Truth or Dare: Investigating Claims Truthfulness with ClaimsKG

Susmita Gangopadhyay, Katarina Boland, Danilo Dessí, Stefan Dietze, Pavlos Fafalios, Andon Tchechmedjiev, Konstantin Todorov, Hajira Jabeen

Truth or Dare: Investigating Claims Truthfulness with ClaimsKG

Susmita Gangopadhyay[1], Katarina Boland[1], Danilo Dessí[1], Stefan Dietze[1,2], Pavlos Fafalios[3], Andon Tchechmedjiev[4], Konstantin Todorov[5] and Hajira Jabeen[1]

[1]*KTS Department, GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany*
[2]*Heinrich-Heine-University Dusseldorf, Dusseldorf, Germany*
[3]*Institute of Computer Science, FORTH-ICS, Greece*
[4]*EuroMov Digital Health in Motion, Univ. Montpellier, IMT Mines Alès, Alès, France*
[5]*LIRMM / University of Montpellier / CNRS, France*

Abstract

Searching and exploring online information is fundamental for our society. However, it is common to find inaccurate information on the Internet, that can quickly spread and be hard to identify. Fortunately, today, many fact-checking sources verify online information to provide online users with a means to recognize its truthfulness. These sources use different languages and scoring systems, which makes fact validation challenging and time-consuming. To address this issue, we propose a new release of ClaimsKG, a knowledge graph of about 59,580 claims, which covers 13 different fact-checking sources and provides a structured way to retrieve verified online claims. ClaimsKG is built using a pipeline that makes use of entity linking and disambiguation tools to fetch entities from DBpedia and an ad-hoc scoring normalization system. ClaimsKG is used as a showcase to provide the public with interesting and verified information about events of our times.

Keywords

Claims, Knowledge Graph, Fact-Checking, Entity-Fishing, RDF

1. Introduction

Fact-checking is the task of assessing the veracity of claims made by the public. Currently, there is worldwide concern over the spread of fake news and how it affects social, political, and economic well-being. In recent times, we have seen misinformation spreading faster than truth [1]. The dissemination of fake news has sparked widespread interest among researchers and evolved active research directions in the field of automatic fact-checking [2], fake news detection [3], or spreading patterns of online discourse [4]. Various fact-checking organizations around the world have employed journalists dedicated to this cause. Large amounts of claims are processed at regular intervals to manually assess their credibility based on sources, facts, and figures. Nevertheless, it is still challenging to grasp the trustworthiness of their content. The reason for this is that fact-checking websites and companies do not express such information in a structured way that might be accessible to the public from a unique entry point and can

Second International Workshop on Linked Data-driven Resilience Research (D2R2'23) co-located with ESWC 2023, May 28th, 2023, Hersonissos, Greece

susmita.gangopadhyay@gesis.org (S. Gangopadhyay); katarina.boland@gesis.org (K. Boland); danilo.dessi@gesis.org (D. Dessí); stefan.dietze@gesis.org (S. Dietze); fafalios@ics.forth.gr (P. Fafalios); andon.tchechmedjiev@mines-ales.fr (A. Tchechmedjiev); konstantin.todorov@lirmm.fr (K. Todorov); hajira.jabeen@gesis.org (H. Jabeen)

be processed by machines to provide a variety of services. For example, among the existing fact-checking sources, *Politifact* uses "correct", *Snopes* uses "Correctly Attributed", while *AFP Factuel* uses "Vrai" to state that a claim reports a truth. This makes the task of interpreting and understanding Internet content difficult. Thus, novel ways to access and use online information are demanded. To fulfill this need, ClaimsKG, an RDF knowledge graph (KG) of fact-checked claims, enabling structured queries about their truth values, authors, dates, related entities, and metadata, was released in 2019 [5]. ClaimsKG makes it easier for users to explore claims in a standardized manner, enabling the discovery and search of fact-checked online information. However, as a result of the dynamic nature of Internet content and fact-checking source websites, ClaimsKG is continuously evolving. This paper presents the latest release of ClaimsKG named *ClaimsKG (Aug2022)* which is generated by a pipeline that periodically harvests data from popular fact-checking sources. Furthermore, claims and their review articles are enriched with entities from DBpedia using a state-of-the-art entity-linking tool. The collected information is described using a specifically designed RDFS model based on well-established vocabularies such as *Schema.org*[1] and *NIF*[2]. Lastly, to simplify the comprehension of claim veracity for users, we introduced a *Normalized Truth Rating*[3] scheme, containing four generic categories: *TRUE, FALSE, MIXTURE,* and *OTHER*. ClaimsKG will be released at regular intervals, maintaining the pipeline updated with state-of-the-art tools and methods, and covering a larger set of fact-checking sources. The contribution of this paper is threefold:

- We present the latest release of ClaimsKG, the largest collection of multilingual claims and associated metadata.
- We describe the novelties of the ClaimsKG construction process, resulting data and release its source code[4,5].
- We present various use cases using federated SPARQL queries to uncover information that would be difficult or impossible to discover without ClaimsKG.

2. Related work

Several studies have utilized machine learning to identify fake news, and one such approach is the Deep Triple Network (DTN) [6] that employs knowledge graphs to aid in detecting fake news, along with triple-enhanced explanations. The DTN utilizes background knowledge graphs, including open knowledge graphs and graphs extracted from news databases, for feature extraction to classify the news article. The work in [7] semantically detects fake news that utilizes relational features, such as sentiment, entities, and facts directly extracted from the text. They demonstrate that the inclusion of semantic features leads to improved accuracy in classifying fake news. Sciclops [8] proposes a method involving extraction, clustering, and contextualization for analyzing scientific claims in social media posts. A recent survey [9] has examined the utilization of semantic KGs in the integration of heterogeneous news information. While it shows that there has been previous work on data provision for claims detection, e.g., the work in [10] which provides a static data set for claims detection, no other dataset

[1]Schema.org: https://schema.org
[2]NIF: https://persistence.uni-leipzig.org/nlp2rdf/
[3]Normalized Truth Rating: https://data.gesis.org/claimskg/ratings.pdf
[4]Extractor source code: https://github.com/claimskg/claimskg-extractor/tree/latest_release
[5]Generator source code: https://github.com/claimskg/claimskg_generator/tree/latest_release

Table 1
Statistics about claims harvested from each website (as of Aug 2022)

Websites	URL	Total Claims	Total Entities	True Claims	False Claims	Mixture Claims	Other Claims
Global	NA	59,580	1,371,271	7,151	30,858	10,790	10,781
Politifact	https://politifact.com	21,450	354,653	2,501	8,353	6,733	3,863
Snopes	https://snopes.com	14,031	481,199	2,843	6,803	2,556	1,829
AFP Factcheck	https://factcheck.afp.com	5,058	151,208	3	4,147	97	811
AFP Factuel(FR)	https://factuel.afp.com	935	18,739	5	627	94	209
Checkyourfact	https://checkyourfact.com	3,971	16,699	233	3,691	4	43
Vishva news	https://www.vishvasnews.com	3,490	8,930	0	2,933	565	0
Fullfact	https://fullfact.org	2,928	6,870	403	729	152	1,644
Truth or Fiction	https://truthorfiction.com	2,908	21,298	853	260	14	1,781
Africacheck	https://africacheck.org	2,854	11,448	197	2,364	258	35
Fatabyyano	https://fatabyanno.net	1,379	101	46	820	4	509
Factograph	https://factograph.info	255	1,201	19	69	144	23
Eu Factcheck	https://eufactcheck.eu	297	5,699	48	48	159	42
Polygraph	https://polygraph.info	24	293,226	0	14	10	0

is mentioned that could be compared to ClaimsKG, which is a verified claim collection of continuous longitudinal nature (i.e., a systematic, ongoing process of claims collection).

3. ClaimsKG (Aug/2022) Overview and Statistics

This section describes the latest release of ClaimsKG (available at https://doi.org/10.7802/2469) and reports its statistics. ClaimsKG contains 59, 580 claims harvested from 13 popular fact-checking sources. The websites are selected based on the International Fact-Checking Network's (IFCN)[6] signatories list and considered only sources referred by the fact-checking community as highly reputable. The list of covered sources is mentioned in Table 1. As part of the new release, we have crawled new claims from the fact-checking sites that were included in the previous release and added *Factograph*, *Fatabyanno*, *Eufactcheck*, *Vishvasnews* and *Polygraph*. This release does not contain claims from *Factscan.ca* since its website is no longer online. However, past harvested data from *Factscan.ca* will still be available in previous ClaimsKG releases. Harvested sources contain claims in various languages like English, French, Russian, Urdu, Hindi, Punjabi, Assamese, Tamil, Malayalam, Gujarati, Telegu, Marathi, Odia, and Bengali, thus making it interesting for a broad audience. The time frame for collected claims ranges from 1996 to August 2022. Since these sources were launched at different points in time, the start year for the earliest claims of each website is different. This allows the study of a multitude of entities through fact-checked claims that are contained by the sources and provides the users with the possibility to study events over a long period. In the latest version, the earliest claim from the year 1996 belongs to the website Snopes. This claim describes "A hostess named Deborah Gail Stone working the America Sings attraction was crushed to death by a rotating wall.", labeled as "True". Towards the end of the pipeline (see Section 4) and after each run we generate statistics, both global and per source, to monitor the health of the extracted data and also to keep track of the recent changes in the websites. Table 1 provides information on the total number of claims collected from each of these sources, the total number of entities mentioned in these claims and their reviews, and the veracity label of claims obtained from

[6]https://www.poynter.org/ifcn/

Table 2
Links to ClaimsKG data and tools

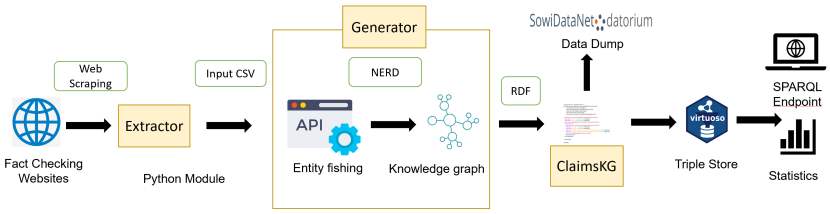
Data/Tools	Links
ClaimsKG Website	https://data.gesis.org/claimskg/
Dataset Download	https://doi.org/10.7802/2469
Previous Versions	https://zenodo.org/record/3518960
DCAT Description	Included in the KG
SPARQL Endpoint	https://data.gesis.org/claimskg/sparql
Claims Ontology	https://data.gesis.org/claimskg/#Data-model
Source Code	https://github.com/claimskg/claimskg-extractor/tree/latest_release https://github.com/claimskg/claimskg_generator/tree/latest_release
ClaimsKG Explorer [11]	https://data.gesis.org/claimskg/explorer

each website. One can find further information such as the percentage of claims that contain date published and author names, the number of entities per review, the number of entities per claim, and the total number of keywords for each fact-checking source on ClaimsKG website (see Table 2). For other information such as the latest data dump, SPARQL query endpoint, usage instructions, and source code, refer to Table 2. The latest release of ClaimsKG supersedes the previous versions as it contains all the claims from the previous versions together with additional claims as well as improved entity annotations.

4. ClaimsKG Pipeline

In this section, we discuss the processing pipeline of ClaimsKG and also discuss the updates for the newest release. The pipeline consists of two major building blocks, namely the Extractor and Generator. The steps of the pipeline are summarized in Figure 1.

Figure 1: Pipeline of ClaimsKG depicting all its modules



4.1. Extractor

In the Extractor module, we perform web scraping of the identified sources. The web-crawling process is different for each source since it is tailor-made and adapted to the structure of each website. For this release, we added five new sources which needed new ad-hoc sub-modules. This module collects the information in a JSON and consolidates it as a CSV file. The data consists of the claim text, its truth value or original rating, the claim body, a link to the claim review from the fact-checking website, the references cited in the claim reviews, the author of the claim, the author of the review, the date of publication of the claim, the date of the review if available, the title of the review article, and a set of topic keywords extracted from the source websites if available. This module's main pipeline and working mechanism to harvest data have remained similar to the previous ClaimsKG-generating pipeline but we have made changes according to newly added sources. This module does not perform any data processing so that the time for harvesting the source websites is not overloaded with additional computation.

4.2. Generator

The output of the Extractor serves as input to the Generator. The Generator performs: i) entity annotation and linking ii) rating normalization, and iii) lifting and serialization.

4.2.1. Entity Annotation and Linking

In this phase, we perform Named Entity Recognition and Disambiguation (NERD) of the claims and their reviews. Differently from the previous pipeline, the entity annotation task is performed within the *Generator* so that it can be applied to the whole downloaded corpus of fact-checked claims at once. This allows a better separation of various modules of the pipeline, making it easy to maintain. One major update in this module is the use of Python Entity Fishing Client (PEFC)[7] instead of TagMe[8]. The reason for this choice is twofold: (i) TagMe code is legacy and we wanted to move to a more recent and easily deployable service, (ii) PEFC supports multilingual claims in ClaimsKG and performs entity recognition and disambiguation against Wikidata in 11 different languages (English, French, German, Italian, Spanish, Arabic, Mandarin, Russian, Japanese, Portuguese, and Farsi). PERC uses GROBID-NER[9], a Named-Entity Recogniser based on GROBID for recognizing the Entities. GROBID-NER has 27 Named Entity classes and is specifically dedicated to supporting the resolution and disambiguation of entities against Knowledge Bases. We use SPARQL queries through the Entity Fishing API to fetch DBpedia entities. We run a local version of the NERDClient, using the latest available dump of Wikipedia and Wikidata on Feb 1, 2022, as reported in the guideline[10].

4.2.2. Rating Normalization

We observed that fact-checking source websites have different rating systems with non-uniform labels. For example, Politifact uses labels like "Pants on Fire" while AFP Factcheck has values like "Misleading", and "Satire". To make the rating uniform, we provide a normalized rating score for all claims in the dataset, alongside the original ratings. We classified sources into four categories *TRUE, FALSE, MIXTURE, OTHER* respectively indicated within ClaimsKG with rating values 3, 1, 2, -1, and only labeled a claim as *TRUE* or *FALSE* if it was completely true or completely false and did not have any ambiguity in their ratings. *MIXTURE* is assigned to claims which hold a degree of both truth and falsehood, such as "half-true", "Truth! But Postponed!", or "misleading". *OTHER* is for claims that do not fit into the TRUE/FALSE or MIXTURE categories and has rating names like "Pending Investigation" or "photo out of context", among several others. The entire approach to rating normalization is similar to what was performed in the previous releases. In the newest version of the generating pipeline, this module has been extended to normalize the ratings of the newly added sources.

4.2.3. Lifting and Serialization

The data model of the KG is available on the ClaimsKG website mentioned in Table 2. We used the *Rdflib*[11] python library to create the model and an abstract RDF graph to then serialize it in

[7] https://github.com/kermitt2/entity-fishing
[8] https://sobigdata.d4science.org/web/tagme/tagme-help
[9] https://github.com/kermitt2/grobid-ner
[10] https://nerd.readthedocs.io/en/latest/build.html
[11] Rdflib: https://rdflib.readthedocs.io/en/stable/

Figure 2: Sample output of a claim from the website TruthorFiction.

```
@prefix ckg: http://data.gesis.org/claimskg .

<ckg/claim_review/001ef100-7cd2-5ef1-830c-6159cc771a88> a schema:ClaimReview ;
    schema:author <ckg/organization/truthorfiction> ;
    schema:datePublished "2019-06-26"^^xsd:date ;
    schema:headline "Does It Cost $750 a Day to House Migrant Children in Camps? –..."@en ;
    schema:inLanguage <ckg/language/English> ;
    schema:itemReviewed <ckg/creative_work/8b1af5ad-48ee-5174-8b42-3b92f62fb3a6> ;
    schema:mentions <ckg/mention/106db6bc-7ccb-53f0-9b42-9f36da92faa8>,
        <ckg/mention/215efb6e-368b-5abe-915f-51c7b362acd1>,
        <ckg/mention/51b15339-47c0-55b9-ad06-fabdb55c36f7>,
    schema:reviewBody "Posted in Fact Checks, Viral ContentTagged...tolerance policy"@en ;
    schema:reviewRating <ckg/rating/normalized/claimskg_TRUE>,
        <ckg/rating/original/truthorfiction_true_> ;
    schema:url <https://www.truthorfiction.com/does-it-cost-750-a-day-to-house-migrant-children-in-camps/> .
```

one of the supported formats *(TTL,n3, XML,nt, pretty-xml,trix, trig, and nquads).* We generate unique URI identifiers as UUIDs that are based on a one-way hash of key attributes for each instance. We present an exemplary claim in Figure 2.

5. Use-cases

The publication of the latest release of ClaimsKG facilitates the uncovering of several relations, patterns, and trends between the entities, claims, and their sources. The *SPARQL query endpoint*[12] allows the fetching of information about specific entities and also supports the execution of federated queries from external knowledge bases like DBpedia. Here we present a few of the use cases as an example of what could be achieved using ClaimsKG.

5.1. ClaimsKG and Corona Virus

The query in Figure 3(a) finds all claims mentioning *Coronavirus*. For each claim, it returns the text, the date, the rating, and the review URL by the fact-checking website. We can further drill down to finding only false claims about Coronavirus for a particular year by adding the "ratingValue=1" filter, as mentioned in Figure 3(b). This query returns, among many claims, the following one "Eating bananas is a preventative against the COVID-19 coronavirus disease." along with its date published *(2020-03-22)* and link to the original fact check webpage https://www.snopes.com/fact-check/bananas-coronavirus/. The result of this query shows how ClaimsKG can bring claims from different fact-checking sources about one topic in one place, which makes it easier and time-saving for the exploration of facts for a user. Searching for the same result manually would have been extremely laborious or unfeasible. A user would have to visit each website and search for claims related to a particular entity or topic (in this case *Coronavirus*), which might or might not be allowed for the source website, and would have to manually look for the claims belonging to a particular veracity label.

5.2. ClaimsKG and Historical Events

The query in Figure 4(a) is an example of fetching trends or important events in history. The query fetches all claims regarding the LGBT community and their corresponding year. The result of this query shows a sudden spike in the year 2018 from 2017 with a number of claims that rises from **21** to **75** (see Figure 4(b)). This spike can be attributed to the event of legalizing same-sex weddings by the Australian Parliament in Dec 2017, resulting in a sudden rise in claims regarding this topic. One could look for interesting entities like "Black lives matter", "Global Warming" and the "Great Recession" which would fetch us claims regarding these topics.

[12]SPARQL query endpoint: https://data.gesis.org/claimskg/sparql

```
PREFIX itsrdf:<https://www.w3.org/2005/11/its/rdf#>
PREFIX schema:<http://schema.org/>
PREFIX dbr:<http://dbpedia.org/resource/>
SELECT ?text ?date ?reviewurl ?rating
WHERE {
  ?claim a schema:CreativeWork ;
           schema:datePublished ?date.
  ?claim schema:text ?text ;
         schema:mentions ?entity .
  ?entity itsrdf:taIdentRef dbr:Coronavirus .
  ?claimReview schema:itemReviewed ?claim;
               schema:reviewRating ?rating;
               schema:url ?reviewurl .
     }
```

(a)

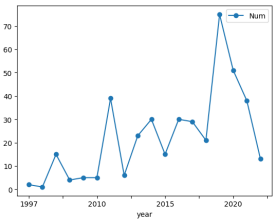
```
PREFIX itsrdf:<https://www.w3.org/2005/11/its/rdf#>
PREFIX schema:<http://schema.org/>
PREFIX dbr:<http://dbpedia.org/resource/>
SELECT ?text ?date ?reviewurl ?ratingName ?ratingValue
WHERE {
  ?claim a schema:CreativeWork ;
           schema:datePublished ?date FILTER(year(?date)=2020)
  ?claim schema:author ?author ;
         schema:text ?text ;
         schema:mentions ?entity .
  ?entity itsrdf:taIdentRef dbr:Coronavirus .
  ?claimReview schema:itemReviewed ?claim ;
               schema:url ?reviewurl .
  ?rating schema:author <http://data.gesis.org/claimskg/organization/claimskg>;
          schema:alternateName ?ratingName ;
          schema:ratingValue ?ratingValue FILTER (?ratingValue = 1) }
```

(b)

Figure 3: (a) SPARQL query to fetch all claims related to Coronavirus. (b) SPARQL query to fetch all false claims relating to Coronavirus for the year 2020.

```
PREFIX itsrdf:<https://www.w3.org/2005/11/its/rdf#>
PREFIX schema:<http://schema.org/>
PREFIX dbr:<http://dbpedia.org/resource/>
SELECT year(?date) as ?year count(?claim) as ?num
WHERE {
   ?claim a schema:ClaimReview ; schema:datePublished ?date.
   ?claim schema:mentions ?entity1 .
   ?entity1 itsrdf:taIdentRef ?entity2Uri .
   VALUES(?entity2Uri){
       (dbr:LGBT_community)(dbr:LGBT_movements)(dbr:LGBT_in_the_United_States)(dbr:LGBT_adoption)
       (dbr:LGBT_people_and_military_services)(dbr:LGBT_people_in_prison)(dbr:LGBT_parenting)
       (dbr:LGBT_rights_by_country_or_territory)(dbr:LGBTQ_Victory_Fund)(dbr:LGBT_culture)
       (dbr:LGBT_rights_opposition)(<http://dbpedia.org/resource/Rainbow_flag_(LGBT)>)(dbr:LGBT)
       (<http://dbpedia.org/resource/The_Advocate_(LGBT_magazine)>)(dbr:LGBT_rights_in_Russia) }
    } GROUP BY year(?date) ORDER BY year(?date)
```

(a)



(b)

Figure 4: (a) SPARQL query to fetch all claims regarding the LGBT community year-wise. (b) Spread of LGBTQ claims from 1996-2022.

```
PREFIX itsrdf: <https://www.w3.org/2005/11/its/rdf#>
PREFIX schema: <http://schema.org/>
PREFIX dbr: <http://dbpedia.org/resource/>
SELECT DISTINCT ?text ?reviewurl ?President  ?ratingName ?ratingValue
WHERE {
      SERVICE <http://dbpedia.org/sparql> {
      ?President <http://dbpedia.org/property/office> "President of the United States"@en .
      ?President <http://dbpedia.org/property/party> <http://dbpedia.org/resource/Democratic_Party_(United_States)> .}
      ?claim a schema:CreativeWork .
      ?claimReview schema:itemReviewed ?claim ;
                   schema:url ?reviewurl .
      ?rating schema:author <http://data.gesis.org/claimskg/organization/claimskg>;
              schema:alternateName ?ratingName ;
              schema:ratingValue ?ratingValue FILTER (?ratingValue = 1)
      ?claim schema:text ?text ;
             schema:mentions ?entity1 .
      ?entity1 <https://www.w3.org/2005/11/its/rdf#taIdentRef> ?President. }
```

Figure 5: SPARQL query to fetch all false claims regarding US Presidents of the democratic party

These queries could be particularly useful for social scientists, who are interested in studying specific phenomena at specific points in time and can find out which information (both true and false) was spreading online.

5.3. ClaimsKG and US Presidents

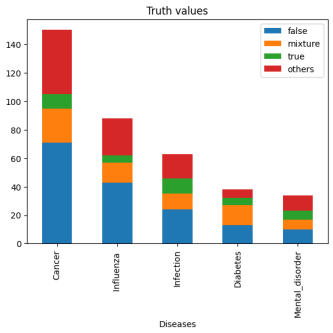
This example demonstrates the interesting ability of ClaimsKG to fetch information from external databases with the help of SPARQL queries. For example, the query in Figure 5 fetches false claims regarding all Presidents of the United States who belonged to the Democratic Party. As the reader can see, the information about which president belonged to the Democratic Party is given by DBpedia and used to explore ClaimsKG. The query outputs the claim text, the review URL, the name of the President as a DBpedia resource and the claim's rating. The results display claims such as "Barack Obama began his presidency 'with an apology tour'." or "Joe Biden calls Pennsylvania voters who don't support him 'chumps'." which were rated as false by the

```
PREFIX itsrdf:<https://www.w3.org/2005/11/its/rdf#>
PREFIX schema:<http://schema.org/>
PREFIX dbr:<http://dbpedia.org/resource/>

SELECT DISTINCT ?text ?reviewurl ?c
WHERE {
      SERVICE <http://dbpedia.org/sparql>{
          ?b <http://dbpedia.org/ontology/disease> ?c
      }
      ?claim a schema:CreativeWork.
      ?claimReview schema:itemReviewed ?claim ;
                   schema:url ?reviewurl .
      ?claim schema:author ?author ; schema:text ?text ; schema:mentions ?entity1 .
      ?entity1 <https://www.w3.org/2005/11/its/rdf#taIdentRef> ?c.

      }
```

(a)



(b)

Figure 6: (a) SPARQL query to fetch all claims regarding diseases mentioned in ClaimsKG. (b) Mention of disease entities and their veracity labels.

fact-checkers published in the fact-checking website *PolitiFact*.

5.4. ClaimsKG and Diseases

In recent years, the spread of diseases and the number of claims regarding them have been on the rise. We tried to have a closer look at this topic. Figure 6(a) fetches claims that contain mentions of any disease in ClaimsKG. The query outputs 806 results along with the source review URL and the disease entity that was mentioned. Disease entities were retrieved from DBpedia. From the results, we filter out the top 5 diseases that were mentioned in these claims which show that *Cancer* was the highest (Table 3). Apart from the diseases mentioned in this table, there were also other diseases like Plague, Strabismus Palliative care, and others that were sparsely mentioned in the Knowledge Graph. After analysis of the veracity labels, we plot a graph of the entities and their truth values in Figure 6(b). The plot shows that *Cancer* is the most discussed and mentioned disease, and almost 47% of the claims made about cancer are false. These included claims like "Drinking hot coconut water kills cancer cells" or "An association of pediatricians "admitted" that HPV vaccine Gardasil causes ovarian "failure" or cancer." There are very few true claims, only 0.06% of the total which included claims like "Four in ten cancer patients lose their life savings after starting treatment." which is highly alarming. There are mentions of other diseases like influenza, infection, diabetes, and mental disorder that are predominantly present in the Knowledge Graph. On a general inspection, it is observable that the number of false claims is always more than claims labeled as *TRUE* or *MIXTURE* for any kind of disease. There is also a considerable percentage of claims that are rated as *OTHER* which did not have any clear rating value associated with them. For example, the claim "Asparagus has miraculous cancer-fighting properties." published in Snopes has a rating of "Unproven" which is clearly neither true nor false. Likewise, *OTHER* claims included claims with original ratings such as "research in progress", "outdated", etc, and sometimes no rating attached to the claim.

Table 3
Top 5 disease entities that are mentioned in ClaimsKG.

Diseases	Counts
http://dbpedia.org/resource/Cancer	150
http://dbpedia.org/resource/Influenza	88
http://dbpedia.org/resource/Infection	63
http://dbpedia.org/resource/Diabetes	38
http://dbpedia.org/resource/Mental_disorder	34

6. Conclusion and Future Work

We present the latest release of ClaimsKG, a knowledge graph of fact-checked claims which enables structured queries about their related metadata. We describe the pipeline changes made in this version and provide detailed statistics of the data. We also demonstrate use cases to show how ClaimsKG can be used for data analysis for various social-science-related research questions. We observe that the NERD tool fails to recognize some entities. This gives us the scope for further improvements. In the future, we aim to enhance the entity annotation capabilities with more focus on the social science domain. We plan to include more sources with diverse languages and perform multilingual entity linking and disambiguation for enhancing the quality of ClaimsKG. We also intend to analyze how the same or similar claims are covered across different sources and the degree of agreement between the fact-checking websites. ClaimsKG will foster new discussions about topics related to specific domains and support users in the exploration of online truthfulness about specific facts.

References

[1] S. Vosoughi, Roy, et. al., The spread of true and false news online, science 359 (2018) 1146–1151.

[2] N. Hassan, B. Adair, J. T. Hamilton, et. al., The quest to automate fact-checking, in: Proceedings of the 2015 computation+ journalism symposium, Citeseer, 2015.

[3] S. Tschiatschek, A. Singla, et al., Fake news detection in social networks via crowd signals, in: Companion proceedings of the web conference 2018, 2018, pp. 517–524.

[4] G. Pennycook, Z. Epstein, M. Mosleh, et. al., Understanding and reducing the spread of misinformation online, ACR North American Advances (2020).

[5] A. Tchechmedjiev, P. Fafalios, K. Boland, et. al., Claimskg: A knowledge graph of fact-checked claims, in: The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, 2019, Proceedings, Part II 18, Springer, 2019, pp. 309–324.

[6] J. Liu, C. Wang, et. al., Dtn: Deep triple network for topic specific fake news detection, Journal of Web Semantics 70 (2021) 100646.

[7] A. Brasoveanu, R. Andonie, Semantic fake news detection: A machine learning perspective, in: 15th International Work-Conference on Artificial Neural Networks, 2019, pp. 656–667.

[8] P. Smeros, et. al., Sciclops: Detecting and contextualizing scientific claims for assisting manual fact-checking, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 1692–1702.

[9] A. L. Opdahl, et. al., Semantic knowledge graphs for the news: A review, ACM Computing Surveys 55 (2022) 1–38.

[10] S. Shaar, et. al., That is a known lie: Detecting previously fact-checked claims, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 3607–3618.

[11] M. Gasquet, D. Brechtel, M. Zloch, A. Tchechmedjiev, K. Boland, P. Fafalios, S. Dietze, K. Todorov, Exploring fact-checked claims and their descriptive statistics, in: ISWC 2019 Satellite Tracks-18th International Semantic Web Conference, 2019.