2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

How Well Did You Locate Me? Effective Evaluation of Twitter User Geolocation

Ahmed Mourad, Falk Scholer, and Mark Sanderson
School of Science and Information Technology
RMIT University
Melbourne, Australia
{ahmed.mourad, falk.scholer, mark.sanderson}@rmit.edu.au

Walid Magdy
School of Informatics
The University of Edinburgh
Edinburgh, UK
wmagdy@inf.ed.ac.uk

***Abstract*—We analyze fifteen Twitter user geolocation models and two baselines comparing how they are evaluated. Our results demonstrate that the choice of effectiveness metric can have a substantial impact on the conclusions drawn from an experiment. We show that for general evaluations, a range of metrics should be reported to ensure that a complete picture of system effectiveness is conveyed.**

I. Introduction

Geolocating Twitter users is needed in many social media-based applications, such as identifying geographic lexical variation [1], managing natural crises [2], gathering news [3], and tracking epidemics [4]. While users can record their location on their profile, more than 34% record fake or sarcastic locations [5]. Twitter allows users to GPS locate their content, however, less than 1% of tweets are geotagged [6]. Inferring user location is therefore an important field of investigation.

The different needs of each geolocation application may require evaluation from several perspectives. Current evaluation practices focus on a few measures [1], which were shown to be biased towards urban locations [7]. For tasks such as searching for sources to cover local news [3], monitoring natural disasters in rural areas [2], or tracking epidemics in rural cities [4], the measures were potentially unsuitable. Evaluation at multiple levels of geographic granularity is also not widely used despite a requirement in some applications: e.g. when identifying eyewitnesses from social media, journalists sometimes aggregate predicted eyewitness locations at different scales (e.g. city, state or country) [8].

The evaluation of geo-inference methods is affected by many factors, such as ground-truth construction, geographic coverage, and how the earth is represented. Analyzing the quality of fifteen geolocation models and two baselines using ten different evaluation measures over four geographic granularities, our study makes two key contributions:

- We standardize the evaluation process for models to ensure a fairness of comparison. We demonstrate that some older models that were previously thought to be uncompetitive perform comparably to recent approaches.
- We examine the influence of social media population bias on the quality of geolocation prediction. We find that multiple metrics and a majority class baseline are required for the evaluation of more complex geolocation models.

We highlight a critical shortcoming of current evaluation of geolocation models: the choice of effectiveness metric may lead a researcher to conflicting conclusions about which system performs better. Systems should be evaluated using consistent measures. Our results also demonstrate the different properties of measures, which can in turn lead to a better understanding of the differences between models, and to better decision-making based on specific application requirements.

II. Related Work

Evaluation of geolocation models has evolved sporadically over the years. *Median* and *Mean* error distances were initially used as intuitive measures of the error distance between an estimated and true location beside accuracy (*Acc*) at the level of states and regions [1]. Accuracy within x miles from the original city was later introduced [9], and at the level of country [5]. Precision and recall were reported at the level of each city and an overall macro-F1 metric [10], which were further extended to consider micro, weighted, and macro averaging techniques at the level of the three metrics [7]. Other research employed a combination of these measures.

Han el al. [6] demonstrated that a multinomial naïve bayes model performs better than logistic regression [11] using city-based representation, Wing et al. [12] demonstrated the opposite using uniform grids. Two different models, each performed better using different geocoding technique.

Social media is known to have a substantial population bias [13]. Not many researchers explored the impact of this bias on either determining the most effective models or evaluation metrics. Rodrigues et al. [10] reported macro precision, recall, and f1-score beside accuracy to accommodate for the distribution imbalance in their dataset. Johnson et al. [14] demonstrated that existing geolocation approaches perform significantly worse for rural areas. They explored different sampling techniques on a US rural-urban county-based dataset and evaluated two models as representatives of the most popular approaches: text and network-based. Although precision (as a function in error distance) and recall were employed instead of accuracy as conventional alternatives, consolidating locations into two classes only (rural-vs-urban) limits the

IEEE/ACM ASONAM 2018, August 28-31, 2018, Barcelona, Spain
978-1-5386-6051-5/18/$31.00 © 2018 IEEE

scalability of the analysis, where most of the recent work relies on datasets with global geographic coverage, and the population bias exists even within each of the two classes.

Two previous studies have analyzed the effectiveness of evaluation metrics of Twitter user geolocation. Jurgens et al. [15] compared nine geolocation models using a standardized evaluation framework. Their evaluation was limited to a network-based geolocation approach using error distance measures (AUC and Median) and a network specific measure, which does not generalize to other approaches, such as the widely-used text-based ones. A recent work pointed out that accuracy measures are biased towards locations with a large population [7]. Although they employed a wide range of metrics, however, their work was limited to a single geolocation model while focusing on the influence of language rather than the effectiveness of the evaluation measures.

We focus on the effectiveness of geolocation evaluation regardless of the underlying geolocation approach or the language of text, which entails generalization challenges that we discuss in the next section. We evaluate the relative performance of fifteen geolocation models and two baselines.

III. Standardized Evaluation

First, alternate metrics are described that address data imbalance. Second, a unified output format and geocoding method are employed to ensure the fairness of comparisons.

A. Alternate Metrics

Much past research treated the problem of geolocating Twitter users as a categorization task. Given the global geographic coverage of such a task (typically thousands of locations), there is an inherent imbalance in the distribution of users over locations. *Acc* and *Acc@161* are biased towards regions with a high population (the majority classes) [14]. Hence, we investigate the conventional measures for multi-class categorization [16], which were included partially [10] and fully [7] in the context of Twitter user geolocation. We consider Precision (P), Recall (R) and F1-score (F1) using Micro (μ) and Macro (M) averaging. *Precision* is more favored in situations such as when journalists are looking for eyewitnesses within a specific city [8]. *Recall* is favored in situations such as when these journalists want to increase the search pool [17]. Both scenarios focus on a single location, where the comparison at the micro and macro levels is essential.

B. Unified Output and Geocoding

When comparing models, we train and test on the same dataset and use models that output the same earth representation. Assume we have two models: A and B. A represents the earth as a grid of cells and B represents the earth as cities, with a city often corresponding to multiple cells in the first representation. A user's home location is identified as cell x inside city Z. Now assume A predicted the location of this user as cell y and B predicted it as city Z. Based on the underlying representation of each model, the prediction of model A will be considered incorrect while the prediction of model B is correct.

To avoid such inconsistency, we unified the output of all the models to GPS coordinates [18], which were resolved to a location using the Google Geocoding API (V3) before evaluation. Using one API ensures fair comparison over the same set of locations (classes) and it also allows evaluation over different granularities. Due to space constraints, we report only performance at city and country level. We also calculated county and state level, but found little difference in results.

IV. Experimental Setup

We examine two sets of systems. The first set (Local) includes four geolocation models and two baselines, trained and tested (over 30k users) locally over the same data collection with free earth representation to evaluate the considered process. The second set (W-NUT) includes eleven submissions from a geolocation shared task [19], which we use to assess the robustness of our proposed evaluation process. Although the published results for participating models were evaluated at city level only, we were also able to infer output at country level based on information released by the W-NUT organizers.

A. Local *Models*

1) Data Collection Method: We employed a geographically global geotagged tweet collection, **TwArchive**, holding content since 2013[1] drawn from the 1% sample Twitter public API stream. We used a 2014 subset spanning nine months. We focus on English tweets only as identified by langid.py.[2] Non-geotagged and duplicate tweets were removed using user id and tweet text. For the sake of a standard evaluation, users with an unresolved home location—based on the model that accepts home locations in the form of cities instead of GPS coordinates [6]—were removed from the dataset. The total number of users and tweets after pre-processing is ~1.5 million and ~3.1 million respectively.

2) Ground Truth: The home location of a user was identified at the geometric median of their geotagged tweets, which has been shown to be more accurate than other approaches [20]. Such a point is the minimum error distance to all locations of a user.

3) Geolocation Inference Models: Four models and two baselines were compared using four classification methods and two statistical methods. The models were chosen based on their availability, reproducibility, and recency.

Rl12 is an adaptive grid-based representation with a trained probabilistic language model per cell [11]. Each cell has the same number of users, but a different geographical area. We employ their best reported parameter values. The output represents the centroid of the predicted cell.

Hn14 locates users to one of 3,709 cities [6]. We re-implemented their system, focusing on the part that uses Location Indicative Words (LIW) drawn from tweets, where mainstream noisy words were filtered out using their best

[1] https://archive.org/details/twitterstream&tab=collection

[2] https://github.com/saffsd/langid.py

reported feature selection method, Information Gain Ratio. The output represents the centre of the predicted city.

Rm16 assigns a user to one of 930 non-overlapping geographic clusters based on the similarity of content [21]. The output represents the median of the predicted cluster.

Lsvm *(Linear SVM)* is a classic approach for imbalanced learning unlike Naïve Bayes. It is a variation of Hn14 by just replacing the classifier. The linear kernel is known to perform well over large datasets within a reasonable time.

Mc *(Majority Class)* is a baseline that always predicts the most frequent class in the training set. It was used as a baseline in previous work [6], [7].

Ss *(Stratified Sampling)* is a baseline which picks a single class randomly biased by the proportion of each class in the training set. Ss is expected to be a strong baseline for a classification task with multiple majority (or close to majority) classes, unlike Mc which originated in binary classification.

Both baselines output a class and not a GPS coordinate. Measures that require a GPS coordinate to measure distance, Acc@161 and mean/median error, were consequently not used to evaluate the baselines.

B. W-NUT Models

W-NUT[3] is a shared task for predicting the location of posts and users from a pre-defined set of cities [19]. We analyze the results of user geolocation for eleven systems submitted by five teams. The top two submissions were based on ensemble learning (Csiro.1) and neural networks (FujiXerox.2), making use of multiple sources of information, including tweets, user self-declared location, timezone values, and other features. One submission used tweet text only (Ibm). Two teams (Aist and Drexel) did not submit a system description.

V. Results

Table I shows results on two sets of systems (Local and W-NUT); PRF (precision, recall, f1-score) are calculated using μ and M averaging; using the output levels city and country. Error distance metrics (Median and Mean) are measured between the home and estimated GPS coordinates of a user. The best scoring systems are highlighted in bold.

We compare which systems are judged best under different evaluations across output levels (i.e. city vs country) and at the same output level (i.e. city or country).

We also compare two forms of evaluation based on metric popularity: most popular metrics (Acc, Acc@161, Median and Mean error distances) and most recent metrics (PRF using μ vs M averaging).

A. Unified output influence using most-popular metrics

The country and city representations are evaluated using two measures: Acc and Acc@161, giving two comparisons. Across those two, the best performing geolocation model is different in 50% and 50% of the comparisons in the Local and W-NUT sets, respectively.

Previous research [6] demonstrated that Hn14 performs better than Rl12 using a city-based representation; though using an alternate representation [12] obtained different results. In terms of accuracy measures, results in the Local section of Table I show that Rl12 and Hn14 are competitive in terms of Acc at the level of city, while Rl12 achieves better results in terms of Acc at the level of country and Acc@161 at both levels. On the other hand, the Lsvm model achieves the best Acc at the level of city only. Hence, standardization enables the comparison of the best performance of each geolocation model regardless of the underlying approach.

We examine the error distance measures to try to understand the observed differences in best systems. There is a gap in performance between the grid based model (Rl12) and the city (Hn14 and Lsvm) or region/cluster (Rm16) based models, see rows 1–3 of Table I. This gap is related to the geographic footprint per unit of the underlying earth representation. Grid-based approaches have lower error distances (as they are calculated from the center of a predicted cell) followed by city-based, and finally region-based approaches, in an ascending order of the geographical area covered by each granularity.

Results in the Local section of Table I show that Mc establishes a strong baseline at the level of country, where it performs much better than Rm16, Lsvm and Ss. Mc is effective here because of the lower number of countries (few hundreds) compared to cities (few thousands). Given the large size of the training set (1.5 million), the sparsity at the country level will be less, still with bias in the distribution, which also explains why the Naïve Bayes based model (Hn14) performs better than Lsvm in this case. In contrast to expectations, the Ss baseline performs much worse than expected, which suggests it should not be considered as a baseline.

B. Imbalance influence using most-recent metrics

The three evaluation measures (PRF) that use the two averaging methods can be compared across city and country giving six μ vs M comparisons. Across those six, the best system is different in 67% and 100% of the comparisons in the Local and W-NUT sets, respectively.

A consistent drop in performance can be seen from μ to M, see columns P_μ to $F1_M$ of Table I. While Rl12 and Hn14 are competitive at the level of Acc, Rl12 tends to have higher precision than Hn14 using micro averaging, and vice versa using macro averaging. Lsvm is another example where Acc is a limited measure when comparing to other systems. While Lsvm achieves the best Acc at the level of city, it tends to have less precision than Rl12 using micro averaging and Hn14 using macro averaging Mc is still competitive at the country level using micro averaging, achieving higher PRF than Rm16 and Lsvm.

If we consider both unified output and imbalance influences, in W-NUT, the Csiro submissions collectively outperform FujiXerox at the level of city across all the evaluation metrics, except for Acc@161 and error distance measures. On the other hand, FujiXerox submissions outperform Csiro at the level of country in terms of accuracy, micro averaging and

[3] https://noisy-text.github.io/2016/geo-shared-task.html

TABLE I

Evaluation based on all metrics at the level of city and country and sorted in a descending order of Acc.

		City								Country									
		Acc	Acc@161	P_μ	R_μ	$F1_\mu$	P_M	R_M	$F1_M$	Acc	Acc@161	P_μ	R_μ	$F1_\mu$	P_M	R_M	$F1_M$	Median	Mean
Local	Lsvm	**0.145**	0.193	0.085	0.068	**0.075**	0.045	**0.040**	**0.039**	0.446	0.448	0.447	0.446	0.447	0.098	0.113	0.099	3656	5936
	Rl12	0.128	**0.228**	**0.114**	0.050	0.070	0.036	0.020	0.023	**0.615**	**0.619**	**0.621**	**0.615**	**0.618**	0.144	**0.138**	**0.133**	**1740**	**3785**
	Hn14	0.127	0.182	0.068	**0.070**	0.069	**0.091**	0.014	0.020	0.599	0.600	0.600	0.600	0.600	**0.241**	0.050	0.068	3128	4489
	Rm16	0.074	0.132	0.030	0.021	0.025	0.007	0.001	0.001	0.315	0.316	0.315	0.315	0.315	0.062	0.015	0.015	5909	5653
	Mc	0.018	0.000	0.018	0.020	0.019	0.000	0.000	0.000	0.523	0.000	0.523	0.524	0.523	0.004	0.007	0.005	—	—
	Ss	0.002	0.000	0.003	0.002	0.002	0.001	0.000	0.000	0.301	0.000	0.302	0.302	0.302	0.007	0.007	0.007	—	—
W-NUT	Csiro.1	**0.529**	0.636	**0.544**	**0.529**	**0.537**	0.545	0.432	0.454	0.798	0.799	0.798	0.798	0.798	0.661	**0.538**	**0.568**	21	1928
	Csiro.2	0.523	0.619	0.544	0.523	0.533	0.555	**0.434**	**0.458**	0.787	0.789	0.788	0.787	0.787	0.653	0.535	0.561	23	2071
	Csiro.3	0.503	0.585	0.529	0.503	0.516	**0.576**	0.422	0.455	0.771	0.773	0.772	0.771	0.771	0.662	0.530	0.560	30	2242
	FujiXerox.2	0.476	0.635	0.481	0.476	0.478	0.358	0.279	0.289	0.866	0.868	0.866	0.866	0.866	**0.692**	0.519	0.562	**16**	1122
	FujiXerox.1	0.464	**0.645**	0.468	0.464	0.466	0.313	0.253	0.253	**0.883**	**0.886**	**0.884**	**0.883**	**0.884**	0.634	0.514	0.542	20	**963**
	FujiXerox.3	0.452	0.629	0.455	0.452	0.453	0.283	0.243	0.237	0.869	0.872	0.869	0.869	0.869	0.621	0.502	0.527	28	1084
	Drexel.3	0.352	0.474	0.367	0.352	0.359	0.348	0.230	0.253	0.686	0.689	0.701	0.686	0.693	0.631	0.494	0.530	262	3124
	Ibm.1	0.225	0.349	0.225	0.225	0.225	0.099	0.049	0.053	0.706	0.707	0.706	0.706	0.706	0.306	0.148	0.169	630	2860
	Aist.1	0.098	0.199	0.103	0.098	0.100	0.123	0.052	0.063	0.562	0.564	0.565	0.562	0.564	0.297	0.107	0.137	1711	4002
	Drexel.1	0.080	0.140	0.082	0.080	0.081	0.062	0.025	0.031	0.354	0.355	0.355	0.354	0.355	0.157	0.072	0.086	5714	6053
	Drexel.2	0.079	0.135	0.082	0.079	0.080	0.056	0.024	0.029	0.435	0.435	0.443	0.435	0.439	0.168	0.072	0.090	4000	6161

error distance measures, and vice versa using macro averaging, except for macro precision (P_M).

VI. Conclusion

We examined the effectiveness of metrics employed in the evaluation of Twitter user geolocation from two key aspects: standardized evaluation process, and compensating bias due to population imbalance through micro vs macro averaging.

A standardized evaluation process, including eight measures over four geographic granularities, allowed the comparison of systems with different earth representations. We demonstrated that different systems were best for different representations. Using one geocoding API ensured a fair comparison and avoided any mismatch of predictions based on different representations. We demonstrated how competitive geolocation models – previously proclaimed inferior – could be equal to state-of-the-art models in terms of accuracy.

Using a micro vs macro comparison revealed the influence of data imbalance across all geographic granularities and the limitations of the most common metrics (accuracy and error distance). The substantial drop in performance using macro averaging showed the quality of user geolocation prediction for applications treating urban and rural locations with the same degree of importance. We therefore suggest using a majority class baseline for the evaluation of more complex models at the level of coarse geographic granularities, state and country in particular, which achieved competitive results.

References

[1] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing, "A latent variable model for geographic lexical variation," in *Proc of EMNLP*, 2010, pp. 1277–1287.

[2] Y. Kryvasheyeu, H. Chen, E. Moro, P. Van Hentenryck, and M. Cebrian, "Performance of social network sensors during hurricane sandy," *PLoS one*, vol. 10, no. 2, 2015.

[3] X. Liu, Q. Li, A. Nourbakhsh, R. Fang, M. Thomas, K. Anderson, R. Kociuba, M. Vedder, S. Pomerville, R. Wudali *et al.*, "Reuters tracer: A large scale system of detecting & verifying real-time news events from twitter," in *Proc of CIKM*. ACM, 2016, pp. 207–216.

[4] M. Dredze, M. J. Paul, S. Bergsma, and H. Tran, "Carmen: A twitter geolocation system with applications to public health," in *AAAI workshop on expanding the boundaries of health informatics using AI (HIAI)*, 2013, pp. 20–24.

[5] B. Hecht, L. Hong, B. Suh, and E. H. Chi, "Tweets from justin bieber's heart: the dynamics of the location field in user profiles," in *Proc of SIGCHI*, 2011, pp. 237–246.

[6] B. Han, P. Cook, and T. Baldwin, "Text-based twitter user geolocation prediction," *JAIR*, pp. 451–500, 2014.

[7] A. Mourad, F. Scholer, and M. Sanderson, "Language influences on tweeter geolocation," in *Proc of ECIR*, 2017.

[8] N. Diakopoulos, M. De Choudhury, and M. Naaman, "Finding and assessing social media information sources in the context of journalism," in *Proc of SIGCHI*, 2012, pp. 2451–2460.

[9] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in *Proc of CIKM*, 2010, pp. 759–768.

[10] E. Rodrigues, R. Assunção, G. L. Pappa, D. Renno, and W. Meira Jr, "Exploring multiple evidence to infer users location in twitter," *Neurocomputing*, vol. 171, pp. 30–38, 2016.

[11] S. Roller, M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge, "Supervised text-based geolocation using language models on an adaptive grid," in *Proc of EMNLP*, 2012.

[12] B. Wing and J. Baldridge, "Hierarchical discriminative classification for text-based geolocation," in *Proc of EMNLP*, 2014.

[13] B. J. Hecht and M. Stephens, "A tale of cities: Urban biases in volunteered geographic information." in *Proc of ICWSM*, vol. 14, 2014, pp. 197–205.

[14] I. Johnson, C. McMahon, J. Schöning, and B. Hecht, "The effect of population and structural biases on social media-based algorithms: A case study in geolocation inference across the urban-rural spectrum," in *Proc of CHI*, 2017.

[15] D. Jurgens, T. Finethy, J. McCorriston, Y. T. Xu, and D. Ruths, "Geolocation prediction in twitter using social networks: a critical analysis and review of current practice," in *Proc of ICWSM*, 2015.

[16] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[17] K. Starbird, G. Muzny, and L. Palen, "Learning from the crowd: collaborative filtering techniques for identifying on-the-ground twitterers during mass disruptions," in *Proc of ISCRAM*, 2012.

[18] D. Jurgens, T. Finethy, C. Armstrong, and D. Ruths, "Everyones invited: A new paradigm for evaluation on non-transferable datasets," in *Proc of ICWSM*, 2015.

[19] B. Han, A. Hugo, A. Rahimi, L. Derczynski, and T. Baldwin, "Twitter geolocation prediction shared task of the 2016 workshop on noisy user-generated text," *WNUT 2016*, 2016.

[20] A. Poulston, M. Stevenson, and K. Bontcheva, "Hyperlocal home location identification of twitter profiles," in *Proc of HT*. ACM, 2017, pp. 45–54.

[21] A. Rahimi, T. Cohn, and T. Baldwin, "pigeo: A python geotagging tool," in *Proc of ACL*, 2016, pp. 127–132.