Missing Counter-Evidence Renders NLP Fact-Checking Unrealistic for Misinformation

Max Glockner[1], **Yufang Hou**[2], **Iryna Gurevych**[1]
[1] Ubiquitous Knowledge Processing Lab (UKP Lab),
Department of Computer Science and Hessian Center for AI (hessian.AI),
Technical University of Darmstadt
[2] IBM Research Europe, Ireland
www.ukp.tu-darmstadt.de, yhou@ie.ibm.com

Abstract

Misinformation emerges in times of uncertainty when credible information is limited. This is challenging for NLP-based fact-checking as it relies on counter-evidence, which may not yet be available. Despite increasing interest in automatic fact-checking, it is still unclear if automated approaches can realistically refute harmful real-world misinformation. Here, we contrast and compare NLP fact-checking with how professional fact-checkers combat misinformation in the absence of counter-evidence. In our analysis, we show that, by design, existing NLP task definitions for fact-checking cannot refute misinformation as professional fact-checkers do for the majority of claims. We then define two requirements that the evidence in datasets must fulfill for realistic fact-checking: It must be (1) sufficient to refute the claim and (2) not leaked from existing fact-checking articles. We survey existing fact-checking datasets and find that all of them fail to satisfy both criteria. Finally, we perform experiments to demonstrate that models trained on a large-scale fact-checking dataset rely on leaked evidence, which makes them unsuitable in real-world scenarios. Taken together, we show that current NLP fact-checking cannot realistically combat real-world misinformation because it depends on unrealistic assumptions about counter-evidence in the data[1].

1 Introduction

According to van der Linden (2022), misinformation is "*false or misleading information masquerading as legitimate news, regardless of intent*". Misinformation is dangerous as it can directly impact human behavior and have harmful real-world consequences such as the Pizzagate shooting (Fisher et al., 2016), interfering in the 2016 democratic US election (Bovet and Makse, 2019), or the promotion of false COVID-19 cures (Aghababaeian et al., 2020). Surging misinformation during the COVID-19 pandemic, coined "infodemic" by WHO (Zarocostas, 2020), exemplifies the danger coming from misinformation. To combat misinformation, journalists from fact-checking organizations (e.g., PolitiFact or Snopes) conduct a laborious manual effort to verify claims based on possible harms and their prominence (Arnold, 2020). However, manual fact-checking cannot keep pace with the rate at which misinformation is posted and circulated. Automatic fact-checking has gained significant attention within the NLP community in recent years, with the goal of developing tools to assist fact-checkers in combating misinformation. For the past few years, NLP researchers have created a wide range of fact-checking datasets with claims from fact-checking organization websites (Vlachos and Riedel, 2014; Wang, 2017; Augenstein et al., 2019; Hanselowski et al., 2019; Ostrowski et al., 2021; Gupta and Srikumar, 2021; Khan et al., 2022). The fundamental goal of fact-checking is, given a *claim* made by a *claimant*, to find a collection of *evidence* and provide a *verdict* about the claim's veracity based

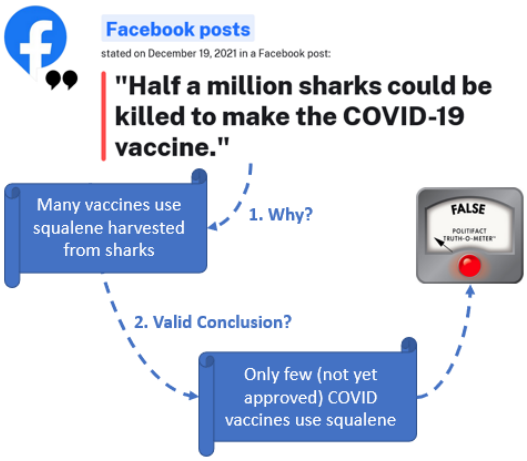

Figure 1: A false claim from PolitiFact. It is unlikely to find counter-evidence. Fact-checkers refute the claim by disproving why it was made.

[1]Code provided at https://github.com/UKPLab/emnlp2022-missing-counter-evidence

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 5916 - 5936
December 7-11, 2022

on the evidence. The underlying technique used by fact-checkers, and journalists in general, to assess the veracity of a claim is called *verification* (Silverman, 2016). In a comprehensive survey, Guo et al. (2022) proposed an NLP fact-checking framework (FCNLP) that aggregates existing (sub)tasks and approaches of automated fact-checking. FCNLP reflects current research trends on automatic fact-checking in NLP and divides the aforementioned process into *evidence retrieval*, *verdict prediction*, and *justification production*.

In this paper, we focus on harmful misinformation claims that satisfied the professional fact-checkers' selection criteria and refer to them as *real-world misinformation*. Our goal is to answer the following research question: **Can evidence-based NLP fact-checking approaches in FCNLP refute novel real-world misinformation?** FCNLP assumes a system has access to counter-evidence (e.g., through information retrieval) to refute a claim. Consider the false claim "*Telemundo is an English-language television network*" from FEVER (Thorne et al., 2018): A system following FCNLP must find counter-evidence contradicting the claim (i.e., *Telemundo is a Spanish company*) to refute the claim. This may require more complex reasoning over multiple documents. We contrast this example to the real-world false claim that "*Half a million sharks could be killed to make the COVID-19 vaccine*" (Figure 1). If true, credible sources would likely report this incident, providing supporting evidence. As it is not, before being fact-checked, there is *no refuting evidence* stating that COVID-19 vaccine production will not kill sharks. Only after *guaranteeing* that the claim relies on the false premise of COVID-19 vaccines using squalene (harvested from sharks), it can be refuted. After the claim's verification, fact-checkers publish reports explaining the verdict and thereby produce counter-evidence. Relying on counter-evidence leaked from such reports is unrealistic if a system is to be applied to new claims.

In this work, we identify gaps between current research on FCNLP and the verification process of professional fact-checkers. Via analysis from different perspectives, we argue that the assumption of the existence of counter-evidence in FCNLP is unrealistic and does not reflect real-world requirements. We hope our analysis sheds light on future research directions in automatic fact-checking. In summary, our major contributions are:

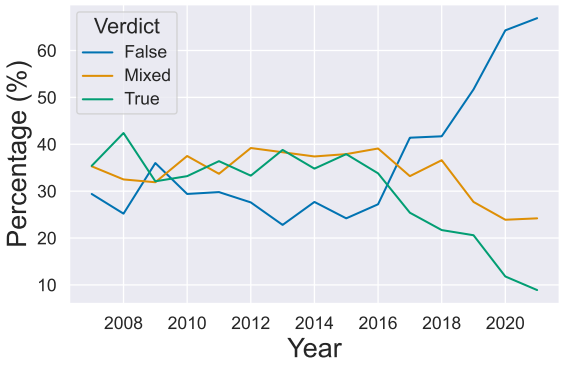

Figure 2: Ratio of verdicts per year (PolitiFact).

- We identify two criteria from the journalistic verification process, which allow overcoming the reliance on counter-evidence (Section 2).
- We show that FCNLP is incapable of satisfying these criteria, preventing the successful verification of most misinformation claims from the journalistic perspective (Section 3).
- We identify two evidence criteria (*sufficient* & *unleaked*) for realistic fact-checking. We find that all existing datasets in FCNLP containing real-world misinformation violate at least one criterion (Section 4) and are hence *unrealistic*.
- We semi-automatically analyze MULTIFC, a large-scale fact-checking dataset to support our findings, and show that models trained on claims from PolitiFact and Snopes (via MULTIFC) rely on leaked evidence.

2 How Humans Fact-check

To motivate our distinct focus on *misinformation*, we investigate what claims professional fact-checkers verify. We crawl 20,274 fact-checked claims from PolitiFact [2] ranging from 2007–2021. Figure 2 shows the ratio of different verdicts [3] per year. After 2016, fact-checkers increasingly select *false* claims as important for fact-checking. In 2021 less than 10% of the selected claims were correct.

Some claims can be refuted via counter-evidence (as required by FCNLP). For example, official statistics can contradict the false claim about the U.S. that "*In the 1980s, the lowest income people had the biggest gains*". If the evidence makes it

[2] https://www.politifact.com/

[3] We conservatively group verdicts "pants on fire" and "false" to *False*, "mostly false" and "half true" to *Mixed* and "mostly true" and "true" to *True*.

	Claim	Based Upon
(1)	If you were forced to use a Sharpie to fill out your ballot, that is voter fraud.	false assumption
(2)	The Biden administration will begin "spying" on bank and cash app accounts starting 2022.	tax legislation
(3)	Barcelona terrorist is cousins with former President Barack Obama.	satire article
(4)	The Democratic health care plan is a government takeover of our health programs.	health care plan
(5)	People in Holland protests against of COVID-19 measures.	protests event

Table 1: Example misinformation claims for source guarantee.

impossible for the claim to be true (e.g., because of mutually exclusive statistics) we refer to the evidence as *global counter-evidence*. Global counter-evidence attacks the textual claim itself without relying on reasoning and sources behind it. In contrast, to refute the claim that "*COVID-19 vaccines may kill sharks*" (Figure 1), fact-checkers did not rely on global counter-evidence specifically proofing that sharks will not be killed to produce COVID-19 vaccines. Neither is it plausible that such counter-evidence exists. Here, the counter-evidence is bound to the claim's underlying (false) reasoning. The claim is only refuted because it follows the false assumption, not because it was disproved. The absence of global counter-evidence is not an exceptional problem for this specific claim but is common among misinformation: Misinformation surges when the high demand for information cannot be met with a sufficient supply of credible answers (Silverman, 2014; FullFact, 2020). Non-credible and possibly false and harmful information fill these deficits of credible information (Golebiewski and Boyd, 2019; Shane and Noel, 2020). The very existence of misinformation often builds on the absence of credible counter-evidence, which in turn, is essential for FCNLP.

Professional fact-checkers refute misinformation even if no global counter-evidence exists, e.g., by rebutting underlying assumptions (Figure 1). Table 1 shows a few false claims built on top of various resources: (1) relies on a false assumption that sharpies invalidate election ballots, (2 & 4) misinterpret official documents or laws, (3) is based on non-credible sources, and (5) changes a topic of a specific event from "*gas extraction*" to "*COVID-19 measures*". Fact-checkers use the reasoning *for* the claim to consider evidence that is, or refers to, the claimant's source: the original tax legislation (2), or alternate (correct) descriptions of protests against gas extraction (5). Here, the content of the evidence alone is often insufficient. The assertion that the claimant's source and the used counter-evidence are identical, or refer to the same event is crucial to refute the claim: Claim (2) is refuted because the tax legislation it relies upon does not support the "spying" claim. However, the document does not specifically refute the claim, and without knowing that the claimant relied on it, it becomes useless as counter-evidence. Similarly, the correct narrative of protests against gas extraction is only mutually exclusive to the false claim (5) of protests against COVID-19 measures when assuring both refer to the identical incident. For similar reasons, the co-reference assumption is critical to the task definition of SNLI (Bowman et al., 2015). After this assertion, mutual exclusiveness is not required to refute the claim: It is sufficient if the claim is not entailed (i.e. incorrectly derived or relies on unverifiable speculations) or based on invalid sources (such as satire) to refute it. Based on these observations we identify two criteria to refute claims if no global-counter evidence exists. We validate their relevance in Section 3:

- ***Source Guarantee:*** The guarantee that identified evidence either constitutes or refers to the claimant's reason for the claim.
- ***Context Availability:*** We broadly consider context as the claim's original environment, which allows us to unambiguously comprehend the claim, and trace the claim and its sources across multiple platforms if required. It is a logical precondition for the source guarantee.

Both criteria are challenging for computers but naturally satisfied by human fact-checkers. Buttry (2014) defines the question *"How do you know that?"* to be at the heart of verification. After selecting a claim, finding provenance and sourcing are the first steps in journalistic verification. Provenance provides crucial information about context and motivation (Urbani, 2020). Journalists must then identify solid sources to compare the claim with (Silverman, 2014; Borel, 2016). Ideally, the claimant provides sources, which must be included and assessed in the verification process. During

verification, journalists rely, if possible, on relevant primary sources, such as uninterpreted and original legislation documents (for claim 2, Table 1). Fact-checking organisations see sourcing as one of the most important parts of their work (Arnold, 2020).

3 Can FCNLP Help Human Verification?

In this section, we first analyze human verification strategies based on an analysis of 100 misinformation claims. We then contrast human verification strategies with FCNLP.

3.1 Human Verification Strategies

We manually analyze 100 misinformation claims[4] from two well-known fact-checking organizations: PolitiFact and Snopes. We randomly choose 50 misinformation claims from each website which contains 25 claims from MULTIFC (a large NLP fact-checking dataset with real-world claims before 2019) and 25 claims from 2020/2021. We extract the URL for each claim and analyze its verification strategy based on the entire fact-checking article. Claims that require the identification of scam webpages, imposter messages, or multi-modal reasoning[5] such as detecting misrepresented, miscaptioned or manipulated images (Zlatkova et al., 2019) were marked as not applicable to FCNLP by nature. In the first round of analysis, we assess whether humans relied on the *source guarantee* to refute the claim. Each claim (and its verification) is unique and can be refuted using different strategies. In the second round of analysis we identify the primary strategy to refute the claim and verify that it is based on the source guarantee. This led us to identify 4 primary human-verification strategies:

1. *Global counter-evidence (GCE):* Counter-evidence via arbitrarily complex reasoning but without the source guarantee.
2. *Local counter-evidence (LCE):* Evidence requires the source guarantee to refute the (reasoning behind) the claim.
3. *Non-credible source (NCS):* Evidence requires the source guarantee to refute the claim based on non-credible sources (e.g. satire).
4. *No evidence assertion (NEA):* The claim is refuted as no (trusted) evidence supports it.

[4]Claims are from the following categories: "*pants on fire*", "*false*" and "*mostly false*".

[5]If a claim can be expressed in text and verified without multi-modal reasoning we consider the verbalized variant of the claim and do not discard it.

Src.	Strategy	MULTIFC	20/21	All	%
yes	LCE	19	16	35	46.7
yes	NCS	9	5	14	18.7
no	GCE	10	10	20	26.7
no	NEA	1	4	5	6.7
no	other	0	1	1	1.3
yes	***all***	**28**	**21**	**49**	**65.3**
no	***all***	**11**	**15**	**26**	**34.7**
all	***all***	**39**	**36**	**75**	100.0

Table 2: Strategies used to refute 75 of 100 misinformation claims with and without source guarantee (**Src.**).

We discard 25 non-applicable claims and show the results of the remaining 75 claims in Table 2. Please refer to Appendix A for more analysis details and examples. In some cases, the selection of one strategy is ambiguous if multiple strategies are applied. In a pilot study to analyze human verification strategies, two co-authors agreed on 9/10 applicable misinformation claims. In general, about two-thirds of the claims were refuted by relying on the source guarantee. In 20 cases fact-checkers refuted the claim by finding global counter-evidence. In one case (*other*), fact-checkers relied entirely on expert statements. In general, experts supported the fact-checkers in identifying and discussing evidence, or strengthened their argument via statements but did not affect the underlying verification strategy.

3.2 NLP Fact Verification

Focusing on evidence-based approaches. Approaches in FCNLP estimate the claim's veracity based on surface cues within the claim (Rashkin et al., 2017; Patwa et al., 2021), assisted with metadata (Wang, 2017; Cui and Lee, 2020; Li et al., 2020; Dadgar and Ghatee, 2021), or using evidence documents. Here, the system uses the stance of the evidence towards the claim to predict the verdict. Verdict labels are often non-binary and include a neutral stance (Thorne et al., 2018), or fine-grained veracity labels from fact-checking organizations (Augenstein et al., 2019). Evidence-based approaches either rely on unverified documents or user comments (Ferreira and Vlachos, 2016; Zubiaga et al., 2016; Pomerleau and Rao, 2017), or assume access to a presumed trusted knowledge base such as Wikipedia (Thorne et al., 2018), scientific publications (Wadden et al., 2020), or search

engine results (Augenstein et al., 2019). In this paper, we focus on trusted evidence-based verification approaches which can deal with the truth changing over time (Schuster et al., 2019). More importantly, they are the most representative of professional fact verification. Effectively debunking misinformation requires stating the corrected fact and explaining the myth's fallacy (Lewandowsky et al., 2020), both of which require trusted evidence.

Global counter-evidence assumption in FCNLP. In FCNLP, evidence retrieval-based approaches assume that the semantic content of a claim is sufficient to find relevant (counter-) evidence in a trusted knowledge base (Thorne et al., 2018; Jiang et al., 2020; Wadden et al., 2020; Aly et al., 2021). This becomes problematic for misinformation that requires the source guarantee to refute the claim. By nature, in this case, the claim and evidence content are distinct and not entailing. Content cannot assert that two different narratives describe the same protests (e.g., Claim 5 in Table 1), or that a non-entailing fact (squalene is harvested from sharks) serves as a basis for the false claim (e.g., Figure 1). The consequence is a circular reasoning problem: Knowing that a claim is false is a precondition to establishing the source guarantee, which in turn is needed to refute the claim. To escape this cycle, one must (a) provide the source guarantee by other means than content (e.g., context), or (b) find evidence that refutes the claim without the source guarantee (global counter-evidence). By relying only on the content of the claim, FCNLP cannot provide the source guarantee and is limited to global counter-evidence, which only accounts for 20% of misinformation claims analyzed in the previous section.

Current FCNLP fails to provide source guarantees. We note that providing the source guarantee goes beyond entity disambiguation, as required in FEVER (Thorne et al., 2018). The self-contained context within claims in FEVER is typically sufficient to disambiguate named entities if required.[6] After disambiguation, the retrieved evidence serves as global counter-evidence.

Recent approaches further add context snippets from Wikipedia (Sathe et al., 2020) or dialogues (Gupta et al., 2022) to resolve ambiguities and cannot provide the source guarantee to break the circular reasoning problem. These snippets differ from the context used by professional fact-checkers who often need to trace claims and their sources across different platforms. Recently, Thorne et al. (2021) annotate more realistic claims w.r.t. multiple evidence passages. They found supporting and refuting passages for the same claim, which prevents the prediction of an overall verdict. Some works collect evidence for the respective claims by identifying scenarios where the *claimant's source* is naturally provided: such as a strictly moderated forum (Saakyan et al., 2021), scientific publications (Wadden et al., 2020), or Wikipedia references (Sathe et al., 2020). However, such source evidence is only collected for true claims. Adhering to the global counter-evidence assumptions of previous work, false claims in these works are generated artificially and do not reflect real-world misinformation.

3.3 Human and NLP Comparison

Our analysis (Table 2) finds fact-checkers only refuted 26% of false claims with global counter-evidence. In all other cases, fact-checkers relied on source guarantees (LCE, NCS) or asserted that no supporting evidence exists (NEA). The verification strategy is not evident given the claim alone but dependent on existing evidence. The claim that "*President Barack Obama's policies have forced many parts of the country to experience rolling blackouts*" is refuted via global counter-evidence (that rolling blackouts had natural causes). The claim that "*90% of rural women and 55% of all women are illiterate in Morocco*" seems verifiable via official statistics. Yet, no comparable statistics exist and the claim is refuted due to relying on a decade-old USAID request report.

We further analyze claims refuted via global counter-evidence, that FCNLP, in theory, can refute. Some claims only require shallow reasoning as directly contradicting evidence naturally exists: A transcript of an interview in which Ron DeSantis was asked about the coronavirus can easily refute the claim "*Ron DeSantis was never asked about coronavirus*". Another case is when information about the claim's veracity already exists, e.g., because those affected by the myth already corrected the claim. Most claims require complex reasoning like legal text understanding or comparing and deriving statistics. Some claims require the definition of some terms first, to make them verifiable. Col-

[6] In the claim "*Poseidon grossed $181,674,817 at the worldwide box office on a budget of $160 million*" it is clear that "Poseidon" refers to the *film*, not an ancient god. (FEVER)

	Dataset	*Claims* Source	False Claims	*Evidence* Unleaked	Sufficient	Ev. Ann.
1	SciFact (Wadden et al., 2020)	Scientific	*generated*	n/a	✓	✓
2	Covid-Fact (Saakyan et al., 2021)	Reddit	*generated*	n/a	✓	✓
3	WikiFactCheck (Sathe et al., 2020)	Wikipedia	*generated*	n/a	✓	✓
4	Fm2 (Eisenschlos et al., 2021)	Game	*generated*	n/a	✓	✓
5	Thorne et al. (2021)	User Queries	*paraphrased*	n/a	✓	✓
6	FaVIQ (Park et al., 2022)	User Queries	*paraphrased*	n/a	✓	*no*
7	LiarPlus (Wang, 2017; Alhindi et al., 2018)	FC Article	*real-world*	*no*	✓	✓
8	PolitiHop (Ostrowski et al., 2021)	FC Article	*real-world*	*no*	✓	✓
9	ClimateFever (Diggelmann et al., 2020)	Web	*real-world*	✓	*no*	✓
10	HealthVer (Sarrouti et al., 2021)	Web	*real-world*	✓	*no*	✓
11	UKP-Snopes (Hanselowski et al., 2019)	FC Article	*real-world*	✓	*no*	✓
12	PubHealth (Kotonya and Toni, 2020b)	FC Article	*real-world*	✓	*no*	*no*
13	WatClaimCheck (Khan et al., 2022)	FC Article	*real-world*	✓	*no*	*no*
14	Baly et al. (2018)	FC Article	*real-world*	*no*	*no*	✓
15	MultiFC (Augenstein et al., 2019)	FC Article	*real-world*	*no*	*no*	*no*
16	X-Fact (Gupta and Srikumar, 2021)	FC Article	*real-world*	*no*	*no*	*no*

Table 3: Overview of NLP fact-checking datasets as realistic test-beds to combat real-world misinformation. We indicate whether humans annotated the stance between claim and evidence (**Ev. Ann.**)

lecting all required global counter-evidence often requires aggregating and comparing different information, possibly under time constraints. Consider the false claim that "*Illegal immigration wasn't a subject that was on anybody's mind until [Trump] brought it up at [his] announcement*": To refute this claim, one must first determine when Trump announced his run for the presidency, then count and compare how often "illegal immigration" was mentioned before and after the announcement.[7]

4 NLP Fact-Checking Datasets

Based on our observations in Section 3 and FcNlp's reliance on global counter-evidence, we hypothesize that evidence in existing fact-checking datasets does not fully satisfy real-world demands. We, hence, investigate how FcNlp's assumptions affect fact-checking datasets and if they constitute realistic test beds for real-world misinformation. For real-world scenarios, datasets must contain real-world misinformation claims and realistic counter-evidence. For evidence, we define the following two requirements:

- *Sufficient:* Evidence must be sufficient to justify the verdict from a human perspective.
- *Unleaked:* Evidence must not contain information that only existed after the claim was verified.

The issue of leaked evidence was also mentioned very recently by Khan et al. (2022). Unlike us, they did not comprehensively analyze existing datasets, evaluate the impact on trained systems (Section 5), or consider the complementary criterion of *sufficient* (counter-) evidence. Relying on leaked evidence is related to the important yet different task of detecting already-verified claims (Shaar et al., 2020), but is unrealistic for novel claims.

We survey NLP fact-checking datasets with natural input claims[8] that assume access to trusted evidence. Table 3 summarizes our survey results. Datasets 1-6 contain no real-world misinformation: False claims are derived from true real-world claims (1-3) or within a gamified setting (4), ensuring that counter-evidence exists. Other works (5 & 6) reformulate real-world user queries, which are linked to Wikipedia articles as (counter-) evidence.

We find that no dataset with real-world misinformation (7-16) satisfies both evidence criteria. We identify four categories: First, datasets that consider (parts of) a fact-checking article as evidence contain sufficient, yet leaked evidence (7 & 8). Second, annotators estimate claim veracity based on evidence such as Wikipedia or scientific publications. The authors find that evidence often only covers parts of these realistic, complex claims, which yield low annotator agreement (9), or a weakened task definition for stances (10).

Third is to rely on the same evidence as fact-

[7] We assume disambiguation requires no source guarantees, and that basic context (date, location, claimant) is known.

[8] See the survey from Guo et al. (2022) for datasets with natural and artificial claims.

Detected via	Claim & Evidence
phrase & snippet URL	**Claim**: *Google Earth Finds SOS From Woman Stranded on Deserted Island* **Evidence Snippet:** The Truth: The story is a **hoax**. ... GOOGLE EARTH FINDS WOMAN TRAPPED ON DESERTED ISLAND FOR 7 YEARS ... other end "How did you find me" to which they replied "Some kid from Minnesota found your SOS sign on Google Earth"; From ***Truth Or Fiction***
phrase	**Claim**: *Country music singer Merle Haggard left his entire estate to an LGBT group.* **Evidence Snippet:** Discover ideas about Country Singers. **Fake news** reports that recently-deceased country music legend Merle Haggard left his entire estate to an LGBT group; From *Pinterest*

Table 4: Examples from MULTIFC of leaked evidence detected via the *snippet URL* (linking to a fact-checking article) or a **phrase** of the evidence snippet.

checkers, termed *premise evidence* by Khan et al. (2022). Here, only UKP-SNOPES (11) provides evidence annotations. Hanselowski et al. (2019) collect and annotate original evidence snippets from the fact-checking article. They find the stance often conflicts with the verdict: Though most claims are false, the majority of evidence is supporting. In 45.5% of cases, annotators found no stance for the professionally selected evidence snippets even though professional fact-checkers considered these snippets important to be included in the article. Due to conflicting and unexplained evidence snippets, we rate this *insufficient* to predict the correct verdict. The human verification process (Section 2) guides the creation of the fact-checking article and can serve as a possible explanation for these problems. Articles link to the claim's context and possibly other similar claims (likely *supporting*). Often (e.g. during COVID-19 (Simon et al., 2020)), claims are not completely fabricated. Fact-checkers identify documents and their interdependence when investigating the claimant's reasoning for the claim (likely not *refuting*). Documents used to disprove the claimant's reasoning may have no or little relevance to the original claim (as in Figure 1). Each step is non-trivial and may rely on numerous documents (or expert statements). Relying on premise evidence without considering the verification process and *how* these documents relate, is insufficient. Both other datasets (12 & 13) in this category provide no annotations and are limited to freely available evidence documents (as opposed to paywalled web pages or e-mails).

Fourth is using a search engine during dataset construction to expand the accessible knowledge. Even when excluding search results that point to the claim's fact-checking article, leaked evidence persists: Different organizations may verify the same claims, or disseminate the fact-checkers verification. Only Baly et al. (2018) provide stance annotation for Arabic claim and evidence pairs. For false claims, they found that only a few documents disagree, and more agree, with the claim. A possible explanation is that misinformation often emerges when trusted information or counter-evidence is scarce. Fact-checking articles fill this deficit. Partially excluding them during dataset generation reduces the found counter-evidence. Lacking counter-evidence is not a problem of the dataset generation, but the underlying nature of misinformation, and should be considered by the task definition. We rate evidence in this category (14-16) leaked and insufficient, and back it up in Section 5.

5 A Case Study of Leaked Evidence

We view MULTIFC (Augenstein et al., 2019), the largest dataset of its group, as an instantiation of FCNLP applied to the real world: It contains real-world claims and professionally assessed verdicts as labels from 26 fact-checking organizations (like Snopes or PolitiFact). The authors use the Google search engine to expand evidence retrieval to the real world during dataset construction. We abstract from the fact that MULTIFC only provides incomplete evidence snippets and consider (if possible) the underlying article in its entirety.

5.1 Quantification of Leaked Evidence

We focus our analysis on 16,244 misinformation claims that we identify via misinformation labels (listed in Appendix B.1). To quantify how many claims in MULTIFC contain leaked evidence, we consider all evidence snippets stemming from a fact-checking article, or discussing the veracity of a claim, as leaked. Table 4 shows examples of leaked evidence that strongly indicates the claim's verdict. The first snippet comes directly from a fact-checking organization (Truth Or Fiction[9]). Only identifying leaked evidence that directly comes

[9] https://www.truthorfiction.com/

Leaked	Claims (Number)	Claims (%)
Url	8,999	55.6%
Phrase	9,656	59.7%
Url or Phrase	11,267	69.7%

Table 5: Absolute and relative number of automatically identified leaked evidence of MULTIFC misinformation.

Categories	*Claim has leaked evidence* All	Leaked	Unleaked
Any	100	32	68
No Stance	37	0	37
No Refuting	63	0	63
Original & Refuting	15	10	5

Table 6: Manual analysis of 100 claims without automatically identified leaked evidence.

Input	All	Leaked	Unleaked	Δ
		Snopes		
Samples	1,014	482	532	
Sn.-Text	29.4/60.4	26.3/66.3	30.1/55.0	+11.3
Sn.-Title	27.3/57.8	23.8/64.6	28.2/51.5	+13.1
Snippets	30.5/60.5	28.7/67.6	30.2/53.7	**+13.9**
Full	**32.7/62.7**	**30.6/68.7**	**33.0/57.2**	+11.5
		PolitiFact		
Samples	2,717	2,111	606	
Sn.-Text	35.5/34.5	38.0/37.2	24.1/24.5	+12.7
Sn.-Title	48.0/47.4	55.1/54.6	21.1/21.6	+33.0
Snippets	52.0/51.3	**59.7**/59.2	22.1/23.0	**+36.2**
Full	**52.6/51.9**	**59.7/59.3**	**25.6/25.9**	+33.4

Table 7: F1-macro/micro scores and difference in F1-micro (Δ) averaged over 3 runs. Inputs are: only snippet texts, titles, entire snippets or claim and snippets (full).

from fact-checking organizations is insufficient: After the publication of the verification report, its content is disseminated via other publishers (such as Pinterest in the second example). We identify leaked evidence snippets using patterns for their source URLs or contained phrases. A complete list of all used patterns is given in Appendix B.2). This requires the evidence to be relevant. Irrelevant articles are insufficient, albeit not leaking. To this end, we manually analyze 100 claims with 230 automatically found leaking evidence snippets. We confirm that 83.9% of the snippets are leaked (details in Appendix B.3). 97/100 of the selected claims contain at least one leaked evidence snippet.

Table 5 lists the number of claims with leaked evidence identified by the pattern-based approach. It detected leaked evidence for 69.7% of misinformation claims. In addition, we manually analyze evidence of 100 misinformation claims for which this approach found no leaked evidence. Misinformation verification often requires multiple evidence documents, rendering a single sufficient evidence snippet unrealistic. We follow Sarrouti et al. (2021) and test if a snippet supports or refutes parts of the claim. Table 6 shows that approximately one-third of the claims contain further leaked evidence. 15 claims have unleaked refuting evidence. In 10 cases this evidence is overshadowed via leaked evidence for the same claim. Most analyzed claims only have non-refuting evidence. Similar to Baly et al. (2018), we found supporting evidence for 40 misinformation claims; for 35 of these claims, the evidence was misinformation and thus supported the claim; for the remaining five claims, the claim became accurate, and the evidence became available at a date later than the claim's creation.

5.2 Impact on Trained Systems

Hansen et al. (2021) found that models in MULTIFC can predict the correct verdict based on the evidence snippets alone. To test if leaked evidence can serve as an explanation, we fine-tune BERT (Devlin et al., 2019) (*bert-base-uncased*) to predict the veracity label of a claim given the evidence snippets with and without a claim. As input to BERT, we separate the claim (when used) from the evidence snippets using a [SEP] token and predict the veracity label based on a linear layer on top of a preceding [CLS] token (Training details in Appendix C.1.). When each evidence snippet is represented via its content only this performs on par with the specialized model introduced by Hansen et al. (2021). We additionally find that the snippet's title carries much signal, and adding it to the input improves the overall performance on PolitiFact. Snippets are concatenated (separated by ";" in the order provided by MULTIFC and truncated after 512 tokens. We experiment on the train-, dev- and test-splits Hansen et al. (2021) extracted from MULTIFC on claims from Snopes and PolitiFact. We test four types of input: only evidence (only title, only text, both), or the complete sample of claim and evidence. For the evaluation (Table 7), we split the test data based on whether a claim contains automatically identified leaked evidence.

On Snopes, the macro-F1 is higher on the unleaked than on the leaked subset. Upon closer inspection, we find that the label distribution on Snopes is heavily skewed towards "false", which

worsens on the leaked subset. Models seem to rely on patterns of leaked evidence to predict the majority label "false" (see Appendix C.2). On the leaked subset, this comes at the cost of incorrect predictions for all other labels, yielding a lower F1-macro. On the larger PolitiFact subset, labels are not much skewed towards a single majority label. Across all experiments, the performance gap signals the reliance on leaked evidence. We confirm the impact of leaked evidence for both datasets by evaluating the model on the *same* instances with leaked *or* unleaked evidence, to avoid the different label distribution distorting the results (Appendix C.3).

6 Related Work

Combat Misinformation After Its Verification. The identified limitations of the previous studies on NLP fact-checking datasets described in Section 4 do not devalue the surveyed datasets and we view them as highly important and useful contributions. These limitations are tied to our specific research question to refute *novel* real-world misinformation. We strongly build on these previous works and view them as crucial starting points to fact-check real-world misinformation. Existing fact-checking articles are highly valuable and automatic methods should utilize them to detect and combat misinformation. Automatic methods specifically using these resources detect misinformation by matching claims with known misconceptions (Hossain et al., 2020; Weinzierl and Harabagiu, 2022) or already verified claims (Vo and Lee, 2020; Shaar et al., 2020; Martín et al., 2022; Hardalov et al., 2022b).

Surveys on Automatic Fact-Checking. Recent work surveyed (aspects of) automated fact-checking and related tasks, including explainability (Kotonya and Toni, 2020a), stance classification (Küçük and Can, 2020; Hardalov et al., 2022a), propaganda detection (Da San Martino et al., 2021), rumor detection on social media (Zubiaga et al., 2018; Islam et al., 2020), fake-news detection (Oshikawa et al., 2020; Zhou and Zafarani, 2020), and automated fact-checking (Thorne and Vlachos, 2018). We refer interested readers to these papers.

Guo et al. (2022) surveyed the state of automatic fact-checking. Based on their work, we zoom in on real-world misinformation to investigate the gap between professional fact-checkers and FCNLP. Recently, Nakov et al. (2021) surveyed tasks to assist humans during the verification. Our work differs in that we focus solely on the automatic verification approach of misinformation. They argue for the need for automatic tools to support *humans* during verification. Similarly, Graves (2018) interviewed expert fact-checkers and computer scientists and conclude, that automatic fact-checking cannot replicate professional fact-checkers in the foreseeable future. Our results confirm the challenging nature of misinformation but also outline why current models have unrealistic expectations, and how humans overcome these problems. We believe this to be important as real-world misinformation is well within the scope of current NLP research.

Towards Human Verification. A possible path forward is to align automatic verification with journalistic verification: Use the claimant's reasoning to find evidence and verify the claim. This relies on the complex task of finding the correct sources (Arnold, 2020). A fruitful but understudied direction may be automated provenance detection (Zhang et al., 2020, 2021). Building systems that can provide source guarantees paves the way for reasoning tasks, such as the detection of logical fallacies (Jin et al., 2022), implicit warrants (Habernal et al., 2018), or propaganda techniques (Da San Martino et al., 2019; Huang et al., 2022). Integrating sufficient context into datasets is non-trivial and may require tracing a claim and its source across multiple platforms. Existing literature shows the heterogeneity of misinformation (Borel, 2016; Wardle et al., 2017; Cook, 2020) and can help to identify small, focused problems that can realistically be translated into NLP. Approaches from computer vision focus on misinformation-specific approaches to detect manipulated or misrepresented images (Zlatkova et al., 2019; Abdelnabi et al., 2022; Musi and Rocci, 2022).

7 Conclusion

In this work, we contrasted NLP fact-checking approaches with how professional fact-checkers combat misinformation. We identified that reliance on counter-evidence hinders current fact-checking systems to refute real-world misinformation. Using MULTIFC we find that most evidence is insufficient, or leaked and exploited by trained models. Moving forward, we suggest to align NLP approaches with the human verification process, and task definitions with smaller and well-defined verification strategies.

Limitations

The scope of this study is restricted to misinformation claims, and their representation as textual statements, that professional fact-checking organizations selected as important to verify. This only represents a fraction of all existing misinformation (Vinhas and Bastos, 2022). Our findings cannot be generalized to other types of misinformation. Process definitions for claim selection and verification differ amongst fact-checkers (Arnold, 2020). The assessed claims for the analysis and experiments are biased to the claim selection criteria, including the domain, language, and geographical biases of Snopes and PolitiFact. Even fact-checkers cannot fully eliminate subjectivity. Nieminen and Sankari (2021) find 11% PolitiFact's verified claims uncheckable. We consider the fact-checkers assessment as the gold standard and adhere to the introduced subjectivity. PolitiFact and Snopes verify claims from English-speaking countries with rich resources and trusted government documents. Fact-checking organizations may rely on different strategies, adapted to different scenarios such as different topics, dissemination of misinformation, or trust and availability of official information.[10] The quantification of leaked evidence is bound to the time-frame, fact-checking organizations, and found evidence of MULTIFC. We did not investigate the influence of different factors such as the fact-checkers language, domain, or popularity, nor did we evaluate different evidence collection strategies. The same restrictions apply to the experimental results. Further, following Hansen et al. (2021) we only consider labels on a veracity scale for the experiments (e.g. excluding "misleading").

Ethics Statement

In this work we only consider publicly available data as provided by fact-checking organizations or MULTIFC, but do not publish it. We do not use any personal data. We note that creating more realistic datasets (including realistic context), as suggested by us, induces ethical challenges as it requires personalized data (e.g. from Twitter or Facebook). We consider this study's goal to reduce harmful misinformation by aligning automatic methods with best-practices from professional fact-checkers as ethically correct. However, even if successfully developed, fact-checking systems are inevitably imperfect. Malicious actors may design claims that exploit the system's weakness to predict the opposite verdict, giving legitimacy to false claims, or discrediting correct claims. Further, malicious actors may develop fact-checking systems under their control. When extended with triggers enabling backdoor attacks (Chen et al., 2021) to control the outcome, these systems can serve as powerful tools to decide what seems true or false.

[10] https://www.poynter.org/fact-checking/2019/heres-how-fact-checking-is-developing-across-africa/

Acknowledgements

We thank the anonymous reviewers for their valuable feedback. We further thank Luke Bates, Tim Baumgärtner and Ilia Kuznetsov for their feedback on this work. This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE and by the European Regional Development Fund (ERDF) and the Hessian State Chancellery – Hessian Minister of Digital Strategy and Development under the promotional reference 20005482 (TexPrax).

References

Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. 2022. Open-Domain, Content-based, Multi-modal Fact-checking of Out-of-Context Images via Online Resources. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14940–14949.

Hamidreza Aghababaeian, Lara Hamdanieh, and Abbas Ostadtaghizadeh. 2020. Alcohol intake in an attempt to fight COVID-19: A medical myth in Iran. *Alcohol*, 88:29–32.

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.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Phoebe Arnold. 2020. The challenges of online fact checking: how technology can (and can't) help. Technical report, FullFact.

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.

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, 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.

Brooke Borel. 2016. *The Chicago guide to fact-checking*. University of Chicago Press.

Alexandre Bovet and Hernán A Makse. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications*, 10(1):1–14.

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.

Steve Buttry. 2014. Verification fundamentals: Rules to live by. *Verification Handbook: A Definitive Guide to Verifying Digital Content for Emergency Coverage*, pages 15–23.

Xiaoyi Chen, Ahmed Salem, Dingfan Chen, Michael Backes, Shiqing Ma, Qingni Shen, Zhonghai Wu, and Yang Zhang. 2021. BadNL: Backdoor Attacks against NLP Models with Semantic-preserving Improvements. In *Annual Computer Security Applications Conference*, pages 554–569.

John Cook. 2020. Deconstructing Climate Science Denial. In *Edward Elgar Research Handbook in Communicating Climate Change*. Edward Elgar Publishing.

Limeng Cui and Dongwon Lee. 2020. CoAID: COVID-19 Healthcare Misinformation Dataset. *arXiv preprint arXiv:2006.00885*.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2021. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4826–4832.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-Grained Analysis of Propaganda in News Article. 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 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Sajad Dadgar and Mehdi Ghatee. 2021. Checkovid: A COVID-19 misinformation detection system on Twitter using network and content mining perspectives. *arXiv preprint arXiv:2107.09768*.

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.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims. In *Tackling Climate Change with Machine Learning workshop at NeurIPS*.

Julian Eisenschlos, Bhuwan Dhingra, Jannis Bulian, 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.

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.

Marc Fisher, John Woodrow Cox, and Peter Hermann. 2016. Pizzagate: From rumor, to hashtag, to gunfire in DC. *Washington Post*, 6:8410–8415.

FullFact. 2020. Framework for information incidents. Technical report, FullFact.

Michael Golebiewski and Danah Boyd. 2019. Data voids: Where missing data can easily be exploited. Technical report, Data & Society Research Institute.

Lucas Graves. 2018. Understanding the Promise and Limits of Automated Fact-Checking. In *Reuters Institute for the Study of Journalism (Reuters Institute for the Study of Journalism Factsheets)*. Reuters Institute for the Study of Journalism.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

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.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. DialFact: A Benchmark for Fact-Checking in Dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801, Dublin, Ireland. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The Argument Reasoning Comprehension Task: Identification and Reconstruction of Implicit Warrants. 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 1930–1940, 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.

Casper Hansen, Christian Hansen, and Lucas Chaves Lima. 2021. Automatic Fake News Detection: Are Models Learning to Reason? 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 80–86, Online. Association for Computational Linguistics.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022a. A Survey on Stance Detection for Mis- and Disinformation Identification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277, Seattle, United States. Association for Computational Linguistics.

Momchil Hardalov, Anton Chernyavskiy, Ivan Koychev, Dmitry Ilvovsky, and Preslav Nakov. 2022b. CrowdChecked: Detecting Previously Fact-Checked Claims in Social Media. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, page (to appear), online. Association for Computational Linguistics.

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.

Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2022. Faking Fake News for Real Fake News Detection: Propaganda-loaded Training Data Generation. *arXiv preprint arXiv:2203.05386*.

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. *Social Network Analysis and Mining*, 10(1):1–20.

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.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. 2022. Logical Fallacy Detection. *arXiv preprint arXiv:2202.13758*.

Kashif Khan, Ruizhe Wang, and Pascal Poupart. 2022. WatClaimCheck: A new dataset for claim entailment and inference. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1293–1304, Dublin, Ireland. Association for Computational Linguistics.

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 (CSUR)*, 53(1).

Stephan Lewandowsky, John Cook, Ullrich Ecker, Dolores Albarracin, Michelle Amazeen, P. Kendou, D. Lombardi, E. Newman, G. Pennycook, E. Porter, D. Rand, D. Rapp, J. Reifler, J. Roozenbeek, P. Schmid, C. Seifert, G. Sinatra, B. Swire-Thompson, S. van der Linden, E. Vraga, T. Wood,

and M. Zaragoza. 2020. *The Debunking Handbook 2020*. OpenBU.

Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020. Toward A Multilingual and Multimodal Data Repository for COVID-19 Disinformation. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 4325–4330.

Alejandro Martín, Javier Huertas-Tato, Álvaro Huertas-García, Guillermo Villar-Rodríguez, and David Camacho. 2022. FacTeR-Check: Semi-automated fact-checking through semantic similarity and natural language inference. *Knowledge-Based Systems*, 251:109265.

Elena Musi and Andrea Rocci. 2022. Staying Up to Date with Fact and Reason Checking: An Argumentative Analysis of Outdated News. In Steve Oswald, Marcin Lewiński, Sara Greco, and Serena Villata, editors, *The Pandemic of Argumentation*, pages 311–330. Springer International Publishing, Cham.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated Fact-Checking for Assisting Human Fact-Checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4551–4558. International Joint Conferences on Artificial Intelligence Organization.

Sakari Nieminen and Valtteri Sankari. 2021. Checking PolitiFact's Fact-Checks. *Journalism Studies*, 22(3):358–378.

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.

Wojciech Ostrowski, Arnav Arora, Pepa Atanasova, and Isabelle Augenstein. 2021. Multi-Hop Fact Checking of Political Claims. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3892–3898. International Joint Conferences on Artificial Intelligence Organization.

Jungsoo Park, Sewon Min, Jaewoo Kang, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. FaVIQ: FAct Verification from Information-seeking Questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5154–5166, Dublin, Ireland. Association for Computational Linguistics.

Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2021. Fighting an infodemic: Covid-19 fake news dataset. In *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 21–29. Springer.

Dean Pomerleau and Delip Rao. 2017. Fake News Challenge. `http://www.fakenewschallenge.org/`.

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.

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 (Volume 1: Long Papers)*, pages 2116–2129, Online. Association for Computational Linguistics.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based Fact-Checking of Health-related Claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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*, pages 6874–6882, Marseille, France. European Language Resources Association.

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.

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.

Tommy Shane and Pedro Noel. 2020. Data deficits: why we need to monitor the demand and supply of information in real time. Technical report, First Draft.

Craig Silverman. 2014. *Verification handbook: An ultimate guideline on digital age sourcing for emergency coverage*. European Journalism Centre.

Craig Silverman. 2016. *Verification handbook: Additional Materials*. European Journalism Centre.

Felix Simon, Philip N. Howard, and Rasmus Kleis Nielson. 2020. Types, sources, and claims of COVID-19 misinformation. Technical report, Reuters Institute for the Study of Journalism.

James Thorne, Max Glockner, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2021. Evidence-based Verification for Real World Information Needs. *arXiv preprint arXiv:2104.00640*.

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, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Shaydanay Urbani. 2020. Verifying Online Information. Technical report, First Draft.

Sander van der Linden. 2022. Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, 28(3):460–467.

Otávio Vinhas and Marco Bastos. 2022. Fact-Checking Misinformation: Eight Notes on Consensus Reality. *Journalism Studies*, 23(4):448–468.

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.

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.

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.

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.

Claire Wardle et al. 2017. Fake news. it's complicated. Technical report, First Draft.

Maxwell Weinzierl and Sanda Harabagiu. 2022. VaccineLies: A natural language resource for learning to recognize misinformation about the COVID-19 and HPV vaccines. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6967–6975, Marseille, France. European Language Resources Association.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

John Zarocostas. 2020. How to fight an infodemic. *The Lancet*, 395(10225):676.

Yi Zhang, Zachary Ives, and Dan Roth. 2020. "Who said it, and Why?" Provenance for Natural Language Claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4416–4426, Online. Association for Computational Linguistics.

Yi Zhang, Zachary Ives, and Dan Roth. 2021. What is Your Article Based On? Inferring Fine-grained Provenance. 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 5894–5903, Online. Association for Computational Linguistics.

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

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 (CSUR)*, 51(2).

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, and Michal Lukasik. 2016. Stance Classification in Rumours as a Sequential Task Exploiting the Tree Structure of Social Media Conversations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2438–2448, Osaka, Japan. The COLING 2016 Organizing Committee.

A Human Misinformation Verification Examples

Table 8 shows further examples for the reasoning over global counter-evidence, or when source guarantees are required. Further, we list examples we considered inapplicable to FCNLP by nature and ergo excluded. Cases include imposter content (1), fake web pages (2), and claims that are about multimodal content (3) or require reasoning over multimodal sources during verification (4). We show further examples of complex and different reasoning via global counter-evidence. This includes: reasoning over scientific documents (5); aggregating counter-examples (6); aggregating distinct donations (7); finding and contextualizing multiple events, including deportation and imprisonment of a murderer, and aligning these events with the political leadership in the U.S. (8); Claims (9–11) require source guarantees to resources like a specific document (9), event (10), or organization (11).

B Leaked Evidence Analysis

B.1 Misinformation Labels

We consider all claims rated with strongly-leaning false verdicts and other verdicts that fall into the misinformation category such as "misleading" as misinformation. Remaining claims are either true (e.g. "verified", "mostly true"), mixed (e.g. "half-true", "outdated") or not clearly applicable to misinformation (e.g. "opinion!", "scam", "full flop"). We provide all considered labels within MULTIFC below.

- *ABC*: in-the-red
- *Africa Check*: incorrect, misleading
- *BOOM Live*: rating: false
- *Check Your Fact*: verdict: false
- *Climate Feedback*: incrorrect, misleading
- *FactScan*: factscan score: false, factscan score: misleading
- *Factly*: false
- *FactCheckNI*: conclusion: false
- *FactCheck.org*: false, distorts the facts, misleading, spins the facts, not the whole story, cherry picks
- *Gossip Cop*: 0, 1, 2, 3
- *Hoax Slayer*: fake news
- *Huffington Post CA*: a lot of baloney
- *MPR News*: false, misleading
- *Observatory*: mostly_false
- *Pandora*: mostly false, false, pants on fire!
- *PesaCheck*: false
- *PolitiFact*: mostly false, false, pants on fire!, fiction
- *Radio NZ*: fiction
- *Snopes*: false, mostly false, miscaptioned, misattributed
- *The Ferret*: mostly false, false
- *The Journal*: we rate this claim false
- *Truth Or Fiction*: fiction!, mostly fiction!, incorrect attribution!, misleading!, inaccurate attribution!
- *VERA Files*: fake, misleading, false
- *Voice of San Diego*: determination: false, determination: huckster propaganda, determination: barely true, determination: misleading
- *Washington Post*: 4 pinnochios, false, not the whole story, needs context

B.2 Automatic Identification of Leaked Evidence

Table 9 shows the URLs used to automatically determine leaked evidence. We consider an evidence snippet as leaked if any URLs of Table 9 is a substring of the snippet's URL. We exclude URLs if they may also cover URLs to news articles. Further, we consider an evidence snippet as leaked, if its lowercased title or text matches any of the regular expressions in Table 10. We identified two commonly made errors:

- **Different Claim:** The approach considered evidence as leaked if it is not relevant to the exact same claim, but connected to the same incident (*"President Obama pushed through the stimulus based on the promise of keeping unemployment under 8 percent."* and *"The president promised that if he spent money on*

#	Year	Misinformation Claim	Strategy
(1)	2020	Tennis star Serena Williams posted a message on social media that began, "I'm sick of COVID-19. I'm sick of black vs. white. I'm sick of Republicans vs. Democrats."	*n/a*
(2)	2010	You can look up anyone's driver's license for free through the 'National Motor Vehicle Licence Organization' web site.	*n/a*
(3)	2016	A photograph shows a newly hatched baby dragon.	*n/a*
(4)	2018	Couple Arrested For Selling 'Golden Tickets To Heaven.	*n/a*
(5)	2021	There is no added safety to the public if you're vaccinated.	*GCE*
(6)	2011	Limiting labor negotiations to only wages is how it is for the most part in the private sector.	*GCE*
(7)	2018	Kathy Manning gave nearly $1 million to liberals	*GCE*
(8)	2018	Democrats let him (cop killer Luis Bracamontes) into our country, and Democrats let him stay.	*GCE*
(9)	2014	A list reproduces Saul Alinsky's rules for "How to Create a Social State.".	*LCE*
(10)	2014	Greg Abbott and his surrogates have referred to women who have been the victims of rape or incest as though somehow what they are confronting is a minor issue.	*LCE*
(11)	2021	During protests over the police in-custody death of George Floyd in the summer of 2020, Kamala Harris donated money to a Minnesota nonprofit that helped protesters who were arrested get out of jail and break more laws.	*LCE*

Table 8: Example claims from Snopes and PolitiFact that professional fact-checkers refuted with global counter-evidence (*GCE)*, local counter-evidence (*LCE*) via the source guarantee, or marked as not-applicable (*n/a)* for FCNLP.

Organization	Template
Africa Check	`africacheck.org/reports`
AFP Fact Check	`factcheck.afp.com`
Check Your Fact	`checkyourfact.com`
Climate Feedback	`climatefeedback.org/claimreview`
Fact or Fiction	`radionz.co.nz/programmes/election17-fact-or-fiction`
FactCheck.org	`factcheck.org`
FactCheckNI	`factcheckni.org`
FACTLY	`factly.in`
FactsCan	`factscan.ca`
Full Fact	`fullfact.org`
Gossip Cop	`gossipcop.com`
Health Feedback	`healthfeedback.org/claimreview`
Hoax Slayer	`hoax-slayer.net`
Lead Stories FactChecker	`hoax-alert.leadstories.com`
PesaCheck	`pesacheck.org`
PolitiFact	`politifact.com`
Snopes	`snopes.com`
Truth or Fiction	`truthorfiction.com`
Washington Post	`washingtonpost.com/news/fact-checker`

Table 9: Used Templates to automatically identify leaked snippets via the URLs.

Regular Expression	Example
`'^false:'`	**FALSE:** Map Shows Results of the 2012 Presidential Election If Only ... A map doesn't show the results of the 2012 election if only people who pay ... FALSE: Map Shows Results of the 2012 Presidential Election If Only Taxpayers Had ... is a map of how the Electoral College vote would look like if ONLY those who ...
`'politifact'`	**PolitiFact**: Testing Kathleen Vinehout claim on Scott Walker, new car ... Dec 20, 2013 ... We check a claim by state Sen. Kathleen Vinehout that Gov. Scott Walker bought "80 new, brand new vehicles" that "we probably don't need.".
`'snopes'`	Real History Blog: The ACLU has NOT filed suit to have all military ... Feb 10, 2010 ... The ACLU has never filed such a suit, says the ACLU. Says **Snopes**, if ... and another suit to end prayer from the military completely. They're ...
`'^debunk'`	**Debunk**ed: Did 'The Simpsons' predict President Donald Trump's ... Feb 9, 2017 ... 'The Simpsons' has predicted a number of world events and an internet rumor said the show predicted the death of Donald Trump. Veuer's Nick ...
`'real story behind'`	**The real story** behind the statistic Trump just used to attack Obamacare; Jun 13, 2017 ... ... tweeted that 2 million people "just dropped out of ObamaCare." 2 million more people just dropped out of ObamaCare. It is in a death spiral.
`'\bfake\b'`	Trump "moron" Harley-Davidson CEO quote: **Fake**.Jun 27, 2018 ... The CEO of Harley-Davidson Did Not Call Donald Trump a "Moron" ... Harley Davidson CEO Matthew S Levatich says: "Our decision to move ...
`'\bhoax\b'`	Eddie Murphy - latest news, breaking stories and comment - The ... All the latest breaking news on Eddie Murphy. Browse The ... Paul Walker tragedy sparks Eddie Murphy Twitter death **hoax** · News · Final film of the Twilight ...
`'\bfalsely\b'`	CNN helpfully fact-checks Donald Trump's tweet about its "way down"; Jun 27, 2017 ... Trump tweeted that "Fake News CNN" had its "Ratings way down!" which he said was due to the network being "caught **falsely** pushing their ...
`'\brumors?\b'`	Did the Obama White House ban Christmas Nativity scenes ... Nov 21, 2018 ... Contrary to "War on Christmas" **rumors**, the Obama White House did not ban Nativity scenes from the premises: ...
`'\bmyths?\b'`	Did Coca-Cola Contain Coke? Here's What History Says; Since I was a little girl, I've heard the **myth** that Coca-Cola used to actually contain cocaine. However, how credible is this rumor? I set out to find if there was any ...
`'\bnot real news\b'`	**NOT REAL NEWS**: A look at what didn't happen this week; Aug 11, 2017 ... NOT REAL: John McCain Says He 'Accidentally' Voted No On Healthcare Repeal ... last month that sank a GOP effort to repeal the Affordable Care Act, ... story purportedly showing the fake senator in handcuffs is actually a ...
`'\bunfounded\b'`	No, 15,000 people did not vote for Harambe in 2016 | PunditFact; Nov 22, 2016 ... Harambe received 15,000 votes in the presidential election. ... Rumors that 15,000 people voted for the dead gorilla Harambe are **unfounded**.
`'fact[ -]check'`	**Fact-check**ing an immigration meme that's been circulating for more ...Jul 5, 2018 ... "More than 66% of ALL births in California are to illegals on Medi-Cal" ... According to Medi-Cal, 50.4 percent of the state's births that year were ...

Table 10: Used regular expressions to automatically identify leaked evidence snippets.

a stimulus program that unemployment would go to 5.7 percent or 6 percent. Those were his words"), or thematically related (*"Cadbury chocolate eggs are infected with HIV-positive blood"* and *"HIV & AIDS infected oranges coming from Libya"*).

- **Discussing Fake News or Fact-Checking:** The approach selects snippets, that discuss fact-checking or fake news from a different perspective, not as a result of verification. This includes opinions or reports complaining about "fake news" being spread, or about the fact-checking process.

B.3 Manual Guidelines

To determine the stance of evidence snippets to a claim, or whether it is leaked or not, we proceed in the following order: We first read the original fact-checking article, to fully comprehend the claim and how fact-checkers refuted it. If the title or text of the evidence snippets provides sufficient information, we decide based on the snippet alone. If we cannot make a clear decision based on the snippet, we consider the original web page. This may be required, as evidence snippets often contain incomplete sentences.

B.3.1 Leaked Evidence Snippets

We consider evidence snippets as leaked if (a) they constitute information that relies on the verification of the same claim, or (b) provides originally unknown information from the claim's future. When relying on content from the claim's verification, we do not require the information to contradict the claim from a human perspective. This often occurs when different pages (such as overview pages) reference the fact-checking article. Such a page may be a clear indication of the verdict in some cases (e.g. if titled "All False Claims by Person A"). In other cases, different interpretations are valid: The statement "We previously fact-checked similar claims that ..." may be seen as neutral or as a give-away that similar claims were refuted. Further, humans cannot judge whether models may rely on latent patterns. An overview page titled "All claims from Person A" may be sufficient for the model if it learned that most claims by Person A are false. To remove this ambiguity, we consider any mention / or information taken form the claim's verification as leaked. We do not strictly consider all evidence that appeared after the verification as leaked. Not all evidence published after the claim's verification, is based not based on the verification. If not, we verify whether it relies on new information that previously did not exist or whether the truth changed. Consider the claim "Khloe Kardashian did give birth over easter." refuted on April 5, 2018. Evidence about her actual birth on April 12, 2018, does not rely on a previous verification but is still considered leaked (new information available). In other cases ("Coca-Cola's "Share A Coke" campaign includes a bottle for the KKK.", March 2, 2016) we consider evidence from March 30, 2016, as unleaked: It correctly reports about the same incident the claim refers to without any mention of the false claim, or its verification.

B.3.2 Stance of Evidence Snippets

For most claims, it is unrealistic to assume a single evidence snippet can refute them entirely. We follow Sarrouti et al. (2021) to allow evidence to support or refute parts of the claim only. We separately mark supporting evidence from the claim's future, as the claim's veracity may have changed. We consider correctly identified counter-evidence as refuting the claim, even when it requires the source guarantee.

C Experiments on MULTIFC

C.1 Training details

For our experiments we use `bert-base-uncased` as provided by Wolf et al. (2020). We represent each evidence snippet e as the title, the text body, or the concatenation of both (depending on the experiment). We concatenate all evidence snippets e_i, separated by a semicolon (e_1; e_2; ...; e_n). We input the concatenation of the claim c and the concatenated evidence, separated by `[SEP]` token, and truncated after 512 tokens: `[CLS]` c `[SEP]` e_1; e_2; ...; e_n `[SEP]`. We predict the label via a linear layer on the [CLS] token. We train all models for `5` epochs with a learning rate of `2e-5` and a batch size of `16`. We select the model with the highest F1 score on the development dataset, evaluated after each epoch. We keep the default parameters for all other values. We always train and evaluate three models using the seeds `(1,2,3)`. We did not fine-tune any hyperparameter. We provide code for reproduction. We run our experiments on a DGX A100.

Gold Label	*Leaked* Ev.-Only	*Leaked* Full	*Leaked* # Pairs	*Unleaked* Ev.-Only	*Unleaked* Full	*Unleaked* # Pairs	*Difference* Δ Ev.-Only	*Difference* Δ Full
true	33.3 ±1.2	33.6 ±3.9	39	45.0 ±0.8	44.9 ±1.7	93	-11.6 ±2.0	-11.3 ±2.5
mostly true	0.0 ±0.0	0.0 ±0.0	10	0.0 ±0.0	0.0 ±0.0	19	0.0 ±0.0	0.0 ±0.0
mixture	18.0 ±3.8	22.4 ±3.0	44	28.6 ±2.3	32.8 ±1.0	81	-10.6 ±5.9	-10.4 ±2.5
mostly false	6.2 ±5.5	10.8 ±4.0	36	4.4 ±3.9	9.4 ±7.2	40	+1.9 ±8.3	+1.5 ±4.1
false	85.8 ±0.4	86.2 ±0.3	353	73.2 ±0.9	77.7 ±2.0	299	+12.5 ±1.1	+8.5 ±1.7

Table 11: F1-score on **Snopes** (via MULITFC) using the evidence-only model using solely evidence (title & text) snippets, and the full model. We report the F1-score for each label on both splits (leaked and unleaked evidence), and their difference.

Gold Label	*Leaked* Ev.-Only	*Leaked* Full	*Leaked* # Pairs	*Unleaked* Ev.-Only	*Unleaked* Full	*Unleaked* # Pairs	*Difference* Δ Ev.-Only	*Difference* Δ Full
True	57.9 ±1.2	58.3 ±1.0	288	29.0 ±1.4	27.3 ±3.7	113	+29.0 ±1.0	+31.0 ±3.7
Mostly True	59.0 ±0.7	58.1 ±0.2	387	20.5 ±1.9	25.8 ±2.4	123	+38.5 ±2.6	+32.3 ±2.5
Half-True	58.3 ±1.3	58.4 ±0.3	405	28.6 ±2.4	30.9 ±2.6	132	+29.8 ±3.6	+27.5 ±2.3
Mostly False	56.2 ±1.2	56.8 ±0.4	364	18.2 ±6.2	17.1 ±0.7	97	+38.0 ±7.1	+39.7 ±0.3
False	58.4 ±0.9	58.8 ±1.1	419	19.0 ±3.0	26.4 ±0.9	102	+39.5 ±2.8	+32.4 ±0.9
Pants on Fire!	68.1 ±0.6	68.0 ±0.4	248	17.6 ±3.7	25.9 ±2.2	39	+50.5 ±3.1	+42.2 ±1.9

Table 12: F1-score on **PolitiFact** (via MULITFC) using the evidence-only model using solely evidence (title & text) snippets, and the full model. We report the F1-score for each label on both splits (leaked and unleaked evidence) and their difference.

C.2 Performance per Label

We show the F1 score for each label and both datasets in Table 11 (Snopes) and Table 12 (PolitiFact). The dataset of Snopes is highly imbalanced towards the majority class "false". The class imbalance is amplified within the leaked subset. The evidence-only model benefits from leaked evidence for (leaning) false claims. The performance drops on samples with evidence exhibiting leaked characteristics when the gold veracity tends towards true. The performance on the unleaked subset is slightly more balanced when comparing different labels. The majority label "false" is still dominating. Yet, the more balanced predictions across all labels yield a higher F1-macro score. On PolitiFact, a majority of claims across all labels contain leaked evidence. Models learn to exploit it for all labels. The best performance gain can be seen over claims misinformation claims. Yet, the more balanced predictions across all labels yield a higher F1-macro score.

C.3 Evaluation on Identical Claims with Different Evidence

To avoid the label distribution to distort the impact of leaked evidence, we evaluate the trained model only on claims that contain leaked *and* unleaked evidence snippets. We separately measure the performance of the evidence-only models on this subset, when only concatenating leaked evidence to the claim, and when only concatenating unleaked evidence to the same claim. The overall metrics on Snopes (Table 13) do not change much. The performance on the individual labels differs (even not always for the better) when comparing the prediction with leaked or unleaked evidence. We find that leaked evidence snippets are strong indicators for the model to predict the label "false". This comes at the cost of a lower recall across all other labels. On PolitiFact evidence snippets show great improvements and double almost every metric (Table 14).

C.4 Comparison with a Claim-Only Baseline

We show the detailed results of the *claim-only* baseline for all claims that contain leaked and unleaked evidence snippets in Table 15. All these samples are considered leaked, as they contain (amongst others) leaked evidence. The results align with our previous observations: Models trained on the Snopes subset rely on the majority class "false", yielding an overall high performance. We compare the results of the claim-only model (which by default cannot benefit from leaked evidence) with the results of the evidence-only model (compare Table 13) on the same subset. The claim-only model outperforms the evidence-only model on unleaked evidence. When the evidence-only model sees leaked evidence, it either performs on

	Leaked			*Unleaked*			
Gold Label	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Support**
true	**31.5**	14.0	19.0	17.0	**33.3**	**22.5**	38
mostly true	0.0	0.0	0.0	0.0	0.0	0.0	10
mixture	**28.1**	23.5	25.4	19.8	**37.1**	**25.8**	44
mostly false	0.0	0.0	0.0	**5.9**	**2.8**	**3.8**	36
false	77.3	**93.5**	**84.6**	**82.4**	73.9	77.9	353
Accuracy		**71.9**			60.5		481
F1-micro		**66.0**			61.6		481
F1-macro		25.8			**26.0**		481

Table 13: Precision recall and F1 of the evidence-only BERT model based on all claims of the **Snopes** dataset that contain leaked and unleaked evidence snippets.

	Leaked			*Unleaked*			
Gold Label	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Support**
true	**58.3**	**53.9**	**56.0**	21.1	23.8	22.3	288
mostly true	**61.4**	**54.5**	**57.0**	26.7	26.6	26.5	385
half-true	**46.7**	**61.6**	**52.8**	24.2	32.3	27.6	404
mostly false	**59.1**	**55.4**	**57.1**	21.8	18.4	19.7	360
false	**60.9**	**58.1**	**59.4**	26.4	28.5	27.4	419
pants on fire!	**75.0**	**63.0**	**68.5**	57.4	22.3	32.1	247
Accuracy		**57.6**			25.8		2103
F1-micro		**57.9**			25.8		2103
F1-macro		**58.5**			25.9		2103

Table 14: Precision recall and F1 of the evidence-only BERT model based on all claims of the **PolitiFact** dataset that contain leaked and unleaked evidence snippets.

	Snopes				*PolitiFact*			
Gold Label	**Precision**	**Recall**	**F1**	**Support**	**Precision**	**Recall**	**F1**	**Support**
true	41.8	33.3	36.9	38	25.6	25.6	25.5	288
mostly true	0.0	0.0	0.0	10	29.8	34.1	31.7	385
half-true / mixture	22.6	28.0	24.9	44	26.7	35.1	30.3	404
mostly false	7.3	1.9	2.9	36	23.0	21.5	22.2	360
false	80.7	88.7	84.5	353	28.2	23.9	25.9	419
pants on fire!	–	–	–	–	54.7	33.3	41.3	247
Accuracy		70.4		481		28.8		2103
F1-micro		67.4		481		28.9		2103
F1-macro		29.8		481		29.5		2103

Table 15: Precision recall and F1 of the *claim-only* BERT model based on all claims containing leaked and unleaked evidence.

par (F1-micro, accuracy) or worse (F1-macro) than the claim-only model. Here, the claim-only model is slightly less biased towards predicting "false", which improves the performance on the remaining labels. Both results indicate the reliance of the model on leaked evidence (even if not for the better). On the PolitiFact subset, the claim-only baseline performs well behind the evidence-only baseline (compare Table 14).