SciClops: Detecting and Contextualizing Scientific Claims for Assisting Manual Fact-Checking

Panayiotis Smeros
EPFL
Lausanne, Switzerland
panayiotis.smeros@epfl.ch

Carlos Castillo
Universitat Pompeu Fabra
Barcelona, Catalunya, Spain
chato@acm.org

Karl Aberer
EPFL
Lausanne, Switzerland
karl.aberer@epfl.ch

ABSTRACT

This paper describes SciClops, a method to help combat online scientific misinformation. Although automated fact-checking methods have gained significant attention recently, they require pre-existing ground-truth evidence, which, in the scientific context, is sparse and scattered across a constantly-evolving scientific literature. Existing methods do not exploit this literature, which can effectively contextualize and combat science-related fallacies. Furthermore, these methods rarely require human intervention, which is essential for the convoluted and critical domain of scientific misinformation.

SciClops involves three main steps to process scientific claims found in online news articles and social media postings: extraction, clustering, and contextualization. First, the extraction of scientific claims takes place using a domain-specific, fine-tuned transformer model. Second, similar claims extracted from heterogeneous sources are clustered together with related scientific literature using a method that exploits their content and the connections among them. Third, check-worthy claims, broadcasted by popular yet unreliable sources, are highlighted together with an enhanced fact-checking context that includes related verified claims, news articles, and scientific papers. Extensive experiments show that SciClops tackles sufficiently these three steps, and effectively assists non-expert fact-checkers in the verification of complex scientific claims, outperforming commercial fact-checking systems.

CCS CONCEPTS

• **Information systems** → **Spam detection**; *Trust*; **Content analysis and feature selection**; *Clustering and classification*.

KEYWORDS

Scientific Claims; Misinformation; Fact-Checking

ACM Reference Format:
Panayiotis Smeros, Carlos Castillo, and Karl Aberer. 2021. SciClops: Detecting and Contextualizing Scientific Claims for Assisting Manual Fact-Checking. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3459637.3482475

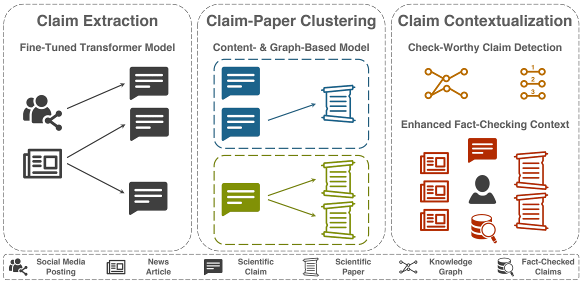


Figure 1: Overview of SciClops including the three methods for extraction (§3), clustering (§4), and contextualization (§5) of scientific claims.

1 INTRODUCTION

Although the amount of news at our disposal seems to be ever-expanding, traditional media companies and professional journalists remain the key to the production and communication of news. The way in which news is disseminated has become more intricate than in the past, with social media playing a fundamental role [12].

The ephemeral, fast-paced nature of social media, the brevity of the messages circulating on them, the short attention span of their users, their preference for multimedia rather than textual content, and in general the fierce competition for attention, has forced journalists to adapt in order to survive in the attention economy [36]. As a consequence, news outlets are increasingly using catchy headlines, as well as outlandish and out-of-context claims that perform well in terms of attracting eyeballs and clicks [46].

When mainstream news media communicate scientific content to the public, the situation is by no means different [48]. Oversimplified scientific claims are rapidly shared in social media, while the scientific evidence that may support or refute them remains absent or locked behind pay-walled journals. For instance, on March 11th, 2020, an article in *The Lancet Respiratory Medicine* theorized that nonsteroidal anti-inflammatory drugs such as Ibuprofen could worsen COVID-19 symptoms [13]. Without referencing explicitly to this article, but motivated by it, the Minister of Health of France posted on Twitter, advising people to avoid Ibuprofen when possible.[1] His message was re-posted nearly 43K times and liked nearly 40K times. In contrast, a World Health Organization's message posted four days later, which insisted Ibuprofen was safe, was re-posted only 7.5K times and liked only 8.5K times.[2]

Fact-checking portals such as ScienceFeedback.co, among others, work closely with domain experts and scientists to debunk misinformation and bring nuance to potentially misleading claims. This remains, however, a labor-intensive and time-consuming task [19].

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

ACM ISBN 978-1-4503-8446-9/21/11…$15.00
https://doi.org/10.1145/3459637.3482475

[1] https://twitter.com/olivierveran/status/1238776545398923264
[2] https://twitter.com/WHO/status/1240409217997189128

Table 1: Approaches for Extraction, Clustering, and Contextualization as proposed by selected references

	Fact-Checking Portals	Hassan et al. [20]	Popat et al. [43]	Jaradat et al. [21]	Shaar et al. [50]	Hansen et al. [18]	Zlatkova et al. [62]	Karagiannis et al. [25]	Pinto et al. [42]	Pavllo et al. [40]	Smeros et al. [53]	Levy et al. [29]	Stab et al. [54]	Patwari et al. [39]	Lippi and Torroni [32]	Jiang et al. [22]	Reimers et al. [44]	Yao et al. [60]	Zhou et al. [61]	Hamilton et al. [17]	Wang et al. [57]	Duong et al. [11]	Kochkina et al. [26]	Shao et al. [51]	Ciampaglia et al. [9]	Nadeem et al. [37]	Gad-Elrab et al. [16]	Chen et al. [7]	**SciClops**
Extraction																													
Weak Supervision	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	-	-	-	-	-	-	-	-	-	-	-	✓
Traditional ML Model	✗	✓	✓	✗	✗	✗	✓	✓	✓	✗	✗	✓	✓	✓	✓	✗	✗	-	-	-	-	-	-	-	-	-	-	-	✓
Neural ML Model	✗	✗	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	-	-	-	-	-	-	-	-	-	-	-	✓
Clustering																													
Text Modality	✓	✓	✗	✗	✗	✗	✗	✗	✓	-	-	-	-	-	-	-	✓	✓	✓	✓	✗	✗	-	-	-	-	-	-	✓
Graph Modality	✗	✗	✗	✗	✗	✗	✓	✗	✓	-	-	-	-	-	-	-	✗	✓	✓	✓	✓	✓	-	-	-	-	-	-	✓
Bipartite Clusters	✗	✗	✗	✗	✗	✗	✗	✗	✗	-	-	-	-	-	-	-	✗	✗	✗	✗	✗	✓	-	-	-	-	-	-	✓
Contextualization																													
Ground-Truth KBs	✓	✓	✓	✗	✓	✗	✓	✓	✗	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	✓	✓	✓	✓	✓	✓
Priority Ranking	✗	✗	✗	✓	✓	✓	✗	✓	✗	-	-	-	-	-	-	-	-	-	-	-	-	-	✗	✗	✗	✓	✗	✗	✓
Scientific Context	✓	✗	✗	✗	✗	✗	✗	✗	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	✗	✗	✗	✗	✗	✗	✓

On the other hand, despite misinformation circulating online exceeding the capacity of manual fact-checking, traditional news outlets are skeptical towards adopting fully-automated methods [52]. Their main concern is that such tools provide poorly-interpretable evidence (according to the journalistic standards), and any false judgment can lead to a downfall of the outlet's reputation. Indeed, even big tech companies were forced to suspend automated fact-checking features due to similar criticism from news outlets [15]. Hence, the consensus regarding the usage of automation in journalism is that it should assist but not replace journalists and news consumers when they validate the veracity of news, enabling the movement onward the era of citizen journalism [38].

Our work focuses on scientific claims in news articles and social media postings. As scientific claims, we consider *sentence-level segments that involve one or more scientific entities and are eligible for fact-checking*. For example, the sentence *"Ibuprofen can worsen COVID-19 symptoms"* is a scientific claim because it involves two scientific entities (*Ibuprofen* and *COVID-19*) and implies a causal relation between them. To increase the coverage of our definition, we bound neither the number of entities nor the type of relation between them. Such non-deterministic definition makes the detection of scientific claims a challenging task, even for human annotators (details in §6.1). To address this task and enable the discovery of complex-structured claims, there is a need for advanced language models which are fine-tuned with domain-specific knowledge.

Once we identify candidate scientific claims, we seek evidence that proves or contradicts them via *contextualization*, i.e., via building an enhanced context of trustworthy information. In the scientific domain, the appropriate context consists of related scientific papers. Grouping similar claims and linking them to related scientific literature is a complex task, to a large extent because of the different nature of the items that we are seeking to connect (i.e., social media postings, news articles, and scientific papers). These contain key passages that determine such connections, but are fundamentally different in terms of: i) verbosity, ranging from character-limited postings to extended scientific papers, and ii) complexity, ranging from a "social media friendly" style of writing to the more formal registry of journalism and academic writing.

Finally, since there is a plethora of controversial claims (especially in the times of a pandemic), there is a need for a check-worthiness ranking that considers the prevalence and the reliability of the broadcasting medium. Providing a scientific context enables non-expert fact-checkers to verify claims with more precision than commercial fact-checking systems, and more confidence since the provided context is fully-interpretable (details in §6.3).

Our Contribution. In this paper we describe *SciClops* (Figure 1), a method to assist manual verification of dubious claims, in scientific fields with open-access literature and limited fact-checking coverage. The technical contributions we introduce are the following:

- pretrained and fine-tuned transformer-based models for scientific claim extraction from news and social media (§3);
- multimodal, joint clustering models for claims and papers that utilize both content and graph information (§4);
- methods for ranking check-worthy claims using a custom knowledge graph, and methods for creating enhanced scientific contexts to assist manual fact-checking (§5); and
- extensive experiments involving expert and non-expert users, strong baselines and commercial fact-checking systems (§6).

2 RELATED WORK

Fact Checking Portals in general (Snopes.com), political (PolitiFact.com), and scientific (ScienceFeedback.co) domains employ specialized journalists who manually: i) detect suspicious claims (extraction), ii) discover variants of these claims published in social and news media (clustering), and iii) find the appropriate prism under which they assess their credibility (contextualization). We summarize some automated methods tackling these steps in Table 1.

Claim Extraction. On *weakly supervised models*, Pavllo et al. [40] and Smeros et al. [53] generate complex rule-based heuristics to extract quotes from, respectively, general and scientific news articles.

On *traditional ML models*, Levy et al. [29] and Stab et al. [54] propose learning models for claim detection and argument mining and introduce publicly available datasets, which we utilize to train our extraction models (details in §6.1). Hassan et al. [20] and Popat et al. [43] propose claim classification models that use the aforementioned fact-checking portals to verify political claims, while Patwari et al. [39] and Lippi and Torroni [32] propose, respectively, an ensemble and a context-independent model for claim extraction. Finally, Zlatkova et al. [62] propose a claim extraction model for images, Karagiannis et al. [25] propose a framework for statistical claims verification, and Pinto et al. [42] propose a method for identifying pairwise relationships between scientific entities.

On *neural ML models*, Jaradat et al. [21] and Shaar et al. [50] detect and rank previously fact-checked claims using deep neural models, while Hansen et al. [18] also train a neural ranking model for check-worthy claims using weak supervision. Furthermore, Jiang et al. [22] use contextualized embeddings to factor fact-checked claims, while Reimers et al. [44] use also contextualized embeddings for claim extraction and clustering. Finally, CheckThat! Lab [2] features claim extraction and check-worthiness tasks which are oriented towards political debates in social media platforms.

While the other approaches cover the cases of political, statistical, and visual claims, our approach provides the first dedicated solution for scientific claims. Given the complex nature of the scientific claims in terms of structure and vocabulary, our approach is based on advanced language models with contextualized embeddings that are fine-tuned with domain-specific knowledge. Furthermore, our approach works with arbitrary input text, e.g., from social media postings, blog posts, or news articles.

Claim-Paper Clustering. Since our data contains multimodal information (the textual representation of claims and papers and the interconnections between them), we present multimodal clustering approaches that combine text and graph data modalities.

Yao et al. [60] propose a unified convolutional network of terms and documents, while Zhou et al. [61] use weighted graphs that encode the attribute similarity of the clustered nodes. Hamilton et al. [17] introduce a methodology for jointly training embeddings based on text and graph information, while Reimers et al. [44] apply a numerical clustering on top of such embeddings. Finally, Wang et al. [57] propose a technique for training network embeddings that preserves the communities (clusters) of a graph, while Duong et al. [11] provide interpretable such embeddings.

In our approach, we jointly cluster scientific claims and referenced papers, using both content and graph information. To the best of our knowledge, this is the first approach that deals with heterogeneous passages in terms of length and vocabulary type, which are also interconnected through a bipartite graph.

Claim Contextualization. In addition to the extraction methods described above, the majority of which also provide contextualization/verification techniques (details in Table 1), Kochkina et al. [26] and Shao et al. [51] propose methods for automatic rumor verification using well-known fact-checking portals. Ciampaglia et al. [9], Nadeem et al. [37], and Chen et al. [7] use Wikipedia for fact-validation, while Gad-Elrab et al. [16] use custom knowledge graphs for generating interpretable explanations for candidate facts.

While other approaches describe this step as *"verification"*, since essentially they lookup a claim in a ground-truth knowledge base, we consider the general case in which claims rarely appear in such knowledge bases. As we observe in §6.3, this is a pragmatic assumption since the majority of the fact-checking effort targets non-scientific topics. As the verification of scientific claims is typically more demanding than other types of claims (e.g., ScienceFeedback.co has built an entire peer-reviewing system for this purpose), we propose a methodology that contextualizes claims based on related scientific literature and ranks them based on the prevalence and the reliability of the broadcasting medium.

3 CLAIM EXTRACTION

We address claim extraction as a classification problem at the sentence level, i.e., we want to distinguish between claim-containing and non-containing sentences. Below, we present the baseline and the advanced extractors that we evaluate in §6.1.

3.1 Baseline Extractors

We implement several baseline extractors that cover most of the related work on claim extraction described in §2: i) two complex heuristics which are used by state-of-the-art *weakly supervised models* [40, 53]; ii) an off-the-shelf classifier trained with standard textual features which is used by state-of-the-art *traditional ML models* [20, 32]; and iii) a transformer model which is used by state-of-the-art *neural ML models* [44, 50].

3.1.1 Grammar-Based Heuristic. The usage of reporting verbs such as "say," "claim," or "report," is a typical element of pattern-matching heuristics for finding claims. Another element is the usage of domain-specific vocabulary; in the scientific context, common verbs in claims include "prove" and "analyze." Thus, we compile a seed set of such verbs, which we extend with synonyms from WordNet [35]. In the following, we refer to this set of reporting verbs as RV.

Scientific claims fundamentally refer to scientific studies, scientists or, more generally, scientific notions. Thus, we employ a shortlist of nouns related to studies and scientists (including "survey" or "researcher"). In the following, we refer to this set of nouns, together with the set of Person and Organization entities, as E.

Finally, to capture the syntactic structure of claims, we obtain part-of-speech tags from the candidate claim-containing sentences. Using this information, we construct a series of complex expressions over *classes of words* such as the following:

$$(root(s) \in RV) \land ((nsubj(s) \in E) \lor (dobj(s) \in E)) \implies (s \in Claims)$$

where s is a sentence, $root(.)$ returns the root verb of the syntactic tree of a sentence, $nsubj(.)$ returns the nominal subject, and $dobj(.)$ the direct object of a sentence.

3.1.2 Context-Based Heuristic. This heuristic is based on a frequent non-syntactic pattern, which is quite evident in our data: if an article is posted on social media, then its central claim is typically re-stated or minimally paraphrased in the postings. We investigate pairs (s, p) of candidate sentences s, extracted from news articles, and postings p, referencing these news articles. Our heuristic has the form:

$$(\exists p : sim(s, p) \cdot pop(p) \geq threshold) \implies (s \in Claims)$$

where $sim(s, p)$ denotes the *cosine similarity* between the embeddings representations of s and p, and $pop(p)$ denotes the normalized popularity of p, i.e., the raw popularity of p over the sum of the popularity of all the p's that refer to s. As popularity, we consider the sum of the *re-postings* and *likes*. Finally, *threshold* is a hyperparameter of our heuristic, which in our implementation is fixed to 0.9, yielding a good compromise of precision and recall. We note that this is the only proposed extractor that is not purely content-based since it also requires contextual information.

3.1.3 Random Forest Classifier. To train this classifier, we apply a standard text-preprocessing pipeline, including stop-words removal and part-of-speech tagging. Then, we transform the candidate claim-containing sentences into embeddings by averaging the word embeddings provided by GloVe [41]. As we see in our evaluation (§6.1), this classifier performs better than the aforementioned baselines; we also note that, compared to the complex transformer models, it is substantially less intensive in terms of computational resources and training time needed.

3.1.4 BERT Model. One of the most successful state-of-the-art approaches to several NLP tasks, including classification, is the *transformer model* [10]. In our implementation we use the well-known model *BERT* and particularly its version named *bert-base-uncased* [59]. The configuration parameters of the model are those suggested in a widely used software release of this model.[3]

As the last layer of the transformer architecture of *BERT* (and the variants we introduce next), we add a standard binary classification layer with two output neurons, which we train using the datasets described in §6.1. During the training, we keep the rest of the layers of the model frozen at their initial parameters.

3.2 Fine-Tuned Transformer Extractors

Since *BERT* is originally trained on the generic corpus of Wikipedia, the word representations it generates are also generic. However, scientific claim extraction is a downstream task, where the model has to recognize patterns of a more narrow domain. Thus, we introduce three variants of *BERT* with domain-specific fine-tuning namely, *SciBERT*, *NewsBERT* and *SciNewsBERT*:

- *SciBERT* is pretrained on top of *BERT* with a corpus from SemanticScholar.org containing ~1*M* papers [3]. *SciBERT* has its own vocabulary that is built to best match the scientific domain.
- *NewsBERT* is a new model that we introduce, built on top of *BERT* and pretrained on a freely-available corpus of ~1*M* headlines published by the Australian Broadcasting Corporation [27].
- *SciNewsBERT* is also a new model that is pretrained like *NewsBERT*, albeit, it is built on top of *SciBERT* instead of *BERT*.

For training *NewsBERT* and *SciNewsBERT* we employ the standard tasks for training *BERT*-like models: i) *Masked Language Modeling*, where the model has to predict the randomly masked words in a sequence of text, and ii) *Next Word Prediction*, where the model has to predict the next word, given a set of preceding words. The hyper-parameters used for training the models are the default proposed by the software release referenced above. Since both *NewsBERT* and *SciNewsBERT* need substantial computational power and training time, we make them publicly available for research purposes (§7).

4 CLAIM-PAPER CLUSTERING

Contextualizing scientific claims requires to connect them with related scientific papers. To achieve this, our approach employs a clustering methodology. The clusters, composed of a mixture of claims and papers, must have high semantic coherence and ideally maintain the connections that exist between some of these claims and papers. These implicit connections are hyperlinks starting from news articles and social media postings containing these claims and ending on referenced papers, forming a sparse bipartite graph.

The clustering methods that we employ are: i) *Content-Based* methods on top of either the raw text or an embeddings representation of the passages, ii) *Graph-Based* methods on top of the bipartite graph between the claims and the papers, or iii) *Hybrid* methods that combine the *Content-Based* and the *Graph-Based* methods. Furthermore, we consider both soft (overlapping) clustering (i.e., passages can belong to more than one cluster), and hard (non-overlapping) clustering (i.e., passages must belong to exactly one cluster). The notation used in this section is summarized in Table 2.

Table 2: Clustering notation. The embeddings dimension (*dim*) of our models is *300*. Matrix *L* has a *1* in position (*c*, *p*), iff a news article or a social media posting containing claim *c* has a hyperlink to paper *p*. Each row of the clustering matrices (*C′* and *P′*) contains the probability of a claim or a paper to belong to a cluster; for hard clustering it is "one-hot", i.e., it has a single non-zero element, and for soft clustering it is a general probability distribution.

Symbol	Description
$C \in \mathbb{R}^{\lvert \text{claims} \rvert \times \text{dim}}$	initial claims matrix
$P \in \mathbb{R}^{\lvert \text{papers} \rvert \times \text{dim}}$	initial papers matrix
$L \in \{0, 1\}^{\lvert \text{claims} \rvert \times \lvert \text{papers} \rvert}$	interconnection matrix
$C' \in [0, 1]^{\lvert \text{claims} \rvert \times \lvert \text{clusters} \rvert}$	final claims clustering matrix
$P' \in [0, 1]^{\lvert \text{papers} \rvert \times \lvert \text{clusters} \rvert}$	final papers clustering matrix
$f_C : C \rightarrow C'$	non-linear neural transformation
$f_P : P \rightarrow P'$	non-linear neural transformation
$\lVert . \rVert_F$	*Frobenius Norm*

4.1 Content-Based Clustering

Our baseline is content-based (topic) clustering. According to this approach, we assume that claims and papers are represented in the same latent space, in which we compute topical joint clusters. This approach does not consider the interconnections (i.e., the bipartite graph) between the claims and the papers.

For topic modeling, we use *Latent Dirichlet Allocation* (*LDA*), an unsupervised statistical model that computes a soft topic clustering of a given set of passages [4]. We also use *Gibbs Sampling Dirichlet Mixture Model* (*GSDMM*), which assumes a hard topic clustering and is more appropriate for small passages such as claims [30]. When the passages are projected in an embeddings space, we use either the generic *Gaussian Mixture Model* (*GMM*), which computes a soft clustering by combining multivariate Gaussian distributions [45], or *K-Means* [33], which computes a hard clustering. Finally, we test these methods with and without reducing the embeddings dimensions using *Principal Component Analysis* (*PCA*) [14].

4.2 Graph-Based Clustering

Since our data is multimodal, an alternative to pure *Content-Based* clustering is pure *Graph-Based* clustering. We define this problem as an optimization problem, introducing an appropriate loss function that we want to minimize. Our goal is to compute the optimal clusters C' and P', and our evaluation criterion is the extent to which C' and P' fit with the interconnection matrix L. Hence, we propose the following loss function:

$$loss = \lVert C' - LP' \rVert_F$$

This loss function is also known as the *Reconstruction Error* and is commonly used in *Linear Algebra* for factorization and approximation problems. By applying this loss function, we force C' and P' to be aligned with L: the claims that appear in a news article should belong to the same cluster as the papers referenced by this article.

A degenerate solution to the problem, if we use only this loss function, is a uniform clustering for both claims and papers. The loss is minimized, but the clustering is useless, because the probability

[3] https://huggingface.co/bert-base-uncased

of any claim and any paper to belong to any cluster is uniform. To overcome this problem, we exploit the following technique that is widely used in image processing [28].

In row-stochastic matrices (i.e., matrices that each row sums to 1), a uniform soft clustering has lower *Frobenius Norm* than a non-uniform clustering. Consequently, any hard clustering has the maximum possible *Frobenius Norm*. Thus, we introduce a regularizer that imposes non-uniformity on the clusters by penalizing low *Frobenius Norms* for C' and P':

$$regularizer = \begin{cases} -\beta\,(\|C'\|_F + \|P'\|_F) & C', P' \in V \\ -\beta\,\|P'\|_F & C' \notin V \\ -\beta\,\|C'\|_F & P' \notin V \end{cases}$$

where V is the set of optimizable variables of our model, and β a hyper-parameter that in our experiments defaults to $\beta = 0.3$. We use a different regularizer in each alternative version of the model that we describe below. These alternative versions have varying flexibility, i.e., either both C' and P' are optimizable variables ($C', P' \in V$), or one of them is fixed, thus not optimizable ($C' \notin V$ or $P' \notin V$). If both of them are fixed ($C', P' \notin V$) then the model has no optimizable variables ($V = \emptyset$). Below we present the alternative versions of the model.

4.2.1 Graph-Based Adaptation. In this alternative (entitled *GBA-CP*), we start with arbitrary cluster assignments for C' and P', which we both optimize based on the loss function. This approach completely ignores the semantic information of C and P and adapts arbitrarily the clusters to the interconnection matrix L. This behavior of *GBA-CP* is confirmed in our experiments (§6.2).

In a less aggressive approach, we fix either C' or P' using one of the *Content-Based* algorithms explained above, and optimize only one clustering (the non-fixed) based on the loss function. We entitle these alternatives as *GBA-C* for optimizing C', and *GBA-P* for optimizing P' .

4.2.2 Graph-Based Transformation. In this alternative (entitled *GBT-CP*), instead of optimizing directly C' and P', we optimize the weights of the non-linear neural transformations f_C and f_P. The architecture of f_C and f_P consists of a hidden layer of neurons with a rectified linear unit (*ReLU*), and a linear *Softmax* classifier that computes the overall cluster-membership distribution. We use the same loss function as above where $C' = f_C(C)$ and $P' = f_P(P)$.

Similarly as above, in a less aggressive approach, we fix C' or P' using a *Content-Based* algorithm, and optimize only the weights of one transformation (f_C or f_P). We entitle these alternatives as *GBT-C* for optimizing f_C, and *GBT-P* for optimizing f_P.

4.3 Hybrid Clustering

The last clustering model that we propose is a *Hybrid* model that combines a *Content-Based* and a *Graph-Based* model. As we point out in our experimental evaluation (§6.2), there is a trade-off between these two approaches in terms of the semantic and interconnection coherence of the computed clusters. Hence, we introduce a tunable model that controls this trade-off.

Our model initializes the clusters C'_{init} and P'_{init} using a *Content-Based* model. Then, it uses an *Alternate Optimization* (*AO*) approach to jointly compute the final C' and P' that adjust best to L. More specifically, it iteratively freezes one of the two clusters and adjusts the other, until they both converge to an optimal state. The loss function of this model is the following:

$$loss = \begin{cases} \gamma\,\|C' - LP'\|_F + (1-\gamma)\,\left\|C' - C'_{init}\right\|_F & C'\text{-optim.} \\ \gamma\,\|C' - LP'\|_F + (1-\gamma)\,\left\|P' - P'_{init}\right\|_F & P'\text{-optim.} \end{cases}$$

where γ is a hyper-parameter that controls the trade-off between *Content-Based* and *Graph-Based* clustering. In our experiments for brevity we present results for three values: *AO-Content* for $\gamma = 0.1$, *AO-Balanced* for $\gamma = 0.5$, and *AO-Graph* for $\gamma = 0.9$.

5 CLAIM CONTEXTUALIZATION

In the previous section, we explain how we construct claim-paper clusterings in an unsupervised fashion. These clusterings give already an initial context for claims since they relate them with relevant scientific literature. In this section, we describe how we rank claims within clusters based on their check-worthiness and how we complement their fact-checking context by discovering (when available) previously verified related scientific claims.

5.1 Check-Worthy Claim Ranking

The check-worthiness of a scientific claim depends on its intent (e.g., whether it implies a causal relation or describes a particular aspect of an entity) and its prevalence (e.g., in news and social media). We construct a custom in-cluster knowledge graph in which we encode the intent of the claims into the topology of the graph and the prevalence of the claims into the weighting of the graph.

In-Cluster Knowledge Graph. We construct a knowledge graph by using terms from a domain-specific vocabulary as nodes. The edges of the graph denote the co-occurrence of two terms in the same claim (e.g., the claim *"Ibuprofen can worsen COVID-19 symptoms"* contributes the edge (*Ibuprofen – COVID-19*)).

Since the dataset we use in our evaluation is health-related (details in §6), we use the vocabulary of *CDC A-Z Index*[4] that includes health terms used by laypeople and professionals. We note that the rest of the methodology is independent of the domain of the dataset, and can be simply adapted by selecting an appropriate vocabulary.

Graph Topology. We distinguish between two types of topologies based on two different intents:

- *Causality-Based* topologies which contain nodes from distinct classes such as: i) *"Diseases and Disorders"* (e.g., *Depression*, *Influenza*, and *Cancer*), and ii) *"Conditions, Symptoms, Medications, and Nutrients"* (e.g., *Pregnancy*, *Fever*, and *Red Meat*). A directed edge between two nodes of a different class denotes, to a certain degree, a causal relation between these nodes [8].
- *Aspect-Based* topologies which focus on the "ego-network" for one particular node (e.g., *"COVID-19"*) and the different aspects regarding this node (e.g., *"Origin"* or *"Mortality Rate"*) [34].

Graph Weighting. The weighting scheme that we employ combines two criteria, namely the *popularity* and the *reputation* of the primary sources (i.e., the social media postings and the news articles) from which the claims were extracted.

The *popularity* of a posting is computed as the sum of the number of re-postings and likes. If multiple postings share the same claim, then their *popularity* is aggregated. Then, Box-Cox transformation

[4] https://www.cdc.gov/az

($\lambda = 0$) [5], to diminish the effect of the long-tail distribution, and Min-Max normalization in the interval $[0, 1]$ are applied.

On the other hand, the *reputation* of a news article is entailed from the reputation of the news outlet that publishes the article. In the context of this paper, we use the outlet scores compiled by the *American Council on Science and Health* (*ACSH*) [1], which we also normalize in the interval $[0, 1]$. News outlets that are not on *ACSH*'s list (i.e., "long-tail" outlets hosting only 13.5% of the total articles in our collection) are assigned a neutral score (0.5).

Since we want to discover claims that are popular and come from low-reputable sources, we linearly combine the two metrics for each edge e, using a tuning parameter θ as follows:

$$weight(e) = \theta\ popularity(e) + (1 - \theta)\ (1 - reputation(e))$$

In our implementation, we slightly favorite low reputation over popularity; thus, we use $\theta = 0.4$.

Claim Ranking. We rank the edges, and consequently the claims, of the *Causality-Based* topologies using the *Betweenness Centrality* metric [6], and the *Aspect-Based* topologies using the *in-Degree* metric. Examples of check-worthy claims in our data include the term pairs: (*Autism – Vaccines*), (*Breast Cancer – Abortion*), and (*Chemotherapy – Cannabis*) (details in §6.3).

5.2 Enhanced Fact-Checking Context

The final step for contextualizing the claims is to relate them (when available) with previously verified claims. To retrieve such claims, we use *ClaimsKG* [55], a knowledge graph that aggregates claims and reviews published using *ClaimReview*[5]. After filtering out, based on the mentioned entities, claims with non-scientific content (i.e., 62.3% of the total claims), we end up with a final set of ~4K scientific claims, out of which 79.8% has been determined to be *False*, and 20.2% has been determined to be *True*. We relate claims by computing their Semantic Textual Similarity [31] and setting an appropriate threshold (0.9 in our experiments).

Our final fact-checking context for scientific claims consists of related scientific papers and news articles from the same cluster, and, if available, related, previously verified claims. As we see in our experiments (§6.3), this enhanced context improves the verification accuracy and confidence of non-expert fact-checkers and helps them outperform commercial fact-checking systems.

6 EXPERIMENTAL EVALUATION

In this section we evaluate the methods for extraction (§6.1), clustering (§6.2), and contextualization (§6.3) of scientific claims.

Raw Dataset. We evaluate all three methods on a state-of-the-art dataset for measuring health-related scientific misinformation [53]. This dataset has the form of a directed graph, from *social media postings* to *news articles* to *scientific papers*, where edges denote a hyperlink connection. The ~50K *social media postings* of the dataset include the text of the postings as well as popularity indicators such as the number of *re-postings* and *likes*. The ~12K *news articles* of the dataset include articles from mainstream news outlets (e.g., theguardian.com or popsci.com), as well as from alternative blogging platforms (e.g., mercola.com or foodbabe.com). Finally, the ~24K *scientific papers* of the dataset include peer-reviewed or gray literature[6] papers hosted at universities, academic publishers, or scientific repositories (e.g., Scopus, PubMed, JSTOR, and CDC). We note that the overall volume of the dataset simulates the typical news coverage on health-related topics for a period of four months.

Table 3: Cross validation of scientific claim extractors. Since, as we explain in §6.1.1, both datasets are balanced, the evaluation metric that we use is *Accuracy* (*ACC*).

		Generic Dataset	**Scientific Dataset**
		ACC	*ACC*
Baseline	*Grammar-Based*	50.4%	52.3%
	Context-Based	49.5%	50.2%
	Random Forest	74.7%	75.6%
	BERT	**82.2%**	81.0%
SciClops	*SciBERT*	81.5%	80.6%
	NewsBERT	82.0%	80.0%
	SciNewsBERT	81.1%	**81.2%**

6.1 Evaluation of Claim Extraction

The evaluation of the extractors is two-fold; first, we validate their accuracy using a widely-used clean and labeled dataset, and then, we use them in a real-world scenario where we apply them on the raw dataset described above, and evaluate them via crowdsourcing.

6.1.1 Training. Since there is no specific training dataset for the task of scientific claim extraction, we use two datasets mainly used for argumentation mining, namely *UKP* [54] and *IBM* [29]. We train our classifiers using the balanced union of the two datasets (~11K positive and negative samples). In the following, we refer to this dataset as the *Generic Dataset* of claims.

We also train our classifiers with a "science-flavored" dataset derived from the *UKP* and *IBM* datasets. Specifically, in this dataset, we oversample claims regarding, e.g., "abortion" and downsample claims regarding, e.g., "school uniforms". We apply this data augmentation by manually processing based on the "general topic" field that exists in both *UKP* and *IBM* datasets. The described dataset is also balanced, containing ~16K positive and negative samples, and in the following, we refer to it as the *Scientific Dataset* of claims.

6.1.2 Cross Validation. We perform a 5-fold cross validation over the datasets described above; the results are shown in Table 3. We observe that the *Heuristic-Based* extractors perform poorly for this task, which confirms that it is a demanding task with many corner cases. Remarkably, the *Context-Based* heuristic, which is domain-agnostic, achieves identical *accuracy* with the *Grammar-Based* heuristic, which contains manually curated grammar rules. We also observe that the *Random Forest* classifier does not perform extremely worse than the *Transformer-Based* models, while being more eco-friendly in terms of resources and training time needed.

The performance of the transformer-based models confirms the fact that they are state-of-the-art in most NLP tasks. However, from this task, we do not see the benefits of the domain-specific pretraining. On the *Generic Dataset*, *BERT*, which is pre-trained on a generic corpus, performs better, while on the *Scientific Dataset*, *SciNewsBERT*, which is pre-trained on a scientific and a news corpus, performs better; nonetheless, their difference is negligible. The real difference among these models is shown in the next experiment.

[5] https://www.claimreviewproject.com

[6] https://en.wikipedia.org/wiki/Grey_literature

Table 4: Crowd Evaluation of scientific claim extraction. Results reported for weak (2 out of 3) annotator agreement (*125* claims - *174* non-claims) and strong (3 out of 3) annotator agreement (*82* claims - *242* non-claims). Since, especially the second set is highly unbalanced, the evaluation metrics that we use are *Precision* (*P*), *Recall* (*R*), and *F1 Score* (*F1*).

		Weak Agreement			**Strong Agreement**		
		P	*R*	*F1*	*P*	*R*	*F1*
Baseline	*Grammar-Based*	51.8%	70.4%	59.9%	40.4%	28.0%	33.1%
	Context-Based	44.6%	49.6%	47.0%	24.5%	45.1%	31.8%
	Random Forest-gen	52.1%	70.4%	59.9%	43.7%	**80.5%**	56.7%
	Random Forest-sci	56.7%	54.4%	55.5%	43.3%	44.8%	44.1%
	BERT-gen	50.8%	50.4%	50.6%	33.5%	68.3%	45.0%
	BERT-sci	78.7%	38.4%	51.6%	79.2%	51.2%	62.2%
SciClops	*NewsBERT-gen*	55.0%	48.8%	51.7%	38.9%	62.2%	47.9%
	NewsBERT-sci	76.9%	40.0%	52.6%	74.2%	56.1%	63.9%
	SciBERT-gen	48.8%	66.4%	56.3%	32.8%	72.0%	45.0%
	SciBERT-sci	48.8%	66.4%	56.2%	**86.5%**	39.1%	53.8%
	SciNewsBERT-gen	49.8%	**80.0%**	**61.3%**	38.8%	78.0%	51.8%
	SciNewsBERT-sci	**84.4%**	30.4%	44.7%	82.7%	52.4%	**64.2%**

6.1.3 Crowd Evaluation. We collect boolean labels for 700 sentences extracted from the raw dataset described above by asking the crowd workers a simple classification question (i.e., whether a given sentence contains a scientific claim or not). We use the platform *Mechanical Turk*, asking input from three independent crowd workers per sentence (57 in total). To ensure high-quality annotations, we employ what the platform calls *Master Workers*, i.e., the most experienced workers with approval rate greater than 80%. Finally, we consider *Strong Agreement* among crowd-workers, the 3 out of 3 agreement, and *Weak Agreement* the 2 out of 3 agreement.

We note that there are 77 out of the 700 sentences for which the majority of the annotators answered *N/A*, because they could not distinguish whether these sentences contain a claim or not. For example, interrogative sentences like *"What? Ibuprofen Can Make You Deaf?"* confused the annotators, while similar affirmative sentences like *"Tylenol PM Causes Brain Damage"* were easily identified as scientific claims. The remaining 623 sentences are divided into two subsets; i) sentences having *Strong Agreement* among annotators, with 82 claims (positive examples) and 242 non-claims (negative examples), and ii) sentences having *Weak Agreement* among annotators, with 125 claims and 174 non-claims.

We observe that especially the subset with *Strong Agreement* is highly unbalanced, which is indeed a realistic scenario if we consider the ratio of claim and non-claim containing sentences in typical news articles. Furthermore, annotators fully agree that a sentence contains a scientific claim for less than 12% of total the sentences, which confirms it is a highly confusing task.

Results. The overall results of the comparison of the extraction models are summarized in Table 4. For all the models, we use the following naming convention: the suffix *-gen* is used to denote that models are trained on the *Generic Dataset* explained in §6.1.1, while suffix *-sci* is used to denote that models are trained on the *Scientific Dataset* also explained in §6.1.1. This convention does not apply to heuristic models that do not require training.

We observe that all the *-gen* models have better or equally good recall as the respective *-sci* models. This happens because *-gen* models have been trained equally towards all the labeled claims and have learned to better recognize the structure of a claim. After analyzing the errors of the models, we noticed that claims with simple structure like *"Repetitive behaviors in autism show sex bias early in life"* were identified more from *-gen* than from *-sci* models. On the other hand, *-sci* models, which have been optimized for the narrow scientific domain, are more selective, hence they show in general better precision than the respective *-gen* models.

Focusing more on the variants of *BERT*, we observe that task-specific pretraining boosts the performance of the model, which is not visible in the first experiment. Specifically, we see that pretraining on both scientific and news domain gives the best results. One illuminative example is the claim *"Galactosides Treat Urinary Tract Infections Without Antibiotics"*, where *Galactosides* is a word that does not appear in the basic vocabulary of *BERT*[7], however, it appears in the extended vocabulary of *SciBERT*[8] and *SciNewsBERT*.

Finally, it is noteworthy that the *Random Forest* model provides quite comparable results to the transformer-based models, while being, as stated above, a much lighter and faster-to-train model.

6.2 Evaluation of Claim-Paper Clustering

Since we construct a bimodal clustering of claims and papers, we evaluate its quality with respect to two axes; a good-quality clustering must contain clusters of semantically related claims and papers (*Semantic Coherence*), and adhere to the implicit connections between these claims and papers (*Interconnection Coherence*).

Semantic Coherence. To measure the semantic coherence of a clustering, we compute a modified version of the *Average Silhouette Width* (*ASW*) [47]. The first modification is that the distance used is not a metric distance (e.g., Euclidean distance) but a semantic distance (Semantic Textual Similarity (*STS*)). The second modification is that we generalize the metric for two (or more) joint clusterings. The original metric computes the average distance between the centroid of each cluster and its elements. In our case, since we have two joint clusterings for claims and papers, we compute the metric for all the combinations of centroids ($\bar{c}$) and elements (e) of each cluster. Thus, the modified *ASW* is computed as follows:

$$ASW(cluster) = \frac{1}{|\,\text{centroids}\,|\cdot|\,\text{cluster}\,|} \sum_{\substack{e\,\in\text{cluster}\\ \bar{c}\,\in\text{centroids}}} STS(e, \bar{c})$$

where *centroids* consists of the claims centroid and the papers centroid of each cluster. Finally, we report the mean *ASW* across all clusters. This cross-computation of the metric allows capturing the semantic coherence of the clusters both individually and jointly.

Interconnection Coherence. To measure the interconnection coherence of the clusterings (i.e., the adaptivity of the clusterings towards the interconnection matrix *L*), we use ideas from link-based recommendation. First, we compute a hard clustering for claims and papers:

$$\begin{aligned} C'_{comp} &= argmax_x(C') \\ P'_{comp} &= argmax_x(P') \end{aligned}$$

Since, as we explain in Table 2, each row of C' and P' contains the probability of a claim or a paper to belong to a cluster, when we compute *argmax* over rows we obtain a hard clustering, while when we compute *argmax* over columns we obtain the cluster centroids. For example, given a single claim c and three clusters cl_0, cl_1, cl_2:

[7] https://cdn.huggingface.co/bert-base-uncased-vocab.txt

[8] https://cdn.huggingface.co/allenai/scibert_scivocab_uncased/vocab.txt

Table 5: Clustering Evaluation. *Semantic Coherence* is measured using the *Average Silhouette Width* (*ASW*), and *Interconnections Coherence* is measured using *Recall@3* (*R@3*).

		clusters=10		clusters=50		clusters=100	
		ASW	R@3	ASW	R@3	ASW	R@3
Content-Based	*LDA*	44.5%	86.8%	63.2%	69.4%	66.6%	69.5%
	GSDMM	42.1%	98.9%	48.5%	86.2%	48.7%	72.4%
	GMM	55.5%	68.9%	67.7%	52.4%	72.8%	45.2%
	PCA/GMM	51.3%	90.0%	66.6%	34.2%	71.7%	28.4%
	K-Means	53.2%	97.9%	**68.9%**	83.4%	73.2%	74.2%
	PCA/K-Means	52.0%	97.6%	66.8%	87.8%	71.2%	75.1%
Graph-Based	*GBA-CP*	38.2%	**100.0%**	40.9%	**100.0%**	44.5%	99.5%
	GBA-C	38.1%	96.7%	44.5%	93.2%	48.7%	92.0%
	GBA-P	40.0%	96.5%	43.0%	93.6%	47.3%	92.3%
	GBT-CP	26.5%	99.6%	27.1%	98.9%	32.1%	71.8%
	GBT-C	37.9%	92.5%	45.0%	59.8%	47.2%	53.8%
	GBT-P	36.4%	88.4%	42.3%	62.4%	43.7%	65.9%
Hybrid	*AO-Content*	54.8%	96.7%	67.9%	90.0%	**73.3%**	92.1%
	AO-Balanced	**56.0%**	99.8%	67.6%	99.6%	72.1%	99.5%
	AO-Graph	55.6%	99.8%	67.3%	**100.0%**	71.8%	**99.8%**

$$c' = [0.1, 0.8, 0, 1] \Rightarrow c'_{comp} = cl_1$$

Next, we use one clustering (e.g., of claims) to recommend possible instances of the other clustering (e.g., of papers). The recommendation is content-agnostic and exploits only the interconnection matrix L. Formally:

$$\begin{aligned} C'_{rec} &= argsort_x(sum_y(L \odot P')) \\ P'_{rec} &= argsort_x(sum_y(L^T \odot C')) \end{aligned}$$

where $\odot$ is the Hadamard (element-wise) product. For the same claim c, papers $p1$ and $p2$, and clusters cl_0, cl_1, cl_2 we have:

$$c \begin{array}{l} \nearrow p_1[0.5, 0.1, 0.4] \\ \\ \searrow p_2[0.1, 0.8, 0.1] \end{array} \Rightarrow c'_{rec} = argsort_x(0.6, 0.9, 0.5) = [cl_1, cl_0, cl_2]$$

To compute the recommendation quality, we utilize the metric of *Recall@k* (*R@k*), which measures the ratio in which the correct cluster is recommended among the top-k results. We report the mean of the *R@k* for the claims and the papers clustering.

Results. The results of the evaluation are shown in Table 5. As we observe, the *Content-Based* (baseline) clustering techniques that use a textual representation of claims and papers (i.e., *LDA* and *GSDMM*), generate clusters with lower *Semantic Coherence* than the ones that use an embeddings representation (i.e., *GMM* and *K-Means*). This is partially explained by a vocabulary mismatch: the language used in papers is more complex and contains more scientific terms than the one used in social and news media (where the claims derive from). Thus, embeddings representations have the advantage of capturing the semantic proximity of topics, even if these topics occur from two heterogeneous vocabularies. Furthermore, we observe that soft clustering techniques (i.e., *LDA* and *GMM*) generate, in general, clusters with higher *Semantic Coherence* than the respective hard clustering techniques (i.e., *GSDMM* and *K-Means*), indicating that the theme of claims and papers is usually multifaceted. Finally, we observe that the dimensionality reduction, performed by *PCA*, is not helpful in the context of this task.

Regarding the *Graph-Based* techniques, we see that they construct clusters with high *Interconnections Coherence* but the lowest *Semantic Coherence*. Not surprisingly, *GBA-CP* achieves the maximum *Interconnections Coherence* since, as we explain in §4.2, it arbitrarily adapts the clusters to the interconnection matrix L.

Overall, we observe that the most robust technique in terms of balance between *Semantic* and *Interconnections Coherence* is the *Hybrid* technique (*AO-Balanced*), which computes a soft clustering based on an embeddings representation and considers both the text and the graph modality of the dataset equally.

6.3 Evaluation of Claim Contextualization

The overall evaluation of our method is performed with an experiment that involves expert and non-expert fact-checkers as well as two state-of-the-art commercial systems. Using SciClops, we extract, cluster, and finally select the *top-40* check-worthy scientific claims in the data collection. The topics of the claims are heterogeneous, covering controversial online discussions such as the usage of therapeutic cannabis in modern medicine, the consumption of small amounts of alcohol during pregnancy, and the effect of vaccines in disorders such as autism.

Claim Post-Processing. We notice that in some of the claims, redundant information that could confuse the fact-checkers is mentioned (e.g., we find the claim "Donald Trump has said vaccines cause autism," in which the scientific question is whether "vaccines cause autism" and not whether Donald Trump made this statement). Thus, to avoid misinterpretations and to mitigate preexisting biases for or against public figures, we replace from these claims all the *Person* and *Organization* entities with indefinite pronouns.

Non-Experts. We employ crowdsourcing workers using the same setup described in §6.1, and ask them to evaluate the *Validity* of each claim in a *Likert Scale* [23] (from "Highly Invalid" to "Highly Valid"). We also ask them to rate their *Effort* to find evidence and their *Confidence* that the evidence they found is correct.

We divide non-experts into one control group of *Non-Experts Without Context*, and two experimental groups of *Non-Experts With Partial Context* and *Non-Experts With Enhanced Context*:

- *Non-Experts Without Context* are shown a bare scientific claim with no additional information, as they would read it online in, e.g., a messaging app or a social media posting.
- *Non-Experts With Partial Context* are shown a scientific claim and its source news article, i.e., the news article from which the claim was extracted.
- *Non-Experts With Enhanced Context* are shown a scientific claim, its source news article, and: i) the top-k news articles where the same or similar claims were found, ii) the top-k most relevant papers, and, if available, iii) the top-k most similar, previously verified claims. To avoid overwhelming this experimental group with redundant information, we set $k = 3$.

Experts. We ask two independent experts to evaluate the validity of the claims. Each expert evaluated all 40 claims independently, and was given the chance to cross-check the ratings by the other expert and revise their own ratings, if deemed appropriate. Overall, we use the average of the two expert ratings as ground-truth.

Commercial Systems. Finally, for the verification of the same scientific claims, we use two commercial systems for fact-checking, namely ClaimBuster [20] and Google Fact Check Explorer[9]:

- *ClaimBuster* is a system used massively by journalists which initially aimed at detecting important factual claims in political

[9] https://toolbox.google.com/factcheck/explorer

Table 6: Left: Root Mean Square Error (*RMSE*) between the scores provided by the *Experts* (ground-truth) and the scores provided by *Non-Experts* and *Commercial Systems*; the last row shows the *RMSE* across *Experts* (lower is better). Right: Verification of two contradictory claims from *CNN* and *MensJournal* by *Non-Experts* and *Commercial Systems*; the last row shows the ground-truth provided by the *Experts*.

	RMSE	CNN Claim	MensJournal Claim
Non-Experts			
Without Context	1.91	*Borderline*	*Borderline*
With Partial Context	1.73	***Valid***	*Valid*
With Enhanced Context	**1.54**	***Valid***	***Highly Invalid***
Commercial Systems			
ClaimBuster	1.74	***Valid***	*Borderline*
Google Fact Check Explorer	2.79	*N/A*	*N/A*
Experts	1.02	***Highly Valid***	***Highly Invalid***

discourses; however, its current architecture allows for investigating any kind of check-worthy claims (details in §2).

- *Google Fact Check Explorer* is also an exploration tool used by journalists to verify claims published using the tagging system of *ClaimReview*; we note that *ClaimReview* is also exploited in the contextualization step of SciClops (details in §5.2).

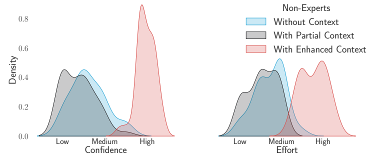
To homogenize the scores of these systems with the scores of the fact-checkers, we quantize them to the aforementioned *Likert Scale*.

Results. Results are summarized in Table 6. Given the ground-truth provided by the experts, we measure the accuracy of the three aforementioned groups of non-experts and the two commercial systems using the *Root Mean Square Error* (*RMSE*).

We observe that *ClaimBuster* performs better than our control group of *Non-Experts Without Context* while providing a solution without human intervention. Furthermore, we observe that *Google Fact Check Explorer* performs poorly, mainly because only 20% of the queried claims were present in the fact-checking portals it monitors (e.g., the claim "*Vaccines cause Autism*" is present in the fact-checking section of *USA Today* [56], while the *Contradictory Claims* described next are absent from all the fact-checking portals).

Finally, regarding the non-expert human fact-checkers, we observe that the more contextual information is available, the more accurately they rate the claims. Indicatively, the *RMSE* of *Non-Experts With Enhanced Context* is only *50%* greater than the *RMSE* across *Experts*. Overall, we see that, when the under-verification claims derive from a narrow scientific domain, **non-expert human fact-checkers, provided with the proper fact-checking context, may outperform state-of-the-art commercial systems**.

Case Study: Contradictory Claims. Within the set of under-verification claims, we noticed two contradictory claims. The first claim opposes the use of therapeutic cannabis for treating *Post-Traumatic Stress Disorder* (*PTSD*) and comes from a mainstream news outlet (*CNN*).[10] The second claim supports the use of cannabis for treating *PTSD* and comes from a popular health blog (*MensJournal*).[11] Current scientific understanding supports the first claim (from *CNN*), but not the second one (from *MensJournal*), as evidenced by a paper of the *Journal of Clinical Psychiatry* [58].



Average Work Time of *Non-Experts* (seconds)		
Without Context	*With Partial Context*	*With Enhanced Context*
191s	*388s*	***898s***

Figure 2: Kernel Density Estimation (*KDE*) of *Confidence* (left) and estimated *Effort* (right), and *Average Work Time* (bottom), of *Non-Experts* verifying claims. Best seen in color.

As we show in Table 6, *ClaimBuster* and all the groups of *Non-Experts* mostly support the claim from *CNN* as valid. Moreover, as discussed above, *Google Fact Check Explorer* provides no answer for these two claims since they are not present in the monitored fact-checking portals. Indeed, only *Non-Experts With Enhanced Context* were able to indicate that the claim from *MensJournal* is invalid, mainly because **SciClops provided a fact-checking context that included the paper from the *Journal of Clinical Psychiatry* which debunks the claim even in its title.**[12]

Case Study: Confidence & Effort. As we observe in Figure 2, *Non-Experts* that were shown the *Enhanced Context* of claims were more confident in their verification, additionally to being more accurate than the other two groups of users, which is partially explained by the fact that the provided context is fully-interpretable (as explained above), thus more trustworthy. However, the same users' self-assessment of their effort as well as their actual work time was higher than the other two groups of users, which is explained by the fact that they had to visit more potential verification sources.

7 CONCLUSIONS

We have presented an effective method for assisting non-experts in the verification of scientific claims. We have shown that transformer models are indeed the state-of-the-art on scientific claim detection, however, they require domain-specific fine-tuning to perform better than other baselines. We have also shown that, by exploiting the text of a claim and its connections to scientific papers, we effectively cluster topically-related claims and papers, as well as that, by building an in-cluster knowledge graph, we enable the detection of check-worthy claims. Overall, we have shown that SciClops can build the appropriate fact-checking context to help non-expert fact-checkers verify complex scientific claims, outperforming commercial systems. We believe that our method complements these systems in domains with sparse or non-existing ground-truth evidence, such as the critical domains of science and health.

Reproducibility. All the data, code, models, as well as expert and crowd annotations used for this paper are publicly available for research purposes in ***http://scilens.epfl.ch***.

Acknowledgments. We would like to thank the external experts: *Dimitra Synefiaridou (Postdoctoral Researcher in Microbiology)* and *Sylvia Kyrilli (Senior Pediatrician)*. This work has been partially supported by: the *Open Science Fund of EPFL*; the *Swiss Academy of Engineering Sciences*; the *HUMAINT project of the Joint Research Centre of European Commission*; the *La Caixa project (LCF/PR/PR16/11110009)*.

[10] *CNN*: "*Marijuana does not treat chronic pain or post-traumatic stress disorder.*" [49]

[11] *MensJournal*: "*Marijuana can help battle depression, anxiety, post-traumatic stress disorder, and even addictions to alcohol and painkillers.*" [24]

[12] *J. of Clinical Psychiatry*: "*Marijuana use is associated with worse outcomes in symptom severity and violent behavior in patients with posttraumatic stress disorder.*" [58]

REFERENCES

[1] Alex Berezow. March 5, 2017. Infographic: The Best and Worst Science News Sites. *American Council on Science and Health* (March 5, 2017). https://acsh.org/news/2017/03/05/infographic-best-and-worst-science-news-sites-10948

[2] Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, Shaden Shaar, and Zien Sheikh Ali. 2020. Overview of CheckThat! 2020: Automatic Identification and Verification of Claims in Social Media. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22-25, 2020, Proceedings (Lecture Notes in Computer Science, Vol. 12260)*, Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika, Stefanos Vrochidis, Hideo Joho, Christina Lioma, Carsten Eickhoff, Aurélie Névéol, Linda Cappellato, and Nicola Ferro (Eds.). Springer, 215–236. https://doi.org/10.1007/978-3-030-58219-7_17

[3] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. 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 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 3613–3618. https://doi.org/10.18653/v1/D19-1371

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (2003), 993–1022. http://jmlr.org/papers/v3/blei03a.html

[5] G. E. P. Box and D. R. Cox. 1964. An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)* 26, 2 (1964), 211–252. http://www.jstor.org/stable/2984418

[6] Ulrik Brandes. 2001. A faster algorithm for betweenness centrality. *The Journal of Mathematical Sociology* 25, 2 (2001), 163–177. https://doi.org/10.1080/0022250X.2001.9990249 arXiv:https://doi.org/10.1080/0022250X.2001.9990249

[7] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. TabFact: A Large-scale Dataset for Table-based Fact Verification. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=rkeJRhNYDH

[8] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*, Jofish Kaye, Allison Druin, Cliff Lampe, Dan Morris, and Juan Pablo Hourcade (Eds.). ACM, 2098–2110. https://doi.org/10.1145/2858036.2858207

[9] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M. Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational Fact Checking from Knowledge Networks. *PLOS ONE* 10, 6 (jun 2015), e0128193. https://doi.org/10.1371/journal.pone.0128193

[10] 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://www.aclweb.org/anthology/N19-1423/

[11] Chi Thang Duong, Quoc Viet Hung Nguyen, and Karl Aberer. 2019. Interpretable node embeddings with mincut loss. *Learning and Reasoning with Graph-Structured Representations Workshop - ICML 2019* (2019).

[12] Elisa Shearer. 10 December 2018. Social media outpaces print newspapers in the U.S. as a news source. *Pew Research Center* (10 December 2018). https://www.pewresearch.org/fact-tank/2018/12/10/social-media-outpaces-print-newspapers-in-the-u-s-as-a-news-source

[13] Lei Fang, George Karakiulakis, and Michael Roth. 2020. Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection? *The Lancet Respiratory Medicine* (2020).

[14] Karl Pearson F.R.S. 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2, 11 (1901), 559–572. https://doi.org/10.1080/14786440109462720

[15] Daniel Funke. January 19, 2018. Google suspends fact-checking feature over quality concerns. *Poynter* (January 19, 2018). https://www.poynter.org/fact-checking/2018/google-suspends-fact-checking-feature-over-quality-concerns

[16] Mohamed H. Gad-Elrab, Daria Stepanova, Jacopo Urbani, and Gerhard Weikum. 2019. ExFaKT: A Framework for Explaining Facts over Knowledge Graphs and Text. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019*, J. Shane Culpepper, Alistair Moffat, Paul N. Bennett, and Kristina Lerman (Eds.). ACM, 87–95. https://doi.org/10.1145/3289600.3290996

[17] William L. Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 1024–1034. http://papers.nips.cc/paper/6703-inductive-representation-learning-on-large-graphs

[18] Casper Hansen, Christian Hansen, Stephen Alstrup, Jakob Grue Simonsen, and Christina Lioma. 2019. Neural Check-Worthiness Ranking with Weak Supervision: Finding Sentences for Fact-Checking. In *Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, Sihem Amer-Yahia, Mohammad Mahdian, Ashish Goel, Geert-Jan Houben, Kristina Lerman, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (Eds.). ACM, 994–1000. https://doi.org/10.1145/3308560.3316736

[19] Naeemul Hassan, Bill Adair, James T Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. 2015. The quest to automate fact-checking. In *Proceedings of the 2015 Computation+ Journalism Symposium*.

[20] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*. ACM, 1803–1812. https://doi.org/10.1145/3097983.3098131

[21] Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. ClaimRank: Detecting Check-Worthy Claims in Arabic and English. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 2-4, 2018, Demonstrations*, Yang Liu, Tim Paek, and Manasi S. Patwardhan (Eds.). Association for Computational Linguistics, 26–30. https://doi.org/10.18653/v1/n18-5006

[22] Shan Jiang, Simon Baumgartner, Abe Ittycheriah, and Cong Yu. 2020. Factoring Fact-Checks: Structured Information Extraction from Fact-Checking Articles. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen (Eds.). ACM / IW3C2, 1592–1603. https://doi.org/10.1145/3366423.3380231

[23] Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *Current Journal of Applied Science and Technology* (2015), 396–403.

[24] Melaina Juntti. 2017. Study: Marijuana Can Help Battle Depression, Anxiety, PTSD, and Addiction. *Men's Journal* (2017). https://www.mensjournal.com/health-fitness/study-marijuana-can-help-battle-depression-anxiety-ptsd-and-addiction-w453012

[25] Georgios Karagiannis, Mohammed Saeed, Paolo Papotti, and Immanuel Trummer. 2020. Scrutinizer: A Mixed-Initiative Approach to Large-Scale, Data-Driven Claim Verification. *Proc. VLDB Endow.* 13, 11 (2020), 2508–2521. http://www.vldb.org/pvldb/vol13/p2508-karagiannis.pdf

[26] Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task Learning for Rumour Verification. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, Emily M. Bender, Leon Derczynski, and Pierre Isabelle (Eds.). Association for Computational Linguistics, 3402–3413. https://www.aclweb.org/anthology/C18-1288/

[27] Rohit Kulkarni. 2018. A Million News Headlines. https://doi.org/10.7910/DVN/SYBGZL

[28] Stamatios Lefkimmiatis, Aurélien Bourquard, and Michael Unser. 2012. Hessian-Based Norm Regularization for Image Restoration With Biomedical Applications. *IEEE Trans. Image Process.* 21, 3 (2012), 983–995. https://doi.org/10.1109/TIP.2011.2168232

[29] Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context Dependent Claim Detection. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, Jan Hajic and Junichi Tsujii (Eds.). ACL, 1489–1500. https://www.aclweb.org/anthology/C14-1141/

[30] Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 2016. Topic Modeling for Short Texts with Auxiliary Word Embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, Raffaele Perego, Fabrizio Sebastiani, Javed A. Aslam, Ian Ruthven, and Justin Zobel (Eds.). ACM, 165–174. https://doi.org/10.1145/2911451.2911499

[31] Matthias Liebeck, Philipp Pollack, Pashutan Modaresi, and Stefan Conrad. 2016. HHU at SemEval-2016 Task 1: Multiple Approaches to Measuring Semantic Textual Similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, Steven Bethard, Daniel M. Cer, Marine Carpuat, David Jurgens, Preslav Nakov, and Torsten Zesch (Eds.). The Association for Computer Linguistics, 595–601. http://aclweb.org/anthology/S/S16/S16-1090.pdf

[32] Marco Lippi and Paolo Torroni. 2015. Context-Independent Claim Detection for Argument Mining. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, Qiang Yang and Michael J. Wooldridge (Eds.). AAAI Press, 185–191. http://ijcai.org/Abstract/15/033

[33] Stuart P. Lloyd. 1982. Least squares quantization in PCM. *IEEE Trans. Information Theory* 28, 2 (1982), 129–136. https://doi.org/10.1109/TIT.1982.1056489

[34] Yukun Ma, Haiyun Peng, and Erik Cambria. 2018. Targeted Aspect-Based Sentiment Analysis via Embedding Commonsense Knowledge into an Attentive LSTM. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press, 5876–5883. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16541

[35] George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (1995), 39–41. https://doi.org/10.1145/219717.219748

[36] Merja Myllylahti. 2018. An attention economy trap? An empirical investigation into four news companies' Facebook traffic and social media revenue. *Journal of Media Business Studies* 15, 4 (2018), 237–253. https://doi.org/10.1080/16522354.2018.1527521

[37] Moin Nadeem, Wei Fang, Brian Xu, Mitra Mohtarami, and James R. Glass. 2019. FAKTA: An Automatic End-to-End Fact Checking System. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, Waleed Ammar, Annie Louis, and Nasrin Mostafazadeh (Eds.). Association for Computational Linguistics, 78–83. https://doi.org/10.18653/v1/n19-4014

[38] Preslav Nakov, David P. A. 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. *CoRR* abs/2103.07769 (2021). arXiv:2103.07769 https://arxiv.org/abs/2103.07769

[39] Ayush Patwari, Dan Goldwasser, and Saurabh Bagchi. 2017. TATHYA: A Multi-Classifier System for Detecting Check-Worthy Statements in Political Debates. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, Ee-Peng Lim, Marianne Winslett, Mark Sanderson, Ada Wai-Chee Fu, Jimeng Sun, J. Shane Culpepper, Eric Lo, Joyce C. Ho, Debora Donato, Rakesh Agrawal, Yu Zheng, Carlos Castillo, Aixin Sun, Vincent S. Tseng, and Chenliang Li (Eds.). ACM, 2259–2262. https://doi.org/10.1145/3132847.3133150

[40] Dario Pavllo, Tiziano Piccardi, and Robert West. 2018. Quootstrap: Scalable Unsupervised Extraction of Quotation-Speaker Pairs from Large News Corpora via Bootstrapping. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*. AAAI Press, 231–240. https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17827

[41] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). ACL, 1532–1543. http://aclweb.org/anthology/D/D14/D14-1162.pdf

[42] José María González Pinto, Janus Wawrzinek, and Wolf-Tilo Balke. 2019. What Drives Research Efforts? Find Scientific Claims that Count!. In *19th ACM/IEEE Joint Conference on Digital Libraries, JCDL 2019, Champaign, IL, USA, June 2-6, 2019*, Maria Bonn, Dan Wu, J. Stephen Downie, and Alain Martaus (Eds.). IEEE, 217–226. https://doi.org/10.1109/JCDL.2019.00038

[43] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media. In *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017*, Rick Barrett, Rick Cummings, Eugene Agichtein, and Evgeniy Gabrilovich (Eds.). ACM, 1003–1012. https://doi.org/10.1145/3041021.3055133

[44] Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and Clustering of Arguments with Contextualized Word Embeddings. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 567–578. https://www.aclweb.org/anthology/P19-1054/

[45] Douglas A. Reynolds. 2009. Gaussian Mixture Models. In *Encyclopedia of Biometrics*, Stan Z. Li and Anil K. Jain (Eds.). Springer US, 659–663. https://doi.org/10.1007/978-0-387-73003-5_196

[46] Md Main Uddin Rony, Naeemul Hassan, and Mohammad Yousuf. 2017. Diving Deep into Clickbaits: Who Use Them to What Extents in Which Topics with What Effects?. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, Sydney, Australia, July 31 - August 03, 2017*, Jana Diesner, Elena Ferrari, and Guandong Xu (Eds.). ACM, 232–239. https://doi.org/10.1145/3110025.3110054

[47] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53 – 65. https://doi.org/10.1016/0377-0427(87)90125-7

[48] Dietram A. Scheufele. 2013. Communicating science in social settings. *Proceedings of the National Academy of Sciences* 110, Supplement 3 (2013), 14040–14047. https://doi.org/10.1073/pnas.1213275110

[49] Susan Scutti. August 14, 2017. Little evidence that marijuana helps chronic pain, PTSD, studies find. *CNN* (August 14, 2017). https://edition.cnn.com/2017/08/14/health/medical-marijuana-pain-ptsd-study/index.html

[50] 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, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 3607–3618. https://www.aclweb.org/anthology/2020.acl-main.332/

[51] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A Platform for Tracking Online Misinformation. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11-15, 2016, Companion Volume*, Jacqueline Bourdeau, Jim Hendler, Roger Nkambou, Ian Horrocks, and Ben Y. Zhao (Eds.). ACM, 745–750. https://doi.org/10.1145/2872518.2890098

[52] Julia Sittmann and Andrew Tompkins. July 17, 2020. The strengths and weaknesses of automated fact-checking tools. *Deutsche Welle* (July 17, 2020). https://www.dw.com/en/the-strengths-and-weaknesses-of-automated-fact-checking-tools/a-53956958

[53] Panayiotis Smeros, Carlos Castillo, and Karl Aberer. 2019. SciLens: Evaluating the Quality of Scientific News Articles Using Social Media and Scientific Literature Indicators. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (Eds.). ACM, 1747–1758. https://doi.org/10.1145/3308558.3313657

[54] Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic Argument Mining from Heterogeneous Sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 3664–3674. https://www.aclweb.org/anthology/D18-1402/

[55] Andon Tchechmedjiev, Pavlos Fafalios, Katarina Boland, Malo Gasquet, Matthäus Zloch, Benjamin Zapilko, Stefan Dietze, and Konstantin Todorov. 2019. ClaimsKG: A Knowledge Graph of Fact-Checked Claims. In *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 11779)*, Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtech Svátek, Isabel F. Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois, and Fabien Gandon (Eds.). Springer, 309–324. https://doi.org/10.1007/978-3-030-30796-7_20

[56] Bayliss Wagner. April 18, 2021. Fact check: Autism diagnosis criteria changes have led to increased rates. *USA Today* (April 18, 2021). https://www.usatoday.com/story/news/factcheck/2021/04/18/fact-check-autism-diagnosis-changes-over-years-account-high-rate/7102414002

[57] Xiao Wang, Peng Cui, Jing Wang, Jian Pei, Wenwu Zhu, and Shiqiang Yang. 2017. Community Preserving Network Embedding. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, Satinder P. Singh and Shaul Markovitch (Eds.). AAAI Press, 203–209. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14589

[58] Samuel T Wilkinson, Elina Stefanovics, and Robert A Rosenheck. 2015. Marijuana use is associated with worse outcomes in symptom severity and violent behavior in patients with posttraumatic stress disorder. *The Journal of clinical psychiatry* 76, 9 (sep 2015), 1174–1180. https://doi.org/10.4088/JCP.14m09475

[59] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *CoRR* abs/1910.03771 (2019). arXiv:1910.03771 http://arxiv.org/abs/1910.03771

[60] Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph Convolutional Networks for Text Classification. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 7370–7377. https://doi.org/10.1609/aaai.v33i01.33017370

[61] Yang Zhou, Hong Cheng, and Jeffrey Xu Yu. 2009. Graph Clustering Based on Structural/Attribute Similarities. *PVLDB* 2, 1 (2009), 718–729. https://doi.org/10.14778/1687627.1687709

[62] 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 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 2099–2108. https://doi.org/10.18653/v1/D19-1216