gesis Leibniz Institute for the Social Sciences

hhu Heinrich Heine Universität Düsseldorf

Geotagging TweetsCOV19: Enriching a COVID-19 Twitter Discourse Knowledge Base with Geographic Information

Dimitar Dimitrov[1], Dennis Segeth[2,*], Stefan Dietze[1,2]

Workshop on Knowledge Graphs for Online Discourse Analysis (KnOD 2022)

April 13, 2022, Lyon, France

[1]GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany

[2]Heinrich-Heine-University Düsseldorf, Germany

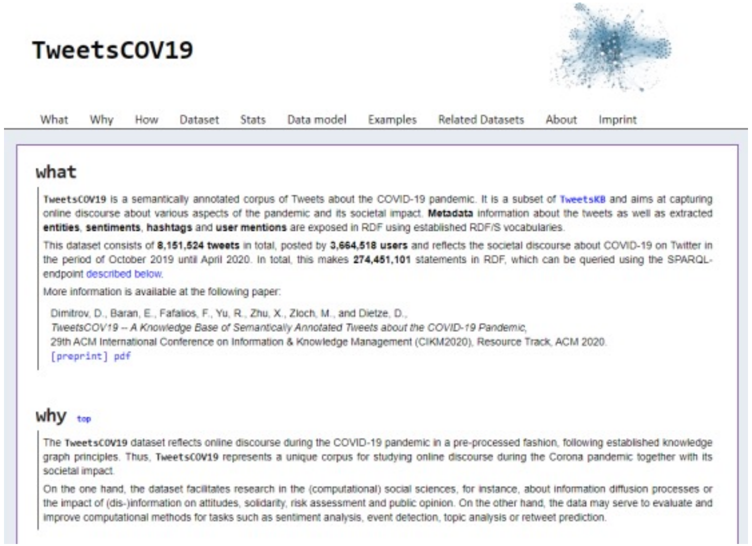
*Work conducted as part of a bachelor thesis

Leibniz Association

@trovdimi

TweetsCOV19

- Public RDF corpus of anonymized COVID-19-related tweets
- Spanning period: October 2019 – April 2020
- More than 8 milion original tweets in English
- Posted by more than 3,6 milion users
- 268 COVID-19 related keywords
- Pre-computed features:
  - Entity extraction and linking (Blanco et al., 2015)
  - Sentiment analysis (Thelwall et al., 2017)
- Dataset is available as N3 and TSV files registered with Zenodo[1]
- Everything about TweetsCOV19 at https://data.gesis.org/tweetscov19



TweetsCOV19

What Why How Dataset Stats Data model Examples Related Datasets About Imprint

what

TweetsCOV19 is a semantically annotated corpus of Tweets about the COVID-19 pandemic. It is a subset of TweetsKB and aims at capturing online discourse about various aspects of the pandemic and its societal impact. **Metadata** information about the tweets as well as extracted **entities**, **sentiments**, **hashtags** and **user mentions** are exposed in RDF using established RDF/S vocabularies.

This dataset consists of **8,151,524 tweets** in total, posted by **3,664,518 users** and reflects the societal discourse about COVID-19 on Twitter in the period of October 2019 until April 2020. In total, this makes **274,451,101** statements in RDF, which can be queried using the SPARQL-endpoint described below.

More information is available at the following paper:

Dimitrov, D., Baran, E., Fafalios, F., Yu, R., Zhu, X., Zloch, M., and Dietze, D.,
TweetsCOV19 -- A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic,
29th ACM International Conference on Information & Knowledge Management (CIKM2020), Resource Track, ACM 2020.
[preprint] pdf

why top

The **TweetsCOV19** dataset reflects online discourse during the COVID-19 pandemic in a pre-processed fashion, following established knowledge graph principles. Thus, **TweetsCOV19** represents a unique corpus for studying online discourse during the Corona pandemic together with its societal impact.

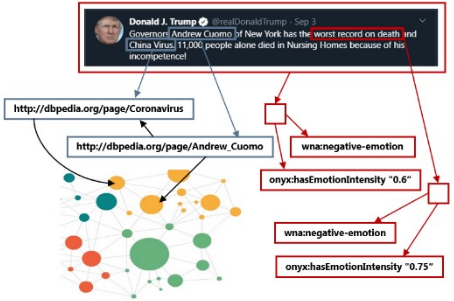
On the one hand, the dataset facilitates research in the (computational) social sciences, for instance, about information diffusion processes or the impact of (dis-)information on attitudes, solidarity, risk assessment and public opinion. On the other hand, the data may serve to evaluate and improve computational methods for tasks such as sentiment analysis, event detection, topic analysis or retweet prediction.

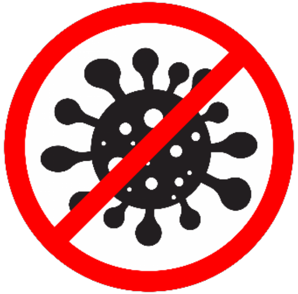
[1] Erdal Baran, & Dimitar Dimitrov. (2020). TweetsCOV19 - A Semantically Annotated Corpus of Tweets About the COVID-19 Pandemic [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3871753

gesis Leibniz Institute for the Social Sciences

Most Research Requires Geotagging

- Interdisciplinary research
  - Discourse Data for Policy (DD4P)
  - Solidarity in the COVID-19 pandemic (SAFE19)
- Spreading of diseases(Sloan et al., 2013)
- Earthquake detection (Sakaki et al. 2010)
- Deriving demographic characteristics (Sloan et al., 2013)





Goal: Enriching knowledge bases with geolocation information

This Work: Status Quo and Problem

- **Status quo** of geotagging
  - Only 1% of tweets are geotagged (Sloan et al., 2013)
  - Variety of pre-trained geotagging models (Lau et al., 2017), (Rahimi et al.,2015) and many others
  - Vocabulary shifts and training data freshness issues (Hombaiaha et al., 2021)

- **RQ:** How do established pre-trained geotagging models perform *compared* to models trained using fresh data, i.e., COVID-19 discourse data?

Approach and Experiments

- Extracting geolocation data from TweetsCOV19
- Geotagging algorithms (DeepGeo vs. GeoLocation)
- Evaluation metric
- Experiment 1: Vocabulary shifts and training data freshness
  - Model accuracy per error distance
  - Influence of tweet length
- Experiment 2: Geo-coverage for TweetsCOV19
  - Unique cities and countries
  - Number of tweets per country

gesis Leibniz Institute for the Social Sciences

Extracting Geolocation Data from TweetsCOV19

- 229,045 tweets from 147.902 unique users
  - 11,311 tweets with populated „geo“ metadata field
  - 217,734 tweets with populated „place“ metadata field
- Dataset is available as a TSV file registered with Zenodo[2]
- Each line contains tweet ID, latitude, longitude, country, state, county, city information

TweetsCOV19 - Geolocation Data

tweetID	latitude	longitude	country	state	county	city
1178823685077118978	34.687331	-82.434848	United States	South Carolina	Anderson County	Piedmont
1178995114640891904	33.841705	-84.487242	United States	Georgia	Cobb County	Vinings
1179019429792899073	28.156842	77.149786	India	Haryana	Gurgaon	Sohna
1179069332858572805	34.271183	-91.351087	United States	Arkansas	Arkansas County	De Witt
1179139369346764800	52.381063	-2.033651	United Kingdom	England	Worcestershire	Barnt Green
1179089789359812608	53.303584	-115.118937	Canada	Alberta		Drayton Valley
1179105986881216512	40.3164361	-79.985697	United States	Pennsylvania	Allegheny County	South Park Township

“geo” – JSON example


```
"geo": {
    "type": "Point",
    "coordinates": [45.4643, 9.1897]
},
```


“place” – JSON example


```
"place": {
    "id": "8eb7d0abedc4817b",
    "url": "https://api.twitter.com/1.1/geo/id/8eb7d0abedc4817b.json",
    "place_type": "city",
    "name": "Greenville",
    "full_name": "Greenville, SC",
    "country_code": "US",
    "country": "United States",
    "contained_within": [],
    "bounding_box": {
        "type": "Polygon",
        "coordinates": [[[-82.434848, 34.687331], [-82.249689, 34.687331],
                         [-82.249689, 34.904552], [-82.434848, 34.904552]]]
    },
    "attributes": {}
},
```


[2]Segeth, Dennis, & Dimitrov, Dimitar. (2021). TweetsCOV19 - Geolocation Data (Part 1, October 2019 - April 2020) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.4986365

Geotagging Algorithms

DeepGeo (Lau et al., 2017)

- DeepGeo predicts the **tweet location**
- DeepGeo is a tweet text-based approach
- Accepts specific attributes from the metadata, i.e., “tweet creation time”, ”account creation time”, “UTC offset”, “timezone”, "location“
- Comes with 12 pre-trained models
- DeepGeo + Noise adds Gaussian noise to sharpen the activation values

GeoLocation (Rahimi et al.,2015)

- GeoLocation predicts the user’s **home location**
- GeoLocationLR: tweet text-based approach
- GeoLocationLP : social network approach
    - unidirected mentions (@user)
- GeoLocation Hybrid: combines GeoLocation LR and LP
    - removed “celebrity“ nodes

gesis Leibniz Institute for the Social Sciences

Evaluation Metric

- *Acc@d* - percentage of predictions with an *error distance (ED)* $\leq d$
- *ED* is the distance in kilometer between the predicted and the true geocoordinates
- *Acc@161*km (~100milles) commonly used (Zhiyuan et al., 2010)
- We experiment with $d \in \{25, 50, 100, 161\}$ km
- To make DeepGeo and GeoLocation comparable, we assign the **predicted user home location** to all user's tweets

$$Acc@d = \frac{|\{s \in S : ED(s) \leq d\}|}{|S|}$$

$$ED(s) = distance(X(s), X^*(s))$$

gesis Leibniz Institute for the Social Sciences

Results: Accuracy per error distance

Model	Prediction Type	Acc@25	Acc@50	Acc@100	Acc@161
DeepGeo TweetsCOV19	Tweet location	12.93	15.2	17.36	18.37
DeepGeo Pre-trained	Tweet location	30.31	45.34	52.63	**55.91**
DeepGeo + Noise TweetsCOV19	Tweet location	**37.05**	42.06	45.66	47.94
DeepGeo + Noise Pre-trained	Tweet location	30.32	45.42	52.33	55.50
GeoLoc LR TweetsCOV19	Home location	2.85	3.71	4.64	5.69
GeoLoc LR Pre-trained	Home location	5.46	7.77	9.81	11.07
GeoLoc LP TweetsCOV19	Home location	1.96	2.66	2.95	3.34
GeoLoc LP Pre-trained	Home location	2.53	3.68	4.64	5.49
GeoLoc Hybrid TweetsCOV19	Home location	5.16	6.64	8.07	9.63
GeoLoc Hybrid Pre-trained	Home location	6.89	9.77	12.28	13.83

Finding: Pre-trained models achieve solid results for Acc@161 while "fresh" ground truth can improve accuracy at Acc@25

gesis Leibniz Institute for the Social Sciences

Results: Influence of tweet length

Model	Prediction Type	short	medium	long
DeepGeo TweetsCOV19	Tweet location	17.71	18.25	**19.13**
DeepGeo Pre-trained	Tweet location	52.02	**58.08**	57.51
DeepGeo + Noise TweetsCOV19	Tweet location	44.78	49.04	**49.88**
DeepGeo + Noise Pre-trained	Tweet location	51.62	**57.55**	57.18
GeoLoc LR TweetsCOV19	Home location	2.73	5.68	**8.01**
GeoLoc LR Pre-trained	Home location	6.65	12.13	**13.51**
GeoLoc LP TweetsCOV19	Home location	0.85	3.62	**5.74**
GeoLoc LP Pre-trained	Home location	3.52	5.92	**6.63**
GeoLoc Hybrid TweetsCOV19	Home location	6.22	10.37	**11.59**
GeoLoc Hybrid Pre-trained	Home location	9.16	14.93	**16.44**

Finding: With small exceptions, longer tweets are easier to geotag

Geo-coverage for TweetsCOV19

- Unique countries and cities (pre-trained)

	DeepGeo	DeepGeo+Noise	GeoLoc LR	GeoLoc LP	GeoLoc Hybrid
Countries	166	166	77	184	184
Cities	2564	2519	741	9165	8434

Finding: GeoLoc Hybrid exhibits the highest number of unique cities and countries

- Number of tweets per country (pre-trained)

# of Tweets	DeepGeo	DeepGeo+Noise	GeoLoc LR	GeoLocLP	GeoLoc Hybrid
France	21K	20K	15.7K	18.4K	29.2K
Germany	28K	28K	21.9K	3K	23.4K
India	444K	446K	385.5K	263.8K	313.3K
Italy	21K	33K	23.6K	5K	27.6K
United Kingdom	1.44M	1.25M	1.09M	411.3K	1.02M
United States	3.14M	3.23M	3.28M	5.04M	3.37M

Finding: GeoLocLP assigns predominantly geolocations in the US and "misses" cities in Germany and Italy

gesis Leibniz Institute for the Social Sciences

Summary: Our Results

1. Language changes faster than locations change their names
2. Fresh ground truth can improve Acc@25 (city-level)
3. DeepGeo outperforms GeoLocation in terms of Acc@d
4. GeoLocation(Hybrid) shows the highest geographic coverage

Take away: Methods and training data-based biases must be stated when enriching knowledge bases

Ethics: Geotagging can violate user privacy!

gesis Leibniz Institute for the Social Sciences

Questions?

Dimitar Dimitrov
GESIS – Leibniz Institute for the Social Sciences
@trovdimi

Dennis Segeth
Heinrich Heine University

Stefan Dietze
GESIS – Leibniz Institute for the Social Sciences & Heinrich Heine University

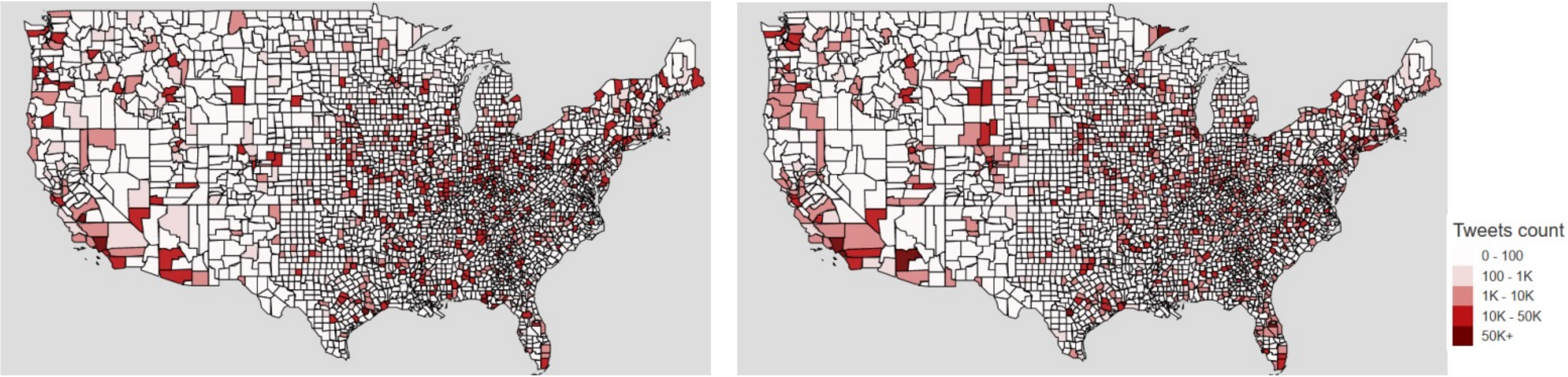
Thank you!

E-Mail: dimitar.dimitrov@gesis.org

Data: https://zenodo.org/record/4986365

gesis Leibniz Institute for the Social Sciences

TweetsCOV19: USA County-level Coverage

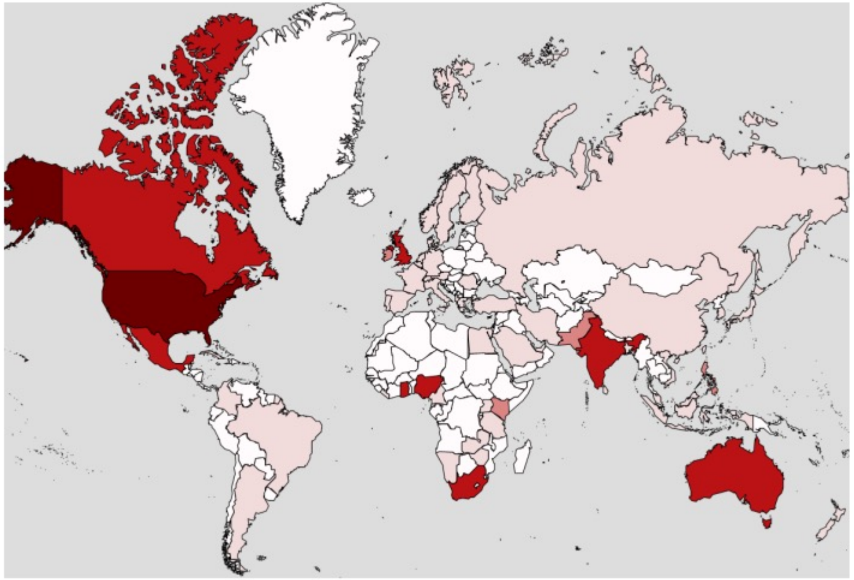


DeepGeo+Noise TweetsCOV19

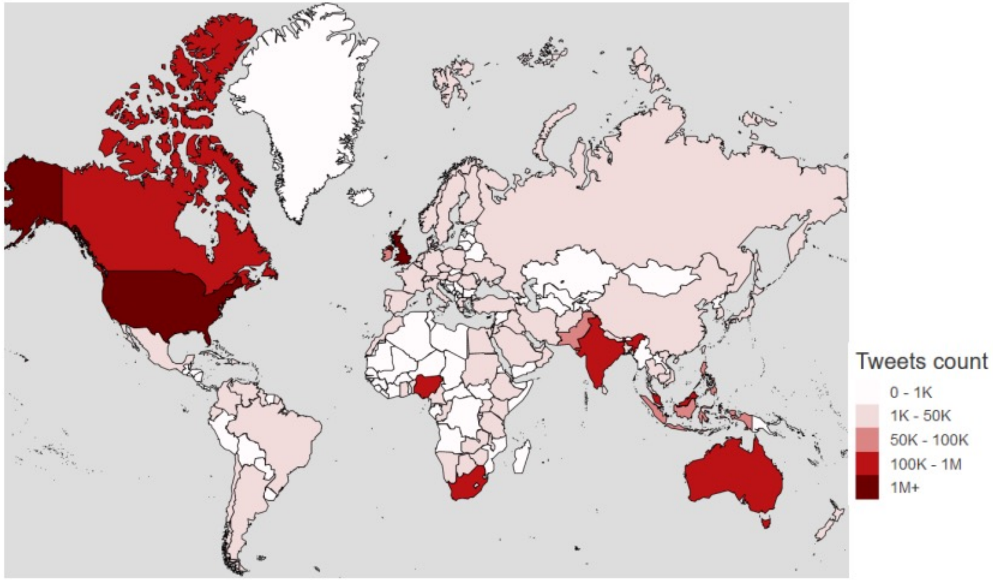
DeepGeo+Noise Pre-trained

gesis Leibniz Institute for the Social Sciences

TweetsCOV19: Global Coverage



DeepGeo+Noise TweetsCOV19



DeepGeo+Noise Pre-trained