International Journal of Data Science and Analytics (2022) 13:299–314
https://doi.org/10.1007/s41060-022-00310-7

REGULAR PAPER

Beyond belief: a cross-genre study on perception and validation of health information online

Chaoyuan Zuo[1] · Kritik Mathur[1] · Dhruv Kela[1] · Noushin Salek Faramarzi[1] · Ritwik Banerjee[1]

Received: 27 April 2021 / Accepted: 31 December 2021 / Published online: 2 February 2022

Abstract
Natural language undergoes significant transformation from the domain of specialized research to general news intended for wider consumption. This transition makes the information vulnerable to misinterpretation, misrepresentation, and incorrect attribution, all of which may be difficult to identify without adequate domain knowledge and may exist even in the presence of explicit citations. Moreover, newswire articles seldom provide a precise correspondence between a specific claim and its origin, making it harder to identify *which* claims, if any, reflect the original findings. For instance, an article stating "Flagellin shows therapeutic potential with H3N2, known as Aussie Flu." contains two claims ("Flagellin … H3N2," and "H3N2, known as Aussie Flu") that may be true or false independent of each other, and it is *prima facie* unclear which claims, if any, are supported by the cited research. We build a dataset of sentences from medical news along with the sources from peer-reviewed medical research journals they cite. We use these data to study *what* a general reader perceives to be true, and *how* to verify the scientific source of claims. Unlike existing datasets, this captures the metamorphosis of information across two genres with disparate readership and vastly different vocabularies and presents the first empirical study of health-related fact-checking across them.

Keywords Natural language processing · Claim extraction · Check-worthiness · Cross-genre information retrieval · Fact-checking · Misinformation

1 Introduction

Health information-seeking behavior is increasingly reliant on the Internet, with the general population trusting online articles significantly more than other media such as radio or television [47,70]. Thus, it is critically important that news articles remain faithful to the medical findings they report. Even more so because information propagated on social media is often sourced from news coverage further upstream [72]. This concern has led to several qualitative and manual assessments of medical news vis-à-vis the original research publications [49–51,79]. In surveys where expert panels have judged the accuracy of reports, nearly half of all media coverage was found to be inaccurate, albeit often due to innocuous enthusiasm [51,95]. These inaccurate statements about medical information have also been attributed to overstating risks [10,36], exaggerated claims [11], and sensationalism [64,73]. Furthermore, when scientific research makes its way out of conferences and journals into mass media, the language in which the information is expressed undergoes drastic changes. The general reader is unprepared for specialist medical language comprehension [29], to the extent that changing the language to one meant for a wider "lay" audience has been treated as a discipline by itself [78]. So, while this change is necessary, it often results in the conversion of highly specific and nuanced scientific claims into what studies on scientific misinformation have termed "sound bites" [46,60].

The reader seldom has the means to determine if the medical information remains accurate after this conversion, and

✉ Chaoyuan Zuo
chzuo@cs.stonybrook.edu

Kritik Mathur
kmathur@cs.stonybrook.edu

Dhruv Kela
dkela@cs.stonybrook.edu

Noushin Salek Faramarzi
nsalekfarama@cs.stonybrook.edu

Ritwik Banerjee
rbanerjee@cs.stonybrook.edu

[1] Stony Brook University, Stony Brook, NY 11794-2424, USA

Table 1 An article citing and (mis-) quoting peer-reviewed research while presenting medical information. General trust in the publisher and the mere existence of the hyperlink (Bold) are powerful markers of credibility. The reader often trusts such information without further verification

Newswire claim
Similarly, the **thyroid drug levothyroxine** should be taken "on an empty stomach, one-half to one hour before breakfast."
Source: www.nytimes.com/2018/09/28/well/live/drug-medication-empty-stomach-prescription.html Published: Sep 28, 2018 *Accessed*: July 14, 2020
Cited research article
Title: Comprehension of Top 200 Prescribed Drugs in the US as a Resource for Pharmacy Teaching, Training and Practice
Source: www.ncbi.nlm.nih.gov/pmc/articles/PMC6025009

relies largely on the *perception* of credibility derived from (1) the authority of the source and (2) the existence of references to other external, seemingly credible, sources. This perception, however, is not always correct [95]. Table 1 shows one such example, where in spite of the general reputation of the news publisher and the existence of a link to an external credible source, the medical claim is entirely unsupported by that cited source.

We introduce a dataset that leverages markers of credibility in online news articles and address the following challenges to vetting health-related claims:

(1) Identify *what* is worth checking, in the specific context of a marker, and
(2) Whether the perception provided by the marker to the general reader is indeed true.

In Sect. 2, we discuss the markers of credibility in online medical news and how we use them to address these two challenges and discuss our study design and our dataset in Sect. 3. We then present our annotations, experiments, and findings for the two challenges in Sects. 4 and 5, respectively, before discussing our work in the context of related technical research in Sect. 6. Finally, we conclude with some remarks about future avenues of research in Sect. 7.

2 Perception and markers of credibility

Given that perceived credibility has an immense impact on how much the reader believes the message [58], its markers have been studied across various disciplines. For electronic media, and text in particular, the *perception* of credibility is based largely on two factors: the authority of the author and referrals to texts from credible external sources [15,22,56,62]. In health-related newswire, embedded citations to peer-reviewed research satisfy both conditions, serving as "proximal cues" in the immediate environment that provide markers of credibility for the reader [24,59].

Multiple studies, such as those conducted by Bråten et al. [8] and Kolstø [32], have demonstrated that contextual markers frequently determine the credibility of the text, more so when readers have low knowledge of the topic and are unable to distinguish between nuanced claims made in technical language. This is typical of medical news, given how difficult it is for a non-specialist to understand the language of medical research literature [29,92]. Thus, the mere existence of citations to peer-reviewed research amplifies the role played by the perception of credibility, since the reader cannot easily verify a claim by reading and understanding the cited peer-reviewed research publication.

It is not uncommon that peer-reviewed research is cited in medical news,[1] and this facilitates our empirical study. Given that a citation is embedded in the text of one document (a news article), and links to another (a research publication), vetting a claim may appear to be a simple task at first. Upon closer inspection, however, it becomes clear that there are major challenges. First, identifying *what to verify* is not obvious. Second, the language of news articles is vastly different from the language of medical research, so determining if a claim is supported by the cited research is a difficult task, akin to information retrieval and natural language entailment across distinct genres.

2.1 Identifying what to verify

Following the guidelines of the TREC 2006 Question Answering Track [31]—and thereafter the knowledge discovery, question-answering, and summarization communities within natural language processing research (see, for instance, Clarke et al. [13], Lin and Zhang [42], Sathiaraj and Triantaphyllou [69])—atomic pieces of domain-specific information are called "information nuggets." In spite of presenting medical findings in lay terms, newswire sentences are often complex, containing multiple such nuggets of information.

Moreover, there is seldom a direct correspondence between the primary claim being reported and the linguistic information in the embedded text. Some core nuggets are *perceived* by the reader as being upheld by the cited research while others play an auxiliary role. As a result, which claim(s) to verify may be *prima facie* unclear. Table 2 shows four examples illustrating some of the hurdles. Not all claims are the primary focus, and not all claims can be verified in light of the cited research. Further, the text spanned by the embedded hyperlink is not necessarily the primary claim, even though

[1] In our data collection, 15.3% of the articles provide such citations (see Sect. 3).

Table 2 Sentences with at least one primary claim worth verifying along with embedded citations (bold). Claims unsupported by the cited research are marked by a red asterisk (*). All sources last accessed on May 5, 2020

(A)	In a research published in the **Journal of the American Academy of Child and Adolescent Psychiatry**, researchers found that less parental warmth and having harsher home environments can contribute to how aggressive children become. [www.inquisitr.com/5115311/parenting-antisocial-behavior-children]
	Primary: less parental … become
	Auxiliary: research published in …Psychiatry
(B)	Flaxseed fiber reportedly helps **balance cholesterol** levels and **lower blood pressure**, among other benefits. [www.medicalnewstoday.com/articles/324604]
	Primary: (a) Flaxseed fiber… cholesterol
	(b) Flaxseed fiber helps… pressure
	Auxiliary: Flaxseed fiber has other benefits (*)
(C)	Health workers have been using a vaccine made by Merck, which has been **shown in field testing** to potentially reduce infection rates, and showed some success during the last outbreak in DRC. [https://time.com/5426847/democratic-republic-congo-ebola-outbreak/time.com/5426847/democratic-republic-congo-ebola-outbreak]
	Primary: The vaccine … reduce infection rates
	Auxiliary: (a) The vaccine is made by Merck (*)
	(b) It showed some success … DRC
(D)	Some experts say they can **enhance cognitive function** and boost problem-solving abilities, while other researchers **point out** that gamers have sedentary lifestyles and can experience mental health issues.
	[www.mentalfloss.com/article/523460/excessive-gaming-might-soon-be-recognized-official-disorder]
	Primary: (a) gaming can … cognitive function
	(b) gaming … problem-solving abilities
	(c) gamers have sedentary lifestyles
	(d) gamers … mental health issues

it is intended to serve as a credibility marker for *that* specific primary claim. For example, in (A), the citation is embedded into the location, while the reader perceives it as a credibility marker for the claim presented in the clausal complement. In many cases, identifying *what* information to check is further obscured because the choice of the text span may not have a clear pattern even within a single sentence, as shown in (D). There, the first embedded citation spans the verb and direct object of one claim while being intended as a marker for two claims, and the second citation spans the action of the "researchers" while being a marker for two additional claims. Syntactic complexities such as shared subjects—"Flaxseed fiber" in (B) and "gamers" in (D)—further add to the complexity.

2.2 Defining what to verify, and how

Whether or not a specific piece of information is worth checking for veracity is based on its perceived importance and debates abound regarding how much of it is a conscious process [85]. Regardless, studies support the subjectivity inherent in answering the question:

2.2.1 Is this piece of information worth checking?

Hanto and Tostrup [25] observe the answer to have higher disagreement across people from different backgrounds and age-groups. Their observations suggest that while those with domain knowledge might view something as common sense, others could find it worth checking.[2] Similarly, Konstantinovskiy et al. [33] find that check-worthiness of an information nugget is subjective and highly dependent on the context. They decouple the identification of information from its importance and its domain, such as "crime" or "health" (ibid. p. 6). Our work, on the other hand, is already specific to "health," and the context is entirely characterized by the citations. Further details of our annotation process are provided in Sect. 3.

Annotating nuggets of information should, of course, also answer the question

2.2.2 Can this piece of information be checked?

Starting with classical accounts in the philosophy of science, verification has been based on the dictum that a statement is upheld by empirical observations [18, p. 121], and the observations are themselves "independent of any subjective interpretation" [23]. However, in the rush to provide scalable fact-checking, relatively less attention has been paid to the quality, authority, or extent of the encyclopedic knowledge used. Fact-checking endeavors largely use public data while trusting the judgment of fact-checkers—automated or human—about the *choice* of evidence [26,27,65,81,88].

Qualitative research in philosophy and journalism had been critical of this due to potential epistemological bias [28,83]. This is especially pertinent for medical information, since public knowledge bases are known to be incomplete or inaccurate [34,35,45], and contradictory findings are plentiful [52]. Thus, the selection of observations is critical in establishing the veracity of a claim.[3] In newswire report-

[2] Hanto and Tostrup [25, p. 117] illustrate this with examples like "Norge har en lang kyst, og det tar minst tre døgn å seile den kysten fra ende tilannen." ("Norway has a long coast, and it will take at least three days to sail from one end to the other.")—a statement deemed check-worthy by people with a background in the humanities, but not by those from the natural sciences.

[3] Since no feasible fact-checker can claim to have investigated *every* datum publicly available, the final verdict—in the true sense of Bayesian

ing medical research, however, the use of hyperlinks as markers of credibility precludes navigating these difficult issues. Checking a claim naturally reduces to verifying it *with respect to* the authoritative context of the peer-reviewed publication explicitly cited by the author making the claim.[4] The next natural question in our pursuit is thus, *how* to check a piece of information. We explain this next, through our study design.

3 Study design and dataset description

Earlier works in identifying check-worthy claims were at the granularity of entire sentences [5,26,54,81]. In contrast, our study identifies specific nuggets of information from within a sentence (which may contain multiple such pieces). In the domain we investigate, the sentences are often longer and syntactically more complex. Figure 1 compares the distribution of the size of sentences in our medical newswire dataset to two other well-known fact-checking datasets, FEVER [81] and CLEF-2019 CheckThat! [5]. Furthermore, we seek to extract those claims that *appear* to be supported by a citation and can indeed be verified on that basis. With this as the backdrop, we break down the problem into a pipeline with two components:

(1) Extraction of check-worthy claims from sentences in medical newswire on the basis of perceived external support of peer-reviewed research, and
(2) Cross-genre claim verification across newswire and medical research literature.

An auxiliary goal in the first step is to understand whether the existence of multiple claims (along with their markers of credibility) in a single sentence makes information extraction harder. While this is not the focus of our work, and the findings may not be applicable to different kinds of markers, the experiments are nevertheless designed to shed some light on this matter.

3.1 Data collection

Given the dearth of empirical studies of medical fact-checking on the basis of specific context from research, we start by building a new dataset. Using the MediaRank project

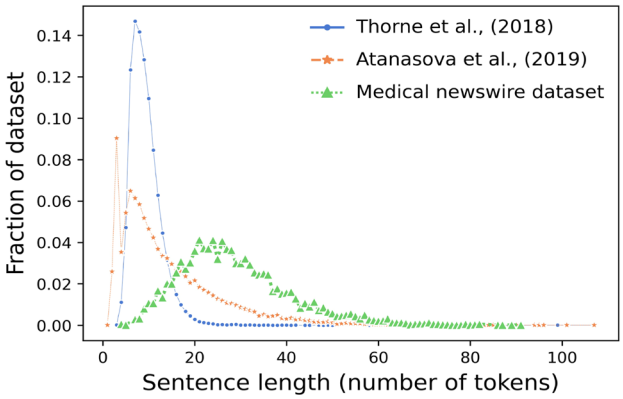

Fig. 1 Compared to other fact-checking datasets (FEVER [81] and CLEF-2019 CheckThat! [5]), sentences in medical newswire are long and complex, often positioning the primary claim(s) within a larger context of other information

[96], we obtain 6,000 news articles from the "Health" category of Google News during April 2018 and augment this with the top 25 RSS feeds in the "Health and Healthy Living" category[5] from November 2018 through April 2019 to get over 34,000 news articles. To exclude articles that do not cite peer-reviewed research, we filter out those without any links to the domains among the top science publications on Alexa or the Wikipedia list of medical journals.[6,7] A further filtering due to paywalls yields 6,195 articles from the initial collection of 40,516 (15.3%). Finally, we remove cases that require anaphora resolution across sentences, since that would introduce significant errors of its own (see Lee et al. [38]). The final dataset comprises 5116 sentences. Of these, 4882, 174, and 60 sentences have 1, 2, and 3 embedded citations, respectively.

3.2 Data annotation

After collecting these data, we first carry out an annotation task to obtain ground truth about what is perceived as verifiable and check-worthy. In this work, it is crucial for the ground truth to reflect how a reader of medical information from news articles, who is *not* a domain expert, perceives the claims made in such articles, along with the provided evidence. In some recent work, crowdsourcing studies have been conducted on medical information to compare the quality of annotation with expert-labeling. These studies—Roitero et al. [67], among others—are based on medical information that is already presented in "lay" terminology.

The annotation in this work, however, has a twofold requirement: (i) general reading comprehension of medical

confirmation—may very well depend on *what* and *how much* was considered as evidence. For a critique, we point the reader to the discussion of "underdetermination" [75].

[4] This aligns with the journalistic ethos of fact-checking organizations. For example, www.factcheck.org/our-process/: "The burden is on the person or organization making the claim to provide the evidence to support it." [Accessed: May 5, 2020].

[5] https://blog.feedspot.com/healthy_living_rss_feeds.

[6] www.alexa.com/topsites/category/Top/Science/Publications.

[7] https://en.wikipedia.org/wiki/List_of_medical_journals.

Table 3 Annotations on medical newswire claims perceived as verifiable and check-worthy, showing the number of sentences with dis/agreements. The two main types of disagreements in sentences with only one embedded hyperlink (bold) to peer-reviewed research are over the (1) inclusion of the post-modifier and (2) scope of the primary claim itself

	No. of embedded citations in a sentence			Total
	1	2	3	
Annotators agree	4757	151	51	4959
Annotators disagree	125	23	9	157
Disagreement (%)	2.56	13.22	15.00	3.07

(1) This may help to prevent delay **sarcopenia**, which is the decline of skeletal muscle tissue with aging

Annotator 1: "This may … with aging" (inclusion of complex post-modifier of sarcopenia)

Annotator 2: "This may … sarcopenia"

(2) Your body starts a fever because the flu virus **doesn't grow** as well at high temperatures, and some immune cells actually work better.

Annotator 1: "body starts … temperatures" (causality perceived as the primary check-worthy claim)

Annotator 2: "flu virus … temperatures" (only the effect perceived as the primary check-worthy claim)

research language and (ii) a lack of expertise in the biomedical sciences. The first requirement precludes annotators with less than university-level education [92]. As such, controlling for both the amount and the domain of education becomes important. The importance of these two factors in the annotation quality of complex tasks has been demonstrated by Kazai et al. [30]. Their work also finds that for complex tasks, the age-group of annotators is strongly correlated with the annotation quality, with the best quality provided by those who are 20–30 years old.

Taking all these factors into account, we choose three non-medical graduate students to work independently based on an annotation guideline document. The annotators are also first tested on sample data for quality assurance of the main annotation task. Furthermore, the information verification labeling is done on a 5-point Likert scale (described in Sect. 5) instead of a hard binary labeling, thereby making the task more tolerant toward annotator differences.

Our annotations are carried out using the BRAT tool [76], and each sentence provided to the annotators is supplemented by (1) the text segments corresponding to the embedded hyperlinks in a sentence, e.g., "balance cholesterol" from Table 2 (B), and (2) the citation URL. They are asked to mark only those claims that they perceive as being verifiable on the basis of the cited publication, and specify which hyperlink offers the perception of support for the claim. Since our data comprise sentences with at most three embedded citations, this specification is 1, 2, or 3. In the event that the perception is unverifiable, the annotation assigns a special value, NO_INFO. In 339 sentences, this value is assigned because a hyperlink takes the reader to an article entirely about the publication venue or the author(s), instead of a peer-reviewed article presenting the medical information reported in newswire.

We achieve excellent inter-annotator agreement, as shown in Table 3. There are disagreements in only 157 out of the 5,116 sentences (3.07%). A breakdown reveals that disagreements increase substantially to 13.22% and 15% when the annotators are given sentences with two and three embedded hyperlinks, respectively—a potential indication that the inclusion of multiple cues within a single sentence makes it difficult for readers to distill the claims and their corresponding evidence. An analysis of our empirical results corroborates this as well (see Sect. 4.2). Table 3 also illustrates two main types of inter-annotator disagreement in syntactically complex sentences, due to either a difference over the scope of the primary claim or the inclusion of post-modifiers.

3.3 Dataset description

The final dataset consists of 4959 sentences obtained from 2828 unique newswire articles published across 304 news agencies. It is not uncommon for an article to cite multiple research publications, and a majority of the citations in our dataset are to peer-reviewed medical research publications on PubMed. The distribution of citations over the sources is shown in Fig. 2.

The citations in newswire data are present as embedded hyperlinks. The position of these hyperlinks within the sentence does not follow any discernible pattern, but there is a slight preference to embed the link in the first half of a sentence, exhibited by 64% sentences in our dataset. In Fig. 3, we provide a more comprehensive picture of the relation between the sentences in our dataset and the check-worthy claims and hyperlinks within those sentences. Even though the sentences are long (as are many check-worthy claims in

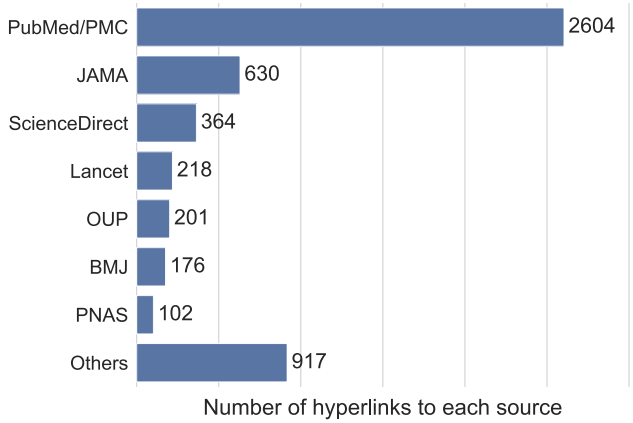

Fig. 2 Distribution of citations over publication sources. Only top ten shown for brevity, including "others"

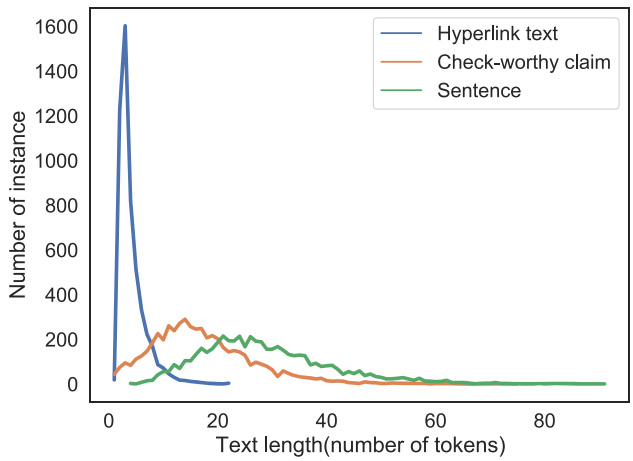

Fig. 3 The distribution of sentence lengths, lengths of the embedded hyperlink text spans, and the length of the check-worthy claims

those sentences), the embedded hyperlink rarely spans more than four words, resulting in the sharp peak.

Finally, before delving into the first component of our pipeline, we contrast two word clouds in Fig. 4, showing the prevalence of redundant words in newswire sentences that contain important medical information. Without the extraction of check-worthy claims from within these sentences, claim verification is likely to suffer due to the frequent occurrence of boilerplate terms like "published," "found," "study," etc.

4 Check-worthy claim extraction

The problem of extracting check-worthy claims from newswire sentences is devised as a sequence labeling task. To serve as the baseline model, we fine-tune pretrained BERT embeddings on our task—which can be done by adding just one output layer [17]. For this, we train for five epochs with a batch size of 32. The maximum sequence length and the learning rate are set to 128 and 5×10^{-5}, respectively. We then use the Flair framework [2] with the BiLSTM + CRF architecture (Fig. 5) for token classification, motivated by its success in flat named entity recognition tasks [77].

Our experiments are categorized into the use of three types of word embeddings. First, we use two classical pretrained word embeddings: the general GloVe model [57] and BioNLP [61] trained on PubMed articles and Wikipedia text. Second, we use contextual word embeddings—Flair [1], and two pretrained models based on BERT's architecture, RoBERTa [44] and BioBERT [37]. Thus, with both the contextual and the classical models, our experiments cover generic embeddings as well as domain-specific ones. Third, for stacked embeddings, we test GloVe + RoBERTa, BioBERT + RoBERTa, and BioBERT + Flair. Finally, predicated on the idea that the piece of text onto which a hyperlink is embedded may be especially significant, we add the positional information of the hyperlink by adding $[\pm 1]$ for each token, depending on whether or not the token is in that piece of text. Since we have no need for nested representations, all the data are tagged using the BIO (acronym for "beginning, inside, outside") scheme, proposed first by Ramshaw and Marcus [63] for phrase chunking tasks.

For all our experiments, we discard the 157 sentences on which annotators disagreed (see Table 3). Our investigation is then divided into

$\mathcal{D}_1$: a collection of the remaining 4,757 sentences with a single citation, and

$\mathcal{D}_2$: the above collection, *plus* the sentences with multiple citations.

In the latter, a sentence with multiple hyperlinks is present multiple times, *once per citation*. Thus, a sentence with two embedded hyperlinks (Table 2 (B), for example) appears twice, as

(1) Flaxseed fiber reportedly helps balance cholesterol levels and lower blood pressure, …

(a) Entire sentences (b) Check-worthy claims

Fig. 4 Prominent words (size proportional to a word's frequency) in **a** complete sentences containing check-worthy claims, and **b** check-worthy claims alone

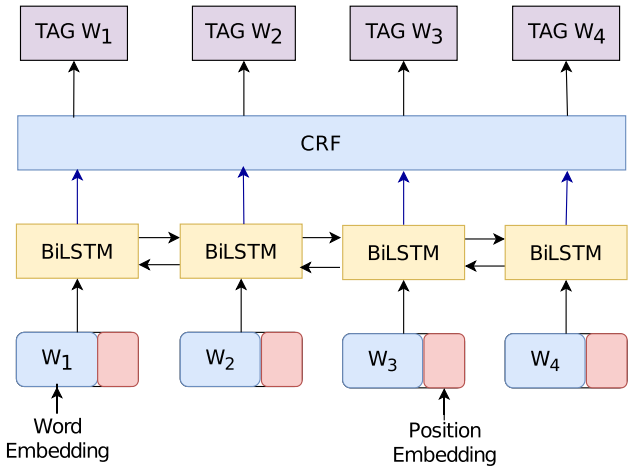

Fig. 5 The BiLSTM+CRF architecture, with pretrained word embeddings serving as the input

Table 4 Training, development, and test sets for claim–extraction. $\mathcal{D}_1$: sentences with a single embedded citation, and $\mathcal{D}_2$: sentences with multiple citations are included, but repeated with one citation per copy

	Number of sentences	
	$\mathcal{D}_1$	$\mathcal{D}_2$
Training	3550	3868
Development	627	695
Test	580	549
Total	4757	5212

(2) Flaxseed fiber reportedly helps balance cholesterol levels and lower blood pressure, …

In both cases, the data are divided into training, development, and test sets (Table 4). Each model (other than the baseline) has two hidden LSTM layers, each with 256 features, and we use identical hyperparameters across the board for mini-batch gradient descent, as follows: a learning rate of 0.1, batch size of 32, and with the maximum number of epochs set to 60.

Table 5 reports the precision, recall, and F_1 score[8]. We present (1) a *strict* evaluation, where a prediction is counted as a true positive if and only if the boundaries are an exact match with the ground truth, and (2) a somewhat *relaxed* evaluation, where incorrect inclusion or exclusion of surrounding punctuation is ignored.

4.1 A discussion and analysis of the results

We choose BERT as the baseline due to its state-of-the-art performance reported by Zuo et al. [99] in the check-worthiness task of the CLEF-2018 CheckThat! Lab [54], and find that on $\mathcal{D}_1$, it has good recall, but poor precision. Thus, even though it outperforms all other classical word embeddings, it will very likely harm subsequent verification by providing claims that were not intended as verifiable with respect to the cited research. In terms of precision and F_1 score, the contextual and stacked embeddings perform better. This is unsurprising, given BERT's performance on similar tasks when the sentences have complex syntax [99, p. 281].

In line with the evaluation methodology in other sequence labeling tasks [41,43], we use approximate randomization of the paired t test [97] to determine statistically significant changes, and reject the null hypothesis if $p \geq 0.05$ across 100 trials.

Adding the hyperlink's position to the BiLSTM-CRF network leads to a marginal performance benefit across all models, but this improvement is not statistically significant. The benefits of non-strict evaluation are not statistically significant either, indicating that in general, the errors may not be attributed to incorrect inclusion/exclusion of surrounding punctuation.

In an attempt to capture the performance of the various models across general and domain-specific embeddings

Table 5 Claim extraction results on the test set (models marked with $*$ use the position embedding), showing the Precision, Recall, and F_1. Pretrained BERT fine-tuned on the training set, but without the BiLSTM-CRF layer, serves as the baseline (marked with ^{b})

Embedding	Strict			Relaxed		
	P	R	F_1	P	R	F_1
			$\mathcal{D}_2$			
Classical pretrained embeddings						
GloVe*	70.0	68.7	69.3	71.1	70.3	71.0
BioNLP*	71.9	68.2	70.0	73.9	70.2	72.0
Contextual embeddings						
Flair*	77.5	74.2	75.8	78.2	75.0	76.6
RoBERTa*	78.5	75.1	76.8	79.3	75.9	77.5
BioBERT*	**79.8**	74.7	77.1	**80.4**	75.3	77.8
Stacked embeddings						
Glove+RoBERTa*	79.4	75.1	77.2	80.2	75.9	78.0
BioBERT+RoBERTa*	79.0	76.0	77.5	79.6	76.6	78.1
BioBERT+Flair*	78.7	**77.4**	**78.0**	79.6	**78.3**	**78.9**
			$\mathcal{D}_1$			
BERT^b	72.3	80.7	76.3	73.7	81.7	77.5
Glove	72.8	74.4	73.6	74.5	76.1	75.2
GloVe*	73.6	73.7	73.6	75.1	75.3	75.2
RoBERTa	79.9	81.1	80.5	81.1	82.4	81.7
RoBERTa*	82.1	81.5	81.8	83.1	82.5	82.8
BioBERT+Flair*	**83.9**	**83.2**	**83.6**	**84.6**	**83.9**	**84.3**

The best results are in bold

[8] The precise definitions we follow are in accordance with the convention set by the CoNLL-2003 shared task on named entity recognition [82].

while maintaining brevity, we report a subset of the models for $\mathcal{D}_2$ in Table 5. The stacked embedding of BioBERT + Flair achieves significantly better performance over the baseline. Relative to other contextual embeddings such as RoBERTa, however, the benefits of stacking a domain-specific embedding are marginal.

4.2 Error analysis

Over 32% of the errors were caused by minor differences such as the inclusion of adverbs and conjunctions (e.g., "also," "even") as part of the claim. A fourth of the errors were Type I errors where semantically similar words from a different part of the sentence were mistakenly identified as a claim worth verifying based on the citations. Another 30% were Type II errors, a large fraction of which were due to only the post-modifiers being identified as the claim while mistakenly excluding the primary entity.

We also find that in spite of the relatively low number of sentences with more than one citation (4.57% of the total dataset), their inclusion causes a statistically significant drop in the performance of every model. Analogously, the inclusion of sentences with two citations had immediately increased the inter-annotator *dis*agreement rate from 2.56% to 13.22%.

Thus, even though state-of-the-art contextual embeddings perform well in this task with just one marker (i.e., a single embedded hyperlink to peer-reviewed research) in a sentence, identifying concrete claims worth verifying on the basis of multiple such markers is, in general, a difficult task and may not be obvious even for human readers.

5 Cross-genre claim verification

In this work, it is reasonable to limit the scope of information verification to the abstract of a peer-reviewed research publication, under the assumption that an abstract is summarizes its main findings. As such, we collect the abstracts instead of the entire publications. Since nearly half the citations in our data collection are not open access, this is arguably a better representation of what the general reader finds readily available for verification. To study whether the claim identified in the news article is, indeed, supported by the cited research, we form claim–abstract pairs and then split the abstract into its sentences, obtaining a set of claim–sentence pairs.

The first step is to obtain ground truth. For the claim–sentence pairs, the following triple serves as the input for an annotator: a newswire sentence **n**, a specific claim **c** in that sentence, and a sentence **s** from the abstract of the cited research publication. The task itself is to score each triple (**n**, **c**, **s**) on the 5-point Likert-type scale shown in Table 6. We opt for this, instead of a binary *true-or-false* rating, in response to the nuanced picture of science misinformation discussed in notable prior work [29,71]. Our initial inspection, too, showed that a binary labeling would be a gross oversimplification of the problem.

Table 6 Likert-type rating scale used in the annotation task for cross-genre claim verification

Score	Relation between the sentence from the abstract of the cited research publication (**s**) and the claim from the news article (**c**)
1	**s** and **c** are completely unrelated, no inference is possible
2	**s** Does not describe the same event as **c**, but there are shared entities (usually, some relevant properties of those entities are being described)
3	**s** Does not describe the same event as **c**, but **c** may still be inferred from **s** (typically based on expert domain-knowledge)
4	**s** Contains some of the information in **c**, but some details are missing and may possibly be inferred (typically based on expert domain-knowledge)
5	**s** Contains all the information in **c**, and thus, **c** can be immediately inferred from **s**

Verifying claims presented in specialized medical language is a grueling task for non-specialists, so we proceed with the ground truth on a subset of the collection $\mathcal{D}_1$, comprising 1,652 triples (corresponding to 203 unique citations). Two raters work independently to create ground-truth scores for the claim–sentence pairs[9]. The claim–abstract pairs are labeled as *supported*, *unsupported*, or *uncertain*. To obtain these labels, we employ a third reviewer with domain-knowledge to adjudicate in case of disagreements. Instead of simply labeling the abstracts vis-à-vis the claims, we design our experiments first at the granularity of sentences to offer some explainability to our models.

5.1 A heuristic approach to labeling

Before diving into the main study, we discuss a natural heuristic for assigning one of three categories to a claim–abstract pair. This follows immediately from the claim–sentence scores:

1. The claim is *supported* by the research if the highest score assigned is 5.
2. The claim is *unsupported* by the research if the highest score assigned is 1 or 2.
3. If the highest score assigned is 3 or 4, the annotators are *uncertain* if the claim may be inferred from the research findings. These may require clarifications from a domain expert.

[9] Strictly speaking, the triple (**n**, **c**, **s**) is scored, but the newswire sentence **n** only provides context to the rater and plays no role in offering scientific support to the claim **c**. So, we refer to this as the score for the claim–sentence pair (**c**, **s**).

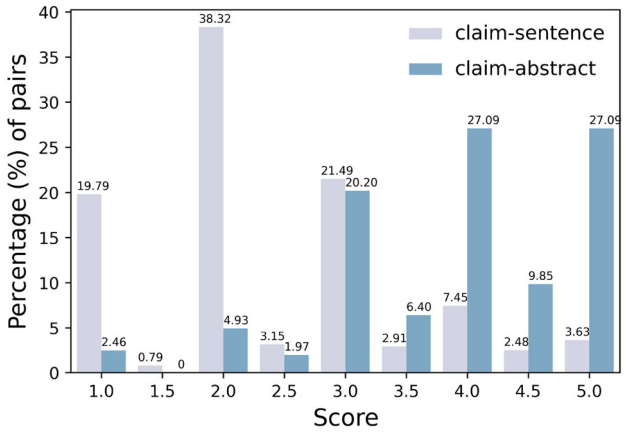

Fig. 6 The distribution of scores indicating how well a claim from newswire is supported by (1) a specific sentence from the abstract of the cited peer-reviewed research, and (2) the entire abstract of that publication

Table 7 Pairs of (1) a claim (italics) perceived by readers as being supported by a cited research, and (2) a sentence from the cited research

(A)	*High protein intake can have mixed results for people with type 2 diabetes*
	Substituting 5% energy intake from vegetable protein for animal protein was associated with a 23% (95% CI: 16, 30) reduced risk of T2D.
(B)	*Split-squats had the highest impact on the gluteus maximus, compared with deadlifts and good-mornings*
	Hamstrings were loaded isometrically during good-mornings but dynamically during deadlifts.
(C)	*Imbalance in intestinal bacteria may cause the inflammation that occurs in people with UC*
	It is believed that genetic factors, host immune system disorders, intestinal microbiota dysbiosis, and environmental factors contribute to the pathogenesis of UC

We assign the mean of the scores given by both annotators to each claim–sentence pair, and the aggregate of this value over all the sentences in an abstract is taken as a measure of how well the claim **c** is supported by the abstract of the cited peer-reviewed research.

5.2 Annotation results and implications

We measure the consistency of the annotation using the weighted kappa coefficient κ_w [14] and achieve a high inter-rater agreement ($\kappa_w = 0.916$). In the 1652 claim–sentence pairs, there were 157 cases where the raters differed, and in each case, they differed by unit score. It is worth noting that, had the raters differed wildly, the heuristic proposed to score claim–abstract pairs would not be meaningful. Figure 6 shows the distributions of the claim–sentence scores and the heuristic-based claim–abstract scores.

On the sentences where both raters agree, these scores have positive and negative skewness (0.438 and −0.270), respectively. This observation of a positive skew is expected for sentences, since even if an abstract supports a claim, a majority of its constituent sentences will not, just by itself, entail the claim.

The relatively low negative skew of the distribution of the claim–abstract pair scores has unexpected implications, however. Only 27.09% of all claim–abstract pairs were given the highest score of 5 by both raters, which rises to 36.95% if we consider the highest score given by at least one rater. On the other hand, 7.39% got a score of 2 or lower (*unsupported*). Given that these are newswire articles that explicitly cite supporting research, this is a surprisingly high number (and rises to over 10% if we include cases where at least one rater thought the cited research does not support the claim). We can also posit that the existence of these citations does not spontaneously offer a means of verification for the general reader, since a little over half (53.69%) the claims left the raters uncertain, likely requiring further explanations from domain experts. (These correspond to the scores 3 and 4 in Table 6.)

The scores also indicate that *if* there is an accurate model to discriminate between supported and unsupported claims, the volume of work for manual fact-checking will reduce by at least a third. Medical domain experts are expensive to employ and are often pressed for time. Thus, there are significant benefits to building such a system for medical information presented in newswire. This provides yet another impetus for our technical experiments (shown in Sect. 5.4).

5.3 A qualitative analysis of disagreements

The most conspicuous inter-rater disagreements are due to the presence of specific partial information in a sentence, while the claim itself is rather general (e.g., Table 7 (A)). In conjunction with several other sentences, however, the sum of these pieces of information does, indeed, entail the claim. There are 48 such cases in our data, and the two raters differ in assigning a score of 3 or 4 to these claim–sentence pairs. On the other hand, disagreements between a score of 2 and 3 lead to 52 pairs having different scores. These differences are mostly caused by the presence of common entities between the claim and the sentence, while at the same time, the language is such that inferring (or rejecting) the claim requires specialized domain knowledge, as shown with example (B) in Table 7.

Similarly, 41 claim–sentence pairs differed in getting a score of 4 or 5 from two annotators. This was often traced to both raters having an adequate understanding of the medical entities, but being unsure of a biologic process in *that specific context*. Table 7 (C) is an exemplary case: One rater was

Table 8 Size of the training, development, and test sets for cross-genre (newswire and medical research literature) verification of claims

	Number of claim–abstract pairs			
	Supported	Unsupported	Uncertain	Total
Training	28	7	45	80
Development	15	3	23	41
Test	25	6	51	82
Total	68	16	119	203

unsure if "inflammation" can be inferred from "pathogenesis," even though the rater knew the meaning of both terms in a general sense. In this particular example, the presence of linguistic hedging ("may" in the claim and "it is believed" in the sentence) aligns perfectly with the fact that inflammation is implicated in pathogenesis, but not a certainty. Thus, a score of 5 is appropriate.

5.4 Experiments and evaluation

Predicting the score of a claim–sentence pair is formulated as regression learning with target set $[1, 5]$ aligned with the Likert-type scale. We use three pretrained models with Transformers [91]: BERT, BioBERT, and XLNet [94]. Given that a claim will bear some semantic similarity to the evidence supporting it [4,48], we fine-tune the three models on the semantic textual similarity (STS) benchmark [12]. For domain-specific knowledge, we also then fine-tune them on the MedSTS dataset [89].

We use an identical set of hyperparameters to train all models with mini-batch gradient descent: a batch size of 32, maximum sequence length of 128 tokens, and a learning rate of 2×10^{-5}. The number of epochs is varied from 3 to 40 and chosen based on the minimizing the mean squared error (MSE) on the development set. Table 8 shows the split into training, development, and test sets.

The ranking models are evaluated in terms of their MSE on the test set. Keeping with the evaluations on prior semantic similarity tasks, we also report Pearson's correlation coefficient (PCC) between a model's predictions and human judgment. We then plug their predictions into the heuristic approach (Sect. 5.1), with the rounding function $f(x) = \frac{1}{2}\lceil 2x \rceil$, where x denotes the regression model's output. We report the micro-averaged accuracy, and given the class imbalance between the three labels, also the weighted average of precision, recall, and F_1 (Table 9).

5.5 Discussion and analysis

Tuning BERT embeddings on the STS and MedSTS datasets did not improve the MSE. BioBERT, which is pretrained on medical corpora, performs significantly better. In this case, fine-tuning on STS leads to mild improvements, but further tuning on MedSTS does not. When fine-tuned on both the STS and MedSTS datasets, the best sentence-ranking results are achieved by XLNet. Here, the best results of BioBERT and XLNet are very similar. The difference, however, lies in that while BioBERT performs well largely due to its domain-specific pretraining, XLNet achieves comparable performance through fine-tuning on a relatively small amount of domain-specific information. Our findings align with prior results on the STS benchmark [94, pp. 7–8] and underscore how entailment is cued by similarity, even across genres with rather distinct vocabularies.

Table 9 Claim–verification results, with the models fine-tuned further on STS (∗) and MedSTS (†), evaluating the ranking of sentences in a cited research by the *mean squared error* (MSE) and *Pearson correlation coefficient* (PCC). The subsequent classification results Precision, Recall, F_1, Accuracy) are shown in italics

	MSE	**PCC**	***P***	***R***	$\boldsymbol{F_1}$	**Acc**
BERT	0.865	0.506	*59.8*	*62.2*	*59.1*	*62.2*
BERT*	0.851	0.536	*57.4*	*58.5*	*57.0*	*58.5*
BERT*†	0.832	0.540	*57.3*	*57.3*	*55.4*	*57.3*
BioBERT	0.518	0.729	*68.8*	*69.5*	*69.1*	*69.5*
BioBERT*	**0.491**	**0.743**	**80.1**	***80.5***	***80.2***	***80.5***
BioBERT*†	**0.490**	**0.743**	**79.0**	***79.3***	***79.0***	***79.3***
XLNet	0.646	0.663	*64.2*	*63.4*	*61.4*	*63.4*
XLNet*	0.522	**0.744**	*66.5*	*62.2*	*63.0*	*62.2*
XLNet*†	**0.489**	**0.748**	*69.8*	67.1	67.6	67.1

Significant results are highlighted in bold

One might ask, *are the standard measures of semantic similarity sufficient to understand cross-genre entailment?* A quick glance only at the results of the regression task may be misleading, since both XLNet and the domain-specific embedding, BioBERT, perform similarly upon tuning with the two benchmark datasets. There is a stark difference in claim–abstract classification results, however, largely due to XLNet consistently assigning higher scores to claim–sentence pairs than BioBERT. This leads to XLNet labeling many claim–abstract pairs as *supported*, while human readers perceived them as *uncertain* (Fig. 7).

Further, when claims contain information relayed in common terms, every model increasingly fails to identify "unsupported" claims. In such cases, the claim is highly similar to several sentences in the research abstract. Medical facts expressed using complex domain-specific terms remain unaffected by this behavior.

We specifically investigate the six cases in our test set where the claim is decidedly *not supported* by the cited research. Both BioBERT and XLNet, when fine-tuned on STS and MedSTS, succeed in all but one datum. Thus, our

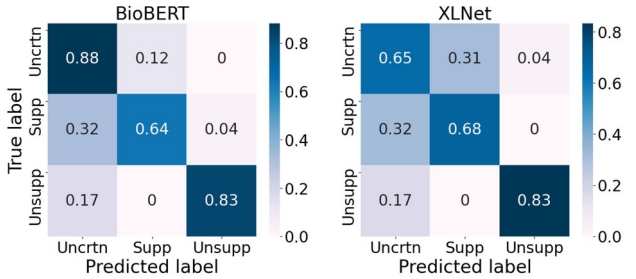

Fig. 7 BioBERT and XLNet (both fine-tuned on STS and MedSTS benchmarks) results, showing the percentage of true labels classified across the three categories

models are capable of learning a general readers perception of citations that seem to offer no support to the presented claim. This, however, does not cover a mistake where the link is clearly incorrect (as we illustrated with Table 1). To investigate whether such mistakes can be accurately detected, we randomly pair claims with abstracts (minus the correct citation). Of these 203 incorrect citations, our best model fails only on one. Most claim–sentence pairs were assigned a score less than two.

5.6 Limitations and future work

Being the first cross-genre study on information verification for medical news obtained directly from research findings, we hope this study opens up a pathway into further research in this direction. The work presented here is necessarily limited to the scope of a single publication, but opens up questions that warrant further and deeper investigations.

Since nearly half the cited research publications are not open access, our confinement to information verification with respect to the abstracts is arguably a reasonable initial approximation of what a general reader often faces if and when wanting to verify a claim. While our approach is in line with a large body earlier work on information extraction from medical corpora—e.g., knowledge discovery from Medline abstracts [84], or the creation of knowledge bases from abstracts for potential downstream use [87]—the use of abstracts instead of entire articles leaves out the possibility that secondary findings not mentioned in an abstract might be reported in newswire. It also raises questions about what details of a study, and to what extent, can possibly be gleaned from abstracts alone. In spite of this limitation, we believe our work—much like the earlier body of research done based on abstracts alone—can still find real utility in future research in cross-genre misinformation detection. Moreover, our technical strategy can already be extended to full research articles, by splitting the article into sentence chunks, and conducting the same sentence-level analysis as discussed in this work.

The use of annotators may lead to a second path of research based on our work. We have used graduate students with no formal medical training in order to mimic a cohort that controls for age, level of education, and domain of education. Our choices have been guided by the findings of crowdsourcing studies on complex tasks as well as reading comprehension studies specifically on medical research articles [30,92]. Other studies, as done by Roitero et al. [67], have shown that annotations on medical information done by non-experts are comparable to those done by experts. But these studies have used data that are already expressed using common terminology. It remains to be seen if similar results are true for medical research language as well. Kazai et al. [30] have also demonstrated that the geographic location and personality type of annotators also have a significant impact on annotation quality. Exploring how information verification is influenced by the various traits of the annotators is beyond the scope of our current work, but we hope that our discussion here kindles further research along this avenue.

6 Related technical work

In Sect. 2, we presented a discussion of prior research on the psychological aspects of trust and credibility of information. Independent of that approach, there has been considerable amount of work in natural language processing, even purely from a computational standpoint, aimed at the detection of misrepresentation of information, or fake news. Here, we present a brief overview of this body of work.

6.1 Coarse fact-checking based on perception

Scientists in various fields—medical research in particular—have decried the misrepresentation of their work in news [71,95]. Despite this, a majority of the research on fact-checking has focused on general knowledge or political narratives. Some rely on fact-checkers who assign a *general* rating to a Web site in its entirety based on an *aggregate perception* of bias and credibility [55]. As we have shown throughout this work, medical fact-checking cannot afford to depend on such perceptions. As to social consequences of misinformation, it is the reader immersed in information consumption whose perception ultimately matters, and this is best understood via ground truth from "lay" readers of news. Various tasks and approaches have been constructed along this line, and multiple datasets have also been put forth [21,27,54,80,81,88].

Neither identifying a claim nor verifying it have, however, been tackled at a granularity finer than entire sentences. Since support may vary across different components of a single sentence, there is a need to distill these nuggets of information.

6.2 Fact-checking with encyclopedic knowledge

Verifying information usually requires external knowledge, and the choice of knowledge may critically affect the *perceived* veracity of information (Sect. 2.2). Particularly for medical information, knowledge bases (KBs) like Wikipedia—as well as domain-specific ones like UMLS—are incomplete and inaccurate even for not so new information [6,35]. Moreover, when novel findings are first disseminated through news, these KBs do not contain that information. Thus, dependence on such KBs or even fact-checking Web sites [53,81,86] is not suitable for verifying new medical information in newswire. Instead, verification straight from a direct authority becomes necessary.

In our work, this authority is marked by a citation, connecting newswire to the medical research literature. There is no prior computational work on fact-checking across these genres, but an *ad rem* comparison to argumentation mining is in order.

6.3 Argumentation mining from text

Given a discourse structure such as persuasive essays or domain-specific texts like Wikipedia articles, claims can be extracted with promising accuracy [20,66]. There is some work in identifying claims across domains as well. For example, Rosenthal and McKeown [68] use relatively similar data stemming from social media to connect claims in blogs to Wikipedia discussions. Others have used discourse-level models across domains [3,16], but the models used in these studies are highly dependent on the genres and do not translate other types of texts [74]. Dusmanu et al. [19] approach cross-genre fact-checking by connecting claims made in Twitter to their news sources. Their work, however, uses a mix of explicit citations with other tweets.

Furthermore, the above body of work remains limited to mining claims at the coarser granularity of sentences, tweets, or entire articles. A few (notably Levy et al. [39] and Levy et al. [40]) delve into claim extraction from complex sentences, but without further investigations into other domains or genres.

6.4 Fake news detection in health-related claims

With the rise of the COVID-19 pandemic, a few datasets about misinformation specifically related to COVID-19 have been put forth. For example, Brennen et al. [9] identify some main types, sources, and claims of such COVID-19 misinformation, using a sample of 225 claims to demonstrate the diversity in false claims.

Within this body, Zhou et al. [98] present data across two different genres—news and Twitter, somewhat in the spirit of our cross-genre data. Their work, however, is on combating the spread of false claims. To this end, they collect 2,029 news articles along with 140,028 tweets to analyze the spread of those articles on social media, and build models to predict the credibility of a claim made in Twitter by using the credibility of news sources. Indeed, a majority of the prior work in pandemic-related misinformation relies on perceived credibility of a source, instead of verifying with respect to a specific scientific authority.

7 Conclusion

In his prescient and widely celebrated critical work, *The Image*, historian Daniel Boorstin writes

> It is more important that a statement be believable than that it be true. [7, p. 289]

To what extent news articles fit his observation remains to be seen, but we have explored the question of what is *believable* by investigating how the non-specialist reader may distill specific information conveyed within long and complex sentences, when provided with embedded citations as markers of credibility. Since recent empirical studies on health-related misinformation have exclusively focused on the COVID-19 pandemic, but have not delved into medical misinformation in general, we chose medical newswire articles for this work.

Medical information is a critically important domain that warrants deeper investigations into credibility, trust, and misrepresentation of information, especially given that people often need help evaluating health information [93]. In light of very difficult readability of medical research [90], the mere existence of source is not enough. So we have vetted those claims against the cited peer-reviewed research and studied *if* and *to what extent* a general reader is able to verify medical claims propagated in news.

This is the first quantitative work in fact-checking and argument mining that investigates medical facts at the critical juncture when they first appear into general public awareness, thereby cutting across two very distinct genres. During our data collection, we discover not all health-related news articles provide citations for the claims they propagate. Moreover, we find that *even with explicit citations*, news articles may sometimes mislead about medical findings. The technical core of our work is the development of models capable of identifying these cases. Due to our choice of domain, the cross-genre nature of our study, and the fine-grained annotation, our dataset can be used to study medical misinformation in explainable ways distinct from previous "fake news" benchmarks. Investigating the perception of credibility by moving beyond easily comprehensible genres is a key step to help readers become more discerning of medical claims that would ordinarily be viewed as believable, and

foster healthy skepticism especially during times of social unrest and emergencies.

Funding This work was supported in part by the Division of Social and Economic Sciences of the U.S. National Science Foundation (NSF) under the award SES-1834597.

Data availability Available at http://github.com/chzuo/jdsa_cross_genre_validation.

Declarations

Conflict of interest The authors declare that they have no conflict of interest that are relevant to the content of this article.

Code availability To be released.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. In: Proc. of the 27th Int. Conf. on Comput. Linguistics, Assoc. for Comput. Linguistics, Santa Fe, New Mexico, USA, pp 1638–1649 (2018)
2. Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., Vollgraf, R.: FLAIR: An easy-to-use framework for state-of-the-art NLP. In: Proc. of the 2019 Conf. of the North Am. Chapter of the Assoc. for Comput. Linguistics (Demonstrations), Assoc. for Comput. Linguistics, Minneapolis, Minnesota, pp 54–59, (2019) https://doi.org/10.18653/v1/N19-4010
3. Al-Khatib, K., Wachsmuth, H., Hagen, M., Köhler, J., Stein, B.: Cross-domain mining of argumentative text through distant supervision. In: Proc. of the 2016 Conf. of the North Am. Chapter of the Assoc. for Comput. Linguistics: Hum. Lang. Technol., Assoc. for Comput. Linguistics, San Diego, California, pp 1395–1404, (2016) https://doi.org/10.18653/v1/N16-1165
4. Alonso-Reina, A., Sepúlveda-Torres, R., Saquete, E., Palomar, M.: Team GPLSI. approach for automated fact checking. In: Proc. of the Second Workshop on Fact Extraction and VERification (FEVER), Assoc. for Comput. Linguistics, Hong Kong, China, pp 110–114, (2019) https://doi.org/10.18653/v1/D19-6617
5. Atanasova, P., Nakov, P., Karadzhov, G., Mohtarami, M., Martino, G.D.S.: Overview of the CLEF-2019 checkthat! lab: Automatic identification and verification of claims. task 1: Check-worthiness. In: Working Notes of CLEF 2019 - Conf. and Labs of the Evaluation Forum, CEUR-WS.org, Lugano, Switzerland, CEUR Workshop Proc., vol 2380 (2019)
6. Azer, S.A.: Is Wikipedia a reliable learning resource for medical students? Evaluating respiratory topics. Adv. Phys. Edu. **39**(1), 5–14 (2015)
7. Boorstin, D.J.: The Image?: A Guide to Pseudo-Events in America. Harper, New York (1962)
8. Bråten, I., Strømsø, H.I., Salmerón, L.: Trust and mistrust when students read multiple information sources about climate change. Learn. Instr. **21**(2), 180–192 (2011)
9. Brennen, J.S., Simon, F., Howard, P.N., Nielsen, R.K.: Types, sources, and claims of covid-19 misinformation. Reuters Inst **7**, 3–1 (2020)
10. Brown, J., Chapman, S., Lupton, D.: Infinitesimal risk as public health crisis: news media coverage of a doctor-patient HIV contact tracing investigation. Social Sci. Med. **43**(12), 1685–1695 (1996)
11. Caulfield, T.: The commercialisation of medical and scientific reporting. PLoS Med. **1**(3), e38 (2004)
12. Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L.: SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In: SemEval-2017, pp. 1–14. ACL, Vancouver, Canada (2017)
13. Clarke, C.L., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I.: Novelty and diversity in information retrieval evaluation. In: Proceedings of the 31st annual international acm sigir conference on research and development in information retrieval, Association for Computing Machinery, New York, NY, USA, SIGIR '08, p 659-666, (2008) https://doi.org/10.1145/1390334.1390446
14. Cohen, J.: Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychol. Bull. **70**(4), 213 (1968)
15. Corritore, C.L., Wiedenbeck, S., Kracher, B., Marble, R.P.: Online trust and health information websites. Int. J. Tech. Hum. Interact. **8**(4), 92–115 (2012)
16. Daxenberger, J., Eger, S., Habernal, I., Stab, C., Gurevych, I.: What is the essence of a claim? cross-domain claim identification. In: Proc. of the 2017 Conf. on Empirical Methods in Nat. Lang. Process., Assoc. for Comput. Linguistics, Copenhagen, Denmark, pp 2055–2066, (2017) https://doi.org/10.18653/v1/D17-1218, https://www.aclweb.org/anthology/D17-1218
17. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the 2019 Conf. of the North Am. Chapter of the Assoc. for Comput. Linguistics: Hum. Lang. Technol., Volume 1 (Long and Short Papers), Assoc. for Comput. Linguistics, Minneapolis, Minnesota, pp 4171–4186, (2019) https://doi.org/10.18653/v1/N19-1423
18. Ducasse, C.J.: Is scientific verification possible in philosophy? Philosophy Sci. **2**(2), 121–127 (1935)
19. Dusmanu, M., Cabrio, E., Villata, S.: Argument mining on twitter: Arguments, facts and sources. In: Proc. of the 2017 Conf. on Empirical Methods in Nat. Lang. Process., Assoc. for Comput. Linguistics, Copenhagen, Denmark, pp 2317–2322, (2017) https://doi.org/10.18653/v1/D17-1245
20. Eger, S., Daxenberger, J., Gurevych, I.: Neural end-to-end learning for comput. argumentation mining. In: Proc. of the 55th Annu. Meet. of the Assoc. for Comput. Linguistics (Volume 1: Long Papers), Assoc. for Comput. Linguistics, Vancouver, Canada, pp 11–22, (2017) https://doi.org/10.18653/v1/P17-1002
21. Ferreira, W., Vlachos, A.: (2016) Emergent: a novel data-set for stance classification. In: Proc, pp. 1163–1168. NAACL, HLT (2016)
22. Flanagin, A., Metzger, M.J.: From encyclopaedia britannica to wikipedia: generational differences in the perceived credibility of online encyclopedia information. Inf. Commun. Soc. **14**(3), 355–374 (2011)
23. Fleck, L.: The Genesis and Development of a Scientific Fact. The Univ. of Chicago Press (1979)
24. Fogg, B.J., Cuellar, G., Danielson, D.: Motivating, influencing, and persuading users: an introduction to captology. Hum Comput Interaction Fundamentals pp 109–122 (2009)
25. Hanto, V., Tostrup, M.: Towards automated fake news classifica—on building collections for claim analysis research. Master's thesis, Norwegian Univ. of Sci. and Technol (2018)

26. Hassan, N., Arslan, F., Li, C., Tremayne, M.: Toward automated fact-checking: detecting check-worthy factual claims by claimbuster. In: Proc. of the 23rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp 1803–1812 (2017a)
27. Hassan, N., Zhang, G., Arslan, F., Caraballo, J., Jimenez, D., Gawsane, S., Hasan, S., Joseph, M., Kulkarni, A., Nayak, A.K., et al.: ClaimBuster: the first-ever end-to-end fact-checking system. Proc VLDB Endowment **10**(12), 1945–1948 (2017)
28. Hochschild, J.L., Einstein, K.L.: Do Facts Matter? Information and Misinformation in Am. Politics. Univ. of Oklahoma Press, Norman, OK (2015)
29. Ioaniddis, J.P.A., Suart, M.E., Brownlee, S., Strite, S.A.: How to survive the medical misinformation mess. Eur. J. Clin. Inv. **47**(11), 795–802 (2017)
30. Kazai, G., Kamps, J., Milic-Frayling, N.: The face of quality in crowdsourcing relevance labels: demographics, personality and labeling accuracy. In: Proceedings of the 21st ACM international conference on information and knowledge management, association for computing machinery, New York, NY, USA, CIKM '12, p 2583-2586, (2012) https://doi.org/10.1145/2396761.2398697
31. Kelly, D., Lin, J.: Overview of the TREC 2006 CiQA Task. SIGIR Forum **41**(1), 107–116 (2007). https://doi.org/10.1145/1273221.1273231
32. Kolstø, S.D.: 'to trust or not to trust pupils,...' ways of judging information encountered in a socio-scientific issue. Int. J. Sci. Ed. **23**(9), 877–901 (2001)
33. Konstantinovskiy, L., Price, O., Babakar, M., Zubiaga, A.: Towards automated factchecking: developing an annotation schema and benchmark for consistent automated claim detection. arXiv preprint (2018) arXiv:180908193
34. Koppen, L., Phillips, J., Papageorgiou, R.: Analysis of reference sources used in drug-related Wikipedia articles. J. Med. Lib. Assoc. **103**(3), 140 (2015)
35. Kupferberg, N., Protus, B.M.: Accuracy and completeness of drug information in Wikipedia: an assessment. J. Med. Lib. Assoc. **99**(4), 310 (2011)
36. Lebow, M.A.: The pill and the press: reporting risk. Obstet. Gynecol. **93**(3), 453–456 (1999)
37. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics (2019). https://doi.org/10.1093/bioinformatics/btz682
38. Lee, K., He, L., Zettlemoyer, L.: Higher-order coreference resolution with coarse-to-fine inference. In: Proc. of the 2018 Conf. of the North Am. Chapter of the Assoc. for Comput. Linguistics: Hum. Lang. Technol., Volume 2 (Short Papers), Assoc. for Comput. Linguistics, New Orleans, Louisiana, pp 687–692, (2018) https://doi.org/10.18653/v1/N18-2108
39. Levy, R., Bilu, Y., Hershcovich, D., Aharoni, E., Slonim, N.: Context dependent claim detection. In: Proc. of COLING 2014, the 25th Int. Conf. on Comput. Linguistics: Technical Papers, Dublin City Univ. and Assoc. for Comput. Linguistics, Dublin, Ireland, pp 1489–1500 (2014)
40. Levy, R., Gretz, S., Sznajder, B., Hummel, S., Aharonov, R., Slonim, N.: Unsupervised corpus—wide claim detection. In: Proc. of the 4th Workshop on Argument Mining, Assoc. for Comput. Linguistics, Copenhagen, Denmark, pp 79–84, (2017) https://doi.org/10.18653/v1/W17-5110
41. Lin, B.Y., Lu, W.: Neural adaptation layers for cross-domain named entity recognition. In: Proceedings of the 2018 conference on empirical methods in natural language processing, association for computational linguistics, Brussels, Belgium, pp 2012–2022, (2018) https://doi.org/10.18653/v1/D18-1226
42. Lin, J., Zhang, P.: Deconstructing nuggets: the stability and reliability of complex question answering evaluation. In: Proceedings of the 30th annual international acm sigir conference on research and development in information retrieval, Association for Computing Machinery, New York, NY, USA, SIGIR '07, p 327-334, (2007) https://doi.org/10.1145/1277741.1277799
43. Liu, J., Chen, Y., Liu, K., Bi, W., Liu, X.: Event extraction as machine reading comprehension. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), Association for Computational Linguistics, Online, pp 1641–1651, (2020) https://doi.org/10.18653/v1/2020.emnlp-main.128
44. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov. V.: RoBERTa: a robustly optimized bert pretraining approach. arXiv preprint (2019) arXiv:190711692
45. London, D.A., Andelman, S.M., Christiano, A.V., Kim, J.H., Hausman, M.R., Kim, J.M.: Is Wikipedia a complete and accurate source for musculoskeletal anatomy? Surg. Radiol. Anat. **41**(10), 1187–1192 (2019)
46. Mainous, A.G.: Perspectives in primary care: Disseminating scientific findings in an era of fake news and science denial. Ann. Fam Med. **16**(6), 490–491 (2018)
47. Medlock, S., Eslami, S., Askari, M., Arts, D.L., Sent, D., de Rooij, S.E., Abu-Hanna, A.: Health information-seeking behavior of seniors who use the internet: a survey. J. Med. Internet Res. **17**(1), e10 (2015)
48. Mohtarami, M., Baly, R., Glass, J., Nakov, P., Màrquez, L., Moschitti, A.: Automatic stance detection using end-to-end memory networks. In: Proc. of the 2018 Conf. of the North Am. Chapter of the Assoc. for Comput. Linguistics: Hum. Lang. Technol., Volume 1 (Long Papers), Assoc. for Comput. Linguistics, New Orleans, Louisiana, pp 767–776, (2018) https://doi.org/10.18653/v1/N18-1070
49. Molitor, F.: Accuracy in science news reporting by newspapers: the case of aspirin for the prevention of heart attacks. Health Commun. **5**(3), 209–224 (1993)
50. Moore, B., Singletary, M.: Scientific sources' perceptions of network news accuracy. J. Q. **62**(4), 816–823 (1985)
51. Moynihan, R., Bero, L., Ross-Degnan, D., et al.: Coverage by the news media of the benefits and risks of medications. New. Eng. J. Med. **342**(22), 1645–1650 (2000)
52. Nagler, R.H., Hornik, R.C.: Measuring media exposure to contradictory health information: a comparative analysis of four potential measures. Commun. Methods Measures **6**(1), 56–75 (2012)
53. Nakashole, N., Mitchell, T.M.: Language-aware truth assessment of fact candidates. In: ACL, ACL, pp 1009–1019 (2014)
54. Nakov, P., Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Màrquez, L., Zaghouani, W., Gencheva, P., Kyuchukov, S., Da San Martino, G.: Overview of the CLEF-2018 Lab on automatic identification and verification of claims in political debates. In: Working Notes of CLEF 2018—Conf. and Labs of the Evaluation Forum, CEUR-WS.org, Avignon, France, CLEF '18 (2018)
55. Nørregaard, J., Horne, B.D., Adalı, S.: NELA-GT-2018: a large multi-labelled news dataset for the study of misinformation in news articles. In: Proc. of the Int. AAAI Conf. on Web and Social Media, Assoc. for the Adv. of Artif. Intell., vol 13, pp 630–638 (2019)
56. Olaisen, J.: Information quality factors and the cognitive authority of electronic information. In: Wormwell, I. (ed.) Information Quality: Definitions and Dimensions, pp. 91–121. Taylor Graham, England (1990)
57. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: EMNLP, ACL, pp 1532–1543 (2014)
58. Petty, R.E., Cacioppo, J.T.: Involvement and persuasion: tradition versus integration. Psychol. Bull. **107**(3), 367–374 (1990)
59. Pirolli, P.: Exploring and finding information. In: Caroll, J. (ed.) HCI Models, Theories and Frameworks: Toward a Multidisciplinary Science, pp. 157–191. Morgan Kauffmann, San Francisco (2003)

60. Pribble, J.M., Goldstein, K.M., Fowler, E.F., Greenberg, M.J., Noel, S.K., Howell, J.D.: Medical news for the public to use? What's on local TV news. Am. J. Manag. Care **12**, 170–176 (2006)
61. Pyysalo, S., Ginter, F., Moen, H., Salakoski, T., Ananiadou, S.: Distributional semantics resources for biomedical text processing. In: Proc. 5th Int. Symposium on Lang.s in Biology and Med., pp 39–44 (2013)
62. Rains, S.A., Karmikel, C.D.: Health information-seeking and perceptions of website credibility: examining web-use orientation, message characteristics, and structural features of websites. Comput. Hum. Behav. **25**(2), 544–553 (2009)
63. Ramshaw, L., Marcus, M.: Text chunking using transformation-based learning. In: Third workshop on very large Corpora, (1995) https://www.aclweb.org/anthology/W95-0107
64. Ransohoff, D.F., Ransohoff, R.M.: Sensationalism in the media: when scientists and journalists may be complicit collaborators. Eff. Clin. Pract. **4**(4), 185 (2001)
65. Rashkin, H., Choi, E., Jang, J.Y., Volkova, S., Choi, Y.: Truth of varying shades: Analyzing language in fake news and political fact-checking. In: EMNLP, pp 2931–2937 (2017)
66. Rinott, R., Dankin, L., Alzate Perez, C., Khapra, M.M., Aharoni, E., Slonim, N.: Show me your evidence - an automatic method for context dependent evidence detection. In: Proc. of the 2015 Conf. on Empirical Methods in Nat. Lang. Process., Assoc. for Comput. Linguistics, Lisbon, Portugal, pp 440–450, (2015) https://doi.org/10.18653/v1/D15-1050, https://www.aclweb.org/anthology/D15-1050
67. Roitero, K., Soprano, M., Portelli, B., Spina, D., Della Mea, V., Serra, G., Mizzaro, S., Demartini, G.: The COVID-19 Infodemic: can the crowd judge recent misinformation objectively? In: Proceedings of the 29th ACM international conference on information & knowledge management, Association for Computing Machinery, New York, NY, USA, CIKM '20, p 1305-1314, (2020) https://doi.org/10.1145/3340531.3412048
68. Rosenthal, S., McKeown, K.: I couldn't agree more: the role of conversational structure in agreement and disagreement detection in online discussions. In: Proc. of the 16th Annu. Meet. of the Special Interest Group on Discourse and Dialogue, Assoc. for Comput. Linguistics, Prague, Czech Republic, pp 168–177, (2015) https://doi.org/10.18653/v1/W15-4625
69. Sathiaraj, D., Triantaphyllou, E.: on identifying critical nuggets of information during classification tasks. IEEE Trans. Knowl. Data Eng. **25**(6), 1354–1367 (2013). https://doi.org/10.1109/TKDE.2012.112
70. Sbaffi, L., Rowley, J.: Trust and credibility in web-based health information: a review and agenda for future research. JMIR **19**(6), e218 (2017)
71. Scheufele, D.A., Krause, N.M.: Science audiences, misinformation, and fake news. PNAS **116**(16), 7662–7669 (2019)
72. Schwitzer, G.: Pollution of health news. BMJ **356**, j1262 (2017)
73. Shuchman, M., Wilkes, M.S.: Medical scientists and health news reporting: a case of miscommunication. Ann. Internal Med. **126**(12), 976–982 (1997)
74. Stab, C., Miller, T., Schiller, B., Rai, P., Gurevych, I.: Cross-topic argument mining from heterogeneous sources. In: Proc. of the 2018 Conf. on Empirical Methods in Nat. Lang. Process., Assoc. for Comput. Linguistics, Brussels, Belgium, pp 3664–3674, (2018) https://doi.org/10.18653/v1/D18-1402, https://www.aclweb.org/anthology/D18-1402
75. Stanford, K.: Underdetermination of Scientific Theory. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy (Winter 2017 Edition), Metaphysics Res. Lab, Stanford Univ., (2017) https://plato.stanford.edu/archives/win2017/entries/scientific-underdetermination/
76. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: brat: a Web-based Tool for NLP-assisted text annotation. In: EACL, ACL, Avignon, France (2012)
77. Straková, J., Straka, M., Hajic, J.: Neural architectures for nested NER through linearization. In: Proc. of the 57th Annu. Meet. of the Assoc. for Comput. Linguistics, Assoc. for Comput. Linguistics, Florence, Italy, pp 5326–5331, (2019) https://doi.org/10.18653/v1/P19-1527, https://www.aclweb.org/anthology/P19-1527
78. Swales, J.M.: Languages for specific purposes. Ann. Rev. Appl. Ling. **20**, 59–76 (2000)
79. Tankard, J.W., Jr., Ryan, M.: News source perceptions of accuracy of science coverage. J. Q. **51**(2), 219–225 (1974)
80. Thorne, J., Vlachos, A.: Automated fact checking: task formulations, methods and future directions. In: Proc. of the 27th Int. Conf. on Comput. Linguistics, Assoc. for Comput. Linguistics, Santa Fe, New Mexico, USA, pp 3346–3359 (2018)
81. Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: FEVER: a large-scale dataset for fact extraction and VERification. In: Proc. of the 2018 Conf. of the North Am. Chapter of the Assoc. for Comput. Linguistics: Hum. Lang. Technol., Volume 1 (Long Papers), Assoc. for Comput. Linguistics, New Orleans, Louisiana, pp 809–819, (2018) https://doi.org/10.18653/v1/N18-1074
82. Tjong, Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In: Proc. of the Seventh Conf. on Nat. Lang. Learning at HLT-NAACL 2003, Assoc. for Comput. Linguistics, Edmonton, Canada, pp 142–147, (2003) https://www.aclweb.org/anthology/W03-0419
83. Uscinski, J.E., Butler, R.W.: The epistemology of fact checking. Crit. Rev. **25**(2), 162–180 (2013)
84. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., Zhao, S.: Applications of machine learning in drug discovery and development. Nat. Rev. Drug Discov. **18**, 463–477 (2019). https://doi.org/10.1038/s41573-019-0024-5
85. Velmans, M.: Is human information processing conscious? Behav. Brain Sci. **14**(4), 651–726 (1991)
86. Vlachos, A., Riedel, S.: Fact Checking: Task definition and dataset construction. In: Proc. of the ACL 2014 Workshop on Lang. Technol. and Comput. Social Sci., Assoc. for Comput. Linguistics, Baltimore, MD, USA, pp 18–22, (2014) https://doi.org/10.3115/v1/W14-2508
87. Voss, E., Boyce, R., Ryan, P., van der Lei, J., Rijnbeek, P., Schuemie, M.: Accuracy of an automated knowledge base for identifying drug adverse reactions. J. Biomed. Inform. **66**, 72–81 (2017). https://doi.org/10.1016/j.jbi.2016.12.005
88. Wang, W.Y.: Liar, liar pants on fire: A new benchmark dataset for fake news detection. In: Proc. of the 55th Annu. Meet. of the Assoc. for Comput. Linguistics (Volume 2: Short Papers), Assoc. for Comput. Linguistics, Vancouver, Canada, pp 422–426, (2017) https://doi.org/10.18653/v1/P17-2067
89. Wang, Y., Afzal, N., Fu, S., Wang, L., Shen, F., Rastegar-Mojarad, M., Liu, H.: MedSTS: a resource for clinical semantic textual similarity. Lang. Res. Eval. **54**, 57–72 (2018)
90. Weeks, W.B., Wallace, A.E.: Readability of British and Am medical prose at the start of the 21st century. BMJ **325**(7378), 1451–1452 (2002)
91. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Brew, J.: HuggingFace's Transformers: State-of-the-art Natural Language Processing. (2019) arXiv:abs/1910.03771
92. Wu, D.T., Hanauer, D.A., Mei, Q., Clark, P.M., An, L.C., Proulx, J., Zeng, Q.T., Vydiswaran, V.G., Collins-Thompson, K., Zheng, K.: Assessing the readability of Clinical Trials.gov. J. Am. Med. Inform. Assoc. **23**(2), 269–275 (2016)

93. Xie, B., Bugg, J.M.: Public library computer training for older adults to access high-quality Internet health information. Libr. Inf. Sci. Res. **31**(3), 155 (2009). https://doi.org/10.1016/j.lisr.2009.03.004
94. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: XLNet: generalized autoregressive pretraining for language understanding. In: NeurIPS, pp 5754–5764 (2019)
95. Yavchitz, A., Boutron, I., Bafeta, A., Marroun, I., Charles, P., Mantz, J., Ravaud, P.: Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. PLoS Med. **9**(9), e1001308 (2012)
96. Ye, J., Skiena, S.: MediaRank: comput. ranking of online news sources. CoRR abs/1903.07581, (2019) arXiv:1903.07581
97. Yeh, A.: More accurate tests for the statistical significance of result differences. In: COLING 2000 Volume 2: The 18th International conference on computational linguistics (2000)
98. Zhou, X., Mulay, A., Ferrara, E., Zafarani, R.: Recovery: a multimodal repository for covid-19 news credibility research. In: CIKM '20: Proc. of the 29th ACM Int. Conf. on Inf. & Knowl. Manag., Assoc. for Comput. Machinery, New York, NY, USA, p 3205-3212, (2020) https://doi.org/10.1145/3340531.3412880
99. Zuo, C., Karakas, A., Banerjee, R.: to check or not to check: syntax, semantics, and context in the language of check-worthy claims. In: Crestani F, Braschler M, Savoy J, Rauber A, Müller H, Losada DE, Bürki GH, Cappellato L, Ferro N (eds) Experimental IR Meets Multilinguality, Multimodality, and Interaction – Proc. of the 10th Int. Conf. of the CLEF Assoc., Springer Int. Publishing, Lugano, Switzerland, Lecture Notes in Comput. Sci., vol 11696, pp 271 – 283 (2019)