content to the users (Qazvinian, Rosengren, Radev, & Mei, 2011). Additionally, metadata related to virality, such as the number of shares (or retweets and re-posts), likes, or comments are also considered when identifying whether a post is a rumor (Gorrell et al., 2019; Zhang, Cao et al., 2021). However, detecting rumors alone is not sufficient to decide whether a claim needs to be fact-checked.

For each text of interest, the key questions fact-checking systems need to address include:

1. Is there a claim to check?
2. Does the claim contain verifiable information?
3. Is the claim checkworthy?
4. Has a trusted source already fact-checked the claim?

Regarding the first criterion – is there a claim – one might ask whether the claim contains a purported fact or an opinion (Hassan, Arslan, Li and Tremayne, 2017). For example, a statement such as *"reggae is the most soulful genre of music"* represents personal preference that is not checkable. In contrast, *"<NAME> won a gold medal in the Olympics"* is checkable by matching <NAME> to the list of all gold medal winners.

Whether the claim contains verifiable information is more challenging. For example, if a claim can only be verified by private knowledge or personal experience that is not broadly accessible, then it cannot be checked (Konstantinovskiy, Price, Babakar, & Zubiaga, 2021). For example, if someone claims to have eaten a certain food yesterday, it is probably impossible to verify beyond their personal testimony.

As this example suggests, the question of whether the claim contains verifiable information depends in large part on what evidence is available for verification. This, in turn, may not be clear until after evidence retrieval is performed. In practice fact-checkers may perform some preliminary research, but mostly try to gauge checkworthiness only based on the claim itself.

In addition, this consideration is only one of many that factors into deciding whether to check a claim. Even a claim that may appear to be unverifiable may still be of such great public interest that it is worth conducting the fact-check. Moreover, even if the fact-check is conducted and ultimately indeterminate (i.e., evidence does not exist either to verify or refute the claim), simply showing that a claim's veracity cannot be determined may still be a valuable outcome.

A claim is deemed *checkworthy* if a claim is of significant public interest or has the potential to cause harm (Hassan, Li, & Tremayne, 2015; Nakov, Da San Martino et al., 2021). For example, a claim related to the effect of a vaccine on the COVID-19 infection rate is more relevant to the public interest and hence more checkworthy than a claim about some philosopher's favorite food.

Claims, like memes, often appear several times and/or across multiple platforms (in the same form or with slight modification) (Arnold, 2020; Leskovec, Backstrom, & Kleinberg, 2009; Nakov, Corney et al., 2021). Fact-checking organizations maintain a growing database claims which have already been fact-checked. Thus, detected claims are compared against databases of already fact-checked claims by trusted organizations (Shaar, Alam, Da San Martino and Nakov, 2021; Shaar, Martino, Babulkov, & Nakov, 2020). Comparing new claims against such databases helps to avoid duplicating work on previously fact-checked claims. This step is also known as *claim matching* (Zeng et al., 2021).

Reports from practitioners argue that if a claim is not checked within the first few hours, a late fact-check does not have much impact on changing the ongoing misinformation narrative (Arnold, 2020; Nakov, Corney et al., 2021). Moreover, limited resources for fact-checking make it crucial for organizations to prioritize the claims to be checked (Borel, 2016). Claims can be prioritized based on their checkworthiness (Nakov et al., 2022; Nakov, Da San Martino et al., 2021). Nakov et al. (2022) note that checkworthiness is determined based on factors such as

1. How urgently a claim needs to be checked?
2. How much harm can a claim cause (Alam, Dalvi et al., 2021; Alam, Shaar et al., 2021; Shaar, Hasanain et al., 2021)?
3. Would the claim require attention from policy makers for addressing the underlying issue?

Note that estimating harms is quite challenging, especially without first having a thorough understanding and measures of harm caused by misinformation (Neumann, De-Arteaga, & Fazelpour, 2022).

The spread of a claim on social media provides another potential signal for identifying public interest (Arnold, 2020). In the spirit of doing "the greatest good for the greatest number", viral claims might be prioritized highly because any false information in them has the potential to negatively impact a large number of people. On the other hand, since fairness considerations motivate equal protections for all people, we cannot serve only the majority at the expense of minority groups (Ekstrand, Das, Burke, Diaz, et al., 2022; Neumann et al., 2022). Moreover, such minority groups may be more vulnerable, motivating greater protections, and may be disproportionately impacted by mis/disinformation (Guo et al., 2022). See Section 3.6 for additional discussion.

3.2. Evidence retrieval

Some sub-tasks in automated fact-checking can be performed without the presence of explicit evidence. For example, the linguistic properties of the text can be used to determine whether it is machine-generated (Rashkin, Choi, Jang, Volkova, & Choi, 2017; Wang, 2017). However, assessing claim veracity without evidence is clearly more challenging (Schuster, Schuster, Shah, & Barzilay, 2020).

Provenance of a claim can also signal information quality; known unreliable source or distribution channels are often repeat offenders in spreading false information.[2] Such analysis of provenance can be further complicated when content is systematically propagated by multiple sources (twitter misinformation bots) (Jones, 2019).

It is typically assumed that fact-checking requires gathering of reliable and trustworthy evidence that provides information to reason about the claim (Graves, 2018a; Li et al., 2016). In some cases, multiple aspects of a claim needs to be checked. A fact-checker would then decompose such a claim into distinct questions and gather relevant evidence for the question (Borel, 2016; Chen et al., 2022). From an information retrieval (IR) perspective, we can conceptualize each of those questions as an "information need" for which the fact-checker must formulate one or more queries to a search engine (Bendersky, Metzler, & Croft, 2012) in order to retrieve necessary evidence.

Evidence can be found across many modalities, including text, tables, knowledge graphs, images, and videos. Various metadata can also provide evidence and are sometimes required to assess the claim. Examples include context needed to disambiguate claim terms, or background of the individual or organization from whom the claim originated.

Retrieving relevant evidence also depends on the following questions (Singh, Das, Li, & Lease, 2021):

1. Is there sufficient evidence available related to a claim?
2. Is it accessible or available in the public domain?
3. Is it in a format that can be read and processed?

As noted earlier in Section 3.1, the preceding claim detection task involves assessing whether a claim contains verifiable information; this depends in part on what evidence exists to be retrieved, which is not actually known until evidence retrieval is performed. Having now reached this evidence retrieval step, we indeed discover whether sufficient evidence exists to support or refute the claim.

Additionally, evidence should be trustworthy, reputable (Lease, 2018; Nguyen, Kharosekar, Lease and Wallace, 2018), and unbiased (Chen, Khashabi, Yin, Callison-Burch and Roth, 2019).

Once evidence is retrieved, *stance detection* assesses the degree to which the evidence supports or refutes the claim (Ferreira & Vlachos, 2016; Nguyen, Kharosekar, Krishnan et al., 2018; Popat, Mukherjee, Yates, & Weikum, 2018). Stance detection is typically formulated as a classification task (or ordinal regression) over each piece of retrieved evidence. Note that some works formulate stance detection as an independent task (Hanselowski et al., 2018; Hardalov et al., 2021).

3.3. Veracity prediction

Given a claim and gathered evidence, *veracity prediction* involves reasoning over the collected evidence and the claim. Veracity prediction can be formulated as a binary classification task (i.e., true vs. false) (Nakashole & Mitchell, 2014; Popat et al., 2018; Potthast, Kieselj, et al., 2018), or as a fine-grained, multi-class task following the journalistic fact-checking practices (Augenstein et al., 2019; Shu, Mahudeswaran, Wang, Lee, & Liu, 2020; Wang, 2017). In some cases, there may not be enough information available to determine the veracity of a claim (Thorne, Vlachos, Christodoulopoulos, & Mittal, 2018).

Note that fact-checking is potentially a recursive process because retrieved evidence may itself need to be fact-checked before it can be trusted and acted upon (Graves, 2018a). This is also consistent with broader educational practices in information literacy[3] in which readers are similarly encouraged to evaluate the quality of information they consume. Such assessment of information reliability can naturally integrate with the veracity prediction task in factoring in the reliability of the evidence along with its stance (Guo et al., 2022; Nguyen, Kharosekar, Lease et al., 2018).

3.4. Explaining veracity prediction

While a social media platform might use automated veracity predictions in deciding whether to automatically block or demote content, the use of fact-checking technology often involves a human-in-the-loop, whether it is a platform moderator, a journalist, or an end-user. When we consider such human-centered use of fact-checking technologies, providing an automated veracity prediction without justifying the answer can cause a system to be ignored or distrusted, or even reinforce mistaken human beliefs in false claims (the "backfire effect" Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012). Explanations and justifications are especially important given the noticeable drop in performance of state-of-the-art NLP systems when facing adversarial examples (Kovatchev et al., 2022). Consequently, automated fact-checking systems intended for human-consumption should seek to explain their veracity predictions in a similar manner to that of existing journalistic fact-checking practices (Uscinski, 2015). A brief point to make is that

[2] https://disinformationindex.org/.

[3] https://en.wikipedia.org/wiki/Information_literacy.

much of the explanation research has focused on explanations for researchers and engineers engaged in system development (types of explanations, methods of generating them, and evaluation regimens). In contrast, we emphasize here explanations for *system users*.

Various types of explanations can be provided, such as through

1. evidence attribution
2. explaining the decision-making process for a fact-check
3. summarizing the evidence
4. case-based explanations

Evidence attribution is the process of identifying evidence or a specific aspect of the evidence (such as paragraphs, sentences, or even tokens of interest) (Lu & Li, 2020; Popat et al., 2018; Shu, Cui, Wang, Lee, & Liu, 2019; Thorne et al., 2018). Furthermore, the relative importance of the evidence can also justify the fact-checking outcome (Nguyen, Kharosekar, Krishnan et al., 2018). Alternatively, a set of rules or interactions to break down parts of the decision-making process can also serve as an explanation (Gad-Elrab, Stepanova, Urbani, & Weikum, 2019; Nguyen, Kharosekar, Lease et al., 2018). Such formulation focuses more on how the evidence is processed to arrive at a decision. Explaining the veracity can also be formulated as a summarization problem over the gathered evidence to explain a fact-check (Atanasova, Simonsen, Lioma and Augenstein, 2020; Kotonya & Toni, 2020b). Finally, case-based explanations can provide the user with similar, human-labeled instances (Das, Gupta, Kovatchev, Lease, & Li, 2022).

3.5. Related tasks

In addition to tasks that are considered central to the automated fact-checking pipeline, some additional tasks bear mentioning as related and complementary to the fact-checking enterprise. Examples of such tasks include propaganda detection (Da San Martino et al., 2020), clickbait detection (Potthast, Köpsel, Stein, & Hagen, 2016), and argument mining (Lawrence & Reed, 2020). Furthermore, some tasks can be formulated independent of the fact-checking pipeline and utilized later to improve individual fact-checking sub-tasks. For example, predicting the virality of social media content (Jain, Garg, & Jain, 2021) can help improve claim detection and claim checkworthiness. Similarly, network analysis on fake news propagation (Shao, Ciampaglia, Flammini, & Menczer, 2016) can help in analyzing provenance.

With an eye toward building more human-centered AI approaches, there are also some tasks that could be applied to help automate parts of the fact-checking process. For example, claim detection might be improved via an URL recommendation engine for content that might need fact-checking (Vo & Lee, 2018). Additionally, fact-checkers could benefit from a predicted score for claim difficulty (Singh et al., 2021). In terms of evidence retrieval and veracity prediction, one might generate fact-checking briefs to aid inexpert fact-checkers (Fan et al., 2020). Instead of summarizing the evidence in general (Section 3.4), one might instead summarize with the specific goal of decision support (Hsu & Tan, 2021).

3.6. Key challenges

Most work in automated fact-checking has been done on veracity prediction, and to a lesser extent, on explanation generation. Recently, we have seen more attention directed toward claim detection and checkworthiness. In contrast, work on evidence retrieval remains less developed.

Claim detection Guo et al. (2022) points out several sources of biases in the claim check-worthiness task. Claims could be of variable interest to different social groups. Additionally, claims that might cause more harm to marginalized groups compared to the general population may not get enough attention. Ideally, models identifying check-worthiness need to overcome any possible disparate impact.

Similar concerns appear in the report by Full Fact (Arnold, 2020). One of the criteria for selecting a claim for fact-checking across several organizations is "Could the claim threaten democratic processes or minority groups?" However, such criterion may be at odds with the concerns of virality. Fact-checking organizations often monitor virality metrics to decide which claims to fact-check (Arnold, 2020; Nakov, Corney et al., 2021). Nevertheless, if a false claim is targeted toward an ethnic minority, such claims may not cross the virality thresholds.

Prioritizing which claims to fact-check requires attention to various demographic traits: content creators, readers, and subject matter. Claim check-worthiness dataset design can thus benefit from consideration of demographics.

Evidence retrieval Evidence retrieval has been largely neglected in the automated fact-checking NLP literature. It is often assumed that evidence is already available, or, coarse-grained evidence is gathered from putting the claims into a search engine (Nguyen, Kharosekar, Krishnan et al., 2018; Popat et al., 2018). However, Hasanain and Elsayed (2022) show in their study that search engines optimized for relevance seldom retrieve evidence most useful for veracity prediction. Although retrieving credible information has been studied thoroughly in IR (Clarke, Rizvi, Smucker, Maistro, & Zuccon, 2020), more work is needed that is focused on retrieving evidence for veracity assessment (Clarke et al., 2020; Lease, 2018).

Veracity prediction and explanation A critical challenge for automated systems is to reason over multiple sources of evidence while also taking source reputation into account. Additionally, explaining a complex reasoning process is a non-trivial task. The notion of model explanations itself is polysemous and evolving in general, not to mention in the context of fact checking. As explainable NLP develops, automated fact-checking tasks also need to evolve and provide explanations that are accessible to human stakeholders yet faithful to the underlying model. For example, case-based explanations are mostly unexplored in automated fact-checking, although working systems have been proposed for propaganda detection (Das et al., 2022).

In many NLP tasks, such as machine translation or natural language inference, the goal is to build fully-automated, end-to-end solutions. However, in the context of fact-checking, state-of-the-art limitations suggest the need for humans-in-the-loop for the foreseeable future. Given this, automated tooling to support human fact-checkers is crucial. However, understanding the fact-checker needs and incorporating those needs in the task formulation has been largely absent from the automated fact-checking literature, with a few notable exceptions (Demartini et al., 2020; Nakov, Corney et al., 2021). Future research could benefit from greater involvement of fact-checkers in the NLP research process and shifting goals from complete automation toward human support.

4. Dataset construction

Corresponding to task formulation (Section 3), our presentation of fact-checking datasets is also organized around claims, evidence, veracity prediction, and explanation. Note that not all datasets have all of these components.

4.1. Claim detection and claim check-worthiness

Claim detection datasets typically contain claims and their sources (documents, social media streams, transcripts from political speeches) (Guo et al., 2022). One form of claim detection is identifying rumors on social media, where datasets are primarily constructed with text from Twitter (Qazvinian et al., 2011; Zubiaga et al., 2018) and Reddit (Gorrell et al., 2019; Lillie, Middelboe, & Derczynski, 2019). Some works provide the claims in the context they appeared on social media (Ma et al., 2016; Zhang, Cao et al., 2021). However, several studies note that most claim detection datasets do not contain enough context. As the discussion of metadata in Section 3 suggests, broader context might include: social media reach, virality metrics, the origin of a claim, and relevant user data (i.e., who posted a claim, how influential they are online, etc.) (Arnold, 2020; Nakov, Corney et al., 2021).

Claim check-worthiness datasets (Atanasova et al., 2018; Barrón-Cedeño et al., 2020; Hassan et al., 2015; Konstantinovskiy et al., 2021; Nakov, Da San Martino et al., 2021; Shaar, Hasanain et al., 2021) filter claims from a source (similar to claim detection, sources include social media feeds and political debate transcripts, among others) by annotating claims based on the checkworthiness criteria (mentioned in Section 3.1). Each claim is given a checkworthiness score to obtain a ranked list. Note that claim detection and checkworthiness datasets may be expert annotated (Hassan et al., 2015) or crowd annotated (Atanasova et al., 2018; Barrón-Cedeño et al., 2020; Konstantinovskiy et al., 2021; Nakov, Da San Martino et al., 2021; Shaar, Hasanain et al., 2021).[4]

The datasets discussed above do not capture multi-modal datasets, and few do. One such dataset is r/Fakeddit (Nakamura, Levy, & Wang, 2020). This dataset contains images and associated text content from Reddit as claims. Misinformation can also spread through multi-modal memes, and tasks such as Facebook (now Meta) *Hateful Memes Challenge* (Kiela et al., 2020) for hate speech suggest what might be similarly done for misinformation detection.

4.2. Evidence

Early datasets in fact-checking provide metadata with claims as the only form of evidence. Such metadata include social media post properties, user information, publication date, source information (Potthast et al., 2018; Wang, 2017). As discussed earlier in Section 3.2, such metadata does not contain the world knowledge necessary to reason about a complex claim. To address the above limitations, recent datasets consider external evidence (Guo et al., 2022).

Evidence is collected differently depending upon the problem setup. For artificial claims, evidence is often retrieved from a single source such as Wikipedia articles (Jiang et al., 2020; Schuster, Fisch, & Barzilay, 2021; Thorne et al., 2018). Domain limited evidence for real-world claims is collected from problem-specific sources, such as academic articles for scientific claims (Kotonya & Toni, 2020b; Wadden et al., 2020), or specific evidence listed in fact-checking websites (Hanselowski, Stab, Schulz, Li, & Gurevych, 2019; Vlachos & Riedel, 2014). Open-domain evidence for real-world claims is usually collected from the web via search engines (Augenstein et al., 2019; Popat et al., 2018).

Recently, there has been more work considering evidence beyond free text. Such formats include structured or semi-structured forms of evidence. Sources include knowledge bases for structured form of evidence (Shi & Weninger, 2016) and semi-structured evidence from semi-structured knowledge bases (Vlachos & Riedel, 2015), tabular data (Chen, Khashabi et al., 2019; Gupta, Mehta, Nokhiz, & Srikumar, 2020), and tables within a document (Aly et al., 2021).

Additionally, there are some retrieval-specific datasets that aim at retrieving credible information from search engines (Clarke et al., 2020). However, such tasks do not incorporate claim checking as an explicit task.

[4] Some of these datasets, such as the CheckThat! datasets, are partially crowd and partially expert annotated.

4.3. Veracity prediction

Evidence retrieval and veracity prediction datasets are usually constructed jointly. Note, in some cases, evidence may be absent from the datasets. Veracity prediction datasets usually do not deal with claim detection or claim checkworthiness tasks separately. Instead, such datasets contain a set of claims that are either artificially constructed or collected from the internet.

Artificial claims in veracity prediction datasets are often limited in scope and constructed for natural language reasoning research (Aly et al., 2021; Chen, Wang et al., 2019; Jiang et al., 2020; Schuster et al., 2021; Thorne et al., 2018). For example, FEVER (Thorne et al., 2018) and HoVer (Jacovi & Goldberg, 2021) obtain claims from Wikipedia pages. Some datasets also implement subject–predicate–object triplets for fact-checking against knowledge bases (Kim & Choi, 2020; Shi & Weninger, 2016).

Fact-checking websites are popular sources for creating veracity prediction datasets based on real claims. Several datasets obtain claims from either a single website or collect claims from many such websites and collate them (Augenstein et al., 2019; Hanselowski et al., 2018; Vlachos & Riedel, 2014; Wang, 2017). Note that such claims are inherently expert annotated. Other sources of claims are social media (Potthast et al., 2018; Shu, Sliva, Wang, Tang, & Liu, 2017), news outlets (Gruppi, Horne, & Adalı, 2021; Horne, Khedr, & Adali, 2018; Nørregaard, Horne, & Adalı, 2019), blogs, discussions in QA forums, or similar user-generated publishing platforms (Mihaylova et al., 2018).

Additionally some fact-checking datasets target domain-specific problems such as scientific literature (Wadden et al., 2020), climate change (Diggelmann, Boyd-Graber, Bulian, Ciaramita, & Leippold, 2020), and public health (Kotonya & Toni, 2020b). Most datasets are monolingual but recent effort have started to incorporate multi-lingual claims (Barnabò et al., 2022; Gupta & Srikumar, 2021).

Early datasets focus on a binary veracity prediction — true or false (Mihalcea & Strapparava, 2009). Recent datasets often adopt an ordinal veracity labeling scheme that mimics fact checkering websites (Augenstein et al., 2019; Vlachos & Riedel, 2014; Wang, 2017). However, every fact-checking website has a different scale for veracity, so datasets that span across multiple websites come with a normalization problem. While some datasets do not normalize the labels (Augenstein et al., 2019), some normalize them post-hoc (Gupta & Srikumar, 2021; Hanselowski et al., 2019; Kotonya & Toni, 2020a).

4.4. Explanation

While an explanation is tied to veracity prediction, only a few datasets explicitly address the problem of explainable veracity prediction (Alhindi, Petridis, & Muresan, 2018; Atanasova, Simonsen et al., 2020; Kotonya & Toni, 2020b). Broadly in NLP, often parts of the input is highlighted to provide an explanation for the prediction. This form of explanations is known as extractive rationale (Kutlu, McDonnell, Elsayed, & Lease, 2020; Zaidan, Eisner, & Piatko, 2007). Incorporating the idea of the extractive rationale, some datasets include a sentence from the evidence along with the label (Hanselowski et al., 2018; Schuster et al., 2021; Thorne et al., 2018; Wadden et al., 2020). Although such datasets do not explicitly define evidence as a form of explanation in such cases, the line between evidence retrieval and explanation blurs if the evidence is the explanation. However, explanations are different from evidence in a few ways. Particularly, explanations need to be concise for user consumption, while evidence can be a collection of documents or long documents. Explanations are user sensitive. Consequently, evidence alone as a form of explanation might have some inherent assumption about the user that might not be understandable for different groups of users (e.g., experts vs. non-experts).

4.5. Challenges

Claims Checkworthiness datasets are highly imbalanced, i.e., the number of checkworthy claims are relatively low compared to non-checkworthy claims (Williams, Rodrigues, & Novak, 2020). Datasets are also not generalizable due to their limited domain-specific context (Guo et al., 2022). Additionally, while existing datasets cover various languages such as English, Arabic, Spanish, Bulgarian, and Dutch, they are primarily monolingual. Consequently, building multilingual checkworthiness predictors is still challenging. Much of the data in check-worthiness datasets is not originally intended to be used in classification. The criteria used by different organizations when selecting which claims to check is often subjective and may not generalize outside of the particular organization.

Some annotation practices can result in artifacts in the dataset. For example, artificially constructed false claims, such as a negation-based false claim in FEVER, can lead to artifacts in models (Schuster et al., 2021). Models do not generalize well beyond the dataset because they might overfit to the annotation schema (Bansal et al., 2019). One way to identify such blind spots is by using adversarial datasets for fact-checking. Such a setting is incorporated in FEVER 2.0 (Thorne & Vlachos, 2019).

Datasets constructed for research may not always capture how fact-checkers work in practice. This leads to limitations in the algorithms built on them. For example, interviews with fact-checkers report that they tend to consider a combination of contents of the posts and associated virality metrics (indicating reach) during fact-checking (Arnold, 2020). However, most fact-checking datasets do not include virality metrics.

Evidence retrieval Some datasets have been constructed by using a claim verbatim as a query and taking the top search results as evidence. However, some queries are better than others for retrieving desired information. Consequently, greater care might be taken in crafting effective queries or otherwise improving evidence retrieval such that resulting datasets are more likely to contain quality evidence for veracity prediction. Otherwise, poor quality evidence becomes a bottleneck for the quality of the models trained at the later stages in the fact-checking pipeline (Singh et al., 2021).

Veracity prediction A key challenge in veracity prediction datasets is that the labels are not homogeneous across fact-checking websites and normalizing might introduce noise.

Explanation Some datasets include entire fact-checking articles as evidence and their summaries as the form of explanation (Atanasova, Simonsen et al., 2020; Kotonya & Toni, 2020b). In such cases, "explanation" components assume an already available fact-checking article. Instead, providing abstractive summaries and explaining the reasoning process over the evidence would be more valuable.

Data generation Recent years have seen an increasing interest in the use of data generation and data augmentation for various NLP tasks (Dhole et al., 2021; Hartvigsen et al., 2022; Kovatchev, Smith, Lee, & Devine, 2021; Liu, Swayamdipta, Smith, & Choi, 2022). The use of synthetic data has not been extensively explored in the context of fact-checking.

5. Automating fact-checking

NLP research in automated fact-checking has primarily focused on building models for different automated fact-checking tasks utilizing existing datasets. In the following section, we highlight the broad modeling strategies employed in the literature, with more detailed discussion related to explainable methods for automated fact-checking.

5.1. General NLP capabilities

Claim detection and checkworthiness While claim detection is usually implemented as a classification task only, claim checkworthiness is typically implemented both as ranking (Nakov, Corney et al., 2021) and classification task (Zeng et al., 2021). As discussed earlier in the task formulation Section 3.1, the broad task of claim detection can be broken down into sub-tasks of identifying claims, filtering duplicate claims, and prioritizing claims based on their checkworthiness. Another instance of identifying claims is detecting rumors in social media streams.

Some early works in rumor detection focus on feature engineering from available metadata the text itself (Aker, Derczynski, & Bontcheva, 2017; Enayet & El-Beltagy, 2017; Zhou, Jain, Phoha, & Zafarani, 2020). More advanced methods for claim detection involve LSTM and other sequence models (Kochkina, Liakata, & Augenstein, 2017). Such models are better at capturing the context of the text (Zubiaga, Liakata, Procter, Wong Sak Hoi, & Tolmie, 2016). Tree-LSTM (Ma, Gao, & Wong, 2018) and Hierarchical attention networks (Guo, Cao, Zhang, Guo, & Li, 2018) capture the internal structure of the claim or the context in which the claim appears. Additionally, graph neural network approaches can capture the related social media activities along with the text (Monti, Frasca, Eynard, Mannion, & Bronstein, 2019).

Similarly, early works in claim-checkworthiness utilize support vector machines using textual features and rank the claims in terms of their priorities (Hassan, Arslan et al., 2017). For example, Konstantinovskiy et al. (2021) build a classification model for checkworthiness by collapsing the labels to checkable vs. non-checkable claim. They build a logistic regression model that uses word embeddings along with syntax based features (parts of speech tags, and named entities). Neural methods such as LSTM performed well in earlier checkworthiness shared tasks (Elsayed et al., 2019). Additionally, Atanasova, Nakov, Màrquez et al. (2019) show that capturing context helps with the checkworthiness task as well. Models such as RoBERTa obtained higher performance in the later edition of the *CheckThat!* shared task (Martinez-Rico, Martinez-Romo, & Araujo, 2021; Williams et al., 2020) for English language claims. Fine-tuning such models for claim detection tasks has become more prevalent for claim checkworthiness in other languages as well (Hasanain & Elsayed, 2020; Williams et al., 2020).

Filtering previously fact-checked claims is a relatively new task in this domain. Shaar et al. (2020) propose an approach using BERT and BM-25 to match claims against fact-checking databases for matching claims with existing databases. Additionally, fine-tuning RoBERTa on various fact-checking datasets resulted in high performance for identifying duplicate claims (Bouziane, Perrin, Cluzeau, Mardas, & Sadeq, 2020). Furthermore, a combination of pretrained model Sentence-BERT and re-ranking with LambdaMART performed well for detecting previously fact-checked claims (Nakov, Da San Martino et al., 2021).

Evidence retrieval and veracity prediction Evidence retrieval and veracity prediction in the pipeline can be modeled sequentially or jointly. Similar to claim detection and checkworthiness models, early works use stylistic features and metadata to arrive at veracity prediction without external evidence (Rashkin et al., 2017; Wang, 2017). Models that include evidence retrieval often use commercial search APIs or some retrieval approach such as TF–IDF, and BM25 (Thorne et al., 2018). Similar to question-answering models, some works adopt a two-step approach. First a simpler model (TF–IDF or BM-25) is used at scale and then a more complex model is used for re-ranking after the initial pruning (Hanselowski et al., 2019; Nie, Wang, & Bansal, 2019; Thorne et al., 2018). Additionally, document vs. passage retrieval, or 2-stage "telescoping" approaches, are adopted where the first stage is retrieving related documents and the second stage is to retrieve the relevant passage. Two stage approaches are useful for scaling up applications as the first stage is more efficient than the second stage. For domain specific evidence retrieval, using domain-bound word embeddings has been shown to be effective (Zeng et al., 2021).

The IR task is not always a part of the process. Instead, it is often assumed that reliable evidence is already available. While this simplifies the fact-checking task so that researchers can focus on veracity prediction, in practice evidence retrieval is necessary and cannot be ignored. Moreover, in practice one must contend with noisy (non-relevant), low quality, and biased search results during inference.

As discussed earlier in Section 3.3, assessing the reliability of gathered evidence may be necessary. If the evidence is assumed to be trustworthy, then it suffices to detect the stance of each piece of evidence and then aggregate (somehow) to induce veracity

(e.g., perhaps assuming all evidence is equally important and trustworthy). However, often one must contend with evidence "in the wild" of questionable reliability, in which case assessing the quality (and bias) of evidence is an important precursor to using it in veracity prediction.

Veracity prediction utilizes textual entailment for inferring veracity over either a single document as evidence or over multiple documents. Dagan, Dolan, Magnini, and Roth (2010) define *textual entailment* as "deciding, given two text fragments, whether the meaning of one text is entailed (can be inferred) from another text." Real-world applications often require reasoning over multiple documents (Augenstein et al., 2019; Kotonya & Toni, 2020b; Schuster et al., 2021). Reasoning over multiple documents can be done either by concatenation (Nie et al., 2019) or weighted aggregation (Nguyen, Kharosekar, Lease et al., 2018). Weighted aggregation virtually re-ranks the evidence considered to filter out the unreliable evidence (Ma, Gao, Joty, & Wong, 2019; Pradeep, Ma, Nogueira, & Lin, 2021). Some approaches also use Knowledge Bases as the central repository of all evidence (Shi & Weninger, 2016). However, evidence is only limited to what is available in the knowledge base (Guo et al., 2022; Zeng et al., 2021). Moreover, a fundamental limitation of knowledge bases is that not all knowledge fits easily into structured relations.

Recent developments in large language models help extend the knowledge base approach. Fact-checking models can rely on pretrained models to provide evidence for veracity prediction (Lee et al., 2020). However, this approach can encode biases present in the language model (Lee, Bang, Madotto, & Fung, 2021).

An alternative approach is to help fact-checkers with downstream tasks by processing evidence. An example of such work is generating summaries over available evidence using BERT (Fan et al., 2020).

Limitations With the recent development of large, pre-trained language models and deep learning for NLP, we see a significant improvement across the fact-checking pipeline. With the introduction of FEVER (Aly et al., 2021; Thorne et al., 2018; Thorne, Vlachos, Cocarascu, Christodoulopoulos, & Mittal, 2019) and *CheckThat!* (Nakov, Da San Martino et al., 2021) we have benchmarks for both artificial and real-life claim detection and verification models. However, even the state-of-the-art NLP models perform poorly on the benchmarks above. For example, the best performing model on FEVER 2018 shared task (Thorne et al., 2018) reports an accuracy of 0.67.[5] Models perform worse on multi-modal shared task FEVEROUS (Aly et al., 2021): the best performing model reports 0.56 accuracy score.[6] Similarly, the best checkworthiness model only achieved an average precision of 0.65 for Arabic claims and 0.224 for English claims in the *CheckThat!* 2021 shared task for identifying checkworthiness in tweets (Nakov, Da San Martino et al., 2021). On the other hand, the best performing model for identifying check-worthy claims in debates reports 0.42 average precision. Surprisingly, Barrón-Cedeño et al. (2020), the top performing model for checkworthiness detection, report an average precision of 0.806 (Williams et al., 2020). For the fact-checking task of CheckThat! 2021 (Nakov, Da San Martino et al., 2021), the best performing model reports a 0.83 macro F1 score. However, the second-best model only reports a 0.50 F1 score. Given this striking gap in performance between the top system vs. others, it would be valuable for future work to benchmark systems on additional datasets in order to better assess the generality of these findings.

It is not easy to make a direct comparison between different methods that are evaluated in different settings and with different datasets (Zeng et al., 2021). Moreover, the pipeline design of automated fact-checking creates potential bottlenecks, e.g., performance on the veracity prediction task on most datasets is dependent on the claim detection task performance or the quality of the evidence retrieved. Extensive benchmarks are required to incorporate all of the prior subtasks in the fact-checking pipeline systematically (Zeng et al., 2021).

Much of AI research is faced with a fundamental trade-off between working with diverse formulations of a problem and standardized benchmarks for measuring progress. This trade-off also impacts automated fact-checking research. While there exist benchmarks such as FEVER and the *CheckThat!*, most models built on those benchmarks may not generalize well in a practical setting. Abstract and tractable formulations of a problem may help us develop technologies that facilitate practical adoption. However, practical adoption requires significant engineering effort beyond the research setting. Ideally, we would like to see automated fact-checking research continue to move toward increasingly realistic benchmarks while incorporating diverse formulations of the problem.

5.2. Explainable approaches

Although the terms *interpretability* and *explainability* are often used interchangeably, and some times defined to be so (Molnar, 2020), we distinguish interpretability vs. explainability similar to Kotonya and Toni (2020a). Specifically, *interpretability* represents methods that provide direct insight into an AI system's components (such as features and variables), often requiring some understanding of the specific to the algorithm, and often built for expert use cases such as model debugging. *Explainability* represents methods to understand an AI model without referring to the actual component of the systems. Note that, in the task formulation section, we have also talked about explaining veracity prediction. The goal of such explanation stems from fact-checker needs to help readers understand the fact-checking verdict. Therefore, explaining veracity prediction aligns more closely with explainability over interpretability. When the distinction between explainability vs. interpretability does not matter, we follow Vaughan and Wallach (2020) in adopting *intelligibility* (Vaughan & Wallach, 2020) as an umbrella term for both concepts.

Sokol and Flach (2019) propose a desiderata for designing user experience for machine learning applications. Kotonya and Toni (2020a) extend them in the context of fact-checking and suggest eight properties of intelligibility: *actionable*, *causal*, *coherent*, *context-full*, *interactive*, *unbiased or impartial*, *parsimonious*, and *chronological*.

Additionally, there are three dimensions specifically for explainable methods in NLP (Jacovi & Goldberg, 2020):

[5] https://fever.ai/2018/task.html.

[6] https://fever.ai/task.html.

1. **Readability:** are explanations clear?
2. **Plausibility:** are explanations compelling or persuasive?
3. **Faithfulness:** are explanations faithful to the model's actual reasoning process?

In comparison with the available intelligibility methods in NLP (Wiegreffe & Marasovic, 2021), only a few are applied to existing fact-checking works. Below, we highlight only commonly observed explainable fact-checking methods (also noted by Kotonya and Toni (2020a)).

Attention-based intelligibility Despite the debate about attention being a reliable intelligibility method (Bibal et al., 2022; Jain & Wallace, 2019; Serrano & Smith, 2019; Wiegreffe & Pinter, 2019), it remains a popular method in existing deep neural network approaches in fact-checking. Attention-based explanations are provided in various forms:

1. highlighting tokens in articles (Popat et al., 2018)
2. highlighting salient excerpts from evidence utilizing comments related to the post (Shu et al., 2019)
3. n-gram extraction using self-attention (Yang et al., 2019)
4. attention from different sources other than the claim text itself, such as the source of tweets, retweet propagation, and retweeter properties (Lu & Li, 2020)

Rule discovery as explanations Rule mining is a form of explanation prevalent in knowledge base systems (Ahmadi, Lee, Papotti, & Saeed, 2019; Gad-Elrab et al., 2019). These explanations can be more comprehensive, but as noted in the previous section, not all statements can be fact-checked via knowledge-based methods due to limitations of the underlying knowledge-base itself. Some approaches provide general purpose rule mining in an attempt to address this limitation (Ahmadi, Truong, Dao, Ortona, & Papotti, 2020).

Summarization as explanations Both extractive and abstractive summaries can provide explanations for fact-checking. Atanasova, Simonsen et al. (2020) provides natural language summaries to explain the fact-checking decision. They explore two different approaches — explanation generation and veracity prediction as separate tasks, and joint training of the both. Joint training performs worse than single training. Kotonya and Toni (2020b) combine abstractive and extractive approaches to provide a novel summarization approach. Brand, Roitero, Soprano, Rahimi, and Demartini (2018) show jointly training prediction and explanation generation with encoder–decoder models such as BART (Lewis et al., 2020) results in explanations that help the crowd to perform better veracity assessment.

Counterfactuals and adversarial methods Adversarial attacks on opaque models help to identify any blind-spots, biases and discover data artifacts in models (Ribeiro, Wu, Guestrin, & Singh, 2020). Shared task FEVER 2.0 (Thorne et al., 2019) asked participants to devise methods for generating adversarial claims to identify weaknesses in the fact-checking methods. Natural language generation models such as GPT-3 can assist in formulating adversarial claims. More control over the generation can come from manipulating the input to natural language generation methods and constraining the generated text within original vocabulary (Niewiński, Pszona, & Janicka, 2019). Atanasova, Wright and Augenstein (2020) generate claims with n-grams inserted into the input text. Thorne and Vlachos (2019) experiment with several adversarial methods such as rule-based adversary, semantically equivalent adversarial rules (or SEARS) (Ribeiro, Singh, & Guestrin, 2018), negation, and paraphrasing-based adversary. Adversarial attacks are evaluated based on the *potency* (correctness) of the example and reduction in system performance. While methods such as SEARS and paraphrasing hurt the system performance, hand-crafted adversarial examples have higher potency score.

Interpretable methods (non-BlackBox) Some fact-checking works use a white-box or inherently interpretable model for fact-checking. Nguyen, Kharosekar, Krishnan et al. (2018) and Nguyen, Kharosekar, Lease et al. (2018) utilize a probabilistic graphical model and build an interactive interpretable model for fact-checking where users are allowed to directly override model decisions. Kotonya, Spooner, Magazzeni, and Toni (2021) propose an interpretable graph neural network for interpretable fact-checking on FEVEROUS dataset (Aly et al., 2021).

Limitations Intelligible methods in NLP and specifically within fact-checking are still in their infancy. Analysis of Kotonya and Toni (2020a) shows that most methods do not fulfill the desiderata mentioned earlier in this section. Specifically, they find that none of the existing models meet the criteria of being actionable, causal, and chronological. They also highlight that no existing method explicitly analyzes whether explanations are impartial. Some forms of explanations, such as rule-based triplets, are unbiased as they do not contain sentences or contain fragments of information (Kotonya & Toni, 2020a).

Some explainable methods address a specific simplified formulation of the task. For example, Atanasova, Simonsen et al. (2020) and Kotonya and Toni (2020b) both assume that expert-written fact-checking articles already exist. They provide explanations as summaries of the fact-checking article. However, in practice, a fact-checking system would not have access to such an article for an unknown claim.

In the case of automated fact-checking, most intelligible methods focus on explaining the outcome rather than describing the process to arrive at the outcome (Kotonya & Toni, 2020a). Moreover, all of the tasks in the fact-checking pipeline have not received equal attention for explainable methods. Kotonya and Toni (2020a) also argue that automatic fact-checking may benefit from explainable methods that provide insight into how outcomes of earlier sub-task in the fact-checking pipeline impact the outcome of later subtasks.

Most explainable NLP works evaluate explanation quality instead of explanation utility or faithfulness. Jacovi and Goldberg (2020) argue for a thorough faithfulness evaluation for explainable models. For example, even though attention-based explanations

may provide quality explanations, they may not necessarily be faithful. Moreover, explanation utility requires separate evaluation by measuring whether explanations improve both (i) human understanding of the model (Hase & Bansal, 2020) and (ii) human effectiveness of the downstream task (Nakov, Corney et al., 2021). Additionally, most intelligible methods establish only one-way communication from the model to humans. Instead, explanations might improve the model and human performance by establishing a bidirectional feedback loop.

5.3. Human-in-the-loop approaches

Human-in-the-loop (HITL) approaches can help scale automated solutions while utilizing human intelligence for complex tasks. There are different ways of applying HITL methods, e.g., delegating sub-tasks to crowd workers (Demartini, Difallah, & Cudré-Mauroux, 2012; Demartini, Trushkowsky, Kraska, Franklin, & Berkeley, 2013; Sarasua, Simperl, & Noy, 2012), active learning (Settles, 2009; Zhang, Lease, & Wallace, 2017), interactive machine learning (Amershi, Cakmak, Knox, & Kulesza, 2014; Joachims & Radlinski, 2007), and decision support systems where humans make the final decision based on model outcome and explanations (Zanzotto, 2019).

While HITL approaches in artificial intelligence are prevalent, only a few recent works employ such approaches in fact-checking. HITL approaches are predominantly more present in the veracity prediction task than other parts of the pipeline. For example, Demartini et al. (2020) propose a HITL framework for combating online misinformation. However, they only consider hybrid approaches for two sub-tasks in the fact-checking pipeline: (a) claim check-worthiness and (b) truthfulness judgment (same as veracity prediction). Below, we discuss the existing HITL approaches by how the system leverages human effort for each sub-task in the fact-checking pipeline.

Claim detection, checkworthiness, and prioritization Social media streams are often monitored for rumors as a part of the claim detection task (Guo et al., 2022). Farinneya, Abdollah Pour, Hamidian, and Diab (2021) apply an active learning-based approach at the claim detection stage for identifying rumors on social media. In-domain data is crucial for traditional supervised methods to perform well for rumor detection (Ahsan, Kumari, & Sharma, 2019), but in real-world scenarios, sufficient in-domain labeled data may not be available in the early stages of development. A semi-supervised approach such as active learning is beneficial for achieving high performance with fewer data points. Empirical results shows that Tweet-BERT, along with the least confidence-based sample selection approach, performs on par with supervised approaches using far less labeled data (Farinneya et al., 2021).

Similarly, Tschiatschek, Singla, Gomez Rodriguez, Merchant, and Krause (2018) propose a HITL approach that aims to automatically aggregate user flags and recommend a small subset of the flagged content for expert fact-checking. Their Bayesian inference-based approach jointly learns to detect fake news and identify the accuracy of user flags over time. One strength of this approach is that the algorithm improves over time in identifying users' flagging accuracy. Consequently, over time this algorithm's performance improves. This approach is also robust against spammers. By running the model on publicly available Facebook data where a majority of the users are adversarial, experiments show that their algorithm still performs well.

Duke's Tech & Check team implemented HITL at the claim check-worthiness layer (Adair & Stencel, 2020). To avoid flagging false check-worthy claims, a human expert would sort claims detected by ClaimBuster (Hassan, Zhang et al., 2017), filter out the ones deemed more important for fact-checkers, and email them to several organizations. In essence, this approach helped fact-checkers prioritize the claims to check through an additional level of filtering. Currently, several published fact-checks on PolitiFact were first alerted by the emails from Tech & Check.

Note that the *CheckThat!* (Atanasova, Nakov, Karadzhov, Mohtarami and Da San Martino, 2019; Barrón-Cedeño et al., 2020; Nakov, Da San Martino et al., 2021; Shaar, Hasanain et al., 2021) is a popular shared task for claim detection, check-worthiness, and prioritization tasks. However such shared tasks often have no submissions that employs HITL methodologies. Shared tasks for HITL approaches could encourage more solutions that can complement the limitations of model-only based approaches.

Evidence retrieval and veracity prediction Most work in HITL fact-checking caters to veracity prediction, and only a few consider evidence retrieval as a separate task. While there is a body of literature on HITL approaches in information retrieval (Chen & Jain, 2013; Demartini, 2015), we know of no work in that direction for fact-checking.

Shabani, Charlesworth, Sokhn, and Schuldt (2021) leverage HITL approaches for providing feedback about claim source, author, message, and spelling (SAMS). Annotators answer four yes/no questions about whether the article has a source, an author, a clear and unbiased message, and any spelling mistake. Furthermore, this work integrates the features provided by humans in a machine learning pipeline, which resulted in a 7.1% accuracy increase. However, the evaluation is performed on a small dataset with claims related to Covid-19. It is unclear if this approach would generalize outside of the domain. Moreover, further human effort can be reduced in this work by automating spell-check and grammar-check. SAMS could be quite limited in real life situations as most carefully crafted misinformation often looks like real news. Model generated fake news can successfully fool annotators (Zellers et al., 2020), and thus SAMS might also fail to flag such fake news.

Qu, Barbera et al. (2021) and Qu, Roitero, Mizzaro, Spina and Demartini (2021) provide an understanding of how human and machine confidence scores can be leveraged to build HITL approaches for fact-checking. They consider explicit self-reported annotator confidence and compute implicit confidence based on standard deviation among ten crowd workers. Model confidence is obtained from bootstrapping (Efron & Tibshirani, 1985) ten different versions of the model and then computing standard deviation over the scores returned by the soft-max layer. Their evaluation shows that explicit crowd and model confidence are poor indicators of accurate classification decisions. Although the crowd and the model make different mistakes, there is no clear signal that confidence is related to accuracy. However, they show that implicit crowd confidence can be a useful signal for identifying when

to engage experts to collect labels. A more recent study shows that a politically balanced crowd of ten is correlated with the average rating of three fact-checkers (Allen, Arechar, Pennycook, & Rand, 2020). Gold, Kovatchev, and Zesch (2019) also find that annotations by a crowd of ten correlate with the judgments of three annotators for textual entailment, which is utilized by veracity prediction models.

A series of studies show that the crowd workers can reliably identify misinformation (Roitero, Soprano and Fan et al., 2020; Roitero, Soprano and Portelli et al., 2020; Soprano et al., 2021). Furthermore, Roitero, Soprano and Portelli et al. (2020) show that crowd workers not only can identify false claims but also can retrieve proper evidence to justify their annotation. One weakness of this study is that it only asks users to provide one URL as evidence. However, in practice, fact-checking might need reasoning over multiple sources of information. Although these studies do not propose novel HITL solutions, they provide sufficient empirical evidence and insights about where crowd workers can be engaged reliably in the fact-checking pipeline.

Nguyen, Kharosekar, Lease et al. (2018) propose joint modeling of crowd annotations and machine learning to detect the veracity of textual claims. The key strength of the model is that it assumes all annotators can make mistakes, which is a possibility as fact-checking is a difficult task. Another strength is that this model allows users to import their knowledge into the system. Moreover, this HITL approach can collect on-demand stance labels from the crowd and incorporate them in veracity prediction. Empirical evaluation shows that this approach achieves strong predictive performance. A follow-up study provides an interactive HITL tool for fact-checking (Nguyen, Kharosekar, Krishnan et al., 2018).

Nguyen et al. (2020) propose a HITL system to minimize user effort and cost. Users validate algorithmic predictions but do so at a minimal cost by only validating the most-beneficial predictions for improving the system. This system provides a guided interaction to the users and incrementally gets better as users engage with it.

It is important to note that research on crowdsourcing veracity judgment is at an early stage. Different factors such as demographics, political leaning, criteria for determining the expertise of the assessors (Bhuiyan, Zhang, Sehat, & Mitra, 2020), cognitive factors (Kaufman, Haupt, & Dow, 2022), and even the rating scale (La Barbera, Roitero, Demartini, Mizzaro, & Spina, 2020) led to different levels of alignment with expert ratings. Bhuiyan et al. (2020) outline research directions for designing better crowd processes specific to different types of misinformation for the successful utilization of crowd workers.

Explaining veracity prediction HITL systems in fact-checking often use veracity explanations to correct model errors. As discussed earlier, Nguyen, Kharosekar, Krishnan et al. (2018) provides an interpretable model that allows users to impart their knowledge when the model is wrong. Empirical evaluation shows that users could impart their knowledge into the system. Similarly, Zhang, Rudra and Anand (2021) propose a method that collects user feedback from explanations. Note that this method explains veracity prediction outcomes based on the evidence retrieved and their stance. Users provide feedback in terms of stance and relevance of the retrieved evidence. The proposed approach employs lifelong learning which enables the system to improve over time. Currently there is no empirical evaluation of this system to identify the effectiveness of this approach.

Although natural language generation models are getting increasingly better (Radford et al., 2019), generating abstractive fact-checking explanations is still in its infancy (Kotonya & Toni, 2020b). HITL methods could be leveraged to write reports justifying fact-checking explanations.

Limitations After reviewing existing HITL approaches across different fact-checking tasks, we also list out several limitations as follow. First, some HITL approaches adopt several interpretable models to integrate human input, but the resulting models do not perform as well as the state-of-the-art deep learning models (Nguyen, Kharosekar, Krishnan et al., 2018; Nguyen, Kharosekar, Lease et al., 2018).

Farinneya et al. (2021) apply HITL approaches to scale up rumor detection from a limited amount of annotated data. Although it performs well to generalize the algorithm for a new topic in a few-shot manner, one of the weaknesses is that data from other domains or topics causes a high variance in model performance. Consequently, in-domain model performance might degrade when out-of-domain data is introduced in model training. This issue may hinder the model's generalizability in practice, especially where a clear demarcation between topic domains may not be possible.

More importantly, there is a lack of empirical studies on how to apply HITL approaches of fact-checking for practical adoption. Although HITL approaches provide a mechanism to engage human in the process of modeling development, several human factors, such as usability, intelligibility, and trust, become important to consider when applying this method in the real-world use case. Fact-checking is a time-sensitive task and requires expertise to process complex information over multiple sources (Graves, 2017). Fact-checkers and policy makers are often skeptical about any automated or semi-automated solutions as this type of research requires human creativity and expertise (Arnold, 2020; Micallef et al., 2022). Therefore, more empirical evidence needs to be found to assess the effectiveness of applying different HITL approaches to automated fact-checking.

6. Existing tools for fact-checking

In the previous section, we reviewed the details of current NLP technologies for fact-checking. Subsequently, we extend our review of automated fact-checking to the HCI literature and discuss existing practices of applying fact-checking into real-world tools that assist human fact-checkers. In brief, there is a lack of holistic review of fact-checking tools from a human-centered perspective. Additionally, we found that the articulation of work between human labor and AI tools is still opaque in this field.

Research questions include but are not limited to: (1) how can NLP tools facilitate human work in different fact-checking tasks? (2) how can we incorporate user needs and leverage human expertise to inform the design of automated fact-checking?

In this section, we examine current real-world tools that apply NLP technologies in different stages of fact-checking and clarify the main use cases of these tools. We argue that more research concerning human factors for building automated fact-checking, such as user research, human-centered design, and usability studies, should be conducted to improve the practical adoption of automated fact-checking. These studies help us identify the design space of applying explainable and HITL approaches for real-world NLP technologies.

6.1. Claim detection and prioritization

The first step in claim detection is sourcing content to possibly check. On end-to-end encrypted platforms, such as WhatsApp, Telegram, and Signal, crowdsourcing-based tip-lines play a vital role in identifying suspicious content that is not otherwise accessible (Kazemi, Garimella et al., 2021). As another example, *Check* from Meedan,[7] a tip-line service tool, also helps fact-checkers monitor fake news for in-house social media. User flagging of suspect content on social media platforms such as Facebook is also a valuable signal for identifying such content, and crowdsourcing initiatives like Twitter's BirdWatch can further help triage and prioritize claims for further investigation.

In the stage of finding and choosing claims to check, fact-checkers assess the fact-checking related quality of a claim and decide whether to fact-check it (Graves, 2017; Micallef et al., 2022). NLP models in claim detection, claim matching, and check-worthiness are useful to assist the above decision-making process. However, integrating them into real-world tools that help fact-checkers prioritize what to check requires more personalized effort. Graves (2018a) points out that it is important to design the aforementioned models to cater to fact-checker organizational interests, stakeholder needs, and changing news trends.

As one of the fact-checking qualities of a claim, checkability can be objectively analyzed by whether a claim contains one or more purported facts that can be verified (Section 3.1). Fact-checkers find it useful to apply models that identify checkable claims to their existing workflow because the model helps them filter irrelevant content and claims that are uncheckable when they are choosing claims to check (Arnold, 2020). ClaimBuster, one of the well-known claim detection tools, is built to find checkable claims from a large scale of text content (Hassan, Arslan et al., 2017). Claim detection can also be integrated into speech recognition tools to spot claims from live speech (Adair, 2020).

Additionally, if a claim has already been fact-checked, fact-checkers can skip it and prioritize claims that have not been checked. As a relatively new NLP task, claim matching has been integrated into some current off-the-shelf search engines or fact-checking tools to help fact-checkers find previously fact-checked claims. For example, Google Fact Check Explorer,[8] can retrieve previously fact-checked claims by matching similar fact-check content to user input queries. Similarly, with Meedan's *Check* if users send a tip with fake news that has been previously fact-checked, the tool further helps fact-checkers retrieve the previous fact-check and send it to users.

Whether or not to fact-check a claim depends on an organization's goals and interests. Tools built for claim detection need to take such interests into account. For example, Full Fact developed a claim detection system that classifies claims into different categories, such as quantity, predictions, correlation or causation, personal experience, and laws or rules of operations (Konstantinovskiy et al., 2021). The claim categories are designed by their fact-checkers to cater to their needs of fact-checking UK political news in a live fact-checking situation. Identifying certain claims, such as quantity, correlation or causation, might be particularly useful for fact-checkers to evaluate the credibility of politician statements and claims. The system also helps tailor fact-checkers' downstream tasks, such as fact-check assignments and automated verification for statistical claims (Nakov, Corney et al., 2021).

Fact-checkers also use social media monitoring tools to find claims to check, such as CrowdTangle, TweetDeck, and Facebook's (unnamed) fact-checking tool, but those tools are not very effective to detect checkable claims. Some fact-checkers reported that only roughly 30% of claims flagged by Facebook's fact-checking tool were actually checkable (Arnold, 2020). A low hanging fruit is to integrate claim detection models into these social media monitoring tools so that it is easier for fact-checkers to identify claims that are both viral and checkable. Additionally, these tools should enable fact-checkers to locate certain figures, institutions, or agencies according to their fact-checking interests and stakeholder needs so that these tools can better identify and prioritize truly check-worthy claims. An important question in implementing those systems is how to measure the virality of a claim and its change over time.

It would also be useful to integrate veracity prediction into previous fact-checking tools because fact-checkers may pay the most attention to claims[9] that are suspect and uncertain (since obviously true or false claims likely do not require a fact-check). However, information or data points that are used to give such predictions should also be provided to fact-checkers. If sources, evidence, propagation patterns, or other contextual information that models use to predict claim veracity can be explained clearly for fact-checkers, they can also triage these indicators to prioritize claims more holistically.

[7] https://meedan.com/check.
[8] https://toolbox.google.com/factcheck/explorer.
[9] https://www.factcheck.org/our-process/.

6.2. Tools for evidence retrieval

After finding and prioritizing which claim to check, fact-checkers investigate claims following three main activities: (1) decomposing claims, (2) finding evidence, and (3) tracing the provenance of claims and their spread. Note that these three activities are intertwined with each other by using different information-seeking tools in the fact-checking process. Fact-checkers search for evidence by decomposing claims into sub questions. Evidence found while investigating a claim may further modify or add to the sub-questions (Singh et al., 2021). By iteratively investigating claims via online search, fact-checkers reconstruct the formation and the spread of a claim to assess its truth (Graves, 2017). In this section, we discuss the utility of existing information-seeking tools, including off-the-shelf search engines and domain-specific databases, that assist fact-checkers in each activity.

Claim decomposition is not a specific activity that qualitative researchers have reported or analyzed in their fact-checking studies, but we can find more details from where fact-checking organizations describe their methodology.[10] and how fact-checkers approach complex claims in their fact-checks[11] Claim decomposition refers to how fact-checkers interpret ambiguous terms of a claim and set the fact-checking boundaries to find evidence. Decomposing claims effectively requires sensitive curiosity and news judgments for fact-checkers that are cultivated through years of practice. Unfortunately, we are not aware of any existing tools that facilitate this process.

Traditional methodology to decompose claims is to ask sub-questions. Recent NLP studies simulate this process by formulating it as a question-answering task (Chen et al., 2022; Fan et al., 2020). Researchers extract justifications from existing fact-checks and crowdsource sub-questions to decompose the claim. For automated-fact-checking, this NLP task might be very beneficial to improve the performance of evidence retrieval by auto-decomposing claims into smaller checkable queries (Chen et al., 2022). Although it is difficult for NLP to match the abilities of professional fact-checkers, it might help scale up the traditional, human fact-checking process. It could also help the public, new fact-checkers, or journalists to more effectively investigate complex claims and search for evidence.

How fact-checkers find evidence is usually a domain-specific reporting process, contacting experts or looking for specific documents from reliable sources (Graves, 2017; Micallef et al., 2022). Instead of conducting random searches online, most fact-checkers include a list of reliable sources in which to look for evidence. Tools that are designed for searching domain datasets can also help fact-checkers to find evidence. For example, Li, Fang, Lou, Li, and Zhang (2021) built an analytical search engine for retrieving the COVID-19 news data and summarizing it in an easy to digest, tabular format. The system can decompose analytical queries into structured entities and extract quantitative facts from news data. Furthermore, if evidence retrieval is accurate enough for in-domain datasets, the system can take a leap further to auto-verify domain-related claims. We provide more detailed use cases of veracity prediction in Section 6.3.

Fact-checkers mainly use off-the-shelf search engines, such as Google, Bing, etc., to trace a claim's origin from publicly accessible documents (Arnold, 2020; Beers, Haughey, Melinda, Arif, & Starbird, 2020). Other digital datasets, such as *LexisNexis* and *InternetArchive*, are also useful for fact-checkers to trace claim origin. To capture the formation and change of a claim, search engines should not only filter unrelated content, but also retrieve both topically and evidentially relevant content. Hasanain and Elsayed (2021) report that most topically relevant pages retrieved from Google do not contain evidential information, such as statistics, quotes, entities, or other types of facts. Additionally, most built-in search engines in social media platforms, such as Twitter and Facebook, only filter "spreadable" content not "credible" content (Alsmadi, Alazzam, & AlRamahi, 2021).

Furthermore, these off-the-shelf search engines do not support multilingual search, so it is difficult for fact-checkers to trace claims if they are translated from other languages (Graves, 2017; Nakov, Corney et al., 2021). NLP researchers have started to use multilingual embedding models to represent claim-related text in different languages and match existing fact-checks (Kazemi, Gaffney, Garimella and Hale, 2021). This work not only helps fact-checkers find previously fact-checked claims more easily from other languages, but also to examine how the claim is transformed and reshaped by the media in different languages and socio-political contexts.

6.3. Domain-specific tools for claim verification

As discussed in Sections 3.2 and 3.3, most verification prediction models are grounded on the collected evidence and the claim. To build an end-to-end claim verification system, NLP developers need to construct domain-specific datasets incorporating both claims and evidence. Different from complex claims that contain multiple arguments and require decomposition, claims that have simple linguistic structure with purported evidence or contain statistical facts can be automatically verified (Nakov, Corney et al., 2021).

Karagiannis, Saeed, Papotti, and Trummer (2020) built CoronaCheck, a search engine that can directly verify Covid-19 related statistical claims by retrieving official data curated by experts (Dong, Du, & Gardner, 2020). Full Fact (The Poynter Institute, 2021) also took a similar approach to verify statistical macroeconomic claims by retrieving evidence from UK parliamentary reports and national statistics. Additionally, Wadden et al. (2020) built a scientific claim verification pipeline by using abstracts that contain evidence to verify a given scientific claim.

[10] https://leadstories.com/how-we-work.html.

[11] https://www.factcheck.org/2021/10/oecd-data-conflict-with-bidens-educational-attainment-claim/ In this fact-check, fact-checkers decompose what President Biden mean by "advanced economies" and "young people". The approach of defining these two terms directly influence their fact-checking results.

However, pitfalls still exist if fact-checkers use these domain-specific verification tools in practice. For example, the CoronaCheck tool cannot check the claim "The Delta variant causes more death than the Alpha variant" simply because the database does not contain fine-grained death statistics for Covid variants. Additionally, checking a statistical or scientific claim might only be a part of the process of checking a more complex claim, which requires fact-checkers to contextualize the veracity of previous statistical or scientific checks. In general, domain-specific tools are clearly valuable to use when available, though in practice they are often incomplete and insufficient on their own to check complex claims.

7. Discussion

In this review, we have (1) horizontally outlined the research of applying NLP technologies for fact-checking from the beginning of task formulation to the end of tools adoption; as well as (2) vertically discussing the capabilities and limitations of NLP for each step of a fact-checking pipeline. We perceive a lack of research that bridges both to assist fact-checkers. Explainable and HITL approaches leverage both human and computational intelligence from a human-centered perspective, but there is a need to provide actionable guides to utilize both methods for designing useful fact-checking tools. In this section, we propose several research directions to explore the design space of applying NLP technologies to assist fact-checkers.

7.1. Distributing work between human and AI for mixed-initiative fact-checking

The practice of fact-checking has already become a type of complex and distributed computer-mediated work (Graves, 2018a). Although Graves (2017) breaks down a traditional journalist fact-checking pipeline into five steps, the real situation of fact-checking a claim is more complicated (Juneja & Mitra, 2022). Various AI tools are adopted dynamically and diversely by fact-checkers to complete different fact-checking tasks (Arnold, 2020; Beers et al., 2020; Micallef et al., 2022).

Researchers and practitioners increasingly believe that future fact-checking should be a mixed-initiative practice in which humans perform specific tasks while machines take over others (Lease, 2020; Nakov, Corney et al., 2021; Nguyen, Kharosekar, Krishnan et al., 2018). To embed such hybrid and dynamic human–machine collaborations into existing fact-checking workflow, the task arrangement between human and AI need to be articulated clearly by understanding the expected outcomes and criteria for each. Furthermore, designing a mixed-initiative tool for different fact-checking tasks requires a more fine-grained level of task definition for human and AI (Lease, 2018, 2020). In Section 5.3, we discuss several studies highlighting the role of humans in the fact-checking workflow, e.g., (a) human experts select check-worthy claims from claim detection tools (Hassan, Zhang et al., 2017) and deliver them to fact-checkers, (b) ask crowd workers to judge reliable claims sources (Shabani et al., 2021), or (c) flag potential misinformation (Roitero, Soprano and Portelli et al., 2020) to improve veracity prediction. All of the above human activities are examples of micro-tasks within a mixed-initiative fact-checking process.

Prior work in crowdsourcing has shown that it is possible to effectively break down the academic research process and utilize crowd workers to partake in smaller research tasks (Vaish, Davis, & Bernstein, 2015; Vaish et al., 2017). Given this evidence, we can also break down sub-tasks of a traditional fact-checking process into more fine-grained tasks. Therefore, key research questions include: (a) How can we design these micro-tasks to facilitate each sub-task of fact-checking, and (b) What are the appropriate roles for human and AI in different micro-tasks?

To effectively orchestrate human and AI work, researchers need to understand the respective roles of human and AI, and how they will interact with one another, because it will directly affect whether humans decide to take AI advice (Cimolino & Graham, 2022). Usually, if AI aims to assist high-stake decision-making tasks, such as recidivism prediction (Veale, Van Kleek, & Binns, 2018) and medical treatments (Cai, Winter, Steiner, Wilcox, & Terry, 2019), considerations of risk and trust will be important factors for people to adopt such AI assistants (Lai, Chen, Liao, Smith-Renner, & Tan, 2021). In the context of fact-checking, if AI directly predicts the verdict of a claim, fact-checkers may be naturally skeptical about how the AI makes such a prediction (Arnold, 2020). On the other hand, if AI only helps to filter claims that are uncheckable, such as opinions and personal experience, fact-checkers may be more willing to use such automation with less concern about how AI achieves it. Deciding whether a claim is true or false is a high-stake decision-making task for fact-checkers, while filtering uncheckable claims is a less important but tedious task that fact-checkers want automation to help with. Therefore, the extent of human acceptance of AI varies according to how humans assess the task assigned to AI, resulting in different human factors, such as trust, transparency, and fairness. Researchers need to specify or decompose these human factors into different key variables that can be measured during the model development process. Given a deep understanding of the task relationship between human and AI, researchers can then ask further research questions on how to apply an explainable approach, or employ a HITL system vs. automated solutions, to conduct fact-checking. Here we list out several specific research topics that contain mixed-initiative tasks, including: (a) assessing claim difficulty leveraging crowd workers, (b) breaking down a claim into a multi-hop reasoning task and engaging the crowd to find information relevant to the sub-claims, and (c) designing micro-tasks to parse a large number of documents retrieved by web search to identify sources that contain the evidence needed for veracity prediction.

7.2. Human-centered evaluation of NLP technology for fact-checkers

We begin this section by proposing key metrics from human factors for evaluating systems (i.e., what to measure and how to measure them): accuracy, time, model understanding, and trust (Section 7.2.1). Following this, we further propose a template for an experimental protocol for human-centered evaluations in fact-checking (Section 7.2.2).

7.2.1. Metrics

Accuracy Most fact-checking user studies assume task accuracy as the primary user goal (Mohseni et al., 2021; Nguyen, Kharosekar, Krishnan et al., 2018). Whereas non-expert users (i.e., social media users or other form of content consumers) might be most interested in the veracity outcome along with justification, fact-checkers often want to use automation and manual effort interchangeably in their workflow (Arnold, 2020; Nakov, Corney et al., 2021). Thus, we need a more fine-grained approach toward measuring accuracy beyond the final veracity prediction accuracy. For fine-grained accuracy evaluation, it is also crucial to capture fact-checker accuracy, particularly for the sub-tasks for which they use the fact-checking tool.

With the assumption that "ground truth" exists for all of the sub-tasks in the fact-checking pipeline, accuracy can be computed by comparing user answers with the ground truth. Note that measuring sub-task level accuracy is trickier than end-to-end fact-checking accuracy. Sub-task level accuracy can be captured by conducting separate experiments for each sub-task. Suppose the point of interest is to understand user performance for detecting *claim-checkworthiness*. In that case, we will need to collect additional data specific to the *claim-checkworthiness* task.

In some cases, it is possible to merge multiple sub-tasks for evaluation purposes. For example, Miranda et al. (2019) evaluate the effectiveness of their tool with journalists by capturing the following two key variables: (a) the relevance of retrieved evidence, and (b) the accuracy of the predicted stance. This method provides essential insight into evidence retrieval, stance detection, and the final fact-checking task. Depending on the tool, the exact detail of this metric will require specific changes according to tool affordances.

Note that both time and accuracy measures need to control for claim properties. For example, if a claim has been previously fact-checked, it would take less time to fact-check such claims. On the other hand, a new claim that is more difficult to assess would require more time.

Model understanding Fact-checkers want to understand the tools they use. For example, Arnold (2020) pointed out that fact-checkers expressed a need for understanding CrowdTangle's algorithm for detecting viral content on various social media platforms. Similarly, Nakov, Corney et al. (2021) observed a need for increased system transparency in the fact-checking tools used by different organizations. Lease (2018) argues that transparency is equally important for non-expert users to understand the underlying system and make an informed judgment. Although this is not a key variable related to user performance, it is important for practical adoption.

To measure understanding, users could be asked to self-report their level of understanding on a Likert-scale. However, simply asking participants if they understand the algorithm is not a sufficient metric. For example, it does not indicate whether participants will be able to simulate tool behavior (Hase & Bansal, 2020). We suggest the following steps for measuring model understanding based on prior work (Cheng et al., 2019).

1. **Decision Prediction.** To capture users' holistic understanding of a tool, users could be provided claims and asked the following: "What label would the tool assign to this claim?"
2. **Alternative Prediction.** Capturing how changes in the input influence the output can also measure understanding, e.g., by asking users how the tool would assign a label to a claim when input parts are changed. Imagine a tool that showed the users the evidence it has considered to arrive at a veracity conclusion. Now, if certain pieces of evidence were swapped, how would that be reflected in the model prediction?

Trust For practical adoption, trust in a fact-checking tool is crucial across all user groups. While model understanding is often positively correlated with trust, understanding alone may not suffice to establish trust. In this domain, fact-checkers and journalists may have less trust in algorithmic tools (Arnold, 2020). On the other hand, there is also the risk of over-trust, or users blindly following model predictions (Mohseni et al., 2021; Nguyen, Kharosekar, Krishnan et al., 2018). To maximize the tool effectiveness, we would want users to neither dismiss all model predictions out of hand (complete skepticism) nor blindly follow all model predictions (complete faith). Instead, it is important to calibrate user trust for the most effective tool usage. We suggest measuring a notion of *calibrated trust* (Lee & See, 2004): how often users abide by correct model decisions and override erroneous model decisions.

To measure calibrated trust, we imagine a confusion matrix shown in Fig. 2. The rows denote correct vs. incorrect model predictions while the columns denote correct vs. incorrect user predictions. A user who blindly followed all model predictions would have their behavior entirely captured by the main (primary) diagonal, whereas a user who skeptically rejected all model predictions would have their behavior captured entirely in the secondary diagonal. The ideal user's behavior would be entirely captured in the first column: accepting all correct model predictions and rejecting all incorrect model predictions. To promote effective human-AI teaming, AI tools should assist their human users in developing strong calibrated trust to appropriately trust and distrust model predictions as each case merits.

Beyond calibrated trust, one could also measure quantitative trust by adopting methodologies from the human–machine trust literature (Lee & Moray, 1992). For example, Cheng et al. (2019) adapted prior work into a 7-point Likert scale. A similar scale can be reused for evaluating trust in a fact-checking tool. For example, we can create five different Likert-scales to measure the agreement (or disagreement) of users with the following statements:

- I understand the fact-checking tool.
- I can predict how the tool will behave.
- I have faith that the tool would be able to cope with the different fact-checking task.
- I trust the decisions made by the tool.
- I can count on the tool to provide reliable fact-checking decisions.

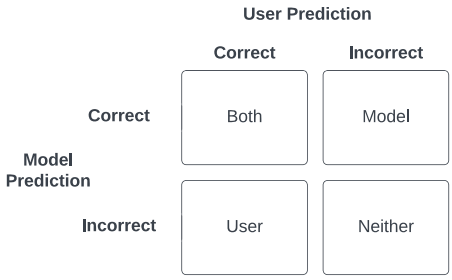

Fig. 2. Confusion Matrix for User Predictions vs. Model Predictions with respect to ground truth (gold). We assume model predictions are provided to the user, who then decides whether to accept or reject the model prediction. The top-left quadrant (*Both*) covers cases where users correctly follow model predictions. The top-right quadrant (*Model*) denotes the cases where the model is correct but users mistakenly reject the model decisions. The bottom-left quadrant *User* denotes the cases where users correctly reject erroneous model predictions. The bottom-right quadrant *Neither* denotes the cases where users incorrectly accept erroneous model predictions. Quantifying user vs. model predictions in this manner enables measurement of *calibrated trust*: how often users abide by correct model decisions and override erroneous model decisions.

Additional factors Individual differences among users might result in substantial variation in experimental outcomes. For example, varying technical literacy (Cheng et al., 2019), any prior knowledge about the claims, and users' political leaning (Thornhill, Meeus, Peperkamp, & Berendt, 2019) might influence user performance on the task while using fact-checking tools. Thus it is valuable to capture these factors in study design. For example:

1. **Technical Literacy:** Users' familiarity with popular technology tools (e.g., recommendation engines, spam detectors) and their programming experience (Cheng et al., 2019) as well as familiarity with existing fact-checking tools.
2. **Media Literacy:** Users' familiarity with (1) the fact-checking process, and (2) fact-checks from popular organizations such as PolitiFact and FactCheck.Org.
3. **Demographics:** Users' education level, gender, age, and political leaning.

Quantitative measures alone are not sufficient as they do not capture certain nuances about how effectively a tool integrates into a fact-checker's workflow. For example, even if users understand and trust the working principle of a tool, it is unclear *why* they do so. Hence, users might be asked a few open-ended questions at the end of the study to gather qualitative insights. Such questions could include:

1. Describe your understanding of the tool. Do any specific aspects of its design seem to assist or detract from your understanding of how it works?
2. Why do you trust or not trust the tool?
3. Would you use this tool beyond this study, and if so, in what capacity?

7.2.2. Experimental protocol

One strategy to capture the aforementioned metrics is to design a mixed-methods study. Here we outline the template for such a study. Imagine the goal were to measure the user performance for fact-checking using a new tool (let us call it *tool A*) compared to an existing tool (*tool B*). Fact-checking tasks in the real world might be influenced by user priors about the claims being checked. Thus, a *within-subject* study protocol may be more appropriate to account for such priors (Shi, Bhattacharya, Das, Lease, & Gwizdka, 2022).

1. **Pre-task**: Users would first be asked to fact-check a set of claims. To do so, first a user would be asked to leverage a pre-existing *tool B* at this stage. Tool B can be replaced with different baselines, depending on the particular use case, ranging from simple web-search by non-expert users to proprietary tools used by fact-checkers and journalists. Users would be asked to think aloud at this stage.
2. **Learning**: At this stage users would familiarize themselves with the new tool (*tool A*). Users would need to fact-check a different set of claims from the first one. Ground truth would also accessible to the user to form a prior about what kind of mistakes a tool might make. Claims here would be selected at random to reflect tool capabilities. Moreover, tool performance metrics would be given to the users as additional information. Users would be encouraged to ask questions about the tool at this stage.
3. **Prediction:** Users would now be asked to fact-check the same claims from step-1 above but this time they are asked to leverage the *tool A*. Users would be asked to think out loud through this stage. Users could simply guess the answers and achieve a high accuracy score. Thus, claims selected for stages (1) & (3) would be a balanced set of claims with an equal distribution of true positive, true negative, false positive, and false negative samples. This idea is adopted from prior work (Hase & Bansal, 2020).
4. **Post-task survey**: Users would now be asked to take a small survey for capturing trust, understanding, technical literacy, media literacy, and demographic information.

5. **Post-task interview**: Upon completion of these steps, users would be interviewed with open-ended questions to gather insights about their understanding and trust in the system.

The measures and study protocol could be useful in the context of evaluating any new fact-checking system compared to an existing system or practices. Specifics might vary depending on the target user group and the tool's intended purpose. Above we use the whole fact-checking pipeline to illustrate our experimental protocol. However this technique can be applied to other sub-tasks of automated fact checking, granted that we have the ground truth of the outcome for that sub-task. For example, let us assume a new claim detection tool has been proposed that takes claims from a tip-line (Kazemi, Garimella et al., 2021). Currently, fact-checkers use an existing claim-matching algorithm to filter out the already fact-checked claim. Now, if we replace *tool B* above with the existing claim-matching algorithm and *tool A* with the proposed claim detection tool, we can utilize the protocol mentioned above. In conclusion, one could evaluate how users perform for claim detection tasks using the new tool compared to the existing ones in terms of their accuracy, time, understanding, and trust.

While we have proposed an ideal, extensive version of an evaluation protocol for evaluating new fact-checking tools, note that the actual protocol used in practice could be tailored according to the time required from the participants and the cost of conducting the experiment.

8. Conclusion

This review highlights the practices and development of the state-of-the-art in using NLP for automated fact-checking, emphasizing both the advances and the limitations of existing task formulation, dataset construction, and modeling approaches. We partially discuss existing practices of applying these NLP tasks into real-world tools that assist human fact-checkers. In recent years we have seen significant progress in automated fact-checking using NLP. A broad range of tasks, datasets, and modeling approaches have been introduced in different parts of the fact-checking pipeline. Moreover, with recent developments in transformers and large language models, the model accuracy has improved across tasks. However, even state-of-the-art models on existing benchmarks – such as FEVER and *CLEF!* – may not yet be ready for practical adoption and deployment.

To address these limitations, we advocate development of hybrid, HITL systems for fact-checking. As a starting point, we may wish to reorient the goals of existing NLP tasks from full automation toward decision support. In contrast with fully-automated systems, hybrid systems instead involve humans-in-the-loop and facilitate human-AI teaming (Bansal, Nushi, Kamar, Horvitz and Weld, 2021; Bansal et al., 2019; Bansal, Wu et al., 2021). Such use of hybrid systems can help (a) scale-up human decision making; (b) augment machine learning capabilities with human accuracy; and (c) mitigate unintended consequences from machine errors. Additionally, we need new benchmarks and evaluation practices that can measure how automated and hybrid systems can improve downstream human accuracy (Fan et al., 2020; Smeros, Castillo, & Aberer, 2021) and efficiency in fact-checking.

CRediT authorship contribution statement

Anubrata Das: Conceptualization, Investigation, Visualization, Writing – original draft, Writing – review & editing. **Houjiang Liu:** Investigation, Writing – original draft. **Venelin Kovatchev:** Supervision, Writing – review & editing. **Matthew Lease:** Funding acquisition, Supervision, Project administration, Writing – review & editing.

Data availability

No data was used for the research described in the article.

Acknowledgments

This research was supported in part by the Knight Foundation, USA, the Micron Foundation, USA, Wipro, and by Good Systems,[12] a UT Austin Grand Challenge to develop responsible AI technologies. The statements made herein are solely the opinions of the authors and do not reflect the views of the sponsoring agencies.

References

Adair, B. (2020). Squash report card: Improvements during state of the union ... and how humans will make our AI smarter - Duke reporters' lab. URL: https://reporterslab.org/squash-report-card-improvements-during-state-of-the-union-and-how-humans-will-make-our-ai-smarter/.

Adair, B., & Stencel, M. (2020). A lesson in automated journalism: Bring back the humans | Nieman journalism lab. https://www.niemanlab.org/2020/07/a-lesson-in-automated-journalism-bring-back-the-humans/. (Accessed on 02/08/2022).

Ahmadi, N., Lee, J., Papotti, P., & Saeed, M. (2019). Explainable fact checking with probabilistic answer set programming. In *Conference on truth and trust online*.

Ahmadi, N., Truong, T.-T.-D., Dao, L.-H.-M., Ortona, S., & Papotti, P. (2020). RuleHub: A public corpus of rules for knowledge graphs. *Journal of Data and Information Quality (JDIQ)*, *12*(4), 1–22.

Ahsan, M., Kumari, M., & Sharma, T. (2019). Detection of context-varying rumors on Twitter through deep learning. *International Journal of Advanced Science and Technology*, *128*, 45–58.

[12] http://goodsystems.utexas.edu/.

Aker, A., Derczynski, L., & Bontcheva, K. (2017). Simple open stance classification for rumour analysis. In *Proceedings of the international conference recent advances in natural language processing, RANLP 2017* (pp. 31–39).

Alam, F., Dalvi, F., Shaar, S., Durrani, N., Mubarak, H., Nikolov, A., et al. (2021). Fighting the COVID-19 infodemic in social media: A holistic perspective and a call to arms. In *ICWSM*.

Alam, F., Shaar, S., Dalvi, F., Sajjad, H., Nikolov, A., Mubarak, H., et al. (2021). Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In *Findings of the association for computational linguistics: EMNLP 2021* (pp. 611–649). Punta Cana, Dominican Republic: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.findings-emnlp.56, URL: https://aclanthology.org/2021.findings-emnlp.56.

Alhindi, T., Petridis, S., & Muresan, S. (2018). Where is your evidence: Improving fact-checking by justification modeling.

Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2020). Scaling up fact-checking using the wisdom of crowds. Preprint at http://dx.doi.org/10.31234/osf.io/9qdza.

Alsmadi, I., Alazzam, I., & AlRamahi, M. A. (2021). An ontological analysis of misinformation in online social networks. http://dx.doi.org/10.48550/arxiv.2102.11362, URL: http://arxiv.org/abs/2102.11362. arXiv:2102.11362.

Aly, R., Guo, Z., Schlichtkrull, M. S., Thorne, J., Vlachos, A., Christodoulopoulos, C., et al. (2021). FEVEROUS: Fact extraction and verification over unstructured and structured information. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 1)*.

Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. *AI Magazine*, *35*(4), 105–120.

Arnold, P. (2020). The challenges of online fact checking: how technology can (and can't) help - full fact. https://fullfact.org/blog/2020/dec/the-challenges-of-online-fact-checking-how-technology-can-and-cant-help/. (Accessed on 09/11/2021).

Atanasova, P., Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Zaghouani, W., Kyuchukov, S., et al. (2018). Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. Task 1: Check-worthiness. arXiv preprint arXiv:1808.05542.

Atanasova, P., Nakov, P., Karadzhov, G., Mohtarami, M., & Da San Martino, G. (2019). Overview of the CLEF-2019 CheckThat! lab: Automatic identification and verification of claims. Task 1: Check-worthiness.. In *CLEF (working notes), Vol. 2380*.

Atanasova, P., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Karadzhov, G., Mihaylova, T., et al. (2019). Automatic fact-checking using context and discourse information. *Journal of Data and Information Quality (JDIQ)*, *11*(3), 1–27.

Atanasova, P., Simonsen, J. G., Lioma, C., & Augenstein, I. (2020). Generating fact checking explanations. In *ACL*.

Atanasova, P., Wright, D., & Augenstein, I. (2020). Generating label cohesive and well-formed adversarial claims. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 3168–3177).

Augenstein, I., Lioma, C., Wang, D., Lima, L. C., Hansen, C., Hansen, C., et al. (2019). MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *EMNLP*.

Bansal, G., Nushi, B., Kamar, E., Horvitz, E., & Weld, D. S. (2021). Is the most accurate AI the best teammate? Optimizing AI for teamwork. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 35* (pp. 11405–11414).

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing, Vol. 7* (pp. 2–11).

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., et al. (2021). Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–16).

Barnabò, G., Siciliano, F., Castillo, C., Leonardi, S., Nakov, P., Da San Martino, G., et al. (2022). FbMultiLingMisinfo: Challenging large-scale multilingual benchmark for misinformation detection. In *2022 international joint conference on neural networks* (pp. 1–8). IEEE.

Barrón-Cedeño, A., Elsayed, T., Nakov, P., Da San Martino, G., Hasanain, M., Suwaileh, R., et al. (2020). Overview of CheckThat! 2020: Automatic identification and verification of claims in social media. In *International conference of the cross-language evaluation forum for European languages* (pp. 215–236). Springer.

Beers, A., Haughey, Melinda, M., Arif, A., & Starbird, K. (2020). Examining the digital toolsets of journalists reporting on disinformation. In *Proceedings of computation + journalism 2020 (C+J '20)* (p. 5). New York, NY, USA: ACM, URL: https://cpb-us-w2.wpmucdn.com/express.northeastern.edu/dist/d/53/files/2020/02/CJ_2020_paper_50.pdf.

Bendersky, M., Metzler, D., & Croft, W. B. (2012). Effective query formulation with multiple information sources. In *Proceedings of the fifth ACM international conference on web search and data mining* (pp. 443–452).

Bhuiyan, M. M., Zhang, A. X., Sehat, C. M., & Mitra, T. (2020). Investigating differences in crowdsourced news credibility assessment: Raters, tasks, and expert criteria. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1–26.

Bibal, A., Cardon, R., Alfter, D., Wilkens, R., Wang, X., François, T., et al. (2022). Is attention explanation? An introduction to the debate. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 3889–3900). Dublin, Ireland: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.acl-long.269, URL: https://aclanthology.org/2022.acl-long.269.

Borel, B. (2016). *The Chicago guide to fact-checking*. University of Chicago Press.

Bouziane, M., Perrin, H., Cluzeau, A., Mardas, J., & Sadeq, A. (2020). Team buster. ai at CheckThat! 2020 insights and recommendations to improve fact-checking. In *CLEF (working notes)*.

Brand, E., Roitero, K., Soprano, M., Rahimi, A., & Demartini, G. (2018). A neural model to jointly predict and explain truthfulness of statements. *ACM Journal of Data and Information Quality (JDIQ)*.

Cai, C. J., Winter, S., Steiner, D., Wilcox, L., & Terry, M. (2019). "Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–24.

Chen, E., & Jain, A. (2013). Improving Twitter search with real-time human computation. *Engineering Blog*, *8*(1), 2013.

Chen, S., Khashabi, D., Yin, W., Callison-Burch, C., & Roth, D. (2019). Seeing things from a different angle:Discovering diverse perspectives about claims. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 542–557). Minneapolis, Minnesota: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N19-1053, URL: https://aclanthology.org/N19-1053.

Chen, J., Sriram, A., Choi, E., & Durrett, G. (2022). Generating literal and implied subquestions to fact-check complex claims. URL: http://arxiv.org/abs/2205.06938. arXiv:2205.06938.

Chen, W., Wang, H., Chen, J., Zhang, Y., Wang, H., Li, S., et al. (2019). TabFact: A large-scale dataset for table-based fact verification. In *International conference on learning representations*.

Cheng, H.-F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., et al. (2019). Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1–12).

Cimolino, G., & Graham, T. N. (2022). Two heads are better than one: A dimension space for unifying human and artificial intelligence in shared control. In *CHI conference on human factors in computing systems*. New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3491102.3517610.

Cinelli, M., Pelicon, A., Mozetič, I., Quattrociocchi, W., Novak, P. K., & Zollo, F. (2021a). Dynamics of online hate and misinformation. *Scientific Reports*, *11*(1), 1–12.

Cinelli, M., Pelicon, A., Mozetič, I., Quattrociocchi, W., Novak, P. K., & Zollo, F. (2021b). Online hate: Behavioural dynamics and relationship with misinformation. arXiv preprint arXiv:2105.14005.

Clarke, C. L., Rizvi, S., Smucker, M. D., Maistro, M., & Zuccon, G. (2020). Overview of the TREC 2020 health misinformation track. In *TREC*.
Da San Martino, G., Cresci, S., Barrón-Cedeño, A., Yu, S., Pietro, R. D., & Nakov, P. (2020). A survey on computational propaganda detection. In *IJCAI*.
Dagan, I., Dolan, B., Magnini, B., & Roth, D. (2010). Recognizing textual entailment: Rational, evaluation and approaches–erratum. *Natural Language Engineering*, *16*(1), 105.
Das, A., Gupta, C., Kovatchev, V., Lease, M., & Li, J. J. (2022). ProtoTEx: Explaining model decisions with prototype tensors. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 2986–2997).
Demartini, G. (2015). Hybrid human–machine information systems: Challenges and opportunities. *Computer Networks*, *90*, 5–13.
Demartini, G., Difallah, D. E., & Cudré-Mauroux, P. (2012). Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st international conference on world wide web* (pp. 469–478).
Demartini, G., Mizzaro, S., & Spina, D. (2020). Human-in-the-loop artificial intelligence for fighting online misinformation: Challenges and opportunities. *The Bulletin of the Technical Committee on Data Engineering*, *43*(3).
Demartini, G., Trushkowsky, B., Kraska, T., Franklin, M. J., & Berkeley, U. (2013). Crowdq: Crowdsourced query understanding. In *CIDR*.
Dhole, K. D., Gangal, V., Gehrmann, S., Gupta, A., Li, Z., Mahamood, S., et al. (2021). Nl-augmenter: A framework for task-sensitive natural language augmentation. arXiv preprint arXiv:2112.02721.
Diggelmann, T., Boyd-Graber, J., Bulian, J., Ciaramita, M., & Leippold, M. (2020). CLIMATE-FEVER: A dataset for verification of real-world climate claims.
Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, *20*(5), 533–534.
Efron, B., & Tibshirani, R. (1985). The bootstrap method for assessing statistical accuracy. *Behaviormetrika*, *12*(17), 1–35.
Ekstrand, M. D., Das, A., Burke, R., Diaz, F., et al. (2022). Fairness in information access systems. *Foundations and Trends® in Information Retrieval*, *16*(1–2), 1–177.
Elsayed, T., Nakov, P., Barrón-Cedeño, A., Hasanain, M., Suwaileh, R., Da San Martino, G., et al. (2019). Overview of the CLEF-2019 CheckThat! lab: automatic identification and verification of claims. In *International conference of the cross-language evaluation forum for European languages* (pp. 301–321). Springer.
Enayet, O., & El-Beltagy, S. R. (2017). 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). Vancouver, Canada: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/S17-2082, URL: https://aclanthology.org/S17-2082.
Fan, A., Piktus, A., Petroni, F., Wenzek, G., Saeidi, M., Vlachos, A., et al. (2020). Generating fact checking briefs. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 7147–7161). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-main.580, Online. URL: https://aclanthology.org/2020.emnlp-main.580.
Farinneya, P., Abdollah Pour, M. M., Hamidian, S., & Diab, M. (2021). Active learning for rumor identification on social media. In *Findings of the association for computational linguistics: EMNLP 2021* (pp. 4556–4565). Punta Cana, Dominican Republic: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.findings-emnlp.387, URL: https://aclanthology.org/2021.findings-emnlp.387.
Ferreira, W., & Vlachos, A. (2016). Emergent: a novel data-set for stance classification. In *NAACL*.
Gad-Elrab, M. H., Stepanova, D., Urbani, J., & Weikum, G. (2019). 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).
Gold, D., Kovatchev, V., & Zesch, T. (2019). Annotating and analyzing the interactions between meaning relations. In *Proceedings of the 13th linguistic annotation workshop* (pp. 26–36).
Gorrell, G., Kochkina, E., Liakata, M., Aker, A., Zubiaga, A., Bontcheva, K., et al. (2019). SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 845–854).
Graves, L. (2017). Anatomy of a fact check: Objective practice and the contested epistemology of fact checking. *Communication, Culture & Critique*, *10*(3), 518–537.
Graves, D. (2018a). Understanding the promise and limits of automated fact-checking.
Graves, L. (2018b). Boundaries not drawn: Mapping the institutional roots of the global fact-checking movement. *Journalism Studies*, *19*(5), 613–631.
Graves, L., & Amazeen, M. A. (2019). Fact-checking as idea and practice in journalism. In *Oxford research encyclopedia of communication*.
Gruppi, M., Horne, B. D., & Adalı, S. (2021). NELA-GT-2020: A large multi-labelled news dataset for the study of misinformation in news articles. arXiv preprint arXiv:2102.04567.
Guo, H., Cao, J., Zhang, Y., Guo, J., & Li, J. (2018). Rumor detection with hierarchical social attention network. In *Proceedings of the 27th ACM international conference on information and knowledge management* (pp. 943–951).
Guo, Z., Schlichtkrull, M., & Vlachos, A. (2022). A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, *10*, 178–206. http://dx.doi.org/10.1162/tacl_a_00454, arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00454/1987018/tacl_a_00454.pdf.
Gupta, V., Mehta, M., Nokhiz, P., & Srikumar, V. (2020). 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, http://dx.doi.org/10.18653/v1/2020.acl-main.210, Online. URL: https://aclanthology.org/2020.acl-main.210.
Gupta, A., & Srikumar, V. (2021). X-FACT: A new benchmark dataset for multilingual fact checking. In *ACL/IJCNLP*.
Hanselowski, A., Avinesh, P. V. S., Schiller, B., Caspelherr, F., Chaudhuri, D., Meyer, C. M., et al. (2018). A retrospective analysis of the fake news challenge stance-detection task. In *COLING*.
Hanselowski, A., Stab, C., Schulz, C., Li, Z., & Gurevych, I. (2019). 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).
Hardalov, M., Arora, A., Nakov, P., & Augenstein, I. (2021). A survey on stance detection for mis- and disinformation identification. arXiv arXiv:2103.00242.
Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., & Kamar, E. (2022). ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 3309–3326). Dublin, Ireland: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.acl-long.234, URL: https://aclanthology.org/2022.acl-long.234.
Hasanain, M., & Elsayed, T. (2020). bigIR at CheckThat! 2020: Multilingual BERT for ranking arabic tweets by check-worthiness. In *CLEF (working notes)*.
Hasanain, M., & Elsayed, T. (2021). Studying effectiveness of web search for fact checking. *Journal of the Association for Information Science and Technology*, http://dx.doi.org/10.1002/asi.24577, https://onlinelibrary.wiley.com/doi/full/10.1002/asi.24577 https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.24577 https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24577.
Hasanain, M., & Elsayed, T. (2022). Studying effectiveness of web search for fact checking. *Journal of the Association for Information Science and Technology*, *73*(5), 738–751.
Hase, P., & Bansal, M. (2020). Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *ACL*.
Hassan, N., Arslan, F., Li, C., & Tremayne, M. (2017). Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1803–1812).
Hassan, N., Li, C., & Tremayne, M. (2015). 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).
Hassan, N., Zhang, G., Arslan, F., Caraballo, J., Jimenez, D., Gawsane, S., et al. (2017). Claimbuster: The first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, *10*(12), 1945–1948.
Horne, B. D., Khedr, S., & Adali, S. (2018). Sampling the news producers: A large news and feature data set for the study of the complex media landscape. In *Twelfth international AAAI conference on web and social media*.

Hsu, C.-C., & Tan, C. (2021). Decision-focused summarization. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 117–132).

Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4198–4205). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.386, Online. URL: https://aclanthology.org/2020.acl-main.386.

Jacovi, A., & Goldberg, Y. (2021). Aligning faithful interpretations with their social attribution. *Transactions of the Association for Computational Linguistics*, *9*, 294–310. http://dx.doi.org/10.1162/tacl_a_00367, arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00367/1923972/tacl_a_00367.pdf.

Jain, K., Garg, A., & Jain, S. (2021). Reconstructing diffusion model for virality detection in news spread networks. In *Research anthology on fake news, political warfare, and combatting the spread of misinformation* (pp. 98–111). IGI Global.

Jain, S., & Wallace, B. C. (2019). Attention is not explanation. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 3543–3556).

Jiang, Y., Bordia, S., Zhong, Z., Dognin, C., Singh, M., & Bansal, M. (2020). Hover: A dataset for many-hop fact extraction and claim verification. In *Proceedings of the 2020 conference on empirical methods in natural language processing: findings* (pp. 3441–3460).

Joachims, T., & Radlinski, F. (2007). Search engines that learn from implicit feedback. *Computer*, *40*(8), 34–40.

Jones, M. O. (2019). The gulf information war| propaganda, fake news, and fake trends: The weaponization of twitter bots in the gulf crisis. *International Journal of Communication*, *13*, 27.

Juneja, P., & Mitra, T. (2022). Human and technological infrastructures of fact-checking. arXiv preprint arXiv:2205.10894.

Karagiannis, G., Saeed, M., Papotti, P., & Trummer, I. (2020). Scrutinizer: A mixed-initiative approach to large-scale, data-driven claim verification. *Proceedings of the VLDB Endowment*, *13*(12), 2508–2521. http://dx.doi.org/10.14778/3407790.3407841.

Kaufman, R. A., Haupt, M. R., & Dow, S. P. (2022). Who's in the crowd matters: Cognitive factors and beliefs predict misinformation assessment accuracy. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 1–18.

Kazemi, A., Gaffney, D., Garimella, K., & Hale, S. A. (2021). Claim matching beyond english to scale global fact-checking. In *ACL-IJCNLP 2021 - 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing, proceedings of the conference* (pp. 4504–4517). http://dx.doi.org/10.18653/v1/2021.acl-long.347, URL: http://arxiv.org/abs/2106.00853. arXiv:2106.00853.

Kazemi, A., Garimella, K., Shahi, G. K., Gaffney, D., & Hale, S. A. (2021). Tiplines to combat misinformation on encrypted platforms: A case study of the 2019 Indian election on WhatsApp. arXiv abs/2106.04726.

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., et al. (2020). The hateful memes challenge: Detecting hate speech in multimodal memes. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), *Advances in neural information processing systems, Vol. 33* (pp. 2611–2624). Curran Associates, Inc., URL: https://proceedings.neurips.cc/paper/2020/file/1b84c4cee2b8b3d823b30e2d604b1878-Paper.pdf.

Kim, J., & Choi, K.-S. (2020). Unsupervised fact checking by counter-weighted positive and negative evidential paths in a knowledge graph. In *Proceedings of the 28th international conference on computational linguistics* (pp. 1677–1686).

Kochkina, E., Liakata, M., & Augenstein, I. (2017). Turing at SemEval-2017 task 8: Sequential approach to rumour stance classification with branch-LSTM. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)* (pp. 475–480). Vancouver, Canada: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/S17-2083, URL: https://aclanthology.org/S17-2083.

Konstantinovskiy, L., Price, O., Babakar, M., & Zubiaga, A. (2021). Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digital Threats: Research and Practice*, *2*(2), 1–16.

Kotonya, N., Spooner, T., Magazzeni, D., & Toni, F. (2021). Graph reasoning with context-aware linearization for interpretable fact extraction and verification. In *Proceedings of the fourth workshop on fact extraction and verification* (pp. 21–30). Dominican Republic: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.fever-1.3, URL: https://aclanthology.org/2021.fever-1.3.

Kotonya, N., & Toni, F. (2020a). Explainable automated fact-checking: A survey. In *COLING*.

Kotonya, N., & Toni, F. (2020b). Explainable automated fact-checking for public health claims. In *EMNLP*.

Kovatchev, V., Chatterjee, T., Govindarajan, V. S., Chen, J., Choi, E., Chronis, G., et al. (2022). Longhorns at DADC 2022: How many linguists does it take to fool a question answering model? A systematic approach to adversarial attacks. In *Proceedings of the first workshop on dynamic adversarial data collection* (pp. 41–52).

Kovatchev, V., Smith, P., Lee, M., & Devine, R. (2021). Can vectors read minds better than experts? Comparing data augmentation strategies for the automated scoring of children's mindreading ability. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)* (pp. 1196–1206).

Kovatchev, V., Smith, P., Lee, M., Traynor, I. G., Aguilera, I. L., & Devine, R. (2020). "What is on your mind?" Automated scoring of mindreading in childhood and early adolescence. In *Proceedings of the 28th international conference on computational linguistics* (pp. 6217–6228).

Kutlu, M., McDonnell, T., Elsayed, T., & Lease, M. (2020). Annotator rationales for labeling tasks in crowdsourcing. *Journal of Artificial Intelligence Research*, *69*, 143–189.

La Barbera, D., Roitero, K., Demartini, G., Mizzaro, S., & Spina, D. (2020). Crowdsourcing truthfulness: The impact of judgment scale and assessor bias. *Advances in Information Retrieval*, *12036*, 207.

Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., & Tan, C. (2021). Towards a science of human-AI decision making: A survey of empirical studies. arXiv preprint arXiv:2112.11471.

Lawrence, J., & Reed, C. (2020). Argument mining: A survey. *Computational Linguistics*, *45*(4), 765–818.

Lease, M. (2018). Fact checking and information retrieval. In *DESIRES* (pp. 97–98).

Lease, M. (2020). Designing human-AI partnerships to combat misinfomation.

LeBeau, C. (2017). Entitled to the facts: A fact-checking role for librarians. *Reference and User Services Quarterly*, *57*(2), 76–78.

Lee, N., Bang, Y., Madotto, A., & Fung, P. (2021). Towards few-shot fact-checking via perplexity. In *Proceedings of the 2021 conference of the north American chapter of the association for computational linguistics: human language technologies* (pp. 1971–1981). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.naacl-main.158, Online. URL: https://aclanthology.org/2021.naacl-main.158.

Lee, N., Li, B. Z., Wang, S., Yih, W.-t., Ma, H., & Khabsa, M. (2020). Language models as fact checkers? In *Proceedings of the third workshop on fact extraction and verification (FEVER)* (pp. 36–41). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.fever-1.5, Online. URL: https://aclanthology.org/2020.fever-1.5.

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–1270.

Lee, J. D., & See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.

Leskovec, J., Backstrom, L., & Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 497–506).

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., et al. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7871–7880).

Li, T., Fang, L., Lou, J. G., Li, Z., & Zhang, D. (2021). AnaSearch: Extract, Retrieve and Visualize Structured Results from Unstructured Text for Analytical Queries. In *WSDM 2021 - proceedings of the 14th ACM international conference on web search and data mining* (pp. 906–909). http://dx.doi.org/10.1145/3437963.3441694.

Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., et al. (2016). A survey on truth discovery. *ACM SIGKDD Explorations Newsletter*, *17*(2), 1–16.

Lillie, A. E., Middelboe, E. R., & Derczynski, L. (2019). Joint rumour stance and veracity prediction. In *Proceedings of the 22nd nordic conference on computational linguistics* (pp. 208–221).

Liu, A., Swayamdipta, S., Smith, N. A., & Choi, Y. (2022). WANLI: Worker and AI collaboration for natural language inference dataset creation.

Lu, Y.-J., & Li, C.-T. (2020). GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 505–514).

Ma, J., Gao, W., Joty, S., & Wong, K.-F. (2019). Sentence-level evidence embedding for claim verification with hierarchical attention networks. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 2561–2571). Florence, Italy: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P19-1244, URL: https://aclanthology.org/P19-1244.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., et al. (2016). Detecting rumors from microblogs with recurrent neural networks. In *IJCAI*.

Ma, J., Gao, W., & Wong, K.-F. (2018). Rumor detection on Twitter with tree-structured recursive neural networks. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 1980–1989). Melbourne, Australia: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P18-1184, URL: https://aclanthology.org/P18-1184.

Marcus, J. (1992). *Mesoamerican writing systems: propaganda, myth, and history in four ancient civilizations*. Princeton University Press Princeton.

Martinez-Rico, J., Martinez-Romo, J., & Araujo, L. (2021). L.: NLP&IR@ UNED at CheckThat! 2021: check-worthiness estimation and fake news detection using transformer models. Faggioli et al.[33].

Micallef, N., Armacost, V., Memon, N., & Patil, S. (2022). True or false: Studying the work practices of professional fact-checkers. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW1), 1–44. http://dx.doi.org/10.1145/3512974.

Mihalcea, R., & Strapparava, C. (2009). The Lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 conference short papers* (pp. 309–312). Suntec, Singapore: Association for Computational Linguistics, URL: https://aclanthology.org/P09-2078.

Mihaylova, T., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Mohtarami, M., Karadzhov, G., et al. (2018). Fact checking in community forums. In *Thirty-second AAAI conference on artificial intelligence*.

Miranda, S., Nogueira, D., Mendes, A., Vlachos, A., Secker, A., Garrett, R., et al. (2019). Automated fact checking in the news room. In *The world wide web conference* (pp. 3579–3583).

Mohseni, S., Yang, F., Pentyala, S., Du, M., Liu, Y., Lupfer, N., et al. (2021). Machine learning explanations to prevent overtrust in fake news detection. In *Proceedings of the international AAAI conference on web and social media, Vol. 15* (pp. 421–431).

Molnar, C. (2020). Interpretable machine learning. Lulu. com.

Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake news detection on social media using geometric deep learning. arXiv preprint arXiv:1902.06673.

Nakamura, K., Levy, S., & Wang, W. Y. (2020). Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the 12th language resources and evaluation conference* (pp. 6149–6157). Marseille, France: European Language Resources Association, URL: https://aclanthology.org/2020.lrec-1.755.

Nakashole, N., & Mitchell, T. (2014). 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).

Nakov, P., Corney, D., Hasanain, M., Alam, F., Elsayed, T., Barrón-Cedeño, A., et al. (2021). Automated fact-checking for assisting human fact-checkers. In *IJCAI*.

Nakov, P., Da San Martino, G., Barrón-Cedeño, A., Zaghouani, W., Míguez, R., Alam, F., et al. (2022). CheckThat! Lab on fighting the COVID-19 infodemic and fake news detection (proposal for a CLEF-2022 lab).

Nakov, P., Da San Martino, G., Elsayed, T., Barrón-Cedeño, A., Míguez, R., Shaar, S., et al. (2021). The CLEF-2021 CheckThat! Lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *ECIR (2)*.

Neely-Sardon, A., & Tignor, M. (2018). Focus on the facts: A news and information literacy instructional program. *The Reference Librarian*, *59*(3), 108–121.

Neumann, T., De-Arteaga, M., & Fazelpour, S. (2022). Justice in misinformation detection systems: An analysis of algorithms, stakeholders, and potential harms. In *2022 ACM conference on fairness, accountability, and transparency* (pp. 1504–1515). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3531146.3533205.

Nguyen, A. T., Kharosekar, A., Krishnan, S., Krishnan, S., Tate, E., Wallace, B. C., et al. (2018). Believe it or not: Designing a human-ai partnership for mixed-initiative fact-checking. In *Proceedings of the 31st annual ACM symposium on user interface software and technology* (pp. 189–199).

Nguyen, A. T., Kharosekar, A., Lease, M., & Wallace, B. C. (2018). An interpretable joint graphical model for fact-checking from crowds. In *Proceedings of the thirty-second AAAI conference on artificial intelligence (AAAI-18)* (pp. 1511–1518). URL: ../papers/nguyen-aaai18.pdf.

Nguyen, T. T., Weidlich, M., Yin, H., Zheng, B., Nguyen, Q. H., & Nguyen, Q. V. H. (2020). Factcatch: Incremental pay-as-you-go fact checking with minimal user effort. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval* (pp. 2165–2168).

Nie, Y., Wang, S., & Bansal, M. (2019). Revealing the importance of semantic retrieval for machine reading at scale. 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. 2553–2566). Hong Kong, China: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D19-1258, URL: https://aclanthology.org/D19-1258.

Niewiński, P., Pszona, M., & Janicka, M. (2019). GEM: Generative enhanced model for adversarial attacks. In *Proceedings of the second workshop on fact extraction and verification* (pp. 20–26).

Nørregaard, J., Horne, B. D., & Adalı, S. (2019). Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In *Proceedings of the international AAAI conference on web and social media, Vol. 13* (pp. 630–638).

Oshikawa, R., Qian, J., & Wang, W. Y. (2020). A survey on natural language processing for fake news detection. In *LREC*.

Popat, K., Mukherjee, S., Yates, A., & Weikum, G. (2018). Declare: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 22–32).

Potthast, M., Kieselj, R., et al. (2018). A stylometric inquiryintohyperpartisanandfakenews. In *Proc of the 56th annual meeting of the association for computational linguistics*. Stroudsburg, PA: ACL, Article 231240.

Potthast, M., Köpsel, S., Stein, B., & Hagen, M. (2016). Clickbait detection. In *European conference on information retrieval* (pp. 810–817). Springer.

Pradeep, R., Ma, X., Nogueira, R., & Lin, J. (2021). Scientific claim verification with vert5erini. In *Proceedings of the 12th international workshop on health text mining and information analysis* (pp. 94–103). Association for Computational Linguistics, online. URL: https://aclanthology.org/2021.louhi-1.11.

Qazvinian, V., Rosengren, E., Radev, D., & Mei, Q. (2011). Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 1589–1599).

Qu, Y., Barbera, D. L., Roitero, K., Mizzaro, S., Spina, D., & Demartini, G. (2021). Combining human and machine confidence in truthfulness assessment. *ACM Journal of Data and Information Quality (JDIQ)*.

Qu, Y., Roitero, K., Mizzaro, S., Spina, D., & Demartini, G. (2021). Human-in-the-loop systems for truthfulness: A study of human and machine confidence.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). 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).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 856–865).
Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4902–4912).
Roitero, K., Soprano, M., Fan, S., Spina, D., Mizzaro, S., & Demartini, G. (2020). Can the crowd identify misinformation objectively? The effects of judgment scale and assessor's background. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval* (pp. 439–448).
Roitero, K., Soprano, M., Portelli, B., Spina, D., Della Mea, V., Serra, G., et al. (2020). The covid-19 infodemic: Can the crowd judge recent misinformation objectively? In *Proceedings of the 29th ACM international conference on information & knowledge management* (pp. 1305–1314).
Sarasua, C., Simperl, E., & Noy, N. F. (2012). Crowdmap: Crowdsourcing ontology alignment with microtasks. In *International semantic web conference* (pp. 525–541). Springer.
Schuster, T., Fisch, A., & Barzilay, R. (2021). 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, http://dx.doi.org/10.18653/v1/2021.naacl-main.52, Online. URL: https://aclanthology.org/2021.naacl-main.52.
Schuster, T., Schuster, R., Shah, D. J., & Barzilay, R. (2020). The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, *46*(2), 499–510.
Serrano, S., & Smith, N. A. (2019). Is attention interpretable? In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 2931–2951).
Settles, B. (2009). Active learning literature survey.
Shaar, S., Alam, F., Da San Martino, G., & Nakov, P. (2021). Assisting the human fact-checkers: Detecting all previously fact-checked claims in a document. arXiv preprint arXiv:2109.07410.
Shaar, S., Hasanain, M., Hamdan, B., Ali, Z. S., Haouari, F., Nikolov, A., et al. (2021). Overview of the CLEF-2021 CheckThat! lab task 1 on check-worthiness estimation in tweets and political debates. In *CLEF*.
Shaar, S., Martino, G. D. S., Babulkov, N., & Nakov, P. (2020). That is a known Lie: Detecting previously fact-checked claims. In *ACL*.
Shabani, S., Charlesworth, Z., Sokhn, M., & Schuldt, H. (2021). SAMS: Human-in-the-loop approach to combat the sharing of digital misinformation. In *CEUR workshop proc.*.
Shao, C., Ciampaglia, G. L., Flammini, A., & Menczer, F. (2016). Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th international conference companion on world wide web* (pp. 745–750).
Shi, L., Bhattacharya, N., Das, A., Lease, M., & Gwizdka, J. (2022). The effects of interactive AI design on user behavior: An eye-tracking study of fact-checking COVID-19 claims. In *Proceedings of the 7th ACM SIGIR conference on human information, interaction and retrieval*. URL: https://utexas.box.com/v/shi-chiir2022.
Shi, B., & Weninger, T. (2016). Discriminative predicate path mining for fact checking in knowledge graphs. *Knowledge-Based Systems*, *104*, 123–133.
Shu, K., Cui, L., Wang, S., Lee, D., & Liu, H. (2019). Defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 395–405).
Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). 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.
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.
Singh, P., Das, A., Li, J. J., & Lease, M. (2021). The case for claim difficulty assessment in automatic fact checking. arXiv preprint arXiv:2109.09689.
Smeros, P., Castillo, C., & Aberer, K. (2021). SciClops: Detecting and contextualizing scientific claims for assisting manual fact-checking. In *Proceedings of the 30th ACM international conference on information & knowledge management*.
Sokol, K., & Flach, P. (2019). Desiderata for interpretability: explaining decision tree predictions with counterfactuals. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 33* (pp. 10035–10036).
Soprano, M., Roitero, K., La Barbera, D., Ceolin, D., Spina, D., Mizzaro, S., et al. (2021). The many dimensions of truthfulness: Crowdsourcing misinformation assessments on a multidimensional scale. *Information Processing & Management*, *58*(6), Article 102710.
The Poynter Institute (2021). Global fact 8 pre-recorded segment 4. URL: https://www.youtube.com/watch?v=gOhPKDaeQxI&t=770s.
Thorne, J., & Vlachos, A. (2019). Adversarial attacks against fact extraction and verification. arXiv preprint arXiv:1903.05543.
Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). 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).
Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., & Mittal, A. (2019). The FEVER2. 0 shared task. In *Proceedings of the second workshop on fact extraction and verification* (pp. 1–6).
Thornhill, C., Meeus, Q., Peperkamp, J., & Berendt, B. (2019). A digital nudge to counter confirmation bias. *Frontiers in Big Data*, *2*, 11.
Tschiatschek, S., Singla, A., Gomez Rodriguez, M., Merchant, A., & Krause, A. (2018). Fake news detection in social networks via crowd signals. In *Companion proceedings of the the web conference 2018* (pp. 517–524).
Uscinski, J. E. (2015). The epistemology of fact checking (is still naïve): Rejoinder to amazeen. *Critical Review*, *27*(2), 243–252.
Vaish, R., Davis, J., & Bernstein, M. (2015). Crowdsourcing the research process. *Collective Intelligence*, *3*.
Vaish, R., Gaikwad, S. N. S., Kovacs, G., Veit, A., Krishna, R., Arrieta Ibarra, I., et al. (2017). Crowd research: Open and scalable university laboratories. In *Proceedings of the 30th annual ACM symposium on user interface software and technology* (pp. 829–843).
Vaughan, J. W., & Wallach, H. (2020). A human-centered agenda for intelligible machine learning. In *Machines we trust: getting along with artificial intelligence*.
Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1–14).
Vicario, M. D., Quattrociocchi, W., Scala, A., & Zollo, F. (2019). Polarization and fake news: Early warning of potential misinformation targets. *ACM Transactions on the Web (TWEB)*, *13*(2), 1–22.
Vlachos, A., & Riedel, S. (2014). Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science* (pp. 18–22).
Vlachos, A., & Riedel, S. (2015). 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). Lisbon, Portugal: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D15-1312, URL: https://aclanthology.org/D15-1312.
Vo, N., & Lee, K. (2018). The rise of guardians: Fact-checking url recommendation to combat fake news. In *The 41st international ACM SIGIR conference on research & development in information retrieval* (pp. 275–284).
Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., et al. (2020). Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 7534–7550).
Wang, W. Y. (2017). "Liar, liar pants on fire": A new benchmark dataset for fake news detection. In *ACL*.
Wiegreffe, S., & Marasovic, A. (2021). Teach me to explain: A review of datasets for explainable natural language processing. In J. Vanschoren, & S. Yeung (Eds.), *Proceedings of the neural information processing systems track on datasets and benchmarks, Vol. 1*. URL: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/698d51a19d8a121ce581499d7b701668-Paper-round1.pdf.
Wiegreffe, S., & Pinter, Y. (2019). Attention is not not explanation. 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. 11–20).

Williams, E., Rodrigues, P., & Novak, V. (2020). Accenture at CheckThat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models. arXiv preprint arXiv:2009.02431.

Yang, F., Pentyala, S. K., Mohseni, S., Du, M., Yuan, H., Linder, R., et al. (2019). Xfake: Explainable fake news detector with visualizations. In *The world wide web conference* (pp. 3600–3604).

Zaidan, O., Eisner, J., & Piatko, C. (2007). Using "annotator rationales" to improve machine learning for text categorization. In *Human language technologies 2007: the conference of the north American chapter of the association for computational linguistics; proceedings of the main conference* (pp. 260–267).

Zanzotto, F. M. (2019). Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, *64*, 243–252.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., et al. (2020). Defending against neural fake news. *Neurips*.

Zeng, X., Abumansour, A. S., & Zubiaga, A. (2021). Automated fact-checking: A survey. *Language and Linguistics Compass*, *15*(10), Article e12438.

Zhang, X., Cao, J., Li, X., Sheng, Q., Zhong, L., & Shu, K. (2021). Mining dual emotion for fake news detection. In *Proceedings of the web conference 2021* (pp. 3465–3476).

Zhang, Y., Lease, M., & Wallace, B. (2017). Active discriminative text representation learning. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 31*.

Zhang, Z., Rudra, K., & Anand, A. (2021). FaxPlainAC: A fact-checking tool based on explainable models with human correction in the loop. In *Proceedings of the 30th ACM international conference on information & knowledge management* (pp. 4823–4827).

Zhou, X., Jain, A., Phoha, V. V., & Zafarani, R. (2020). Fake news early detection: A theory-driven model. *Digital Threats: Research and Practice*, *1*(2), 1–25.

Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, *53*(5), 1–40.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2018). Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys*, *51*(2), 1–36.

Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G., & Tolmie, P. (2016). Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS One*, *11*(3), Article e0150989.