Document Geocoding for Inferring Spatial Familiarity Within Reddit Text

Paddy Smith
School of Geography, University of Leeds, UK

Myles Gould
School of Geography, University of Leeds, UK

Ed Manley
School of Geography, University of Leeds, UK

Abstract

In this paper, we reframe social media user geolocation as a task of inferring a user's spatial familiarity. We propose an LLM-based approach that identifies places a user is most familiar, including the strength of familiarity. We evaluate this approach using two Reddit evaluation corpora constructed from subreddit user-flair tags: Reddit-London, where users are associated with London boroughs, and Reddit-USA, where users are associated with US states. We compare these datasets with GeoText benchmark dataset. The results show that LLMs can recover place-associated labels from social media text, achieving the strongest performance on Reddit-USA state prediction and moderate performance on fine-grained London borough prediction. We further show that inferred familiar places exhibit clear spatial dependency. Using polygon-based metrics, familiar places inferred from Reddit-London users have a mean distance to labeled boundary of 3.48 km, while Reddit-USA shows stronger alignment with state labels compared to GeoText, with higher inside-label share, and lower spatial dispersion. These findings suggest spatial familiarity provides a useful alternative framing for social media user geolocation, and that Reddit flair-based labels are more closely aligned with expressed spatial familiarity than home location labels in GeoText benchmark dataset.

2012 ACM Subject Classification Computing methodologies → Natural language processing

Keywords and phrases spatial familiarity, document geocoding, geoparsing, LLMs

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.25

Category Short Paper

Declaration on GenAI/LLM use During the preparation of this work, the authors used GPT-5.5, for improving structure of manuscript and supporting the coding of experiments. Use of Gen-AI and has been reviewed and verified by the authors.

1 Introduction

Identifying the location of social media users is useful across a range of domains, including disaster response, public health [11] and event detection [15]. In urban research, such approaches have been used to examine mobility patterns [9], social ties [17] and mapping place boundaries [18]. With the sparsity of available geotagged social media data [4], and the unreliability of self-declared location [8], there is a large body of literature that attempts to infer user location automatically from text [10, 6, 13].

A common approach to identifying social media user location is document geocoding, in which user-generated text is combined into a single document, and then analysed to infer a geographic location, typically interpreted as "home location" [12]. The primary assumption is that users will employ location or geo- indicative words, such as dialects, place names, cultural references, that implicitly or explicitly encode an association with a particular location [6, 1].

17th International Conference on Spatial Information Theory (COSIT 2026).
Editors: Sabine Timpf, Gabriele Filomena, Armand Kapaj, Rui Zhu, Nicholas Giudice, and Ed Manley;
Article No. 25; pp. 25:1–25:10

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

This approach assumes that the inferred location corresponds to where the user resides, based on their social media activity. However, users may not explicitly mention location-indicative terms in posts, or refer to places from locations they do not reside in [16]. Place references may be made for a variety of reasons, such as travel, past experiences or personal interests. For example, frequent references that indicate Boston, USA, may reflect support for a sports team, rather than physical location in the city. Despite this, many ground-truth datasets are created based on this assumption, such as GeoText [3] and Twitter-US [14], where user location is defined by the geotag of their first tweet [16]. The reliance on "home location" can therefore produce misleading conclusions about the relationship between text and user location, impacting models trained and evaluated on such data.

This limitation is more pronounced in discussion-oriented platforms such as Reddit, where users can engage in discussions, often without explicitly mentioning their location. Unlike other social media platforms, Reddit lacks geotagged data and often contains rich, descriptive conversations about place that reflect varying levels of familiarity. In this context, inferring a single "home location" is often neither feasible or appropriate. Instead, it is more meaningful to consider a user's spatial familiarity, meaning a *"close acquaintance with an environment or its elements"* [5]. It is about the extent a place is known [2], through direct exposure via regular interactions, or indirect exposure to information about places [5].

In this paper, we reconceptualise social media user geolocation as a task of inferring user's familiarity with place, rather than predicting a single home location. We propose an LLM-based approach that identifies a users familiar places across multiple spatial scales, along with the strength of this familiarity. We examine the spatial distributions of users spatial familiarity, evaluating the extent to which they exhibit spatial dependency. We evaluate this approach using proxy ground-truth signals derived from user-generated metadata (i.e., subreddit user flair) and a benchmark GeoText dataset.

The main contributions of this paper are threefold: (1) we reconceptualise social media user geolocation as spatial familiarity inference rather than home location; (2) we demonstrate a methodology for constructing spatial familiarity evaluation corpora from subreddit flair tags; and (3) we provide evidence that spatial familiarity extracted from social media users exhibits spatial dependency, with Reddit text showing tighter spatial clustering around ground-truth labels than GeoText home location labels.

2 Methodology

2.1 Reddit Data

Evaluating LLMs ability to infer Reddit user's spatial familiarity is challenging, due to the lack of ground-truth data [7]. Currently, there are no datasets that have gold-standard location for Reddit comments, or users. In addition, there is a very limited amount of location data attached to Reddit, with users not required to input their own location.

In order to address this, we utilise the "user flair" tag, which are self-attributed labels that are specific to individual subreddits. Often in place-based subreddits, these flairs indicate a user's association with a particular place. For example, in r/London it is common for users to tag a London borough, whilst r/AskAnAmerican users may tag their associated US state. Although these tags do not guarantee the user is located within this area, it can be presumed that the user has a high likelihood of familiarity with this place.

To build a spatial familiarity evaluation corpora, we collect Reddit comments and posts from r/London and r/AskAnAmerican subreddits using the Reddit API[1]. These subreddits are chosen as they provide opportunity to evaluate spatial familiarity inference at two different spatial scales. Users are identified if they have a London borough or US state as their flair tag, or a place within a London borough or state which is subsequently matched manually. Then, comments and posts are collected across the subreddit, and other subreddits they may have contributed to (i.e., r/LondonCycling, r/NewEngland). For each user, their comments and posts are combined together to form a document. Reddit user documents are cleaned to remove unusual characters and whitespaces in the text, and truncated to 35,000 characters, to fit within the LLM context window.

This produces two Reddit-datasets; Reddit-London containing 204 users with London borough labels, and Reddit-USA, containing 239 users with US state labels. LLMs are then evaluated based on their ability to identify the London borough or US state in which the user is associated in (via their flair tag) using the user document. As this is a classification task, precision, recall and F1-score are used to evaluate model performance.

2.2 GeoText

GeoText, a widely used benchmark data within social media user geolocation tasks [10], is also used in this study. We evaluate the LLM-based approach using Geotext, due to the sparsity of benchmark datasets available. GeoText is compromised of 9,475 Twitter users from the US [3]. The location of the user is defined as the geographic coordinates of their first Tweet. GeoText has typically been used to evaluate social media user geolocation methods, by computing distance error metrics between the gold-standard coordinates and predicted coordinates. However, distance-based metrics do not capture if LLMs are able to identify levels of user spatial familiarity. Therefore, we adapt GeoText by assigning each user to a US state, by intersecting their location with state boundaries. An LLM must then predict a US state that the user is based within. The assumption is that the user would have some familiarity with the state, and the LLM can infer these based on the geo-indicative words and place mentions.

2.3 Model inference

We implement two open-source LLMs; `qwen2.5-72B`[2] and `Llama3.3-70B-Instruct`[3], that have been quantized and are available via HuggingFace. We also deployed `claude-sonnet-4.6`[4] on the Reddit-based datasets to test a larger parameter model. Two different prompt are constructed, giving instructions to estimate spatial familiarity for Reddit-London and Reddit-USA (Appendix A.1) and home location for GeoText (Appendix A.2). For this paper, only zero-shot prompt approaches are experimented with. The model is instructed to output in a predefined JSON schema. Each user is assigned a predicted location, a familiarity score, and then a list of places they are familiar with.

[1] https://redditinc.com/policies/data-api-terms
[2] https://huggingface.co/Qwen/Qwen2.5-72B-Instruct-AWQ
[3] https://huggingface.co/casperhansen/llama-3.3-70b-instruct-awq
[4] https://www.anthropic.com/news/claude-sonnet-4-6

2.4 Evaluating spatial dependency

During model inference, the models also predict a set of familiar places for each user, including the location of these places and an estimated familiarity score. These inferred places are used to assess whether a user's place familiarity is spatially dependent on their ground-truth location labels. The spatial dependency analysis uses polygon-based distance metrics. This is appropriate because the ground-truth labels in this study refer to administrative areas, namely London boroughs and US states. For each inferred familiar place, we calculate its distance to the user's labeled administrative boundary. Places located inside the labeled borough or state are assigned a distance of zero, while places outside the labeled area are assigned their distance to the nearest point on the boundary.

For each user, we calculate polygon mean distance to anchor (polygon MDA), defined as the mean distance between a user's inferred familiar places and their labeled polygon. We also calculate a share of familiar places located inside the labeled area and the mean distance of outside-label places from the labeled boundary. Lower polygon distance and higher inside-label share indicate stronger spatial dependency between inferred place familiarity and user's ground-truth label. Additionally, spatial dispersion is computed as the mean pairwise distance between all familiar places for a given user. Higher dispersion indicates a user's familiar places are spread across a wider geographic area, whereas lower dispersion indicates concentrated local familiarity.

Finally, kernel density estimation (KDE) is used to visualize user's spatial familiarity on different aggregate spatial scales. For each user, the LLM-extracted familiar places are aggregated to produce a density surface, weighted by the familiarity score assigned to each place.

3 Results

3.1 Predicting user location labels

We first predict user's location labels for both Reddit-London, Reddit-USA and GeoText datasets. For Reddit-London and Reddit-USA, we are predicting the associated London borough or US state, whereas for GeoText we are predicting the US state as a home location. Model performance for this task is shown in Table 1.

Table 1 Model evaluation for predicting ground-truth location labels for Reddit-London, Reddit-USA and GeoText users.

Model	Dataset	Resolved Users	Precision	Recall	F1-score
Qwen2.5-72B	Reddit-London	99%	0.63	0.5	0.52
Llama3.3-70B	Reddit-London	94%	0.57	0.48	0.5
Claude-Sonnet-4.6	Reddit-London	90%	0.71	0.58	0.61
Qwen2.5-72B	Reddit-USA	93%	0.78	0.78	0.77
Llama3.3-70B	Reddit-USA	96%	0.86	0.88	0.85
Claude-Sonnet-4.6	Reddit-USA	100%	0.9	0.89	0.88
Qwen2.5-72B	GeoText	76%	0.42	0.33	0.36
Llama3.3-70B	GeoText	66%	0.46	0.37	0.39

High F1-scores are achieved across the models for the Reddit-USA dataset. Therefore the task of predicting the US state flair-tag based on the Reddit user document was performed accurately and with minimal errors. In contrast, the same models performed with less precision

when predicting US states for the GeoText users. This may highlight the misalignment between the user document and the assigned home location in the GeoText dataset. For the Reddit-London dataset, the models achieved modest precision and recall scores. Despite the granularity of London boroughs, the models were relatively successfully in correcting labeling a London borough for the Reddit user. As expected, the larger parameter `claude-sonnet-4.6` shows the highest performance on the Reddit-London and Reddit-USA datasets.

3.2 Spatial dependency

Next, we examine whether the familiar places inferred for each user are spatially dependent on the user's ground-truth location label using polygon-based metrics (Table 2). Across all datasets, Reddit-London showed the lowest polygon MDA (3.48km) and outside-boundary distance (4.47km), however had a lower inside-label share (33%). Therefore often familiar places mentioned by a user are nearby to their assigned London borough, but not necessarily within.

Table 2 Polygon-based spatial dependency metrics for Reddit-London, Reddit-USA and GeoText.

Dataset	Polygon MDA (km)	Inside-label share (%)	Outside-boundary distance (km)	Spatial dispersion (km)
Reddit-London	3.48	33	4.47	5.49
Reddit-USA	164.28	80	291.41	263.03
GeoText-state	890.12	49	1081.12	695.34

Reddit-USA shows a different pattern, with a much larger share of inferred familiar places inside labeled states (80%). However, polygon MDA and outside-boundary distance are substantially larger than for Reddit-London, reflecting the broader geographic scale of US states. This indicates stronger within-label alignment, but a much larger spatial extent for familiar places.

GeoText can be compared more directly with Reddit-USA, as they both have US states as user location labels. GeoText shows a weaker spatial dependency, with a lower inside-label share (49%), higher polygon MDA (890.12 km), and larger outside-boundary distance (1081.12 km). GeoText also has a higher spatial dispersion, suggesting familiar places inferred from GeoText users are more geographically spread out. These results support the argument that Reddit flair-based labels are more closely aligned with expressed place familiarity, while GeoText home-location labels capture a different and less spatially concentrated relationship between users and places.

We next use KDE visualizations to examine the spatial structure underlying the spatial dependency analysis. For Reddit-London, we show boroughs that are most and least spatially clustered in terms of familiar places (Figure 1). For Reddit users associated with Islington and Ealing, familiar places commonly discussed are more spatially concentrated within or around the borough boundary. In comparison, users associated with Waltham Forest and Bromley commonly reference more dispersed familiar places, spanning multiple boroughs and extending into central London. This helps explain why Reddit-London has a low polygon MDA despite a relatively low inside-label share, as many familiar places fall outside the labeled borough, but remain nearby within the wider urban context.

We then compare KDEs for selected US states across Reddit-USA and GeoText (Figure 2). States are selected where both datasets contain sufficient users and familiar place mentions. Comparing the datasets, Reddit-USA familiar places are substantially more concentrated within or near the labeled state boundary. In contrast, GeoText familiar places are more

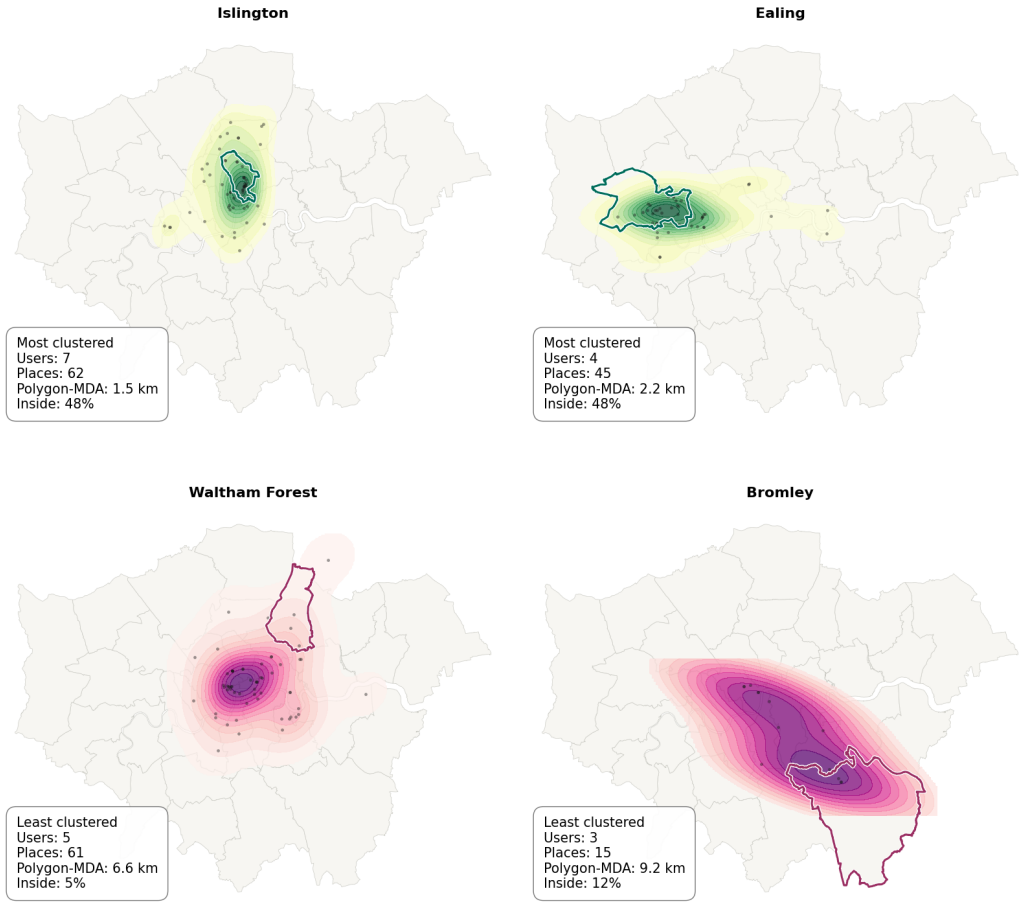

Figure 1 Spatial familiarity density for selected London boroughs. KDE surfaces are weighted by LLM familiarity scores, with each point denoting an inferred familiar place.

spatially dispersed and less consistently aligned with the assigned home-state label. For example, users assigned to Texas in GeoText reference familiar places across a wider national geography, including California and New York. Whereas, Reddit-USA users show a stronger connection of familiar places within Texas. These patterns support the polygon-based metrics, which show higher inside-label share and lower outside-boundary distance for Reddit-USA compared to GeoText.

4 Discussion & Conclusion

In this short paper we have considered the limitations of identifying "home location" of social media users, and instead demonstrated an alternative spatial familiarity approach. The potential application of this approach is demonstrated on Reddit data, a social media dataset which lacks user location information, however is increasingly being used for urban analysis. We illustrate an approach to constructing a place familiarity evaluation corpora using subreddit user flair-tags. We provide preliminary evidence that LLMs can support fine-grained user location inference (i.e., London boroughs), as opposed to coordinates or larger spatial units. We also demonstrate the potential for LLMs estimating continuous levels of spatial familiarity and identifying sets of locally relevant places associated with individual users.

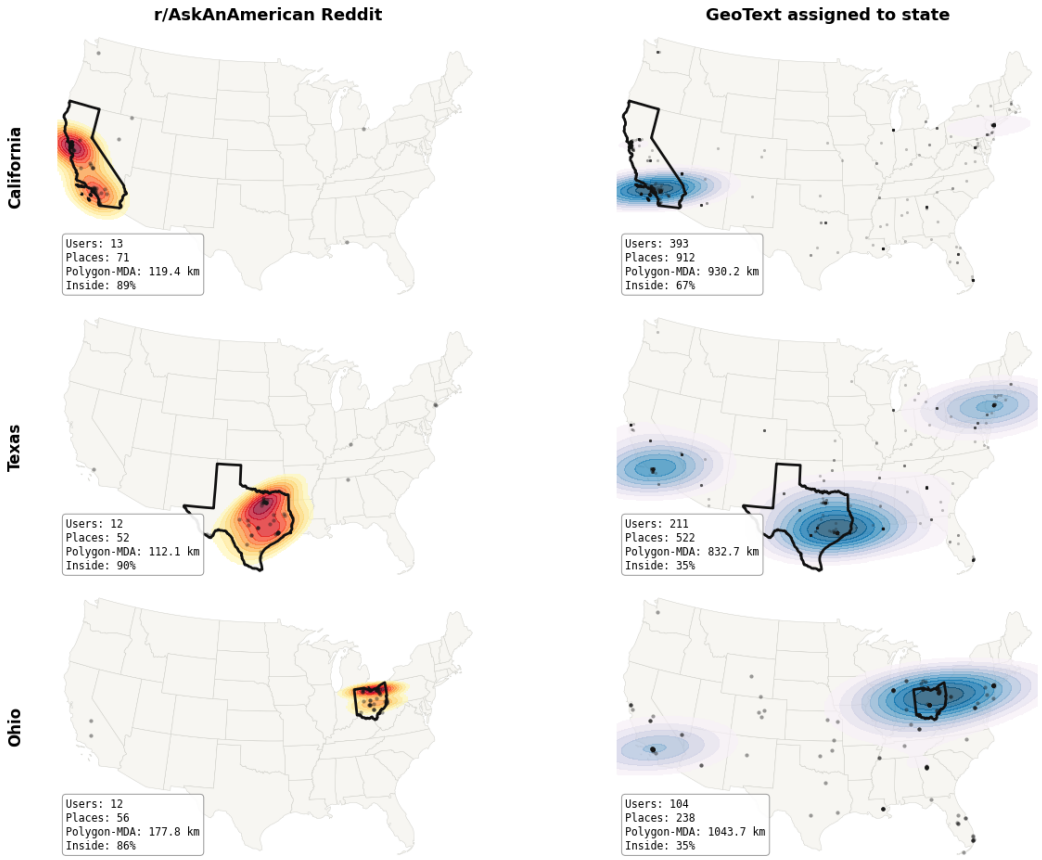

Figure 2 Spatial familiarity density for selected US states in Reddit-USA and GeoText. KDE surfaces are weighted by LLM familiarity scores, with each point denoting an inferred familiar place.

Further work is required to validate model inferences to assess how closely LLM predictions align with actual spatial familiarity. In particular, the Reddit-based evaluation corpus presents several limitations. Firstly, the flair-tags are self-reported therefore unverified and might not reflect an individuals own familiarity. Secondly, the size of our evaluation corpora is relatively small, providing opportunity to increase dataset size by incorporating other place-based Subreddits (r/LosAngeles, r/Australia). Finally, the familiarity scores and locally relevant places are not independently verified, therefore cannot assess if they are meaningful. Future work could compare these scores to survey-based scores, or other urban perception datasets.

We also demonstrate that Reddit user spatial familiarity exhibits clear spatial dependency, forming coherent geographic patterns. Reddit-USA shows a polygon mean distance to anchor of 164.28km, compared to 890.12km for GeoText users, suggesting that social media text from Reddit closer reflects the geographic context indicated by the ground-truth label. In contrast, the higher spatial dispersion in GeoText challenges the validity of home location as a ground-truth proxy, as the geographic content of user text frequently does not align with the registered home location. These findings suggest that Reddit flair-based labels may represent a higher quality ground-truth signal for place familiarity evaluation than geotag-derived home locations.

References

1. Benjamin Adams and Krzysztof Janowicz. On the Geo-Indicativeness of Non-Georeferenced Text. *Proceedings of the International AAAI Conference on Web and Social Media*, 6(1):375–378, August 2012. doi:10.1609/icwsm.v6i1.14309.
2. David J. Chalmers and Robert G. Knight. The Reliability of Ratings of the Familiarity of Environmental Stimuli: A Generalizability Analysis. *Environment and Behavior*, 17(2):223–238, March 1985. doi:10.1177/0013916585172003.
3. Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. A Latent Variable Model for Geographic Lexical Variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010.
4. Deen Freelon. Computational Research in the Post-API Age. *Political Communication*, 35(4):665–668, October 2018. doi:10.1080/10584609.2018.1477506.
5. Nathan Gale, Reginald G. Golledge, William C. Halperin, and Helen Couclelis. EXPLORING SPATIAL FAMILIARITY. *The Professional Geographer*, 42(3):299–313, August 1990. doi:10.1111/j.0033-0124.1990.00299.x.
6. Bo Han, Paul Cook, and Timothy Baldwin. Geolocation Prediction in Social Media Data by Finding Location Indicative Words. In *Proceedings of COLING 2012: Technical Papers, Pages 1045–1062,*, 2012.
7. Keith Harrigian. Geocoding Without Geotags: A Text-based Approach for reddit. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 17–27, Brussels, Belgium, 2018. Association for Computational Linguistics. doi:10.18653/v1/W18-6103.
8. Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. 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*, pages 237–246, Vancouver BC Canada, May 2011. ACM. doi:10.1145/1978942.1978976.
9. Raja Jurdak, Kun Zhao, Jiajun Liu, Maurice AbouJaoude, Mark Cameron, and David Newth. Understanding Human Mobility from Twitter. *PLOS ONE*, 10(7):e0131469, July 2015. doi:10.1371/journal.pone.0131469.
10. Xiangyang Luo, Yaqiong Qiao, Chenliang Li, Jiangtao Ma, and Yimin Liu. An overview of microblog user geolocation methods. *Information Processing & Management*, 57(6):102375, November 2020. doi:10.1016/j.ipm.2020.102375.
11. Hanjia Lyu, Long Chen, Yu Wang, and Jiebo Luo. Sense and Sensibility: Characterizing Social Media Users Regarding the Use of Controversial Terms for COVID-19. *IEEE Transactions on Big Data*, 2020. doi:10.48550/arXiv.2004.06307.
12. Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Home Location Identification of Twitter Users. *ACM Transactions on Intelligent Systems and Technology*, 5(3):1–21, October 2014. doi:10.1145/2528548.
13. Afshin Rahimi, Duy Vu, Trevor Cohn, and Timothy Baldwin. Exploiting Text and Network Context for Geolocation of Social Media Users, 2015. doi:10.48550/arXiv.1506.04803.
14. Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. Supervised Text-based Geolocation Using Language Models on an Adaptive Grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012.
15. Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, pages 851–860, Raleigh North Carolina USA, April 2010. ACM. doi:10.1145/1772690.1772777.
16. Zhaomin Xiao and Eduardo Blanco. Are People Located in the Places They Mention in Their Tweets? A Multimodal Approach. In *Proceedings of the 29th International Conference on Computational Linguistics*, 2022.

17 Sarita Yardi and Danah Boyd. Tweeting from the Town Square: Measuring Geographic Local Networks. *Proceedings of the International AAAI Conference on Web and Social Media*, 4(1):194–201, May 2010. doi:10.1609/icwsm.v4i1.14016.

18 Junjun Yin, Aiman Soliman, Dandong Yin, and Shaowen Wang. Depicting urban boundaries from a mobility network of spatial interactions: A case study of Great Britain with geo-located Twitter data. *International Journal of Geographical Information Science*, 31(7):1293–1313, July 2017. doi:10.1080/13658816.2017.1282615.

A LLM Prompts

A.1 Prompt for estimating place familiarity within Reddit-London

```
You are analysing Reddit comments written by a single user from London related subreddits.
Your goal is to estimate levels of place familiarity, which is the extent an
individual knows their surroundings. Familiarity with a place does not necessarily mean a
user resides in that place.
### TASK
Assess the users level of familiarity with London, and identify specific London borough.
Additionally, identify any specific local places (e.g., stations, parks, landmarks) the
user appears familiar with. For each such place, provide its name, a familiarity score
(0-1), and its approximate coordinates (latitude, longitude).
There is no limit on the number of places you may list.
### How to identify familiarity?
- Repeated references to specific neighbourhoods or places within a borough
- Detailed local knowledge (routes, transport, pricing, safety, comparisons)
- Personal or experiential language ("I live near...", "my area...", "I always go to...")

Return ONLY valid JSON in this format:
{{
  "PredictedBorough": "BOROUGH_NAME",
  "ApproximateLocation": {{
    "lat": 0.0,
    "lon": 0.0
  }},
  "CityFamiliarityScore": 0.0,
  "LocalPlaceFamiliarity": [
    {{
      "place": "PLACE_NAME",
      "familiarity": 0.0,
      "lat": 0.0,
      "lon": 0.0
    }}
  ],
  "Confidence": 0.0,
  "Evidence": "brief explanation of decision"
}}
User comments:
{document}
```

A.2 Prompt for estimating home location of Twitter users in GeoText

```
You are analysing Tweets written by a single user in the United States.
TASK: Predict which state the user most likely lives in.
```

```
Return ONLY valid JSON in this format:
{{
    "PredictedState": "STATE_NAME",
    "Confidence": 0.0
}}
Rules:
- Confidence must be between 0 and 1
- Output JSON only
- No explanation or markdown

User comments:
{document}
```