ClaimsKG: A Live Knowledge Graph of Fact-Checked Claims

Andon Tchechmedjiev[1], Pavlos Fafalios[2], Katarina Boland[3], Stefan Dietze[3,4], Benjamin Zapilko[3], Konstantin Todorov[5]

[1] LGI2P, IMT Mines-Ales, France
[2] L3S Research Center, Leibniz University of Hanover, Germany
[3] GESIS - Leibniz Institute for the Social Sciences, Germany
[4] Heinrich-Heine-University Düsseldorf, Germany
[5]LIRMM / University of Montpellier / CNRS, France
andon.tchechmedjiev@mines-ales.fr, fafalios@L3S.de, {stefan.dietze, benjamin.zapilko, katarina.boland}@gesis.org, konstantin.todorov@lirmm.fr

Abstract. Various research areas at the intersection of computer and social sciences require a ground truth of contextualized claims labelled with their truth values in order to facilitate supervision, validation or reproducibility of approaches dealing, for example, with fact-checking or analysis of societal debates. So far, no reasonably large, up-to-date and queryable corpus of structured information about claims and related metadata is publicly available. In an attempt to fill this gap, we introduce ClaimsKG, a live knowledge graph of fact-checked claims, which facilitates structured queries about their truth values, authors, dates, journalistic reviews and other kinds of metadata. ClaimsKG is generated through a semi-automated pipeline, which harvests data from popular fact-checking websites on a regular basis, annotates claims with related entities from DBpedia, and lifts the data to RDF using an RDF/S model that makes use of established vocabularies. In order to harmonise data originating from diverse fact-checking sites, we introduce normalised ratings as well as a simple claims coreference resolution strategy. The current knowledge graph, extensible to new information, consists of 28,383 claims published since 1996, amounting to 6,606,032 triples.

Keywords: Claims; Fact-checking; Societal debates; Knowledge Graphs

1 Introduction

The spread of controversies, biased discourse and falsehoods on the Web has become an increasingly important issue, from both a societal as well as a research perspective [1, 2]. Recently, a wide range of interdisciplinary research directions are being explored in this broad area, which often rely on a ground truth of labelled claims. Such works include investigations into the spreading pattern of false claims on Twitter [1], pipelines for discovering and finding the stance of claim-relevant Web documents [3], approaches for classifying sources of news, such as Web pages, PLDs, users or posts [4], or research into fake news detection [5] and automatic fact-checking [6]. In all these cases, the availability of a labelled ground truth, consisting of claims, their corresponding metadata and, in particular, their truth values (or ratings) supported by one or more people or organizations is essential in order to enable supervision of machine learning methods, reproduction and explainability of the results, and to facilitate fair evaluation and follow-up work. In addition, as documented by the aforementioned works, claims are usually not considered in isolation, but in a context. Thus, reproducing such research requires not only archiving claims and their truth values, but also their related documents, such as journalistic claim reviews, the associated entities and time-frames that can be linked to particular events, accounting in that way for the continuous evolution of Web content.

To our knowledge, no reasonably large and up-to-date corpus of structured information about claims and their context has been made publicly available. We attempt to fill this gap by introducing ClaimsKG, a live knowledge graph (KG) of fact-checked claims, which facilitates structured queries about their truth values and other kinds of metadata, constructed and published following the W3C recommendations and best practices. We define a claim as a *verifiable statement supported by one or more people or organizations in a particular point in time.* ClaimsKG is generated through a semi-automated pipeline, which periodically harvests data from popular fact-checking websites. The *claims* and their *reviews* (articles written by fact-checkers that accompany a claim and explain its context and veracity judgement) are annotated with related *entities* from DBpedia, and all data are lifted into RDF using a dedicated RDF/S model (dubbed *Claims*), which is based on established vocabularies such as schema.org and NIF. In order to harmonise data originating from diverse fact-checking sites, we introduce a normalised truth ratings scheme, as well as a simple claim matching strategy. ClaimsKG enables advanced exploration and information discovery, e.g., via queries such as *"find all false claims by D. Trump in 2017 that also mention FBI"*, or *"find top 5 politicians per month involved in false claims"*, as well as exploitation of data from various sources via federated SPARQL queries, e.g., *"retrieve all claims mentioning journalists"*. In addition, we provide a Web interface for exploration and search of the graph enabling users from outside of the computer science community to retrieve information or sample data from our resource. The dataset, as of April 2019, consists of 28,383 claims published since 1996, amounting to 6,606,032 triples in our KG.

In summary, we provide (1) the *Claims* data model for representing fact-checked claims and associated information, (2) an open-source pipeline for crawling and extracting data from fact-checking websites and for lifting these data following the *Claims* model, (3) an openly available dynamic large-scale KG of claims and associated metadata and (4) a Web interface for search and exploration of the resource. In the following section, we provide general information about the resource and links for access. We detail on the KG generation process in Section 3. We introduce our *Claims* model in Section 4, while use-cases and queries are discussed in Section 5 along with an overview of the exploratory user interface. We review related work in Section 6 before concluding.

2 ClaimsKG in a Nutshell

ClaimsKG consists of data extracted from a number of fact-checking websites—claims, dates, authors, journalistic or fact-checkers' reviews of the claims, verdicts (or ratings)—where claims are annotated with DBpedia entities mentioned in the claim text or the bodies of their reviews. To select fact-checking domains, we relied on the International Fact-Checking Network's (IFCN) signatories list,[1] admitting only sources considered by the fact-checking community as highly reputable. At this stage, we only consider information in English from six sources: *africacheck.org, checkyourfact.com, factscan.ca, politifact.com, snopes.com, truthorfiction.com*. Note that ClaimsKG is extensible to new websites, however, the information extraction process may vary from one website to anther due to structural specificities of the sources (cf. Section 3). Key facts and links related to ClaimsKG are given in Table 1. The KG is currently accessible from a Virtuoso triplestore with a SPARQL endpoint and downloadable as a Zenodo dump. All represented entities (claims, reviews, etc.) are assigned resolvable identifiers following the W3C best practices (see Section 3 for an example). The dataset has a DCAT description and is released for free distribution under a Creative Commons Attribution 4.0 licence.[2] The graph can be also accessed through ClaimsKG's official page, which displays detailed up-to-date statistics and a set of example SPARQL queries. All tools developed for the KG's creation are made available as open source on GitHub. Table 2 offers some general and per-source coverage statistics for the data, in particular the coverage of key properties. In order to account for emerging claims, the dataset is updated monthly. Note that manual interaction is required between the different steps of our data life-cycle, described in the next section.

Table 1: Key links to ClaimsKG data and tools.

ClaimsKG website	`https://data.gesis.org/claimskg/site`
Dataset DOI	`https://doi.org/10.5281/zenodo.2628745`
DCAT description	Included in the KG
Zenodo dump	`https://zenodo.org/record/2628745`
SPARQL endpoint	`https://data.gesis.org/claimskg/sparql`
The *Claims* ontology	`https://data.gesis.org/claimskg/site/#model`
Exploratory interface	`https://data.gesis.org/claimskg/explorer`
ClimsKG pipeline source code	`https://github.com/claimskg`
TagMe entity annotation tool	`https://tagme.d4science.org/tagme/`

[1] `https://ifcncodeofprinciples.poynter.org/signatories`
[2] `https://creativecommons.org/licenses/by/4.0/`

Table 2: Claim metadata coverage and statistics (as of April 2019)

Property \Web source	Global	Snopes	Politifact	AfricaCheck	TruthOrFiction	CheckYorFact	FactScan
Number of claims	28,383	10,685	15,743	560	778	492	125
Claim text	100%	100%	100%	100%	100%	100%	100%
Claim author	93.55%	100%	100%	0.0%	0.0%	0.0%	100%
Claim date published	92.13%	96.25%	100%	0.0%	0.0%	0.0%	98.4%
Claim with references (≥1)	86.47%	99.82%	75.85%	96.96%	100%	99.59%	100%
Claim with keywords (≥1)	93.78%	95.40%	100%	99.46%	0%	0.0%	100%
Claim with entities (≥1)	99.71%	99.89%	100%	98.21%	99.61%	92.88%	100%
Claim review URL	100%	100%	100%	100%	100%	100%	100%
Claim review title	100%	100%	100%	100%	100%	100%	100%
Claim review author	100%	100%	100%	100%	100%	100%	100%
Claim review date published	100%	100%	100%	100%	0%	100%	100%
Claim review language	100%	100%	100%	100%	100%	100%	100%
Claim review with entities (≥1)	66.94%	74.14%	58.70%	80.17%	76.96%	96.95%	96.0%
Claim rating	100%	100%	100%	100%	100%	100%	100%
Exact claim matches	87	38	49	0	0	0	0
True claims	3,725	1,311	2,255	60	97	0	2
False claims	11,068	6,002	4,663	209	191	0	3
Mixture claims	10,420	1,798	8,564	0	56	2	0
Other claims	3,170	1,574	261	291	434	490	120

3 Generating ClaimsKG

ClaimsKG is built through a pipeline, which periodically crawls popular fact-checking sites, normalises ratings and entity mentions, reconciles identical claims and lifts the data onto the specifically developed *Claims* model, described in Section 4. Hereafter, we detail the technical steps of the pipeline, summarised in Figure 1. Links to its open-source components are given in Table 1.

Extracting claims and metadata. The *Claims extractor* crawls the identified fact-checking domains and collects the information in a JSON or Microdata format to consolidate a large multi-sourced data set (as a CSV file). The collected data consist mainly of: (a) the textual statement of the claim; (b) its truth-value or rating (both the normalised and the original one); (c) a link to the claim review from the fact-checking website; (d) the references cited in the claim reviews; (e) the entities extracted from the claim body and from the review body; (f) the author of the claim and the author of the review; (g) the date of publication of the claim and that of the review; (h) the title of the review article; (i) a set of keywords extracted from the websites acting as topics (e.g. "abortion").

Note that the extraction process is tailored individually to the structure of each of the different fact-checking websites, resulting in a set of website-specific extractors. The statistics generated at each run of the pipeline (globally and per domain) allow to monitor the "health" of the extracted data by detecting potential issues that may be related to changes of the structures of the respective fact-checking websites that may have occurred between two runs of the pipeline.

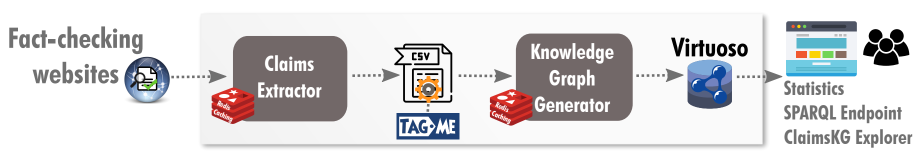


Fig. 1: Overall architecture of the ClaimsKG pipeline.

Entity annotation of the claims. As part of data processing, we annotate the entities (e.g., names of persons, organisations, locations, etc.), mentioned in the texts of the claims and their reviews, using the TagMe tool [7]. TagMe allows to automatically identify entities in a text and link each of them to a Wikipedia page and a DBpedia URI. It is known to achieve particularly good results when annotating short texts, which is the case for the statements in this domain, although we also annotate the body of the claim reviews. We run a local version of TagMe, allowing to update the database regularly, using the latest available dump of Wikipedia (October 2018). We performed all the annotations using the optimal parameters described in [7]. We evaluated our updated TagMe model on the YAGO CONLL-TestB ground truth dataset [8] and obtained an accuracy of 75.5%, which is in line with state of the art performance reported for TagMe.[3]

Normalization of ratings. Each of the fact-checking websites has its own labels describing the truthfulness of the claims, with different discrete textual values of ratings, sometimes also matched to numerical values. While some sites have a controlled vocabulary of possible truth values, others apply an open-ended rating schema. For example, *Truth or Fiction* has a large number of non-uniform labels, such as "truth & misleading" or "reported as fiction". In order to harmonize our dataset, alongside the original ratings, we also provide a normalized rating score, applied across all claims contained in the dataset. For each of the sources, we have summarised the distribution of rating values and have then assigned them to a conservative and coarser-grained set of labels that correspond to the least common denominator between all the classifications of the individual sites. Given the varied rating schemes, where individual labels often are hard to objectively apply or interpret, we opted for a simple rating scheme consisting of four basic categories,[4] that can be mapped in a consensual way to all existing rating schemes: `TRUE`, `FALSE`, `MIXTURE`, `OTHER`. The two extreme cases of a claim being proven true or false are captured by `TRUE` and `FALSE`, while `MIXTURE` characterises something on a truth scale or that holds both a degree of truth and a degree of falsehood. For anything that does not fall into this spectrum, we chose `OTHER` as a fallback. While the `TRUE`/`FALSE` ratings are straightforward, `MIXTURE` conflates a very large number of possible truth values, as diverse as "downplayed" or "mostly true". For `OTHER`, we have rating names such as "half-flip", "scam" or "research in progress".

Lifting and Serialisation. We created a Python 3.6 script to read the extracted claims as a CSV file in the extraction step, and then create the corresponding KG following the data model described in Section 4. We used the `rdflib` library to create the model and an abstract RDF graph to then serialize it in one of the many supported formats. All the caching needs of the generation process are met with a Redis server. We generate unique URI identifiers as UUIDs based on a one-way hash of key attributes for each instance. For example, for a particular claim in our KG,[5] we have: `http://data.gesis.org/claimskg/`

[3] `http://nlpprogress.com/english/entity_linking.html`

[4] We provide full correspondence tables here: `https://goo.gl/Ykus98`.

[5] `https://www.snopes.com/fact-check/was-megyn-kelly-fired-from-nbc/`

creative_work/5f7e8c65-3d8b-57da-bab9-eb3a373bd2ab that can be looked up in Virtuoso. The triplification package is made available under an open-source licence on GitHub along with documentation and usage examples and will be updated regularly with the latest improvements (link given in Table 1).

Handling simple claim coreferences. A certain number of identical claims is present within the websites, published at different dates, with possibly varying reviews (e.g., the same claim published at a later date than the original publication will have an updated review). For that reason, instead of fusing these claims, we have opted for establishing `owl:sameAs` links among them. We have implemented a simple approach to identify these claims, which aims to ensure 100% precision of the discovered links (exact matches). We exploit the text of the claims and the text of the claim titles only, which are normalised (lowercase; remove all " and ' characters and certain stop-words, such as "said", "claimed", etc) and then apply an identity string similarity measure on these texts. This resulted in 38 `owl:sameAs` links on claims from Snopes and 49 from Politifact.

4 The Claims Data Model

Our data model, depicted in Fig. 2, exploits terms from established vocabularies, specifically schema.org,[6] NLP Interchange Format (NIF),[7] and Internationalization Tag Set (ITS).[8] The selection of the vocabularies was based on the following objectives: i) avoiding schema violations, ii) enabling data interoperability through term reuse, iii) having stable identifiers, persistent hosting and open license, iv) being supported by a community, v) being extensible (ability to easily extend ClaimsKG with more data).

The core elements of our model are the *claim* and the *claim review*. To represent them, we make use of schema.org. Following Google's suggestion for Web markup of claims,[9] a claim is of type `schema:CreativeWork` and a claim review of type `schema:ClaimReview`. An instance of `schema:ClaimReview` is connected to an instance of `schema:CreativeWork` through the property `schema:itemReviewed`. A claim (instance of `schema:CreativeWork`) is associated with the actual text of the claim, a date (when the claim was uttered), an author (who uttered the claim), as well as with keywords (tags acting as topics) and one or more citations (URLs of resources, e.g., a tweet or a video). Since there might be many instances of the same claim coming from the same or different fact-checking sites, two claims can be connected through a `owl:sameAs` property.

A claim review is associated with metadata, in particular its author, its publication date, its language, its URL (pointing to the full text of the review), its full text (optionally, according to copyright restrictions), as well as with a title and one or more truthfulness assessments (ratings). The assessment is of

[6] https://schema.org/
[7] http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core
[8] https://www.w3.org/TR/its20/
[9] https://developers.google.com/search/docs/data-types/factcheck

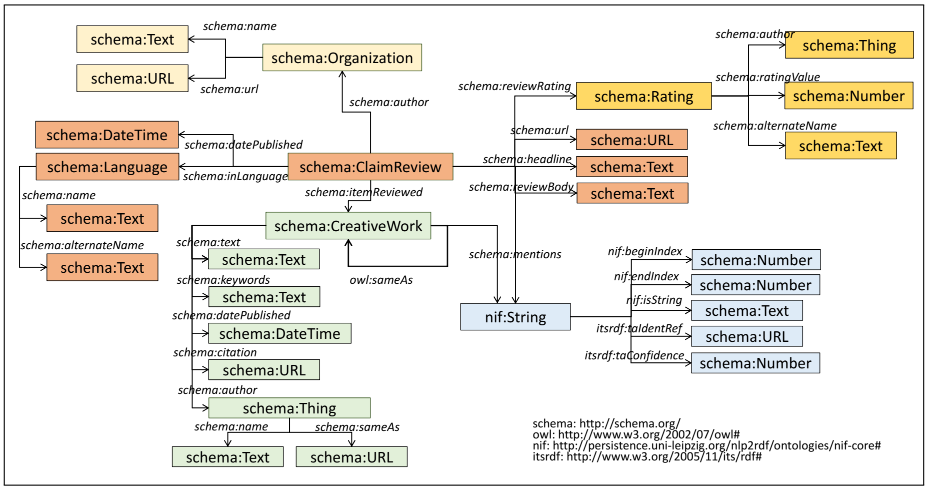


Fig. 2: The *Claims* data model.

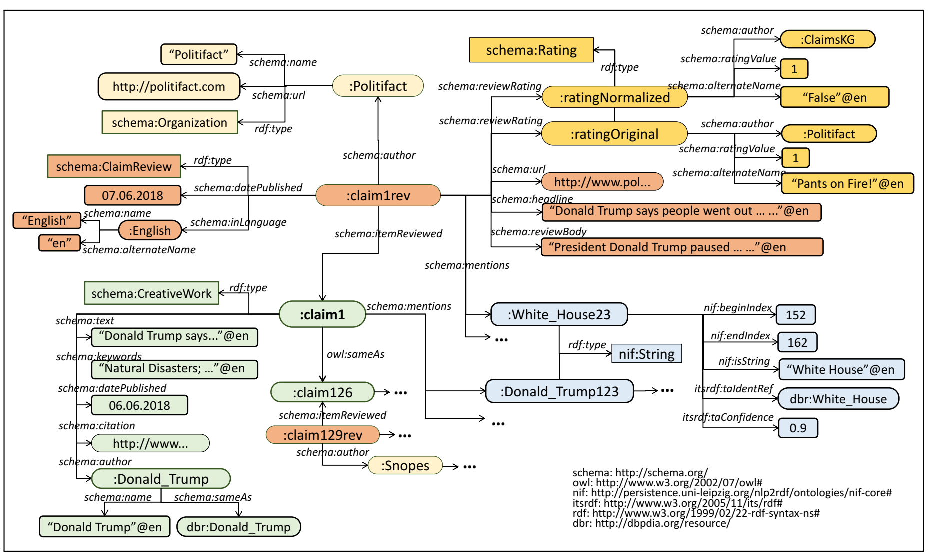


Fig. 3: Instantiation of the *Claims* model for a claim sourced from Politifact made by Donald Trump on June 6, 2018.

type `schema:Rating` and is connected with the review through the property `schema:reviewRating`. A rating is represented through three properties: author (who provids the rating), rating value (a number in a pre-specified range, e.g., 1 to 5), and alternate name (a textual label of the rating value, e.g., "false"). The *author* property allows to provide more than one ratings for the same instance of a claim review.

Both a claim and a claim review can be associated with one or more *entity mentions*, i.e., names of entities mentioned in the short text of the claim or in the claim review. To describe this information, we make use of the NIF and ITS vocabularies, which provide classes and properties to describe the result of natural language processing tools applied on texts or documents. An instance of an entity mention is described through five properties: `nif:beginIndex` (starting position in the text), `nif:endIndex` (ending position in the text), `nif:isString` (a word or sequence of words representing the entity), `itsrdf:taIdentRef` (the identity of the mentioned entity), and `itsrdf:taConfidence` (the confidence that the entity has been disambiguated correctly). Depending on the specific requirements with respect to precision and recall, data consumers can select suitable confidence ranges to consider when querying the data.

Figure 3 depicts an example of a claim review by Politifact for a claim made by Donald Trump on June 6, 2018.[10] We notice that there are two instances of `schema:Rating`, one for the original rating by Politifact and one for the normalized rating provided by our KG. Apart from metadata information, we also see that the review mentions the entity name "White House", which probably (with confidence 0.9/1.0) corresponds to the official residence of the US President.

5 Use Cases and Exploitation

Use-cases and queries. The publication of structured data about a large collection of claims allows to uncover explicit and implicit relations between claims, entities, and sources. A number of existing fact-checking applications rely on linking claims to fact-checked statements in a database (e.g. `https://fullfact.org/automated/`, [9, 10]). By combining claims and ratings from multiple portals and providing a unified structure, ClaimsKG facilitates these efforts. Moreover, the data can be used to enable supervision of machine learning models to support the advancement of automatic fact-checking algorithms [4,11]. In addition to supplying a large number of claims, ClaimsKG enables advanced entity-centric search and exploration and information discovery exploiting data from various sources via federated SPARQL queries, which allow, for example, to query entities belonging to a specific group (e.g. politicians or journalists) and create complex queries using both metadata from claims and extracted entities. We provide different examples on our website (see Table 1 for a link). By exploiting the claim metadata and extracted entities, we can run complex queries that combine different types of information. The query in Fig. 4 requests

[10] `http://www.politifact.com/texas/statements/2018/jun/07/donald-trump/donald-trump-says-people-went-out-their-boats-watc/`

all *false* claims of *2017* mentioning *Donald Trump* and *Climate change*. For each claim, the query returns its *text*, *date*, as well as the URL of its *review* by a fact-checking site. The query returns the following claim: *"Donald Trump signed an executive order naming climate change as a threat 'both to the economy and national security."'* (2017-02-01). In a similar way, we can generate a sample of claims based on certain criteria and use it in other tasks, e.g., for evaluation or training by automated fact checking approaches [4, 11]. Such a sample can be easily produced through a concise SPARQL query over ClaimsKG.

```
SELECT ?text ?date ?reviewurl WHERE {
 ?claim a schema:CreativeWork ; schema:datePublished ?date FILTER(year(?date)=2017)
 ?claim schema:author ?author ; schema:text ?text ; schema:mentions ?entity1, ?entity2 .
 ?entity1 itsrdf:taIdentRef dbr:Climate_change .
 ?entity2 itsrdf:taIdentRef dbr:Donald_Trump .
 ?claimReview schema:itemReviewed ?claim ; schema:reviewRating ?rating ; schema:url ?reviewurl .
 ?rating schema:author <http://data.gesis.org/claimskg/organization/claimskg> ;
         schema:alternateName ?ratingName ;
         schema:ratingValue ?ratingValue FILTER (?ratingValue = 1) }
```

Fig. 4: SPARQL query requesting *false* claims of 2017 mentioning both *Donald Trump* and *Climate change*.

Going beyond the computer science domain, ClaimsKG can be a valuable resource supporting (computational) social scientific research investigating, for example, societal debates and agenda-setting. Agenda-setting theory refers to the influence of mass media on the public's focus of attention [12]. While first-level agenda-setting relates to inserting topics, events or entities into the public discourse thereby regulating societal priorities, second-level agenda-setting is about increasing the salience of specific features or attributes of entities in the discourse. This is also referred to as frame-setting. With the web evolving into a platform where every citizen may become a publisher, express their views and reach out to a large audience, citizens are now able to play a more active role in influencing the public discourse [13]. Online debates about political issues typically exhibit the pattern of a few dominant ideological positions emerging, with different groups expressing different viewpoints and often referring to a disparate set of information sources [14] which, in turn, may focus on different attributes and frames for a given topic. Using ClaimsKG, an exploratory search on a topic and related entities may be performed in order to gain insights on relevant viewpoints, attributes and actors. Also, the KG allows tracking differences over time and in relation to specific events and relating views of specific actors to ideological positions. To illustrate, consider the 2012 incident of the neighbourhood watch coordinator George Zimmerman shooting 17-year-old African-American high school student Trayvon Martin, the incident that later gave rise to the Black Lives Matter movement [15]. The query given in Fig. 5 retrieves all claims mentioning Trayvon Martin or George Zimmerman, yielding 68 claims in total with 8 claims rated true, 33 false, and 24 mixture. The distribution of truth values hints at this being a highly controversial topic with potentially highly po-

```
SELECT ?text ?reviewurl ?rating WHERE {
 ?claim a schema:CreativeWork ; schema:text ?text ; schema:mentions ?entity1 .
 ?entity1 itsrdf:taIdentRef ?entity2Uri
 FILTER (?entity2Uri IN  (dbr:Trayvon_Martin, dbr:George_Zimmerman))
 ?claimReview schema:itemReviewed ?claim ; schema:reviewRating ?rating ; schema:url ?reviewurl .
}
```

Fig. 5: SPARQL query requesting all claims mentioning *Trayvon Martin or George Zimmerman.*

larized viewpoints. Central to the debate is the aspect of racism; some framing the incident as an example of racist violence against black people,[11] some seeing race as an overemphasized point in the Zimmerman trial [12] and others framing the Black Lives Matter debate as racist against white people.[13] An attribute frequently mentioned in these claims is the "stand your ground" law. The query in Fig. 6 retrieves the top-10 entities connected to it revealing a strong association to the Trayvon Martin case.

```
SELECT ?entity2Uri WHERE {
 ?claim a schema:CreativeWork ; schema:text ?text ; schema:mentions ?entity1, ?entity2 .
 ?entity1 itsrdf:taIdentRef dbr:Stand-your-ground_law .
 ?entity2 itsrdf:taIdentRef ?entity2Uri FILTER (?entity2Uri != dbr:Stand-your-ground_law)
} GROUP BY ?entity2Uri LIMIT 10
```

Fig. 6: SPARQL query requesting the top-10 entities mentioned in claims together with *Stand Your Ground law.*

The query in Fig. 7 illustrates the tracking of changes over time and discovering important events. While for 2015, 2017 and 2019 three mentions per year are found and one for 2018, there is a striking peak in 2016 with 17 mentions. This aligns with the incident of law enforcement officers being shot during a Black Lives Matter protest march in July 2016 which reopened a heated debate about the movement. In fact, analysis of the respective claims reveals the emergence of frames attributing violent and disruptive behaviour to the Black Lives Matter movement.[14]

Note that the discussed scenarios represent only starting points for further analyses. Early announcements of the KG via mailing lists have attracted interest by researchers working on truth discovery.

[11] https://www.politifact.com/florida/statements/2013/jul/24/jesse-jackson/homicides-blacks-have-tripled-stand-your-ground-wa/

[12] https://www.politifact.com/florida/statements/2013/jul/17/tweets/look-statistic-blacks-and-murder/

[13] https://www.snopes.com/fact-check/keith-passmore-murder/

[14] https://www.politifact.com/wisconsin/statements/2016/dec/02/sean-duffy/donald-trump-backer-sean-duffy-links-attacks-polic/, https://www.politifact.com/wisconsin/statements/2017/apr/17/sheriff-david-clarke-us-senate/pro-sheriff-david-clarke-group-says-clarke-called-/

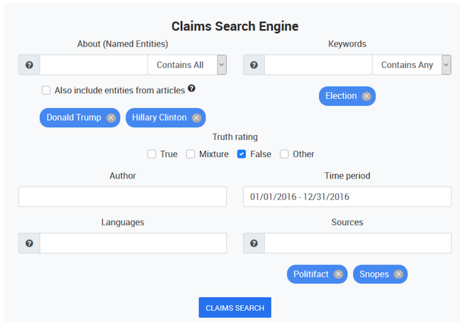
```
SELECT year(?date) as ?year count(?claim) as ?num WHERE {
 ?claim a schema:CreativeWork ; schema:datePublished ?date FILTER(year(?date)>=2012)
 ?claim schema:author ?author ; schema:text ?text ; schema:mentions ?entity .
 ?entity itsrdf:taIdentRef dbr:Black_Lives_Matter } GROUP BY year(?date) ORDER BY year(?date)
```

Fig. 7: SPARQL query requesting the number of claims mentioning *Black Lives Matter* by year.

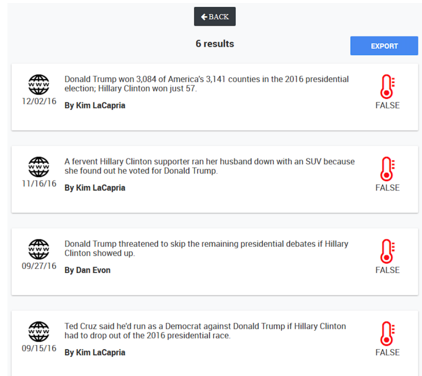
ClaimsKG Explorer. In order to facilitate data access for researchers from outside of the computer science domain, journalists and other interested users, we provide ClaimsKG Explorer (link in Table 1), a user-friendly, Web-based interface to query and interact with ClaimsKG, e.g. to conduct exploratory search over the graph. The application sends HTTP requests to the ClaimsKG SPARQL endpoint and provides information through a Web user interface (UI). The user is given the possibility to filter his/her search with respect to a number of facets, such as a set of entities contained (in conjunction or disjunction) in the text of the claim or its review (by default the application looks for entities contained in both, but the user is given the possibility to restrict the search to entities only contained in the claim body), a number of keywords, as well as the normalized claim verdict (true, false, mixture or other), as shown in the screenshots in Fig. 8a. In addition, the user is allowed to filter the retrieved claims according to the timeframe of interest, the authors, the sources (fact-checking websites) or the language (currently unavailable, but work in progress will incorporate non-English websites). After clicking on CLAIMS SEARCH, the user is provided a list of clickable claims ordered by their date of publication (most recent ones on top), as shown in Fig. 8b. The search result corresponding to the selected criteria can be exported as a csv or an rdf file and reused in a particular scenario by clicking on the EXPORT button on the results page. Clicking on a particular claim allows to access its webpage on the fact-checking website that hosts it, as well as retrieve the list of related (mentioned) entities, references and additional information.

6 Related Work

As outlined above, ground truth data in the form of labelled (and contextualised) claims is necessary for a number of interdisciplinary research problems. Among the most prominent use-case scenarios is provided by the task of automatic fact-checking, which has been of growing interest for the AI community. A number of approaches have been proposed to extract check-worthy pieces of information that can be qualified as claims [6] and to further assess their veracity automatically [16]. The majority of these approaches can be classified either as *reference* or *machine learning approaches*. The former model claims computationally to achieve structured representations [9, 10, 17–19], allowing for their comparison to certified facts contained in knowledge bases or for the application of data/graph mining techniques on the graphs. The latter rely on data in the



(a) The Explorer's search engine.



(b) Results

Fig. 8: The ClaimsKG Explorer user interface

form of labelled claims in order to train and apply machine learning models. The current section provides an overview of datasets for training and/or evaluation covering these two families of approaches. We categorise these datasets according to the process of their collection.

Extracting gold standard from web sources. Alongside ClaimsKG, a number of approaches rely on extracting data form fact-checking websites, allowing to collect large amounts of claims together with veracity annotations of high quality that stem from serious journalistic work. Among the first initiatives in that field is introduced in [20]. The data extraction process is manual, resulting in merely 221 statements collected from Channel 4 and PolitiFact. Since the two sources do not apply the same rating scale, similarly to our approach, the authors map the two respective sets of labels to a common scale consisting of five categories: "true", "mostly true", "half true", "mostly false" and "false". The *Liar* benchmark [21] collects 12.8K Politifact claims over 10 years. Additional information is stored regarding the *speaker*, the *context*, the *label* and the *justification*. The fact-checking approach presented in [22] relies on a data set of approximately 10K claims, also crawled from Politifact in 2016, while [23] and [4] have collected, respectively, 1K and 5K claims from Snopes. Wikipedia's lists of proven hoaxes[15] and fictitious people[16] have been used in [4, 24] in order to generate ground truth labels, resulting in 157 claims labelled as "fake". In parallel, the authors also collect around 4.8K labelled claims from Snopes, published before Faburary 2016. The Clef-2018 challenge includes the fact-checking task *Check that!* [25]. The benchmark data (150 claims) consists of sentences collected from debates from the 2016 US Presidential Election Campaign, as well as from other political speeches during and after the campaign. Evaluations are obtained from FactCheck.org articles resulting in labels of the kind "factually true", "half-true", or "false".

---

[15] `https://en.wikipedia.org/wiki/Listofhoaxes#Provenhoaxes`
[16] `https://en.wikipedia.org/wiki/List_of_fictitious_people`

The *Emergent* data set[17] results from collecting claims from various web sources, such as Politifact, Snopes and Twitter accounts such as @Hoaxalizer, particularly dedicated to rumors and hoaxes [26]. The data set contains 300 claims with three types of annotations ("true", "false", "unverified") provided by journalists. A large scale collection of tweet news (126K stories) labelled on the basis of significant degree of agreement among several fact-checking sites (Snopes, Politifact, Factcheck, Truth-or-Fiction, Hoax-slayer and Urbanlegends) is used in [16]. As a result, a sample of news stories, assigned one of three possible labels: "true", "false" or "mixed", is created. However, in addition to the news labelling process not being transparent, the data is only available upon request for the purposes of reproducing the reported experiments.

Crowdsourced or manually annotated data sets. Crowdsourcing techniques allow for the extraction and labelling of relatively large sets of claims, but the resulting assessments are presumably less trustworthy as compared to those provided by fact-checking organisations. The Open Domain Deception Dataset contains "freely contributed truths and lies" [27]. By using the Amazon Mechanical Turk, each worker has been asked to freely formulate 7 one-sentence truths and lies. After cleansing the collection, the final dataset consists of 7.2K sentences provided by 512 unique contributors, for whom demographic data is also collected and made available. The SemEval'17 challenge dataset contains 5.5K crowdsourced annotated claims [28], while the FEVER dataset, with 185K entries, extracts claims from Wikipedia. Semantics-preserving sentence altering techniques are then applied and the resulting claims are annotated by the crowd [29]. Finally, the approach in [30] relies on a dataset of 250 manually annotated claims.

Automatic annotations. Fact verification methods can be applied in order to construct automatically ground truth datasets for fact-checking. An example is the fake-news dataset,[18] containing text and metadata scraped from 244 websites that have been identified as untrustworthy by the BS Detector Chrome Extension tool.[19] This is naturally a much less reliable approach, as compared to the ones cited above, since the imperfections of the verification system are propagated onto the annotated data.

Knowledge graphs for fact-checking. Finally, we focus on KGs that have been used in a number of fact verification approaches from the *reference* group. Usually, a statement is modelled as a triple and is verified on the basis of properties of the paths involving elements of that triple in existing KGs, considered as ground truth. We find several well-established KGs that have been used for that purpose, such as DBpedia [18] or Wikipedia infoboxes [17]. The Knowledge Vault [31] and the KnowMore [32] resources or the Voldemort KG [33] rely on structured markup annotations in order to match them to established KGs or

[17] `http://www.emergent.info`
[18] `https://www.kaggle.com/mrisdal/fake-news`
[19] `http://bsdetector.tech`

perform graph completion. This process implies the verification of the truthfulness of statements and the production of reliable factual information that can be used as ground truth.

Positioning ClaimsKG is entirely based on data from a number of established fact-checking websites and therefore falls into the first category of datasets presented above. With its more than 28K claims, it is, to our knowledge, the largest resource of its kind so far made available but also one archiving the largest spectrum of metadata categories. The open-source tools for its regeneration and update that we provide will allow for it to grow in size over time. In contrast to existing approaches, we model claims by the help of a specifically designed for that purpose RDF/S data model, fostering re-usability and extensibility. As compared to KG-based *reference* approaches, by collecting dynamically data from fact-checking websites, we focus on information of particular interest for the verification of newly emerging statements that are not available in Wikipedia or established KGs. ClaimsKG can be used as both training and evaluation data, allowing users (researchers in computer science or computational sociology, for example) to compile thematic samples of it by the help of structured queries, or by using the web application (cf. Section 5). Beyond the purposes of fact-checking, this is expected to foster research and data-driven studies in different areas of social and computational social science, as discussed in our use-case scenarios.

7 Conclusion and Future Work

We have introduced ClaimsKG, a live knowledge graph of fact-checked claims, which facilitates structured queries of related metadata, such as their truth values, authors or time of release. ClaimsKG is generated through a semi-automated pipeline, which harvests data from popular fact-checking sites on a regular basis, lifts data into a specifically developed for that purpose model and annotates claims with related entities from DBpedia. The KG is expected to provide support to research in the areas of fact-checking, stance detection and multiple topics related to the analysis of societal debates, where a quality ground truth of labelled claims is required in order to facilitate supervision, validation or reproducibility of research methods. In that, ClaimKG is the first of its kind publicly available large corpus of structured information.

There are several limitations of the current KG that are the focus of ongoing and short-term efforts. The development of an advanced claim matching approach and its evaluation is among them. We are working on building a gold-standard dataset of claim-pairs annotated with respect to different relatedness categories in order to evaluate the process and to provide training data to fine-tune state-of-the-art deep language modelling approaches such as BERT [34] to our matching task. We also intend to extend the content of our graph to other fact-checking websites and languages, enabling multi- and cross-lingual information retrieval and approaches for fact verification. With respect to augmenting ClaimsKG with additional claims, we also intend to harvest semi-structured schema.org markup[20] of claims from Web pages by exploiting data fusion pipelines developed as part of prior work [35]. Regarding the exploratory Web interface, in the future we aim to support the execution of federated queries that integrate information from external KGs like DBpedia (e.g., for enabling queries like the one in listing 2), as well as the inclusion of a statistical observatory allowing to extract distributions and correlations of different entities, topics and claims.

References

1. S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.
2. H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–36, 2017.
3. X. Wang, C. Yu, S. Baumgartner, and F. Korn, "Relevant document discovery for fact-checking articles," in *WWW*, pp. 525–533, 2018.
4. K. Popat, S. Mukherjee, J. Strötgen, and G. Weikum, "Where the truth lies: Explaining the credibility of emerging claims on the web and social media," in *WWW*, pp. 1003–1012, 2017.
5. S. Tschiatschek, A. Singla, M. Gomez Rodriguez, A. Merchant, and A. Krause, "Fake news detection in social networks via crowd signals," in *WWW*, pp. 517–524, 2018.
6. N. Hassan, B. Adair, J. T. Hamilton, C. Li, M. Tremayne, J. Yang, and C. Yu, "The quest to automate fact-checking," *world*, 2015.
7. P. Ferragina and U. Scaiella, "Tagme: On-the-fly annotation of short text fragments (by wikipedia entities)," in *ACM ICIKM*, pp. 1625–1628, ACM, 2010.
8. J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum, "Robust disambiguation of named entities in text," in *EMNLP*, (Stroudsburg, PA, USA), pp. 782–792, ACL, 2011.
9. B. Walenz, Y. Wu, S. Song, E. Sonmez, E. Wu, K. Wu, P. K. Agarwal, J. Yang, N. Hassan, A. Sultana, *et al.*, "Finding, monitoring, and checking claims computationally based on structured data," in *Computation+Journalism*, 2014.
10. Y. Wu, P. K. Agarwal, C. Li, J. Yang, and C. Yu, "Toward computational fact-checking," *Proceedings of the VLDB Endowment*, vol. 7, no. 7, pp. 589–600, 2014.
11. N. Hassan, F. Arslan, C. Li, and M. Tremayne, "Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster," in *ACM SIGKDD*, pp. 1803–1812, ACM, 2017.
12. D. A. Scheufele, "Agenda-setting, priming, and framing revisited: Another look at cognitive effects of political communication," *Mass Communication and Society*, vol. 3, no. 2-3, pp. 297–316, 2000.
13. W. L. Bennett and B. Pfetsch, "Rethinking Political Communication in a Time of Disrupted Public Spheres," *J. of Communication*, vol. 68, pp. 243–253, 04 2018.
14. M. Smith, B. Shneiderman, L. Rainie, and I. Himelboim, "Mapping twitter topic networks: From polarized crowds to community clusters," 02 2014.
15. L. Hon, "Social media framing within the million hoodies movement for justice," *Public Relations Review*, vol. 42, 12 2015.

[20] Claims annotated using `https://schema.org/ClaimReview`

16. S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.
17. G. L. Ciampaglia, P. Shiralkar, L. M. Rocha, J. Bollen, F. Menczer, and A. Flammini, "Computational fact checking from knowledge networks," *PloS one*, 2015.
18. P. Shiralkar, A. Flammini, F. Menczer, and G. L. Ciampaglia, "Finding streams in knowledge graphs to support fact checking," in *ICDM*, pp. 859–864, IEEE, 2017.
19. A. Vlachos and S. Riedel, "Identification and verification of simple claims about statistical properties," in *CEMNLP*, pp. 2596–2601, ACL, 2015.
20. A. Vlachos and S. Riedel, "Fact checking: Task definition and dataset construction," in *Language Technologies and Comp. Social Science*, pp. 18–22, 2014.
21. W. Y. Wang, "Liar, liar pants on fire: A new benchmark dataset for fake news detection," *AMACL*, pp. 422–426, 2017.
22. H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: Analyzing language in fake news and political fact-checking," in *EMNLP*, pp. 2931–2937, 2017.
23. J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks.," in *IJCAI*, pp. 3818–3824, 2016.
24. K. Popat, S. Mukherjee, J. Strötgen, and G. Weikum, "Credibility assessment of textual claims on the web," in *ACM ICIKM*, pp. 2173–2178, ACM, 2016.
25. A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, L. Màrquez, P. Atanasova, W. Zaghouani, S. Kyuchukov, G. Da San Martino, and P. Nakov, "Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims, task 2: Factuality," in *CLEF*, CEUR-WS, 2018.
26. W. Ferreira and A. Vlachos, "Emergent: a novel data-set for stance classification," in *NAACL-HLT*, pp. 1163–1168, 2016.
27. V. Pérez-Rosas and R. Mihalcea, "Experiments in open domain deception detection," in *CEMNLP*, pp. 1120–1125, 2015.
28. G. Gorrell, K. Bontcheva, L. Derczynski, E. Kochkina, M. Liakata, and A. Zubiaga, "Rumoureval 2019: Determining rumour veracity and support for rumours," *Semantic Evaluation*, pp. 60–67, 2017.
29. J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "Fever: a large-scale dataset for fact extraction and verification," *NAACL-HLT*, pp. 809–819, 2018.
30. T. Mihaylova, P. Nakov, L. Marquez, A. Barron-Cedeno, M. Mohtarami, G. Karadzhov, and J. Glass, "Fact checking in community forums," *AAAI*, pp. 879–886, 2018.
31. X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang, "Knowledge vault: A web-scale approach to probabilistic knowledge fusion," in *ACM SIGKDD*, pp. 601–610, ACM, 2014.
32. R. Yua, U. Gadirajua, B. Fetahua, O. Lehmbergb, D. Ritzeb, and S. Dietzea, "Knowmore-knowledge base augmentation with structured web markup," *Semantic Web J., IOS Press*, 2017.
33. A. Tonon, V. Felder, D. E. Difallah, and P. Cudré-Mauroux, "Voldemortkg: Mapping schema. org and web entities to linked open data," in *ISWC*, pp. 220–228, Springer, 2016.
34. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *ArXiv e-prints*, vol. arXiv cs.CL 1810.04805, 2018.
35. R. Yu, U. Gadiraju, B. Fetahu, O. Lehmberg, D. Ritze, and S. Dietze, "Knowmore knowledge base augmentation with structured web markup," *Semantic Web Journal, IOS Press*, 2019.