

A Practical Guide for the Effective Evaluation of Twitter User Geolocation

AHMED MOURAD, School of Computer Science and Information Technology, RMIT University, Australia

FALK SCHOLER, School of Computer Science and Information Technology, RMIT University, Australia

WALID MAGDY, School of Informatics, The University of Edinburgh, United Kingdom

MARK SANDERSON, School of Computer Science and Information Technology, RMIT University, Australia

Geolocating Twitter users—the task of identifying their home locations—serves a wide range of community and business applications such as managing natural crises, journalism, and public health. Many approaches have been proposed for automatically geolocating users based on their tweets; at the same time, various evaluation metrics have been proposed to measure the effectiveness of these approaches, making it challenging to understand which of these metrics is the most suitable for this task. In this paper, we propose a guide for a standardized evaluation of Twitter user geolocation by analyzing fifteen models and two baselines in a controlled experimental setting. Models are evaluated using ten metrics over four geographic granularities. We use rank correlations to assess the effectiveness of these metrics.

Our results demonstrate that the choice of effectiveness metric can have a substantial impact on the conclusions drawn from a geolocation system experiment, potentially leading experimenters to contradictory results about relative effectiveness. We show that for general evaluations, a range of performance metrics should be reported, to ensure that a complete picture of system effectiveness is conveyed. Given the global geographic coverage of this task, we specifically recommend evaluation at micro versus macro levels to measure the impact of the bias in distribution over locations. Although a lot of complex geolocation algorithms have been applied in recent years, a majority class baseline is still competitive at coarse geographic granularity. We propose a suite of statistical analysis tests, based on the employed metric, to ensure that the results are not coincidental ¹.

CCS Concepts: • **General and reference** → **Evaluation**;

Additional Key Words and Phrases: Twitter, User Geolocation, Effective Evaluation, Statistical Analysis

ACM Reference Format:

Ahmed Mourad, Falk Scholer, Walid Magdy, and Mark Sanderson. 2019. A Practical Guide for the Effective Evaluation of Twitter User Geolocation. *ACM Trans. Soc. Comput. 1*, 1, Article 1 (January 2019), 23 pages. <https://doi.org/10.1145/3352572>

1 INTRODUCTION

Geolocating Twitter users is needed in many social media-based applications, such as identifying geographic lexical variation [Eisenstein et al. 2010; Han et al. 2014], managing natural crises [Kryvasheyev et al. 2015], gathering news [Liu

¹The code for the evaluation framework detailed in this article can be found on: <https://bitbucket.org/amourad/geoloceval.git>

Authors' addresses: Ahmed Mourad, School of Computer Science and Information Technology, RMIT University, 124 La Trobe Street, Melbourne, VIC, 3000, Australia, ahmed.mourad@rmit.edu.au; Falk Scholer, School of Computer Science and Information Technology, RMIT University, 124 La Trobe Street, Melbourne, VIC, 3000, Australia, falk.scholer@rmit.edu.au; Walid Magdy, School of Informatics, The University of Edinburgh, 10 Crichton Street, Edinburgh, EH8 9AB, United Kingdom, wmagdy@inf.ed.ac.uk; Mark Sanderson, School of Computer Science and Information Technology, RMIT University, 124 La Trobe Street, Melbourne, VIC, 3000, Australia, mark.sanderson@rmit.edu.au.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2019 Association for Computing Machinery.

Manuscript submitted to ACM

Manuscript submitted to ACM

et al. 2016; Schwartz et al. 2015; Zubiaga et al. 2013], and tracking epidemics [Broniatowski et al. 2013; Dredze et al. 2013]. While users can record their location on their profile, Hecht et al. [2011] reported that more than 34% record fake or sarcastic locations. Twitter allows users to GPS locate their content, however, Han et al. [2014] reported that less than 1% of tweets are geotagged. Inferring user location is therefore an important field of investigation.

Each geolocation application has different needs, which might require evaluation from several perspectives. However, current evaluation practices focus on a few measures introduced by Eisenstein et al. [2010]. These measures were shown to be biased towards densely populated (urban) locations [Mourad et al. 2017], e.g. the accuracy over urban locations will dominate the overall measure. Such measures may be unsuitable to evaluate applications that treat urban and rural locations with the same degree of importance: e.g. searching for sources to cover local news [Liu et al. 2016; Schwartz et al. 2015; Starbird et al. 2012], monitoring natural disasters in rural areas [Kryvasheyev et al. 2015], or tracking epidemics in rural cities [Dredze et al. 2013].

Moreover, evaluation at multiple levels of geographic granularity is not widely used despite it being required by some applications. For instance, Diakopoulos et al. [2012], in determining requirements from journalists for identifying eyewitnesses from social media, found that aggregating predicted eyewitness location at different scales was requested, e.g. city, state or country. Similarly, Dredze et al. [2013] presented a geolocation prediction system (Carmen) for influenza surveillance, which predicts a structured location at different granularities.

The evaluation of geo-inference methods is affected by many factors, such as dataset availability, pre-processing, 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 proposes a guide for the evaluation of Twitter user geolocation through the following contributions:

- We standardize the evaluation process for models to ensure the 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. In particular, we find that a wide range of metrics and a majority class baseline should be used for the evaluation of more complex geolocation models.
- We assess the effectiveness of current evaluation metrics using rank correlations. We demonstrate that the ranking of user geolocation systems varies based on the evaluation metric and geographic granularity. In some cases, some of the most common evaluation metrics are redundant and should not be used simultaneously.
- We validate the effectiveness of the proclaimed state-of-the-art geolocation systems using statistical significance testing. We propose a suite of statistical significance tests suitable for the task at hand, based on the employed metric.
- We study the degree to which metrics can lead to contradictory, yet statistically significant results, concluding that systems should be evaluated using a range of measures.

This paper builds upon our own previously-published work [Mourad et al. 2018] with more statistical analysis (effectiveness and significance) through the last three contributions. Our results 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.

2 RELATED WORK

Zheng et al. [2018] surveyed previous research on the geolocation of Twitter users. They reviewed and summarized all geolocation methods and evaluation metrics employed from an empirical perspective. In this work, we present the metrics from two different perspectives. We briefly introduce the different approaches of inferring a user’s location in §2.1. In §2.2, we discuss how evaluation metrics for Twitter user geolocation evolved over time. This presentation explains the original intuition behind each metric, reveals the decisions taken by subsequent researchers and the impact of such choices on the evaluation process. We also chart the limitations and commonality of each metric. Examination of biases in social media are detailed in §2.3. In §2.4, we survey efforts to standardize the evaluation process of Twitter user geolocation and assess the effectiveness of the evaluation metrics employed.

2.1 Geolocation Methods

Previous research inferred the location of a Twitter user from different sources of information, namely tweet-text, user’s social-network (e.g. followers, following, mentions) and meta-data (e.g. profile location, tweet timezone). Most geolocation methods rely on the first two sources and hence are known in research as text-based and network-based approaches. Text-based methods tend to address geolocation inference as a classification task. They rely on identifying location indicative words (aka local words) over a predefined set of locations (e.g. administrative regions or grid cells). Location sparsity is, therefore, a limitation of text-based approach. The intuition behind network-based methods is that a user is geographically close to their friends. However, if a user is not covered in the training network, a geolocation model will not be able to infer their location. Hence, recent research is focusing on a hybrid approach which combines both approaches.

Jurgens et al. [2015b] constructed a benchmark for the network-based approach. They re-implemented the state-of-the-art models, back at the time, and made it publicly available to the research community. In the process to do that, they constructed their own dataset to train the models and ensure fairness of comparison. Recent research, however, still prefer to reconstruct benchmark datasets which were created by the text-based research to evaluate their models, than constructing their own datasets and retraining the available models. We, therefore, choose to focus on text-based approaches to set a reliable benchmark process for the task of Twitter user geolocation regardless of the underlying approach. We highlight the pitfalls of reconstructing Twitter datasets and comparing to results reported in previous research.

2.2 Geolocation Metric Evolution

Table 1 details a chronological ordering of Geolocation Metrics, which we initially overview and then describe in more detail.

2.2.1 Overview. Evaluation of geolocation models was initially measured using *Median* and *Mean* error distances between an estimated and true location [Eisenstein et al. 2010]. The researchers also used accuracy (*Acc*) at the level of states (49) and regions (4). The choice of spatial granularity was influenced by the use of ground truth datasets, which were drawn from the US (the country with the majority of Twitter users in 2010), and for a better interpretability of the results compared to error distance.

Several metrics based on accuracy and/or error distance were introduced. Backstrom et al. [2010] evaluated performance based on the fraction of predictions within x kilometers from the true location using a Cumulative Distribution Function (*CDF*) for all values of x within 10,000 km. Accuracy within x miles from the original city was introduced by

Cheng et al. [2010], as was accuracy within the top k cities ($Acc@k$), and at the level of country by [Hecht et al. 2011]. Priedhorsky et al. [2014] introduced three new metrics, based on the error distance and the probability of estimation.

Rodrigues et al. [2016] reported precision and recall at the level of each city and an overall macro-F1 metric, which was further extended by Mourad et al. [2017] to consider micro, weighted, and macro averaging techniques at the level of the three metrics. Other research employed a combination of these measures, as described in Table 1. Given that $Acc@k$ [Cheng et al. 2010] and the metrics introduced by Priedhorsky et al. [2014] were employed only in their respective research, they were not presented in the table.

2.2.2 Accuracy Error. Cheng et al. [2010] showed empirically that 30% of users are placed within 10 miles of their true location, and 51% within 100 miles after exploring a range from 0 to 4,000 miles.

Subsequent research used the (perhaps) arbitrarily chosen range of 100 miles (161 km) to measure accuracy ($Acc@161$) [Han et al. 2014; Roller et al. 2012; Wing and Baldrige 2014]. Note, the variance in accuracy with respect to the range was tested on a dataset limited to the US. Using a population-based global earth representation,² the average distance between cities and their neighbours was found to be in the range of 32–46 miles [Mourad et al. 2017], less than half the 100 miles threshold. A system which predicts the location of a user two cities away from his/her home location could be as accurate as a system which predicted the location one city away from the true location. This choice of the tolerance distance questions the appropriateness of $Acc@161$ as a measure that suits global geographic models. A somewhere arbitrary threshold is also found in the metric AUC , introduced by Jurgens et al. [2015b], which quantifies the graph generated by a CDF into a single number. This number is generated using the range value of 10,000km.

Error distance measures (Mean and Median) can be more accurate than $Acc@161$ because they are measured based on the raw estimations of geolocation models without any approximation (discretization, e.g. map to a region such as a city or country). However, they can exhibit a large variability on the measured results and limit evaluation at multiple levels of geographic granularity, which is required by some geolocation applications as mentioned in §1. On the other hand, metrics based on accuracy and error distance (e.g. $Acc@161$, CDF , and AUC) strongly depend on the distance thresholds that are selected.

2.2.3 Dataset Availability. Table 1 (columns #Users and #Tweets) shows a large disparity in the sizes of test datasets. Although Twitter provides access to the public data generated by users, the terms of service limits the sharing of this data to only tweet IDs. Any attempt to reconstruct a dataset used in previous research will be subject to decay, i.e. some tweets will disappear because they have been deleted. In an effort to solve this issue, two approaches were proposed.

First, Jurgens et al. [2015a] proposed an evaluation framework where the dataset is hosted by a single operator. An experimenter submits a request to the host along with a code. However, the cost to the host of maintaining this service, the difficulty of the development process for the experimenter, and the unprotected intellectual property—the ownership of the code—meant this proposal was not taken up.

Second, Han et al. [2016] provided a dataset of tweet IDs for a user geotagging shared task (named WORLD). However, one of the participant teams pointed out that the re-constructed dataset was missing ~25% of the data [Jayasinghe et al. 2016]. Systems are therefore highly likely to be trained on different datasets, based on the time they were re-constructed. Subsequent research [Miura et al. 2017] highlighted the same issue using two different datasets (UTGeo and WORLD).

2.2.4 Earth-representation. The importance of measures was illustrated when two different models were each found to perform better using different reverse-geocoding technique. Han et al. [2014] demonstrated that a multinomial

²<https://github.com/tq010or/acl2013>

Table 1. An overview of past work. Precision, recall and f1-score are combined in the column PRF. For datasets, names in bold represent the original dataset, empty #Users and #Tweets cells means the size of the reconstructed dataset was not reported in the respective work, and Scope refers to the geographical coverage. For testset, percent is the percentage of users in the testset to the whole collection; #Tpu is the minimum number of tweets per test user.

	Evaluation Metrics					Datasets				Testset	
	Acc	Acc@161	Median	Mean	PRF	Name	#Users	#Tweets	Scope	#Users	#Tpu
Eisenstein et al. [2010]	✓		✓	✓		GeoText	9.5k	380k	US	1.9k (20%)	
Backstrom et al. [2010]		CDF				BACKSTROM			US		
Cheng et al. [2010]	✓	0-4k	✓	✓		Cheng	135k	4M	US	5k (3.7%)	1000+
Wing and Baldrige [2011]			✓	✓		GeoText			US		
Roller et al. [2012]		✓	✓	✓		GeoText			US		
Ahmed et al. [2013]			✓	✓		UTGeo	449k	38M	Nth Am	10k (2.22%)	
Han et al. [2014]	✓	✓	✓	✓		GeoText			US		
						UTGeo	1.4M	12M	Nth Am		10+
Wing and Baldrige [2014]		✓	✓	✓		WORLD			Global	10k (0.71%)	
		✓	✓	✓		UTGeo			Nth Am		10+
			✓	✓		WORLD			Global		
Priedhorsky et al. [2014]			✓	✓		GeoText	9.5k	380k	US	1.9k (20%)	
Jurgens et al. [2015b]		Auc	✓	✓		Jurgens					
Rodrigues et al. [2016]	✓				✓	Rodrigues	11.8k		Brazil		
Han et al. [2016] [W-NUT]	✓		✓	✓		WORLD	1.4M	12M	Global	10k (0.71%)	10+
			✓	✓		UTGeo			Nth Am	10k	
Rahimi et al. [2017, 2016, 2018]		✓	✓	✓		WORLD	1.4M	12M	Global	10k (0.71%)	10+
	✓	✓	✓	✓		UTGeo	279k	23.8M	Nth Am	10k	
Miura et al. [2017]	✓	✓	✓	✓		WORLD	782k	9.03M	Global	10k	10+
	✓	✓	✓	✓	✓	WORLD	947k		Global		
Mourad et al. [2017]						TwArchive	1.5M		Global		
						GeoText	9.5k	>370k	US	1.9k (20%)	
Do et al. [2017]		✓	✓	✓		UTGeo	450k	38M	Nth Am	10k	
						WORLD	1.4M	12M	Global	10k	
		✓	✓	✓		GeoText	9.5k	380k	US	1.9k (20%)	
Ebrahimi et al. [2018]		✓	✓	✓		UTGeo	450k	38M	Nth Am	10k	
						WORLD	1.4M	12M	Global	10k	

naïve bayes model with feature selection performs better than logistic regression [Roller et al. 2012] using city-based representation, Wing and Baldrige [2014] demonstrated the opposite using uniform grids.

2.3 Underlying bias

Social media is known to have substantial population biases [Mislove et al. 2011]. They relied on the US census data to reveal the sampling bias in Twitter data based on the demographics of Twitter users, namely geographic distribution, gender and race/ethnicity. Not many researchers explored the impact of this bias on either determining the most effective models or evaluation metrics. Focusing on the geographic bias over the urban-rural spectrum, [Hecht and Stephens 2014] explored three of the most common sources for geotagged information, — Twitter, Flickr and Foursquare. They showed that there is a population bias towards urban regions.

The first attempt to assess the influence of population bias on the existing models — Twitter user geolocation — was done by [Pavalanathan and Eisenstein 2015]. They explored the influence of Twitter user demographics — gender and age — on the tasks of detecting lexical variation over geographic regions and text-based Twitter user geolocation, yet relying on accuracy per category (e.g. male vs female) for evaluation. Johnson et al. [2017] further explored the impact of geographic bias on the latter task. They differentiated between population bias, and structural bias introduced by algorithmic design. To assess the impact of each of these biases, they explored different sampling techniques on a US rural-urban county based dataset. They demonstrated that existing geolocation approaches perform significantly worse for rural areas than for urban.

Relying on an external gazetteer (US census data, which might not be available for other countries), consolidating geographic regions into two classes only (rural-vs-urban) and finally evaluation of individual categories (e.g. accuracy of male vs female) limits the scalability of the analysis. Most of the recent work relies on datasets with global geographic coverage, with hundreds and thousands of classes, and ignores the existence of biases while designing or evaluating their models. We, therefore, believe that focusing on an enhanced and scalable evaluation metrics (macro averaging in specific) should come first; to reveal such biases and assess their impact on the design of geolocation algorithms.

2.4 Comparing Geolocation Evaluation Metrics

Studies have analyzed the effectiveness of evaluation metrics of Twitter user geolocation. Jurgens et al. [2015b] conducted a comparative analysis of 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. More recent work by Mourad et al. [2017] pointed out that accuracy measures are biased towards locations with a large population. Although they employed a wide range of metrics, their work was limited to a single geolocation model while focusing on the influence of language rather than the effectiveness of the evaluation measures.

In this paper, we focus on the effectiveness of geolocation evaluation regardless of the underlying geolocation approach or the language of text, which entails generalization challenges discussed in the next section. We evaluate the relative performance of fifteen geolocation models and two baselines using all the metrics in Table 1.

3 STANDARDIZED EVALUATION

In considering how to build a standardized evaluation, first, alternate metrics are described that address data imbalance. Second, we examine significance tests to assess the statistical differences between the geolocation models under study. Finally, a unified output format and reverse-geocoding method are employed to assure the fairness of comparisons.

3.1 Evaluation 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) [Johnson et al. 2017]. Hence, we investigate conventional measures for multi-class categorization [Sebastiani 2002; Sokolova and Lapalme 2009], which were included partially by Rodrigues et al. [2016] and fully by Mourad et al. [2017] 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 [Diakopoulos et al. 2012]. *Recall* is favored in situations such as when these journalists want to increase the search pool [Starbird et al. 2012]. Both scenarios focus on a single location, where comparison at the micro and macro levels is essential.

Evaluation metrics are categorized into three groups. Continuous evaluation is based on the estimated GPS coordinates (p) of a user (u), and represented by median and mean error distances from the original gps-point (\hat{p}). Discrete evaluation is based on the resolved locations of a user (l and \hat{l} are the predicted and true locations, respectively), and represented by accuracy, precision, recall, and f1-score using micro and macro averaging. Mixed evaluation is based on a combination of continuous and discrete metrics, and represented by accuracy within 100 miles of the true location.

The evaluation metrics considered in this study are defined as:

Continuous evaluation.

$$\begin{aligned} ErrorDistance(u) &= great_circle\{\hat{p}, p\} \\ Median &= median_{i=0}^{n_{users}-1}\{ErrorDistance(u_i)\} \\ Mean &= \frac{1}{n_{users}} \sum_{i=0}^{n_{users}-1} \{ErrorDistance(u_i)\} \end{aligned}$$

Discrete evaluation.

$$\begin{aligned} Acc &= \frac{1}{n_{users}} \sum_{i=0}^{n_{users}-1} 1(l_i = \hat{l}_i) \\ P_M &= \frac{1}{n_{locations}} \sum_{i=0}^{n_{locations}-1} P(l = l_i, \hat{l} = l_i) \\ R_M &= \frac{1}{n_{locations}} \sum_{i=0}^{n_{locations}-1} R(l = l_i, \hat{l} = l_i) \\ F_M &= \frac{1}{n_{locations}} \sum_{i=0}^{n_{locations}-1} F_\beta(l = l_i, \hat{l} = l_i) \end{aligned}$$

For micro averaging, all precision, recall and f1-score are identical to accuracy [Pedregosa et al. 2011].

Mixed evaluation.

$$Acc@161(p, \hat{p}) = \frac{1}{n_{users}} \sum_{i=0}^{n_{users}-1} 1(ErrorDistance u_i \leq 161km)$$

3.2 Significance Tests

Dror et al. [2018] highlighted the importance of applying statistical significance tests in the field of Natural Language Processing (NLP) to ensure that the experimental results are not coincidental. Given the range of NLP tasks and effectiveness metrics that can be applied, different statistical tests are needed. Based on the decision tree algorithm provided by Dror et al. [2018] for statistical significance test selection, we choose a combination of parametric (t-test) and sampling-free non-parametric tests (sign test, and Wilcoxon). Given the large size of our dataset, parametric tests are applicable because the test statistic follows the normal distribution, and sampling-free tests are computationally less expensive than sampling-based non-parametric tests.

Although Dror et al. [2018] surveyed a large number of NLP papers on different tasks, they did not consider the category frequencies of the datasets. We, therefore, follow the recommendation of Yang and Liu [1999] who considered the appropriate choice of significance tests to measure the statistical differences between categorization models trained on datasets with skewed category distributions. Two types are considered: **micro** and **macro tests**.

The **micro tests** considered in this study are the micro sign test (s) and proportions z-test (p) [Yang and Liu 1999]. The former is a binomial test for comparing two systems, A and B, based on binary decisions for all user/location pairs. The latter is used for measures which are proportions: accuracy, precision, and recall. The z-test computation for precision and recall is based on performance scores using micro averaging.

The **macro tests** include macro sign test (S), macro t-test (T), and macro t-test after rank transformation (T' , a.k.a Wilcoxon) [Yang and Liu 1999]. Macro tests were originally based on F1 scores per category as a unit measure, but we employed them for precision and recall as well. The S-test is a binomial sign test used to compare two systems, A and B, based on the paired F1 values for individual locations. While the S-test reduces the influence of outliers, it may be insensitive in performance comparison because it ignores the magnitude of differences between F1 values. Insensitivity issues are resolved in T by considering the absolute differences between paired F1 values in a relevance t-test. However, T becomes sensitive when F1 values are unstable, specifically for low frequency locations. Finally, the Wilcoxon T' provides a compromise between S-test and T by considering the rank differences between paired F1 values for individual locations.

We use two-sided versions of the tests, as they avoid prior expectation about the direction of the effect and are more conservative.

3.3 Unified Output and Reverse-Geocoding

When comparing models, it is necessary to train and test on the same dataset and to use models that output the same earth representation. Figure 1 shows an example of an unfair comparison between models producing different outputs on the same dataset. Assume we have two models: A and B . A represents the earth as polygons (green outlined cells) and B represents the earth as cities. A user's home location is identified as polygon x inside city Z (the orange area). Now assume A predicted the location of this user as 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.

In order to avoid such inconsistency, we unified the output of all the models to be GPS coordinates as suggested by [Jurgens et al. 2015a]. We additionally resolved the coordinates to a location using a single online reverse-geocoding API³ before evaluation. Using a single reverse-geocoding API not only guarantees a fair comparison over the same set of

³There is a trade-off between replicability, efficiency and cost, when choosing a reverse-geocoding API. An offline reverse-geocoding would be fast, but requires implementation and sharing the code base. On the other hand, online reverse-geocoding is easy to consume, but limited by a specific number of

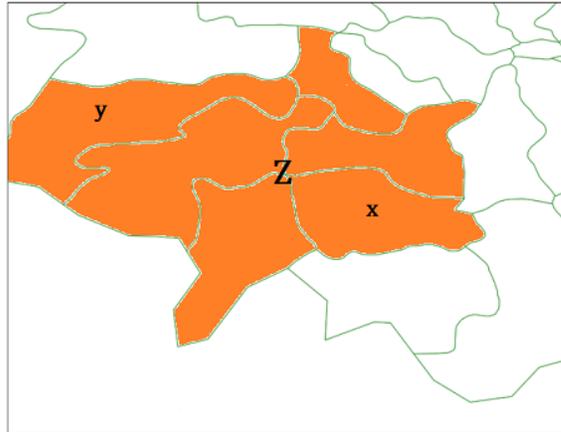


Fig. 1. Example of unfair comparison between systems with different underlying earth representations. Cell x is the home location of a user and cell y is the predicted location by system A. The orange cells represent the home and predicted city (Z) of a user by system B.

locations (classes), it also allows evaluation over different granularities. In this work, we report the model performance at city and country level. We calculated county and state level as well, but trends are consistent.

4 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, to assess the robustness of our proposed metrics [Han et al. 2016]. Although the published results for participating models were evaluated at city level only, we were able to infer output at country level based on information released by W-NUT organizers.

4.1 Local Models

4.1.1 Data Collection Method. We employed a geographically global geotagged tweet collection, **TwArchive**, holding content since 2013⁴ 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` [Lui and Baldwin 2012]. Non-geotagged and duplicate tweets were removed using user id and tweet text. For the sake of a standard evaluation, users with unresolved home location—based on the model that accepts home locations in the form of cities instead of GPS coordinates [Han et al. 2014]—were removed from the dataset. The total number of users and tweets after pre-processing is ~ 1.5 million and ~ 3.1 million respectively.

4.1.2 Ground Truth. The home location of a user was identified as the geometric median of their geotagged tweets [Jurgens 2013]. Such a point is the minimum error distance to all locations of a user. The median has been shown to be more accurate in identifying the home location of a user at a finer granularity than other approaches [Poulston et al. 2017]. The distance between any two GPS points is measured using the great circle distance method.

requests per day. Free APIs have a small limit, e.g. 2.5k requests per day for Nominatim, While commercial APIs have a larger limit, e.g. 100k requests per day for Google Reverse-Geocoding API V3. It took two weeks to reverse-geocode our local dataset of the size 1.5M users, using Google API.

⁴<https://archive.org/details/twitterstream&tab=collection>

4.1.3 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.

Roller et al. [2012] (RL12) proposed an adaptive grid-based representation with a trained probabilistic language model per cell. Each cell has the same number of users, but a different geographical area. We employ their best reported parameter values for constructing the grid to retrain their model⁵ on our local dataset. The output represents the centroid of the predicted cell.

Han et al. [2014] (HN14) locates users to one of 3,709 cities. 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 reported feature selection method, Information Gain Ratio. The output represents the centre of the predicted city.

Rahimi et al. [2016] (RM16) assigns a user to one of 930 non-overlapping geographic clusters based on the similarity of content. Their geotagging tool, Pigeo⁶, allowed retraining their text-based model on our local dataset. The output represents the median of the predicted cluster.

Linear SVM (Lsvm) is a classic approach for imbalanced learning unlike Naïve Bayes. It is a variation of HN14 by replacing the classifier. The linear kernel is known to perform well over large datasets within a reasonable time.

Majority Class (Mc) is a baseline that always predicts the most frequent class in the training set. *Yang [1999]* pointed out that in the case of a low average training instances per category (which applies here) the *majority class trivial classifier* tends to outperform all non-trivial classifiers. It was used as a baseline in previous work [*Han et al. 2014; Mourad et al. 2017*].

Stratified Sampling (Ss) 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 were implemented using scikit dummy classifier [*Pedregosa et al. 2011*] and output a class, 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.

4.2 W-NUT Models

W-NUT⁷ is a shared task for predicting the location of posts and users from a pre-defined set of cities [*Han et al. 2016*]. We analyze the results of eleven systems in the user geolocation prediction task (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 description of their submissions.

5 RESULTS

Table 2 details the results of our experiments on two sets of systems (LOCAL and W-NUT) across all metrics mentioned in Table 1; 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 for each metric are highlighted in bold.

⁵https://github.com/utcompling/textgrounder/wiki/RollerEtAl_EMNLP2012

⁶<https://github.com/afshinrahimi/pigeo>

⁷<https://noisy-text.github.io/2016/geo-shared-task.html>

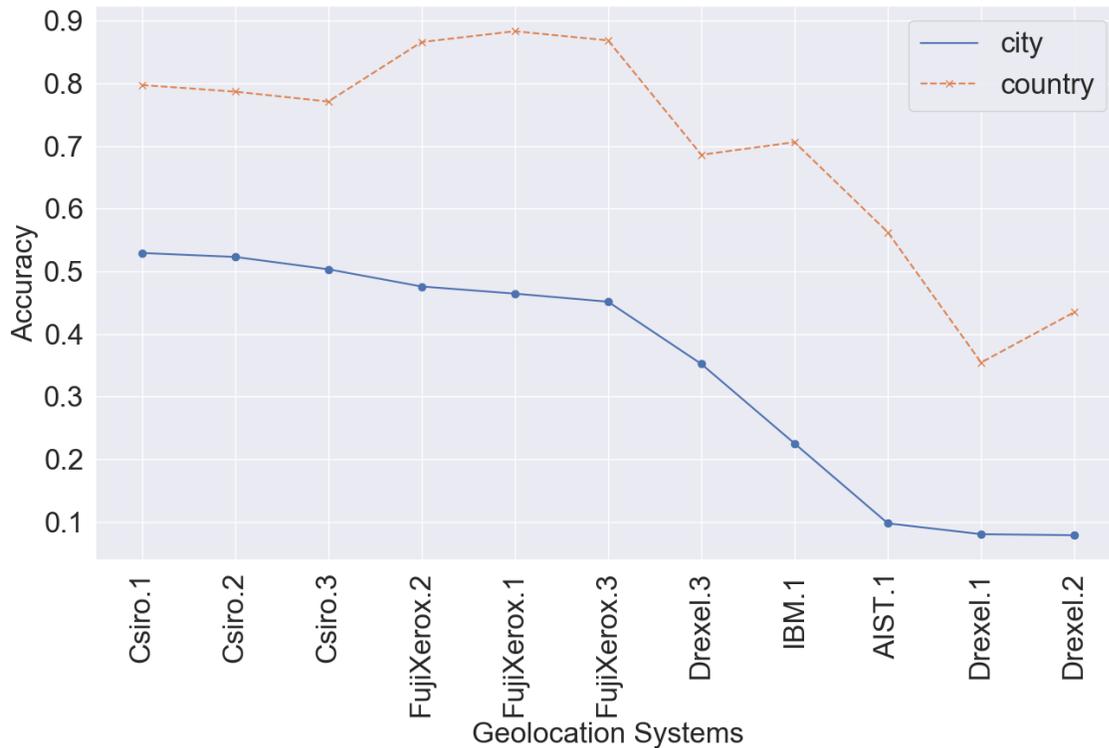


Fig. 2. Evaluation of W-NUT based-on accuracy at the levels of city and country, ordered by city in a descending order.

We first compare which systems are judged best under different evaluations, next we examine rank correlations of systems, and finally study significant differences. For each experiment, we compare across output levels (i.e. city vs country) and at the same output level (i.e. city or country).

5.1 Best system

We compare two forms of evaluation based on metric commonality as shown in Table 1: most common metrics (Acc, Acc@161, Median and Mean error distances) and alternate metrics (PRF using μ vs M averaging).

5.1.1 Unified output influence using most-common metrics. The country and city representations are evaluated using two measures: Acc and Acc@161, which report different best performing geolocation models in the LOCAL and W-NUT sets at the city level, respectively. In terms of accuracy measures, results in the LOCAL section of Table 2 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 .

To further illustrate the differences found when using city and country representations, the W-NUT systems, measured using Acc, are shown in Figure 2. Standardization enables the comparison of the best performance of each geolocation model.

Table 2. Evaluation based on all metrics at the level of city and country and sorted in a descending order of Acc.

	City				Country				Median	Mean									
	Acc	Acc@161	P_μ	R_μ	$F1_\mu$	P_M	R_M	$F1_M$			P_μ	R_μ	$F1_\mu$	P_M	R_M	$F1_M$			
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	0.788	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	

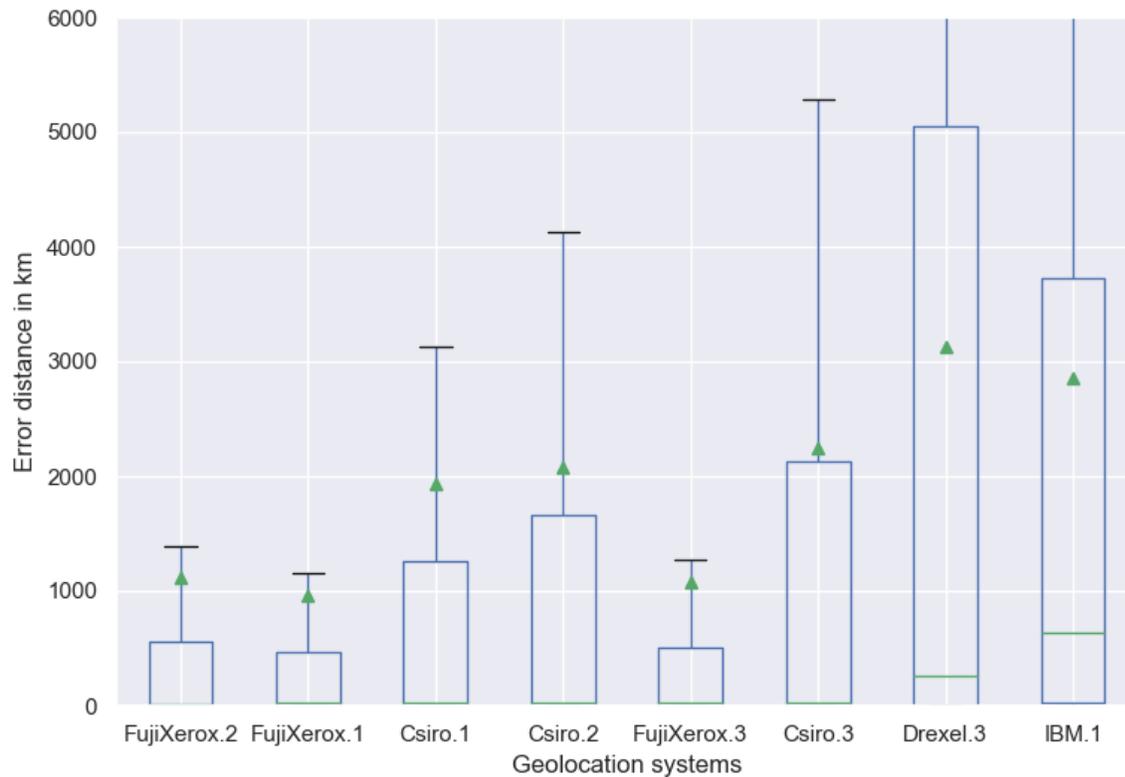


Fig. 3. Evaluation of W-NUT based-on error distance metrics (Median and Mean) in km.

We examine the error distance measures to try to understand the observed differences in best LOCAL system. 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 Table 2. This gap is related to the geographic footprint per unit of the underlying earth representation. Grid-based approaches tend to have the lowest error distances (because they are calculated from the center of the predicted cell), followed by city-based, and finally region-based approaches.

For the W-NUT shared task, we observe that the FUJIXEROX submissions tend to have slightly better Acc@161 at the level of city than the CSIRO submissions (see Table 2). At the level of country, however, the FUJIXEROX submissions achieve much better results than CSIRO, which is correlated to the gap in the mean error distance (in favor of FUJIXEROX models) despite having competitive median error distance, as we will show later. Note that the original WNUT shared task did not evaluate the participating systems at the level of country.

The distribution and results for the error distance measures are represented in more detail using a box plot in Figure 3. The green triangles represent the mean error distance for each system. An upper threshold distance of 6,000 km was applied and the worst three systems (AIST.1, DREXEL.1, and DREXEL.2) were excluded so that details can be seen. We can observe the large variance in 50-75% percentile between FUJIXEROX submissions and CSIRO. Previous research [Han et al. 2014; Melo and Martins 2017] promoted the usage of median error distance to evaluate user geolocation because it is more robust to outliers than the mean, and easy to interpret the results in comparison to accuracy. However, the

boxplot quantifies the variance in error distance, and 25% can not be considered as outliers in this case. The mean error distance therefore is a more effective measure than the median in this context. FUJIXEROX and CSIRO submissions have competitive results in terms of the median error distance, while FUJIXEROX submissions are much better in terms of the mean error distance and have less variance in their estimations.

Results in the LOCAL section of Table 2 show that the two baselines for the locally deployed geolocation models (Mc and Ss) perform poorly at the level of city. In contrast, Mc establishes a strong baseline at the level of country, where it performs much better than RM16, Lsvm and Ss. Mc is effective at the level of country 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. The Ss baseline performs poorly, which suggests it should not be considered as a baseline. At this stage, the use of a simple Mc baseline and Acc did not reveal the influence of imbalance as [Yang 1999] suggested. Therefore, we consider evaluation using different averaging techniques and alternative measures to provide a better insight into the influence of imbalance.

5.1.2 Imbalance Influence using alternate 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 2. While RL12 and HN14 are competitive at the level of Acc, RL12 tend 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, yet has higher recall achieving the best F1-score among all systems in LOCAL. Mc is still competitive at the country level using micro averaging, achieving higher PRF than RM16 and Lsvm.

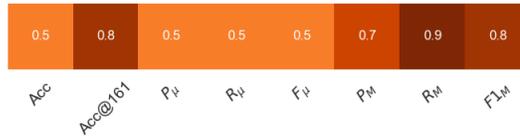
If we consider both unified output and imbalance influences, in W-NUT, collectively the CSIRO submissions outperform FUJIXEROX at the level of city across all the evaluation metrics, except for Acc@161 and error distance measures. On the other hand, FUJIXEROX submissions outperform CSIRO at the level of country in terms of accuracy, micro averaging and error distance measures, and vice versa using macro averaging, except for macro precision (P_M).

To summarize the best system analysis, we demonstrated (in §5.1.1) that unifying the output format and reverse-geocoding locations before evaluation are essential to ensure the fairness of comparison. A majority class baseline is recommended at the country level in the case of using Acc or micro averaging method. The alternate metrics (macro averaging in specific) should be used to evaluate the influence of data imbalance on the quality of geolocation prediction. The question now is: how to quantify the effectiveness of using different evaluation measures?

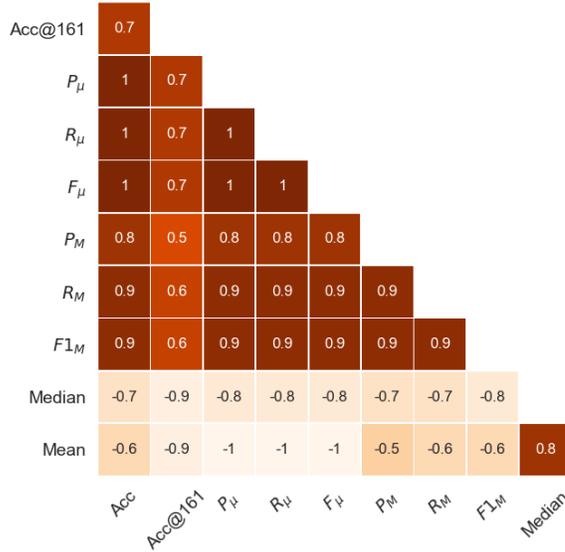
5.2 Rank correlations

Kendall's τ is a correlation measure that quantifies the agreement between two ranked lists. We calculated τ_B for all combinations of the employed metrics, see Figure 4. Note, because the optimal value for distance metrics is 0 and the optimal value for the other metrics is 1, the optimal correlation between those two is -1; the optimal correlation between the non-distance metrics is 1. As the LOCAL collection only includes four non-baseline systems, the range of τ_B values is limited, we therefore focus our analysis on the W-NUT data.

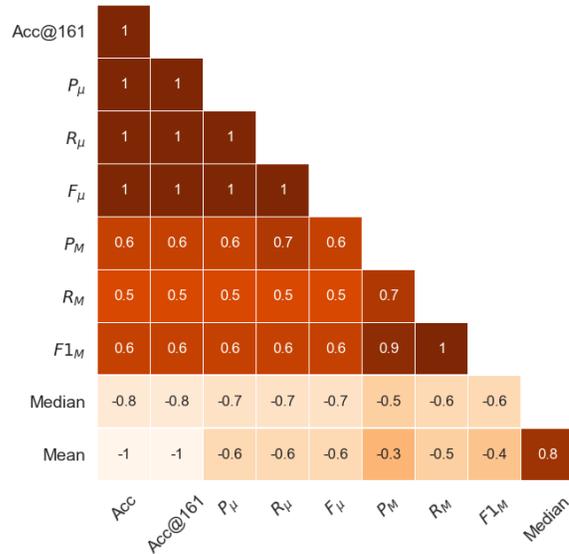
A strong correlation of any metric across different geographic granularities indicates the consistency of such a measure in ranking geolocation models. On the contrary, a strong correlation between any two metrics at the same level



(a) Rank correlations across City and Country. Median and Mean error distances are excluded because geographic granularity is not applicable.



(b) Rank correlations at the level of City.



(c) Rank correlations at the level of Country.

Fig. 4. Kendall's τ_B rank correlations between pairs of effectiveness metrics for the W-NUT collection, $p \leq 0.05$.

of geographic granularity (e.g. city) indicates less benefit from using both metrics at the same time. Hence, a moderate or weak correlation suggests using both measures is important so a more complete picture of system effectiveness is conveyed.

Considering city vs country (Figure 4a), we observe a weak correlation between ranking models across city and country using the commonly used Acc and micro averaging measures. Using macro averaging measures, a strong correlation exists, similarly for Acc@161. This finding suggests that macro measures and Acc@161 are more robust for comparison across geographic granularities.

Considering micro vs macro at the city level (Figure 4b), we observe strong correlations across the three micro and macro measures (0.8, 0.9, 0.9). Acc@161, median and mean error distances also have mutual strong correlations. On the other hand, Acc, median and mean error distances have weak correlations. This contrast in correlations, therefore, suggests not relying solely on measures driven from the error distance (Median, Mean, Acc@161) because they depend on the underlying earth representation, i.e. grid-based representation will always achieve better results than city and cluster based representations in terms of these metrics, even if the accuracy of city and cluster based models are better.

Considering micro vs macro at the level of country (Figure 4c), we observe moderate correlations across the three micro and macro metrics (0.6, 0.5, 0.6). The most common metrics (Acc, Acc@161, Median and Mean) and micro averaging metrics tend to have strong correlations. On the other hand, they tend to have moderate correlations with macro averaging metrics, except for the Median error distance. Therefore, a combination of micro and macro metrics or most common metrics and macro metrics is recommended.

5.3 Statistical Significance

As was apparent from the system effectiveness scores in Table 2, some of the results occurred within a close range. Statistical significance tests are therefore important to establish confidence that differences are not just due to chance.

Following [Moffat et al. 2012], the outcome of a significance test will be categorized into one of two classes. Given two models A and B, calculated for two metrics X and Y, suppose that significance tests are run on the models' outputs using both metrics. If they show statistically significant results for both metrics that system A is better than system B (or vice versa), that would be considered a statistically significant active agreement (SSA). Statistical significant differences, but with contradicting superiority on systems, would be considered a statistically significant active disagreement (SSD).

Figures 5a and 5b summarize the results of the statistical significance tests for the W-NUT collection at the city and country levels, respectively. Each figure summarizes the significant agreements and disagreements. The diagonal values represent the percentage of systems pairs that are significantly different based on a single metric (discriminative power), the values above the diagonal show the percentage of SSA, the values below the diagonal show the percentage of SSD. As can be seen, there are many more agreements than disagreements.

Considering city vs country, we observe that the discriminative power of the evaluation metrics (on the diagonal) and the percentage of SSA at the city level (above the diagonal) are always over 80% (see Figure 5a). They are much lower at the level of country for all comparisons involving macro tests (31–78%, see Figure 5b), which suggests that there is no huge difference in performance between geolocation models. On the contrary, the percentage of SSD (below the diagonal) at the level of country is much lower than at the city level. These results support the importance of using macro metrics for cross granularity evaluation suggested in the previous section.

Considering precision and recall, at the city level, the discriminative power (on the diagonal) using macro averaging is better than micro; macro averaging is able to capture more statistically significant differences for both precision and recall. At the country level, the opposite is true: micro measures are more discriminative than macro. The percentage of

s- R_{aw}	98	93	89	93	89	87	87	86	86	86	93	91	91
p- Acc	0	93	91	94	86	84	84	82	82	82	91	89	91
p- P_{μ}	0	0	93	91	86	84	84	82	82	82	89	86	87
p- R_{μ}	0	0	0	93	86	84	84	82	82	82	91	89	91
S- $F1_M$	5.5	3.6	0	3.6	96	89	93	91	93	93	89	87	89
T- $F1_M$	0	0	0	0	1.8	91	89	86	87	87	87	86	87
T'- $F1_M$	3.6	1.8	0	1.8	0	0	93	89	91	91	87	86	87
S- P_M	7.3	5.5	1.8	5.5	0	0	0	94	93	93	84	84	84
T- P_M	11	9.1	5.5	9.1	1.8	0	0	0	98	98	86	84	86
T'- P_M	11	9.1	5.5	9.1	1.8	0	0	0	0	98	86	84	86
S- R_M	0	0	0	0	3.6	0	1.8	5.5	7.3	7.3	96	89	91
T- R_M	0	0	0	0	3.6	0	1.8	3.6	5.5	5.5	0	91	91
T'- R_M	0	0	0	0	3.6	0	0	0	0	0	0	0	96
s- R_{aw}	p- Acc	p- P_{μ}	p- R_{μ}	S- $F1_M$	T- $F1_M$	T'- $F1_M$	S- P_M	T- P_M	T'- P_M	S- R_M	T- R_M	T'- R_M	

(a) City-level

s- R_{aw}	98	96	94	96	66	56	36	66	56	58	64	58	31
p- Acc	0	96	96	98	66	56	36	66	56	58	64	58	31
p- P_{μ}	0	0	94	96	66	56	36	66	56	58	64	58	31
p- R_{μ}	0	0	0	96	66	56	36	66	56	58	64	58	31
S- $F1_M$	1.8	1.8	0	1.8	67	62	42	67	58	60	67	58	31
T- $F1_M$	1.8	1.8	0	1.8	0	60	42	58	58	58	60	58	31
T'- $F1_M$	1.8	1.8	0	1.8	0	0	66	38	38	38	40	38	31
S- P_M	7.3	9.1	7.3	9.1	0	0	0	78	58	62	62	58	33
T- P_M	1.8	1.8	0	1.8	0	0	0	0	64	58	60	58	31
T'- P_M	1.8	1.8	0	1.8	0	0	0	0	0	64	58	58	31
S- R_M	1.8	1.8	0	1.8	0	0	0	0	1.8	0	67	58	31
T- R_M	1.8	1.8	0	1.8	0	0	0	0	0	0	0	60	33
T'- R_M	1.8	1.8	0	1.8	0	0	0	0	0	0	0	0	64
s- R_{aw}	p- Acc	p- P_{μ}	p- R_{μ}	S- $F1_M$	T- $F1_M$	T'- $F1_M$	S- P_M	T- P_M	T'- P_M	S- R_M	T- R_M	T'- R_M	

(b) Country-level

Fig. 5. Significant agreements and disagreements, $p = 0.05$. W-NUT: 11 systems, 55 system pairs. Micro tests are s- R_{aw} , p- Acc , p- P_{μ} , and p- R_{μ} , while the rest represent Macro tests. Significance tests abbreviations stand for: s→sign-test, p→proportions z-test, S→macro sign-test, T→macro t-test, and T'→Wilcoxon test.

SSA involving micro metrics (first four columns), above the diagonal, are observed to be higher than macro metrics. The percentage of SSA involving macro metrics can drop down to 30.9%. The level of disagreements (SSD) are generally low or zero. However, the occurrence rate is sometimes as high as 10.9% (for example, for $T-P_M$ and s-*Raw* at the city level), similarly for tests involving macro metrics. These are cases where experiments would have led to contradictory conclusions about statistically significant differences in system effectiveness, simply based on the metric that was chosen for evaluation. For general evaluation, a macro-micro statistical significance comparison is recommended.

6 DISCUSSION AND LIMITATIONS

Datasets built using Twitter cannot be fully shared and are practically irreproducible because they are subject to decay over time. This challenge will persist, unless Twitter changes their policy and the end-users give their consent to make use of their data. Sharing datasets [Han et al. 2016], therefore, is not feasible. Building centralized frameworks where all researchers submit their systems [Jurgens et al. 2015a], isn't practical as well. Hence, every researcher will likely need to create their own datasets, which will normalize the impact of confounding factors, such as data decay, pre-processing, and ground-truth construction. However, this step requires other researchers to share their systems with the ability to retrain their models.

The global geographic coverage of social media means that datasets are naturally imbalanced in terms of locations, with bias towards big cities. Given that classification is a common approach to predict the location of a Twitter user, it is important to highlight the large number of classes (thousands) involved in the learning process. For general evaluation such as in WNUT shared-task Han et al. [2016] or applications treating urban and rural locations with the same degree of importance, a macro versus micro evaluation should be employed to address the limits of the most common metrics (accuracy and error distance). A majority class baseline is also recommended at the level of coarse geographic granularities, state and country in particular, as it achieved competitive results. Finally, we encourage researchers to report the probability of their predictions/estimations, as opposed to binary classification outputs, to allow for assessing the effectiveness of more evaluation metrics, such as CDF (§ 2.2) and AUC (§ 2.4).

With the large number of explored metrics, Kendall's τ rank correlation test is recommended to quantify the agreement between pairs of metrics. Our results showed that Acc@161, and macro metrics are more consistent and highly correlated across different granularities in comparison to Acc and micro metrics. We demonstrated that error distance metrics (Median, and Mean) and Acc@161 are dependent on the underlying earth representation. While they are highly correlated at the same geographic granularity, they do not convey different information (redundant) and error distance measures are insensitive to evaluation at several geographic granularities. Hence, they should not be used as sole measures for evaluation, which is still the common practice [Bakerman et al. 2018; Ebrahimi et al. 2018; Miura et al. 2017; Rahimi et al. 2018], specially using Acc@161 at fine granularities (city and county). A combination of macro metrics (precision, recall and f1-score) and either micro metrics or accuracy and error metrics are recommended for evaluation.

Statistical significance tests at micro and macro levels were employed to assess the effectiveness of the evaluation metrics at both levels. Using SSA and SSD to summarize the outcome, our results revealed the disparity in agreements and disagreements between tests based on the chosen evaluation metric and geographic granularity. The SSAs between micro and macro tests are higher (better) at the level of city than country. The SSDs are higher (worse) at the level of city than country. To the best of our knowledge, only few recent works applied statistical analysis [Miura et al. 2017], and choosing the right tests can be challenging. Statistical significance testing is essential to draw robust conclusions about the state-of-the-art. In the context of multi-classification and data imbalance, we recommend this list of two-sided

tests: i. Micro sign test (s) and proportions z-test (p) for micro evaluation using raw predictions, accuracy, precision and recall. ii. Macro sign test (S), macro t-test (T), and Wilcoxon test for macro evaluation using precision, recall and f1-score.

The choice of evaluation metrics should be justified by the needs of the applications and the underlying earth representation. A standardized evaluation process, which unified the output format, allowed the comparison of systems with different earth representations. We demonstrated that different systems were found to be best for different underlying representations using an evaluation process including eight measures. Unlike previous research [Jurgens et al. 2015b], evaluation after resolving the location of the unified output using a single reverse-geocoding API allowed evaluation over four geographic granularities and ensured a fair comparison using the same set of locations and avoided the mismatch of predictions based on different representations although they refer to the same location. We demonstrated how competitive geolocation models—previously proclaimed to be inferior—could compete with state-of-the-art models in terms of accuracy.

A major limitation to this work is not extending our evaluation process to network-based approaches, and more importantly recent hybrid methods that rely on deep learning. User coverage is an essential network specific metric to evaluate the percentage of test users with a predicted location [Jurgens et al. 2015b]. If a user does not have social ties, a network-based geolocation model will not be able to predict a location. While hybrid approaches consider network information for training, they evaluated their performance against text-based approaches using error distance measures for two reasons. First, they rely on datasets constructed by text-based research. Second, they always predict a location for a user; rely on text as a fallback if a user is disconnected. The challenge here is to address the user coverage aspect when evaluating text-based against network-based approaches. In this case, recall could be a potential metric.

7 GEOLOCEVAL

Geocoding is the process of linking a document (e.g. Wikipedia article, web page, social media entity, etc.) to a location on earth. Geocoding serves a wide range of applications. With ever increasing quantities of social media content, many applications exploit such data. Examples include: dialectology (the study of geographic lexical variation of a language); regional sentiment analysis; monitoring public health; managing natural crises; and the search for eyewitnesses by journalists. Document geocoding has been an active research area over the last decade, resulting in hundreds of publications, geocoding systems and datasets [Melo and Martins 2017; Mourad et al. 2018; Zheng et al. 2018]. Comparison of such systems share the same challenges of Twitter user geolocation. We, therefore, share our evaluation framework with the research community, hoping researchers will employ in their future geocoding research.

GeoLocEval is an open source python package ⁸ to evaluate the performance of a given set of geocoding systems. The input is a list of JSON files, one for each system to be compared. Each file contains geolocations expressed in the most generic format: GPS coordinates, as in Listing 1. This format is compatible with the Twitter geolocation prediction shared task at the level of tweets and users [Han et al. 2016], known as WNUT.

```
{ "doc_id": {"lon": "x", "lat": "y"}, }
```

Listing 1. JSON Input Format

The GPS coordinates are expanded using a single geocoding API. Results are exported to a JSON file, as in Listing 2.

⁸<https://bitbucket.org/amourad/geoloceval.git>

```

{
  "483049821":
  {
    "geocoding_system_1":
    {
      "doc_id": "483049821",
      "lon": -74.0344411626724,
      "lat": 40.74801738664574,
      "country": "United States",
      "county": "Hudson County",
      "state": "New Jersey",
      "city": "Hoboken",
      "error_dist": 15137.622354338771
    },
  }
}

```

Listing 2. JSON Output Format

7.1 Geocoding APIs

GeoLocEval supports two of the most common geocoding APIs used in previous research:

- Nominatim: a free OpenStreetMap based geocoder.
- GoogleV3: is a commercial API with a higher number of requests per day compared to Nominatim.

Each API supports four administrative levels, namely city, county, state and country. GeoLocEval caches all the resolved GPS coordinates to reduce the number of requests.

7.2 Evaluation Process

We follow the process presented in § 3, by first comparing systems under different evaluations and geographic granularity, next examining rank correlations of systems, and finally studying significant differences. All the generated results are exported to a text file.

8 CONCLUSION AND FUTURE WORK

The work in this paper examined the effectiveness of metrics employed in the evaluation of Twitter user geolocation from three key aspects: standardized evaluation process, compensating bias due to population imbalance through micro vs macro averaging, and comprehensive statistical analysis. We proposed a practical guide to follow for an effective evaluation of each aspect based on thorough experiments and analysis encompassing fifteen geolocation models and two baselines in a controlled environment.

A recommended practical guide for any new research on Twitter user geolocation includes: i) creating its own dataset, ii) sharing its geolocation model with the ability to be retrained by the research community, iii) using a unified output format (GPS coordinates), iv) using a single reverse-geocoding API for discrete evaluation of all the geolocation models considered, v) employing a combined set of evaluation metrics at the micro and macro levels, vi) quantifying the agreement between the evaluation metrics through rank correlation and vii) verifying the conclusions by conducting the recommended statistical significance tests.

This work was initially motivated by [Gao and Sebastiani \[2015\]](#) who changed the perspective of evaluating sentiment analysis after many years of research. They argued that any study dealing with sentiment analysis is usually interested

in the sentiment at the aggregate level of classes, not at the individual level. Quantification-specific evaluation metrics therefore should be used instead of classification metrics, based on the goal of the applications. Since Twitter user geolocation applications do not have a unified goal as sentiment analysis, we focused on experimental evaluation using a wide range of metrics as a vital step that leads to application-specific evaluation. For future work, we would like to investigate the evaluation of geolocation models analytically. Instead of anticipating the needs of the applications, we are interested in collaboration with domain experts, such as journalists, or humanitarians to develop the needs and evaluation metrics in the context of a specific task.

Evaluation of geolocation models on datasets with different characteristics or domains to ensure their consistent performance is a common practice. Rahimi et al. [2018] evaluated their models on three Twitter datasets with different geographic coverage and size. Mourad et al. [2017] evaluated their model on Twitter datasets for thirteen different languages. Wing and Baldrige [2014] evaluated their models on six datasets from different domains, namely Twitter, Wikipedia and Flickr. In this paper, we measured the statistical significance of the differences between geolocation models evaluated on the same dataset. For future work, we would like to extend our geolocation evaluation guide to include the replicability analysis for statistical significance analysis over multiple datasets [Dror et al. 2017].

REFERENCES

- Amr Ahmed, Liangjie Hong, and Alexander J Smola. 2013. Hierarchical geographical modeling of user locations from social media posts. In *Proceedings of the 22nd international conference on World Wide Web*. 25–36.
- Lars Backstrom, Eric Sun, and Cameron Marlow. 2010. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World Wide Web*. 61–70.
- Jordan Bakerman, Karl Pazdernik, Alyson Wilson, Geoffrey Fairchild, and Rian Bahran. 2018. Twitter geolocation: A hybrid approach. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 12 (2018), 34.
- David A Broniatowski, Michael J Paul, and Mark Dredze. 2013. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. *PLoS one* 8 (2013), e83672.
- 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*. 759–768.
- Nicholas Diakopoulos, Munmun De Choudhury, and Mor Naaman. 2012. Finding and assessing social media information sources in the context of journalism. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 2451–2460.
- Tien Huu Do, Duc Minh Nguyen, Evaggelia Tsiligianni, Bruno Cornelis, and Nikos Deligiannis. 2017. Multiview Deep Learning for Predicting Twitter Users’ Location. *arXiv preprint arXiv:1712.08091* (2017).
- Mark Dredze, Michael J Paul, Shane Bergsma, and Hieu Tran. 2013. Carmen: A twitter geolocation system with applications to public health. In *Proceedings of the AAAI workshop on expanding the boundaries of Health Informatics using AI*, Vol. 23. 20–24.
- Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. 2017. Replicability Analysis for Natural Language Processing: Testing Significance with Multiple Datasets. *Transactions of the Association of Computational Linguistics* 5 (2017), 471–486.
- Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker’s guide to testing statistical significance in natural language processing. In *Proceedings of the 56th annual meeting of the Association for Computational Linguistics*. 1383–1392.
- Mohammad Ebrahimi, Elaheh ShafieiBavani, Raymond Wong, and Fang Chen. 2018. Twitter user geolocation by filtering of highly mentioned users. *Journal of the Association for Information Science and Technology* 69 (2018), 879–889.
- Jacob Eisenstein, Brendan O’Connor, Noah A Smith, and Eric P Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. 1277–1287.
- Wei Gao and Fabrizio Sebastiani. 2015. Tweet sentiment: From classification to quantification. In *Proceedings of the 2015 IEEE/ACM international conference on Advances in Social Networks Analysis and Mining 2015*. 97–104.
- Bo Han, Paul Cook, and Timothy Baldwin. 2014. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research* 49 (2014), 451–500.
- Bo Han, Afshin Rahimi, Leon Derczynski, and Timothy Baldwin. 2016. Twitter geolocation prediction shared task of the 2016 workshop on noisy user-generated text. In *Proceedings of the 2nd Workshop on Noisy User-generated Text*. 213–217.
- 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.
- Brent J Hecht and Monica Stephens. 2014. A Tale of Cities: Urban Biases in Volunteered Geographic Information.. In *Proceedings of the eighth international AAAI Conference on Web and Social Media*. 197–205.

- Gaya Jayasinghe, Brian Jin, James Mchugh, Bella Robinson, and Stephen Wan. 2016. CSIRO Data61 at the WNUT geo shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text*. 218–226.
- Isaac Johnson, Connor McMahon, Johannes Schöning, and Brent Hecht. 2017. The Effect of Population and Structural Biases on Social Media-based Algorithms: A Case Study in Geolocation Inference Across the Urban-Rural Spectrum. In *Proceedings of the 2017 CHI conference on Human Factors in Computing Systems*. 1167–1178.
- David Jurgens. 2013. That’s What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships.. In *Proceedings of the seventh international AAAI Conference on Web and Social Media*. 273–282.
- David Jurgens, Tyler Finethy, Caitrin Armstrong, and Derek Ruths. 2015a. Everyone’s invited: A new paradigm for evaluation on non-transferable datasets. In *Proceedings of the ninth international AAAI Conference on Web and Social Media*. 8–17.
- David Jurgens, Tyler Finethy, James McCorriston, Yi Tian Xu, and Derek Ruths. 2015b. Geolocation prediction in Twitter using social networks: a critical analysis and review of current practice. In *Proceedings of the ninth international AAAI Conference on Web and Social Media*. 188–197.
- Yury Kryvasheyev, Hao-hui Chen, Esteban Moro, Pascal Van Hentenryck, and Manuel Cebrian. 2015. Performance of social network sensors during Hurricane Sandy. *PLoS one* 10 (2015), e0117288.
- Xiaomo Liu, Quanzhi Li, Armineh Nourbakhsh, Rui Fang, Merine Thomas, Kajs Anderson, Russ Kociuba, Mark Vedder, Steven Pomerville, Ramdev Wudali, et al. 2016. Reuters Tracer: A Large Scale System of Detecting & Verifying Real-Time News Events from Twitter. In *Proceedings of the 25th ACM international Conference on Information and knowledge Management*. 207–216.
- Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL-2012 System Demonstrations*. 25–30.
- Fernando Melo and Bruno Martins. 2017. Automated geocoding of textual documents: A survey of current approaches. *Transactions in GIS* 21 (2017), 3–38.
- Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. 2011. Understanding the demographics of twitter users. In *Proceedings of the fifth international AAAI Conference on Weblogs and Social Media*. 554–557.
- Yasuhide Miura, Motoki Taniguchi, Tomoki Taniguchi, and Tomoko Ohkuma. 2017. Unifying text, metadata, and user network representations with a neural network for geolocation prediction. In *Proceedings of the 55th annual meeting of the Association for Computational Linguistics*. 1260–1272.
- Alistair Moffat, Falk Scholer, and Paul Thomas. 2012. Models and metrics: IR evaluation as a user process. In *Proceedings of the seventeenth Australasian Document Computing Symposium*. 47–54.
- Ahmed Mourad, Falk Scholer, and Mark Sanderson. 2017. Language influences on tweeter geolocation. In *Proceedings of the European Conference on Information Retrieval*. 331–342.
- Ahmed Mourad, Falk Scholer, Mark Sanderson, and Walid Magdy. 2018. How Well Did You Locate Me? Effective Evaluation of Twitter User Geolocation. In *Proceedings of the 2018 IEEE/ACM international conference on Advances in Social Networks Analysis and Mining*. 437–440.
- Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Confounds and consequences in geotagged Twitter data. *arXiv preprint arXiv:1506.02275* (2015).
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
- Adam Poulston, Mark Stevenson, and Kalina Bontcheva. 2017. Hyperlocal Home Location Identification of Twitter Profiles. In *Proceedings of the 28th ACM conference on Hypertext and Social Media*. 45–54.
- Reid Priedhorsky, Aron Culotta, and Sara Y Del Valle. 2014. Inferring the origin locations of tweets with quantitative confidence. In *Proceedings of the 17th ACM conference on Computer Supported Cooperative Work & Social Computing*. 1523–1536.
- Afshin Rahimi, Timothy Baldwin, and Trevor Cohn. 2017. Continuous Representation of Location for Geolocation and Lexical Dialectology using Mixture Density Networks. In *Proceedings of the 2017 conference on Empirical Methods in Natural Language Processing*. 167–176.
- Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2016. pigeo: A python geotagging tool. In *Proceedings of the ACL-2016 System Demonstrations*. 127–132.
- Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2018. Semi-supervised User Geolocation via Graph Convolutional Networks. In *Proceedings of the 56th annual meeting of the Association for Computational Linguistics*. 2009–2019.
- Erica Rodrigues, Renato Assunção, Gisele L Pappa, Diogo Renno, and Wagner Meira Jr. 2016. Exploring multiple evidence to infer users’ location in Twitter. *Neurocomputing* 171 (2016), 30–38.
- Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldrige. 2012. 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*. 1500–1510.
- Raz Schwartz, Mor Naaman, and Rannie Teodoro. 2015. Editorial Algorithms: Using Social Media to Discover and Report Local News.. In *Proceedings of the ninth international AAAI Conference on Web and Social Media*. 407–415.
- Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *Comput. Surveys* 34 (2002), 1–47.
- Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management* 45 (2009), 427–437.
- Kate Starbird, Grace Muzny, and Leysia Palen. 2012. Learning from the crowd: collaborative filtering techniques for identifying on-the-ground Twitterers during mass disruptions. In *Proceedings of 9th international conference on Information Systems for Crisis Response and Management*. 1–10.

- Benjamin Wing and Jason Baldrige. 2014. Hierarchical discriminative classification for text-based geolocation. In *Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing*. 336–348.
- Benjamin P Wing and Jason Baldrige. 2011. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies*. 955–964.
- Yiming Yang. 1999. An evaluation of statistical approaches to text categorization. *Information retrieval* 1 (1999), 69–90.
- Yiming Yang and Xin Liu. 1999. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and Development in Information Retrieval*. 42–49.
- Xin Zheng, Jialong Han, and Aixin Sun. 2018. A survey of location prediction on Twitter. *IEEE Transactions on Knowledge and Data Engineering* (2018), 1652–1671.
- Arkaitz Zubiaga, Heng Ji, and Kevin Knight. 2013. Curating and contextualizing twitter stories to assist with social newsgathering. In *Proceedings of the 2013 international conference on Intelligent User Interfaces*. 213–224.