TweetsCOV19 - A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic

Dimitar Dimitrov[1], Erdal Baran[1], Pavlos Fafalios[2], Ran Yu[1], Xiaofei Zhu[3], Matthäus Zloch[1], and Stefan Dietze[1,4,5]
[1]GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany
[2]Institute of Computer Science, FORTH-ICS, Heraklion, Greece
[3]Chongqing University of Technology, Chongqing, China
[4]Heinrich-Heine-University Düsseldorf, Germany
[5] L3S Research Center, Hannover, Germany
{dimitar.dimitrov,erdal.baran,ran.yu,matthaeus.zloch,stefan.dietze}@gesis.org
fafalios@ics.forth.gr
zxf@cqut.edu.cn

ABSTRACT

Publicly available social media archives facilitate research in the social sciences and provide corpora for training and testing a wide range of machine learning, NLP and information retrieval methods. With respect to the recent outbreak of COVID-19, online discourse on Twitter reflects public opinion and perception related to the pandemic itself as well as mitigating measures and their societal impact. Understanding such discourse, its evolution and interdependencies with real-world events or (mis)information can foster valuable insights. On the other hand, such corpora are crucial facilitators for computational methods addressing tasks such as sentiment analysis, event detection or entity recognition. However, obtaining, archiving and semantically annotating large amounts of tweets is costly. In this paper, we describe *TweetsCOV19*, a publicly available knowledge base of currently more than 8 million tweets, spanning the period Oct'19-Apr'20. Metadata about the tweets as well as extracted entities, hashtags, user mentions, sentiments, and URLs are exposed using established RDF/S vocabularies, providing an unprecedented knowledge base for a range of knowledge discovery tasks. Next to a description of the dataset and its extraction and annotation process, we present an initial analysis, use cases and usage of the corpus.

ACM Reference Format:
Dimitar Dimitrov[1], Erdal Baran[1], Pavlos Fafalios[2], Ran Yu[1], Xiaofei Zhu[3], Matthäus Zloch[1], and Stefan Dietze[1,4,5]. 2020. TweetsCOV19 - A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Social web platforms have emerged as a primary forum for online discourse. Such user-generated content can be seen as a comprehensive documentation of societal discourse of immense historical value for future generations [5] and as an important resource for contemporary research. On the one hand, research in the computational social sciences relies on social media data to gain novel insights, for instance, about the spreading pattern of false claims on Twitter [32] or prevalent biases observable in online discourse [6]. On the other hand, computational methods at the intersection of NLP, machine learning and information retrieval rely on social web corpora for training and evaluating methods for tasks such as sentiment analysis [22], the classification of sources of news, such as Web pages, PLDs, users or posts [24], or fake news detection [31].

Twitter specifically has been recognized as an important data source facilitating research focused on insights or methods related to online discourse. In particular during the recent COVID-19 pandemic, online discourse on Twitter has proved crucial to facilitate an understanding of the impact of the pandemic, implemented measures, societal attitudes and perceptions in this context and, most importantly, the interdependencies between public opinion and relevant political actions, policies, media events or scientific discoveries. Recent corpora include a multilingual dataset of COVID-19-TweetIDs [7] consisting of more than 129 million tweet IDs, or a tweet corpus with sentiment annotations released by Lamsal [20]. Next to datasets focused on COVID-19 as a whole, datasets on other related topics have been created, for instance, about vaccines [23] covering sentiment-annotated tweets since June 2017 mentioning vaccine-related keywords.

However, given the legal and computational challenges involved in processing, reusing and publishing data crawled from Twitter, existing corpora usually consist of either raw metadata (such as tweet-ids, user ids, publishing dates) [15] or very limited and only partially precomputed features, such as georeferences [25]. In addition, corpora tend to be tailored towards a technical audience, limiting reuse by non-technical research disciplines lacking the skills and infrastructure for large-scale data processing.

Whereas entity-centric access and exploration methods are crucial to facilitate exploration of large Twitter archives, *TweetsKB*

[1] consistently applies W3C data sharing standards to publish a long-term Twitter archive and precomputed features including disambiguated entities and sentiments in the form of an extensible and easy to access knowledge graph, turning it into a comprehensive knowledge base of online discourse. The pipeline and dataset described in [9] introduced a knowledge base of RDF data for more than 1.5 billion tweets spanning almost 5 years exposed using established vocabularies in order to facilitate a variety of multi-aspect data exploration scenarios. Building on the prior release of *TweetsKB* in 2018, this work provides the following contributions:

- **Extension of TweetsKB.** Building on a continuous Twitter crawl and a parellelised annotation pipeline, we expand TweetsKB with data from April 2018 up to now, including additional metadata of about 486 million tweets, adding up to an unprecedented corpus of more than 63 billion triples describing more than 2 billion tweets starting from February 2013. To the best of our knowledge, TweetsKB is the largest publicly available Twitter archive and the only dataset consistently providing a knowledge graph of tweet metadata and precomputed features about entities and sentiments. Next to adding additional data based on our enrichment and data lifting pipeline (Section 2), we also extend both the applied schema and enrichment pipeline in order to include additional features (shared URLs).
- **Extraction and publishing of *TweetsCOV19*, a knowledge graph of COVID-19-related online discourse**. Taking advantage of TweetsKB and related infrastructure, we extract TweetsCOV19, a unique corpus of COVID-19-related online discourse. By applying a well-designed seed list (Section 3.1), we extract a TweetsKB subset spanning the period OctâĂŹ19-AprâĂŹ20 and apply the same feature extraction and data publishing methods as for TweetsKB. This results in a dataset containing more than 270 million triples describing metadata for about 8.1 million tweets from 3.6 million twitter users. Data is accessible as downloadable dumps following the N3 format and can be queried online through a dedicated, HTTP-accessible SPARQL endpoint. An easy to process TSV file is provided in addition.
- **Initial descriptive data analysis, use cases, tasks and reuse.** Next to providing basic statistics about TweetsKB in general, we provide an initial analysis and exploration of the TweetsCOV19 data (Section 3.2) in order to facilitate an understanding and reuse of the dataset. In order to facilitate and document reuse and impact of the data, we introduce a number of use cases, discuss prior use (Section 4) of the data, for instance, to facilitate research in the social sciences, as well as additional computational tasks facilitated by TweetsCOV19. Among others, these include the task of predicting tweet virality, posed as computation challenge for the *CIKM2020 AnalytiCup*.

Given the fact that all Twitter corpora are prohibited from republishing actual tweet texts, precomputed features which reflect content and semantics of individual tweets, such as mentioned entities, hash-tags, or URLs, together with expressed sentiments provides a unique foundation for studying online discourse and its evolution over time. To the best of our knowledge, TweetsCOV19 is the only COVID-19-related dataset available as public knowledge graph of tweets metadata and semantic annotations following established vocabularies and Web data sharing standards.

[1]https://data.gesis.org/tweetskb

2 CONSTRUCTING A KNOWLEDGE BASE OF TWITTER DISCOURSE

Whereas the processing of TweetsCOV19, described in Section 3, builds on TweetsKB, here, we describe the construction process of TweetsKB as a general, large-scale knowledge base of Twitter discourse. Note that, next to updating the corpus with crawled data after the previous release, improvements were made to the processing pipeline for this release compared to the extraction process described in [9].

TweetsKB is a public RDF corpus containing a unique collection of more than 2 billion semantically-annotated tweets spanning more than 7 years (February 2013 - April 2020). Metadata about the tweets as well as extracted entities, sentiments, hashtags and user mentions are exposed using established RDF/S vocabularies, forming a large knowledge graph of tweet-related data and allowing the expression of structured (SPARQL) queries that satisfy complex/analytical information needs (Section 4). TweetsKB is generated through the following steps: i) harvesting, ii) filtering, iii) cleaning, iv) semantic annotation and metadata extraction, vi) data lifting (using a dedicated RDF/S model). Below we describe these steps.

Harvesting, filtering, cleaning. Tweets are continuously harvested through the public Twitter streaming API since January 2013, accumulating more than 9.5 billion tweets up to now (May 2020). While all data is being archived locally and on restricted servers, TweetsKB is based on the cleaned-up English-language subset. As part of the filtering step, we eliminate re-tweets and non-English tweets, reducing the number of tweets to about 2.3 billion tweets. In addition, we remove spam through a Multinomial Naive Bayes (MNB) classifier, trained on the HSpam dataset which has 94% precision on spam labels [27] removing an additional 10% of tweets.

Semantic annotation and metadata extraction. Adhering to the Twitter license terms, the text of each tweet is not republished itself but only tweet IDs which may be rehydrated for specific purposes. In addition, full-text is exploited for extracting and disambiguating mentioned entities (*entity linking*), as well as for extracting the magnitude of the expressed positive and negative sentiments (*sentiment analysis*). Relying on the experimental motivation and prior work in [9], for *entity linking*, we exploit Yahoo FEL [3]. FEL has shown particularly cost efficient performance on the task of linking entities from short texts to Wikipedia and is fast and lightweight, being well-suited to run over billions of tweets in a distributed manner. We trained the FEL model using a Wikipedia dump of April 2020 and we set a confidence threshold of -3 which has been shown empirically to provide annotations of good quality (favoring precision). We also store the confidence score of each extracted entity to facilitate data consumers to set confidence thresholds which suit their use cases and requirements when working with our precomputed annotations. The quality of the entity annotations produced by FEL over tweets was evaluated

Table 1: Descriptive statistics of *TweetsKB and TweetsCOV19* .

feature	TweetsKB total	TweetsKB unique	TweetsKB ratio of tweets with at least one feature	TweetsCOV19 total	TweetsCOV19 unique	TweetsCOV19 ratio of tweets with at least one feature
hashtags	739,642,147	52,244,423	0.19	3,653,928	566,308	0.30
mentions	1,072,723,250	116,499,222	0.35	5,363,449	1,251,963	0.40
entities	2,575,861,358	1,919,083	0.58	11,537,537	331,307	0.70
non-neutral sentiment	1,047,840,159	-	0.54	4,478,603	-	0.55

in [9], demonstrating high precision (86%) and an overall satisfactory performance (F1 = 54%).

For sentiment analysis, we used SentiStrength [30], a robust and efficient tool for sentiment strength detection on social web data. SentiStrength assigns both a positive and a negative score to a short text, to account for both types of sentiments that can be expressed at the same time. The value of a positive (negative) sentiment ranges from +1 (-1) for no positive (no negative) to +5 (-5) for extremely positive (extremely negative). We provide an evaluation of the quality of sentiment annotations produced by SentiStrength over tweets in [9], demonstrating a reasonable performance, in particular in distinguishing stronger sentiments.

Entity and sentiment annotations are accompanied by the following metadata extracted from the tweets: *tweet id*, *post date*, *username* (user who posted the tweet), *favourite* and *retweet count* (at the time of fetching the tweet), *hashtags* (words starting with #), and *user mentions* (words starting with @). Starting from April 2018, we also extract the *URLs* included in the tweets. For ensuring data privacy, we anonymize usernames to ensure that tweets for particular users can be aggregated but users not identified.

Data lifting. We generate RDF triples in the N3 format using the data model described in [9], which exploits terms from established vocabularies, most notably SIOC core ontology [4], ONYX [26], and schema.org [12]. The selection of vocabularies was based on the following objectives: i) avoiding schema violations, ii) enabling data interoperability through term reuse, iii) having dereferenceable URIs, iv) extensibility. During lifting, we normalize sentiment scores in the range $[0, 1]$ using the formula: $score = (|sentimentValue| - 1)/4)$. For this release, we extended the data model described in [9] with one additional property (*schema:citation*) which refers to a URL mentioned in the tweet. Given that roughly 21% of tweets contain URLs, providing means to analyse shared URLs and Pay-Level-Domains (PLDs) provides additional opportunities for a range of research questions and tasks, for instance, with respect to the spreading of misinformation.

Availability and access. Table 1 summarizes descriptive statistics of the dataset. TweetsKB currently contains approximately 62.23 billions of triples describing online discourse on Twitter. About 46% of the tweets have no sentiment, i.e. the score is zero for both the positive and the negative sentiment. FEL extracted at least one entity for 58% of the tweets, while the average number of entities per tweet is 1.26. 19% of the tweets contain at least one hashtag and 35% at least one user mention. Finally, 21% of the tweets from April 2018 to April 2020 contain at least one URL.

The full *TweetsKB* is available as N3 files (split by month) through the Zenodo data repository (DOI: 10.5281/zenodo.573852),[2] under a *Creative Commons Attribution 4.0* license. For demonstration purposes, we have also set up a public SPARQL endpoint, currently containing a subset of about 5% of the dataset[3]. Example queries and more information are available through *TweetsKB*'s home page.[4] The source code used for triplifying the data is available as open source on GitHub[5].

3 THE TWEETSCOV19 DATASET

In this section, we describe the extraction of the TweetsCOV19 dataset[6] — a subset of TweetKB containing tweets related to COVID-19, which captures online discourse about various aspects of the pandemic and its societal impact.

3.1 Extraction Procedure & Availability

To extract the dataset, we compiled a seed list of 268 COVID-19-related keywords[7]. The seed list is an extension of the seed list[8] of Chen *et al.* [7] and allows a broader view on the societal discourse on COVID-19 in Twitter. We conducted full text filtering on the cleaned full-text of English tweets (Section 2) and retain all tweets containing at least one of the keywords in the seed list. We consider only original tweets and no retweets. Our corpus contains 16,266,285 occurrences of seed terms where "ppe", the acronym for personal protective equipment such as face masks, eye protection, and gloves, is the most frequently matching keyword (cf. Table 2). We applied the same process to extract relevant metadata and semantically enrich each tweet as described in Section 2. To simplify analysis of the posted URLs, we resolved all shortened URLs.

The final TweetsCOV19 dataset consists of 8,151,524 original tweets posted by 3,664,518 users captured during Oct'19-Apr'20. New data will be incrementally added to the corpus. The current state of the full dataset is available in two formats: (i) as a text file with tabular separated values (tsv) and (ii) as RDF triples in N3 format (cf. Section 2). The N3 version of the dataset consists of 274, 451, 101 RDF triples accessible through a dedicated SPARQL-endpoint[9] and as downloadable dumps[10]. All data is available under a *Creative Commons Attribution 4.0* license.

Applications and use of the data are described in greater detail in Section 4. In Section 6, we provide a more thorough comparison of TweetsCOV19 and related datasets.

[2]https://zenodo.org/record/573852

[3]https://data.gesis.org/tweetskb/sparql (Graph IRI: http://data.gesis.org/tweetskb)
[4]https://data.gesis.org/tweetskb
[5]https://github.com/iosifidisvasileios/AnnotatedTweets2RDF
[6]https://data.gesis.org/tweetscov19
[7]https://data.gesis.org/tweetscov19/keywords.txt
[8]https://github.com/echen102/COVID-19-TweetIDs/blob/master/keywords.txt
[9]https://data.gesis.org/tweetscov19/sparql (Graph IRI: http://data.gesis.org/tweetscov19)
[10]https://zenodo.org/record/3871753

Table 2: Top five matching keywords, mentions, hashtags, and pay level domains of *TweetsCOV19*.

keywords	frequency	mentions	frequency	hashtags	frequency	PLDs	frequency
ppe	3,368,192	realdonaldtrump	41,839	covid19	160,585	twitter.com	251,839
coronavirus	2,363,080	narendramodi	13,039	coronavirus	148.317	www.youtube.com	99,505
covid	2,308,054	pmoindia	12,701	covid_19	27,049	www.instagram.com	50,846
corona	1,513,195	jaketapper	9,836	stayhome	26,542	www.nytimes.com	30,892
covid19	1,498,386	who	9,776	china	23,602	www.theguardian.com	26,737

The source code used for triplifying the data is available as open source on GitHub[11].

3.2 Initial Data Analysis

In this section, we present a preliminary and non-exhaustive analysis of the TweetsCOV19 dataset in order to facilitate an understanding of the data and captured features. Table 1 shows descriptive statistics of the TweetsCOV19 corpus. Comparing the ratio of tweets with at least one feature across TweetsKB and TweetsCOV19, we observe constantly higher numbers (at least 5%) for all features with non-natural sentiment being the sole exception (about 1%). The TweetsCOV19 dataset contains 2,148,490 URLs from 1,645,394 distinct pay level domains. Compared to the TweetsKB dataset from which the TweetsCOV19 dataset has been extracted, we observe that about 25% (21%) of tweets in TweetsCOV19 (TweetsKB) contain at least one URL (cf. Section 2). The higher proportion of URLs seems intuitive given that for emerging topics such as COVID-19, sharing informational resources is one of the primary motivations. Politicians, journalists, and health organisations are the most frequent user mentions, with *@realdonaldtrump* being by far the most frequently mentioned twitter user, while the most used hashtags are *#covid19* and *#coronavirus* (cf. Table 2). Apart form URLs to Twitter and other social media platforms, news outlets appear to be primary information sources. Although the TweetsCOV19 dataset contains data from Oct'19 to Apr'20, our next analysis concentrates on the period from Jan'20 to Apr'20 where the topic starts dominating social media. Figure 1 presents a comparison of hashtag popularity over time for *#coronavirus* vs. *#covid19* and *#hydroxychloroquine* vs. *#vaccine*. The hashtag *#coronavirus* is present for the whole period and shows a small initial peak just before the emergence of *#covid19* in the beginning of Feb'20. While *#vaccine* is a topic that has been receiving attention on social media even before the COVID-19 crisis, *#hydroxychloroquine* gained first popularity as a possible drug for treating COVID-19 patients. Nevertheless, mentions of both terms seem to be strongly correlated. Table 3 shows the top five most frequently recognized entities per month for the period Jan'20 to Apr'20. The entity "Coronavirus_disease_2019" experiences a drastic boost in Mar'20 and Apr'20 and is the most frequent overall in TweetsCOV19.

Sentiment features are shown in Figure 2(a) and (b), illustrating sentiments of tweets containing important user mentions, *i.e.*, Donald Trump (@realDonaldTrump) and the World Health Organization (WHO) (@who). While a systematic detection and interpretation of events is out of scope of this paper, the fluctuation of the sentiment in the figures may be better understood in the context of an excerpt from the timeline of major events as compiled by the Washington Post[12], highlighting Trumps mention of the COVID-19 outbreak in his state of the union address (4 February 2020) or the United States Department of Health and Human Services (HHS) announcing its first efforts to rapidly develop a coronavirus vaccine in cooperation with pharmaceutical industry representatives (18 February 2020) .

Another view on these events is provided by the sentiment of tweets containing URLs to BreitbartâĂŤa politically far-right-wing associated news media (cf. Figure 2(c))âĂŤand CNNâĂŤa left-wing associated media (cf. Figure 2(d)). We observe the strongest fluctuations in the sentiment, with the biggest divergence of sentiment to tweets sharing links to Breitbart articles in the week of Trump's State of the Union address (4 February 2020). The strongly diverging spikes, both positivity and negativity increase at the same, suggest controversiality and polarisation of the address, what may deserve further investigation into aspects associated with those sentiments. The opposite pattern can be observed around Feb. 22 for the sentiment of tweets mentioning WHO, showing an more positive scores with respect to both positive and negative sentiments.

4 USE CASES & IMPACT

This section describes usage and use cases of TweetsCOV19.

4.1 Exploitation Scenarios & Example Queries

Next to downloading the dumps, one can directly explore the full dataset or develop applications that make use the data through HTTP requests and the SPARQL endpoint, e.g., for retrieving specific data of interest, or for offering a user friendly interface on top of the endpoint, e.g. one similar to [11].[13] In this section, we introduce a few basic queries to illustrate the use of the endpoint.

Consider, for instance, that a user wants to investigate online discussions around the president of the United States in a specific time period, e.g., April 2020. The SPARQL query in Figure 3 retrieves the top-5 entities co-occurring with the entity *Donald Trump* in tweets of April 2020 ('top' in terms of number of tweets mentioning the entities). The query returns the following 5 entities: *China* (1,103 tweets), *Coronavirus disease 2019* (1,042 tweets), *Hydroxychloroquine* (703 tweets), *Disinfectant* (436 tweets), *President of the United States* (369 tweets). Individual entities of interest such as (*Hydroxychloroquine*) may be explored further. For instance, the SPARQL query in Figure 4 retrieves the number of tweets per day in April 2020 mentioning the entity *Hydroxychloroquine*. The results show a very high increment on April 6-7 (from around 500 tweets before April 6 to 2,324 in April 6 and 2,105 in April 7), suggesting that some significant event related to this entity took place during

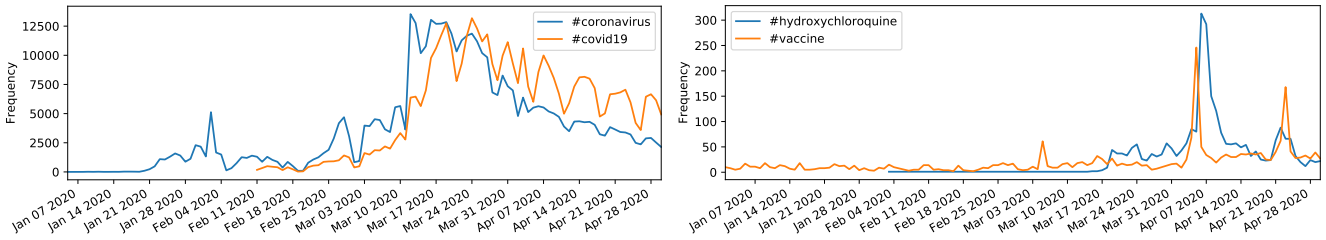
[11] https://github.com/iosifidisvasileios/AnnotatedTweets2RDF

[12] https://www.washingtonpost.com/politics/2020/04/20/what-trump-did-about-coronavirus-february

[13] Offering such a user-friendly interface is beyond the scope of this paper but considered future work.

Table 3: Entities over time. The table shows the top five entities (confidence level -2) and their frequency per month in the TweetsCOV19 dataset since the beginning of 2020. COVID-19* is used as shortcut for Coronavirus_disease_2019.

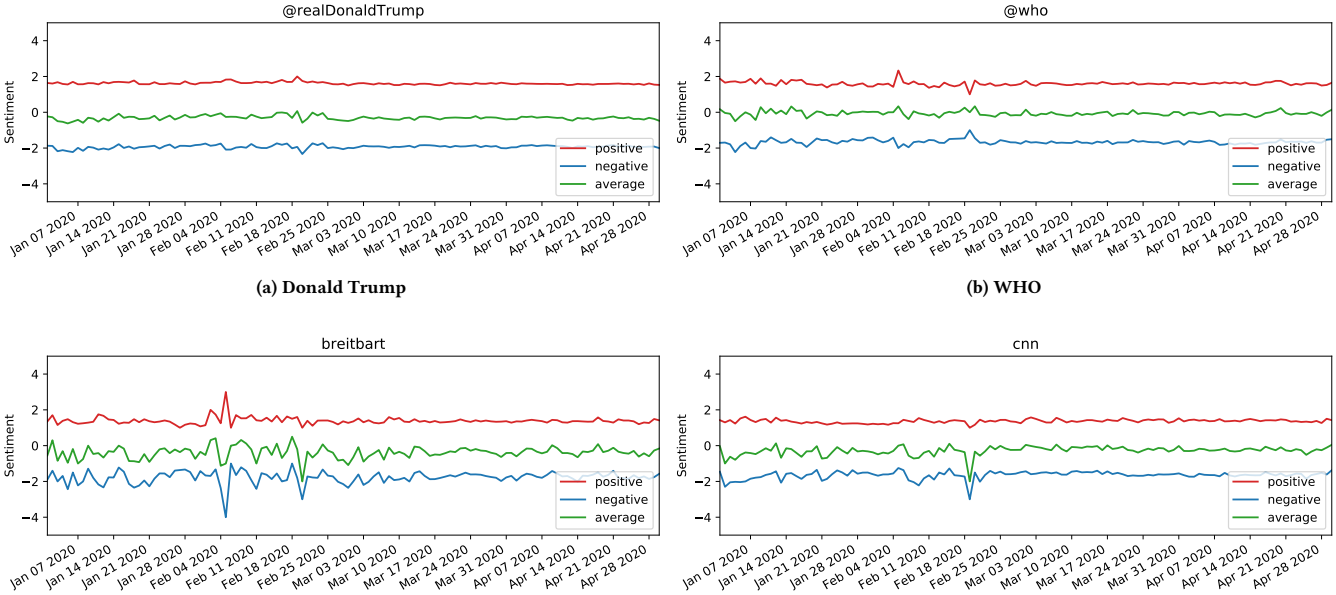
Jan'20		Feb'20		Mar'20		Apr'20	
entity	frequency	entity	frequency	entity	frequency	entity	frequency
Wuhan	10,147	Wuhan	10,494	COVID-19*	178,396	COVID-19*	200,342
Iran	5,905	COVID-19*	4,999	Social_distancing	66,176	Social_distancing	52,323
BTS	5,014	BTS	4,513	Italy	22,164	India	18,992
What's_Happening!!	4,899	What's_Happening!!	3,431	Wuhan	16,804	Hydroxychloroquine	15,820
Twitter	4,105	Twitter	3,351	India	15,822	Wuhan	14,478



(a) #coronavirus vs. #covid19

(b) #hydroxychloroquine vs. #vaccine

Figure 1: Hashtag usage over time. The figure shows a comparison of hashtag popularity over time for (a) the two most popular hashtags *#coronavirus* and *#covid19*, and for (b) *#hydroxychloroquine* vs. *#vaccine*.



(a) Donald Trump

(b) WHO

(c) Breitbart

(d) CNN

Figure 2: Sentiment over time. The figure shows the sentiment of tweets mentioning (a) Donald Trump and (b) WHO, and containing URLs to (c) Breitbart—a politically far-right-wing associated news media—and (d) CNN—a left-wing associated media.

this period. To explore this further, the SPARQL query in Figure 5 retrieves the top URLs included in tweets of 6-7 April 2020 that mention the entity *Hydroxychloroquine*. Headlines of the top 3 URLs are: *"Detroit rep says hydroxychloroquine, Trump helped save her life amid COVID-19 fight"*[14] (54 tweets), *"Trump's Aggressive Advocacy of Malaria Drug for Treating Coronavirus Divides Medical*

[14]https://www.freep.com/story/news/local/michigan/detroit/2020/04/06/democrat-karen-whitsett-coronavirus-hydroxychloroquine-trump/2955430001

Community"[15] (53 tweets), and *"Scoop: Inside the epic White House fight over hydroxychloroquine"*[16] (30 tweets). These news articles provide indicators about the event during 6-7 April 2020 triggering a spike in popularity for the *hydroxychloroquine* topic.

That way, exploration of entities, events, sentiments and topics is facilitated through simple HTTP requests, whereas future work aims at providing a more user-friendly interface.

```
select ?otherEntity (count(?tweet) as ?count) where {
  ?tweet schema:mentions ?entity ; dc:created ?date
                FILTER(year(?date) = 2020 and month(?date) = 4) .
  ?entity a nee:Entity ; nee:hasMatchedURI dbr:Donald_Trump .
  ?tweet schema:mentions ?entity2 .
  ?entity2 a nee:Entity ; nee:hasMatchedURI ?otherEntity
                FILTER (?otherEntity != dbr:Donald_Trump)
} group by ?otherEntity order by desc(?count) LIMIT 5
```

Figure 3: Top entities co-occurring with the entity *Donald Trump* during April 2020.

```
select (day(?date) as ?day) (count(?tweet) as ?count) where {
  ?tweet schema:mentions ?entity ; dc:created ?date
              FILTER(year(?date) = 2020 and month(?date) = 4) .
  ?entity a nee:Entity ; nee:hasMatchedURI dbr:Hydroxychloroquine .
} group by day(?date) order by day(?date)
```

Figure 4: Number of tweets per day in April 2020 mentioning the entity *Hydroxychloroquine*.

```
select ?url (count(?tweet) as ?count) where {
  ?tweet schema:mentions ?entity ; dc:created ?date
    FILTER(year(?date)=2020 and month(?date)=4 and (day(?date)=6 || day(?date)=7)) .
  ?entity a nee:Entity ; nee:hasMatchedURI dbr:Hydroxychloroquine .
  ?tweet schema:citation ?url
} group by ?url order by desc(?count)
```

Figure 5: Top URLs mentioned in tweets of 6 April 2020 together with the entity *Hydroxychloroquine*.

4.2 CIKM2020 AnalytiCup

CIKM 2020 AnalytiCup[17] is an open competition involving exciting data challenges aimed at members of the industry and academia interested in information retrieval, machine learning or NLP. The CIKM 2020 AnalytiCup is to be held in conjunction with the CIKM2020 conference during October 2020. The TweetsCOV19 dataset presented in this paper will be used in the *COVID-19 Retweet Prediction Challenge*[18]. The goal of this challenge is to predict the popularity and virality of COVID-19-related tweets in terms of the number of their retweets. To build the model, participants are allowed to use only the semantic annotations and metadata of the original tweets as provided in the TweetsCOV19 dataset. Retweeting—re-posting original content without any change—is a popular function in online social networks and amplifies the spread of original messages. Understanding retweet behavior is useful and has many practical applications, e.g. (political) audience design and marketing [17, 28], tracking of (fake) news and misinformation [21, 32], social event detection [13]. In particular, when designing campaigns of high societal impact and relevance, for instance when handling communication through emergencies such as hurricane warnings [18] and health-related campaigns about breast cancer screening [8], being able to predict future popularity of tweets is crucial.

This makes retweet prediction a crucial task when studying online information diffusion processes where TweetsCOV19 has the capacity to shape the understanding of such processes through its features such as entities, URLs, sentiments.

4.3 Other Usage & Impact

The TweetsKB and TweetsCOV19 datasets are currently used to support interdisciplinary research in various fields. TweetsKB is currently used to shape the understanding of solidarity discourse in the context of migration, e.g. as part of the SOLDISK project[19]. In addition, ongoing joint work with media and communication studies researchers[20] uses TweetsKB to investigate the societal impact of the ongoing Corona pandemic and most importantly, acceptance and trust for mitigating measures, the individual risk assessment and the impact of specific media events or information campaigns on related discourse and solidarity within society. In this context, in particular the impact of misinformation on solidarity and attitudes is being explored, taking advantage of the provided metadata together with additional metadata such as shared URLs and claims conveyed as part of these. Additional use cases are the joint exploration of means to extract statistically representative data for federal statistical agencies such as DESTATIS[21] as a way to complement traditional data gathering instruments, such as survey programmes, which are not well-suited to capture societal discourse or dynamic interdependencies. From a methodological perspective, recent work is concerned with using the corpus as training/testing data for stance detection tasks.

Among the lessons learned so far is the fact that, despite all data preprocessing and enrichment aimed at simplifying re-use and interpretation of the data, data consumers tend to depend on support from computer and data scientists to handle and analyse the data. While in some cases the key issue is handling data at such a scale, in other cases, interpreting serialisation formats (such as JSON or N3) or vocabularies poses challenges for users. In addition, data quality problems related to the underlying data as well as preprocessed features call for highly collaborative projects where expertise with respect to data characteristics and computational methods contributes to addressing higher level research questions.

5 SUSTAINABILITY, MAINTENANCE & EXTENSIBILITY

With respect to ensuring long-term sustainability, two aspects are of crucial importance: (i) maintenance and sustainability of the corpus and enrichment pipeline and (ii) maintenance of a user base and network. In order to ensure long-term sustainability, GESIS as research data infrastructure organisation exploits its technical expertise in hosting robust research data services has taken over the TweetsKB corpus with this recent update and hosts and maintains

[15] https://www.nytimes.com/2020/04/06/us/politics/coronavirus-trump-malaria-drug.html

[16] https://www.axios.com/coronavirus-hydroxychloroquine-white-house-01306286-0bbc-4042-9bfe-890413c6220d.html

[17] https://cikm2020.org/analyticup

[18] http://data.gesis.org/covid19challenge

[19] https://www.uni-hildesheim.de/soldisk/en/project-description

[20] https://www.phil-fak.uni-duesseldorf.de/en/kmw/professur-i-prof-dr-frank-marcinkowski/research-areas

[21] https://www.destatis.de/EN

both TweetsKB and TweetsCOV19. A user base of social scientists and computer scientists currently exploit the corpus through various use cases, projects and initiatives (Section 4). Maintenance of the corpus will be facilitated through the continuous process of crawling 1% of all tweets (running since January 2013) through the public Twitter API. In order to cater for downtimes and ensure that historic data is available for all time periods, redundant crawlers have been set up since March 2019. Storage of raw API output is currently handled through both, secure local GESIS storage services as well as the HDFS cluster at L3S Research Center.

The annotation and triplification process (Section 2) will be periodically repeated in order to incrementally expand the corpus and ensure its currentness, one of the requirements for many of the envisaged use cases of the dataset. While this will permanently increase the population of the dataset, the schema itself is extensible and facilitates the enrichment of tweets with additional information, for instance, to add information about the users involved in particular interactions (retweets, likes) or additional information about involved entities or references/URLs.

Whereas the use of Zenodo for depositing the dataset, as well as its registration at `datahub.ckan.io`, makes it citable and findable, we are currently exploring additional means, e.g. GESIS-hosted research data portals and registries to further publish and disseminate the dataset or particular subsets.

Next to facilitating reuse of *TweetsKB* itself, we also publish the source code used for triplifying the data (see Footnote 11), to enable third parties establishing and sharing similar corpora, for instance, focused Twitter crawls for certain topics. By following established W3C principles for data sharing and through the use of persistent URIs, both the schema as well as the corpus itself can be extended and linked. Current work, for instance, is concerned with computing stances of tweets towards claims, such as the ones public in *ClaimsKG*[22] [29] and explicitly capture stances as metadata.

TweetsCOV in particular will be updated continuously with the next release scheduled together with the submission deadline of CIKM2020 AnalytiCup. This allows us to utilise data from the period since this release as testing data for challenge participants. A user base emerged gradually throughout the past years, most importantly through enabling non-computer scientists to interact and analyse the data4). In addition, the corpus will be further advertised through interdisciplinary networks and events (like the Web Science Trust[23]) or the CIKM2020 AnalytiCup[17].

6 RELATED WORK

A number of Twitter-related datasets have emerged, to enable research in different fields such as NLP or the social sciences. Some datasets contain only information filtered from raw twitter stream data, for instance, to extract subsets of relevance to particular events[24] while others include annotations, such as mentioned entities [9], or manually curated labels, e.g. sentiments[25] to enable supervised machine learning approaches.

Since the COVID-19 pandemic started around January 2020, several Twitter related datasets have been released for academic use, including one stream API and 13 datasets related to Covid-19 discussion on Twitter, summarised at https://data.gesis.org/tweetscov19/. The COVID-19 streaming API from Twitter[26] returns tweets filtered based on 590 COVID-19 related keywords and hashtags (snapshot of terms on May 13, 2020) in the legacy enriched native response format[27]. For the majority of the 13 datasets, tweets are harvested and filtered from the Twitter stream based on mentions of COVID-19-related keywords and hashtags [2, 15, 16, 23, 23, 25]. The number of keywords and hashtags range from 3 [16] to 800 [25]. Some of the datasets further apply language filters [1, 10, 20, 33] or other requirements such as the availability of location information [19]. Instead of filtering from Twitter streaming data, authors of ArCOV-19 [14] collect tweets returned by the Twitter standard search API[28] when using COVID-19 related keywords (e.g. Corona) as queries and written in Arabic.

All datasets contain IDs and publishing dates of tweets that can be used to rehydrate tweets, i.e. to acquire actual tweet content of the tweets. Some also contain further information of tweets such as the publishing time[2], user ID [10, 15, 25, 33], geo-location [19, 23, 25] and retweet information [33]. A few datesets contain automatic annotations such as frequent used terms [2], sentiment scores per tweet [20], geo-location inferred from tweets [15] or places mentioned [23, 25]. The starting date of the data collection varies, with the earliest available dataset providing data starting from January 1, 2020 [1]. The Vaccine Sentiment Tracking dataset [23] which is intended for sentiment analysis on vaccine related topics even dates back to June 29, 2017. Most of the found datasets are being updated regularly. The number of tweets contained in the 13 datasets range from 747,599 [14] to over 524 million [25] by the time of this study, i.e. 20 May, 2020.

The filtering criteria (e.g. keywords and selected user accounts) of all datasets we discovered are transparent. All datasets are available as *csv*, *tsv*, *json* or plain text files for downloading.

TweetsCOV19 differs from existing datasets as: 1) it is extracted from a permanent crawl (TweetsKB) spanning more that 7 years – facilitating to trace keywords and topics over extended periods of time - and will be continuously updated, 2) it has rich semantic annotations – entities and original URLs mentioned in tweets, sentiment scores of entities, 3) the data is published following FAIR/W3C standards and established vocabularies and can be accessed in various ways – downloadable data dump as tab separated text files and RDF triples in N3 format, and a live SPARQL-endpoint. In particular, given that legal constraints prevent the republication of actual tweet text, precomputed features which reflect the semantics of tweets are both a distinctive feature of our dataset and a crucial requirement for efficiently analysing online discourse.

7 CONCLUSIONS

As part of this work, we have introduced a significant update of TweetsKB and introduced TweetsCOV19, a knowledge graph of Twitter discourse on COVID-19 and its societal impact. TweetsCOV19

[22] https://data.gesis.org/claimskg
[23] http://www.webscience.org
[24] https://digital.library.unt.edu/ark:/67531/metadc1259406
[25] https://data.world/crowdflower/weather-sentiment
[26] https://developer.twitter.com/en/docs/labs/covid19-stream
[27] https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object
[28] https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets

differs from existing related corpora through it unique set of precomputed semantic features and consistently applied knowledge graph principles, facilitating exploration and analysis of online discourse even without costly feature computation or rehydration of tweets. In addition, we have introduced a number of use cases from various disciplines, currently exploiting the corpus for deriving insights or evaluating computational methods for various tasks.

Future work will take advantage of the extensible knowledge graph nature of the corpus to incrementally add further contextual information, for instance, through computation of stances towards claims taking advantage of ClaimsKG [29] and stance detection methods, through classifying tweets or users based on their shared URLs and claims or, more specifically, to detect discourse related to biased claims and misinformation.

ACKNOWLEDGMENTS

We would like to thank our colleagues at L3S Research Center (Germany) and Humboldt University Berlin (Germany) involved in initialising and running the long-term Twitter crawl underlying TweetsKB and TweetsCOV19.

REFERENCES

[1] Sarah Alqurashi, Ahmad Alhindi, and Eisa Alanazi. 2020. Large arabic twitter dataset on covid-19. *arXiv preprint arXiv:2004.04315* (2020).

[2] Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Katya Artemova, Elena Tutubalina, and Gerardo Chowell. 2020. *A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration.* https://doi.org/10.5281/zenodo.3831406

[3] Roi Blanco, Giuseppe Ottaviano, and Edgar Meij. 2015. Fast and space-efficient entity linking for queries. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. 179–188.

[4] John G Breslin, Stefan Decker, Andreas Harth, and Uldis Bojars. 2006. SIOC: an approach to connect web-based communities. *Intern. Journal of Web Based Communities* 2, 2 (2006).

[5] Axel Bruns and Katrin Weller. 2016. Twitter as a first draft of the present: and the challenges of preserving it for the future. In *8th ACM Conference on Web Science*.

[6] Abhijnan Chakraborty, Johnnatan Messias, Fabricio Benevenuto, Saptarshi Ghosh, Niloy Ganguly, and Krishna P Gummadi. 2017. Who makes trends? understanding demographic biases in crowdsourced recommendations. In *Eleventh International AAAI Conference on Web and Social Media*.

[7] Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health Surveill* 6, 2 (29 May 2020), e19273. https://doi.org/10.2196/19273

[8] Jae Eun Chung. 2017. Retweeting in health promotion: Analysis of tweets about Breast Cancer Awareness Month. *Computers in Human Behavior* 74 (2017), 112–119.

[9] Pavlos Fafalios, Vasileios Iosifidis, Eirini Ntoutsi, and Stefan Dietze. 2018. Tweetskb: A public and large-scale rdf corpus of annotated tweets. In *European Semantic Web Conference*. Springer, 177–190.

[10] Zhiwei Gao, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2020. NAIST COVID: Multilingual COVID-19 Twitter and Weibo Dataset. (2020). arXiv:2004.08145 [cs.SI]

[11] Malo Gasquet, Darlene Brechtel, Matthaus Zloch, Andon Tchechmedjiev, Katarina Boland, Pavlos Fafalios, Stefan Dietze, and Konstantin Todorov. 2019. Exploring Fact-checked Claims and their Descriptive Statistics. In *ISWC 2019 Satellite Tracks-18th International Semantic Web Conference*.

[12] Ramanathan V Guha, Dan Brickley, and Steve Macbeth. 2016. Schema.org: evolution of structured data on the web. *Commun. ACM* 59, 2 (2016), 44–51.

[13] Manish Gupta, Jing Gao, ChengXiang Zhai, and Jiawei Han. 2012. Predicting future popularity trend of events in microblogging platforms. *Proceedings of the American Society for Information Science and Technology* 49, 1 (2012), 1–10.

[14] Fatima Haouari, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed. 2020. ArCOV-19: The First Arabic COVID-19 Twitter Dataset with Propagation Networks. *arXiv preprint arXiv:2004.05861* (2020).

[15] Xiaolei Huang, Amelia Jamison, David Broniatowski, Sandra Quinn, and Mark Dredze. 2020. *Coronavirus Twitter Data: A collection of COVID-19 tweets with automated annotations.* https://doi.org/10.5281/zenodo.3735015 http://twitterdata.covid19dataresources.org/index.

[16] Daniel Kerchner and Laura Wrubel. 2020. Coronavirus Tweet Ids. https://doi.org/10.7910/DVN/LW0BTB

[17] Eunice Kim, Yongjun Sung, and Hamsu Kang. 2014. Brand followersâĂŹ retweeting behavior on Twitter: How brand relationships influence brand electronic word-of-mouth. *Computers in Human Behavior* 37 (2014), 18–25.

[18] Marina Kogan, Leysia Palen, and Kenneth M Anderson. 2015. Think local, retweet global: Retweeting by the geographically-vulnerable during Hurricane Sandy. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 981–993.

[19] Rabindra Lamsal. 2020. Coronavirus (COVID-19) Geo-tagged Tweets Dataset. https://doi.org/10.21227/fpsb-jz61

[20] Rabindra Lamsal. 2020. Coronavirus (COVID-19) Tweets Dataset. https://doi.org/10.21227/781w-ef42

[21] Cristian Lumezanu, Nick Feamster, and Hans Klein. 2012. # bias: Measuring the tweeting behavior of propagandists. In *Sixth International AAAI Conference on Weblogs and Social Media*.

[22] Abhilash Mittal and Sanjay Patidar. 2019. Sentiment Analysis on Twitter Data: A Survey. In *Proceedings of the 2019 7th International Conference on Computer and Communications Management* (Bangkok, Thailand) *(ICCCM 2019)*. Association for Computing Machinery, New York, NY, USA, 91âĂŞ95. https://doi.org/10.1145/3348445.3348466

[23] Martin M Müller and Marcel Salathé. 2019. Crowdbreaks: Tracking health trends using public social media data and crowdsourcing. *Frontiers in public health* 7 (2019).

[24] 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 *WWW*. 1003–1012.

[25] Umair Qazi, Muhammad Imran, and Ferda Ofli. 2020. GeoCoV19: A Dataset of Hundreds of Millions of Multilingual COVID-19 Tweets with Location Information. *ACM SIGSPATIAL Special* 12, 1 (2020).

[26] J Fernando Sánchez-Rada and Carlos A Iglesias. 2016. Onyx: A linked data approach to emotion representation. *Information Processing & Management* 52, 1 (2016), 99–114.

[27] Surendra Sedhai and Aixin Sun. 2015. Hspam14: A collection of 14 million tweets for hashtag-oriented spam research. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 223–232.

[28] Stefan Stieglitz and Linh Dang-Xuan. 2012. Political communication and influence through microblogging–An empirical analysis of sentiment in Twitter messages and retweet behavior. In *2012 45th Hawaii International Conference on System Sciences*. IEEE, 3500–3509.

[29] A. Tchechmedjiev, P. Fafalios, K. Boland, M. Gasquet, M. Zloch, B. Zapilko, S. Dietze, and K. Todorov. 2019. ClaimsKG: A Live Knowledge Graph of Fact-Checked Claims. In *18th International Semantic Web Conference (ISWC 19)* (Auckland, New Zealand). https://stefandietze.wordpress.com/publications/

[30] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2012. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology* 63, 1 (2012), 163–173.

[31] Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. 2018. Fake News Detection in Social Networks via Crowd Signals. In *WWW*. 517–524.

[32] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[33] IBRAHIM SABUNCU; ZEYNEP YUREK. 2020. Corona Virus (COVID-19) Turkish Tweets Dataset. https://doi.org/10.21227/0wf0-0792