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

Dimitar Dimitrov
dimitar.dimitrov@gesis.org
GESIS - Leibniz Institute for the Social Sciences
Cologne, Germany

Dennis Segeth
segeth.dennis@outlook.de
Heinrich Heine University
Düsseldorf, Germany

Stefan Dietze
stefan.dietze@gesis.org
GESIS - Leibniz Institute for the Social Sciences
Cologne, Germany
Heinrich Heine University
Düsseldorf, Germany

ABSTRACT

Various aspects of the recent COVID-19 outbreak have been extensively discussed on online social media platforms and, in particular, on Twitter. Geotagging COVID-19-related discourse data on Twitter is essential for understanding the different discourse facets and their regional relevance, including calls for social distancing, acceptance of measures implemented to contain virus spread, anti-vaccination campaigns, and misinformation. In this paper, we aim at enriching TweetsCOV19—a large COVID-19 discourse knowledge base of more than 20 million tweets—with geographic information. For this purpose, we evaluate two state-of-the-art Geotagging algorithms: (1) DeepGeo—predicting the tweet location and (2) GeoLocation—predicting the user location. We compare pre-trained models with models trained on context-specific ground truth geolocation data extracted from TweetsCOV19. Models trained on our context-specific data achieve more than 6.7% improvement in Acc@25 compared to the pre-trained models. Further, our results show that DeepGeo outperforms GeoLocation and that longer tweets are, in general, easier to geotag. Finally, we use the two geotagging methods to study the distribution of tweets per country in TweetsCOV19 and compare the geographic coverage, *i.e.*, the number of countries and cities each algorithm can detect.

CCS CONCEPTS

• **Information systems** → *Data extraction and integration.*

KEYWORDS

geotagging, evaluation, COVID-19, Twitter, discourse

ACM Reference Format:
Dimitar Dimitrov, Dennis Segeth, and Stefan Dietze. 2022. Geotagging TweetsCOV19: Enriching a COVID-19 Twitter Discourse Knowledge Base with Geographic Information. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3487553.3524623

WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9130-6/22/04…$15.00
https://doi.org/10.1145/3487553.3524623

1 INTRODUCTION

The World Health Organization (WHO) officially declared the Coronavirus disease 2019 (COVID-19) a pandemic on March 11, 2020 [17]. At this point, more than 118.000 cases had been officially confirmed, with a total of 4291 deaths spread over 114 countries. Spreading of the virus most commonly occurs from person to person during close contact [2]. The COVID-19 Twitter discourse has proved crucial to developing an understanding of the impact of the pandemic, the usefulness of implemented measures, societal attitudes, and perceptions in this context. Furthermore, understanding the COVID-19-related discourse, its evolution and interdependence between public opinion and relevant political actions, media events or scientific discoveries has been perceived as valuable and led to the collection of a number of different COVID-19 discourse datasets [7, 12, 14]. Specifically, TweetsCOV19 represents a knowledge base of COVID-19-related online discourse, covering information about more than 20 million tweets archived between October 2019 to December 2020 [7]. Metadata information about the tweets as well as extracted entities, sentiments, hashtags and user mentions are exposed in RDF using established RDF/S vocabularies, in order to provide an easy-to-reuse knowledge base of COVID-related online discourse.

One major shortcoming of Twitter discourse data is the very low percentage of geotagged tweets, as only 1% of tweets are estimated to be geotagged [16]. Geotagging Twitter data and user behavior to specific geographic regions can help not only to implement measures to fight the pandemic on the local level but also to gain valuable insights on the effectiveness of implemented measures retrospectively. Geotagging algorithms can predict one of three different types of locations, *i.e.*, the *home location* of a user, which represents the residential place, the *tweet location*, and the *mentioned location*. Based on the used approach, geotagging algorithms can be divided into text-based [3, 4, 9, 10, 13], network-based [5, 6, 11], and hybrid approaches [15, 18].

Problem and objectives. While state-of-the-art geotagging models tend to be pre-trained on prior Twitter discourse, discourse and used vocabulary have drastically changed since the start of the COVID-19 pandemic, introducing a significant vocabulary shift. In this paper, we aim at studying the performance of established geotagging solutions on COVID-19-related Twitter discourse.

Data and approach. More specifically, we compare the performance of two geotagging algorithms DeepGeo [13] and GeoLocation [15] using Acc@d (*cf.* Section 2). To study the effects of vocabulary change, we compare the performance of pre-trained

models and models trained on context-specific ground truth data extracted from TweetsCOV19. We also study the effect of tweet text length on the model performance. Finally, we measure the country and city coverage of the models and the tweet volume from a country-level perspective (*cf.* Section 3). To visualize differences between the maps different models produce, we use Reverse Geocoder[1] to map latitude-longitude coordinates to the nearest town/city.

Contributions and findings. While it is evident that language models used for NLP/NLU tasks benefit significantly from frequent model updates [1], we show that also geotagging model performance can be improved considerably through training on context-specific data. More specifically, models trained on recent context-specific discourse data achieve more than 6.7% improvement in Acc@25 compared to the pre-trained models, underlining the need for training and tuning geotagging models towards the specific use case and context at hand, i.e., knowledge graph enrichment. One could assume that training on a large English Twitter data corpus is sufficient. However, our results suggest that different corpora have very different characteristics, and in the geolocation case at hand, COVID-19 discourse is different from non-COVID-19 or pre-COVID-19 discourse, requiring geolocation models to be trained on COVID-19 discourse data. To facilitate the development of COVID-19-specific geolocation models, we extract and publish geolocation ground truth data from TweetsCOV19. The full dataset is available through the Zenodo data repository (DOI:10.5281/zenodo.4986365)[2].

2 EVALUATION OF GEOTAGGING ALGORITHMS FOR ENRICHING TWEETSCOV19

Although there are many geotagging algorithms that can be used to enrich a knowledge base such as TweetsCOV19, in this work, we focus on DeepGeo and GeoLocation as these algorithms provide pre-trained models that can be used almost out of the box. Further, these algorithms represent established approaches for predicting tweet location and user home location. Additionally, the selected approaches are methodologically diverse, *i.e.*, DeepGeo follows a neural network approach while GeoLocation uses Logistic Regression (LR) trained on tweet text, a Label Propagation (LP) over a @-mention network and a hybrid solution. Next, we describe our experimental setup, including ground truth creation, training procedure and the evaluation metric . We conduct two experiments: (i) we study the accuracy of the models at different distances and (ii) the influence of the tweet text.

Experimental Setup. DeepGeo [13] combines the tweet text and tweet metadata to predict the tweet location. DeepGeo processes the following features: tweet text, tweet time, user UTC offset, user timezone, user account creation time and self-reported user location. Overall there are 12 pre-trained models of which we will use two (i) DeepGeo and (ii) DeepGeo + Noise. DeepGeo + Noise adds Gaussian noise to sharpen the activation values in order to counteract the random noise. GeoLocation [15] predicts the user home location in the form of coordinates. We will use three different versions of GeoLocation: (i) GeoLoc LR uses only the tweet content. (ii) GeoLoc LP follows a social network approach. An edge between two users in the network corresponds to an @-mention. The network is undirected and edges have weights corresponding to the number of mentions. For isolated users, the median of all coordinates in the training data is used as predictions. (iii) GeoLoc Hybrid combines GeoLoc LR and GeoLoc LP. In addition, for GeoLoc Hybrid, we use the "remove celebrity" feature to ignore nodes in the graph with a weight of 15 or more in order to increase the accuracy. The intuition for using this feature is to relax the assumption that mentioning a user indicates geographic proximity, which does not necessarily hold for celebrity nodes as every user independent of her location may mention, *e.g.*, Donald Trump. GeoLocation training examples are concatenations of all tweets from a user and not individual tweets as for DeepGeo.

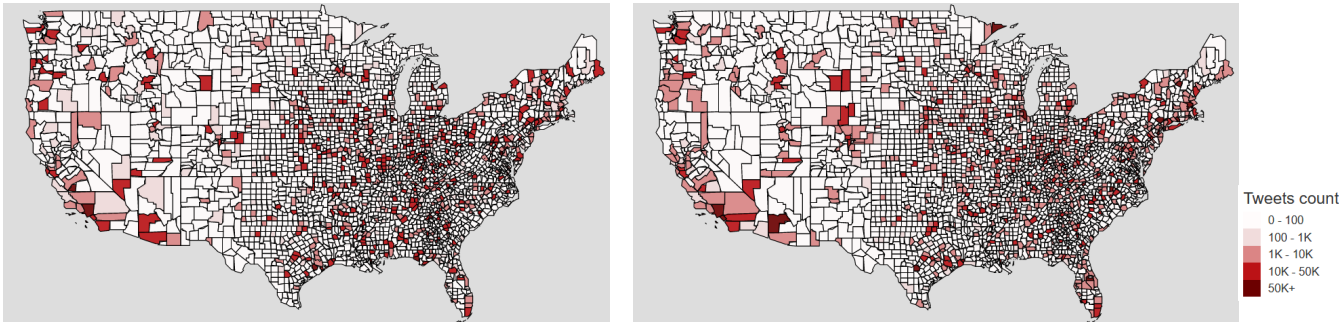
We compare the pre-trained DeepGeo and GeoLocation models with their versions trained on TweetsCOV19 data using the models' hyper-parameters as suggested by the authors. To this end, we hydrated the TweetsCOV19 dataset and collected the text and metadata of a total of 6.8 million tweets. 1.2 million tweets were no longer available for hydration. The resulting loss is equally distributed over the month. In total, we extracted 229K tweets, 11K with point coordinates in the "geo" metadata field, and 217K with polygon coordinates in the "place" metadata field that can serve as ground through data. To ensure a fair comparison between the algorithms, we trained them using the same training and test data. For that purpose, we first grouped all tweets by user and split them randomly, with 80% of the tweets for training and 20% for testing. While grouping the data by user is not critical for training DeepGeo, this ensures no user "leakage" when training GeoLocation. The context-specific and the pre-trained models are evaluated on the test data from TweetsCOV19. To compare the algorithms, we use Acc@d, determines the percentage of predictions with an error distance less then d and is defined as $Acc@d = \frac{|\{s \in S : ED(s) \leq d\}|}{|S|}$. Hereby, the error distance is defined as $ED(s) = distance(X(s), X^*(s))$. While we provide results for $d \in \{25, 50, 100, 161\}$ kilometers, it is common to choose 161 kilometers (100 miles) in order to capture so-called near misses [4]. As GeoLoc predicts the home location of a user, we assign this prediction to all user's tweets to make GeoLoc comparable to DeepGeo that predicts a location at the tweets-level.

Accuracy per error distance. In Table 1 we see that, as expected, the accuracy increases with increasing d independent of the model and data used for training. The highest Acc@161 (55.91%) is achieved by the pre-trained version of DeepGeo + Noise. Interestingly, for Acc@25, DeepGeo + Noise—the model trained on context-specific ground truth data from TweetsCOV19—predicted 37.05% of the tweets correctly, outperforming the pre-trained DeepGeo models by more than 6.7%. This finding suggests that using context-specific ground truth data for training DeepGeo produces the most reliable results for an analysis on a city level—a property particularly useful for identifying locations where people are concerned about COVID-19, allowing for subsequent reasoning. In Figure 1 we see how tweet distributions differ for USA at county-level for pre-trained models and for models trained using TweetsCOV19 data. We also observe

[1] https://pypi.org/project/reverse_geocoder
[2] https://zenodo.org/record/4986365

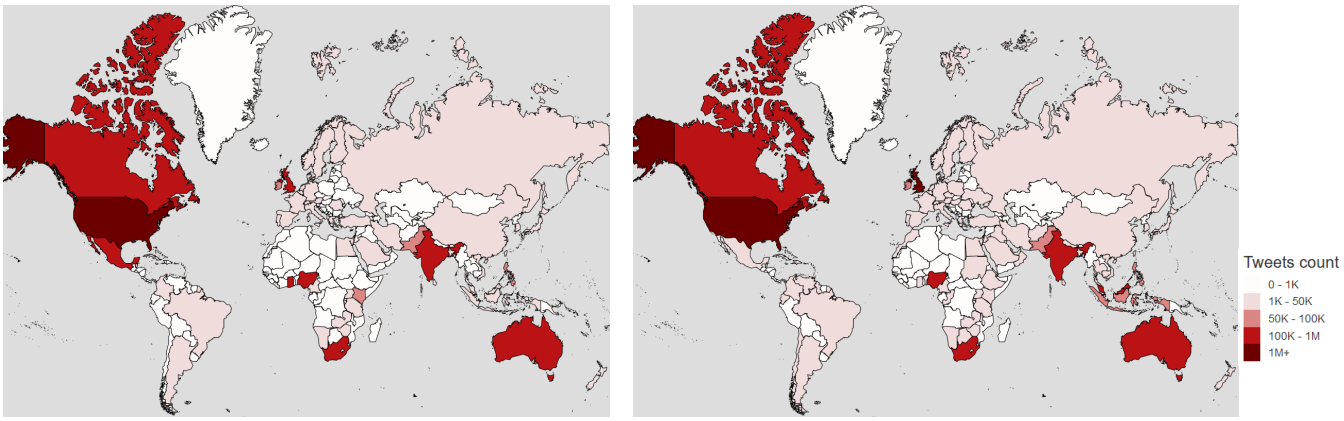
Model	Prediction Type	Acc@25	Acc@50	Acc@100	Acc@161
DeepGeo TweetsCOV19	Tweet location	12.93	15.2	17.36	18.37
DeepGeo Pre-trained	Tweet location	30.31	45.34	52.63	**55.91**
DeepGeo + Noise TweetsCOV19	Tweet location	**37.05**	42.06	45.66	47.94
DeepGeo + Noise Pre-trained	Tweet location	30.32	45.42	52.33	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

Table 1: Accuracy for different d values.



(a) DeepGeo+Noise TweetsCOV19 **(b) DeepGeo+Noise Pre-trained**

Figure 1: USA tweet distribution at county for (a) DeepGeo+Noise TweetsCOV19 and (b) DeepGeo+Noise Pre-trained.



(a) DeepGeo+Noise TweetsCOV19 **(b) DeepGeo+Noise Pre-trained**

Figure 2: Global coverage of TweetsCOV19 for (a) DeepGeo+Noise TweetsCOV19 and (b) DeepGeo+Noise Pre-trained.

that all pre-trained GeoLocation models show higher accuracy than the models trained on context-specific ground truth data.

Accuracy per tweet length. Since our geotagging algorithms process the tweet texts, we want to examine the influence of the text length on the accuracy of pre-trained models and models trained with context-specific data. We categorized the tweets into three categories: "short" for tweets up to 120 characters, "medium" for tweets from 121 to 230 characters and "long" for tweets from 131 to 280 characters. The number of characters per category was selected so that each group has nearly the same number of tweets. For the GeoLocation models, the average tweet length per user was calculated and used to split the users in the respective category. In Table 2, we see that in general higher accuracy values are achieved for longer tweets regardless of the model and training data used. There are only two exceptions for the pre-trained versions of DeepGeo for which the tweets of medium length score higher.

3 GEOTAGGING TWEETSCOV19

Apart from insights from TweetsCOV19 that can help fight the virus at the city-level, the country-level perspective captured in

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**

Table 2: Acc@161 per tweet text length.

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

Table 3: Number of unique countries and cities in TweetsCOV19 for pre-trained model version.

# 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

Table 4: Number of tweets in TweetsCOV19 for selected countries for pre-trained model version.

TweetsCOV19 is also interesting. Based on our accuracy analysis, the most suitable models for this purpose are the pre-trained DeepGeo and GeoLocation versions as these perform better in less fine-grain geotagging scenarios, suitable to distinction at country-level. In Figure 2, we see the global tweet distribution of the TweetsCOV19 dataset. DeepGeo+Noise and GeoLoc Hybrid produce visually similar distributions and assign tweets and user locations mainly to countries using English as an official language, which is not surprising since TweetsCOV19 contains only tweets written in English. Table 3 shows the number of unique cities and countries for all model versions. While for GeoLoc LR we have low numbers of cities and countries compared to DeepGeo and DeepGeo+Noise, for GeoLocLP and GeoLoc Hybrid we see a much higher number of unique cities and comparable number of unique countries. GeoLocLP predicted over 5 million of 6.8 million tweets to the US (*cf.* Table 4), which can be due to a high amount of isolated users in the social network graph. Notably, GeoLoc LP has very low number of tweets for countries such as Italy and Germany, resulting in a low number of user mentions from these countries within the training dataset. Other geolocation models show similar distribution results for the selected countries.

4 DISCUSSION AND CONCLUSION

This work presented an evaluation of DeepGeo and GeoLocation geotagging algorithms with the goal of enriching TweetsCOV19—a COVID-19-related Twitter knowledge base—with geographic information. Overall, DeepGeo outperforms GeoLocation. This suggests that predicting user home location is the more difficult task, probably due to the inherent user mobility. This result is consistent for the pre-trained versions and for the TweetsCOV19 version of the models, which rules out performance differences due to better training data alone for the pre-trained variants. DeepGeo + Noise is able to achieve more than 6.7% improvement for Acc@25 when trained using TweetsCOV19 ground truth data and presents a promising solution for city-level analysis of COVID-19-related Twitter discourse data. To gain country-level insights about the COVID-19 discourse in TweetsCOV19, we used the pre-trained versions of DeepGeo and GeoLocation as they showed better Acc@161. Most of the tweets in TweetCOV19 were geomapped to countries with English as an official language, such as the United Kingdom, USA and India. In terms of detected countries and cities, GeoLoc Hybrid showed the highest coverage. The observed differences in the coverage may be explained through the different corpora used for pre-training DeepGeo and GeoLocation. Such method and training data-based biases present crucial points to be considered when geotagging a specific dataset. The insights from our work are relevant not only for geotagging COVID-19 discourse Twitter data but also for geotagging larger and more diverse corpora such as TweetsKB [8]. As future work, we plan to extend our evaluation to measure region-specific performance, geospatial biases in the country vs. city coverage for the TweetsCOV19 model versions, the geotagging runtime and incorporate mentioned location geotagging approaches. We hope that our analyses help the further development of context-specific models.

REFERENCES

[1] Spurthi Amba Hombaiah, Tao Chen, Mingyang Zhang, Michael Bendersky, and Marc Najork. 2021. Dynamic Language Models for Continuously Evolving Content. (2021), 2514–2524.

[2] CDC. 2020. *How COVID-19 Spreads.* https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/how-covid-spreads.html

[3] Swarup Chandra, Latifur Khan, and Fahad Bin Muhaya. 2011. Estimating Twitter User Location Using Social Interactions–A Content Based Approach. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. 838–843. https://doi.org/10.1109/PASSAT/SocialCom.2011.120

[4] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. Association for Computational Linguistics, Toronto, Canada, 759–768. https://dl.acm.org/doi/10.1145/1871437.1871535

[5] Ryan Compton, David Jurgens, and David Allen. 2014. Geotagging one hundred million Twitter accounts with total variation minimization. In *2014 IEEE International Conference on Big Data (Big Data)*. 393–401. https://doi.org/10.1109/BigData.2014.7004256

[6] Clodoveu A Davis Jr, Gisele L Pappa, Diogo Rennó Rocha De Oliveira, and Filipe de L. Arcanjo. 2011. Inferring the location of twitter messages based on user relationships. *Transactions in GIS* 15, 6 (2011), 735–751.

[7] Dimitar Dimitrov, Erdal baran, Pavlos Fafalios, Ran Yu, Xiaofei Zhu, Matthäus Zloch, and Stefan Dietze. 2020. TweetsCOV19 - A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*. Association for Computing Machinery, Ireland, 2991–2998. https://dl.acm.org/doi/10.1145/3340531.3412765

[8] Pavlos Fafalios, Vasileios Iosifidis, Eirini Ntoutsi, and Stefan Dietze. 2018. TweetsKB: A Public and Large-Scale RDF Corpus of Annotated Tweets. In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings (Lecture Notes in Computer Science, Vol. 10843)*. Springer, 177–190. https://doi.org/10.1007/978-3-319-93417-4_12

[9] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H Chi. 2011. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 237–246.

[10] Mans Hulden, Miikka Silfverberg, and Jerid Francom. 2015. Kernel Density Estimation for Text-Based Geolocation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (Austin, Texas). AAAI Press, 145–150.

[11] David Jurgens. 2013. That's what friends are for: Inferring location in online social media platforms based on social relationships. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 7.

[12] Rabindra Lamsal. 2020. Design and analysis of a large-scale COVID-19 tweets dataset. *Applied Intelligence* (2020).

[13] Jey Han Lau, Lianhua Chi, Khoi-Nguyen Tran, and Trevor Cohn. 2017. End-to-end Network for Twitter Geolocation Prediction and Hashing. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Asian Federation of Natural Language Processing, Taipei, Taiwan, 744–753. https://www.aclweb.org/anthology/I17-1075/

[14] Umair Qazi, Muhammad Imran, Ferda Ofli, and Filipe Arcanjo. 2020. GeoCoV19: a dataset of hundreds of millions of multilingual COVID-19 tweets with location information. *SIGSPATIAL Special* 12, 1 (2020).

[15] Afshin Rahimi, Duy Vu, Trevor Cohn, and Timothy Baldwin. 2015. Exploiting Text and Network Context for Geolocation of Social Media Users. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, 1362–1367. https://www.aclweb.org/anthology/N15-1153.pdf

[16] Luke Sloan, Jeffrey Morgan, William Housley, Matthew Williams, Adam Edwards, Pete Burnap, and Omer Rana. 2013. Knowing the tweeters: Deriving sociologically relevant demographics from Twitter. *Sociological research online* 18, 3 (2013), 74–84.

[17] WHO. 2020. *Coronavirus disease (COVID-19) pandemic.* https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/novel-coronavirus-2019-ncov

[18] Benjamin Wing and Jason Baldridge. 2014. Hierarchical discriminative classification for text-based geolocation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 336–348.