TweetsKB: A Public and Large-Scale RDF Corpus of Annotated Tweets

Pavlos Fafalios(✉), Vasileios Iosifidis, Eirini Ntoutsi, and Stefan Dietze

L3S Research Center, University of Hannover, Hannover, Germany
{fafalios,iosifidis,ntoutsi,dietze}@L3S.de

Abstract. Publicly available social media archives facilitate research in a variety of fields, such as data science, sociology or the digital humanities, where Twitter has emerged as one of the most prominent sources. However, obtaining, archiving and annotating large amounts of tweets is costly. In this paper, we describe *TweetsKB*, a publicly available corpus of currently more than 1.5 billion tweets, spanning almost 5 years (Jan'13–Nov'17). Metadata information about the tweets as well as extracted entities, hashtags, user mentions and sentiment information are exposed using established RDF/S vocabularies. Next to a description of the extraction and annotation process, we present use cases to illustrate scenarios for entity-centric information exploration, data integration and knowledge discovery facilitated by *TweetsKB*.

Keywords: Twitter · RDF · Entity linking · Sentiment analysis
Social media archives

Resource type: Dataset
Permanent URL: https://doi.org/10.5281/zenodo.573852.

1 Introduction

Social microblogging services have emerged as a primary forum to discuss and comment on breaking news and events happening around the world. Such user-generated content can be seen as a comprehensive documentation of the society and is of immense historical value for future generations [4].

In particular, Twitter has been recognized as an important data source facilitating research in a variety of fields, such as data science, sociology, psychology or historical studies where researchers aim at understanding behavior, trends and opinions. While research usually focuses on particular topics or entities, such as persons, organizations, or products, entity-centric access and exploration methods are crucial [31].

However, despite initiatives aiming at collecting and preserving such user-generated content (e.g., the Twitter Archive at the Library of Congress [33]),

A. Gangemi et al. (Eds.): ESWC 2018, LNCS 10843, pp. 177–190, 2018.
https://doi.org/10.1007/978-3-319-93417-4_12

the absence of publicly accessible archives which enable entity-centric exploration remains a major obstacle for research and reuse [4], in particular for non-technical research disciplines lacking the skills and infrastructure for large-scale data harvesting and processing.

In this paper, we present *TweetsKB*, a public corpus of RDF data for a large collection of anonymized tweets. *TweetsKB* is unprecedented as it currently contains data for more than 1.5 billion tweets spanning almost 5 years, includes entity and sentiment annotations, and is exposed using established vocabularies in order to facilitate a variety of multi-aspect data exploration scenarios.

By providing a well-structured large-scale Twitter corpus using established W3C standards, we relieve data consumers from the computationally intensive process of extracting and processing tweets, and facilitate a number of data consumption and analytics scenarios including: (i) time-aware and entity-centric exploration of the Twitter archive [6], (ii) data integration by directly exploiting existing knowledge bases (like DBpedia) [6], (iii) multi-aspect entity-centric analysis and knowledge discovery w.r.t. features like entity popularity, attitude or relation with other entities [7]. In addition, the dataset can foster further research, for instance, in entity recommendation, event detection, topic evolution, and concept drift.

Next to describing the annotation process (entities, sentiments) and the access details (Sect. 2), we present the applied schema (Sect. 3) as well as use case scenarios and update and maintenance procedures (Sect. 4). Finally, we discuss related works (Sect. 5) and conclude the paper (Sect. 6).

2 Generating TweetsKB

TweetsKB is generated through the following steps: (i) tweet archival, filtering and processing, (ii) entity linking and sentiment extraction, and (iii) data lifting. This section summarizes the above steps while the corresponding schema for step (iii) is described in the next section.

2.1 Twitter Archival, Filtering and Processing

The archive is facilitated by continuously harvesting tweets through the public Twitter streaming API since January 2013, accumulating more than 6 billion tweets up to now (December 2017).

As part of the filtering step, we eliminate re-tweets and non-English tweets, which has reduced the number of tweets to about 1.8 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 [25]. This removed about 10% of the tweets, resulting in a final corpus of 1,560,096,518 tweets. Figure 1 shows the number of tweets per month of the final dataset.

For each tweet, we exploit the following metadata: tweet id, post date, user who posted the tweet (username), favourite and retweet count (at the time of

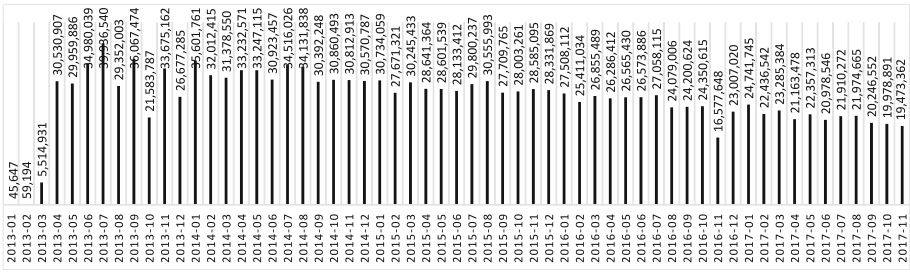

Fig. 1. Number of tweets per month of the *TweetsKB* dataset.

fetching the tweet[1]). We also extract hashtags (words starting with #) and user mentions (words starting with @). For the sake of privacy, we anonymize the usernames and we do not provide the text of the tweets (nevertheless, one can still apply user-based aggregation and analysis tasks). However, actual tweet content and further information can be fetched through the tweet IDs.

2.2 Entity Linking and Sentiment Extraction

For the *entity linking* task, we used Yahoo's FEL tool [1]. FEL is very fast and lightweight, and has been specially designed for linking entities from short texts to Wikipedia/DBpedia. We set a confidence threshold of -3 which has been shown empirically to provide annotations of good quality, while we also store the confidence score of each extracted entity. Depending on the specific requirements with respect to precision and recall, data consumers can select suitable confidence ranges to consider when querying the data.

In total, about 1.4 million distinct entities were extracted from the entire corpus, while the average number of entities per tweet is about 1.3. Figure 2 shows the distribution of the top-100,000 entities. There are around 15,000 entities with more than 10,000 occurrences, while there is a long tail of entities with less than 1,000 occurrences. Regarding their type, Table 1 shows the distribution of the top-100,000 entities in some popular DBpedia types (the sets are not disjoint). We notice that around 20% of the entities is of type *Person* and 15% of type *Organization*.

For *sentiment analysis*, we used SentiStrength, a robust tool for sentiment strength detection on social web data [28]. SentiStrength assigns both a positive and a negative score to a short text, to account for both types of sentiments expressed at the same time. The value of a positive sentiment ranges from $+1$ for no positive to $+5$ for extremely positive. Similarly, negative sentiment ranges from -1 (no negative) to -5 (extremely negative). We normalized both scores in the range $[0, 1]$ using the formula: $score = (|sentimentValue| - 1)/4)$. About

[1] By exploiting the tweet IDs, one can retrieve the latest favourite and retweet counts (however, only in case the corresponding tweets have not been deleted and are still publicly accessible).

Table 1. Overview of popular entity types of the top-100,000 entities.

DBpedia type	Number of distinct entities
http://dbpedia.org/ontology/Person	21,139 (21.1%)
http://dbpedia.org/ontology/Organisation	14,815 (14.8%)
http://dbpedia.org/ontology/Location	8,215 (8.2%)
http://dbpedia.org/ontology/Athlete	5,192 (5.2%)
http://dbpedia.org/ontology/Artist	3,737 (3.7%)
http://dbpedia.org/ontology/City	2,563 (2.6%)
http://dbpedia.org/ontology/Event	510 (0.5%)
http://dbpedia.org/ontology/Politician	208 (0.2%)

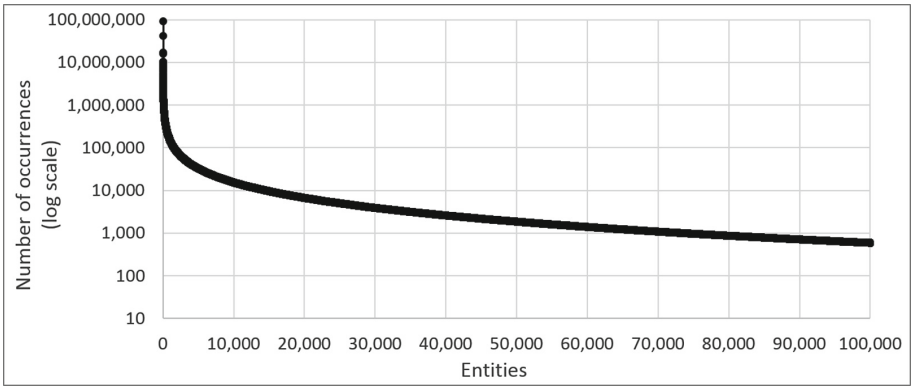

Fig. 2. Distribution of top-100,000 entities.

788 million tweets (50%) have no sentiment (*score* = 0 for both positive and negative sentiment).

Quality of Annotations. We evaluated the quality of the *entity annotations* produced by FEL using the ground truth dataset provided by the 2016 NEEL challenge of the 6th workshop on "Making Sense of Microposts" (#Microposts2016)[2] [16]. The dataset consists of 9,289 English tweets of 2011, 2013, 2014, and 2015. We considered all tweets from the provided training, dev and test files, without applying any training on FEL. The results are the following: *Precision* = 86%, *Recall* = 39%, *F1* = 54%. We notice that FEL achieves high precision, however recall is low. The reason is that FEL did not manage to recognize several difficult cases, like entities within hashtags and nicknames, which are common in Twitter due to the small number of allowed characters per tweet. Nevertheless, FEL's performance is comparable to existing approaches [15,16].

[2] http://microposts2016.seas.upenn.edu/.

Regarding *sentiment analysis*, we evaluated the accuracy of SentiStrength on tweets using two ground truth datasets: SemEval2017[3] (Task 4, Subtask A) [18], and TSentiment15[4] [8]. The SemEval2017 dataset consists of 61,853 English tweets of 2013–2017 labeled as positive, negative, or neutral. We run the evaluation on all the provided training files (of 2013–2016) and the 2017 test file. SentiStrength achieved the following scores: *AvgRec = 0.54* (recall averaged across the positive, negative, and neutral classes [24]), $F1^{PN} = 0.52$ (F1 averaged across the positive and negative classes), $Accuracy = 0.57$. The performance of SentiStrength is good considering that this is a multi-class classification problem. Moreover, the user can achieve higher precision by selecting only tweets with high positive or negative SentiStrength score. Regarding TSentiment15, this dataset contains 2,527,753 English tweets of 2015 labeled only with positive and negative classes (exploiting emoticons and a sentiment lexicon [8]). SentiStrength achieved the following scores: $F1^{PN} = 0.80$, $Accuracy = 0.91$. Here we notice that SentiStrength achieves very good performance.

2.3 Data Lifting and Availability

We generated RDF triples in the N3 format applying the RDF/S model described in the next section. The total number of triples is more than 48 billion. Table 2 summarizes the key statistics of the generated dataset. The source code used for triplifying the data is available as open source on GitHub[5].

Table 2. Key statistics of *TweeetsKB*.

Number of tweets	1,560,096,518
Number of distinct users	125,104,569
Number of distinct hashtags	40,815,854
Number of distinct user mentions	81,238,852
Number of distinct entities	1,428,236
Number of tweets with sentiment	772,044,599
Number of RDF triples	48,207,277,042

TweetsKB is available as N3 files (split by month) through the Zenodo data repository (DOI: 10.5281/zenodo.573852)[6], under a *Creative Commons Attribution 4.0* license. The dataset has been also registered at `datahub.ckan.io`[7].

[3] http://alt.qcri.org/semeval2017/task4/.
[4] https://l3s.de/~iosifidis/TSentiment15/.
[5] https://github.com/iosifidisvasileios/AnnotatedTweets2RDF.
[6] https://zenodo.org/record/573852.
[7] https://datahub.ckan.io/dataset/tweetskb.

Sample files, example queries and more information are available through *TweetsKB*'s home page[8]. For demonstration purposes, we have also set up a public SPARQL endpoint, currently containing a subset of about 5% of the dataset[9].

2.4 Runtime for Annotation and Triplification

The time for annotating the tweets and generating the RDF triples depends on several factors including the dataset volume, the used computing infrastructure as well as the available resources and the load of the cluster during the analysis time. The Hadoop cluster used for creating *TweetsKB* consists of 40 computer nodes with a total of 504 CPU cores and 6,784 GB RAM. The most time consuming task is entity linking where we annotated on average 4.8M tweets per minute using FEL, while SentiStrength annotated almost 6M tweets per minute. Finally, for the generation of the RDF triples we processed 14M tweets per minute on average.

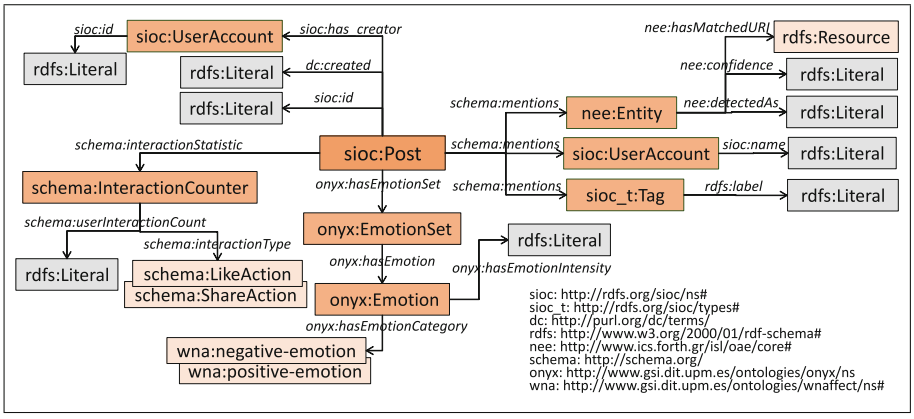

Fig. 3. An RDF/S model for describing metadata and annotation information for a collection of tweets.

3 RDF/S Model for Annotated Tweets

Our schema, depicted in Fig. 3, exploits terms from established vocabularies, most notably SIOC (Semantically-Interlinked Online Communities) core ontology [3] and schema.org [17]. The selection of the vocabularies was based on the following objectives: (i) avoiding schema violations, (ii) enabling data interoperability through term reuse, (iii) having dereferenceable URIs, (iv) extensibility. Next to modeling data in our corpus, the proposed schema can be applied over

[8] http://l3s.de/tweetsKB/.
[9] http://l3s.de/tweetsKB/endpoint/ (Graph IRI: http://l3s.de/tweetsKB/).

any annotated social media archive (not only tweets), and can be easily extended for describing additional information related to archived social media data and extracted annotations.

A tweet is associated with six main types of elements: (1) *general tweet metadata*, (2) *entity mentions*, (3) *user mentions*, (4) *hashtag mentions*, (5) *sentiment scores*, (6) *interaction statistics* (values expressing how users have interacted with the tweet, like favourite and retweet count). We use the property schema:mentions from schema.org[10] for associating a tweet with a mentioned entity, user or hashtag. We exploit schema.org due to its wide acceptance and less strict domain/range bindings which facilitate reuse and combination with other schemas, by avoiding schema violations.

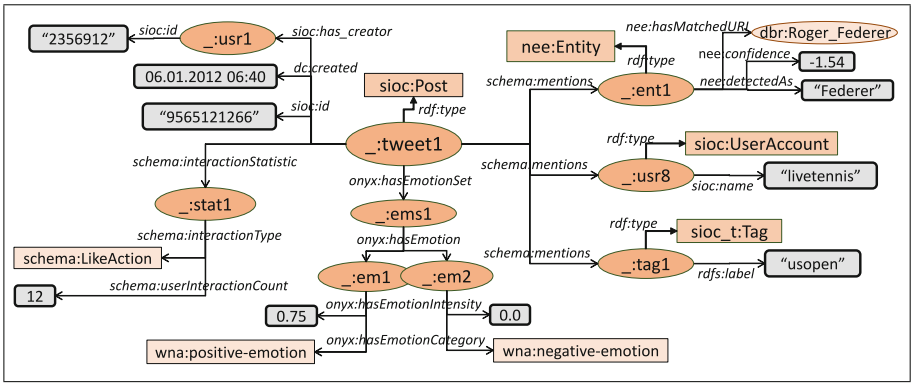

Fig. 4. Instantiation example of the RDF/S model.

For general *metadata*, we exploit SIOC as an established vocabulary for representing social Web data[11]. The class sioc:Post represents a tweet, while sioc:UserAccount a Twitter user.

An *entity mention* is represented through the Open NEE (Named Entity Extraction) model [5] which is an extension of the Open Annotation data model [23] and enables the representation of entity annotation results. For each recognized entity, we store its surface form, URI and confidence score. A *user mention* simply refers to a particular sioc:UserAccount, while for *hashtag mentions* we use the class sioc_t:Tag of the SIOC Types Ontology Module[12].

For expressing *sentiments*, we use the Onyx ontology[13] [22]. Through the class onyx:EmotionSet we associate a tweet with a set of emotions (onyx:Emotion). Note that the original domain of property onyx:hasEmotionSet is owl:Thing, which is

[10] http://schema.org/.
[11] Specification available at: http://rdfs.org/sioc/spec/.
[12] http://rdfs.org/sioc/types\#.
[13] https://www.gsi.dit.upm.es/ontologies/onyx/.

compatible with our use as property of sioc:Post. The property onyx:hasEmotionCategory defines the emotion type, which is either negative-emotion or positive-emotion as defined by the WordNet-Affect Taxonomy[14] and is quantified through onyx:hasEmotionIntensity.

Finally, for representing aggregated *interactions*, we use the class InteractionCounter of schema.org. We distinguish schema:LikeAction (for the favourite count) or schema:ShareAction (for the retweet count) as valid interaction types.

Figure 4 depicts a set of instances for a single tweet. In this example, the tweet mentions one user account (@livetennis) and one hashtag (#usopen), while the entity name *"Federer"* was detected, referring probably to the tennis player *Roger Federer* (with confidence score −1.54). Moreover, we see that the tweet has a positive sentiment of 0.75, no negative sentiment, while it has been marked as "favourite" 12 times.

4 Use Cases and Sustainability

4.1 Scenarios and Queries

Typical scenarios facilitated by *TweetsKB* include:

Advanced Exploration and Data Integration. By exploiting tweet metadata, extracted entities, sentiment values, and temporal information, one can run sophisticated queries that can also directly (at query-execution time) integrate information from external knowledge bases like DBpedia. For example, Listing 1 shows a SPARQL query obtaining popular tweets in 2016 (with more than 100 retweets) mentioning *German politicians* with strong negative sentiment (≥0.75). The query exploits extracted entities, sentiments, and interaction statistics, while it uses query federation to access DBpedia for retrieving the list of German politicians and their birth place.

```
SELECT DISTINCT ?tweetID ?sentNegScore ?retweetCount ?politician ?birthPlace WHERE {
  SERVICE <http://dbpedia.org/sparql> {
    ?politician dc:subject dbc:German_politicians ; dbo:birthPlace ?birthPlace }
  ?tweet a sioc:Post ; dc:created ?date ; sioc:id ?tweetID FILTER(year(?date) = 2016) .
  ?tweet schema:mentions ?entity . ?entity a nee:Entity ; nee:hasMatchedURI ?politician .
  ?tweet schema:interactionStatistic ?stat . ?stat schema:interactionType schema:ShareAction .
  ?stat schema:userInteractionCount ?retweetCount FILTER(?retweetCount > 100) .
  ?tweet onyx:hasEmotionSet ?emotSet . ?emotSet onyx:hasEmotion ?emot .
  ?emot onyx:hasEmotionCategory wna:negative-emotion ;
        onyx:hasEmotionIntensity ?sentNegScore FILTER (?sentNegScore >= 0.75) }
```

Listing 1. SPARQL query for retrieving popular tweets in 2016 mentioning German politicians with strong negative sentiment.

Listing 2 shows a query that combines extracted entities with hashtags. The query requests the top-50 hashtags co-occurring with the entity *Refugee* (http://dbpedia.org/resource/Refugee) in tweets of 2016. The result contains, among others, the following hashtags: #auspol, #asylum, #Nauru, #Greece, #LetThemStay, #BringThemHere.

[14] http://www.gsi.dit.upm.es/ontologies/wnaffect/.

```
SELECT ?hastagLabel (count(distinct ?tweet) as ?num) WHERE {
  ?tweet dc:created ?date FILTER(year(?date) = 2016) .
  ?tweet schema:mentions ?entity .
  ?entity a nee:Entity ; nee:hasMatchedURI dbr:Refugee .
  ?tweet schema:mentions ?hashtag.
  ?hashtag a sioc:Tag ; rdfs:label ?hastagLabel
} GROUP BY ?hastagLabel ORDER BY DESC(?num) LIMIT 50
```

Listing 2. SPARQL query for retrieving the top-50 hashtags co-occurring with the entity *Refugee* in tweets of 2016.

Temporal Entity Analytics. The work in [7] has proposed a set of measures that allow studying how entities are reflected in a social media archive and how entity-related information evolves over time. Given an entity and a time period, the proposed measures capture the following entity aspects: *popularity*, *attitude* (predominant sentiment), *sentimentality* (magnitude of sentiment), *controversiality*, and *connectedness* to other entities (entity-to-entity connectedness and k-network). Such time-series data can be easily computed by running SPARQL queries on *TweetsKB*. For example, the query in Listing 3 retrieves the monthly popularity of *Alexis Tsipras* (Greek prime minister) in Twitter in 2015 (using Formula 1 of [7]). The result of this query shows that the number of tweets increased significantly in June and July, likely to be caused by the Greek bailout referendum that was held in July 2015, following the bank holiday and capital controls of June 2015.

```
SELECT ?month xsd:double(?cEnt)/xsd:double(?cAll)
WHERE {
{ SELECT (month(?date) AS ?month) (count(?tweet) AS ?cAll) WHERE {
   ?tweet a sioc:Post ; dc:created ?date FILTER(year(?date) = 2015)
  } GROUP BY month(?date) }
{ SELECT (month(?date) AS ?month) (count(?tweet) AS ?cEnt) WHERE {
   ?tweet a sioc:Post ; dc:created ?date FILTER(year(?date) = 2015) .
   ?tweet schema:mentions ?entity .
   ?entity a nee:Entity ; nee:hasMatchedURI dbr:Alexis_Tsipras
  } GROUP BY month(?date) }
} ORDER BY ?month
```

Listing 3. SPARQL query for retrieving the monthly popularity of *Alexis Tsiprats* (Greek prime minister) in tweets in 2015 (using Formula 1 of [7]).

Time and Social Aware Entity Recommendations. Recent works have shown that entity recommendation is time-dependent, while the co-occurrence of entities in documents of a given time period is a strong indicator of their relatedness during that period and thus should be taken into consideration [29,32]. By querying *TweetsKB*, we can find entities of a specific type, or having some specific characteristics, that co-occur frequently with a query entity in a specific time period, a useful indicator for temporal prior probabilities when implementing time- and social-aware entity recommendations. For example, the query in Listing 4 retrieves the top-5 politicians co-occurring with *Barack Obama* in tweets of summer 2016. Here one could also follow a more sophisticated approach, e.g., by also considering the inverse tweet frequency of the top co-occurred entities.

```
SELECT ?politician (count(distinct ?tweet) as ?num) WHERE {
  SERVICE <http://dbpedia.org/sparql> {
     ?politician a dbo:Politician }
  ?tweet a sioc:Post ; dc:created ?date FILTER(?date >= "2016-06-01"^^xsd:date &&
                                               ?date <= "2016-08-30"^^xsd:date) .
  ?tweet schema:mentions ?entity .
  ?entity a nee:Entity ; nee:hasMatchedURI dbr:Barack_Obama .
  ?tweet schema:mentions ?entityPolit.
  ?entityPolit nee:hasMatchedURI ?politician FILTER (?politician != dbr:Barack_Obama)
} GROUP BY ?politician ORDER BY DESC(?num) LIMIT 5
```

Listing 4. SPARQL query for retrieving the top-5 politicians co-occurring with *Barack Obama* in tweets of summer 2016.

Data Mining and Information Discovery. Data mining techniques allow the extraction of useful and previously unknown information from raw data. By querying *TweetsKB* we can generate time series for a specific entity of interest modeling the temporal evolution of the entity w.r.t. different tracked dimensions like sentiment, popularity, or interactivity. Such multi-dimensional time-series can be used in a plethora of data mining tasks like entity forecasting (predicting entity-related features) [21], network-analysis (find communities and influential entities) [19], stream mining (sentiment analysis over data streams) [9,27], or change detection (e.g., detection of critical time-points) [11].

Thus, research in a range of fields is facilitated through the public availability of well-annotated Twitter data. Note also that the availability of publicly available datasets is a requirement for the data mining community and will allow not only the development of new methods but also for valid comparisons among existing methods, while existing repositories, e.g., UCI[15], lack of big, volatile and complex data.

4.2 Sustainability, Maintenance and Extensibility

The dataset has seen adoption and facilitated research in inter-disciplinary research projects such as ALEXANDRIA[16] and AFEL[17], involving researchers from a variety of organizations and research fields [6,7,10,30]. With respect to ensuring long-term sustainability, we anticipate that reuse and establishing of a user community for the corpus is crucial. While the aforementioned activities have already facilitated access and reuse, the corpus will be further advertised through interdisciplinary networks and events (like the Web Science Trust[18]). Besides, the use of Zenodo for depositing the dataset, as well as its registration at `datahub.ckan.io`, makes it citable and web findable.

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 and storing obtained data within the local Hadoop cluster at L3S Research Center. The annotation and triplification process (Sect. 2) will be

[15] http://archive.ics.uci.edu/ml/.
[16] http://alexandria-project.eu/.
[17] http://afel-project.eu/.
[18] http://www.webscience.org/.

periodically (quarterly) 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. Depending on the investigated research questions, it is anticipated that this kind of enrichment is essential, at least for parts of the corpus, i.e. for specific time periods or topics.

Next to the reuse of *TweetsKB*, we also publish the source code used for triplifying the data (see Footnote 5), to enable third parties establishing and sharing similar corpora, for instance, focused Twitter crawls for certain topics.

5 Related Work

There is a plethora of works on modeling social media data as well as on semantic-based information access and mining semantics from social media streams (see [2] for a survey). There are also Twitter datasets provided by specific communities for research and experimentation in specific research problems, like the "Making Sense of Microposts" series of workshops [15,16], or the "Sentiment Analysis in Twitter" tasks of the International Workshop on Semantic Evaluation [13,18]. Below we discuss works that exploit Semantic Web technologies for representing and querying social media data.

Twarql [12] is an infrastructure which translates microblog posts from Twitter as Linked Data in real-time. Similar to our approach, Twarql extracts entity, hashtag and user mentions, and the extracted content is encoded in RDF. The authors tested their approach using a small collection of 511,147 tweets related to iPad[19]. SMOB [14] is a platform for distributed microblogging which combines Social Web principles and Semantic Web technologies. SMOB relies on ontologies for representing microblog posts, hubs for distributed exchanging information, and components for linking the posts with other resources. TwitLogic [26] is a semantic data aggregator which provides a set of syntax conventions for embedding various structured content in microblog posts. It also provides a schema for user-driven data and associated metadata which enables the translation of microblog streams into RDF streams. The work in [20] also discusses an approach to annotate and triplify tweets. However, none of the above works provides a large-scale and publicly available RDF corpus of annotated tweets.

6 Conclusion

We have presented a large-scale Twitter archive which includes entity and sentiment annotations and is exposed using established vocabularies and standards. Data includes more than 48 billion triples, describing metadata and annotation

[19] The dataset is not currently available (as of March 15, 2018).

information for more than 1.5 billion tweets spanning almost 5 years. Next to the corpus itself, the proposed schema facilitates further extension and the generation of similar, focused corpora, e.g. for specific geographic or temporal regions, or targeting selected topics.

We believe that this dataset can foster further research in a plethora of research problems, like event detection, topic evolution, concept drift, and prediction of entity-related features, while it can facilitate research in other communities and disciplines, like sociology and digital humanities.

Acknowledgements. The work was partially funded by the European Commission for the ERC Advanced Grant ALEXANDRIA under grant No. 339233 and the H2020 Grant No. 687916 (AFEL project), and by the German Research Foundation (DFG) project OSCAR (Opinion Stream Classification with Ensembles and Active leaRners).

References

1. Blanco, R., Ottaviano, G., Meij, E.: Fast and space-efficient entity linking for queries. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. ACM (2015)
2. Bontcheva, K., Rout, D.: Making sense of social media streams through semantics: a survey. Semant. Web **5**(5), 373–403 (2014)
3. Breslin, J.G., Decker, S., Harth, A., Bojars, U.: SIOC: an approach to connect web-based communities. Int. J. Web Based Commun. **2**(2), 133–142 (2006)
4. Bruns, A., Weller, K.: Twitter as a first draft of the present: and the challenges of preserving it for the future. In: 8th ACM Conference on Web Science (2016)
5. Fafalios, P., Baritakis, M., Tzitzikas, Y.: Exploiting linked data for open and configurable named entity extraction. Int. J. Artif. Intell. Tools **24**(02), 42 (2015)
6. Fafalios, P., Holzmann, H., Kasturia, V., Nejdl, W.: Building and querying semantic layers for web archives. In: ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2017, Toronto, Ontario, Canada (2017)
7. Fafalios, P., Iosifidis, V., Stefanidis, K., Ntoutsi, E.: Multi-aspect entity-centric analysis of big social media archives. In: Kamps, J., Tsakonas, G., Manolopoulos, Y., Iliadis, L., Karydis, I. (eds.) TPDL 2017. LNCS, vol. 10450, pp. 261–273. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67008-9_21
8. Iosifidis, V., Ntoutsi, E.: Large scale sentiment learning with limited labels. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1823–1832. ACM (2017)
9. Iosifidis, V., Oelschlager, A., Ntoutsi, E.: Sentiment classification over opinionated data streams through informed model adaptation. In: Kamps, J., Tsakonas, G., Manolopoulos, Y., Iliadis, L., Karydis, I. (eds.) TPDL 2017. LNCS, vol. 10450, pp. 369–381. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67008-9_29
10. Kowald, D., Pujari, S.C., Lex, E.: Temporal effects on hashtag reuse in Twitter: a cognitive-inspired hashtag recommendation approach. In: Proceedings of the 26th International Conference on World Wide Web, pp. 1401–1410. International World Wide Web Conferences Steering Committee (2017)
11. Liu, S., Yamada, M., Collier, N., Sugiyama, M.: Change-point detection in time-series data by relative density-ratio estimation. Neural Netw. **43**, 72–83 (2013)

12. Mendes, P.N., Passant, A., Kapanipathi, P.: Twarql: tapping into the wisdom of the crowd. In: 6th International Conference on Semantic Systems. ACM (2010)
13. Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., Stoyanov, V.: SemEval-2016 task 4: sentiment analysis in Twitter. In: SemEval@ NAACL-HLT, pp. 1–18 (2016)
14. Passant, A., Bojars, U., Breslin, J.G., Hastrup, T., Stankovic, M., Laublet, P., et al.: An overview of SMOB 2: open, semantic and distributed microblogging. In: ICWSM, pp. 303–306 (2010)
15. Rizzo, G.: Making sense of microposts (# Microposts2015) named entity rEcognition and linking (NEEL) challenge (2015)
16. Rizzo, G., van Erp, M., Plu, J., Troncy, R.: Making sense of microposts (#Microposts2016) named entity rEcognition and linking (NEEL) challenge (2016)
17. Ronallo, J.: HTML5 microdata and schema.org. Code4Lib J. (16) (2012)
18. Rosenthal, S., Farra, N., Nakov, P.: SemEval-2017 task 4: sentiment analysis in Twitter. In: Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval 2017, pp. 502–518 (2017)
19. Rossi, M.-E.G., Malliaros, F.D., Vazirgiannis, M.: Spread it good, spread it fast: identification of influential nodes in social networks. In: Proceedings of the 24th International Conference on World Wide Web, pp. 101–102. ACM (2015)
20. Sahito, F., Latif, A., Slany, W.: Weaving Twitter stream into linked data a proof of concept framework. In: International Conference on Emerging Technologies (2011)
21. Saleiro, P., Soares, C.: Learning from the news: predicting entity popularity on Twitter. In: Boström, H., Knobbe, A., Soares, C., Papapetrou, P. (eds.) IDA 2016. LNCS, vol. 9897, pp. 171–182. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46349-0_15
22. Sánchez-Rada, J.F., Iglesias, C.A.: Onyx: a linked data approach to emotion representation. Inf. Process. Manag. **52**(1), 99–114 (2016)
23. Sanderson, R., Ciccarese, P., Van de Sompel, H., Bradshaw, S., Brickley, D., Castro, L.J.G., Clark, T., Cole, T., Desenne, P., Gerber, A., et al.: Open annotation data model. W3C Community Draft (2013)
24. Sebastiani, F.: An axiomatically derived measure for the evaluation of classification algorithms. In: Proceedings of the 2015 International Conference on The Theory of Information Retrieval, pp. 11–20. ACM (2015)
25. Sedhai, S., Sun, A.: HSpam14: a collection of 14 million tweets for hashtag-oriented spam research. In: SIGIR ACM (2015)
26. Shinavier, J.: Real-time #SemanticWeb in <= 140 chars. In: Proceedings of the Third Workshop on Linked Data on the Web, LDOW 2010 at WWW 2010 (2010)
27. Spiliopoulou, M., Ntoutsi, E., Zimmermann, M.: Opinion stream mining. Encycl. Mach. Learn. Data Min. 1–10 (2016)
28. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment strength detection for the social web. J. Am. Soc. Inf. Sci. Technol. **63**(1), 163–173 (2012)
29. Tran, N.K., Tran, T., Niederée, C.: Beyond time: dynamic context-aware entity recommendation. In: Blomqvist, E., Maynard, D., Gangemi, A., Hoekstra, R., Hitzler, P., Hartig, O. (eds.) ESWC 2017. LNCS, vol. 10249, pp. 353–368. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58068-5_22
30. Tran, T., Tran, N.K., Hadgu, A.T., Jäschke, R.: Semantic annotation for microblog topics using Wikipedia temporal information. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 97–106 (2015)
31. Weikum, G., Spaniol, M., Ntarmos, N., Triantafillou, P., Benczúr, A., Kirkpatrick, S., Rigaux, P., Williamson, M.: Longitudinal analytics on web archive data: it's about time! In: Biennial Conference on Innovative Data Systems Research (2011)

32. Zhang, L., Rettinger, A., Zhang, J.: A probabilistic model for time-aware entity recommendation. In: Groth, P., Simperl, E., Gray, A., Sabou, M., Krötzsch, M., Lecue, F., Flöck, F., Gil, Y. (eds.) ISWC 2016. LNCS, vol. 9981, pp. 598–614. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46523-4_36
33. Zimmer, M.: The Twitter archive at the library of congress: challenges for information practice and information policy. First Monday **20**(7) (2015)