# Syntactical Text Analysis to Disambiguate between Twitter Users' In-situ and Remote Location

Helen N. Serere [1], Bernd Resch [1,2]

[1] *University of Salzburg, Department of Geoinformatics, Schillerstrasse 30, Salzburg, 5020 Austria*
[2] *Harvard University, Center for Geographic Analysis, MA 02138, Cambridge, Index, USA*

**Abstract**

The precision of text-based location inference models, which aim to identify a tweets' point of origin through analysing the post's text, is strongly influenced by differing location mentions. This particularly concerns the description of remote locations, i.e., locations that do not coincide with the user's location when posting a tweet. To filter out remote location mentions keyword filtering, temporal information matching and rule-based matching approaches have been used. However, these methods fail to take into account the tweets' syntax and hence produce low performance. We propose an advanced Named Entity Recognition model that not only extracts location entities but distinguishes between remote and in-situ location mentions based on the texts' surrounding grammatical cues. We train our algorithm on a base spaCy model which exhibits moderate performance on a relatively small training size. Preliminary results show that our approach outperforms similar studies and suggest the possibility of distinguishing between in-situ and remote location mentions with higher precision upon further refinement of the study design.

**Keywords**

Named Entity Recognition, Location inference, Tweet text, spaCy

## 1. Introduction

With the automatic disabling of location sharing, less than 3% of generated tweets are coordinate geotagged [1], that is, have a latitude and longitude value corresponding to the user's location when posting a tweet. This small percentage of geotagged posts limits the sample size of posts that can be used in spatial analysis, thereby compromising the representativeness of the Twitter population [2].

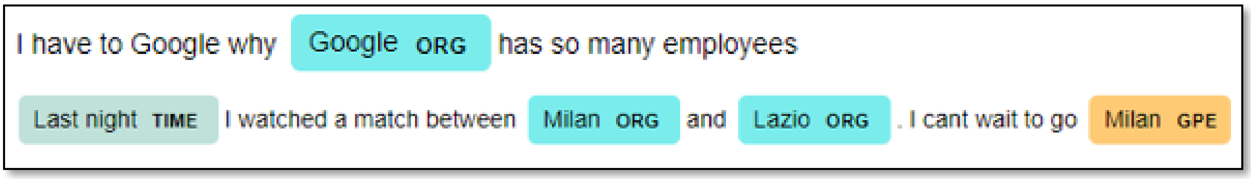
Text-based location inference models have been developed to increase the percentage of geotagged posts by inferring the tweets' point of origin. These developed models have reported precision values ranging between 55% and 85% within a 50 km radius of the tweets' point of origin [3–5].

Several factors can be attributed to the reduction in precision values [6–10]. In this paper, we address the reduction of precision values as a result of unfiltered remote location mentions. In the context of this paper, remote location is used to refer to any location that does not coincide with the tweets' point of origin. For example, South Carolina would be a remote location in the post, '*I may move to South Carolina by the end of the semester*' because the tweet refers to a distant location rather than an in-situ location (tweet location / location of the user when writing the post).

Remote location mentions have been filtered in the past by using keywords [3,11,12], rule based matching [13–15], classifiers [16] and machine learning models [17,18]. However the developed methods reported relatively lower precision values on the location inference models which suggests that they miss a percentage of remote mentions.

GeoExT 2023: First International Workshop on Geographic Information Extraction from Texts at ECIR 2023, April 2, 2023, Dublin, Ireland.
EMAIL: helenngonidzashe.serere@plus.ac.at (H. N. Serere); bernd.resch@plus.ac.at (B. Resch)
ORCID: 0000-0002-3494-7337 (H.N. Serere); 0000-0002-2233-6926 (B. Resch)

We propose to integrate a custom model into an existing spaCy [19] syntax based Named Entity Recognition (NER) model to distinguish between remote and in-situ location mentions. Similar to the existing spaCy[2] models, our custom model needs to be able to distinguish locations based on the sentence syntax. Example posts in **Figure 1** show the capability of spaCy in distinguishing between named entities regardless of the entities being written alike. By adopting spaCy's base models, we can create a customised model for location distinction.



I have to Google why Google ORG has so many employees

Last night TIME I watched a match between Milan ORG and Lazio ORG . I cant wait to go Milan GPE

**Figure 1:** Example of spaCy recognized entities highlighted by spaCy's displaCy visualizer.

## 2. Approach

We design our analysis following four main steps highlighted in **Figure 2**. In the subsection below, we provide more details for each of the four steps.

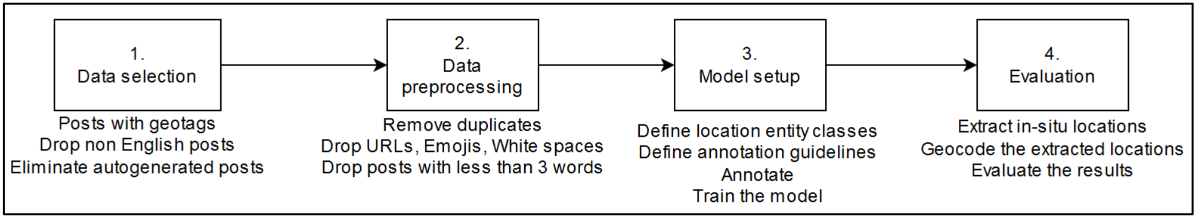

1. Data selection
Posts with geotags
Drop non English posts
Eliminate autogenerated posts
2. Data preprocessing
Remove duplicates
Drop URLs, Emojis, White spaces
Drop posts with less than 3 words
3. Model setup
Define location entity classes
Define annotation guidelines
Annotate
Train the model
4. Evaluation
Extract in-situ locations
Geocode the extracted locations
Evaluate the results

**Figure 2**: Overall design workflow showing the main components of the methodology.

1. **Data selection:** To evaluate the robustness of our developed approach, we restricted our dataset to posts which contained a coordinate geotag on a worldwide scale. We discarded auto-generated posts following the procedure outlined in [14].
2. **Data preprocessing:** We deleted URLs, emojis, and extra spaces because they did not aid value to our analysis. The elimination of these characters resulted in a simplified annotation process.
3. **Model setup:** We annotated 4,028 tweets using three entities classes: in-situ, remote and unclear. The annotation was done by a single annotator with an in depth knowledge of the objective. The annotation guidelines used are as follows:
   i. **A location was annotated as in-situ if:**
      a. The author clearly states that there are in that location at the time of sending the post. For instance, "*It feels good to be back here in* ***Ohio*** *[in-situ] after my six months internship abroad….*"
      b. The author attaches a location at the end of a post that does not include any mention to a past or future event. For instance: "*This is the kind of thing I like to see in my basement. @* ***Kitchener, Ontario*** *[in-situ]".*
   ii. **A location was annotated as remote if:**
      a. There is clear evidence that the author was not in the stated location at the time of sending the post: For instance: "*Popped over to* ***Budapest*** *[remote] last week for a couple of days. We really lucked out with the weather! It was absolutely gorgeous and made me really excited about the arrival of spring*."
   iii. **A location was annotated as unclear if:**
      a. There was no evidence that suggests that the location is either in-situ or remote: e.g. "*The croissants are DEFINITELY better in* ***France*** *[unclear]*" or in

[2] https://spacy.io/

the post, *"I need someone I can travel with from **Ferndale** [unclear] to **Bryanston** [unclear] twice a week"*.

b. The author attaches a location at the end of a post that includes a mention to a past or future event. e.g. *"Me last night…. @ **San Diego, California** [unclear]".*

c. A location is attached without any surrounding text. e.g. *"@ **Open Arms Christian Fellowship** [unclear]".*

d. The location follows the structure: Just posted a photo / video @. e.g. *"Just posted a photo @ **Irving, Texas** [unclear]".*

We excluded all location names that were used metonymically e.g. the country names Tanzania and Nigeria in the post: "***Tanzania*** *has reportedly started exploring a Central Bank Digital Currency (CBDC). The country is following the footsteps of* **Nigeria**, *which Launched its own digital currency last month…"* To prevent multiple location labeling, we annotated locations in their full totality including any linking terms e.g. "*I'm at* ***The Village of River Oaks in Houston, TX*** *[in-situ]".*

We trained our syntax model on a version 3.1.0 empty spaCy high accuracy English model officially abbreviated as '*en_core_web_trf'*. We used 80% (*3,222/4,028*) of the annotated tweets for training and the remaining 20% (*806/4,028*) for testing.

4. **Evaluation***:* The aim of our syntax based model is to distinguish between in-situ and remote location mentions so as to as to obtain higher precision values when inferring tweets' points of origin in non-geotagged posts. Therefore, after training our syntax model we evaluated our overall method design by inferring in-situ locations from a random sample of geotagged tweets.

   Our evaluation followed three steps. First, we extracted in-situ locations from a random sample of 88,732 pre-processed coordinate geotagged tweets. Second, we geocoded the extracted in-situ locations using the Google Maps geocoding API. The geocoder returned, for each geocoded in-situ location, the centroid coordinates and northeast and southwest coordinate pairs of the location's bounds. In our third and finally step, we compared each locations' geocoded coordinates against the tweets' attached geotagged coordinates.

   We defined two approaches for the comparison. In **Approach 1**, we computed the geodesic displacement between the geocoded centroid coordinates and the geotagged coordinates. To account for geographical scale, in **Approach 2** we generated a bounding box from the geocoded northeast and southwest coordinates of each in-situ location and counted the number of geotagged points found within the bounding box of the corresponding geocoded location.

## 3. Discussion of preliminary findings

The overall F1 score of our model was **77.8%.** The model's performance was high for in-situ location entities (precision 86.2%, recall 88.2% and F1 score 87.2%) compared to the remote location entity (precision 54.7%, recall 43.9% and F1 score 48.7%) and unclear location entities (precision 62.4%, recall 43.8% and F1 score 51.5%). The low performance of the model on the remote and unclear location entities can be attributed to the low number of posts with a remote and unclear location in the training dataset, respectively. Of the 4,028 annotated tweets, 44.2% contained an in-situ location, 8.2% a remote location and 15.6% an unclear location.

Compared to a similar study of [18], our trained model returned a higher F1 score for in-situ location mentions (**87.2%**) than the reported best performing model (**74.0%**) . The authors reported much higher model performances (87.0%) for posts with a low evidence of being in-situ. Since we divided what would equate the low evidence of in-situ location mentions into remote and unclear entities, it is not justifiable to compare our results with the authors' findings.

After passing a random sample of pre-processed English tweets to our trained syntax model, we extracted in-situ locations from 31.7% (28,129/88,732) of the posts. Of the extracted locations, 84.9% (23,869/28,129) were successfully geocoded by the Google Maps API. Figure 3 shows the obtained results from applying the two approaches described in section 4. Using Approach 1, only 8.8% of the in-situ locations were geocoded to a radius of greater than or equal to 50 km from the tweets' geotagged coordinate. This result could be due to the model extracting remote locations as in-situ, which is probable given that the model's in-situ location precision value was only 86.2%. Another reason could be the presence of high granular locations defined as in-situ locations for example USA.

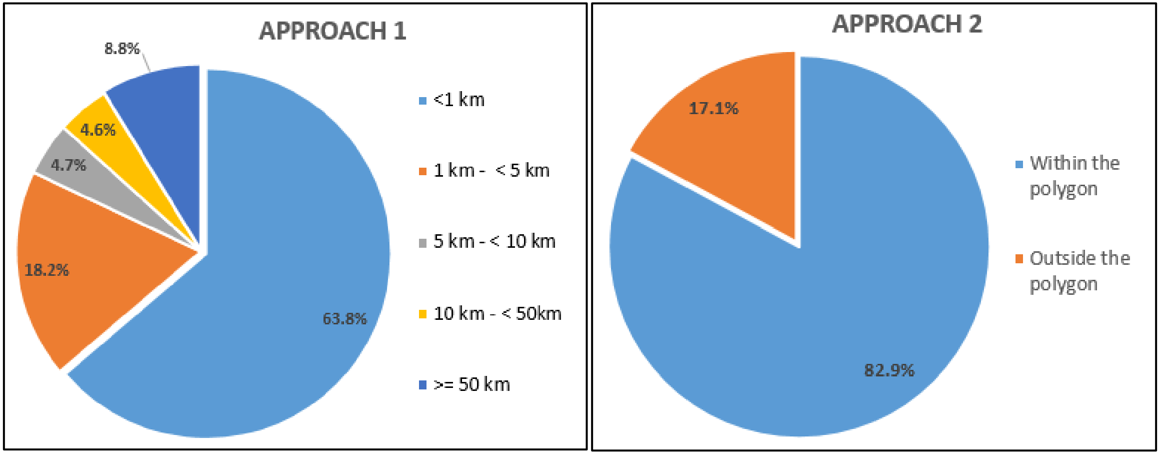


*Figure 3: Performance of developed method using evaluation Approach 1 (left) and Approach 2 (right).*

To cater for high spatial granularity, we used **Approach 2** which considers the bounding box of the defined area. However, using approach 2, surprisingly resulted in an even higher percentage of posts (17.1%) classified outside the geocoded bounding box. This result suggests probable limitations to bounds defined by the Google Maps geocoder. For instance, in the post "*I can't believe I am standing right in front of the **Eiffel Tower*** [in-situ]", the tweets' geotagged position might be falling outside of the Eiffel Tower's bounding box, according to Google Maps' defined bounding box, which then lowers the percentage of posts counted within the polygon. This theory, however, needs to be investigated further perhaps by using manual validation.

In Table 1 we show a comparison of our syntax based approach to previous studies which inferred in-situ locations from tweet text. Overall, our syntax model was able to outweigh most of the studies with the exception of the entity prioritization method [12] for the 10 km and 50 km radius values. By eliminating limitations surrounding the development of our approach such as, the small annotation size, use of a single annotator, low percentage of remote locations in the training data etc. the performance of our syntax model can be greatly improved.

*Table 1: Comparison of location inference results with previous studies*

| **Precision @ 1 km radius** | **(%)** | **Precision @ 10 km radius** | **(%)** | **Precision @ 50 km radius** | **(%)** |
|---|---|---|---|---|---|
| Staking approach [20] | 22 | Staking approach [20] | 37 | Stacking approach [20] | 54 |
| Keyword association [21] | 18 | Keyword association [21] | 45 | Label propagation [4] | 65 |
| Bayes model [22] | 44.4 | Ranking algorithm [14] | 60 | Ranking algorithm [14] | 83 |
| Entity prioritization [12] | 61.9 | Entity prioritization [12] | **86.1** | Entity prioritization [12] | **92.1** |
| **Syntax model** | **63.8** | **Syntax model** | 82.0 | **Syntax model** | 91.2 |

## 4. Conclusion

The aim of our paper was to customize a syntax model that can distinguish between in-situ and remote locations. Our preliminary results show high performance of our developed syntax model in comparison to related studies. However, further refinement of the study design is needed to improve the overall model performance especially with regards to the extraction of remote location mentions.

## 5. References

[1] Huang B, Carley KM. A large-scale empirical study of geotagging behavior on Twitter. Proc. 2019 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min., New York, NY, USA: ACM; 2019, p. 365–73. https://doi.org/10.1145/3341161.3342870.

[2] Karami A, Kadari RR, Panati L, Nooli SP, Bheemreddy H, Bozorgi P. Analysis of Geotagging Behavior: Do Geotagged Users Represent the Twitter Population? ISPRS Int J Geo-Information 2021;10:373. https://doi.org/10.3390/ijgi10060373.

[3] Serere HN, Resch B, Havas CR, Petutschnig A. Extracting and Geocoding Locations in Social Media Posts: A Comparative Analysis. GI_Forum 2021;9:167–73. https://doi.org/10.1553/giscience2021_02_s167.

[4] Apreleva S, Cantarero A. Predicting the location of users on Twitter from low density graphs. 2015 IEEE Int. Conf. Big Data (Big Data), IEEE; 2015, p. 976–83. https://doi.org/10.1109/BigData.2015.7363848.

[5] Pontes T, Vasconcelos M, Almeida J, Kumaraguru P, Almeida V. We know where you live: Privacy characterization of foursquare behavior. UbiComp'12 - Proc 2012 ACM Conf Ubiquitous Comput 2012:898–905.

[6] Middleton SE, Kordopatis-Zilos G, Papadopoulos S, Kompatsiaris Y. Location extraction from social media: Geoparsing, location disambiguation, and geotagging. ACM Trans Inf Syst 2018;36. https://doi.org/10.1145/3202662.

[7] Inkpen D, Liu J, Farzindar A, Kazemi F, Ghazi D. Location detection and disambiguation from twitter messages. J Intell Inf Syst 2017;49:237–53. https://doi.org/10.1007/s10844-017-0458-3.

[8] Gritta M, Pilehvar MT, Collier N. Which Melbourne? Augmenting geocoding with maps. ACL 2018 - 56th Annu Meet Assoc Comput Linguist Proc Conf (Long Pap 2018;1:1285–96. https://doi.org/10.18653/v1/p18-1119.

[9] Gritta M, Pilehvar MT, Limsopatham N, Collier N. What's missing in geographical parsing? Lang Resour Eval 2018;52:603–23. https://doi.org/10.1007/s10579-017-9385-8.

[10] Hahmann S, Purves R, Burghardt D. Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes. J Spat Inf Sci 2014;9:1–36. https://doi.org/10.5311/JOSIS.2014.9.185.

[11] Steiger E, de Albuquerque JP, Zipf A. An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data. Trans GIS 2015;19:809–34. https://doi.org/10.1111/tgis.12132.

[12] Serere HN, Resch B, Havas CR. Enhanced geocoding precision for location inference of tweet text using spaCy, Nominatim and Google Maps. A comparative analysis of the influence of data selection. PLoS One 2023;18:e0282942. https://doi.org/10.1371/journal.pone.0282942.

[13] Vu HQ, Li G, Law R, Zhang Y. Travel Diaries Analysis by Sequential Rule Mining. J Travel Res 2018;57:399–413. https://doi.org/10.1177/0047287517692446.

[14] Laylavi F, Rajabifard A, Kalantari M. A Multi-Element Approach to Location Inference of Twitter: A Case for Emergency Response. ISPRS Int J Geo-Information 2016;5:56. https://doi.org/10.3390/ijgi5050056.

[15] Karagoz P, Oguztuzun H, Cakici R, Ozdikis O, Onal KD, Sagcan M. Extracting Location Information from Crowd-sourced Social Network Data. Eur Handb Crowdsourced Geogr Inf 2016:195–204. https://doi.org/10.5334/bax.o.

[16] Ribeiro S, Pappa GL. Strategies for combining Twitter users geo-location methods. Geoinformatica 2018;22:563–87. https://doi.org/10.1007/s10707-017-0296-z.

[17] Priedhorsky R, Culotta A, Del Valle SY. Inferring the Origin Locations of Tweets with Quantitative Confidence. Proc 17th ACM Conf Comput Support Coop Work Soc Comput 2014;23:1523–36. https://doi.org/10.1145/2531602.2531607.

[18] Lamsal R, Harwood A, Read MR. Where did you tweet from? Inferring the origin locations of tweets based on contextual information. 2022 IEEE Int. Conf. Big Data (Big Data), IEEE; 2022, p. 3935–44. https://doi.org/10.1109/BigData55660.2022.10020460.

[19] Honnibal M and M, Landeghem I and Van, Adriane S and B. spaCy: Industrial-strength Natural Language Processing in Python. Zenodo 2020. https://doi.org/10.5281/zenodo.1212303.

[20] Schulz A, Hadjakos A, Paulheim H, Nachtwey J, Mühlhäuser M. A multi-indicator approach for geolocalization of tweets. Proc. 7th Int. Conf. Weblogs Soc. Media, ICWSM 2013, 2013, p. 573–82.

[21] Ikawa Y, Enoki M, Tatsubori M. Location inference using microblog messages. Proc. 21st Int. Conf. companion World Wide Web - WWW '12 Companion, New York, New York, USA: ACM Press; 2012, p. 687. https://doi.org/10.1145/2187980.2188181.

[22] Lee K, Ganti RK, Srivatsa M, Liu L. When twitter meets foursquare: Tweet location prediction using foursquare. MobiQuitous 2014 - 11th Int Conf Mob Ubiquitous Syst Comput Netw Serv 2014:198–207. https://doi.org/10.4108/icst.mobiquitous.2014.258092.