

A Survey on Automated Fact-Checking

Zhijiang Guo*, Michael Schlichtkrull*, Andreas Vlachos

Department of Computer Science and Technology

University of Cambridge

{zg283,mss84,av308}@cam.ac.uk

Abstract

Fact-checking has become increasingly important due to the speed with which both information and misinformation can spread in the modern media ecosystem. Therefore, researchers have been exploring how fact-checking can be automated, using techniques based on natural language processing, machine learning, knowledge representation, and databases to automatically predict the veracity of claims. In this paper, we survey automated fact-checking stemming from natural language processing, and discuss its connections to related tasks and disciplines. In this process, we present an overview of existing datasets and models, aiming to unify the various definitions given and identify common concepts. Finally, we highlight challenges for future research.

1 Introduction

Fact-checking is the task of assessing whether claims made in written or spoken language are true. This is an essential task in journalism, and is commonly conducted manually by dedicated organizations such as PolitiFact. In addition to *external* fact-checking, *internal* fact-checking is also performed by publishers of newspapers, magazines, and books prior to publishing in order to promote truthful reporting. Figure 1 shows an example from PolitiFact, together with the evidence (summarized) and the verdict.

Fact-checking is a time-consuming task. To assess the claim in Figure 1, a journalist would need to search through potentially many sources to find job gains under Trump and Obama, evaluate the reliability of each source, and make a comparison. This process can take professional fact-checkers several hours or days (Hassan et al., 2015; Adair et al., 2017). Compounding the problem, fact-checkers often work under strict and

* Equally Contributed.

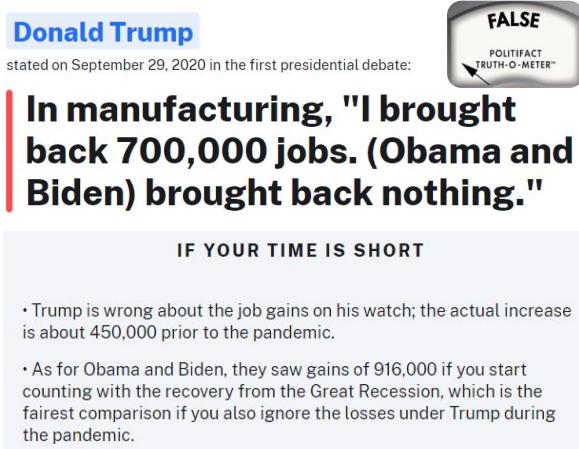


Figure 1: An example of a fact-checked statement. Referring to the manufacturing sector, Donald Trump said “*I brought back 700,000 jobs. Obama and Biden brought back nothing.*” The fact-checker gave the verdict *False* based on the collected evidence.

tight deadlines, especially in the case of internal processes (Borel, 2016; Godler and Reich, 2017), and some studies have shown that less than half of all published articles have been subject to verification (Lewis et al., 2008). Given the amount of new information that appears and the speed with which it spreads, manual validation is insufficient.

Automating the fact-checking process has been discussed in the context of computational journalism (Flew et al., 2010; Cohen et al., 2011; Graves, 2018), and has received significant attention in the artificial intelligence community. Vlachos and Riedel (2014) proposed structuring it as a sequence of components – identifying claims to be checked, finding appropriate evidence, producing verdicts – that can be modelled as natural language processing (NLP) tasks. This motivated the development of automated pipelines consisting of sub-tasks that can be mapped to tasks well-explored in the NLP community. Advances were made possible by the development of datasets, consisting of either claims collected from fact-checking websites, e.g. Liar (Wang, 2017), or purpose-made for

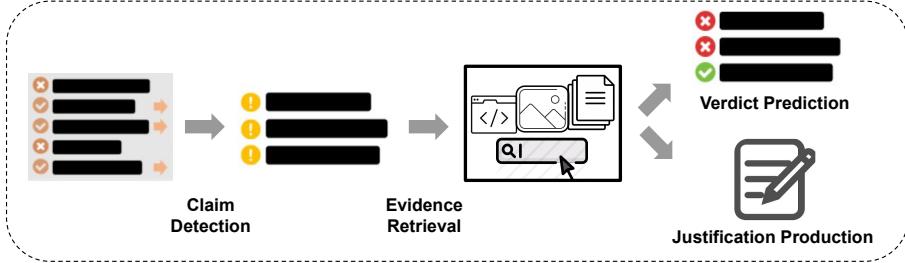


Figure 2: A natural language processing framework for automated fact-checking.

research, e.g. FEVER (Thorne et al., 2018a).

A growing body of research is exploring the various tasks and subtasks necessary for the automation of fact-checking, and to meet the need for new methods to address emerging challenges. Early developments were surveyed in Thorne and Vlachos (2018), which remains the closest to an exhaustive overview of the subject. However, their proposed framework does not include work on determining *which* claims to verify (i.e. claim detection), nor does their survey include the recent work on producing explainable, convincing verdicts (i.e. justification production).

Several recent papers have surveyed research focusing on individual components of the task. Zubiaga et al. (2018) and Islam et al. (2020) focus on identifying rumours on social media, Küçük and Can (2020) and Hardalov et al. (2021) on detecting the stance of a given piece of evidence towards a claim, and Kotonya and Toni (2020a) on producing explanations and justifications for fact-checks. Finally Nakov et al. (2021a) surveyed automated approaches to assist fact-checking by humans. While these surveys are extremely useful in understanding various aspects of fact-checking technology, they are fragmented and focused on specific subtasks and components; our aim is to give a comprehensive and exhaustive birds-eye view of the subject as a whole.

A number of papers have surveyed related tasks. Lazer et al. (2018) and Zhou and Zafarani (2020) surveyed work on fake news, including descriptive work on the problem, as well as work seeking to counteract fake news through computational means. A comprehensive review of NLP approaches to fake news detection was also provided in Oshikawa et al. (2020). However, fake news detection differs in scope from fact-checking, as the former focuses on assessing news articles, and includes labelling items based on aspects not related to veracity, such as satire detection (Oshikawa et al., 2020; Zhou and Zafarani, 2020). Furthermore, other factors – such as the audience

reached by the claim, and the intentions and forms of the claim – are often considered. These factors also feature in the context of propaganda detection, recently surveyed by Da San Martino et al. (2020b). Unlike these efforts, the works discussed in this survey concentrate on assessing veracity of general-domain claims. Finally, Shu et al. (2017) and da Silva et al. (2019) surveyed research on fake news detection and fact-checking with a focus on social media data, while this survey covers fact-checking across domains and sources, including newswire, science, etc.

In this survey, we present a comprehensive and up-to-date survey of automated fact-checking, unifying various definitions developed in previous research into a common framework. We begin by defining the three stages of our fact-checking framework – claim detection, evidence retrieval, and claim verification, the latter consisting of verdict prediction and justification production. We then give an overview of the existing datasets and modelling strategies, taxonomizing these and contextualizing them with respect to our framework. We finally discuss key research challenges that have been addressed, and give directions for challenges which we believe should be tackled by future research. We accompany the survey with a repository,¹ which lists the resources mentioned in our survey.

2 Task Definition

Figure 2 shows a NLP framework for automated fact-checking consisting of three stages: (i) *claim detection* to identify claims that require verification; (ii) *evidence retrieval* to find sources supporting or refuting the claim; (iii) *claim verification* to assess the veracity of the claim based on the retrieved evidence. Evidence retrieval and claim verification are sometimes tackled as a single task referred to as *factual verification*, while claim detec-

¹www.github.com/Cartus/Automated-Fact-Checking-Resources

tion is often tackled separately. Claim verification can be decomposed into two parts that can be tackled separately or jointly: *verdict prediction*, where claims are assigned truthfulness labels, and *justification production*, where explanations for verdicts must be produced.

2.1 Claim Detection

The first stage in automated fact-checking is claim detection, where claims are selected for verification. Commonly, detection relies on the concept of check-worthiness. Hassan et al. (2015) defined check-worthy claims as those for which the general public would be interested in knowing the truth. For example, “*over six million Americans had COVID-19 in January*” would be check-worthy, as opposed to “*water is wet*”. This can involve a binary decision for each potential claim, or an importance-ranking of claims (Atanasova et al., 2018; Barrón-Cedeño et al., 2020). The latter parallels standard practice in internal journalistic fact-checking, where deadlines often require fact-checkers to employ a triage system (Borel, 2016).

Another instantiation of claim detection based on check-worthiness is rumour detection. A rumour can be defined as an unverified story or statement circulating (typically on social media) (Ma et al., 2016; Zubiaga et al., 2018). Rumour detection considers language subjectivity and growth of readership through a social network (Qazvinian et al., 2011). Typical input to a rumour detection system is a stream of social media posts, whereupon a binary classifier has to determine if each post is rumourous. Metadata, such as the number of likes and re-posts, is often used as features to identify rumours (Zubiaga et al., 2016; Gorrell et al., 2019; Zhang et al., 2021).

Check-worthiness and rumourousness can be subjective. For example, the importance placed on countering COVID-19 misinformation is not uniform across every social group. The check-worthiness of each claim also varies over time, as countering misinformation related to current events is in many cases understood to be more important than countering older misinformation (e.g. misinformation about COVID-19 has a greater societal impact in 2021 than misinformation about the Spanish flu). Furthermore, older rumours may have already been debunked by journalists, reducing their impact. Misinformation that is harmful to marginalized communities may also be

judged to be less check-worthy by the general public than misinformation that targets the majority. Conversely, claims *originating from* marginalised groups may be subject to greater scrutiny than claims originating from the majority; for example, journalists have been shown to assign greater trust and therefore lower need for verification to stories produced by male sources (Barnoy and Reich, 2019). Such biases could be replicated in datasets that capture the (often implicit) decisions made by journalists about which claims to prioritize.

Instead of using subjective concepts, Konstantinovskiy et al. (2021) framed claim detection as whether a claim makes an assertion about the world that is checkable, i.e. whether it is verifiable with readily available evidence. Claims based on personal experiences or opinions are uncheckable. For example, “*I woke up at 7 am today*” is not checkable because appropriate evidence cannot be collected; “*cubist art is beautiful*” is not checkable because it is a subjective statement.

2.2 Evidence Retrieval

Evidence retrieval aims to find information beyond the claim – e.g. text, tables, knowledge bases, images, relevant metadata – to indicate veracity. Some earlier efforts do not use any evidence beyond the claim itself (Wang, 2017; Rashkin et al., 2017; Volkova et al., 2017; Dungs et al., 2018). Relying on surface patterns of claims without considering the state of the world fails to identify well-presented misinformation, including machine-generated claims (Schuster et al., 2020). Recent developments in natural language generation have exacerbated this issue (Radford et al., 2019; Brown et al., 2020), with machine-generated text sometimes being perceived as more trustworthy than human-written text (Zellers et al., 2019). In addition to enabling verification, evidence is essential for generating verdict justifications to convince users of fact-checks.

Stance detection can be viewed as an instantiation of evidence retrieval, which typically assumes a more limited amount of potential evidence and predicts its stance towards the claim. For example, Ferreira and Vlachos (2016) used news article headlines from the Emergent project² as evidence to predict whether articles supported, refuted or

²www.cjr.org/tow_center_reports/craig_silverman_lies_damn_lies_viral_content.php

merely reported a claim. The Fake News Challenge (Pomerleau and Rao, 2017) further used entire documents, allowing for evidence from multiple sentences. More recently, Hanselowski et al. (2019) filtered out irrelevant sentences in the summaries of fact-checking articles to obtain fine-grained evidence via stance detection. While both stance detection and evidence retrieval in the context of claim verification are classification tasks, what is considered evidence in the former is broader, including for example a social media post responding “@AJENews @germanwings yes indeed :-(.” to a claim (Gorrell et al., 2019).

A fundamental issue is that not all available information is trustworthy. Most fact-checking approaches implicitly assume access to a trusted information source such as encyclopedias (e.g. Wikipedia (Thorne et al., 2018a)) or results provided (and thus vetted) by search engines (Augenstein et al., 2019). *Evidence* is then defined as information that can be retrieved from this source, and *veracity* as coherence with the evidence. For real-world applications, evidence must be curated through the manual efforts of journalists (Borel, 2016), automated means (Li et al., 2015), or their combination. For example, Full Fact uses tables and legal documents from government organisations as evidence.³

2.3 Verdict Prediction

Given an identified claim and the pieces of evidence retrieved for it, verdict prediction attempts to determine the veracity of the claim. The simplest approach is binary classification, e.g. labelling a claim as true or false (Nakashole and Mitchell, 2014; Popat et al., 2016; Potthast et al., 2018). When evidence is used to verify the claim, it is often preferable to use supported/refuted (by evidence) instead of true/false respectively, as in many cases the evidence itself is not assessed by the systems. More broadly it would be dangerous to make such strong claims about the world given the well-known limitations (Graves, 2018).

Many versions of the task employ finer-grained classification schemes. A simple extension is to use an additional label denoting a lack of information to predict the veracity of the claim (Thorne et al., 2018a). Beyond that, some datasets and systems follow the approach taken by journalistic

fact-checking agencies, employing multi-class labels representing degrees of truthfulness (Wang, 2017; Alhindi et al., 2018; Shahi and Nandini, 2020; Augenstein et al., 2019).

2.4 Justification Production

Justifying decisions is an important part of journalistic fact-checking, as fact-checkers need to convince readers of their interpretation of the evidence (Uscinski and Butler, 2013; Borel, 2016). Debunking purely by calling something *false* often fails to be persuasive, and can induce a “backfire”-effect where belief in the erroneous claim is reinforced (Lewandowsky et al., 2012). This need is even greater for automated fact-checking, which may employ black-box components. When developers deploy black-box models whose decision-making processes cannot be understood, these artefacts can lead to unintended, harmful consequences (O’Neil, 2016). Developing techniques that explain model predictions has been suggested as a potential remedy to this problem (Lipton, 2018), and recent work has focused on the generation of *justifications* (see Kotonya and Toni’s (2020a) survey of explainable claim verification). Research so far has focused on justification production for claim verification, as the latter is often the most scrutinized stage in fact-checking. Nevertheless, explainability may also be desirable and necessary for the other stages in our framework.

Justification production for claim verification typically relies on one of four strategies. First, attention weights can be used to highlight the salient parts of the evidence, in which case justifications typically consist of scores for each evidence token (Popat et al., 2018; Shu et al., 2019; Lu and Li, 2020). Second, decision-making processes can be designed to be understandable by human experts, e.g. by relying on logic-based systems (Gad-Elrab et al., 2019; Ahmadi et al., 2019); in this case, the justification is typically the derivation for the veracity of the claim. Finally, the task can be modelled as a form of summarization, where systems generate textual explanations for their decisions (Atanasova et al., 2020b). While some of these justification types require additional components, we did not introduce a fourth stage in our framework as in some cases the decision-making process of the model is self-explanatory (Gad-Elrab et al., 2019; Ahmadi et al., 2019).

A basic form of justification is to show which

³www.fullfact.org/about/frequently-asked-questions

Dataset	Type	Input	#Inputs	Evidence	Verdict	Sources	Lang
CredBank (Mitra and Gilbert, 2015)	Worthy	Aggregate	1,049	Meta	5 Classes	Twitter	En
Weibo (Ma et al., 2016)	Worthy	Aggregate	5,656	Meta	2 Classes	Twitter/Weibo	En/Ch
PHEME (Zubiaga et al., 2016)	Worthy	Individual	330	Text/Meta	3 Classes	Twitter	En/De
RumourEval19 (Gorrell et al., 2019)	Worthy	Individual	446	Text/Meta	3 Classes	Twitter/Reddit	En
DAST (Lillie et al., 2019)	Worthy	Individual	220	Text/Meta	3 Classes	Reddit	Da
Suspicious (Volkova et al., 2017)	Worthy	Individual	131,584	✗	2/5 Classes	Twitter	En
CheckThat20-T1 (Barrón-Cedeño et al., 2020)	Worthy	Individual	8,812	✗	Ranking	Twitter	En/Ar
CheckThat21-T1A (Nakov et al., 2021b)	Worthy	Individual	17,282	✗	2 Classes	Twitter	Many
Debate (Hassan et al., 2015)	Worthy	Statement	1,571	✗	3 Classes	Transcript	En
ClaimRank (Gencheva et al., 2017)	Worthy	Statement	5,415	✗	Ranking	Transcript	En
CheckThat18-T1 (Atanasova et al., 2018)	Worthy	Statement	16,200	✗	Ranking	Transcript	En/Ar
CitationReason (Redi et al., 2019)	Checkable	Statement	4,000	Meta	13 Classes	Wikipedia	En
PolitiTV (Konstantinovskiy et al., 2021)	Checkable	Statement	6,304	✗	7 Classes	Transcript	En

Table 1: Summary of claim detection datasets. Input can be a set of posts (aggregate) or an individual post from social media, or a statement. Evidence include text and metadata. Verdict can be a multi-class label or a rank list.

pieces of evidence were used to reach a verdict. However, a justification must also explain *how* the retrieved evidence was used, explain any assumptions or commonsense facts employed, and show the reasoning process taken to reach the verdict. Presenting the evidence returned by a retrieval system can as such be seen as a rather weak baseline for justification production, as it does not explain the process used to reach the verdict. There is furthermore a subtle difference between evaluation criteria for evidence and justifications: good evidence facilitates the production of a correct verdict; a good justification accurately reflects the reasoning of the model through a readable and plausible explanation, *regardless* of the correctness of the verdict. This introduces different considerations for justification production, e.g. *readability* (how accessible an explanation is to humans), *plausibility* (how convincing an explanation is), and *faithfulness* (how accurately an explanation reflects the reasoning of the model) (Jacovi and Goldberg, 2020).

3 Datasets

Datasets can be analysed along three axes aligned with three stages of the fact-checking framework (Figure 2): the input, the evidence used, and verdicts and justifications which constitute the output. In this section we bring together efforts that emerged in different communities using different terminologies, but nevertheless could be used to develop and evaluate models for the same task.

3.1 Input

We first consider the inputs to claim detection (summarized in Table 1) as their format and content influences the rest of the process. A typical in-

put is a social media post with textual content. Zubiaga et al. (2016) constructed PHEME based on source tweets in English and German that sparked a high number of retweets exceeding a predefined threshold. Derczynski et al. (2017) introduced the shared task RumourEval using the English section of PHEME; for the 2019 iteration of the shared task, this dataset was further expanded to include Reddit and new Twitter posts (Gorrell et al., 2019). Following the same annotation strategy, Lillie et al. (2019) constructed a Danish dataset by collecting posts from Reddit. Instead of considering only source tweets, subtasks in CheckThat (Barrón-Cedeño et al., 2020; Nakov et al., 2021b) viewed every post as part of the input. A set of auxiliary questions, such as “*does it contain a factual claim?*”, “*is it of general interest?*”, were created to help annotators identify check-worthy posts. Since an individual post may contain limited context, other works (Mitra and Gilbert, 2015; Ma et al., 2016; Zhang et al., 2021) represented each claim by a set of relevant posts, e.g. the thread they originate from.

The second type of textual input is a document consisting of multiple claims. For Debate (Hassan et al., 2015), professionals were asked to select check-worthy claims from U.S. presidential debates to ensure good agreement and shared understanding of the assumptions. On the other hand, Konstantinovskiy et al. (2021) collected checkable claims from transcripts by crowd-sourcing, where workers labelled claims based on a predefined taxonomy. Different from prior works focused on the political domain, Redi et al. (2019) sampled sentences that contain citations from Wikipedia articles, and asked crowd-workers to annotate them based on citation policies.

Dataset	Input	#Inputs	Evidence	Verdict	Sources	Lang
CrimeVeri (Bachenko et al., 2008)	Statement	275	✗	2 Classes	Crime	En
Politifact (Vlachos and Riedel, 2014)	Statement	106	Text/Meta	5 Classes	Fact Check	En
StatsProperties (Vlachos and Riedel, 2015)	Statement	7,092	KG	Numeric	Internet	En
Emergent (Ferreira and Vlachos, 2016)	Statement	300	Text	3 Classes	Emergent	En
CreditAssess (Popat et al., 2016)	Statement	5,013	Text	2 Classes	Fact Check/Wiki	En
PunditFact (Rashkin et al., 2017)	Statement	4,361	✗	2/6 Classes	Fact Check	En
Liar (Wang, 2017)	Statement	12,836	Meta	6 Classes	Fact Check	En
Verify (Baly et al., 2018)	Statement	422	Text	2 Classes	Fact Check	Ar/En
CheckThat18-T2 (Barrón-Cedeño et al., 2018)	Statement	150	✗	3 Classes	Transcript	En
Snopes (Hanselowski et al., 2019)	Statement	6,422	Text	3 Classes	Fact Check	En
MultiFC (Augenstein et al., 2019)	Statement	36,534	Text/Meta	2-27 Classes	Fact Check	En
Climate-FEVER (Diggelmann et al., 2020)	Statement	1,535	Text	4 Classes	Climate	En
SciFact (Wadden et al., 2020)	Statement	1,409	Text	3 Classes	Science	En
PUBHEALTH (Kotonya and Toni, 2020b)	Statement	11,832	Text	4 Classes	Fact Check	En
COVID-Fact (Saakyan et al., 2021)	Statement	4,086	Text	2 Classes	Forum	En
X-Fact (Gupta and Srikumar, 2021)	Statement	31,189	Text	7 Classes	Fact Check	Many
cQA (Mihaylova et al., 2018)	Answer	422	Meta	2 Classes	Forum	En
AnswerFact (Zhang et al., 2020)	Answer	60,864	Text	5 Classes	Amazon	En
NELA (Horne et al., 2018)	Article	136,000	✗	2 Classes	News	En
BuzzfeedNews (Potthast et al., 2018)	Article	1,627	Meta	4 Classes	Facebook	En
BuzzFace (Santia and Williams, 2018)	Article	2,263	Meta	4 Classes	Facebook	En
FA-KES (Salem et al., 2019)	Article	804	✗	2 Classes	VDC	En
FakeNewsNet (Shu et al., 2020)	Article	23,196	Meta	2 Classes	Fact Check	En
FakeCovid (Shahi and Nandini, 2020)	Article	5,182	✗	2 Classes	Fact Check	Many

Table 2: Summary of factual verification datasets with natural inputs. KG denotes knowledge graphs. CheckThat18 has been extended later (Hasanain et al., 2019; Barrón-Cedeño et al., 2020; Nakov et al., 2021b). NELA has been updated by adding more data from more diverse sources (Nørregaard et al., 2019; Gruppi et al., 2020, 2021)

Next, we discuss the inputs to factual verification. The most popular type of input to verification is textual claims, which is expected given they are often the output of claim detection. These tend to be sentence-level statements, which is a practice common among fact-checkers in order to include only the context relevant to the claim (Mena, 2019). Many existing efforts (Vlachos and Riedel, 2014; Wang, 2017; Hanselowski et al., 2019; Augenstein et al., 2019) constructed datasets by crawling real-world claims from dedicated websites (e.g. Politifact) due to their availability (see Table 2). Unlike previous work that focus on English, Gupta and Srikumar (2021) collected non-English claims from 25 languages.

Others extract claims from specific domains, such as science (Wadden et al., 2020), climate (Diggelmann et al., 2020), and public health (Kotonya and Toni, 2020b). Alternative forms of sentence-level inputs, such as answers from question answering forums, have also been considered (Mihaylova et al., 2018; Zhang et al., 2020). There have been approaches that consider a passage (Mihalcea and Strapparava, 2009; Pérez-Rosas et al., 2018) or an entire article (Horne et al., 2018; Santia and Williams, 2018; Shu et al., 2020) as input. However, the implicit assumption that every claim in it is either factually correct or in-

correct is problematic, and thus rarely practised by human fact-checkers (Uscinski and Butler, 2013).

In order to better control the complexity of the task, efforts listed in Table 3 created claims artificially. Thorne et al. (2018a) had annotators mutate sentences from Wikipedia articles to create claims. Following the same approach, Khouja (2020) and Nørregaard and Derczynski (2021) constructed Arabic and Danish datasets respectively. Another frequently considered option is subject-predicate-object triples, e.g. (*London*, *city_in*, *UK*). The popularity of triples as input stems from the fact that they facilitate fact-checking against knowledge bases (Ciampaglia et al., 2015; Shi and Weninger, 2016; Shiralkar et al., 2017; Kim and Choi, 2020) such as DBpedia (Auer et al., 2007), SemMedDB (Kilicoglu et al., 2012), and KBox (Nam et al., 2018). However, such approaches implicitly assume the non-trivial conversion of text into triples.

3.2 Evidence

A popular type of evidence often considered is metadata, such as publication date, sources, user profiles, etc. However, while it offers information complementary to textual sources or structural knowledge which is useful when the latter are unavailable (Wang, 2017; Potthast et al., 2018), it

Dataset	Input	#Inputs	Evidence	Verdict	Sources	Lang
KLinker (Ciampaglia et al., 2015)	Triple	10,000	KG	2 Classes	Google/Wiki	En
PredPath (Shi and Weninger, 2016)	Triple	3,559	KG	2 Classes	Google/Wiki	En
KStream (Shiralkar et al., 2017)	Triple	18,431	KG	2 Classes	Google/Wiki	En
UFC (Kim and Choi, 2020)	Triple	1,759	KG	2 Classes	Wiki	En
LieDetect (Mihalcea and Strapparava, 2009)	Passage	600	✗	2 Classes	News	En
FakeNewsAMT (Pérez-Rosas et al., 2018)	Passage	680	✗	2 Classes	News	En
FEVER (Thorne et al., 2018a)	Statement	185,445	Text	3 Classes	Wiki	En
HOVER (Jiang et al., 2020)	Statement	26,171	Text	2 Classes	Wiki	En
WikiFactCheck (Sathe et al., 2020)	Statement	124,821	Text	2 Classes	Wiki	En
VitaminC (Schuster et al., 2021)	Statement	488,904	Text	3 Classes	Wiki	En
TabFact (Chen et al., 2020)	Statement	92,283	Table	2 Classes	Wiki	En
InfoTabs (Gupta et al., 2020)	Statement	23,738	Table	3 Classes	Wiki	En
Sem-Tab-Fact (Wang et al., 2021)	Statement	5,715	Table	3 Classes	Wiki	En
FEVEROUS (Aly et al., 2021)	Statement	87,026	Text/Table	3 Classes	Wiki	En
ANT (Khouja, 2020)	Statement	4,547	✗	3 Classes	News	Ar
DanFEVER (Nørregaard and Derczynski, 2021)	Statement	6,407	Text	3 Classes	Wiki	Da

Table 3: Summary of factual verification datasets with artificial inputs. Google denotes Google Relation Extraction Corpora, and WSDM means the WSDM Cup 2017 Triple Scoring challenge.

does not provide evidence grounding the claim.

Textual sources, such as news articles, academic papers, and Wikipedia documents, are one of the most commonly used types of evidence for fact-checking. Ferreira and Vlachos (2016) used the headlines of selected news articles, and Pomerleau and Rao (2017) used the entire articles instead as the evidence for the same claims. Instead of using news articles, Alhindi et al. (2018) and Hanselowski et al. (2019) extracted summaries accompanying fact-checking articles about the claims as evidence. Documents from specialized domains such as science and public health have also been considered (Wadden et al., 2020; Kotonya and Toni, 2020b; Zhang et al., 2020).

The aforementioned works assume that evidence is given for every claim, which is not conducive to developing systems that need to retrieve evidence from a large knowledge source. Therefore, Thorne et al. (2018a) and Jiang et al. (2020) considered Wikipedia as the source of evidence and annotated the sentences supporting or refuting each claim. Schuster et al. (2021) constructed VitaminC based on factual revisions to Wikipedia, in which evidence pairs are nearly identical in language and content, with the exception that one supports a claim while the other does not. However, these efforts restricted world knowledge to a single source (Wikipedia), ignoring the challenge of retrieving evidence from heterogeneous sources on the web. To address this, other works (Popat et al., 2016; Baly et al., 2018; Augenstein et al., 2019) retrieved evidence from the Internet, but the search results were not annotated. Thus, it is possible that irrelevant information is present in the

evidence, while information that is necessary for verification is missing.

Though the majority of studies focus on unstructured evidence (i.e. textual sources), structured knowledge has also been used. For example, the truthfulness of a claim expressed as an edge in a knowledge base (e.g. DBpedia) can be predicted by the graph topology (Ciampaglia et al., 2015; Shi and Weninger, 2016; Shiralkar et al., 2017). However, while graph topology can be an indicator of plausibility, it does not provide conclusive evidence. A claim that is not represented by a path in the graph, or that is represented by an unlikely path, is not necessarily false. The knowledge base approach assumes that true facts relevant to the claim are present in the graph; but given the incompleteness of even the largest knowledge bases, this is not realistic (Bordes et al., 2013; Socher et al., 2013).

Another type of structural knowledge is semi-structured data (e.g. tables), which is ubiquitous thanks to its ability to convey important information in a concise and flexible manner. Early work by Vlachos and Riedel (2015) used tables extracted from Freebase (Bollacker et al., 2008) to verify claims retrieved from the web about statistics of countries such as population, inflation, etc. Chen et al. (2020) and Gupta et al. (2020) studied fact-checking textual claims against tables and info-boxes from Wikipedia. Wang et al. (2021) extracted tables from scientific articles and required evidence selection in the form of cells selected from tables. Aly et al. (2021) further considered both text and table for factual verification, while explicitly requiring the retrieval of evidence.

3.3 Verdict & Justification

The verdict in early efforts (Bachenko et al., 2008; Mihalcea and Strapparava, 2009) is a binary label, i.e. *true/false*. However, fact-checkers usually employ multi-class labels to represent degrees of truthfulness (e.g. *true*, *mostly-true*, *mixture*, etc),⁴ which were considered by Vlachos and Riedel (2014) and Wang (2017). Recently, Augenstein et al. (2019) collected claims from different sources, where the number of labels vary greatly, ranging from 2 to 27. Due to the difficulty of mapping veracity labels onto the same scale, they didn’t attempt to harmonize them across sources. On the other hand, other efforts (Hanselowski et al., 2019; Kotonya and Toni, 2020b; Gupta and Srikumar, 2021) performed normalization by post-processing the labels based on rules to simplify the veracity label. For example, Hanselowski et al. (2019) mapped *mixture*, *unproven*, and *undetermined* onto *not enough information*.

Unlike prior datasets that only required outputting verdicts, FEVER (Thorne et al., 2018a) expected the output to contain both sentences forming the evidence and a label (e.g. support, refute, not enough information). Later datasets with both natural (Hanselowski et al., 2019; Wadden et al., 2020) and artificial claims (Jiang et al., 2020; Schuster et al., 2021) also adopted this scheme, where the output expected is a combination of multi-class labels and extracted evidence.

Most existing datasets do not contain textual explanations provided by journalists as justification for verdicts. Alhindi et al. (2018) extended the Liar dataset with summaries extracted from fact-checking articles. While originally intended as an auxiliary task to improve claim verification, these justifications have been used as explanations (Atanasova et al., 2020b). Recently, Kotonya and Toni (2020b) constructed the first dataset which explicitly includes gold explanations. These consist of fact-checking articles and other news items, which can be used to train natural language generation models to provide post-hoc justifications for the verdicts. However, using fact-checking articles is not realistic, as they are not available during inference, which makes the trained system unable to provide justifications based on retrieved evidence.

4 Modelling Strategies

We now turn to surveying modelling strategies for the various components of our framework. The most common approach is to build separate models for each component and apply them in pipeline fashion. Nevertheless, joint approaches have also been developed, either through end-to-end learning or by modelling the joint output distributions of multiple components.

4.1 Claim Detection

Claim detection is typically framed as a classification task, where models predict whether claims are checkable or check-worthy. This is challenging, especially in the case of check-worthiness: rumourous and non-rumourous information is often difficult to distinguish, and the volume of claims analysed in real-world scenarios – e.g. all posts published to a social network every day – prohibits the retrieval and use of evidence. Early systems employed supervised classifiers with feature engineering, relying on surface features like Reddit karma and up-votes (Aker et al., 2017), Twitter-specific types (Enayet and El-Beltagy, 2017), named entities and verbal forms in political transcripts (Zuo et al., 2018), or lexical and syntactic features (Zhou et al., 2020).

Neural network approaches based on sequence- or graph-modelling have recently become popular, as they allow models to use the context of surrounding social media activity to inform decisions. This can be highly beneficial, as the ways in which information is discussed and shared by users are strong indicators of rumourousness (Zubiaga et al., 2016). Kochkina et al. (2017) employed an LSTM (Hochreiter and Schmidhuber, 1997) to model branches of tweets, Ma et al. (2018) used Tree-LSTMs (Tai et al., 2015) to directly encode the structure of threads, and Guo et al. (2018) modelled the hierarchy by using attention networks. Recent work explored fusing more domain-specific features into neural models (Zhang et al., 2021). Another popular approach is to use Graph Neural Networks (Kipf and Welling, 2017) to model the propagation behaviour of a potentially rumourous claim (Monti et al., 2019; Li et al., 2020; Yang et al., 2020a).

Some works tackle claim detection and claim verification jointly, labelling potential claims as *true rumours*, *false rumours*, or *non-rumours* (Buntain and Golbeck, 2017; Ma et al.,

⁴www.snopes.com/fact-check-ratings

2018). This allows systems to exploit specific features useful for both tasks, such as the different spreading patterns of false and true rumours (Zubiaga et al., 2016). Veracity predictions made by such systems are to be considered preliminary, as they are made without evidence.

4.2 Evidence Retrieval & Claim Verification

As mentioned in Section 2, evidence retrieval and claim verification are commonly addressed together. Systems mostly operate as a pipeline consisting of an evidence retrieval module and a verification module (Thorne et al., 2018b), but there are exceptions where these two modules are trained jointly (Yin and Roth, 2018).

Claim verification can be seen as a form of Recognizing Textual Entailment (RTE; Dagan et al. 2010; Bowman et al. 2015), predicting whether the evidence supports or refutes the claim. Typical retrieval strategies include commercial search APIs, Lucene indices, entity linking, or ranking functions like dot-products of TF-IDF vectors (Thorne et al., 2018b). Recently, dense retrievers employing learned representations and fast dot-product indexing (Johnson et al., 2017) have shown strong performance (Lewis et al., 2020; Maillard et al., 2021). To improve precision, more complex models – for example stance detection systems – can be deployed as second, fine-grained filters to re-rank retrieved evidence (Thorne et al., 2018b; Nie et al., 2019b,a; Hanselowski et al., 2019). Similarly, evidence can be re-ranked implicitly during verification in late-fusion systems (Ma et al., 2019; Schlichtkrull et al., 2021). An alternative approach was proposed by Fan et al. (2020), who retrieved evidence using question generation and question answering via search engine results. Some work avoids retrieval by making a *closed-domain assumption* and evaluating in a setting where appropriate evidence has already been found (Ferreira and Vlachos, 2016; Chen et al., 2020; Zhong et al., 2020a; Yang et al., 2020b; Eisenschlos et al., 2020); this, however, is unrealistic. Finally, Allein et al. (2021) took into account the timestamp of the evidence in order to improve veracity prediction accuracy.

If only a single evidence document is retrieved, verification can be directly modelled as RTE. However, both real-world claims (Augenstein et al., 2019; Hanselowski et al., 2019; Kotonya and Toni, 2020b), as well as those created for research

purposes (Thorne et al., 2018a; Jiang et al., 2020; Schuster et al., 2021) often require reasoning over and combining multiple pieces of evidence. A simple approach is to treat multiple pieces of evidence as one by concatenating them into a single string (Luken et al., 2018; Nie et al., 2019a), and then employ a textual entailment model to infer whether the evidence supports or refutes the claim. More recent systems employ specialized components to aggregate multiple pieces of evidence. This allows the verification of more complex claims where several pieces of information must be combined, and addresses the case where the retrieval module returns several highly-related documents all of which *could* (but might not) contain the right evidence (Yoneda et al., 2018; Zhou et al., 2019; Ma et al., 2019; Liu et al., 2020; Zhong et al., 2020b; Schlichtkrull et al., 2021).

Some early work does not include evidence retrieval at all, performing verification purely on the basis of surface forms and metadata (Wang, 2017; Rashkin et al., 2017; Dungs et al., 2018). Recently Lee et al. (2020) considered using the information stored in the weights of a large pre-trained language model – BERT (Devlin et al., 2019) – as the only source of evidence, as it has been shown competitive in knowledge base completion (Petroni et al., 2019). Without explicitly considering evidence such approaches are likely to propagate biases learned during training, and render justification production impossible (Lee et al., 2021; Pan et al., 2021).

4.3 Justification Production

Approaches for justification production can be separated into three categories, which we examine along the three dimensions discussed in Section 2.4 – readability, plausibility, and faithfulness. First, some models include components that can be analysed as justifications by human experts, primarily attention modules. Popat et al. (2018) selected evidence tokens that have higher attention weights as explanations. Similarly, co-attention (Shu et al., 2019; Lu and Li, 2020) and self-attention (Yang et al., 2019) were used to highlight the salient excerpts from the evidence. Wu et al. (2020b) further combined decision trees and attention weights to explain which tokens were salient, and how they influenced predictions. Recent studies have shown the use of attention as explanation to be problematic. Some tokens with

high attention scores can be removed without affecting predictions, while some tokens with low (non-zero) scores turn out to be crucial (Jain and Wallace, 2019; Serrano and Smith, 2019; Pruthi et al., 2020). Explanations provided by attention may therefore not be sufficiently faithful. Furthermore, as they are difficult for non-experts and/or those not well-versed in the architecture of the model to grasp, they lack readability.

Another approach is to construct decision-making processes that can be fully grasped by human experts. Rule-based methods use Horn rules and knowledge bases to mine explanations (Gad-Elrab et al., 2019; Ahmadi et al., 2019), which can be directly understood and verified. These rules are mined from a pre-constructed knowledge base, such as DBpedia (Auer et al., 2007). This limits what can be fact-checked to claims which are representable as triples, and to information present in the (often manually curated) knowledge base.

Finally, some recent work has focused on building models which – like human experts – can generate textual explanations for their decisions. Atanasova et al. (2020b) used an extractive approach to generate summaries, while Kotonya and Toni (2020b) adopted the abstractive approach. A potential issue is that such models can generate explanations that do not represent their actual veracity prediction process, but which are nevertheless plausible with respect to the decision. This is especially an issue with abstractive models, where hallucinations can produce very misleading justifications (Maynez et al., 2020). Also, the model of Atanasova et al. (2020b) assumes fact-checking articles provided as input during inference, which is unrealistic.

5 Related Tasks

Misinformation and Disinformation Misinformation is defined as constituting a claim that contradicts or distorts common understandings of verifiable facts (Guess and Lyons, 2020). On the other hand, disinformation is defined as the subset of misinformation that is deliberately propagated. This is a question of intent: disinformation is meant to deceive, while misinformation may be inadvertent or unintentional (Tucker et al., 2018). Fact-checking can help detect misinformation, but not distinguish it from disinformation. A recent survey (Alam et al., 2021) proposed to integrate both factuality and harmfulness into a frame-

work for multi-modal disinformation detection. Although misinformation and conspiracy theories overlap conceptually, conspiracy theories do not hinge exclusively on the truth value of the claims being made, as they are sometimes proved to be true (Sunstein and Vermeule, 2009). A related problem is *propaganda detection*, which overlaps with disinformation detection, but also includes identifying particular techniques such as appeals to emotion, logical fallacies, whataboutery, or cherry-picking (Da San Martino et al., 2020b).

Propaganda and the deliberate or accidental dissemination of misleading information has been studied extensively. Jowett and O’Donnell (2019) address the subject from a communications perspective, Taylor (2003) provides a historical approach, and Goldman and O’Connor (2021) tackle the related subject of epistemology and trust in social settings from a philosophical perspective. For fact-checking and the identification of misinformation by journalists, we direct the reader to Silverman (2014) and Borel (2016).

Detecting Previously Fact-checked Claims

While in this survey we focus on methods for verifying claims by finding the evidence rather than relying on previously conducted fact checks, misleading claims are often repeated (Hassan et al., 2017); thus it is useful to detect whether a claim has already been fact-checked. Shaar et al. (2020) formulated this task recently by as ranking, and constructed two datasets. The social media version of the task then featured at the shared task CheckThat! (Barrón-Cedeño et al., 2020; Nakov et al., 2021b). This task was also explored by Vo and Lee (2020) from a multi-modal perspective, where claims about images were matched against previously fact-checked claims. More recently, Sheng et al. (2021) and Kazemi et al. (2021) constructed datasets for this task in languages beyond English. Hossain et al. (2020) detected misinformation by adopting a similar strategy. If a tweet was matched to any known COVID-19 related misconceptions, then it would be classified as misinformative. Matching claims against previously verified ones is a simpler task that can often be reduced to sentence-level similarity (Shaar et al., 2020), which is well studied in the context of textual entailment. Nevertheless, new claims and evidence emerge regularly. Previous fact-checks can be useful, but they can become outdated and potentially misleading over time.

6 Research Challenges

Choice of Labels The use of fine-grained labels by fact-checking organisations has recently come under criticism (Uscinski and Butler, 2013). In-between labels like “*mostly true*” often represent “meta-ratings” for composite claims consisting of multiple elementary claims of different veracity. For example, a politician might claim improvements to unemployment and productivity; if one part is true and the other false, a fact-checker might label the full statement “*half true*”. Noisy labels resulting from composite claims could be avoided by intervening at the dataset creation stage to manually split such claims, or by learning to do so automatically. The separation of claims into *truth* and *falsehood* can be too simplistic, as true claims can still mislead. Examples include cherry-picking, where evidence is chosen to suggest a misleading *trend* (Asudeh et al., 2020), and technical truth, where true information is presented in a way that misleads (e.g. “*I have never lost a game of chess*” is also true if the speaker has never played chess). A major challenge is integrating analysis of such claims into the existing frameworks. This could involve new labels identifying specific forms of deception, as is done in propaganda detection (Da San Martino et al., 2020a), or a greater focus on producing justifications to show *why* claims are misleading (Atanasova et al., 2020b; Kotonya and Toni, 2020b).

Sources & Subjectivity Not all information is equally trustworthy, and sometimes trustworthy sources contradict each other. This challenges the assumptions made by most current fact-checking research relying on a single source considered authoritative, such as Wikipedia. Methods must be developed to address the presence of disagreeing or untrustworthy evidence. Recent work proposed integrating credibility assessment as a part of the fact-checking task (Wu et al., 2020a). This could be done for example by assessing the agreement between evidence sources, or by assessing the degree to which sources cohere with known facts (Li et al., 2015; Dong et al., 2015; Zhang et al., 2019). Similarly, check-worthiness is a subjective concept varying along axes including target audience, recency, and geography. One solution is to focus solely on objective checkability (Konstantinovskiy et al., 2021). However, the practical limitations of fact-checking (e.g. the deadlines of journalists

and the time-constraints of media consumers) often force the use of a triage system (Borel, 2016). This can introduce biases regardless of the intentions of journalists and system-developers to use objective criteria (Uscinski and Butler, 2013; Uscinski, 2015). Addressing this challenge will require the development of systems allowing for real-time interaction with users to take into account their evolving needs.

Dataset Artefacts & Biases Synthetic datasets constructed through crowd-sourcing are common (Zeichner et al., 2012; Hermann et al., 2015; Williams et al., 2018). It has been shown that models tend to rely on biases in these datasets, without learning the underlying task (Gururangan et al., 2018; Poliak et al., 2018; McCoy et al., 2019). For fact-checking, Schuster et al. (2019) showed that the predictions of models trained on FEVER (Thorne et al., 2018a) were largely driven by indicative claim words. The FEVER 2.0 shared task explored how to generate adversarial claims and build systems resilient to such attacks (Thorne et al., 2019). Alleviating such biases and increasing the robustness to adversarial examples remains an open question. Potential solutions include leveraging better modelling approaches (Utama et al., 2020a,b; Karimi Mahabadi et al., 2020; Thorne and Vlachos, 2021), collecting data by adversarial games (Eisenschlos et al., 2021), or context-sensitive inference (Schuster et al., 2021).

Multimodality Information (either in claims or evidence) can be conveyed through multiple modalities such as text, tables, images, audio, or video. Though the majority of existing works have focused on text, some efforts also investigated how to incorporate multimodal information, including claim detection with misleading images (Zhang et al., 2018), propaganda detection over mixed images and text (Dimitrov et al., 2021), and claim verification for images (Zlatkova et al., 2019; Nakamura et al., 2020). Monti et al. (2019) argued that rumours should be seen as signals propagating through a social network. Rumour detection is therefore inherently multimodal, requiring analysis of both graph structure and text. Available multimodal corpora are either small in size (Zhang et al., 2018; Zlatkova et al., 2019) or constructed based on distant supervision (Nakamura et al., 2020). The construction of large-scale annotated datasets paired with evidence beyond

metadata will facilitate the development of multimodal fact-checking systems.

Multilinguality Claims can occur in multiple languages, often different from the one(s) evidence is available in, calling for multilingual fact-checking systems. While misinformation spans both geographic and linguistic boundaries, most work in the field has focused on English. A possible approach for multilingual verification is to use translation systems for existing methods (Dementieva and Panchenko, 2020), but relevant datasets in more languages are necessary for testing multilingual models’ performance within each language, and ideally also for training. Currently, there exist a handful of datasets for factual verification in languages other than English (Baly et al., 2018; Lillie et al., 2019; Khouja, 2020; Shahi and Nandini, 2020; Nørregaard and Derczynski, 2021), but they do not offer a cross-lingual setup. More recently, Gupta and Srikanth (2021) introduced a multilingual dataset covering 25 languages, but found that adding training data from other languages did not improve performance. How to effectively align, coordinate, and leverage resources from different languages remains an open question. One promising direction is to distill knowledge from high-resource to low-resource languages (Kazemi et al., 2021).

Faithfulness A significant unaddressed challenge in justification production is faithfulness. As we discuss in Section 4.3, some justifications – such as those generated abstractively (Maynez et al., 2020) – may not be faithful. This can be highly problematic, especially if these justifications are used to convince users of the validity of model predictions (Lertvittayakumjorn and Toni, 2019). Faithfulness is difficult to evaluate for, as human evaluators and human-produced gold standards often struggle to separate highly plausible, unfaithful explanations from faithful ones (Jacovi and Goldberg, 2020). In the model interpretability domain, several recent papers have introduced strategies for testing or guaranteeing faithfulness. These include introducing formal criteria which models should uphold (Yu et al., 2019), measuring the accuracy of predictions after removing some or all of the predicted non-salient input elements (Yeh et al., 2019; DeYoung et al., 2020; Atanasova et al., 2020a), or disproving the faithfulness of techniques by counterexample (Jain and

Wallace, 2019; Wiegreffe and Pinter, 2019). Further work is needed to develop such techniques for justification production.

From Debunking to Early Intervention and Prebunking

The prevailing application of automated fact-checking is to discover and intervene against circulating misinformation, also referred to as debunking. Efforts have been made to respond quickly after the appearance of a piece of misinformation (Monti et al., 2019), but common to all approaches is that intervention takes place *reactively* after misinformation has already been introduced to the public. NLP technology could also be leveraged in *proactive* strategies. Prior work has employed network analysis and similar techniques to identify key actors for intervention in social networks (Farajtabar et al., 2017); using NLP, such techniques could be extended to take into account the information shared by these actors, in addition to graph-based features (Nakov, 2020; Mu and Aletras, 2020). Another direction is to disseminate countermessaging before misinformation can spread widely; this is also known as *pre-bunking*, and has been shown to be more effective than post-hoc debunking (van der Linden et al., 2017; Roozenbeek et al., 2020; Lewandowsky and van der Linden, 2021). NLP could play a crucial role both in early detection and in the creation of relevant countermessaging. Finally, training people to *create* misinformation has been shown to increase resistance towards false claims (Roozenbeek and van der Linden, 2019). NLP could be used to facilitate this process, or to provide an adversarial opponent for gamifying the creation of misinformation. This could be seen as a form of dialogue agent to educate users, however there are as of yet no resources for the development of such systems.

7 Conclusion

We have reviewed and evaluated current automated fact-checking research by unifying the task formulations and methodologies across different research efforts into one framework comprising claim detection, evidence retrieval, verdict prediction, and justification production. Based on the proposed framework, we have provided an extensive overview of the existing datasets and modelling strategies. Finally, we have identified vital challenges for future research to address.

Acknowledgements

Zhijiang Guo, Michael Schlichtkrull and Andreas Vlachos are supported by the ERC grant AVeriTec (GA 865958), The latter is further supported by the EU H2020 grant MONITIO (GA 965576). The authors would like to thank Rami Aly, Christos Christodoulopoulos, Nedjma Ousidhoum, and James Thorne for useful comments and suggestions.

References

- Bill Adair, Chengkai Li, Jun Yang, and Cong Yu. 2017. Progress toward “the holy grail”: The continued quest to automate fact-checking. In *Proceedings of the 2017 Computation+Journalism Symposium*.
- Naser Ahmadi, Joohyung Lee, Paolo Papotti, and Mohammed Saeed. 2019. Explainable fact checking with probabilistic answer set programming. In *Proceedings of the 2019 Truth and Trust Online Conference (TTO 2019), London, UK, October 4-5, 2019*.
- Ahmet Aker, Leon Derczynski, and Kalina Bontcheva. 2017. Simple open stance classification for rumour analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 31–39, Varna, Bulgaria. INCOMA Ltd.
- Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2021. A survey on multi-modal disinformation detection. *arXiv preprint arXiv:2103.12541*.
- Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.
- Liesbeth Allein, Isabelle Augenstein, and Marie-Francine Moens. 2021. Time-Aware Evidence Ranking for Fact-Checking. *Web Semantics*.
- Rami Aly, Zhijiang Guo, M. Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, O. Cocarascu, and Arpit Mittal. 2021. FEVEROUS: Fact Extraction and VERification over unstructured and structured information. *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*.
- Abolfazl Asudeh, H. V. Jagadish, You (Will) Wu, and Cong Yu. 2020. On detecting cherry-picked trendlines. *Proceedings of the VLDB Endowment*, 13(6):939–952.
- Pepa Atanasova, Lluís Márquez, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. 2018. Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. task 1: Check-worthiness. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, volume 2125 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020a. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.
- Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020b. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.
- Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. 2007. DBpedia: A nucleus for a web of open data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer.
- Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen.

2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Joan Bachenko, Eileen Fitzpatrick, and Michael Schonwetter. 2008. Verification and implementation of language-based deception indicators in civil and criminal narratives. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 41–48, Manchester, UK. Coling 2008 Organizing Committee.

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Márquez, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 21–27, New Orleans, Louisiana. Association for Computational Linguistics.

Aviv Barnoy and Zvi Reich. 2019. The When, Why, How and So-What of Verifications. *Journalism Studies*, 20(16):2312–2330.

Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, Shaden Shaar, and Zien Sheikh Ali. 2020. Overview of CheckThat! 2020: Automatic identification and verification of claims in social media. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22-25, 2020, Proceedings*, volume 12260 of *Lecture Notes in Computer Science*, pages 215–236. Springer.

Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Márquez, Pepa Atanasova, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. 2018.

Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. task 2: Factuality. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, volume 2125 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Kurt D. Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008*, pages 1247–1250. ACM.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795.

Brooke Borel. 2016. *The Chicago Guide to Fact-checking*. University of Chicago Press.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on*

*Neural Information Processing Systems 2020,
NeurIPS 2020, December 6-12, 2020, virtual.*

Cody Buntain and Jennifer Golbeck. 2017. [Automatically identifying fake news in popular twitter threads](#). In *2017 IEEE International Conference on Smart Cloud (SmartCloud)*, pages 208–215. IEEE.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. [TabFact: A large-scale dataset for table-based fact verification](#). In *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia.

Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PloS one*, 10(6):e0128193.

Sarah Cohen, Chengkai Li, Jun Yang, and Cong Yu. 2011. [Computational journalism: A call to arms to database researchers](#). In *CIDR 2011, Fifth Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 9-12, 2011, Online Proceedings*, pages 148–151. www.cidrdb.org.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020a. [SemEval-2020 task 11: Detection of propaganda techniques in news articles](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020b. [A survey on computational propaganda detection](#). In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 4826–4832. ijcai.org.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2010. [Recognizing textual entailment: Rational, evaluation and approaches](#). *Natural Language Engineering*, 16(1):105.

Daryna Dementieva and A. Panchenko. 2020. [Fake news detection using multilingual evidence](#). In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 775–776.

Leon Derczynski, Kalina Bontcheva, Maria Likata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. [SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours](#). In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. [ERASER: A benchmark to evaluate rationalized NLP models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Thomas Diggelmann, Jordan L. Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. [CLIMATE-FEVER: A dataset for verification of real-world climate claims](#). *CoRR*, abs/2012.00614.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. [Detecting propaganda techniques in memes](#). 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)*, pages 6603–6617, Online. Association for Computational Linguistics.

- Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Luges, Shaohua Sun, and Wei Zhang. 2015. **Knowledge-based trust: Estimating the trustworthiness of web sources.** *Proceedings of the VLDB Endowment*, 8(9):938–949.
- Sebastian Dungs, Ahmet Aker, Norbert Fuhr, and Kalina Bontcheva. 2018. **Can rumour stance alone predict veracity?** In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3360–3370, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Julian Eisenschlos, Bhuwan Dhingra, Jannis Builan, Benjamin Börschinger, and Jordan Boyd-Graber. 2021. **Fool Me Twice: Entailment from Wikipedia gamification.** In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 352–365, Online. Association for Computational Linguistics.
- Julian Eisenschlos, Syrine Krichene, and Thomas Müller. 2020. **Understanding tables with intermediate pre-training.** In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 281–296, Online. Association for Computational Linguistics.
- Omar Enayet and Samhaa R. El-Beltagy. 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)*, pages 470–474, Vancouver, Canada. Association for Computational Linguistics.
- Angela Fan, Aleksandra Piktus, Fabio Petroni, Guillaume Wenzek, Marzieh Saeidi, Andreas Vlachos, Antoine Bordes, and Sebastian Riedel. 2020. **Generating fact checking briefs.** In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7147–7161, Online. Association for Computational Linguistics.
- Mehrdad Farajtabar, Jiachen Yang, Xiaojing Ye, Huan Xu, Rakshit Trivedi, Elias Khalil, Shuang Li, Le Song, and Hongyuan Zha. 2017. **Fake news mitigation via point process based intervention.** In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1097–1106. PMLR.
- William Ferreira and Andreas Vlachos. 2016. **Emergent: a novel data-set for stance classification.** In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168, San Diego, California. Association for Computational Linguistics.
- Terry Flew, Christina Spurgeon, Anna Daniel, and Adam Swift. 2010. The promise of computational journalism. *Journalism Practice*, 6:157–171.
- Mohamed H. Gad-Elrab, Daria Stepanova, Jacopo Urbani, and Gerhard Weikum. 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, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019*, pages 87–95. ACM.
- Pepa Gencheva, Preslav Nakov, Lluís Márquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. **A context-aware approach for detecting worth-checking claims in political debates.** In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 267–276, Varna, Bulgaria. INCOMA Ltd.
- Yigal Godler and Zvi Reich. 2017. Journalistic evidence: Cross-verification as a constituent of mediated knowledge. *Journalism*, 18(5):558–574.
- Alvin Goldman and Cailin O’Connor. 2021. Social Epistemology. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Spring 2021 edition. Metaphysics Research Lab, Stanford University.
- Genevieve Gorrell, Ahmet Aker, Kalina Bontcheva, Leon Derczynski, Elena Kochkina, Maria Liakata, and Arkaitz Zubia. 2019. **SemEval-2019 task 7: RumourEval, determining rumour veracity and support for**

- rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019*, pages 845–854. Association for Computational Linguistics.
- Lucas Graves. 2018. Understanding the promise and limits of automated fact-checking. *Reuters Institute for the Study of Journalism*.
- Maurício Gruppi, Benjamin D. Horne, and Sibel Adali. 2020. **NELA-GT-2019: A large multi-labelled news dataset for the study of misinformation in news articles.** *CoRR*, abs/2003.08444.
- Maurício Gruppi, Benjamin D. Horne, and Sibel Adali. 2021. **NELA-GT-2020: A large multi-labelled news dataset for the study of misinformation in news articles.** *CoRR*, abs/2102.04567.
- Andrew M. Guess and Benjamin A. Lyons. 2020. Misinformation, disinformation, and online propaganda. In Nathaniel Persily and Joshua A. Tucker, editors, *Social media and democracy: the state of the field, prospects for reform*, pages 10–33. Cambridge University Press.
- Han Guo, Juan Cao, Yazi Zhang, Junbo Guo, and Jintao Li. 2018. **Rumor detection with hierarchical social attention network.** In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, pages 943–951. ACM.
- Ashim Gupta and Vivek Srikumar. 2021. **X-Fact: A new benchmark dataset for multilingual fact checking.** In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682, Online. Association for Computational Linguistics.
- Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Srikumar. 2020. **INFOTABS: Inference on tables as semi-structured data.** In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2309–2324, Online. Association for Computational Linguistics.
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. **Annotation artifacts in natural language inference data.** In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.
- Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. **A richly annotated corpus for different tasks in automated fact-checking.** In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503, Hong Kong, China. Association for Computational Linguistics.
- Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. A survey on stance detection for mis- and disinformation identification. *ArXiv*, abs/2103.00242.
- Maram Hasanain, Reem Suwaileh, Tamer El-sayed, Alberto Barrón-Cedeño, and Preslav Nakov. 2019. **Overview of the CLEF-2019 CheckThat! lab: Automatic identification and verification of claims. task 2: Evidence and factuality.** In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019*, volume 2380 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. **Detecting check-worthy factual claims in presidential debates.** In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*, pages 1835–1838. ACM.
- Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. 2017. **ClaimBuster: The first-ever end-to-end fact-checking system.** *Proceedings of the VLDB Endowment*, 10(12):1945–1948.

- Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. *Teaching machines to read and comprehend*. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. *Long short-term memory*. *Neural Computation*, 9(8):1735–1780.
- Benjamin D. Horne, Sara Khedr, and Sibel Adali. 2018. *Sampling the news producers: A large news and feature data set for the study of the complex media landscape*. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, pages 518–527. AAAI Press.
- Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. *COVIDLies: Detecting COVID-19 misinformation on social media*. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.
- Md. Rafiqul Islam, Shaowu Liu, Xianzhi Wang, and Guandong Xu. 2020. *Deep learning for misinformation detection on online social networks: a survey and new perspectives*. *Soc. Netw. Anal. Min.*, 10(1):82.
- Alon Jacovi and Yoav Goldberg. 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*, pages 4198–4205, Online. Association for Computational Linguistics.
- Sarthak Jain and Byron C. Wallace. 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)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.
- Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. *HoVer: A dataset for many-hop fact extraction and claim verification*. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. *Billion-scale similarity search with gpus*. *CoRR*, abs/1702.08734.
- Garth S. Jowett and Victoria O’Donnell. 2019. *Propaganda & Persuasion*, 7th edition. SAGE Publications.
- Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. *End-to-end bias mitigation by modelling biases in corpora*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716, Online. Association for Computational Linguistics.
- Ashkan Kazemi, Kiran Garimella, Devin Gaffney, and Scott Hale. 2021. *Claim matching beyond English to scale global fact-checking*. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4504–4517, Online. Association for Computational Linguistics.
- Jude Khouja. 2020. *Stance prediction and claim verification: An Arabic perspective*. In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 8–17, Online. Association for Computational Linguistics.
- Halil Kilicoglu, Dongwook Shin, Marcelo Fiszman, Graciela Rosemblat, and Thomas C. Rindflesch. 2012. *SemMedDB: a pubmed-scale repository of biomedical semantic predications*. *Bioinform.*, 28(23):3158–3160.
- Jiseong Kim and Key-sun Choi. 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*, pages 1677–1686, Barcelona, Spain (Online). International Committee on Computational Linguistics.

- Thomas N. Kipf and Max Welling. 2017. [Semi-supervised classification with graph convolutional networks](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- Elena Kochkina, Maria Liakata, and Isabelle Augenstein. 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)*, pages 475–480, Vancouver, Canada. Association for Computational Linguistics.
- Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2021. Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digital Threats: Research and Practice*, 2(2):1–16.
- Neema Kotonya and Francesca Toni. 2020a. [Explainable automated fact-checking: A survey](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Neema Kotonya and Francesca Toni. 2020b. [Explainable automated fact-checking for public health claims](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.
- Dilek Küçük and Fazli Can. 2020. [Stance detection: A survey](#). *ACM Computing Surveys*, 53(1):12:1–12:37.
- David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science*, 359(6380):1094–1096.
- Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. 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*, pages 1971–1981, Online. Association for Computational Linguistics.
- Nayeon Lee, Belinda Z. Li, Sinong Wang, Wentau Yih, Hao Ma, and Madian Khabsa. 2020. [Language models as fact checkers?](#) In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 36–41, Online. Association for Computational Linguistics.
- Piyawat Lertvittayakumjorn and Francesca Toni. 2019. [Human-grounded evaluations of explanation methods for text classification](#). 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)*, pages 5195–5205, Hong Kong, China. Association for Computational Linguistics.
- Stephan Lewandowsky, Ullrich K.H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. 2012. [Misinformation and Its Correction: Continued Influence and Successful Debiasing](#). *Psychological Science in the Public Interest, Supplement*, 13(3):106–131.
- Stephan Lewandowsky and Sander van der Linden. 2021. [Countering misinformation and fake news through inoculation and prebunking](#). *European Review of Social Psychology*, 0(0):1–38.
- Justin Matthew Wren Lewis, Andy Williams, Robert Arthur Franklin, James Thomas, and Nicholas Alexander Mosdell. 2008. The quality and independence of british journalism. *Mediawise*.
- Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wentau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Jiawen Li, Yudianto Sujana, and Hung-Yu Kao. 2020. [Exploiting microblog conversation](#)

structures to detect rumors. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5420–5429, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2015. A survey on truth discovery. *SIGKDD Explor.*, 17(2):1–16.

Anders Edelbo Lillie, Emil Refsgaard Middelboe, and Leon Derczynski. 2019. Joint rumour stance and veracity prediction. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 208–221, Turku, Finland. Linköping University Electronic Press.

Sander van der Linden, Anthony Leiserowitz, Seth Rosenthal, and Edward Maibach. 2017. Inoculating the public against misinformation about climate change. *Global Challenges*, 1(2):1600008.

Zachary C. Lipton. 2018. The mythos of model interpretability. *Commun. ACM*, 61(10):36–43.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351, Online. Association for Computational Linguistics.

Yi-Ju Lu and Cheng-Te Li. 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*, pages 505–514, Online. Association for Computational Linguistics.

Jackson Lukens, Nanjiang Jiang, and Marie-Catherine de Marneffe. 2018. QED: A fact verification system for the FEVER shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 156–160, Brussels, Belgium. Association for Computational Linguistics.

Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. 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, pages 2561–2571, Florence, Italy. Association for Computational Linguistics.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9–15 July 2016*, pages 3818–3824. IJCAI/AAAI Press.

Jing Ma, Wei Gao, and Kam-Fai Wong. 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)*, pages 1980–1989, Melbourne, Australia. Association for Computational Linguistics.

Jean Maillard, Vladimir Karpukhin, Fabio Petroni, Wen-tau Yih, Barlas Oguz, Veselin Stoyanov, and Gargi Ghosh. 2021. Multi-task retrieval for knowledge-intensive tasks. 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)*, pages 1098–1111, Online. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Paul Mena. 2019. Principles and boundaries of fact-checking: Journalists’ perceptions. *Journalism Practice*, 13(6):657–672.

Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic

recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 309–312, Suntec, Singapore. Association for Computational Linguistics.

Tsvetomila Mihaylova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Mitra Mohtarami, Georgi Karadzhov, and James R. Glass. 2018. Fact checking in community forums. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5309–5316. AAAI Press.

Tanushree Mitra and Eric Gilbert. 2015. CRED-BANK: A large-scale social media corpus with associated credibility annotations. In *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, May 26-29, 2015*, pages 258–267. AAAI Press.

Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M. Bronstein. 2019. Fake news detection on social media using geometric deep learning. *CoRR*, abs/1902.06673.

Yida Mu and Nikolaos Aletras. 2020. Identifying twitter users who repost unreliable news sources with linguistic information. *PeerJ Computer Science*, 6.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 6149–6157. European Language Resources Association.

Ndapandula Nakashole and Tom M. Mitchell. 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)*, pages 1009–1019, Baltimore, Maryland. Association for Computational Linguistics.

Preslav Nakov. 2020. Can we spot the "fake news" before it was even written? *CoRR*, abs/2008.04374.

Preslav Nakov, David P. A. Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021a. Automated fact-checking for assisting human fact-checkers. *CoRR*, abs/2103.07769.

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, and Thomas Mandl. 2021b. The CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II*, volume 12657 of *Lecture Notes in Computer Science*, pages 639–649. Springer.

Sangha Nam, Eun-Kyung Kim, Jiho Kim, Yoosung Jung, Kijong Han, and Key-Sun Choi. 2018. A korean knowledge extraction system for enriching a kbox. In *COLING 2018, The 27th International Conference on Computational Linguistics: System Demonstrations, Santa Fe, New Mexico, August 20-26, 2018*, pages 20–24. Association for Computational Linguistics.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019a. Combining fact extraction and verification with neural semantic matching networks. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6859–6866. AAAI Press.

Yixin Nie, Songhe Wang, and Mohit Bansal. 2019b. 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), pages 2553–2566, Hong Kong, China. Association for Computational Linguistics.

Jeppe Nørregaard and Leon Derczynski. 2021. **DanFEVER: claim verification dataset for danish**. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics, NoDaLiDa 2021, Reykjavik, Iceland (Online), May 31 - June 2, 2021*, pages 422–428. Linköping University Electronic Press, Sweden.

Jeppe Nørregaard, Benjamin D. Horne, and Sibel Adali. 2019. **NELA-GT-2018: A large multi-labelled news dataset for the study of misinformation in news articles**. In *Proceedings of the Thirteenth International Conference on Web and Social Media, ICWSM 2019, Munich, Germany, June 11-14, 2019*, pages 630–638. AAAI Press.

Cathy O’Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, USA.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. **A survey on natural language processing for fake news detection**. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6086–6093, Marseille, France. European Language Resources Association.

Liangming Pan, Wenhui Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. **Zero-shot fact verification by claim generation**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 476–483. Association for Computational Linguistics.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. **Automatic detection of fake news**. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. **Language models as knowledge bases?** 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)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. **Hypothesis only baselines in natural language inference**. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Dean Pomerleau and Delip Rao. 2017. The fake news challenge: Exploring how artificial intelligence technologies could be leveraged to combat fake news. *Fake News Challenge*.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. **Credibility assessment of textual claims on the web**. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, pages 2173–2178. ACM.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 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*, pages 22–32, Brussels, Belgium. Association for Computational Linguistics.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. **A stylometric inquiry into hyperpartisan and fake news**. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia. Association for Computational Linguistics.

Danish Pruthi, Mansi Gupta, Bhuvan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020.

Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793, Online. Association for Computational Linguistics.

Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 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*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Miriam Redi, Besnik Fetahu, Jonathan T. Morgan, and Dario Taraborelli. 2019. Citation Needed: A taxonomy and algorithmic assessment of wikipedia’s verifiability. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 1567–1578. ACM.

Jon Roozenbeek and Sander van der Linden. 2019. The fake news game: actively inoculating against the risk of misinformation. *Journal of Risk Research*, 22(5):570–580.

Jon Roozenbeek, Sander van der Linden, and Thomas Nygren. 2020. Prebunking interventions based on the psychological theory of "inoculation" can reduce susceptibility to misinformation across cultures. *The Harvard Kennedy School Misinformation Review*, 1(2).

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-Fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th*

International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, pages 2116–2129. Association for Computational Linguistics.

Fatima K. Abu Salem, Roaa Al Feel, Shady Elbasuoni, Mohamad Jaber, and May Farah. 2019. FA-KES: A fake news dataset around the syrian war. In *Proceedings of the Thirteenth International Conference on Web and Social Media, ICWSM 2019, Munich, Germany, June 11-14, 2019*, pages 573–582. AAAI Press.

Giovanni C. Santia and Jake Ryland Williams. 2018. BuzzFace: A news veracity dataset with facebook user commentary and egos. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, pages 531–540. AAAI Press.

Aalok Sathe, Salar Ather, Tuan Manh Le, Nathan Perry, and Joonsuk Park. 2020. Automated fact-checking of claims from wikipedia. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 6874–6882. European Language Resources Association.

Michael Sejr Schlichtkrull, Vladimir Karpukhin, Barlas Oguz, Mike Lewis, Wen-tau Yih, and Sebastian Riedel. 2021. Joint verification and reranking for open fact checking over tables. 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)*, pages 6787–6799, Online. Association for Computational Linguistics.

Tal Schuster, Adam Fisch, and Regina Barzilay. 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*, pages 624–643, Online. Association for Computational Linguistics.

Tal Schuster, Roei Schuster, Darsh J. Shah, and Regina Barzilay. 2020. The limitations of stylometry for detecting machine-generated fake

- news. *Computational Linguistics*, 46(2):499–510.
- Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. 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)*, pages 3419–3425, Hong Kong, China. Association for Computational Linguistics.
- Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.
- Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618, Online. Association for Computational Linguistics.
- Gautam Kishore Shahi and Durgesh Nandini. 2020. FakeCovid – a multilingual cross-domain fact check news dataset for covid-19. In *Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media*.
- Qiang Sheng, Juan Cao, Xueyao Zhang, Xirong Li, and Lei Zhong. 2021. Article reranking by memory-enhanced key sentence matching for detecting previously fact-checked claims. 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)*, pages 5468–5481, Online. Association for Computational Linguistics.
- Baoxu Shi and Tim Weninger. 2016. Discriminative predicate path mining for fact checking in knowledge graphs. *Knowl. Based Syst.*, 104:123–133.
- Prashant Shiralkar, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. 2017. Finding streams in knowledge graphs to support fact checking. In *2017 IEEE International Conference on Data Mining, ICDM 2017, New Orleans, LA, USA, November 18-21, 2017*, pages 859–864. IEEE Computer Society.
- Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. dEFEND: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 395–405. ACM.
- Kai Shu, Deepak Mahadeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fake-NewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8(3):171–188.
- Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD Explor.*, 19(1):22–36.
- Fernando Cardoso Durier da Silva, Rafael Vieira, and Ana Cristina Bicharra Garcia. 2019. Can machines learn to detect fake news? A survey focused on social media. In *52nd Hawaii International Conference on System Sciences, HICSS 2019, Grand Wailea, Maui, Hawaii, USA, January 8-11, 2019*, pages 1–8. ScholarSpace.
- Craig Silverman. 2014. *Verification Handbook: An Ultimate Guideline on Digital Age Sourcing for Emergency Coverage*. European Journalism Centre.
- Richard Socher, Danqi Chen, Christopher D. Manning, and Andrew Y. Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 926–934.
- Cass R Sunstein and Adrian Vermeule. 2009. Conspiracy theories: Causes and cures. *Journal of Political Philosophy*, 17(2):202–227.
- Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic

- representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, Beijing, China. Association for Computational Linguistics.
- Philip M. Taylor. 2003. *Munitions of the mind: A history of propaganda from the ancient world to the present era*, 3rd edition. Manchester University Press.
- James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- James Thorne and Andreas Vlachos. 2021. Elastic weight consolidation for better bias inoculation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 957–964, Online. Association for Computational Linguistics.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. 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)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.
- James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.
- James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2019. The FEVER2.0 shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 1–6, Hong Kong, China. Association for Computational Linguistics.
- Joshua A. Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*.
- Joseph E. Uscinski. 2015. The epistemology of fact checking (is still naïve): Rejoinder to amazeen. *Critical Review*, 27(2):243–252.
- Joseph E. Uscinski and Ryden W. Butler. 2013. The epistemology of fact checking. *Critical Review*, 25(2):162–180.
- Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020a. Mind the trade-off: Debiasing NLU models without degrading the in-distribution performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8717–8729, Online. Association for Computational Linguistics.
- Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020b. Towards debiasing NLU models from unknown biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7597–7610, Online. Association for Computational Linguistics.
- Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.
- Andreas Vlachos and Sebastian Riedel. 2015. Identification and verification of simple claims about statistical properties. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2596–2601, Lisbon, Portugal. Association for Computational Linguistics.

- Nguyen Vo and Kyumin Lee. 2020. Where are the facts? searching for fact-checked information to alleviate the spread of fake news. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7717–7731, Online. Association for Computational Linguistics.
- Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 647–653, Vancouver, Canada. Association for Computational Linguistics.
- David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.
- Nancy Xin Ru Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. 2021. SemEval-2021 task 9: Fact verification and evidence finding for tabular data in scientific documents (SEM-TAB-FACTS). In *Proceedings of the 15th International Workshop on Semantic Evaluation, SemEval@ACL/IJCNLP 2021, Virtual Event / Bangkok, Thailand, August 5-6, 2021*, pages 317–326. Association for Computational Linguistics.
- William Yang Wang. 2017. “Liar, Liar Pants on Fire”: A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.
- Sarah Wiegreffe and Yuval Pinter. 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)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.
- Lianwei Wu, Yuan Rao, Xiong Yang, Wanzhen Wang, and Ambreen Nazir. 2020a. Evidence-aware hierarchical interactive attention networks for explainable claim verification. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 1388–1394. International Joint Conferences on Artificial Intelligence Organization. Main track.
- Lianwei Wu, Yuan Rao, Yongqiang Zhao, Hao Liang, and Ambreen Nazir. 2020b. DTCA: Decision tree-based co-attention networks for explainable claim verification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1024–1035, Online. Association for Computational Linguistics.
- Fan Yang, Shiva K. Pentyala, Sina Mohseni, Mengnan Du, Hao Yuan, Rhema Linder, Eric D. Ragan, Shuiwang Ji, and Xia (Ben) Hu. 2019. XFake: Explainable fake news detector with visualizations. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 3600–3604. ACM.
- Xiaoyu Yang, Yuefei Lyu, Tian Tian, Yifei Liu, Yudong Liu, and Xi Zhang. 2020a. Rumor detection on social media with graph structured adversarial learning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 1417–1423. ijcai.org.
- Xiaoyu Yang, Feng Nie, Yufei Feng, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2020b. Program enhanced fact verification with verbalization and graph attention network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7810–7825, Online. Association for Computational Linguistics.

- Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Sugala, David I. Inouye, and Pradeep Ravikumar. 2019. [On the \(in\)fidelity and sensitivity of explanations](#). In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 10965–10976.
- Wenpeng Yin and Dan Roth. 2018. [TwoWingOS: A two-wing optimization strategy for evidential claim verification](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 105–114. Association for Computational Linguistics.
- Takuma Yoneda, Jeff Mitchell, Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. [UCL machine reading group: Four factor framework for fact finding \(HexaF\)](#). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 97–102, Brussels, Belgium. Association for Computational Linguistics.
- Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. 2019. [Rethinking cooperative rationalization: Introspective extraction and complement control](#). 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)*, pages 4094–4103, Hong Kong, China. Association for Computational Linguistics.
- Naomi Zeichner, Jonathan Berant, and Ido Dagan. 2012. [Crowdsourcing inference-rule evaluation](#). In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 156–160, Jeju Island, Korea. Association for Computational Linguistics.
- Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. [Defending against neural fake news](#). In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9051–9062.
- Daniel Yue Zhang, Lanyu Shang, Biao Geng, Shuyue Lai, Ke Li, Hongmin Zhu, Md. Tanvir Al Amin, and Dong Wang. 2018. [Faux-Buster: A content-free fauxtography detector using social media comments](#). In *IEEE International Conference on Big Data, Big Data 2018, Seattle, WA, USA, December 10-13, 2018*, pages 891–900. IEEE.
- Wenxuan Zhang, Yang Deng, Jing Ma, and Wai Lam. 2020. [AnswerFact: Fact checking in product question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2407–2417, Online. Association for Computational Linguistics.
- Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. [Mining dual emotion for fake news detection](#). In *WWW ’21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 3465–3476. ACM / IW3C2.
- Yi Zhang, Zachary Ives, and Dan Roth. 2019. [Evidence-based trustworthiness](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 413–423, Florence, Italy. Association for Computational Linguistics.
- Wanjun Zhong, Duyu Tang, Zhangyin Feng, Nan Duan, Ming Zhou, Ming Gong, Linjun Shou, Daxin Jiang, Jiahai Wang, and Jian Yin. 2020a. [LogicalFactChecker: Leveraging logical operations for fact checking with graph module network](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6053–6065, Online. Association for Computational Linguistics.
- Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020b. [Reasoning over semantic-level graph for fact checking](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6170–6180, Online. Association for Computational Linguistics.
- Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. [GEAR: Graph-based evidence aggregating and reasoning for fact verification](#).

In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901, Florence, Italy. Association for Computational Linguistics.

Xinyi Zhou, Atishay Jain, Vir V. Phoha, and Reza Zafarani. 2020. **Fake news early detection: A theory-driven model.** *Digital Threats: Research and Practice*, 1(2).

Xinyi Zhou and Reza Zafarani. 2020. **A survey of fake news: Fundamental theories, detection methods, and opportunities.** *ACM Computing Surveys*, 53(5):109:1–109:40.

Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. 2019. **Fact-checking meets fauxtography: Verifying claims about images.** 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)*, pages 2099–2108, Hong Kong, China. Association for Computational Linguistics.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. **Detection and resolution of rumours in social media: A survey.** *ACM Computing Surveys*, 51(2):32:1–32:36.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. **Analysing how people orient to and spread rumours in social media by looking at conversational threads.** *PloS one*, 11(3):e0150989.

Chaoyuan Zuo, Ayla Karakas, and Ritwik Banerjee. 2018. **A hybrid recognition system for check-worthy claims using heuristics and supervised learning.** In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, volume 2125 of *CEUR Workshop Proceedings*. CEUR-WS.org.