SPARQL querying for validating the usage of automatically georeferenced social media data as human sensors for air quality

Stelios Andreadis
Information Technologies Institute
Centre for Research and Technology Hellas
Thessaloniki, Greece
andreadisst@iti.gr

Mirette Elias
Fraunhofer IAIS and University of Bonn
Bonn, Germany
melias@uni-bonn.de

Thanassis Mavropoulos
Information Technologies Institute
Centre for Research and Technology Hellas
Thessaloniki, Greece
mavrathan@iti.gr

Charis Papadopoulos
Information Technologies Institute
Centre for Research and Technology Hellas
Thessaloniki, Greece
chapapadopoulos@iti.gr

Nick Pantelidis
Information Technologies Institute
Centre for Research and Technology Hellas
Thessaloniki, Greece
pantelidisnikos@iti.gr

Ilias Gialampoukidis
Information Technologies Institute
Centre for Research and Technology Hellas
Thessaloniki, Greece
heliasgj@iti.gr

Stefanos Vrochidis
Information Technologies Institute
Centre for Research and Technology Hellas
Thessaloniki, Greece
stefanos@iti.gr

Ioannis Kompatsiaris
Information Technologies Institute
Centre for Research and Technology Hellas
Thessaloniki, Greece
ikom@iti.gr

***Abstract*—The problem of air pollution is one of the countless topics discussed on social media on an everyday basis. This rich, crowdsourced information can be exploited to assess the air quality of urban areas, using humans as sensors. Nevertheless, the majority of social media data are falsely geotagged or completely lack geoinformation, which is an essential attribute, while the reliability of the air pollution events reported by online citizens has to be proven. The scope of this work is to present a framework that collects Twitter messages in German that refer to the atmosphere, automatically georeferences them, and finally validates them through semantic representation and SPARQL queries in order to associate them with real measurements of air quality sensors. The georeferencing models are evaluated against state-of-the-art works and the proposed framework is validated in a near-six-month scenario in Germany.**

***Index Terms*—air quality, social media, georeferencing, semantic representation, SPARQL querying**

I. Introduction

Air pollution is a critical issue everywhere around the world. For instance, between 2000 and 2017, the population of the 27 member countries of the European Union was exposed to $PM_{2.5}$ and O_3 levels widely exceeding the WHO limit values for the protection of human health [1]. A first step towards dealing with the problems of air pollution is to monitor the air quality. The growth in social media daily use has led to a rich and up-to-date crowdsourced information, which can augment existing air pollution monitoring data, collected by sensors, as well as perception data that traditionally require expensive surveys [2]. However, the involvement of crowdsourced data can present certain shortcomings. Considering that location is an indispensable feature when it comes to air quality monitoring, social media data often lack geoinformation or even provide false geolocations [3]. In addition, the reliability of using humans as sensors is questionable and needs to be validated by discovering links between social data and real observation data from sensors.

In this work we propose a framework to tackle the aforementioned issues. Social media posts that discuss air quality are retrieved from Twitter and are automatically georeferenced by a novel model in German. The georeferenced tweets are then represented as semantic triples and linked through SPARQL queries to sensor-based measurements that are also semantically represented. The georeferencing model has been

978-1-6654-7822-9/22/$31.00 ©2022 IEEE

evaluated against related work, while the complete framework has been validated in a near-six-month scenario.

The remainder of the paper is organised as follows. In Section II we discuss relevant works that deal with the usage of social media in air quality assessment and other georeferencing techniques. Section III disassembles the proposed framework to its four core components, while Section IV presents the results of the evaluation of the georeferencing model and the validation of the framework in German tweets. Finally, Section V concludes the article and suggests some future work.

II. Related work

In the past decade several works have focused on correlating pollution-related social media posts to official data, such as in-situ sensor measurements. The authors of [2] proved a high correlation between 93 million messages from China's social media service Weibo and real particle pollution levels, while [4] demonstrated that the aforementioned platform can be used to monitor the dynamics of air pollution to some extent. More recently, [5] proposed a long short-term memory (LSTM) approach for predicting air quality, which incorporated public Weibo data and was found especially suitable for extreme short-term social events. In addition, [6] assessed the feasibility of using Twitter to monitor outdoor air pollution in London, by comparing tweets and validating them against established air monitoring stations.

To leverage the information found in crowdsourced data, a process known as *georeferencing* [7] is employed. Initially the location from where a tweet has been posted or the place names mentioned within need to be retrieved, before geolocating the post with the respective coordinates. The term used to refer to the identification of location-related entities in texts is *geoparsing* [8], with most modern systems relying [9] on a Natural Language Processing (NLP) task called Named Entity Recognition (NER) to retrieve the candidate toponyms. State-of-the-art NER approaches mainly utilise Transformer-based architectures [10] to achieve human-like results, as is evident in [11] for English and in [12] for German. The succeeding step in the process is *geolocating* the retrieved entities, by assigning a specific geographic location with the respective coordinates, as reported in [13].

Semantic web technologies are used to add meaning to the machine-processed data. The goal of semantic web is to transfer human knowledge into a machine-readable form (i.e. RDF[1]) to perform analytics. Knowledge Graph is a knowledge base that models data using graph representation to link data from various domains, generate new knowledge, and inspect recurring patterns that can be used in simulation and prediction models (i.e. using artificial intelligence and deep learning algorithms). The first step of creating a knowledge graph is representing the entities and relationships of the domain in an ontology [14]. In our study the domains are air quality assessment (e.g. [15], [16]) and social media data. Afterwards, the datasets are mapped to the ontology to generate the knowledge graph. This knowledge graph can be linked to other Linked Open Data (LOD) for further analytics. In this study, the social media data are linked to LOD about sensor real-time air quality data of a city.

[1]https://www.w3.org/RDF/

III. The proposed framework

The overall structure of the proposed framework is illustrated in Fig. 1, while its underlying components are presented in the following subsections. In short, air-quality-related tweets are retrieved with keyword-based search (Section III-A) and are enhanced with geographic information by the automated georeferencing (Section III-B). The georeferenced tweets as well as georeferenced measurements from in-situ air quality sensors are semantically represented and stored to a knowledge base (Section III-C), which allows SPARQL querying for associating social data and sensor observations (Section III-D).

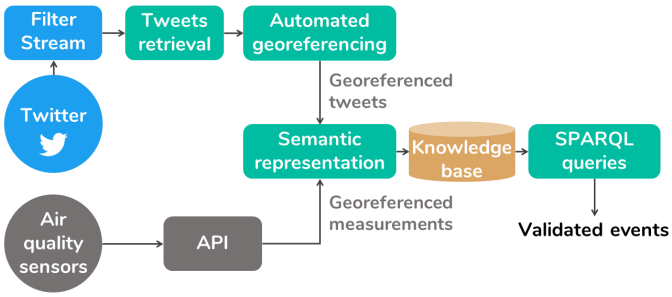

Fig. 1. The structure of the proposed framework

A. Social media data retrieval

Since Twitter's *Search Tweets* endpoints allow to search for tweets up to seven days old and we required a longer period to be examined, we have exploited the *Filter stream* endpoint instead, which allows to stream public tweets from the platform in real-time through an open connection, and utilised it continuously for almost six months (from September 6, 2021 to February 16, 2022). Out of the available operators that filter the data to be retrieved, we have selected the *keyword* operator that matches a word or phrase within the body of a tweet. The keywords that were chosen, referred to particulate matters, air pollutants, and the quality of air in general. Taking into consideration that the framework would be validated for the country of Germany, the keywords were in German and the near-six-month consuming procedure resulted in a collection of more than 29 thousand tweets. The exact number of collected tweets as well as the most mentioned keywords can be seen in Table I.

B. Automated georeferencing

With the goal of retrieving the place names mentioned in the aforementioned tweets, which is considered a sequence labeling task, a state-of-the-art NER implementation was adopted, based on the XLM-RoBERTa (XLM-R) language model, as described in [17]. The transformer-based multilingual masked language model [18] was pre-trained with texts originating from 100 languages. In order to improve performance over the

TABLE I
NUMBER OF COLLECTED AND GEOREFERENCED TWEETS AND THE MOST MENTIONED KEYWORDS

Collected	Georeferenced	Top Five Keywords (% over georeferenced) *translation*
29,132	2,948 (10.12%)	Feinstaub (70.18%) *particulate matter* Luftqualitt (12.18%) *air quality* PM10 (5.43%) gesunde Luft (4.41%) *healthy air* Schwefeldioxid (2.41%) *sulfur dioxide*

original XLM model, XLM-R drew inspiration from RoBERTa [19], in the sense that it was trained for a longer time on more data, i.e. more than 2TBs of filtered CommonCrawl data.

Since the tweets collection relates to the German language, a respective dataset was needed to support the fine-tuning process that would ultimately result in the recognition and retrieval of the mentioned locations. Thus, the well established CoNLL 2003 [20] dataset was utilised, with the available BIO-annotated labels comprising *locations*, *organisations*, *persons* and *miscellaneous* types of named entities. Although slightly better results can be obtained individually for German, a robust multilingual approach was more suitable for this framework, as it facilitates scaling with other languages.

Before feeding the collected data to the georeferencing pipeline, basic preprocessing is performed to filter out special symbols, such as "#" and "@" that are often found in tweets, and other unnecessary information, such as URLs, whose presence might impact the model's performance. XLM-R's output consists of a JSON file with lexical units labelled as locations, which subsequently get enhanced with the respective bounding boxes and exact location point, retrieved via the OpenStreetMap API. In case of multiple results for an entity, the most popular is selected. When multiple locations are present in the same tweet and a sub-region of a bigger area is involved, then the smaller, more precise bounding box/location point is retained and the larger, more general one, is discarded.

Out of the 29 thousand tweets that have been collected, circa 3 thousand have been automatically georeferenced with the above implementation (Table I). In order to allow other researchers to further dig into these data, extract more knowledge and even utilise them for other methodologies (e.g. deep learning for air quality estimation), the data is made openly available to the research community on a public GitHub repository[2]. In full compliance with Twitter's Developer Agreement and Policy[3], only the tweet IDs are released (instead of the actual tweets), together with the georeferencing outcome and the keywords based on which they have been retrieved.

[2]https://github.com/MKLab-ITI/air-quality-tweets-de
[3]https://developer.twitter.com/en/developer-terms/agreement-and-policy

C. Semantic representation

In this study we focus on two domains: the social media data and the air pollutant data collected from in-situ sensors. In order to semantically define these domains and study their relations, we developed an ontology to define the concepts and relations of these domains. The ontology, illustrated in Figure 2, represents the *Tweet* class and the *EventObservation* class, in our case, the air pollutant observations. Each *Tweet* has properties (i.e. *hasLanguage*, *hasID*, *hasDateTime*) and location that is defined with latitude and longitude. The ontology extends and reuses the GeoSPARQL Ontology[4] to describe geolocations and the Semantic Sensor Network Ontology[5] to describe sensor data representation. We mapped the German, georeferenced tweets from JSON to the *Tweets Event Ontology* and generated the knowledge graph in RDF (as the example shown in Listing 1, using RMLMapper processor[6].

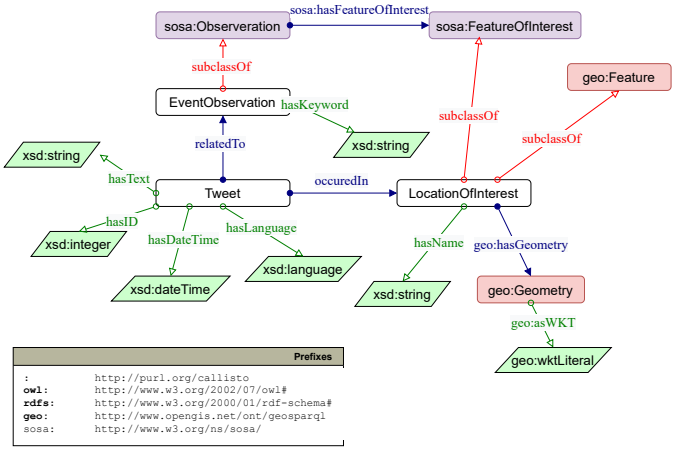

Fig. 2. Tweets Event Ontology

```
<https://purl.org/callisto/Tweet/123> a tweets:Tweet;
  tweets:hasDateTime "2021-11-21T01:13:00Z";
  tweets:hasText """ Achtung Frankfurt ! 02:12 21.11.2021
  Feinstaubwert hoch ....- Feinstaub und ...""";
  tweets:occuredIn <https://purl.org/callisto/location/123>;
  tweets:relatedTo <https://purl.org/callisto/event/11> .

<https://purl.org/callisto/event/11> a tweets:Event;
  tweets:hasKeyword "Feinstaub" .

<https://purl.org/callisto/location/123> a tweets:Location;
  geosparql:asWKT Point (8.67795,50.1249136 8.6820917,
  50.1106444);
  tweets:hasLocationName "Frankfurt, Hesse, Germany",
  "Nordend West, Innenstadt 3, Frankfurt, Hesse, Germany".
```

Listing 1. RDF representation

D. Querying

We use SPARQL[7] language to query the knowledge graphs, i.e. RDF. Listing 2 represents a query that counts the number of tweets that occurred in the cities of *Germany* and has the observations *PM10* or *particulate matter*. The query keeps

[4]http://www.opengis.net/ont/geosparql
[5]https://www.w3.org/TR/vocab-ssn/
[6]https://github.com/RMLio/rmlmapper-java
[7]https://www.w3.org/TR/rdf-sparql-query/

track of the number of times this observation has occurred in each location over time. This query is linked to the air quality sensor data to retrieve the actual measurements of PM10 in the German cities within an interval of time and compare it with the fluctuation of tweets. Table II shows a sample of the query results of tweets which were located in Frankfurt during the examined semester and address PM10 and particulate matter.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX tweets: <https://purl.org/callisto/>
prefix airquality: <https://purl.org/callisto/>

SELECT DISTINCT ?keyword ?loc_name ?dt
       (COUNT(?tweet) AS ?tweetsCount)
WHERE {
    ?tweet a tweets:Tweet.
    ?tweet tweets:hasDateTime ?time.
    ?tweet tweets:relatedTo ?event.
    ?event tweets:hasKeyword ?keyword.
    FILTER (?keyword in ("Feinstaub","PM10" )).
    ?tweet tweets:occuredIn ?location.
    ?location tweets:hasLocationName ?loc_name .
    FILTER regex(str(?loc_name), "Germany").
    BIND( (xsd:string(Month(xsd:dateTime(?time))) + '-' +
    xsd:string(YEAR(xsd:dateTime(?time))) ) AS ?dt).
    {
     #location of the tweets
      SELECT ?ob ?month ?value WHERE {
      ?station airquality:hasCode ?code.
      ?station airquality:inCity ?city.
      ?station airquality:hasObservation ?ob.
      ?ob airquality:hasMonth ?month.
      ?ob airquality:hasValue ?value.
      FILTER regex(str(?loc_name), ?city).}
     }
}
GROUP BY ?keyword ?loc_name ?dt
HAVING(COUNT(?tweet) >1)
ORDER BY ?keyword ?loc_name ASC(?dt) DESC(?tweetsCount)
```

Listing 2. SPARQL query

TABLE II
SAMPLE OF THE PARTICULATE MATTERS OBSERVATION

Location	Date	No. of Tweets
Frankfurt, Hesse, Germany	Sep-21	37
Frankfurt, Hesse, Germany	Oct-21	0
Frankfurt, Hesse, Germany	Nov-21	107
Frankfurt, Hesse, Germany	Dec-21	224
Frankfurt, Hesse, Germany	Jan-22	2
Frankfurt, Hesse, Germany	Feb-22	0

IV. Evaluation & validation

A. Georeferencing evaluation

Since this work heavily relies on the geoinformation of the social media data and a georeferencing model has been specifically trained, we decided to include in this paper its evaluation. To evaluate the performance of the chosen NER model, the metrics that were used include the *Precision*, *Recall* and *F1-score* measures. Table III provides details on the performance achieved in the NER task with the current model implementation and how it compares to other state-of-the-art approaches, as reported in the respective papers. In addition, Table IV presents how the model performed in a small, custom dataset of 50 German sentences (from the tweets collection) of our own annotation. Focus has been placed on the dedicated location class instead of the overall F1-score for all classes, since it is the main area of interest for the georeferencing tool.

TABLE III
NER RESULTS ON THE CONLL2003 (DE) DATASET

Model	F1-score
Lample et al. (2016)	78.76
Akbik et al. (2018)	88.27
mBERT	82.82
Flair	92.31
FLERT XLM-R	92.23
ACE + document-context	91.7
XLM-R-Base	84.60
XLM-R	85.81

TABLE IV
XLM-R PERFORMANCE ON CUSTOM DATASET OF 50 SENTENCES

Precision	Recall	F1-score
73.11	93.15	81.92

B. Framework validation

In order to represent the air quality sensor data and link it with the Tweets knowledge graph, we added air quality concepts to the ontology (i.e. *AirQualityStation*, *AirQualityMeasurments*, *AirQualityObservation*) to find correlations and analyse the data. The dataset from the German Federal Environment Agency (Umwelt Bundesamt[8]) contains historical data of the air quality measurements in each city of Germany in the form of CSV format. The air quality parameters are collected from several sensors distributed within the cities.

When we queried the air quality data, we retrieved the value of PM10 in the examined dates (September 2021 - February 2022), in particular the annual mean (μg/m^3) and number of daily values above 50 μg/m^3 measured (20, 4 respectively) in Frankfurt by the station Kassel Fünffensterstraße. When analysing the query results, we found in some cities the number of tweets that discusses air quality, PM10, and particulate matter increased as well as the PM10 measures were high at this date, such as in Frankfurt city in December 2021. However, deeper analysis can take place to find correlations and extract more knowledge from this data integration.

V. Conclusion & future work

In this paper we presented a framework that retrieves social media data from Twitter with German keywords about air quality, georeferences them with a state-of-the-art NER implementation and maps the dataset to RDF in order to support SPARQL queries for linking to sensor-based observations, showing how semantic web and knowledge graph can integrate data between different domains and allow discovering correlation between open datasets. The georeferencing model has been proved to perform satisfactorily, with room for

[8]https://www.umweltbundesamt.de/en/data/air/air-data

improvement, while the framework has been validated for the PM10 concentration in Frankfurt, Germany.

As future steps, more languages will be supported by the framework, while concerning the bounding box selection for each retrieved location, a more robust disambiguation strategy will be in place instead of choosing the most popular OpenStreetMap API response. In addition, more open datasets will be explored and linked to the ontologies to find correlations and provide more analytics. Finally, SPARQL queries will be enriched with GeoSPARQL functions in order to analyse data with respect to geospatial data (e.g. distance and overlap between locations).

ACKNOWLEDGMENT

This work was supported by the project CALLISTO, funded by the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 101004152.

REFERENCES

[1] P. Sicard, E. Agathokleous, A. De Marco, E. Paoletti, and V. Calatayud, "Urban population exposure to air pollution in Europe over the last decades," *Environmental Sciences Europe*, vol. 33, no. 1, pp. 1–12, 2021.

[2] S. Wang, M. J. Paul, and M. Dredze, "Social media as a sensor of air quality and public response in China," *Journal of medical Internet research*, vol. 17, no. 3, p. e22, 2015.

[3] A. Kruspe, M. Häberle, E. J. Hoffmann, S. Rode-Hasinger, K. Abdulahhad, and X. X. Zhu, "Changes in Twitter geolocations: Insights and suggestions for future usage," *arXiv preprint arXiv:2108.12251*, 2021.

[4] W. Jiang, Y. Wang, M.-H. Tsou, and X. Fu, "Using social media to detect outdoor air pollution and monitor air quality index (AQI): a geotargeted spatiotemporal analysis framework with Sina Weibo (Chinese Twitter)," *PloS one*, vol. 10, no. 10, p. e0141185, 2015.

[5] W. Zhai and C. Cheng, "A long short-term memory approach to predicting air quality based on social media data," *Atmospheric Environment*, vol. 237, p. 117411, 2020.

[6] Y. Hswen, Q. Qin, J. S. Brownstein, and J. B. Hawkins, "Feasibility of using social media to monitor outdoor air pollution in London, England," *Preventive medicine*, vol. 121, pp. 86–93, 2019.

[7] E. Amitay, N. Har'El, R. Sivan, and A. Soffer, "Web-a-where: geotagging web content," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 2004, pp. 273–280.

[8] J. Gelernter and S. Balaji, "An algorithm for local geoparsing of microtext," *GeoInformatica*, vol. 17, no. 4, pp. 635–667, 2013.

[9] M. Gritta, M. T. Pilehvar, and N. Collier, "A pragmatic guide to geoparsing evaluation," *Language resources and evaluation*, vol. 54, no. 3, pp. 683–712, 2020.

[10] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 1, pp. 50–70, 2020.

[11] X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, and J. Li, "Dice loss for data-imbalanced NLP tasks," *arXiv preprint arXiv:1911.02855*, 2019.

[12] J. Zöllner, K. Sperfeld, C. Wick, and R. Labahn, "Optimizing small BERTs Trained for German NER," *Information*, vol. 12, no. 11, 2021. [Online]. Available: https://www.mdpi.com/2078-2489/12/11/443

[13] S. Andreadis, G. Antzoulatos, T. Mavropoulos, P. Giannakeris, G. Tzionis, N. Pantelidis, K. Ioannidis, A. Karakostas, I. Gialampoukidis, S. Vrochidis *et al.*, "A social media analytics platform visualising the spread of COVID-19 in Italy via exploitation of automatically geotagged tweets," *Online Social Networks and Media*, vol. 23, p. 100134, 2021.

[14] H. S. Pinto and J. P. Martins, "Ontologies: How can They be Built?" *Knowledge and Information Systems*, vol. 6, no. 4, pp. 441–464, 2004.

[15] J. Cuenca, F. Larrinaga, and E. Curry, "DABGEO: A reusable and usable global energy ontology for the energy domain," *Journal of Web Semantics*, vol. 61-62, p. 100550, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1570826820300020

[16] M. M. Oprea, "AIR_POLLUTION_Onto: an Ontology for Air Pollution Analysis and Control," in *Artificial Intelligence Applications and Innovations III*, Iliadis, Maglogiann, Tsoumakasis, Vlahavas, and Bramer, Eds. Boston, MA: Springer US, 2009, pp. 135–143.

[17] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.

[18] A. Conneau and G. Lample, "Cross-lingual language model pretraining," *Advances in neural information processing systems*, vol. 32, 2019.

[19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[20] E. F. Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," *arXiv preprint cs/0306050*, 2003.