Influence of Geographic Biases on Geolocation Prediction in Twitter

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Ahmed Mourad

B.Sc. Computer Engineering, Cairo University

School of Science

College of Science, Engineering, and Health

RMIT University

October 31, 2019

Supervisors

Assoc. Prof. Falk Scholer

Prof. Mark Sanderson

Dr. Walid Magdy

Declaration

I certify that except where due acknowledgement has been made, the work is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program; any editorial work, paid or unpaid, carried out by a third party is acknowledged; and, ethics procedures and guidelines have been followed.

Ahmed Mourad

October 31, 2019

Acknowledgement

I would like to thank my supervisors for all the support. Special thanks to my family and friends for helping me go through this overwhelming experience till the end.

Contents

List of Figures

List of Tables

Abstract

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. While users can record their location on their profiles, more than 34% record fake or sarcastic locations. Twitter allows users to GPS locate their content, however, less than 1% of tweets are geotagged. Therefore, inferring user location has been an important field of investigation since 2010. This thesis investigates two of the most important factors which can affect the quality of inferring user location: (i) the influence of tweet-language; and (ii) the effectiveness of the evaluation process.

Previous research observed that Twitter users writing in some languages appeared to be easier to locate than those writing in others. They speculated that the geographic coverage of a language (*language bias*) — represented by the number of locations where the tweets of a specific language come from — played an important role in determining location accuracy. So important was this role that accuracy might be largely predictable by considering language alone. In this thesis, I investigate the influence of language bias on the accuracy of geolocating Twitter users. The analysis, using a large corpus of tweets written in thirteen languages and a re-implemented state-of-the-art geolocation model back at the time, provides a new understanding of the reasons behind reported performance disparities between languages. The results show that data imbalance in the distribution of Twitter users over locations (*population bias*) has a greater impact on accuracy than language bias. A comparison between *micro* and *macro* averaging demonstrates that existing evaluation approaches are less appropriate than previously thought. The results suggest both averaging approaches should be used to effectively evaluate geolocation.

Many approaches have been proposed for automatically geolocating users; 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 thesis, I provide a standardized evaluation framework for geolocation systems. The framework is employed to analyze fifteen Twitter user geolocation models and two baselines in a controlled experimental setting. The models are composed of the re-implemented model and a variation of it, two locally retrained open source models and the results of eleven models submitted to a shared task. Models are evaluated using ten metrics — out of fourteen employed in previous research — over four geographic granularities. Rank correlations and thorough statistical analysis are used to assess the effectiveness of these metrics. The 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. For general evaluations, a range of performance metrics should be reported, to ensure that a complete picture of system effectiveness is conveyed. Although a lot of complex geolocation algorithms have been applied in recent years, a majority class baseline is still competitive at coarse geographic granularity. A suite of statistical analysis tests is proposed, based on the employed metric, to ensure that the results are not coincidental.

Publications

Parts of this thesis are based on the following publications. Chapter 3 is largely based on [Mourad et al. 2017], and Chapter 5 is largely based on [Mourad et al. 2018; 2019]. A public evaluation tool is also presented in [Mourad et al. 2019].

(1) A. Mourad, F. Scholer, and M. Sanderson. Language influences on tweeter geolocation. In *Proceedings of the European Conference on Information Retrieval*, pages 331–342, 2017

(2) A. Mourad, F. Scholer, M. Sanderson, and W. Magdy. 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*, pages 437–440, 2018

(3) A. Mourad, F. Scholer, W. Magdy, and M. Sanderson. A practical guide for the effective evaluation of Twitter user geolocation. *Transactions of the Association for Computational Linguistics*, 0:1–23, 2019

CHAPTER 1

Introduction

Social media has become an integral part of our daily lives, providing a continuously growing stream of data. Allowing users to geolocate their content (e.g. by enabling GPS on their smart phones) adds a geographical context that serves a wide range of applications. Researchers analyze this data to either serve existing community needs, *applied research*, or learn and define new models that might fulfil future needs, *basic research*. In between, basic research is initially conducted to promote the task before being revisited in the context of applications. This thesis bridges the gap between theory and practice in the context of Twitter geolocation — the process of linking a tweet or a user to a real-world location. In particular, we focus on the influence of the geographic coverage of a language (*language bias*) — represented by the number of locations where the tweets of a language come from — on the accuracy of geolocation, and the effectiveness of evaluation based on the generic needs of applications. Findings of the latter can be extended to the broader domain of document geolocation — the process of linking a document to a real location on earth — including other social networks such as Wikipedia, Facebook, Flickr, etc.

1.1 Motivation

Geo-spatial analysis of social media, specifically Twitter, drew the attention of the research community over the last decade, making use of the explosion in the user base of this micro-blogging site. Researchers started to exploit this large stream of data to serve a wide range

of community and business needs. Twitter geolocation is a service needed for many social media-based applications, such as finding an eyewitness to an event, managing natural crises, personalizing regional advertisements, and studying the evolution of languages. While users can record their location on their Twitter profile, Hecht et al. [2011] reported that $> 34\%$ record fake or sarcastic locations. Twitter also allows users to GPS locate their content, however, Han et al. [2014] reported that $< 1\%$ of tweets are geotagged. Twitter geolocation based on features derived from tweet and profile content is therefore a field of investigation, which is the main topic of this thesis.

Knowing the location of Twitter entities (user or tweet) can help to manage natural crises like earthquakes by detecting them promptly, estimate their trajectories and warning users before being hit. It moreover helps to assess the damage caused by natural disasters like hurricanes and floods, to prioritize the needed aid, and to guide humanitarian assistance on the ground. In the public health domain, it helps to track epidemics like influenza and Ebola, and monitor population health, to an accurate level on a par with reports generated by governmental centers. Social media helps journalists in gathering news, especially breaking news like the Boston marathon bombing and Brussels terrorist attack. Knowing the location of Twitter users is vital to the process of validating this news and detecting eyewitnesses.

Dialectology is another application which utilizes geotagged tweets, language bias and the geographic lexical variance of language in social media to study the evolution of languages over time and geography. Leveraging the location of tweets, linguists are able to study the overlap of languages in multilingual regions (e.g. French versus English in Canada), different dialects of the same language (e.g. Egyptian versus Levantine Arabic), or even regional dialects within the same country (e.g. Central and Southern dialects in Spain).

Language bias in social media can also contribute to enhancing the accuracy of Twitter geolocation. The geographic lexical variance of languages (dialects) and topics of interest (sports or music) help to identify the location of Twitter entities. While *language* bias is demonstrated in previous research to be beneficial, Twitter geolocation is also influenced by *population* bias. Twitter datasets are skewed towards populous (ur-

ban) regions at the cost of rural areas. Population bias, therefore, decreases the value of language bias to Twitter geolocation. If rural and urban regions share the same dialect and topic interests, a Twitter geolocation system will be biased towards the urban region. The degree of bias differs from one language to another. For instance, English is the most spoken language in the world with several dialects (e.g. American, British, Malaysian, etc.) and tens of big cities (e.g. Los Angeles, Boston, Chicago, London, Kuala Lumpur), while Turkish is relatively limited to dialects and regions within Turkey.

Previous research focused on the development of geolocation models which employ state of the art machine learning algorithms, and harness all the possible sources of information, such as tweets, user profiles and social networks. On the other hand, language and population biases, and evaluation drew much less attention. English is the most prominent language in existing research, with only few works examining the influence of language bias on geolocation accuracy. A limited set of evaluation measures was considered, which in return does not provide a sufficient understanding of the differences between geolocation models when taking into account different types of biases. Moreover, none of the previous evaluation measures were driven by the specific needs of an application. Therefore, the way these geolocation models will perform in real world scenarios is unpredictable, and not guaranteed to fulfil these needs. For a decision maker, it will be hard to decide which model is better for which application.

1.2 Aim and Scope

Twitter geolocation is an active field of research. However, evaluation has been ineffective in the context of its applications [Diakopoulos et al. 2012, Starbird et al. 2012, Dredze et al. 2013, Schwartz et al. 2015, Kryvasheyeu et al. 2015, Liu et al. 2016]. Motivated by this observation, this thesis examines the evaluation of Twitter geolocation from two aspects, namely language influence and evaluation effectiveness. Among various Twitter geolocation tasks, one task — Twitter user geolocation — is considered, to bridge the gap between the anticipated performance of geolocation models in theory and practice.

Language Influence. Language is a valuable source of location information that

helps to enhance the accuracy of Twitter user geolocation, known as the text-based approach. Twitter users writing in some languages appeared to be easier to locate than those writing in others [Han et al. 2014, Priedhorsky et al. 2014]. Existing research speculated that the geographic coverage of a language (*language bias*) or distribution of users (*population bias*) played an important role in determining geolocation accuracy [Han et al. 2014, Johnson et al. 2017]. So important was this role that accuracy might be largely predictable by considering language alone. The aim of this research is to quantify the influence of *language bias* on Twitter user geolocation, in comparison to *population bias* and *dataset size*, among other features. To achieve this, we limit our focus to text-based geolocation approaches.

Urban and Rural Evaluation. Each geolocation application has different needs, which might require evaluation from several perspectives. Current evaluation practices focus on a few measures introduced in early literature. These measures were shown, in our Language Influence experiment, to be skewed towards densely populated (urban) locations, 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 [Starbird et al. 2012, Schwartz et al. 2015, Liu et al. 2016], monitoring natural disasters in rural areas [Kryvasheyeu et al. 2015], or tracking epidemics in rural cities [Dredze et al. 2013]. This research anticipates the importance of urban and rural evaluation quantitatively, using an external resource as an indicator of journalistic information needs.

Evaluation Effectiveness. 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. This research aims to develop a standard geolocation evaluation framework that helps decision makers with choosing a geolocation model suitable to their needs. This framework takes into account population bias in Twitter datasets, evaluation

at several geographic granularities, and other features that might influence the fairness of evaluation.

Geolocation Evaluation Framework. Twitter user geolocation is a special task of the broader domain of document geolocation. While our evaluation framework (GeoLocEval) is tested on a single task, its utility extends to other tasks given that they share the same challenges. GeoLocEval provides a fair benchmark process for the advancement of research on document geolocation including tweets, Wikipedia articles, and Facebook entities, among other social media networks.

1.3 Contributions

In this work, we measure the influence of language and population biases on the evaluation of Twitter user geolocation, and how it impacts the effectiveness of the evaluation process for the generic domain of document geolocation in social media.

For language influence, we conduct an evaluation of the features that impact the accuracy of a Twitter user geolocation technique. Our contributions are:

- We conduct a preliminary study on one language other than English to evaluate the impact of the scale of a comparable dataset on the accuracy of geolocation. Based on our observations, we propose an alternative set of measures that addresses the limitations of current evaluation approaches.

- We analyze a multi-lingual corpus of tweets written in thirteen languages to understand ten features (including language and population biases) that cause performance disparities between per-language geolocation models, and quantify their influence. Based on the analysis, we propose an alternative perspective and framework for the evaluation of geolocation that we conjecture is more closely aligned with the range of real-world problems for which geolocation is of interest.

To validate the utility of our evaluation framework to real-world applications, we employ a quantitative approach to anticipate the information needs of an application. Our contribution are:

- We use a global database of that contains over 6 million events to anticipate the information needs of a news-media application.
- We anticipate the importance of urban and rural evaluation through a statistically principled process that considers the distribution of these events over cities.
- We reproduce an existing research on web page language identification based on URLs only, which allows conducting the analysis in the previous step on events per language.

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 accuracy of fifteen geolocation models and two baselines, using ten different evaluation measures over four geographic granularities, our study proposes a process for an effective 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 accuracy 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, typical evaluation metrics are found to be 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.

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.

1.4 Thesis Outline

The rest of the thesis is structured as follows:

Chapter 2 reviews previous work related to Twitter geolocation. It presents downstream geolocation applications while highlighting their different information needs. Sources of location information fall into explicit (e.g. location field in user profile) and implicit (e.g. user's network of friends) indicators. Twitter entities whose location can be inferred are tweet, user and named entities mentioned in the tweet. An orthogonal factor to all geolocation tasks is earth representation which can be an administrative region (e.g. city), GPS coordinates, a geographic grid, or a list of candidate locations (e.g. landmarks). We survey the different approaches for Twitter geolocation which broadly fall into text-based, network-based and hybrid models. A detailed description of the text-based systems employed through out the thesis is provided. Then, we discuss geographic biases that exist in Twitter datasets as identified by existing research. Population and language biases are the two types that influence the accuracy of geolocation. Lastly, we review the evaluation of Twitter geolocation, covering the definitions of metrics employed in previous research and the research gaps that we address in this thesis.

In **Chapter 3**, we examine the influence of language on the accuracy of Twitter user geolocation using state-of-the-art geolocation model which utilizes language only. First, we conduct a pilot study on a large Arabic dataset at a scale comparable to an exiting benchmark English dataset. Based on the analysis, we propose alternative measures to reveal the influence of language bias in comparison to population bias. In particular, we employ different averaging techniques to evaluation measures at fine and coarse ge-

ographic granularities. Then, we extend the analysis to two multi-lingual collections of thirteen languages. Statistical methods are employed to measure the correlation between ten features which might influence the accuracy of geolocation and evaluation measures. Features are related to the scale of a monolingual dataset (e.g. number of users, tweets, and cities), the lexical nature of a language (the number of location indicative words used to train a geolocation models), and error distance measures (e.g. the average distance between neighbouring cities). Finally, we test the consistency of our hypothesis over the two data collections.

Chapter 4 reviews the proposed measures in the previous chapter, averaging techniques in particular, in the context of application needs. We first introduce the notion of urban and rural evaluation as defined in existing research. Then we follow a quantitative approach to anticipate the necessity of rural and urban evaluation in the context of journalism using a global database of events. In particular, we measure the similarity of population bias between social media and events data, and apply power-law analysis for a better understanding of the influence of rural regions. Lastly, we examine the consistency of the analysis over seven languages in each data set.

Chapter 5 focuses on the effectiveness of evaluation using fifteen geolocation models and two baselines in a controlled experimental setting, in comparison to a single geolocation model in Chapter 3. In particular, we employ a statistically principled process for evaluation given the classification nature of the task, and highlight the importance of researcher-owned datasets, unified output format and reverse-geocoding of locations to the fairness of evaluation.

Chapter 6 describes the common challenges between Twitter and document geolocation and the utility of our proposed evaluation framework. Then we describe in details the different components of this framework.

Chapter 7 summarizes the findings of this thesis, and discusses some ideas to steer geolocation research towards fulfilling the needs of applications.

CHAPTER 2

Literature Review

In this chapter, we review the previous literature related to Twitter geolocation. Downstream geolocation applications, sources of location information, the granularity of geolocating a Twitter entity and different geographic levels which might be required by applications are presented in Section 2.1. Different approaches for Twitter geolocation, especially geolocation systems employed throughout the thesis, are discussed in Section 2.2. Section 2.3.1 introduces geographic biases in Twitter based on two factors, namely population (in relation to census data), and language. Lastly, the evaluation of Twitter geolocation is discussed in Section 2.4, to cover the formal definitions of each evaluation metric and the limitations or challenges which we address in this thesis.

2.1 Twitter Geolocation

Social networks, such as Twitter, Facebook, and Flickr, have become an integral part of our daily lives and a platform to share text, photos and videos. Billions of active users on these social networks generate a vast amount of data. Allowing users to geolocate their content (e.g. by enabling GPS on their smart phones) adds a geographical context that serves a wide range of applications and location-based services. Document geolocation — the process of linking a document to a real-world location — has, therefore, become an active field of research. In this thesis, we focus on geolocation in Twitter, as the most popular source of research datasets, to approach applications such as managing natural

crisis, tracking epidemics, finding eyewitnesses, and studying the evolution of languages over time and geography among other applications.

2.1.1 Applications

Managing natural disasters, also known as emergency management, is one of the earliest domains to make use of the surge of social media. Sakaki et al. [2010] utilised geotagged Twitter posts to detect earthquake occurrence promptly in Japan. By estimating the center of the earthquake and its trajectory, their real-time system would send a warning to other regions before being hit by it. Starbird et al. [2012] studied the social life of information on Twitter during the flooding of the Red River in the US and Canada. Users profiles and geotagged tweets allowed them to prioritize the information generated by users influenced by the event and track the travel of information to further users. Similarly, Kryvasheyeu et al. [2015] investigated the flow of information before, during and after Hurricane Sandy while giving more weight to the information generated by users geolocated within the affected area. Realizing the power of social media, a program in Jakarta supported by the government[1] utilises Twitter to assess the damage of floods and guide humanitarians on the ground. In order to maximise the benefit out of this information, they would send specific instructions to the citizens on how to report on flood levels and activate geo-tagging. Twitter geolocation is not only beneficial to large-scale incidents, but also to every day small-scale incidents like car crashes or fires in the neighbourhood [Schulz and Ristoski 2013, Schulz et al. 2015; 2016].

In the public health domain, Burton et al. [2012] assessed the reliability of location information extracted from Twitter to surveil diseases. The value and accuracy of this information is important to study and monitor population health. Dredze et al. [2013] developed a Twitter geolocation system, with application to public health, that infers the location of a Twitter user (Carmen). They later deployed their system to track influenza during 2012–2013 [Broniatowski et al. 2013]. Carmen's influenza prevalence estimates were strongly correlated with surveillance data provided by two governmental centers in the United States.

[1] https://bit.ly/2PZG3sF

In the context of journalism, eyewitness detection [Diakopoulos et al. 2012, Starbird et al. 2012, Morstatter et al. 2014] and news gathering [Zubiaga et al. 2013, Schifferes et al. 2014, Schwartz et al. 2015, Liu et al. 2016] from Twitter are two common research fields. For instance, the Boston marathon bombing and Brussels attack appeared on social media first and then picked up by news media. Journalists, therefore, rely on social media to complement their stories by information posted through eyewitnesses or find breaking news. However, this comes at the cost of filtering irrelevant information and verifying the vast amount of data on social media. Geolocation of content lies at the heart out of this process. Diakopoulos et al. [2012] developed a system to find and assess the verity of eyewitnesses found through social media. Based on the feedback from journalists who reviewed the system, knowing the location of the eyewitness and their friends is vital for this process. Reuters Tracer is another real-time news gathering system, which detects and verifies events at a global scale [Liu et al. 2016].

Dialectology in social media, the relation between geographical location and language, has been an active field of research and one of the most important applications as well. Eisenstein et al. [2010] investigated the geographic linguistic variation of English in the United States. They analyzed the topics of interest, such as sports or music, in the geotagged tweets of each state. They demonstrated that geographic regions have different topic interests and users tend to use regionally affiliated terms. Mocanu et al. [2013] explored the mapping of world languages through geotagged Twitter data. They surveyed worldwide linguistic indicators, such as the linguistic homogeneity of different countries, the touristic seasonal patterns within countries and the geographical distribution of different languages in multilingual regions. Later, Gonçalves and Sánchez [2014] focused on Spanish to characterize its varieties on a global scale. Using a large dataset of 50 millions geotagged tweets, they found that Spanish language can be split into two super dialects, and one of them can be further split into five local dialects. Pavalanathan and Eisenstein [2015] compared the data acquisition techniques of geotagged Twitter data to quantify the demographic biases they introduce and how these demographic variables interact with geography to affect language use.

2.1.2 Sources of location information

Explicit Indicators

Since 2006, Twitter allows users to explicitly set their profile location, indicating their home location. This location is in the form of free text. Approximately 30% of users set their home location to a place that can be resolved to the country level.[2] Hecht et al. [2011] also reported that more than 34% record fake or sarcastic locations. In November 2009, Twitter introduced the feature of geotagging tweets with the exact coordinates. Twitter applications like Foursquare utilised this feature to allow users to check-in in places that represent a geographic area on the venue, neighborhood, or town scale (e.g restaurant, stadium, museum). As Foursquare became the most popular location-based application with more than seven billion check-ins at 65 million places. As a result, Twitter partnered with Foursquare since March, 2015 to allow users include their location in tweets.[3] However, approximately $1-2\%$ of Tweets are geo-tagged using either method, according to Twitter. Similar percentages are reported by Han et al. [2014]. Research, therefore, studied other alternatives to predict the geolocation of Twitter users or tweets, while utilising explicit indicators to build the ground-truth datasets.

Implicit Indicators

Research on geolocation prediction in Twitter relies on three alternative types of information. A tweet content might contain geographical references (e.g. Melbourne), use words that are location indicative (e.g. mountains), or include some dialectal words which can be linked to geographical regions (e.g. arvo – afternoon in Australia). Timezone of the tweet is a meta-data information which can be used as a geographical reference, among other attributes. Lastly, a valuable source of geographical information is the user's network of friends, who might have set their profile home location.

[2] https://developer.twitter.com/en/docs/tweets/data-dictionary/guides/tweet-timeline
[3] https://fortune.com/2015/03/23/twitter-foursquare-data/

2.1.3 Geolocation Granularity

Twitter geolocation can be inferred at three levels of granularity, namely user, tweet, and mentioned locations. User geolocation is the task of inferring the home location of a user. Tweet geolocation infers the location where the tweet is posted. Recognition of mentioned locations refers to identifying location named entities in the tweet message and linking it to a real-world location.

Each task has its own set of applications, which might overlap in some cases. For instance, user geolocation enables applications like public health monitoring [Dredze et al. 2013, Broniatowski et al. 2013], public opinion polling estimation and dialectology [Eisenstein et al. 2010], which are interested in the aggregate information of a user over an extended period of time and a coarse geographic granularity (e.g. state), where moving around is not a problem. Tweet geolocation allows applications like managing natural disasters, and identifying small-scale local incidents [Schulz and Ristoski 2013, Schulz et al. 2015; 2016], which rely on real-time information at exact locations (e.g. GPS coordinates or street) to send the required aid. Recognition of mentioned locations is beneficial to applications such as venue recommendation, recommending similar restaurants or music concerts. Other applications like eyewitnesses detection [Diakopoulos et al. 2012, Morstatter et al. 2014, Starbird et al. 2012] and news gathering [Liu et al. 2016] require geolocation at the level of both users and tweets. First, user geolocation is required to filter out noise, i.e. users away from the region (e.g. city) of interest. Second, tweet geolocation is involved to decide whether the user is an eyewitness or not based on the geographic proximity to the event (e.g. tweeting while in the same street or neighbourhood).

Figure 2.1 shows the popularity of Twitter geolocation tasks in existing research. A list of 80 geolocation research papers are considered, including work cited in the most recent Twitter geolocation survey paper [Zheng et al. 2018] and a few papers that have been published afterwards. Obviously, user geolocation is the most popular research task compromising almost half of the existing studies. Only few papers considered user and tweet geolocation in the same study.

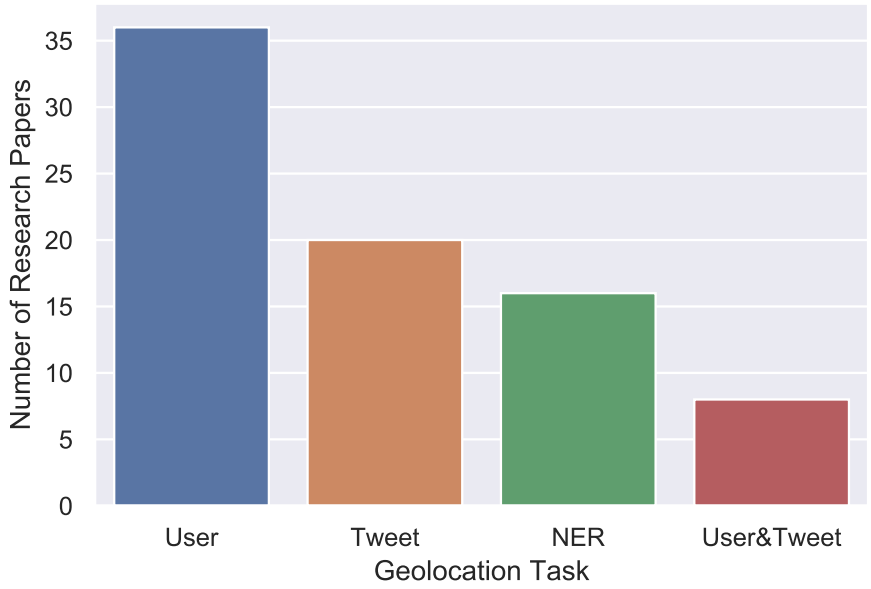


Figure 2.1: Twitter geolocation task popularity in existing research. A total of 80 research papers are considered.

2.1.4 Geographic Granularity

An orthogonal factor common between all tasks is the way a location is represented, based on the requirements of the applications. Generally, there are four levels of geographic granularity:

- Administrative regions, i.e. city, county, state, and country
- Geographical Grids, i.e. the earth is partitioned into cells of equal (uniform) or varying sizes (adaptive)
- GPS Coordinates, i.e. longitude and latitude
- List of candidate locations (venues), i.e a specific business, landmark, or point-of-interest (POI)

The first three granularities are common in the tasks of user and tweet geolocation, while the last is the only representation for mentioned locations prediction.

Geographical granularity is involved into two stages of any geolocation system, namely the input and output of the trained models. A system can combine two different representations. For instance, some studies represent earth as administrative regions [Eisenstein

et al. 2010, Han et al. 2014, Rodrigues et al. 2016] or geographical grids [Roller et al. 2012, Wing and Baldridge 2014, Rahimi et al. 2018] while training their models, and then converts the output to GPS coordinates that represent the center of the predicted region. For the sake of evaluation, a set of compared geolocation systems should have the same output format, while the input format can be different. More details on this is presented in Chapter 5.

The work in Chapter 3 is based on city representation. We employ the framework provided by Han et al. [2014], which has a total of 3,709 cities representing the whole world. It was generated using a seed list of publicly available Geonames[4] dataset which provides a detailed and complex structural city-level metadata. To reduce the dimensionality of cities, they iteratively collapsed cities which are adjacent to one another within the same administrative region. Cities within 50 km of the city with the largest population in a region are collapsed into this city. Finally, all cities with a population of less than 100K are removed. The work in Chapter 5 employs grid representation, beside city representation. Roller et al. [2012], Rahimi et al. [2016] provided geolocation systems based on adaptive grids with varying cell sizes, but uniform number of users within each cell.

2.2 Geolocation Approaches

Previous research inferred the location of Twitter entities 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 the interplay of the language used to write tweets and geography. Location sparsity is, therefore, an inherent challenge. 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.

[4]http://www.geonames.org/

In this section, we focus mainly on reviewing text-based geolocation methods of Twitter users, which have been used in the next two chapters (3 and 5), from a machine learning perspective.

Han et al. [2014] predicted the home city (§2.1.4) of a Twitter user based on his/her tweets, which are concatenated in one document. Due to the noisy and shorthand language used in Twitter, the full text of tweets includes a lot of common words with no geospatial value. To solve this problem, they introduced the idea of finding Location Indicative Words (LIW) via feature selection. Instead of using full text, they constructed a lexicon of words highly associated with cities. Thorough experiments were conducted using several methods of feature selection, including chi^2, Ripley's K statistic, and Information Gain Ratio (IGR), to generate the LIW lexicon. Using a multinomial naïve bayes prediction model, IGR was reported to achieve the best performance. To conduct the experiments in chapter 3, we re-implemented their approach based on IGR, which is defined as follows. Given a set of unique terms $\mathbf{t}$, a set of cities(classes) $\mathbf{c}$, IG for a term $t \in \mathbf{t}$ is calculated as:

$$\begin{aligned} IG(t) = &-\sum_{i=1}^{n} P(c_i) log P(c_i) \\ &+P(t) \sum_{i=1}^{n} P(c_i|t) log P(c_i|t) \\ &+P(t) \sum_{i=1}^{n} P(c_i|\bar{t}) log P(c_i|\bar{t}) \end{aligned} \tag{2.1}$$

where the first operand represents the entropy of classes; cities in this case; which is constant for all terms. Second is the average information gain of each unique term based on that split. Last is the average information gain achieved in absence of each term. However, one of the drawbacks of IG is overfitting attributes that can take on a large number of distinct values. For example, a word that appears rarely in the corpus associated with a random city would mistakenly identify that city. Information Gain Ratio (IGR) reduces this problem by taking the prior probabilities of words presence and absence, aka Intrinsic Value (IV), into account. Both are calculated as:

$$IV(t) = -P(t)logP(t) - P(\bar{t})logP(\bar{t})$$
$$IGR(t) = IG(t)/IV(t) \tag{2.2}$$

Priedhorsky et al. [2014] proposed an approach to estimate the location of a tweet based on the text of the user location string and tweet, among other features. A Gaussian Mixture Model (GMM) is fit for each unique n-gram to model its geographic distribution. The location of a new tweet is inferred by combining the previously trained GMMs for the n-grams it contains, using weights inferred from data. Unlike previous research, location estimates are multi-modal probability distributions, rather than points or regions. Although the considered geolocation granularity is at the level of tweets, their work is considered here because they evaluated the impact of a language on geolocation, more details are discussed in Section 2.3.3.

To conduct the experiments in Chapter 5, few other geolocation models are considered based on the availability of their code-base with the ability to retrain the models on our local datasets ([Roller et al. 2012, Han et al. 2014, Rahimi et al. 2016]), or their results if evaluated on the same dataset ([Chi et al. 2016, Miura et al. 2016, Jayasinghe et al. 2016]).

Roller et al. [2012] predicted the home cell of a Twitter user in an adaptive-grid representation (§2.1.4) based on tweets content. Training documents are constructed by concatenating tweets on two stages, user then cell; each grid cell is represented by a document. Then a unigram probability distribution θ_{c_i} is constructed over the vocabulary of each cell c_i. To infer the location of a user, the similarity of his/her tweets (concatenated into one document) to each grid cell is computed using Kullback-Leibler divergence (Equation 2.3). The location of the most similar grid cell is assigned to the user.

$$KL(\theta_{c_i}||\theta_{c_j}) = \sum_k \theta_{c_i}(k) log \frac{\theta_{c_i}(k)}{\theta_{c_j}(k)} \tag{2.3}$$

Rahimi et al. [2016] built a tool that predicts the location of a Twitter user based on a hybrid geolocation approach which utilizes his/her text and network information.

Similar to Roller et al., the text-based approach predicted the home cell of a user in an adaptive-grid representation based on tweets content. Each user is represented as a vector of term frequency, inverse document frequency, and l_2 normalisation. A stochastic gradient descent classifier is trained using log loss and elastic net regularisation as the cost function. The network-based regression model estimates the home coordinates of a user based on the weighted median coordinates of all his/her connections after several iterations to propagate locations across the constructed graph.

Han et al. [2016] organised a shared task for Twitter geolocation prediction with the Workshop on Noisy User-generated Text (WNUT). The task included geolocation at the level of a tweet and a user. It adopted a multi-class classification problem, to predict the location label (i.e. city) for each tweet/user. Five teams participated with twenty-one submissions. Although we consider the results of all the submissions in Chapter 5, only three of the five teams described their approach, which we review here focusing on geolocation at the level of a user.

Chi et al. [2016] (IBM) created a multinomial naïve bayes classifier trained on location indicative words (similar to Han et al. [2014]) and additional textual features extracted from the tweet, such as city/country mentions, hashtags and user mentions. Miura et al. [2016] (FUJIXEROX) employed a simple neural networks structure with fully connected layers trained on tweets and various user metadata, namely location, description and timezone. Each source of information was represented as an averaged weighted vector. Then a concatenated vector of all sources of information were fed into a stochastic gradient decent classifier using cross entropy loss as the cost function.

Jayasinghe et al. [2016] (CSIRO) adopted an ensemble technique combining four complementary approaches: heuristics based on metadata , label propagation, text classification, and information retrieval. Heuristic methods rank the list of cities based on the prior probability of each of the metadata, namely location mentions, location field, timezone, mobile operating system (iPhone and Android), and location based on URL and IP look up. Label propagation methods predict the location based on the neighbours of a user (network-based). Text classification is a timezone variant of Rahimi et al. [2016], where they trained a classifier per timezone, instead of a single global classifier. Lastly, Infor-

mation retrieval approach predicts locations based on tweets similarity using the Apache Solr search engine.

For more details about other approaches, we refer the reader to the following survey papers. Ajao et al. [2015] conducted a small scale survey of techniques applied to infer the location of Twitter users based on the sources of location information (§2.1.2). Melo and Martins [2017] surveyed text-based methods of user geolocation with focus on the geographic granularity (§2.1.4). Zheng et al. [2018] conducted a more comprehensive survey of previous research on Twitter geolocation considering sources of information, geolocation granularity (§2.1.3) and geographic granularity.

2.3 Biases in Geolocated Data

In this section, we discuss several types of bias in social media datasets. Population bias, in relation to census data, is described in Section 2.3.1. Language bias, based on the linguistic geographic variations, is described in Section 2.3.2. We summarize the related work utilizing language bias to enhance Twitter geolocation in Section 2.3.3.

2.3.1 Population Bias

Social media is known to have a substantial *population bias*, i.e. the distribution of tweets or users based on several demographic characteristics such as gender, race, education, socioeconomic status and geographic regions, varies extensively. This inherent bias in social media datasets raised concerns about its influence on social media-based algorithms.

Previous research used two different methodologies to quantify and understand population bias in social media. Some used a qualitative approach to collect the demographic data of social media users through surveys, while others followed a quantitative approach which analyzes social media data on a large scale. To decide whether a bias exists or not, the collected/analyzed data is usually assessed against census data provided by governments or against the Twitter population.

Pew Research Center (PRC) conducts annual surveys of social media usage in the

United States since 2005. According to their 2018 survey[5], there are substantial differences in Twitter use by: age, which is used by 45% of $18-24$ and 6% of 65+ age groups; income, which is used by 20% of $< \$30k$ and 32% of $\$75k+$ earners; education, which is used by 18% of high school or less and 32% of college or higher; and geography, which is used by 29% of urban and 17% of rural populations.

Mislove et al. [2011] analyzed a Twitter dataset of over 3 million users to examine whether Twitter is a representative sample of the society or not. Utilizing the United States Census data at the county level, they compared the Twitter population to the US population along three axes, namely geography, gender and race/ethnicity. They demonstrated that rural counties in the mid-west and south (e.g. in Chicago and Dallas) are significantly underrepresented in Twitter, while urban counties in the east coast and west coast (e.g in San Francisco and Boston) are oversampled; a significant male bias existed with 71% of the examined users had a male name; and the distribution of race/ethnicity is highly geographically-dependent with Hispanic users undersampled in the south-west, African-Americans undersampled in the south and mid-west, and Caucasians oversampled in many urban counties. While the geographic bias still exists, the gender bias has decreased over time as they expected. According to the PRC survey in 2018, Twitter in the United States is adopted by 23% of males and 24% of females.

Similarly, Longley et al. [2015] investigated biases across age, gender, and ethnicity, focusing on the Greater London area. Using a similar approach to Mislove et al. [2011], they exploited the forename-surname and profile data of 158,375 Twitter users to identify their demographic information. A comparison to the 2011 UK census reveals an over-representation of males and younger adult age groups (15–29). A comparison to the 2007 Electoral Register suggests that all ethnic groups except some White (non-English speakers) and Chinese categories are under-represented.

Focusing on geography over the urban-rural spectrum, Hecht and Stephens [2014] explored three of the most common sources for geotagged information — Twitter, Flickr and Foursquare — at a level of detail greater than simple adoption rates and content coverage [Mislove et al. 2011]. Considering the United States' population, they collected a

[5]https://pewrsr.ch/2HibNa0

Twitter dataset of 56.7 million geotagged tweets from 1.6 million users and a Foursquare dataset which contains 11.1 million check-ins from approximately 122,000 users. Utilizing a geostatistics-based approach, they compared the spatial distributions of social media users to the US Census data at county level — similar to Mislove et al.. However, they addressed two key challenges related to the mobility of users (e.g. excluding tourists visiting rural counties), and the spatial autocorrelation between some attributes of each data source and the population (e.g. the tendency of increase in Twitter activity per user in urban regions). They demonstrated that there is a population bias towards urban regions at the expense of rural ones, e.g. there are 24.4 times more Foursquare users and 5.3 times more tweets per capita in urban areas than rural ones.

For a stronger methodological investigation, Malik et al. [2015] carried out statistical analysis to investigate whether Twitter users with geotagged tweets are randomly distributed over the United States' population. They collected a Twitter dataset of 145 million tweets from 2.6 million users. Unlike previous research, they combined Twitter data with Census data representing the smallest geographic units to have roughly comparable population sizes, i.e. could be an administrative region or a venue such as an airport or a park. However for comparison with previous research, they considered the same populations for race as Mislove et al. [2011], age as Longley et al. [2015], gender as Mislove et al. [2011], Longley et al. [2015] and urban/rural spectrum as Hecht and Stephens [2014]. They demonstrated that users with geotagged tweets are not representative of the US population, with established biases towards younger users, users in urbanized areas (especially on the east and west coasts of the US), Hispanic/Latino and Black users.

Pavalanathan and Eisenstein [2015] compared the demographic biases over age, gender and geography in geotagged Twitter data. Unlike previous research, they also considered the bias introduced by the two most common data acquisition techniques, namely geotagged tweets (GPS-tagged) and the location field in a user profile (self-reported). Focusing on the ten largest metropolitan areas in the United States at the county level, they demonstrated that the urban bias still exists in the GPS-tagged data. Considering acquisition techniques, GPS-tagged data tend to include younger users and significantly more women in comparison with self-reported data. Moreover, they explored the influence

of these biases on two research task, which we will discuss in the next section.

Sloan and Morgan [2015] examined the representativeness of geotagged Twitter users in relation to Twitter population and users who enable location services (i.e. GPS), but do not geotag their tweets. Utilizing two Twitter datasets, they explored the demographic differences over age, gender, class, the language in which tweets are written, and the user-interface language. They demonstrated the existence of statistically significant differences for all demographic characteristics, albeit with substantial differences in the magnitude by factor. For instance, the differences were very small as in the range of $0.1 - -0.8\%$ by gender, and $0.55 - -0.82$ years by age. On the other hand, the biggest differences were related to language. For example, user tweeted in Russian were the least to enable geolocation (18.2%), although they were not in the lowest group for rate of geotagging (2%), in contrast to Korean users (28.9% and 0.4%, respectively). At the end, they suggested that geotagged Twitter users are not representative of either the Twitter population or users who enable location services.

External gazetteer (mainly Census Data) played a crucial role in revealing population bias in social media. Unlike other demographic factors (e.g. gender, age, and geography), language has very limited external resources to help identify such bias. Language bias identification, therefore, relies on the interplay of language usage and geography bias in social media.

2.3.2 Language Bias

Language bias in social media can be revealed using three different approaches. In the field of dialectology (described in § 2.1.1), linguists followed a *traditional methodological* approach to study the evolution of languages over geography and time based on interviews and questionnaires with a small number (typically, a few hundred) of participants [Labov et al. 2008]. However, this approach is naturally limited in scope, based on the choice of participants and the small number of identified lexical variations, and time consuming.

With the rise of social media, researchers followed a *computational linguistic* approach that leverages the massive amount of data generated by users, to investigate the geographic

linguistic variation of languages. Eisenstein et al. [2010] presented a model that reasons jointly about latent topics and geographical regions, known as topic modelling. Focusing on English in the United States, they showed that geographic regions have different topic interests, such as sports or music. To evaluate this approach, they assessed their model's ability to predict the geographic location of Twitter users based on their text alone. However, this approach does not differentiate between two specific types of regionally affiliated terms, namely non-standard (slang) and entity (landmark) terms.

Gonçalves and Sánchez [2014] proposed a *hybrid* approach combining traditional and computational methods. They focused on Spanish because it is one of the most spoken languages in the world and spatially distributed across several continents. To determine the major local varieties of Spanish, they filtered their initial dataset of 50 million geotagged tweets based on a list of concepts and utterances, such as 'popcorn', 'car', 'bus', etc., selected from a database of lexical variants in major Spanish-speaking cities [Tinoco and Ueda 2007]. The remaining 750,000 geotagged tweets were aggregated geographically into a uniform grid. Lastly, they applied computational methods to identify the dominant word for each concept in each geographic cell. This approach was able to identify two Spanish super dialects (in main American and Spanish cities) and five local dialects (in rural areas and small towns).

Eisenstein et al. [2010] demonstrated the value of topic modelling to enhancing the quality of Twitter user geolocation for English. Gonçalves and Sánchez [2014] established the value of language bias in social media to study the evolution of Spanish as well. However, their approach was not evaluated on Twitter geolocation. Only few researchers examined the influence of language bias on Twitter geolocation for languages other than English.

2.3.3 Language Influences on Twitter Geolocation

In the early years of Twitter (2006–2010), most of the users were located in the United States and English was the prominent language. This bias naturally influenced research. The majority of geolocation research relied on datasets of English tweets that cover limited

geographic regions (e.g. 3–4 locations), mainly in the United States.

To the best of our knowledge, only two prior works have evaluated the impact of a language on geolocating users [Han et al. 2014] and tweets [Priedhorsky et al. 2014]. Both claimed that locating users/tweets writing/written in languages with restricted regional coverage were easier to geolocate than those writing in widely used languages.

Priedhorsky et al. [2014] examined the effect of a language as a feature in a multilingual model trained on a dataset of 13M geotagged tweets, showing that language is a valuable feature in geolocation prediction models. However, they did not evaluate their models on a per language basis.

Using a multilingual dataset of 23M geotagged tweets, Han et al. [2014] showed that training separate per language models lead to higher accuracy. They noted that for some languages, geolocation accuracy was higher than for others. To explore user distribution in the geographical region of that language, the authors measured the entropy of tweeters in cities on a per language basis. However, they did not correlate entropy with an evaluation measure, neither did they examine other features of languages that might impact on evaluation. Moreover, all languages other than English in their dataset are under represented. At the time the English subset had more than 1M users, few thousands (1–10) of users tweeted in other languages, per each. As a result, the number of users per city tweeting in a specific language might not reflect the real geographical coverage of this language.

As Twitter popularity increased, researchers started to have more interest in transition to a global scale. In this thesis, we investigate the influence of language on Twitter geolocation.

2.4 Evaluation

In this section, the evaluation metrics of Twitter geolocation are presented from two different perspectives, see Section 2.4.1. First, each metric is defined based on the research work where it first appeared and the intuition behind it, while highlighting any subsequent modifications. Second, the evolution of evaluation metrics over time is presented beside the popularity of each metric. This chronological presentation reveals some of the decisions

taken by subsequent researchers and the impact of such choices on the evaluation process, including the limitations of accuracy based metrics (§2.4.2), the availability of previous Twitter datasets (§2.4.3), geographic granularity (§2.4.4), accounting for population bias (§2.4.5), and comparing metrics to assess the added value of each (§2.4.6).

2.4.1 Metrics

Evaluation metrics are categorized into three groups based on the output representation and the application requirements. In continuous evaluation, locations are represented as GPS coordinates. Discrete evaluation treats locations as either administrative regions, grid cells or a list of candidate venues. A hybrid evaluation combines the metrics of continuous and discrete evaluation.

Continuous Evaluation

Evaluation of geolocation in Twitter was initially measured using *Median* and *Mean* error distances between the estimated and true locations [Eisenstein et al. 2010]. Some studies [Han et al. 2014, Melo and Martins 2017] promoted the usage of median error distance for evaluation because it is more robust to outliers than the mean. One study [Schulz et al. 2013] employed Mean Squared error as well. For unified notations, let p be the true GPS coordinates of a user (u) or tweet, and $\hat{p}$ be the estimated GPS-point. Error distance metrics are defined as:

$$ErrDist(u) = d\{p, \hat{p}\}$$

$$MedianErrDist = median_{i=1}^{n_{\text{users}}}\{ErrDist(u_i)\}$$

$$MeanErrDist = \frac{1}{n_{\text{users}}} \sum_{i=1}^{n_{\text{users}}} \{ErrDist(u_i)\}$$

$$MeanSqErrDist = \frac{1}{n_{\text{users}}} \sum_{i=1}^{n_{\text{users}}} \{ErrDist^2(u_i)\}$$

Discrete Evaluation

Researchers also used accuracy (Acc) at different geographic granularities (e.g. city, state, country, or list of venues) for a better interpretability of the results compared to error distance. For instance, Eisenstein et al. [2010] measured accuracy of classification by state and by region of the United States. The choice of spatial granularity in the majority of previous research 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). Let (l) and ($\hat{l}$) be the true and predicted locations of a user (u), tweet, or recognised location. Accuracy is then defined as the percentage of correct predictions:

$$Acc = \frac{1}{n_{\text{users}}} \sum_{i=1}^{n_{\text{users}}} 1(l_i = \hat{l}_i)$$

Precision, Recall and F1-score are common alternative measures for accuracy in the context of classification tasks. Geolocation systems might not be able to infer the location of a user or tweet in some cases where it relies on the user's social network that does not exist in the dataset. Let (l) = null if the system can not make any prediction. Precision calculates the percentage of correctly inferred locations, while Recall represents the percentage of users or tweets for which the geolocation system was able to infer a location [Davis Jr et al. 2011]:

$$Precision = \frac{1}{n_{\text{l}\neq\text{null}}} \sum_{i=1}^{n_{\text{users}}} P(l = l_i, \hat{l} = l_i)$$

$$Recall = \frac{1}{n_{\text{users}}} \sum_{i=1}^{n_{\text{users}}} 1(l \neq null)$$

Note that Recall definition introduced by Davis Jr et al. is different from the survey paper, which considered only correct predictions instead of non-null [Zheng et al. 2018]. In some studies, Recall is also referred to as the post coverage of the network [Jurgens et al. 2015b]. F1-score, the harmonic mean of Precision and Recall, is mainly employed in the task of mentioned locations recognition. Rodrigues et al. [2016] reported Precision

and Recall for each of the ten different Brazilian cities in their dataset and overall macro-averaging scores, which we further extended in Chapter 3 to consider micro, weighted, and macro averaging techniques at the level of the three metrics.

Hybrid Evaluation

Several metrics based on accuracy and/or error distance were introduced for a better insight into the distribution of location estimation errors. Backstrom et al. [2010] evaluated performance visually based on the fraction of predictions within d kilometers from the true location using a Cumulative Distribution Function (CDF) for all values of d within 10,000 km. Similarly, Cheng et al. [2010] introduced accuracy within 100 miles or 161 km from the original city. For generality, *Acc@d* is defined as the percentage of users or tweets with their error distance less than the d threshold:

$$Acc@d(p, \hat{p}) = \frac{1}{n_{\text{users}}} \sum_{i=1}^{n_{\text{users}}} 1(ErrDist(u_i) \leq d)$$

In some cases, geolocation models rank a list of predicted locations in a decreasing order of confidence. This gives geolocation systems the opportunity to identify a good candidate location, even if the first prediction is wrong. In light of this, Cheng et al. [2010] introduced accuracy within the top k locations (*Acc@k*), which is the same as *Acc@d*, but over the location in the $top - k$ estimations with the least error distance to the actual location. Kinsella et al. [2011] defined *Acc@k* slightly different as the percentage of predicted locations which lie within k hops of the correct location. If $k = 1$, direct neighbours only are accepted as correct locations. In this case, *Acc@k* requires the knowledge of the geographical boundaries of regions.

Figure 2.2 shows the popularity of the evaluation metrics employed in existing Twitter geolocation research. For Twitter user and tweet geolocation tasks, Median and Mean are the most popular evaluation metrics, followed by Acc and Acc@d, while Precision, Recall and F1-Score are rarely used. For named entity recognition, Precision, Recall and F1-Score are dominating, while Acc is used once and the rest of the metrics never used.

Table 2.1 details a chronological ordering of Geolocation Metrics. Other research

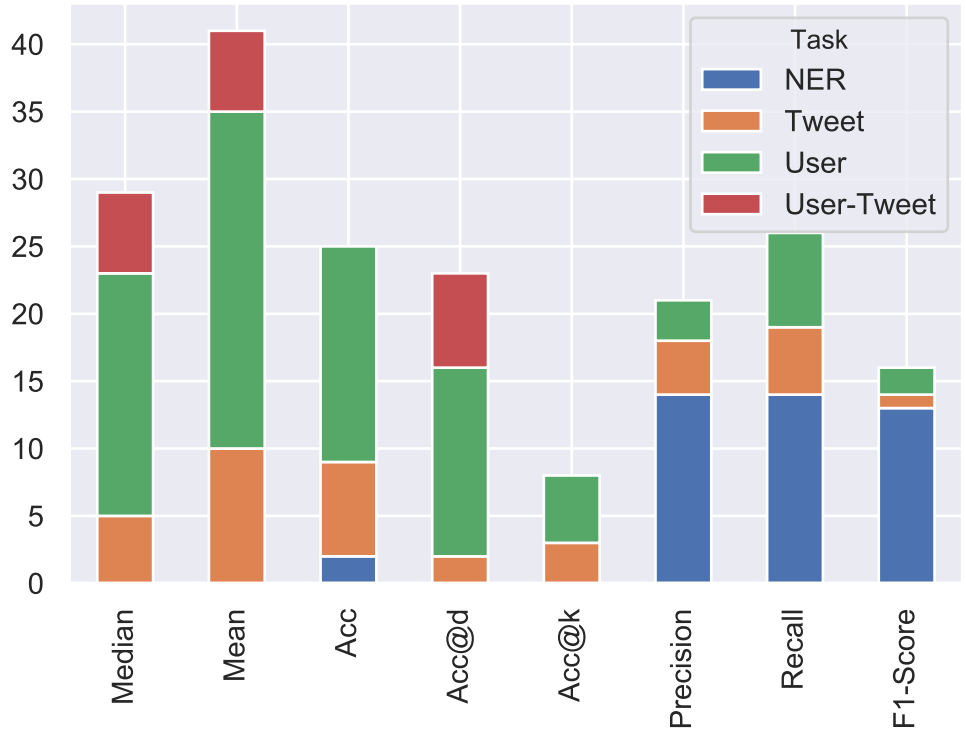


Figure 2.2: Twitter geolocation metric popularity in existing research. A total of 80 research papers are considered.

employed a combination of these measures, as described in this table. 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.

2.4.2 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*) [Roller et al. 2012, Han et al. 2014, Wing and Baldridge 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,[6] the average distance between cities and their neighbours was found to be in the range of 32–46 miles (§3.4.3), less than half the 100 miles threshold. A system which predicts the location

[6] https://github.com/tq010or/acl2013

Table 2.1: 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 Baldridge [2011]			✓	✓		GeoText			US		
Roller et al. [2012]			✓	✓		GeoText			US		
		✓	✓	✓		**UTGeo**	449k	38M	Nth Am	10k (2.22%)	
Ahmed et al. [2013]				✓		GeoText			US		
Han et al. [2014]	✓	✓	✓			UTGeo			Nth Am		
						World	1.4M	12M	Global	10k (0.71%)	10+
Wing and Baldridge [2014]		✓	✓	✓		UTGeo			Nth Am		
						World			Global		10+
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+
Rahimi et al. [2016; 2017; 2018]		✓	✓	✓		UTGeo			Nth Am	10k	
						World	1.4M	12M	Global	10k (0.71%)	10+
Miura et al. [2017]	✓	✓	✓	✓		UTGeo	279k	23.8M	Nth Am	10k	
						World	782k	9.03M	Global	10k	10+
Do et al. [2017]		✓	✓	✓		GeoText	9.5k	>370k	US	1.9k (20%)	
						UTGeo	450k	38M	Nth Am	10k	
						WORLD	1.4M	12M	Global	10k	
Ebrahimi et al. [2018]		✓	✓	✓		GeoText	9.5k	380k	US	1.9k (20%)	
						UTGeo	450k	38M	Nth Am	10k	
						WORLD	1.4M	12M	Global	10k	

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. 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.

2.4.3 Data Decay

Table 2.1 (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 (named WORLD) for a user geotagging shared task. 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).

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.

2.4.4 Mismatching Geographic Granularity

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

Additional challenge related to earth-representation in shared tasks is that the organisers provide a set of pre-defined locations for all participants without providing the library for geocoding locations or reverse-geocoding coordinates [Han et al. 2016]. This implicitly enforces a specific earth-representation. A participating team decided to use a different earth-representation to train their models [Jayasinghe et al. 2016]. First, they had to map the provided set of locations to their local geocoding library. Then, they mapped the results back to the original set of locations.

2.4.5 Accounting for Population Bias

With the inherent population bias in social media data across different demographic factors, and the rise of large scale studies of human behavior, Ruths and Pfeffer [2014] emphasized the importance of correcting for these biases. They highlighted the best practices for managing analytical bias at the levels of data collection and social media-based methods. With the growing interest in complex machine learning algorithms, models tend to employ hundreds and thousands of features which overfits the training data. To avoid this overfitting, the p-value for classifiers based on the number of features involved should be reported; beside the usage of a development data set for feature engineering, and an independent test data set for final evaluation. Multiple hypothesis testing — the use of more data sets in a single study — would guarantee that the successful findings of a study are not the result of random chance. Forcing the sharing of methods at publication time would resolve the incompatibility of methods and data (as mentioned in Section 2.4.3). However, few researchers followed best practices to explore the impact of this bias on either determining the most effective models or evaluation metrics for social media-based algorithms.

Culotta [2014] investigated methods to quantify and control for Twitter population bias in relation to census data. They estimated several health statistics (e.g. obesity, diabetes) of the top 100 counties in the United States. Using standard survey re-weighting to correct for gender and race biases, they balanced their data based on the mismatch between Twitter demographics and the US Census data. For instance, if the female population is under-represented to the half, every female user in the Twitter dataset will be counted twice. Their results demonstrated that adjusting for population bias improved the accuracy of predictions.

Working on the same task of health statistics, Landeiro and Culotta [2016] examined how to control for a *confounding variable* (e.g. gender) that influences both the input to classification algorithms (e.g. text data) and the *target variable* (e.g. cancer). For example, suppose the task is to predict whether a user has cancer or not and that smoking was found to cause cancer in the general population, but not when breaking population into

males and females. Considering the case where the distribution of health status by gender is imbalanced, Landeiro and Culotta [2016] addressed two challenges. First, the noise introduced by distinctive language used by men (if any), but not related to their health status. Second, the change in the sign of correlation when considering the general population versus sub-populations based on gender. On three text classification tasks, namely Twitter user geolocation, IMDB movie review prediction, and Twitter party affiliation prediction, they demonstrated that using back-door adjustment improved robustness of classifiers.

Pavalanathan and Eisenstein [2015] explored the influence of Twitter user demographics — gender, age and geography — on the tasks of geographic lexical variation and text-based Twitter user geolocation prediction. Utilising two types of geographic lexical variables, namely non-standard (slang) and entity (landmark) terms, they demonstrated that GPS-tagged data include more non-standard words, young people use significantly more slang words, and men tend to mention entities significantly more than women for the age group of 30 or older. Resampling based on the county census data to account for the overrepresented urban bias did not impact their lexical analysis. GPS-based ground truth datasets tend to have more geotagged tweets than self-reported technique, making Twitter user geolocation prediction insignificantly easier. Geolocating men and old users are significantly more accurate then females and young users, respectively.

Johnson et al. [2017] further explored the impact of geographic bias across the urban-rural spectrum on Twitter geolocation at the levels of user and tweet. Following the standard practices to build the ground-truth for each of the tasks, they collected two Twitter datasets focusing on the United States at county level. 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. Two open sourced geolocation inference systems [Priedhorsky et al. 2014, Jurgens 2013] were considered to cover the most common approaches for inference (text-based and network-based, respectively). In line with previous results, they found that urban counties are overrepresented relative to census data. To reveal the algorithmic bias, the two geoinference approaches were evaluated at the level of two geographic categories separately (urban and

rural) in comparison to the overall precision. They demonstrated that existing geolocation approaches perform significantly worse for rural areas than for urban. While correcting for the population bias (stratified-sampling) had no effect on the text-based model (in line with Pavalanathan and Eisenstein [2015]), the network-based model was significantly less urban-biased. Similarly, oversampling to match the general population, and training and testing on each geographic category separately showed that the text-based algorithm was still prune to urban-bias, while the rural network-based model outperformed the urban model. They concluded that both population and structural biases contribute to the algorithmic bias of Twitter geolocation methods.

While relying on an external gazetteer (e.g. US census data) is helpful to reveal the demographic bias in social media, it has two major limitations. First, census data might not be available for all countries with the same degree of accuracy as in the US. Second, some demographic factors, such as the language or dialect, are not represented. Other methods, such as consolidating geographic regions into two classes only (rural-vs-urban) or evaluation of individual categories (e.g. accuracy of male vs female) limits the scalability of these approaches to reveal the influence of algorithmic biases. Most of the recent work relies on datasets with global geographic and language coverage, with hundreds and thousands of classes and features, but 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.

2.4.6 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. In Chapter 3, we pointed out that accuracy measures are biased towards

locations with a large population. Although a wide range of metrics was employed, the work was limited to a single geolocation model while focusing on the influence of language rather than the effectiveness of the evaluation measures. Chapter 5, therefore, focused on the effectiveness of geolocation evaluation regardless of the underlying geolocation approach or the language of text. The relative performance of fifteen geolocation models and two baselines were evaluated using all the metrics in Table 2.1.

2.5 Summary

In this chapter, we described downstream geo-spatial applications that benefit from Twitter geolocation at several granularities. In specific, we highlighted the different information needs required by such applications. Twitter datasets are substantially biased based on geographic regions and language, influencing the value of existing geolocation systems. This thesis aims to improve the effectiveness of evaluating geolocation systems, taking into account the information needs of applications and the inherent biases in social media datasets. It concentrates on the influence of language bias on the quality of geolocation in Chapter 3, Chapter 4, the effectiveness of the evaluation process for geolocation systems taking into account the geographic bias in Chapter 5, and Chapter 6.

CHAPTER 3

Language Influences on User Geolocation

In Chapter 2, we showed that English is the most prominent language in the datasets of previous research, with only a few researchers having considered the language in which a tweet is written as a feature for Twitter geolocation. However, the datasets used were of different sizes in comparison to English. In this chapter, we investigate the influence of language on the accuracy of geolocating Twitter users. To conduct our study, we required the following: collections of users of comparable size on which to measure location accuracy (Section 3.1), a geolocation system and evaluation measures (Section 3.2). In Section 3.3, we first perform a preliminary study on one language other than English (Arabic in this case) to validate the language influence hypothesis in previous research and explore alternative evaluation metrics. Based on the outcome, we then extend the experiment to a multi-lingual corpus of tweets written in thirteen languages to have a better understanding of the features (including language) that cause performance disparities between per-language geolocation models, Section 3.4.

3.1 Data

The main objective of this chapter is to examine the language influence on text-based geolocation prediction on large-scale datasets. Han et al. [2014] investigated the influence of 10 different languages on prediction accuracy. However, all languages except English spanned only a few hundreds users in their dataset, named WORLD. First, we chose one

of these under-represented languages which they claimed is geographically limited, Arabic, to conduct a preliminary study using a dataset named ArQAT. Second, we employed a third multi-lingual dataset named TwArchive, with comparable user sizes per language. All three datasets have geographic global coverage and are mainly composed of geotagged tweets.

ArQAT is a collection of Arabic geotagged tweets gathered in collaboration with Qatar University[1] via the Twitter public Streaming API[2] from 25 May, 2014 to 19 Aug, 2014. Global coverage is guaranteed by not forcing any geographic restrictions. Due to the limitations of the Twitter Streaming API, ArQAT was collected over two stages to maximize the retrieved sample within a reasonable time. First, random sampling from the Streaming API was used, followed by filtering of non-Arabic Tweets based on the language reported by Twitter. This resulted in 2M Arabic tweets collected over 9 days. Second, a list of 400 top-frequent Arabic words was extracted from this collection. Finally, the terms in this list were used to search for tweets on the live stream of Twitter. The resulting dataset contained over 1 billion Arabic tweets, of which 10.8M are geotagged, from 365.5K unique users.

WORLD is a multi-lingual dataset spanning five months from late 2011 to early 2012 [Han et al. 2014]. Originally WORLD contained 2.1M users and 23M geotagged tweets. The number of tweets and users in the English dataset alone, WORLD-en, was 12M and 1.4M, respectively. In reconstructing it from the tweet IDs released by the authors, 27% and 30% of users and tweets, were deleted. The decay in the reconstructed dataset is expected as discussed earlier (§ 2.4.3). However, this decay will not influence reproducing the geolocation system published by Han et al. or change their findings, as demonstrated in 3.3.

TwArchive is a public multi-lingual dataset holding over four years of content drawn from the 1% sample Twitter public API stream provided by the Internet Archive Library[3]. We used a 2014 subset spanning nine months. We use the TwArchive dataset to inves-

[1]This work was made possible by NPRP grant# NPRP 6- 1377-1-257 from the Qatar National Research Fund (a member of Qatar Foundation).

[2]https://dev.twitter.com/streaming/overview

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

tigate how the variance in performance between different languages changes as dataset size increases, and how consistent the factors that cause variance in performance between languages across different data collections are.

3.2 Experimental Setup

Data Pre-processing

Languages in *the multi-lingual* collections were separated using langid.py[4] [Lui and Baldwin 2012]. Languages studied are Arabic (ar), English (en), Spanish (es), French (fr), Indonesian (id), Italian (it), Korean (ko), Malaysian (ms), Dutch (nl), Portuguese (pt), Russian (ru), Thai (th), and Turkish (tr). Text was tokenised using a Twitter-specific tokeniser [Han and Baldwin 2011]. Arabic text was normalized using Tashaphyne[5] and an Arabic social media normalizer [Darwish et al. 2012]. Normalization changed only the orthography of Arabic words. Use of the extra systems was necessary to reduce the sparsity of words. All non-alphabetical tokens, and tokens with length < 3 characters, were removed.

Non-geotagged and duplicate tweets (using user id and tweet text) were removed. Duplicate tweets were identified using exact match of text. Cities with fewer than fifty location indicative words were removed to ensure a representative sample of words per city. Each user was assigned a home city based on their geotagged tweets. A publicly available offline search library[6] was employed to either reverse-geocode a given GPS coordinate to the corresponding city, or [none]. A user's home city is the one associated with the simple majority of their tweets; in a tie, the first city is chosen. Users with an unresolved home city (i.e. [none]) were removed from the dataset. Users eligible for testing are required to have at least ten geotagged tweets. All previous processing steps were adopted from previous work [Han et al. 2014] for a fair comparison, except for the Arabic normalization. Common pre-processing steps for Twitter datasets such as handling retweets or using

[4] An open source language identification tool, trained over 97 languages, and tested over six European languages with an accuracy of 0.94. Predictions with confidence ≥ 0.5 only are considered.

[5] http://pythonhosted.org/Tashaphyne/

[6] https://github.com/tq010or/acl2013

Table 3.1: Number of users, test users with at least ten geotagged tweets (percentage), tweets, cities, number of cities with eligible test users and countries after preprocessing.

		en	**es**	**it**	**pt**	**id**	**nl**	**fr**	**ms**	**ko**	**ru**	**ar**	**th**	**tr**
#Users	ArQAT	–	–	–	–	–	–	–	–	–	–	328k	–	–
	WORLD	947k	242k	118k	111k	103k	94k	79k	64k	36k	29k	28k	27k	24k
	TwArchive	1.5M	541k	119k	284k	225k	59k	136k	136k	22k	73k	94k	49k	211k
#TestUsers (%)	ArQAT	–	–	–	–	–	–	–	–	–	–	32	–	–
	WORLD	14	10	3	12	8	8	3	7	6	15	16	12	10
	TwArchive	3	2	0.7	6	0.3	0.4	3	0.2	0.9	2	3	4	1
#Tweets	ArQAT	–	–	–	–	–	–	–	–	–	–	8.5M	–	–
	WORLD	6.2M	1.2M	267k	670k	423k	381k	198k	222k	122k	196k	215k	156k	108k
	TwArchive	3.1M	1.1M	162k	836k	317k	74k	295k	179k	32k	147k	207k	127k	351k
#Cities	ArQAT	–	–	–	–	–	–	–	–	–	–	2k	–	–
	WORLD	2.9k	2.2k	2.1k	1.8k	1.9k	2k	2k	1.6k	1.1k	894	881	413	1.3k
	TwArchive	3.2k	2.3k	2.2k	1.9k	2k	2k	2.2k	1.7k	1.7k	1k	1.6k	727	1.6k
#TestCities	ArQAT	–	–	–	–	–	–	–	–	–	–	1k	–	–
	WORLD	1.7k	849	343	404	272	112	274	180	132	299	277	90	159
	TwArchive	1.4k	460	94	233	33	89	174	57	52	204	188	58	99
#Countries	ArQAT	–	–	–	–	–	–	–	–	–	–	152	–	–
	WORLD	169	151	150	132	145	140	154	125	96	94	90	64	116
	TwArchive	173	159	156	139	147	148	164	142	129	107	139	80	147

similarity measures for detecting duplicate tweets are not considered in this study.

Table 3.1 shows that for all languages, users are spread over thousands of cities and tens of countries. ArQAT is spread over 2k cities with Arabic tweets coming from English speaking countries such as the United States and the United Kingdom. Although the TwArchive time span is almost double that of WORLD, the number of tweets fell by around a half for some languages (e.g. *en*, *it*, *nl* and *ko*). We speculate that, between 2011 and 2014, users have become more concerned about revealing their location. On the other hand, the number of users increased across all languages, which aligns with the continuous increase of Twitter active users base since 2010[7]. These changes led to a drop in the average number of geotagged tweets per user from 5 in WORLD to 2 in TwArchive.

One of the key questions we explore in this work is what defines the difficulty of a language? Is it the geographic coverage represented by the number of cities and countries where users come from? Table 3.1 demonstrates that all languages are spread over thousands of cities and hundreds of countries, in contrast to the perception that some languages are geographically limited [Han et al. 2014]. This argument is also supported by the number of multi-lingual users as shown in Figure 3.1. Around 25% of the total number of users in the WORLD data collection (for all languages) post in more than one

[7]http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

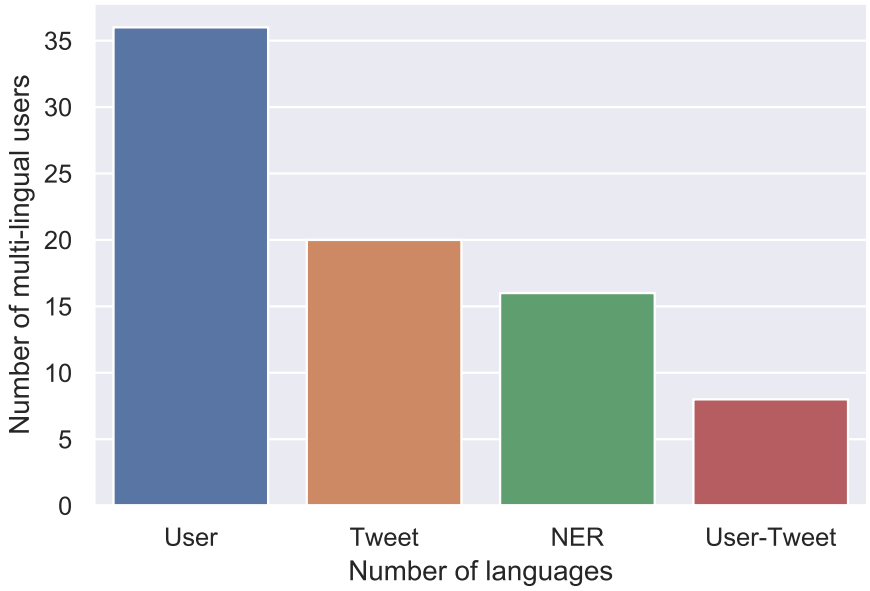


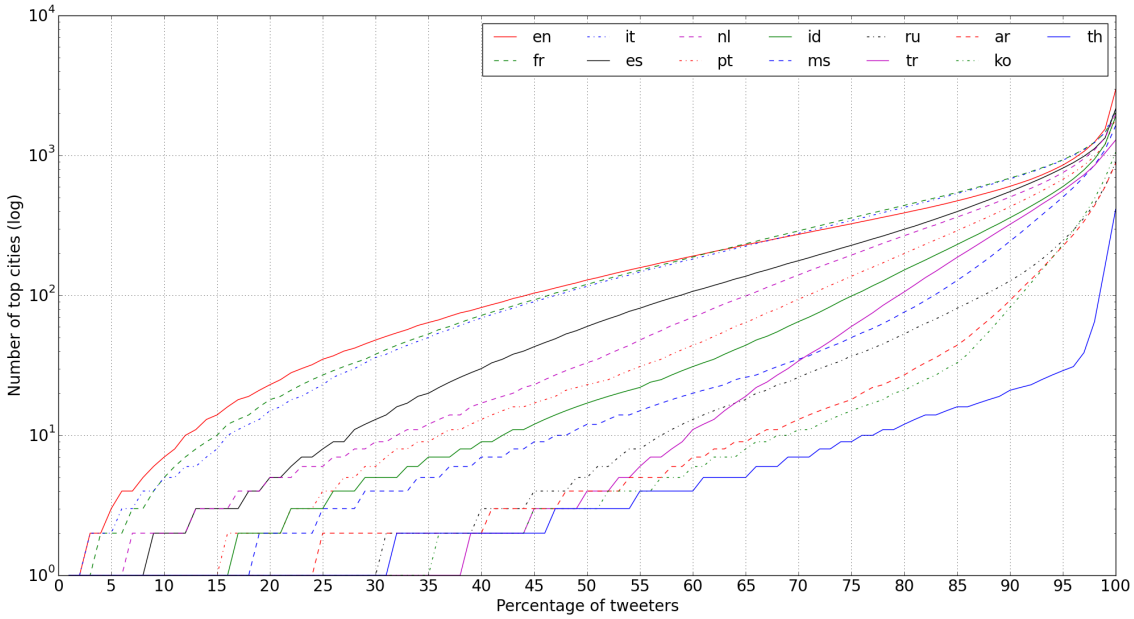
Figure 3.1: Number of multi-lingual users for the top 13 languages in the WORLD dataset. They totally represent 25% of the total number of users.

language, with the absolute majority tweeting in 2 languages (64%).

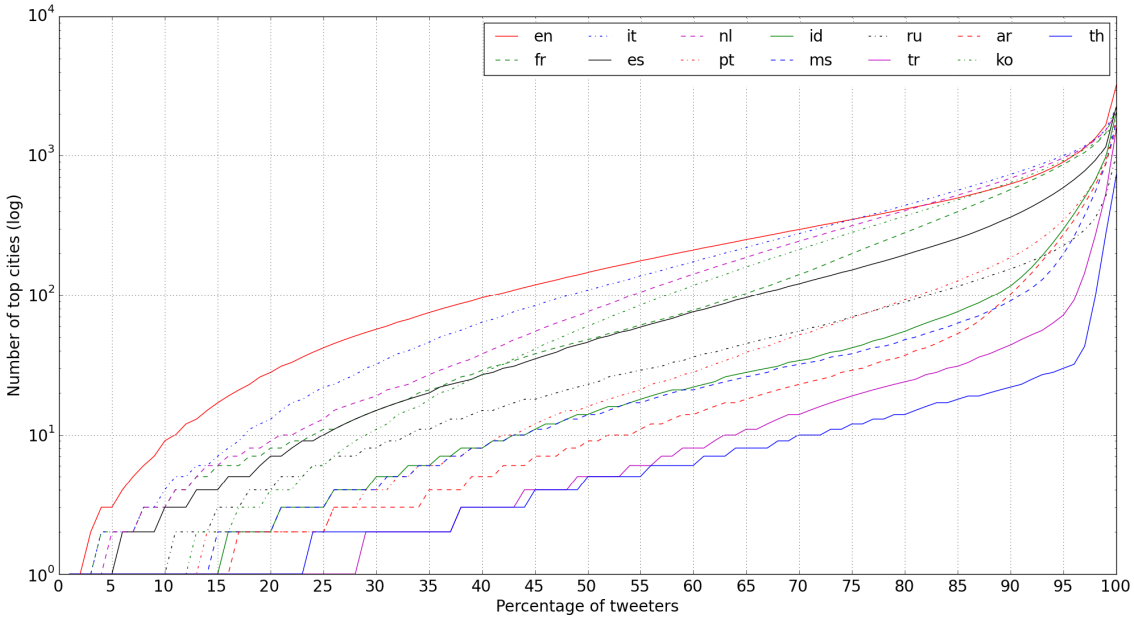
For a better understanding of the data balance based on city partitioning, the cumulative distribution of users over cities is shown, per language, in Figure 3.2. Examining where the plot lines intersect the x-axis in Figure 3.2a, we see that for *en*, *fr* and *it*, no single city contained more than 4% of all users for that language. For languages, such as *id*, and *ar*, one city contained 15–25% of users. For languages, such as *tr*, *ko*, *th* and *ru*, one city contained more than 30% of users. TwArchive was found to be slightly more balanced where for five languages no single city contained more than 5% of users, for six languages one city contained 10-16% of users, and for two languages one city contained more than 20% of users, see (3.2b).

Text-based User Geolocation Model

From the existing text-based geolocation approaches, this work is based on the research that addressed language influence [Han et al. 2014], which 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



(a) WORLD



(b) TwArchive

Figure 3.2: Users' cumulative distribution over cities in WORLD and TwArchive.

using their best reported feature selection method, Information Gain Ratio, as detailed in Section 2.2. Then a Multinomial Naïve Bayes (MNB) prediction model per language was trained using scikit-learn [Pedregosa et al. 2011]. Only MNB was considered for fair comparison with previous research focusing on the language influence. Further machine learning techniques are examined in Chapter 5.

Most Common Evaluation Measures

To measure accuracy, we initially considered the three most common evaluation metrics drawn from past work, as shown in Section 2.4: **(1) Acc,** city-level accuracy; **(2) Acc@161,** accuracy within 161 km (100 miles)[8]; **(3) Median,** median error distance between predicted and actual cities (km).

3.3 Large-Scale Preliminary Analysis for Arabic

In this section, the same Twitter user geolocation prediction model is trained on English and Arabic datasets separately. In particular, we investigate how the approach of measuring language influence introduced in previous research scales to another language of a comparable dataset size. The datasets considered are ArQAT, and the Arabic (ar) and English (en) subsets of WORLD. Baselines are first discussed, and our re-implementation of the Han et al. [2014] geolocation model is validated over the datasets considered for this preliminary study. Next, the trained geolocation models are evaluated based on the most common measures employed in previous research. Lastly, alternative evaluation measures are proposed to reveal the influence of the data imbalance over locations discussed in the previous section (§3.2).

3.3.1 Baselines

Our geolocation prediction model trained on the ArQAT dataset is compared to two baselines: (1) Model trained on the Arabic WORLD-ar dataset. (2) Model trained on the English WORLD-en dataset. Both baselines are tested on the same number of users in the original study by Han et al. [2014], which are 164 for WORLD-ar and 10k for WORLD-en. The model trained on ArQAT is tested on 10k users to validate the hypothesis of previous research. Both baselines are based on our implementation of a single Twitter user geolocation model Han et al., which scores unigrams extracted from the training sets after post-processing using IGR (Equations 2.1 and 2.2) to identify Location Indicative Words

[8]Accuracy within 161km might not be an effective evaluation measure from a language comparison perspective because of the arbitrary choice of a 161km threshold on a U.S dataset as discussed in §2.4.2, however as it has been used in past work, we use it here.

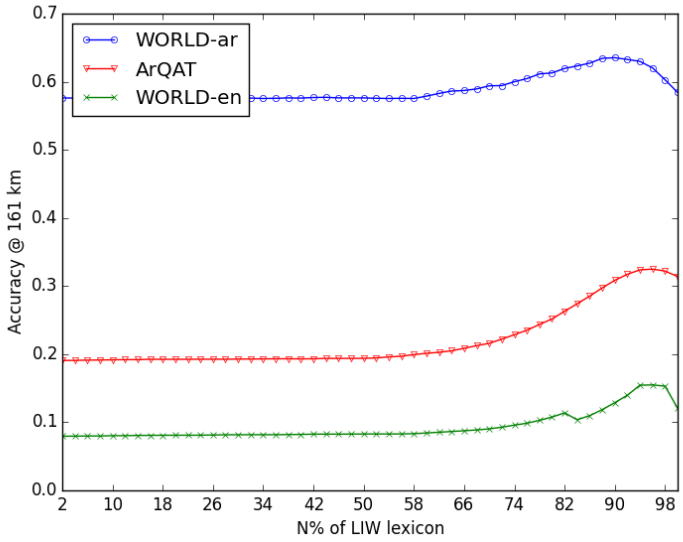


Figure 3.3: Acc@161 for incremental use of top N% of LIW

(LIW). The initial numbers of unique terms in ArQAT, WORLD-en, and WORLD-ar are 1.2M, 548k, and 138k respectively. This large difference between the Arabic and English unique terms could be due to the fact that Arabic is a morphologically rich language, ArQAT has 69M occurrences of terms compared to 10M in WORLD-en, besides having more than 34 variations of dialects[9].

Following the approach of [Han et al. 2014] to determine the best cut-off threshold of LIW, the model is trained incrementally with the top N% features in steps of 2%. Results in Figure 3.3 show that 90%, 96%, and 96%; for WORLD-ar, ArQAT and WORLD-en, respectively, achieve the best results in terms of *Acc@161* using 10 samples cross-validation. This aligns with Han et al.'s outcome indicating that best results could be achieved by using a large subset of the LIW lexicon to overcome underfitting, in case of small top N%, yet remove noisy, common and even stop words causing overfitting (§2.2).

3.3.2 Results

Considering the WORLD-ar–ArQAT comparison, note that the two datasets are different in terms of content and scale. However, this comparison is still important to validate

[9]http://www.ethnologue.com/browse/names

Table 3.2: Evaluation of the prediction models using the most common metrics.

	Train	Test	Acc	Acc@161	Acc@C	MedArr
WORLD-ar	28k	164	0.53	0.660	0.850	7
WORLD-en	937k	10k	0.114	0.228	0.826	922
ArQAT	318k	10k	0.255	0.369	0.708	390

the previous hypothesis of the language influence when the dataset scale for the same language changes. The results in Table 3.2 show a large drop in performance. There is a decrease in *Acc* and *Acc@161* by almost 20 points and median error distance increases from 21km to 391km. The large variance between ArQAT and WORLD-ar in terms of the number of cities in each dataset and the range (min-max) of cities frequencies (see columns #Cities and RCF in Table 3.2), indicates that ArQAT is highly imbalanced, which explains the relatively low performance. In other words, WORLD-ar has a better coverage and balanced distribution because 30% of the cities encompass 95% of the tweets. This result suggests that the users distribution over cities of social media datasets, correlated with the dataset size, has an impact on the quality of estimation, and the outcomes of small scale experiments can't be generalized.

Despite the global geographic coverage of ArQAT and WORLD-ar in terms of number of cities (see Table 3.1), the Arabic language user prediction models are more accurate than the English model as shown in Table 3.2. The large difference in the number of cities, in favour of English, could support the hypothesis that geolocation prediction is much easier for languages with restricted geographic coverage like Arabic. That could be true based on the most common evaluation metrics including, *Acc*, *Acc@161* and MedArr error distance. However, these evaluation metrics do not reflect the impact of imbalance revealed earlier from a machine learning categorization perspective.

3.3.3 Considering Alternative Measures

Accuracy, as an evaluation measure, is insensitive to skewed training datasets [Yang and Liu 1999]. *Acc* and *Acc@161* are, therefore, heavily influenced by the accuracy of the geolocation system on a limited number of cities. As long as the system geolocates correctly

on a few well populated cities, the accuracy will be high.

Evaluation measures are designed to estimate how well a system will do in a particular task. In the introduction, we stated that one use of a geolocation system is finding eyewitnesses. It is perhaps worth asking if the geographic distribution of eyewitnesses needed by a news organization will match the skewed distribution of Twitter users reflected in the bias of accuracy measures. In this section, we explore alternative measures commonly used to evaluate classifiers when data is imbalanced [Sebastiani 2002]. We compare the way that different measures are affected by the different features of languages described above. We first describe the averaging methods, and then the measures to consider.

Averaging

When considering data imbalance, it is important to examine different averaging techniques:

1. ***Micro*** **(μ)** calculates the metric globally on absolute predictions regardless of the city (class); i.e. all that matters is whether the prediction is correct or wrong. This is the default averaging technique used to calculate the overall accuracy of previous geolocation prediction models.

2. ***Weighted*** **(W)** calculates the metric for each class and finds the average, weighted by the frequency of each city in the training dataset.

3. ***Macro*** **(M)** calculates the metric for each city and takes an unweighted mean. It is the most appropriate for evaluating how classifiers behave on cities with a small number of users, rather than *micro* averaging, which is influenced by populous cities.

Measures

Although Precision (P) and recall (R), together with different averaging techniques, are the most common measures used in text categorization to evaluate the effectiveness of classifiers [Sebastiani 2002, Yang 1999], they were never considered in prior *user* geolocation prediction work before this study, and rarely used for *tweet* geolocation. Sometimes

Table 3.3: Results of geolocation prediction model in terms of precision, recall and F1-score at zero error tolerance and within 161 km.

		Prec	P@161	Recall	R@161	F1	F1@161
WORLD-ar	Weighted	0.430	0.558	0.530	0.661	0.459	0.584
	Micro	0.530	0.661	0.530	0.661	0.530	0.661
	Macro	0.153	0.221	0.149	0.209	0.143	0.200
WORLD-en	Weighted	0.152	0.436	0.114	0.233	0.064	0.212
	Micro	0.114	0.233	0.114	0.233	0.114	0.233
	Macro	0.047	0.234	0.016	0.108	0.014	0.122
ArQAT	Weighted	0.179	0.367	0.255	0.369	0.191	0.311
	Micro	0.255	0.369	0.255	0.369	0.255	0.369
	Macro	0.018	0.124	0.020	0.089	0.014	0.090

precision is favored (e.g. when journalists are looking for eyewitnesses within a specific city [Diakopoulos et al. 2012]); at other times *recall* (e.g. when journalists are looking for eyewitnesses on the ground and want to increase the search pool because eyewitnesses are rare in that case [Starbird et al. 2012]). Both scenarios focus on a single location.

Precision and recall using macro averaging are defined as follows, in comparison to the micro definitions in Section 2.4. Results of geolocation based on the three measures and averaging techniques are shown in Table 3.3.

$$
\begin{aligned}
P_M &= \frac{1}{\#locations} \sum_{i=0}^{\#locations-1} P(l = l_i, \hat{l} = l_i) \\
R_M &= \frac{1}{\#locations} \sum_{i=0}^{\#locations-1} R(l = l_i, \hat{l} = l_i) \\
F_M &= \frac{1}{\#locations} \sum_{i=0}^{\#locations-1} F_{\beta=1}(l = l_i, \hat{l} = l_i)
\end{aligned} \tag{3.1}
$$

We observe that there is a gap in performance between micro and weighted, and macro averages for the three datasets. Ignoring the class type (city) in micro averaging achieves the best results as expected, where the correct estimation of cities with large number of users normalize the negative impact of cities misclassified due to lack of enough training information. This explains the misleading high accuracy of ArQAT although it is highly imbalanced. Then comes the weighted average with slightly less performance because the correct prediction of cities with high frequency is rewarded by increasing their weights while cities with low frequency are penalised by decreasing their weights.

Table 3.4: Confusion matrix of top 5 countries, namely Saudi-Arabia (SA), Kuwait (KW), Egypt (EG), United Arab of Emirates (AE), and United States (US), in terms of #users in ArQAT testset where rows are the actual locations and columns are the estimated ones. The diagonal in bold is the true positives.

	SA	KW	EG	AE	US
SA	**6746**	317	33	0	0
KW	348	**897**	14	0	0
EG	168	37	**680**	0	0
AE	429	193	13	**28**	0
US	139	59	9	0	**0**

The interesting observation is the macro averaging which still measures performance per city, but calculates the average uniformly. Macro results show that the WORLD-en model is slightly better than ArQAT. As such, we find that the language so far has less influence on the complexity of the geolocation problem; i.e. users speaking languages other than English share the same difficulty in predicting their locations even if one language is more geographically global than the other.

We further analyzed the inter connection and flow of mis-predicted labels at the level of countries sharing the same language yet still different dialects. A confusion matrix in Table 3.4 of the top five countries shows that the top two are dominating and hindering the performance of the geolocation model. Saudi Arabia gains over 60% of the true negatives while Kuwait gains 30%. The high skewness of Emirates to Saudi Arabic and Kuwait, respectively, could be justified by the close culture and languages of the three countries. However, this can not be the case in the United States which is completely skewed towards other Arabic speaking countries; e.g. a Saudi user living in the US and still tweeting in Saudi dialect will be predicted in Saudi Arabia. This observation is even emphasized in the top three countries Saudi Arabia, Kuwait and Egypt with exchangeable false negatives. A naive assumption of considering all users who are misclassified at the level of country as expats, would leave us with a substantial subset of 20%. This phenomenon not only introduces a new factor which has a strong influence on geolocation prediction, but further questions the validity of the hypothesis that language identification has an influence on the performance of a multi-lingual geolocation prediction model [Han et al. 2014].

Table 3.5: Languages rank correlation τ_β between pairs of evaluation metrics.

	WORLD		TwArchive	
	Acc@161	Median	Acc@161	Median
Acc	0.00	-0.31	0.15	0.15
Acc@161	–	0.03	–	0.13

3.4 Multi-lingual Analysis

In this section, more languages are examined for a better understanding of the language effect. A geolocation model is trained per language in WORLD and TwArchive. Following machine learning common practices, a 90–10% split of the dataset into training and test sets is typically recommended when large amounts of data are available. However, not all languages in WORLD and TwArchive have enough users with at least ten geotagged tweets; to be eligible for testing. Table 3.1 shows the number of users eligible for testing (#TestUsers) per language in both datasets. We observe a large gap in the percentage of test users across languages (3–16% and 0.2–6 in WORLD and TwArchive respectively). To address this issue, a test split of 2% was used whenever applicable to satisfy the requirement of having ten geotagged tweets per test user and for a fair comparison between languages across data collections. The remaining users were included in the training set (98%). In cases where the number of eligible test users is more than 2%, such as in WORLD, 10 samples were generated for training and testing and then the average was calculated for more reliable results. If the number of eligible test users is less than or equal to 2%, such as the majority of languages in TwArchive, a single train and test split is employed.

Table 3.1 also shows the number of cities with eligible test users (#TestCities) per language in both datasets. WORLD tends to have more cities with eligible test users than TwArchive. We employ stratified sampling for languages with more than 2% of eligible test users to maintain the distribution of cities in the test split.

Before that, we measure the agreement of the most common metrics on how they rank the accuracy of the geolocation system across the users of each language, 13 geolocation models per dataset. Kendall's Tau-b was used to measure the correlation between the

Table 3.6: Influence of dataset size, in terms of the *slope* of a linear regression model, on the evaluation measures for six languages in TwArchive.

	en	**es**	**pt**	**fr**	**ar**	**tr**
Acc	0.02	0.06	0.07	0.04	0.03	0.01
Acc@161	0.04	0.07	0.09	0.04	0.04	0.02
Median	-7.34	-1.17	-1.26	-0.31	-0.86	-0.10

ranks, see Table 3.5. There is no statistically significant rank correlation between any pair: the measures appear to be examining different aspects of geolocation. All three measures are therefore considered for evaluation beside the proposed alternative measures, namely precision, recall and f1-score using micro, weighted and macro averaging techniques.

A range of features may influence geolocation accuracy. Although Han et al. [2014] speculated that distribution of users was the reason for accuracy variation, many other differences were present in the language datasets they studied: the sets were of notably different sizes, written in different languages, and each contained different numbers of users, tweets, and cities. Therefore, the features we explore are dataset size, a preliminary test of the impact of the language, and a range of individual features such as entropy and number of users.

3.4.1 Dataset Size

We focus on the six languages that have sufficient users eligible for testing (each has at least ten geotagged tweets as mentioned in §3.2): two of which the geolocation system has low accuracy (*en* and *fr*), two with moderate accuracy (*es* and *pt*), and two with high accuracy (*ar* and *tr*). From each of the language sets, we randomly sample subsets of users in decrements of 10%, from 100% down to 10%. Ten samples of each subset were created, and an average was taken. Figure 3.4 depicts the results per language with an increasing percentage of the dataset beside the fitted regression model. Note that the *Median* error graph (bottom right) is using a log scale because of the gap in error distance estimations between English and other languages. Table 3.6 shows that for *Acc*, there is a weak positive relationship between the number of users and accuracy. We chose a *slope*, over a correlation measure, because it estimates the expected gain in accuracy with the

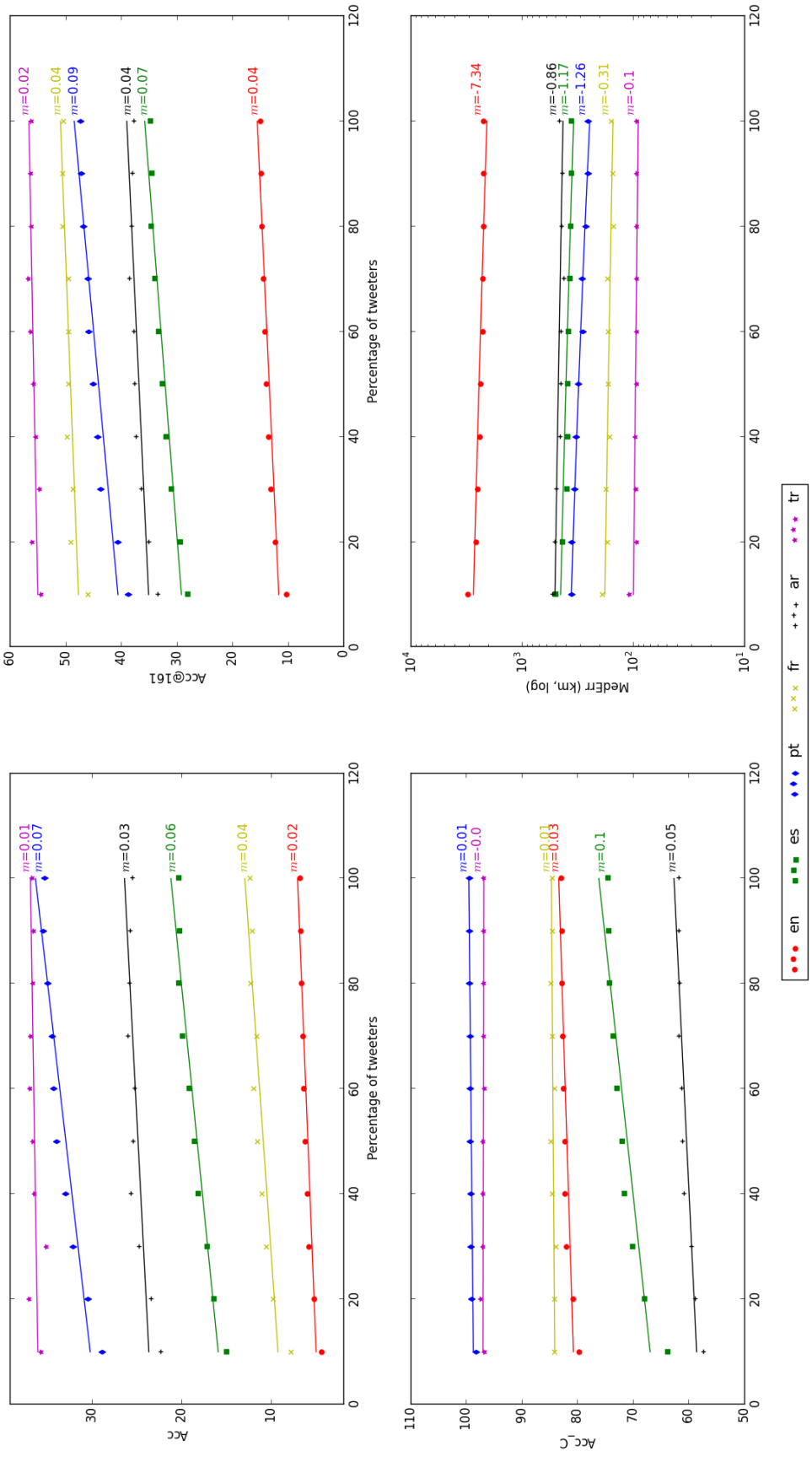


Figure 3.4: Influence of dataset size, in terms of the number of *users*, on the evaluation measures for six languages in TwArchive.

Table 3.7: Accuracy of geolocation for the 13 languages in WORLD and TwArchive.

		en	**es**	**pt**	**fr**	**ar**	**tr**	**id**	**it**	**nl**	**ru**	**ms**	**th**	**ko**
Acc	WORLD	0.11	0.29	0.31	0.13	0.49	0.54	0.4	0.15	0.25	0.33	0.41	0.43	0.45
	TwArchive	0.07	0.20	0.35	0.12	0.26	0.37	0.28	0.12	0.23	0.09	0.37	0.28	0.25

increase in dataset size. While there is some variation across languages, the gradient of the slope is consistently small. The same pattern was found with *Acc@161*, while for the *Median*, the measure tends to decrease (improve) as the number of users increases. The fact that the slope of the linear regression model is greater suggests that *Median* is more affected by the scale of the dataset than the accuracy measures. Hence, the *Median* is not an appropriate measure in the case of small datasets.

3.4.2 Preliminary Examination of Language

In past work, Han et al. [2014] noted that users writing in some languages were easier to geolocate than those writing in others. We speculated that there may be something inherent in the way that tweets are written in each of the languages that causes the differences in geolocation accuracy. Because we had access to two collections covering the same 13 languages, we examined the relative geolocation accuracy per language across the two collections, shown in Table 3.7. Although the two collections vary in the number of users, the previous result showed the impact of dataset scale was small. Therefore, if the language of tweets was impacting on accuracy, the relative accuracy across the two collections might be expected to be similar.

To determine the degree of agreement between the languages in the collections, we ranked the 13 languages by geolocation accuracy and calculated Kendall's Tau-b between the two rankings. We found a statistically significant but moderate correlation of 0.46. The relative geolocation accuracy for a language changed notably across the two collections. The low correlation strongly suggests that differences in geolocation accuracy across languages are influenced by a property other than the actual language of the tweets.

3.4.3 Correlation with Individual Features

In order to measure the impact of collection and user/tweet features on geolocation accuracy per language, we measured the Pearson Correlation Coefficient between feature values and the relative accuracy of languages. The features used were entropy of users distributed across all cities and a subset of cities, the total number of cities, the total number of users, the number of LIWs per language, and the number of tweets. Both collections were used. In addition to Pearson, the coefficient of determination (R^2) was used to measure the explanatory power of the model. The results are shown in Table 3.8.

As can be seen, entropy has the strongest correlation with all three evaluation measures. Entropy over only the cities that had eligible test users with 10+ geotagged tweets (entropy.test) was also calculated, and generally resulted in a higher correlation than entropy measured across all possible cities. For TwArchive, number of cities that had eligible test users correlated strongest with *Median*.

Considering the average number of tweets per eligible test user, if this number increases, accuracy should also increase, since users reveal more information about their location [Cheng et al. 2010]. The correlations with this feature appeared to contradict past work by being negative, however, they were not significant; note that the range of tweets per user here was substantially smaller than the range Cheng et al. [Cheng et al. 2010] examined. The number of location indicative words in a lexicon normalized by the number of tweets per language was also found not to correlate strongly with accuracy. The results shown earlier on the impact of dataset size (Table 3.8) can also be seen here, as the number of users and tweets per language correlate most strongly with *Median*, compared to the other evaluation measures.

Average distance measures were found to have a weak correlation with *Acc@161*. By measuring the average distance between neighboring cities, it was found to be in the range of 52–74km (significantly less than the arbitrarily chosen 161km). This means users mistakenly predicted in 2–3 cities away from their home location would still be considered correct. An appropriate choice of the threshold distance to measure accuracy should be strongly correlated to at least the average distance between neighbour cities covered in

Table 3.8: Pearson Correlation between features and evaluation metrics; ($*$ and † denote statistical significance with $p \leq 0.05$ and $p \leq 0.01$, respectively).

Feature	Acc				Acc@161				Median			
	WORLD		TwArchive		WORLD		TwArchive		WORLD		TwArchive	
	r	r^2	r	r^2	r	r^2	r	r^2	r	r^2	r	r^2
Entropy	**-0.87**†	**0.76**	-0.69†	0.47	-0.62*	0.38	-0.29	0.08	0.52	0.27	0.43	0.19
#Cities	-0.76†	0.57	-0.40	0.16	-0.57*	0.32	-0.26	0.07	0.54	0.30	0.57*	0.32
Entropy.test	-0.83†	0.69	-0.70†	0.49	**-0.85**†	**0.73**	**-0.79**†	**0.62**	**0.82**†	**0.68**	0.89†	0.79
#Cities.test	-0.55*	0.30	-0.51	0.26	-0.67*	0.45	-0.55*	0.30	0.81†	0.66	**0.93**†	**0.87**
Avg #tweets.test	-0.47	0.22	-0.51	0.26	-0.34	0.12	-0.10	0.01	0.34	0.12	0.12	0.01
#LIW words	0.40	0.16	0.37	0.14	–	–	–	–	–	–	–	–
#users	-0.57*	0.32	-0.39	0.15	-0.54	0.29	-0.46	0.21	0.76†	0.58	0.87†	0.76
#Tweets	-0.51	0.26	-0.38	0.15	-0.51	0.26	-0.47	0.22	0.76†	0.58	0.87†	0.75
Avg dist	–	–	–	–	0.12	0.01	0.51	0.26	-0.33	0.11	-0.30	0.09
Nbr avg dist	–	–	–	–	-0.46	0.21	-0.22	0.05	0.55*	0.31	0.53	0.28

the dataset.

In summary, the correlation with different features showed that the distribution of users has a greater impact on the accuracy of geolocation prediction than other features, especially geographical coverage. This is a different result described in previous research. It also shows that *Acc@161* is not an appropriate measure.

3.4.4 Results Considering the Alternative Measures

The results in the previous section showed that the distribution of users across cities (entropy) is a strong predictor of the accuracy of geolocation for different languages. However, the measures *Acc* and *Acc@161* are both heavily influenced by the accuracy of the geolocation system on a limited number of cities as indicated in section 3.2. We, therefore, employ the alternative measures introduced in the pilot study. First, we define an extra geolocation baseline model. Second, we discuss the results.

Baselines

Yang [Yang 1999] pointed out that in the case of a very low average training instances per category (which applies here) the *majority class trivial classifier* tends to outperform all non-trivial classifiers. We therefore start by comparing our geolocation system against the Majority Class (MC) baseline.

Table 3.9: Comparison between Majority Class (MC) and Multinomial Naïve Bayes (MNB) models, in terms of *micro* precision (P_μ) and *macro* precision (P_M), for the top 13 languages in WORLD.

	en	**es**	**pt**	**fr**	**ar**	**tr**	**id**	**it**	**nl**	**ru**	**ms**	**th**	**ko**
MC P_μ	0.02	0.12	0.23	0.10	0.39	0.54	0.27	0.09	0.16	0.34	0.25	0.32	0.45
MNB P_μ	0.11	0.29	0.31	0.13	0.49	0.54	0.40	0.15	0.25	0.33	0.41	0.43	0.45
MC P_M	0.000	0.000	0.001	0.000	0.004	0.007	0.002	0.000	0.003	0.003	0.002	0.008	0.006
MNB P_M	0.047	0.027	0.036	0.033	0.059	0.027	0.079	0.018	0.077	0.006	0.086	0.267	0.046

Results

The first row of Table 3.9 shows that P_μ of MC for languages with the majority of users originating from one city tend to match or outperform the MNB classifier, i.e. *tr*, *ru* and *ko*, in the WORLD data collection. For instance, a MC model for users posting in Russian would fail to predict the location of any user outside Moscow, although 70% of the users are located in other cities (inside and outside Russia). The same pattern applies to TwArchive with one more biased language, than WORLD: Thai (th).

To evaluate classifiers at the level of each city, rather than overall performance, we compare precision based on *macro* averaging in the last two rows of Table 3.9. In contrast to P_μ, P_M shows that MNB classifiers outperform the MC for all languages.

The results of the MC classifiers for languages like *tr*, *ru* and *ko* at the high end of the range of P_μ reflect the data imbalance for such languages, where the data is biased towards a single city as shown in Figure 3.2. However, the results of the MC classifiers for other languages like *en*, *fr* and *it* at the low end of the range don't reflect the influence of data imbalance on the quality of geolocation, with other languages in between. To address this problem, we compare P_μ, to P_M, which shows an expected drop in performance in Table 3.9. In the case of *ru*, an MNB geolocation model would have a high accuracy of 33%, while having a poor average precision at the level of each city (0.6%). This contrast between *micro-macro* indicates the measures evaluate geolocation from different perspectives.

Correlation with Individual Features

Entropy was shown to have the highest correlation with *Acc* compared to other features. Here, we measure the correlation between the proposed alternative measures, using differ-

Table 3.10: Correlation between features and precision using different averages; (* and †) denote statistical significance with $p \leq 0.05$ and $p \leq 0.01$, respectively.

Feature	Micro				Weighted				Macro			
	WORLD		TwArchive		WORLD		TwArchive		WORLD		TwArchive	
	r	r^2	r	r^2	r	r^2	r	r^2	r	r^2	r	r^2
Entropy	**-0.87**†	**0.75**	-0.69†	0.47	-0.79†	0.62	-0.78†	0.61	-0.49	0.24	-0.63*	0.40
#Cities	-0.76†	0.58	-0.40	0.16	-0.64*	0.41	-0.42	0.18	-0.46	0.21	-0.43	0.18
Entropy.test	-0.82†	0.67	-0.70†	0.49	-0.74†	0.54	-0.52	0.27	-0.34	0.12	-0.49	0.24
#Cities.test	-0.54	0.29	-0.51	0.26	-0.44	0.19	-0.32	0.10	-0.24	0.06	-0.36	0.13
#users	-0.56*	0.32	-0.39	0.15	-0.36	0.13	-0.21	0.05	-0.14	0.02	-0.27	0.07
#Tweets	-0.50	0.25	-0.38	0.15	-0.30	0.09	-0.20	0.04	-0.11	0.01	-0.29	0.09

Table 3.11: Languages rank correlation τ_β for micro (μ), weighted (W), and macro (M) averaging; (* and †) denote statistical significance with $p \leq 0.05$ and $p \leq 0.01$, respectively.

(a) Across averaging techniques

	Precision				Recall			
	WORLD		TwArchive		WORLD		TwArchive	
	W	M	W	M	W	M	W	M
μ	0.41†	-0.08	0.38	0.08	1.00†	0.08	1.00†	0.15
M	0.00	–	0.08	–	0.05	–	0.15	–

(b) Across data collections

Precision			Recall		
μ	W	M	μ	W	M
0.46*	0.13	0.00	0.46*	0.49*	0.03

ent averaging techniques, and the same set of features, excluding the poor ones. Correlations for the two data collections (WORLD and TwArchive) are displayed in Table 3.10. The *micro* columns are analogous to accuracy reported earlier in Table 3.9.

In contrast to *Acc* and P_μ, entropy is not as strong an indicator of how well a geolocation model performs on the *macro* level. The moderate insignificant correlation between entropy and P_M aligns with the fact that *macro*-averaging should be independent of the distribution of users across cities, i.e. all cities are treated uniformly. *Macro*-averaging generally has the lowest correlation with the different features. The same pattern applies to recall.

From a language perspective, we observed that the ranking of languages differs from one averaging technique to another and also from precision to recall. For instance, comparing *micro* to *macro* precision, *th* remained among the top ranks while *tr* dropped to the bottom behind *en*. To measure the degree of agreement, we measured the τ_β correlations for all direct combinations of data collection, precision, recall, *micro*, *weighted* and *macro*, see Table 3.11.

For *precision*, the *micro* and *weighted* averages have a statistically significant, but *moderate* rank correlation in WORLD. In contrast, the *micro* and *weighted* averages for

recall coincide, in both data collections. *Micro* and *macro* averages did not have a significant rank correlation. Finally, at the level of data collections, *micro* (precision and recall), and *weighted* recall have a statistically significant, albeit moderate, rank correlation.

The difference in precision between *micro* and *macro* averaging suggests that all languages are affected by the data imbalance. *Micro* averaging is biased towards big cities, while *macro* averaging assumes that all cities contribute equally to the metric. Some languages are still easier than others, but not because they are the only languages biased towards a small set of cities, and/or their usage is geographically limited to a specific region. All languages have a bias towards a small number of big cities; the difference between languages like *en* and *fr* compared to *ru* and *tr* is the number of big cities. For instance, the top 10 cities for *en* and *fr* in WORLD have a comparable number of users (1–4%) of the total number, while the top city in *ru* and *tr* has more than 30% of users and the second city drops down to less than 10% of users.

In the end, the choice of which averaging technique to use in taking decisions depends on the application. However in the general case, we recommend using the *weighted* average instead of *micro* because it limits the dominance of big cities while maintaining their importance. At the same time, it reduces the potentially misleading evaluation when comparing languages.

3.5 Summary

We studied the features that might influence the accuracy of a system that geolocates Twitter users. Examining two large tweet collections covering thirteen languages, we found substantial variation in accuracy across languages, a result that has been observed before but not studied or explained.

We showed that the distribution of users over cities is strongly correlated to accuracy. Past work suggested that a lack of geographical coverage of certain languages may also be a factor, however, all the languages we studied were found to have a global coverage.

The results presented in this chapter can be used for future test set design. The scale of a test set was found to have little influence on accuracy. However, the distribution

of users was a strong influence. Although a geolocation system could potentially ground users to one of few thousand cities, the skewed distribution present in the test sets meant that accuracy was influenced by only a few tens of cities. Current testing approaches are not as geographically broad-ranging as one might imagine or expect. A consequence of the current testing regime is that a simplistic baseline, which grounds to one city per language, was measured to be as accurate as a state of the art system for more than one language.

To overcome such dataset limitations, we proposed using *macro* averaging. The contrast between it and *micro* averaging revealed that data imbalance affects all languages, even one that is extensively used, such as English. Our analysis demonstrated that reporting both *micro* and *macro* averaging, or using a *weighted* average, provides valuable additional insight.

The proposed micro-macro evaluation is driven by the machine learning perspective of evaluating imbalanced classification tasks. In Chapter 4, we assess the need for micro-macro evaluation from the viewpoint of applications.

CHAPTER 4

Urban and Rural Evaluation

Existing research highlighted the concerns about population bias in social media and the importance of accounting for such biases while using social media to understand human behavior or developing and evaluating social media-based algorithms. The wide range of applications built on top of Twitter geolocation systems require evaluation from several perspectives. In the previous chapter, we proposed to contrast micro and macro evaluation, from a theoretical perspective, to address the issue of the skewed data distribution towards urban regions while investigating language influences on Twitter user geolocation. In this chapter, we review micro (urban) and macro (rural) evaluation in the context of journalism due to the availability of a news dataset that can represent its information needs. In Section 4.1, we introduce the notion of urban and rural bias in social media and the different approaches for gathering the information needs of geolocation applications. Data is described in Section 4.2. The process to assess the need for urban and rural evaluation is presented in Section 4.3, and the results are discussed in Section 4.4.

4.1 Overview

The effectiveness of any evaluation metric depends on the information needs of the application. While Twitter geolocation serves a wide range of applications, we explore the importance of urban and rural evaluation in the context of journalism. The urban-rural bias in social media has attracted a growing amount of attention in the literature, with

focus on the U.S. due to the significant size of the rural population [Hecht and Stephens 2014, Johnson et al. 2017]. To categorize locations along the urban-rural spectrum, they utilized the U.S. National Center for Health Statistics' Urban-Rural Classification Scheme for Counties [Ingram and Franco 2012], which assigns each county a code on a scale from 1 (most urban) to 6 (most rural). These codes represent large-metropolitan counties with population of 1 million or more to non-metropolitan counties with at least 10,000 population.

Information needs can be gathered in two ways. A quantitative approach can anticipate the information needs from an external data source such as social media. For instance, the distribution of news and Twitter users over the same set of locations (e.g. cities) would give an indication of the importance of urban and rural areas. If the majority of news is coming from urban cities, then micro evaluation, the current most popular method, should be considered. Macro evaluation is recommended otherwise, to reduce the influence of urban cities in dominating the evaluation process. A qualitative approach, on the other hand, will require conducting surveys with the users of the applications, journalists in this case, to gather the information needs. A geolocation system, then, is developed in an iterative manner. While the latter approach is more effective in formalising the information needs, it is a time-consuming process. We, therefore, choose the former approach to anticipate the information needs of news media applications, in the context of urban and rural evaluation, based on a global catalog of worldwide events that appeared in the news media.

4.2 Data

To anticipate the information needs of geolocation applications in the context of journalism, we need two datasets: a dataset that is used for developing Twitter geolocation models, which is TwArchive (see Section 3.1), and a new dataset for events that appeared in the news media, which is called GDELT.

GDELT[1] is a global event database supported by Google which monitors the broad-

[1] https://www.gdeltproject.org/data.html

cast, print, and web news from around the world in over 100 languages, and identifies people, locations, organizations and other information. For this work, we consider GDELT 1.0 Event Database, which contains over a quarter billion records of over 364 million distinct events from almost every corner of the earth since January 1979. GDELT 1.0 updates daily and records are stored by the date the event was found in the world's news media rather than the date it occurred. To align with our social media analysis in §3.1, we consider events from February to December, 2014. Given the massive size of GDELT, the only way to query the dataset is through Google BigQuery.

Each record in GDELT 1.0 contains 58 fields.[2] However, we use only 4 fields, which contain the information required for this study. The fields considered here are defined as follows:

- **GlobalEventID** is a global unique identifier assigned to each event. An event can be referenced across multiple articles in the dataset.
- **Action latitude** and **Action longitude** are the GPS coordinates of the event. According to the CAMEO taxonomy[3], each event has two actors involved and an action (what actor 1 did to actor 2). These fields capture the location information closest to the point in the event description that contains the actual statement of action. Events with no GPS fields are ignored. This field is going to be used later to identify the language of the article.
- **Source URL** is the URL of the news article the event was found in. If an event was mentioned in multiple articles, only one of the URLs is provided.

GDELT provides an administrative location for each event (e.g. city). However, the geocoding system used in §5.1.3, to identify the home location of Twitter users in the TwArchive dataset, is a customized geocoding library which merges Geonames neighbouring cities of less than 50,000 population [Han et al. 2014]. This means the sets of locations identified in GDELT and TwArchive might be discordant. For a fair comparison

[2]http://data.gdeltproject.org/documentation/GDELT-Data_Format_Codebook.pdf
[3]https://www.gdeltproject.org/data/documentation/CAMEO.Manual.1.1b3.pdf

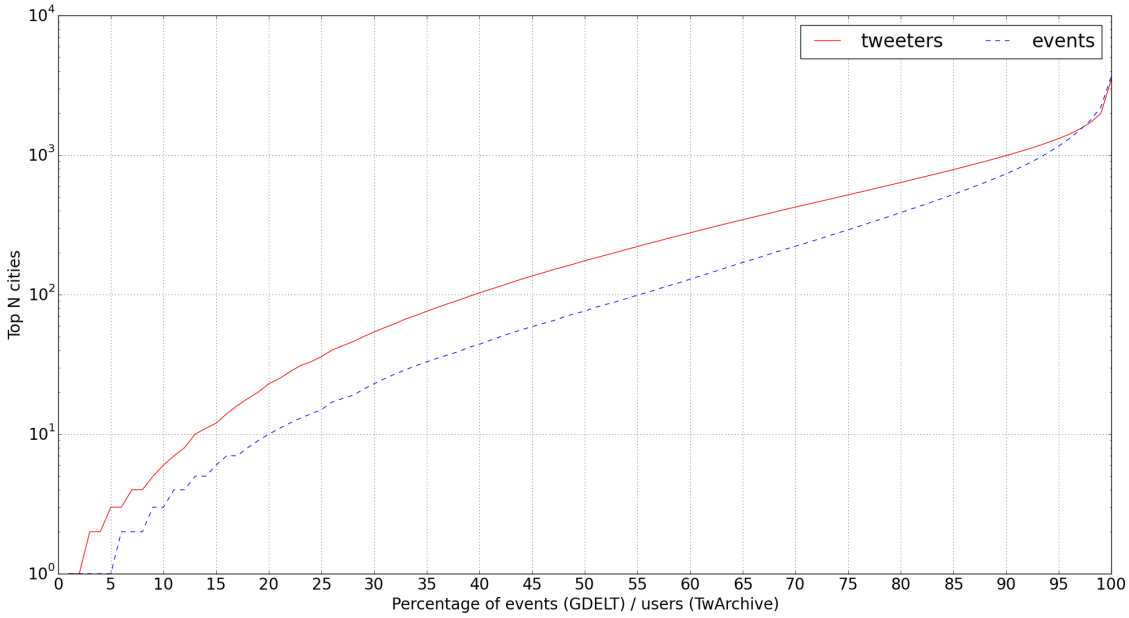


Figure 4.1: Distribution of GDELT events and TwArchive users over 3,600 cities.

between GDELT and TwArchive, we reverse-geocode the GPS points associated with an event using the same geocoding system.

Social media datasets are known to have a substantial population bias towards urban regions. In order to be able to anticipate the information needs via the events database, GDELT has to be checked for data imbalance. Figure 4.1 shows the distribution of GDELT events and TwArchive users over a set of 3,686 cities. We observe that both datasets are skewed towards a small number of cities. GDELT (blue dotted line) is more skewed than TwArchive (red solid line). The top 100 cities (y-axis) cover 55% of events and 40% of users.

4.3 Urban and Rural Evaluation

In this section, we introduce the methodology of anticipating the geospatial needs of journalistic applications, while focusing on urban and rural evaluation at the level of cities. In Section 4.3.1, we quantify the degree of imbalance in our datasets and estimate how important rural cities are, in comparison to urban cities. Not all cities have sufficient information (events or users) to be considered for analysis. In Section 4.3.2, we identify the minimum amount of information required for a city with rare events or users to be

considered for evaluation. Datasets for all languages have a substantial bias towards urban cities (see 3.4.4). To anticipate the needs per language, the event language is identified based on URLs only in Section 4.3.3.

4.3.1 Similarity of Data Imbalance

We assess the importance of urban and rural evaluation based on the similarity between the distributions of events (GDELT) and Twitter users (TwArchive) over the same set of cities. Hence, two similarity measures are employed to quantify the difference in skewness between both datasets, and explore the importance of urban and rural cities based on their ranking in each dataset.

Jensen–Shannon Divergence (JSD) is employed to measure the similarity between the probability distributions of events and users over cities. JSD is based on Kullback-Leibler Divergence (KLD), with some notable difference that it is symmetric and always has a finite value. Let X be a discrete random variable (city) and let p_1 and p_2 be two probability distributions of X. KLD and JSD are defined as follows [Lin 1991]:

$$\begin{aligned} KLD(p_1, p_2) &= \sum_{x \in X} p_1(x) log \frac{p_1(x)}{p_2(x)} \\ JSD(p_1, p_2) &= KLD(p_1, p_2) + KLD(p_2, p_1) \\ &= \sum_{x \in X} (p_1(x) - p_2(x)) log \frac{p_1(x)}{p_2(x)} \end{aligned} \tag{4.1}$$

where p_1 and p_2 are the probability distributions of *aligned* cities in TwArchive and GDELT. A value of zero means the two probability distributions are dissimilar, while a value of one means they are identical (a city has the same probability in both datasets).

Kendall's Tau-b rank correlation is calculated to measure the agreement between the rankings of cities in both datasets based on the number of events or users in a descending order. A value of -1 means perfect inversion, a value of zero indicates no correlation, and a value of 1 means perfect agreement (cities have the same degree of importance in both datasets).

4.3.2 Power-Law Distributions in Twitter and Events Data

Clauset et al. [2009] demonstrated that the population of cities in census data follows power-law distributions. Hence, we conjecture that the populations of cities in social media and events data follow a power-law distribution. A quantity x (city population in this case) obeys a power-law if it is drawn from probability distribution

$$p(x) \quad \alpha \quad x^{-\alpha}, \tag{4.2}$$

where α is a constant parameter of the distribution. Generally, the power-law behavior cannot hold for all values of x. The power-law applies only for values greater than x_{min}. In such cases, the tail of the distribution — the part of the distribution representing large but rare events — follows a power-law. Clauset et al. proposed the following steps for analyzing power-law data in a principled statistical manner:

1. Fit a power-law distribution to a given dataset and estimate the parameters α and x_{min} using the method of maximum likelihood.

2. Test whether the power-law is a plausible fit to the data. Calculate the Kolmogorov-Smirnov (KS) statistic to test the goodness-of-fit, which generates a p-value that quantifies the plausibility of the hypothesis. The power-law hypothesis is accepted only if the resulting p-value is greater than 0.01, based on the python package[4] used to conduct this analysis [Alstott et al. 2014].

3. Compare the power-law with alternative hypotheses via a likelihood ratio test, which generates a positive or negative ratio (R) depending on which model is better, or zero in the event of a tie. The likelihood ratio test gives a p-value that indicates whether the observed sign of R is statistically significant. The sign is a reliable indicator of which model is the better fit to the data if the p-value is less than 0.1. For each alternative, if the calculated R is significantly different from zero, then a positive sign indicates that the alternative is favored over the power-law.

[4] https://pypi.org/project/powerlaw/

In the context of urban and rural evaluation, the power-law analysis helps to better understand the tail of the distribution for TwArchive (users) and GDELT (events) datasets. The parameter x_{min}, in particular, determines the range over which power-law behaviour holds. The lower bound for population (users or events) in rural cities is represented by x_{min}; i.e. cities with x_{min} users or events are considered rural yet collectively have enough population to influence any analysis conducted on this data. Any city with population x_{min} or more should be considered in the urban and rural evaluation study.

4.3.3 Event Language Identification Based on URLs

While GDELT is a multi lingual dataset, the language of each event/article is not available as a field for events before February 19^{th}, 2015. Although the source URL of the event is available, the traditional approach of identifying the language of 6.5 million URLs based on their content is time consuming. Baykan et al. [2008] proposed an efficient approach that allows to obtain this information purely based on a URL without having to download the page. This approach is useful for crawlers of web search engines and scalable construction of high-quality web corpora [Biemann et al. 2013]. Hence, we employ this approach to identify the language of events in GDELT. For the rest of the section, we describe the dataset, features and classification algorithms.

Data. We use the Open Directory Project (ODP)[5] dataset which provides more than 3 million URLs for 81 languages. Originally, Baykan et al. considered five languages, namely English, German, French, Italian and Spanish. Given that TwArchive contained thirteen languages, we consider nine more languages, namely Arabic, Dutch, Indonesian, Korean, Malaysian, Portuguese, Russian, Thai, and Turkish. However, ODP has no URLs for Indonesian and Malaysian. For all datasets, we removed URLs belonging to multiple languages. Table 4.1 gives details about the total sizes (second column) of the remaining twelve languages. The full set we downloaded includes a total of 2.8 million URLs, of which 1.7 million are English and only 2 thousand are Korean. Given the number of magnitudes that the sizes differ, we exclude the smallest four datasets, namely Portuguese, Thai,

[5]ODP has been deactivated since March, 2017. An archive is available here: https://dmoz-odp.org/

Table 4.1: The Open Directory Project dataset.

Language	Total Size	Train-Full	Test-Full	Train-Balanced	Test-Balanced
English	1,738,131	1,330,978	443,020	39,605	13,075
German	410,568	315,596	105,559	39,462	13,218
French	183,912	146,689	49,123	39,380	13,300
Italian	126,479	101,212	33,769	39,591	13,089
Spanish	105,383	83,932	28,148	39,344	13,336
Russian	104,646	79,956	26,633	39,616	13,064
Dutch	69,459	54,315	18,073	39,478	13,202
Turkish	52,021	39,369	13,311	39,448	13,232
Portuguese	11,223	–	–	–	–
Thai	6,081	–	–	–	–
Arabic	4,052	–	–	–	–
Korean	2,163	–	–	–	–

Arabic and Korean. Note that German is not part of the TwArchive dataset, but we included it in training for performance reasons as we will discuss later.

Features. Baykan et al. extracted features from URLs using three different methods: unigrams, trigrams, and custom-made features. Here we employ their best reported method, which is unigrams. Each URL is split into tokens at any punctuation mark, number or non-letter character. Tokens of length less than two and stop words, namely, 'www', 'index', 'html', 'htm', 'http', and 'https' are then removed.

Classification Algorithms. Baykan et al. experimented with four algorithms, namely Multinomial Naïve Bayes (MNB), Decision Trees, Relative Entropy and Maximum Entropy. Here we employ their best reported algorithm, which is MNB. For most settings, they trained one-vs-the-rest classifiers (i.e. a binary classifier per language) instead of one multi-way classifier. Given the data imbalance in the distribution of URLs over language in the ODP dataset, we apply two modifications to the original work: i. use Linear Support Vector Classification (Linear SVC) algorithm; ii. apply two settings of training. Balanced training has the same number of URLs for each language. Full training uses the full data for each language. A fixed percentage is sampled for testing (26%). Sizes of the different settings are shown in Table 4.1, columns 3–6.

4.4 Results

In this section, we present the results of urban and rural evaluation at the level of multilingual collections of GDELT and TwArchive in §4.4.1, and per-language in §4.4.2.

4.4.1 Unified Multilingual Results

We begin by analyzing the similarity of data imbalance, and then power-law distributions.

Similarity of Data Imbalance

Jensen–Shannon divergence between GDELT and TwArchive was found to be of value 0.47, which indicates a moderate similarity in terms of skewness. A rank correlation of -0.014 indicates no association, albeit statistically insignificant (p-$value$=0.23). This means the ranking of cities in both datasets based on population is different. The overlap between the two lists of the top 100 cities is 25, with a rank correlation of value 0.03 (no correlation), albeit statistically insignificant (p-$value$=0.65).

Manually inspecting the top 100 cities in both datasets, the majority are urban cities (especially in TwArchive) with few rural cities (mainly in GDELT). An example for a rural city in the top 100 cities with events in GDELT is Simferopol, a Ukrainian city on the border with Russia. In 2014, more than 70 thousand events or articles originated from that city. However, it comes down the list of ranked cities with users in TwArchive at position 1,737 with only a few hundred users. While Ukranian is not one of the languages considered for geolocation in Section 2.3.3, Russian is the closest language. Looking into the performance of the Russian geolocation model at city level, the accuracy of geolocating Twitter users in Simferopol is zero. This means that journalists will find it hard to locate any Twitter users in this hot region using geolocation systems.

Table 4.2 shows another example, but focusing on the rank mapping between the top 10 cities in GDELT and TwArchive, which are mainly capital (urban) cities. All cities come at a relatively low rank in TwArchive with a much smaller number of Twitter users in comparison to the number of events. For instance, Washington D.C., US is the top city in GDELT with over a million events (ranked at 1), but ranked as 165 in TwArchive with

Table 4.2: Top 10 cities in GDELT (events) and their rank-map in TwArchive (users).

City	GDELT	TwArchive
Washington, D.C., US	1	165
Moscow, RU	2	65
Gaza, PS	3	1,252
City of London, GB	4	11
Jerusalem, IL	5	1,559
Kiev, UA	6	303
Beijing, CN	7	1,382
Delhi, IN	8	505
Tehran, IR	9	1,610
Cairo, EG	10	225

only 6,913 users. The precision and recall of geolocating users living in Washington D.C. and tweeting in English are, therefore, very low, at 8% and 3%, respectively. Geolocation of users is even worse for some capital cities such as Gaza, Palestine and Kiev, Ukraine, which were mistakenly assigned by the corresponding geolocation system to capitals of neighbouring countries. By checking the confusion matrices for the geolocation models of users tweeting in Arabic and Russian, all tweeters in Gaza were geolocated in Amman, Jordan, and these in Kiev were geolocated in Moscow, Russia. City of London, GB is the only urban city in Table 4.2 where its importance is almost aligned between GDELT (ranked at 4) and TwArchive (ranked at 11). The recall of geolocating users living in the City of London and tweeting in English is at 80%, better than for Washington D.C..

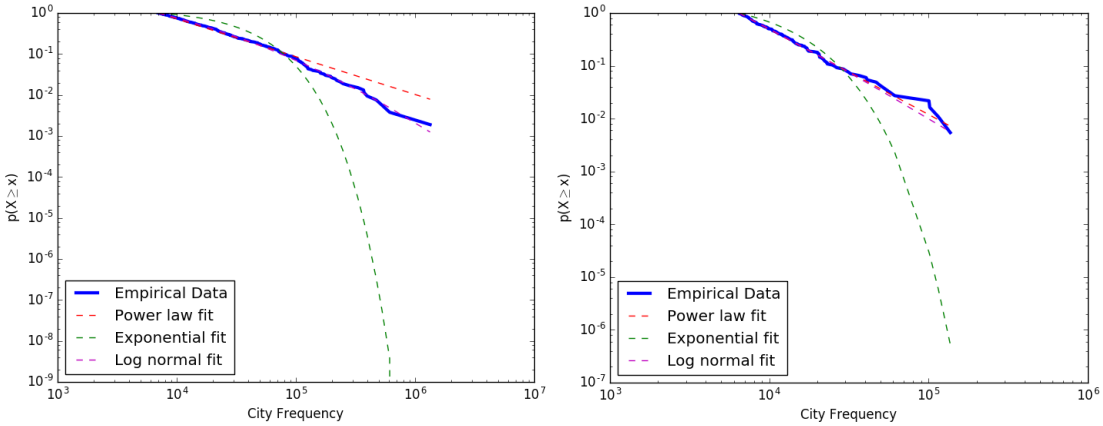
Power-law Analysis

Table 4.3 shows results from the fitting of a power-law distribution to each of our datasets using the methods described in Section 4.3.2. The reported x_{min} (second column) values for both datasets indicate that a city needs to have $\geq$ 6,000 users or events, which maps to $\approx$ 500 cities out of more than 3,000 cities in each of the datasets. The p-values (third column) indicate that both GDELT and TwArchive are consistent with a power-law distribution.

Table 4.3 shows the results of likelihood ratio test comparing the best-fit power-laws for each of our datasets to the alternative distributions, log-normal and exponential. We can rule out the exponential distribution as a possible fit in both datasets. The results

Table 4.3: Tests of power-law behavior in GDELT and TwArchive. For each dataset, we give x_{min} and p-value for the fit to the power-law model and likelihood ratios for the alternatives. We also quote p-values for the significance of each of the likelihood tests. Statistically significant p-values are denoted in **bold**. Positive values of the log-likelihood ratios indicate that the power-law model is favored over the alternatives.

	Power-law		Log-normal		Exponential	
Dataset	x_{min}	p	LR	p	LR	p
GDELT	6,983	0.05	-2.45	**0.01**	4.06	**4.8E-05**
TwArchive	6,442	0.03	-0.27	0.79	2.95	**0.003**



(a) GDELT (b) TwArchive

Figure 4.2: Complementary cumulative distribution function of city frequency for GDELT and TwArchive.

for log-normal are more ambiguous. For GDELT, log-normal is favored over power-law. On the other hand, log-normal is favoured over power-law for TwArchive, but the accompanying p-value is large enough that the result cannot be trusted. Figure 4.2 depicts the estimated distributions. The fitted models represent the Complementary Cumulative Distribution Function (CCDF) of city frequencies based on the number of events in GDELT (Figure 4.2a) and users in TwArchive (Figure 4.2b).

4.4.2 Per-Language Results

Here we introduce the results of identifying the language of events in GDELT, and then present the results of power-law analysis.

Table 4.4: Web page Language Identification based on URLs from the Open Directory Project dataset.

	German			English			Spanish			French		
Classifier	P	R	F1	P	R	F1	P	R	F1	P	R	F1
MNB-B	0.25	**0.99**	0.40	**0.95**	0.28	0.43	**0.99**	0.44	0.61	**0.97**	0.56	0.71
MNB-F	0.65	0.92	0.76	0.87	**0.96**	**0.91**	0.96	0.28	0.43	0.95	0.51	0.66
LSVC-B	0.28	0.98	0.43	0.86	0.39	0.54	0.95	**0.50**	**0.65**	0.94	**0.59**	**0.73**
LSVC-F	**0.84**	0.90	**0.87**	0.87	**0.96**	**0.91**	0.90	0.47	0.62	0.90	**0.59**	0.71
	Italian			Dutch			Russian			Turkish		
Classifier	P	R	F1	P	R	F1	P	R	F1	P	R	F1
MNB-B	**0.97**	0.69	0.81	**0.97**	0.85	0.90	**1.00**	0.84	0.91	0.99	0.62	0.76
MNB-F	0.90	0.63	0.74	0.93	0.63	0.75	0.98	0.68	0.81	0.99	0.44	0.61
LSVC-B	0.96	**0.72**	**0.82**	0.96	**0.87**	**0.91**	0.99	**0.87**	**0.92**	**1.00**	**0.63**	**0.77**
LSVC-F	0.88	0.69	0.77	0.86	0.83	0.85	0.97	0.85	0.91	0.97	0.62	0.76

Event Language Identification Based on URLs

Table 4.4 show results for the language identification models of events described in Section 4.3.3, namely MNB trained on a balanced sample of ODP per language (MNB-B), MNB trained on the full dataset (MNB-F), LSVC-B, and LSVC-F. MNB-B achieves the best precision for all languages, except for German and Turkish, although the latter is competitive with other models. MNB-F mainly enhanced the precision of German and recall of English at the cost of recall for other languages. LSVC-B, in comparison to MNB-B, enhanced the recall for all languages while preserving competitive precision. LSVC-F has the same influence as MNB-F. In summary, MNB-B achieves the best precision and LSVC-B achieves the best recall and f1-scores.

To identify the language of events in GDELT, we favor precision (MNB-B) over recall (LSVC-B) because it contains enough data where missing out on some events will not affect the analysis. Looking at the confusion matrix of MNB-B in Table 4.5, we observer that vast majority of misclassified URLs in the trained model are assigned to German. Given that German is not included into the power-law analysis, it will have no effect on the results.

Table 4.5: Confusion matrix of Web page Language Identification using Multinomial Naïve Bayes classifier and a balanced ODP dataset. A row is the actual language and a column is the predicted language. The diagonal in bold is the true positives.

	German	English	Spanish	French	Italian	Dutch	Russian	Turkish
German	**13,084**	19	10	25	50	11	4	1
English	9,170	**3,596**	31	84	56	54	20	14
Spanish	7,239	32	**5,912**	35	31	7	6	30
French	5,376	38	17	**7,446**	76	320	7	2
Italian	4,031	11	7	14	**9,021**	3	0	0
Dutch	1,908	13	3	34	6	**11,233**	0	2
Russian	2,004	24	3	4	4	2	**11,006**	3
Turkish	5,009	7	4	3	1	4	5	**8,199**

Power-Law Analysis

Table 4.6 shows the results of likelihood ratio test comparing the best-fit power-laws for each language in our datasets to the alternative distributions. Results are consistent with the multilingual power-law analysis. Exponential distribution is ruled out as a possible fit for all languages in both datasets. The results for log-normal are more ambiguous. Log-normal is favored over power-law for French and Turkish in GDELT, and Russian in TwArchive. On the other hand, log-normal is favoured over power-law (or vice-versa) for the remaining cases, but the accompanying p-value is large enough that the result cannot be trusted.

We observe a large variance in x_{min} values across languages in both datasets, which influences the number of cities that the fitted model covers. However, the majority of models cover hundreds of cities, not only a few number of urban cities. Hence, we consider evaluation at the level of both urban and rural regions.

4.5 Summary

In this chapter, we assessed the need for urban and rural evaluation of Twitter geolocation systems from the viewpoint of applications. We used a global database of worldwide events (GDELT) as an indicator of the importance of rural cities. Our geotagged social media dataset (TwArchive) and GDELT were found to be similarly skewed towards a small set

Table 4.6: Tests of power-law behavior for each language in GDELT and TwArchive. For each dataset, we give x_{min} and p-value for the fit to the power-law model. We also quote p-values for the significance of each of the likelihood tests. Statistically significant p-values are denoted in **bold**. Positive values of the log-likelihood ratios indicate that the power-law model is favored over the alternatives.

		Power-law			Log-normal		Exponential	
Dataset	Language	x_{min}	p	#cities	LR	p	LR	p
GDELT	English	12,958	0.04	314	-1.57	0.12	2.98	**0.00**
	Spanish	23	0.04	259	-0.99	0.32	3.34	**0.00**
	French	28	0.03	714	-1.66	**0.10**	5.31	**0.00**
	Italian	31	0.04	88	+0.05	0.96	2.07	**0.04**
	Dutch	12	0.04	190	-0.40	0.69	4.00	**0.00**
	Russian	19	0.03	701	-0.73	0.46	4.17	**0.00**
	Turkish	4	0.06	538	-1.93	**0.05**	5.90	**0.00**
TwArchive	English	4,063	0.05	68	+0.17	0.87	2.01	**0.04**
	Spanish	765	0.05	154	-1.05	0.29	2.53	**0.01**
	French	31	0.02	593	-0.82	0.41	7.76	**0.00**
	Italian	214	0.03	120	+0.06	0.95	2.75	**0.01**
	Dutch	38	0.02	315	-0.03	0.97	5.14	**0.00**
	Russian	72	0.07	163	-1.68	**0.09**	2.18	**0.03**
	Turkish	3	0.04	866	+0.40	0.69	9.25	**0.00**

of cities. On other hand, there is a disagreement between estimating the importance of each city based on the population of events and Twitter users. GDELT tends to have a mix of urban and rural cities at the top of the list, while TwArchive is skewed towards urban cities. The distribution of events and users over cities in both datasets were found to follow a power-law distribution, even when datasets were broken down by language. The estimated lower bound of power-law indicates that hundreds and thousands of cities can be fitted within the model. Given that these cities are spread over the spectrum of urban and rural cities, we employ urban and rural evaluation metrics in the following chapters.

Having focused on language influence using a single geolocation method Chapter 3, we will consider evaluating other geolocation inference techniques in Chapter 5, making use of a wide range of open source frameworks. Considering the data imbalance problem, we focus on developing more robust evaluation framework, examining the different location representations and statistical analysis of the differences in performance between geolocation techniques.

CHAPTER 5

Effective Evaluation of Twitter Geolocation

In Chapter 3, we demonstrated that evaluation metrics play an important role in understanding the influence of language, among other features, on the task of Twitter user geolocation. However, the proposed evaluation metrics—mainly to address the issue of the skewed data distribution towards urban regions—were tested on a single geolocation model using different languages. In Chapter 4, we demonstrated the need for urban and rural evaluation from the viewpoint of applications. In this chapter, we focus on the effectiveness of evaluation using fifteen geolocation models and two baselines in a controlled experimental setting. In Section 5.1, the standardized evaluation process is discussed in detail, covering the evaluation metrics, the significance tests required to validate the observations and assess the effectiveness of these metrics, and the unified output and reverse-geocoding required to guarantee a fair evaluation. The geolocation models and baselines considered are described in Section 5.2. Results are discussed in Section 5.3. In Section 5.4, a practical guide is proposed, while discussing some of the limitations that have not been addressed.

5.1 Standardized Evaluation

In considering how to carry out standardized evaluation, first, 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 ensure the fairness of comparisons.

5.1.1 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] in the context of Twitter user geolocation, which were included partially by Rodrigues et al. [2016] and fully in the previous chapter. 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.

5.1.2 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. 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) introduced by 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.[1]

[1] https://bit.ly/2F2pZB0

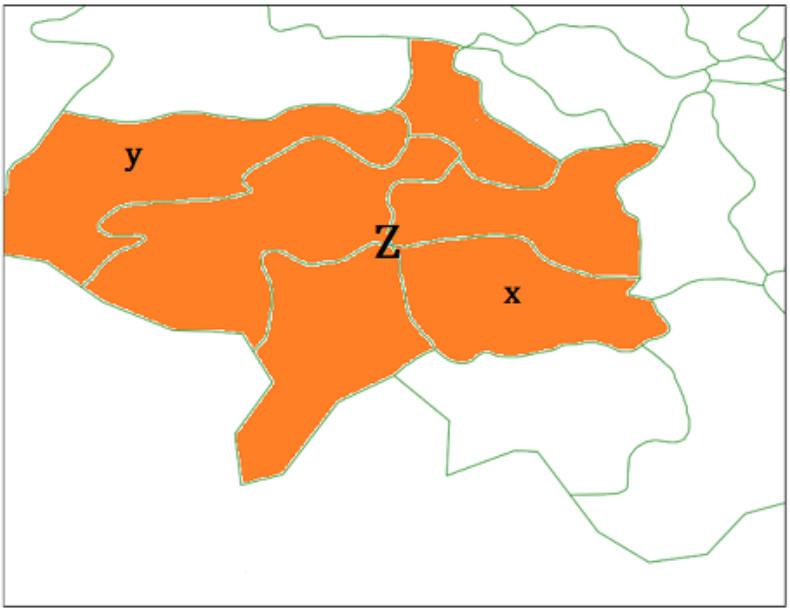


Figure 5.1: Example of unfair comparison between systems with different underlying earth representations. Cell x is the home location of a user and cell y is the predicted location by system A. The orange cells represent the home and predicted city (Z) of a user by system B.

5.1.3 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 5.1 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 API before evaluation. There 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 requests per day. Free APIs have a small limit, e.g. 2,500 requests per day for Nominatim, while commercial APIs have a larger limit, e.g. 100,000 requests per day for Google Reverse-Geocoding API V3. It took two weeks to reverse-geocode our local dataset of the size 1.5 million users, using Google API. Using a single reverse-geocoding API not only guarantees a fair comparison over the same set of 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.

5.2 Experimental Setup

We examine two sets of systems. The first set (LOCAL) includes four geolocation models and two baselines, trained and tested (over 30,000 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.

5.2.1 Local Models

Ground Truth

We employed the geographically global geotagged tweet collection **TwArchive**, described in Section 3.1. We focused on English tweets only. 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.

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 model[2] 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, Pigeo[3], 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 [Fan et al. 2008].

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] and in Chapter 3.

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

[2]https://github.com/utcompling/textgrounder/wiki/RollerEtAl_EMNLP2012
[3]https://github.com/afshinrahimi/pigeo

a classification task with multiple majority (or close to majority) classes [Pedregosa et al. 2011], unlike MC which originated in binary classification.

Both baselines were implemented using scikit dummy classifier 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.

5.2.2 W-NUT Models

W-NUT[4] 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.

5.3 Results

Table 5.1 details the results of our experiments on two sets of systems (LOCAL and W-NUT) across all metrics mentioned in Table 2.1; precision, recall, and 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.

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).

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

Table 5.1: Evaluation based on precision (P), recall (R) and f1-score (F1), using micro (μ) and macro (M) averaging, 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	$F1_M$	Acc	Acc@161	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	0.787	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

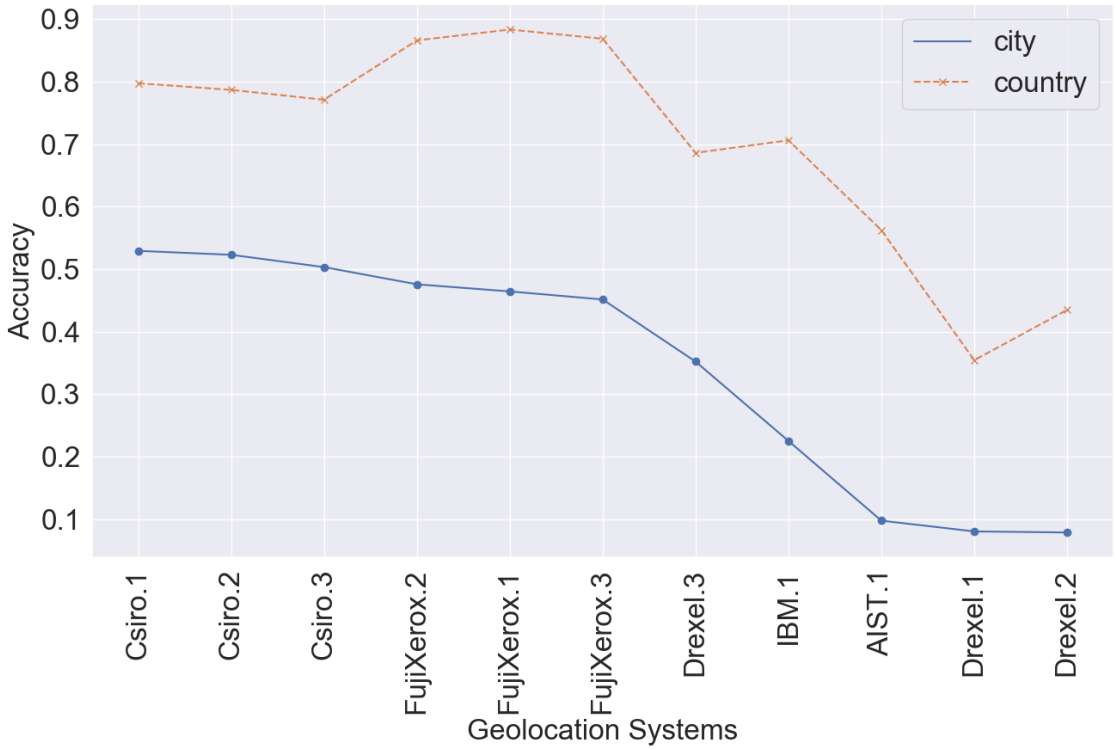


Figure 5.2: Evaluation of W-NUT based-on accuracy at the levels of city and country, ordered by city in a descending order.

5.3.1 Best Geolocation System

We compare two forms of evaluation based on metric commonality as shown in Table 2.1: most common metrics (Acc, Acc@161, Median and Mean error distances) and alternate metrics (PRF using μ vs M averaging).

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 5.1 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 representa-

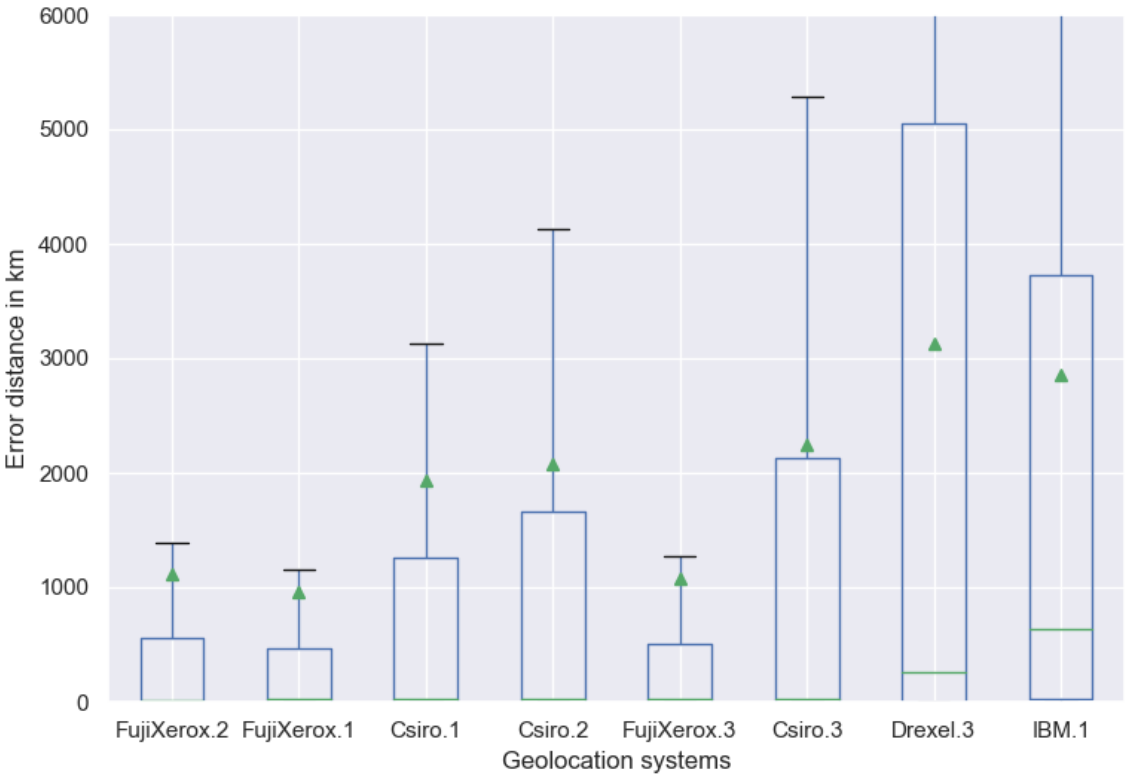


Figure 5.3: Evaluation of W-NUT based-on error distance metrics (Median and Mean) in km.

tions, the W-NUT systems, measured using Acc, are shown in Figure 5.2. Standardization enables the comparison of the best performance of each geolocation model.

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 5.1. This gap is related to the geographic area 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 5.1). 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 5.3. 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 because it penalizes geolocation systems with high error distances. 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 5.1 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.

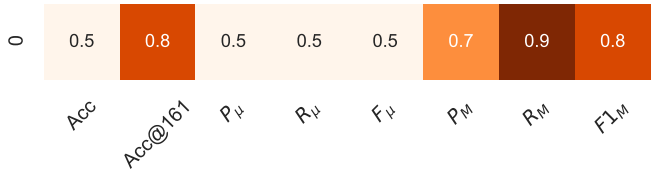
Imbalance Influence Using Alternate Metrics

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.

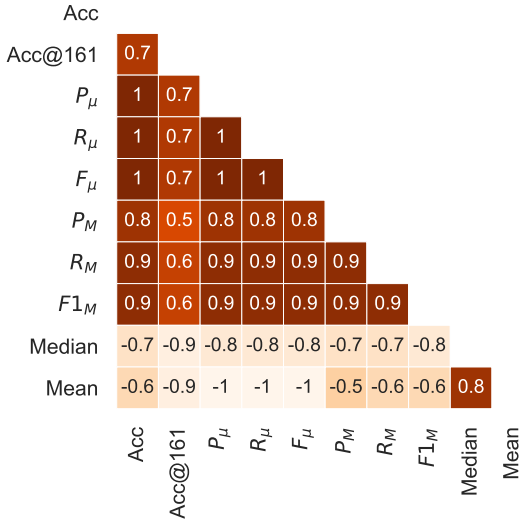
A consistent drop in performance can be seen from μ to M, see columns P_μ to $F1_M$ of Table 5.1. While RL12 and HN14 are competitive at the level of Acc, RL12 tend to have higher precision than HN14 using micro averaging, and vice versa using macro averaging. LSVM is another example where Acc is a limited 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).

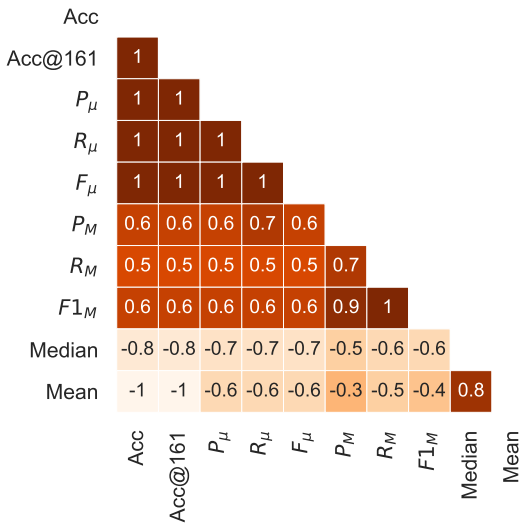
To summarize the best system analysis, we demonstrated (in §5.3.1) that unifying the output format and reverse-geocoding locations before evaluation are essential to ensure the fairness of comparison. A majority class baseline is recommended at the country level in the case of using Acc or micro averaging 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?



(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.

Figure 5.4: Kendall's τ_β rank correlations between pairs of effectiveness metrics for the W-NUT collection, *p-value* ≤ 0.05 for all correlations.

5.3.2 Comparing Metrics

Kendall's τ is a correlation measure that quantifies the agreement between two ranked lists. We calculated τ_β for all combinations of the employed metrics, see Figure 5.4. 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_β 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 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 5.4a), 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 5.4b), 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 5.4c), 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.

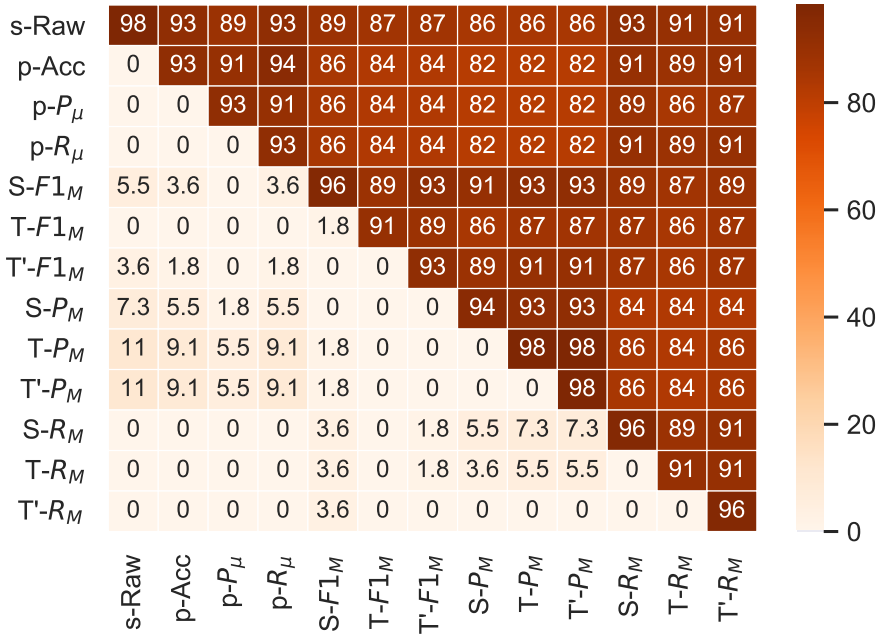
5.3.3 Statistical Significance

As was apparent from the system effectiveness scores in Table 5.1, 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.

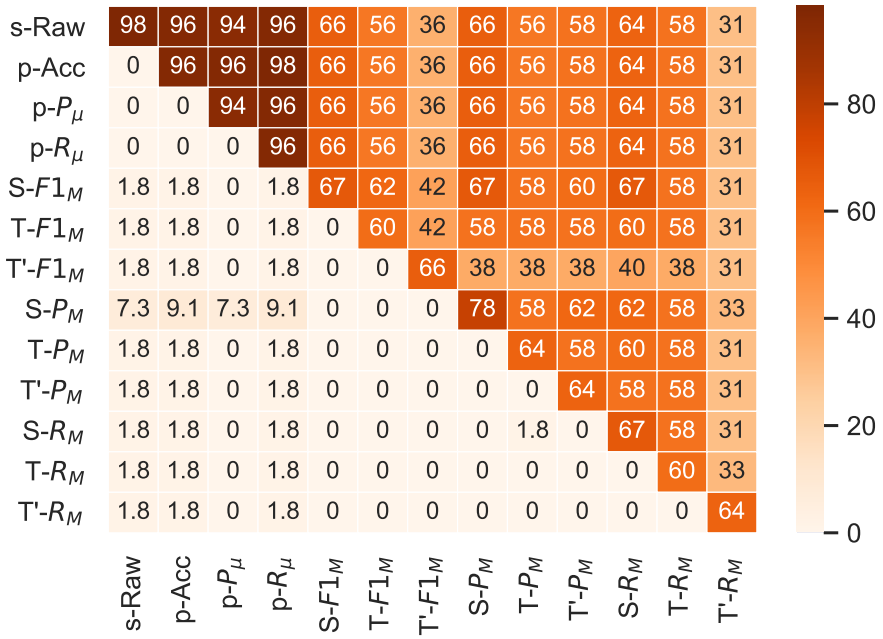
Following [Moffat et al. 2012], the outcome of a significance test will be categorized into one of two classes. Given two models A and B, calculated for two metrics X and Y, suppose that significance tests are run on the models' outputs using both metrics. If they show statistically significant results 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 5.5a and 5.5b 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 5.5a). They are much lower at the level of country for all comparisons involving macro tests (31–78%, see Figure 5.5b), which suggests that there is not a large difference in performance between geolocation models. On the contrary,



(a) City-level



(b) Country-level

Figure 5.5: Significant agreements and disagreements, *p-value* = 0.05 for tests. W-NUT: 11 systems, 55 system pairs. Micro tests are s-Raw, p-Acc, p-P_μ, and p-R_μ, while the rest represent Macro tests. Significance tests abbreviations stand for: s is sign-test, p is proportions z-test, S is macro sign-test, T is macro t-test, and T' is Wilcoxon test.

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 SSA involving micro metrics (first fours 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.

5.4 Discussion and Limitations

Datasets built using Twitter cannot be fully shared and are practically not reproducible 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], is not practical as well because of the high effort exerted by the host to maintain the submissions. 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 (§ 2.4.1) and AUC (§ 2.4.6).

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 (so are 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 [Rahimi et al. 2018, Ebrahimi et al. 2018, Bakerman et al. 2018, Miura et al. 2017], specially using Acc@161 at fine granularities (city and county). A combination of macro metrics (precision, recall and f1-score), 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.

5.5 Summary

In this chapter, we 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 geolocation includes:

1. creating its own dataset,
2. sharing its geolocation model with the ability to be retrained,
3. using a unified output format (GPS coordinates),
4. using a single reverse-geocoding API for discrete evaluation of all the geolocation models considered,
5. employing a combined set of evaluation metrics at the micro and macro levels, and different geographic granularities
6. quantifying the agreement between the evaluation metrics through rank correlation, and
7. verifying the conclusions by conducting the recommended statistical significance tests.

In Chapter 6, we describe the framework employed to conduct the analysis in this chapter, while highlighting its utility to the broader context of document geolocation.

CHAPTER 6

Geolocation Evaluation Framework

In the previous chapter, we proposed a guideline for evaluating Twitter user geolocation systems. Based on the challenges identified and our extensive statistical analysis, we demonstrated the effectiveness of the proposed evaluation framework. In this chapter, we highlight how this framework is useful for the broader discipline of document geolocation. In Section 6.1, the task of document geolocation is defined, along the challenges involved when social media is the source of documents. In Section 6.2, the different approaches of evaluation design are discussed, while justifying the choice of the approach adopted in this framework. Lastly, the design of the proposed framework is introduced in Section 6.3. This framework helps utilize the large volume of non-geotagged social media data to promote geo-spatial research.

6.1 Document Geolocation

Geolocation is the process of linking a document (e.g. Wikipedia article, web page, social media entity, etc.) to a location on earth. Geocoding is a fundamental sub-task of geolocation in which a location is resolved to an address. Based on the input format, geocoding is categorised into two tasks, namely: geocoding, the process of resolving a location from textual information; and reverse-geocoding, the process of resolving a location from GPS-coordinates.

Prior research has adopted two main approaches to construct gold standard data for geolocation applications. In a traditional manual approach, initial addresses are extracted from patients' records collected by health organisations. The quality of gold standard in this case is high because the addresses represent real locations. In an automated approach, initial addresses are extracted from the location field of user profiles on social media. These addresses are then geocoded based on the textual information describing the location. The quality of gold standard data in this case relies on the quality of the geocoders and the input of social media users.

Prior research focused on evaluating the quality of geocoding systems. Goldberg et al. [2013] demonstrated that geocoding systems handle erroneous text with varying degrees of quality, and have different sets of location labels and administrative regions. They proposed an evaluation framework for comparing geocoding systems. This framework empowers researchers and decision makers to choose the correct geocoding system for the particular needs of their applications. However, it does not address the broader task of document geolocation.

Although social media networks, Twitter in specific, provide a large volume of data, only 30% of the users provide textual location information; and only 1% of tweets are geotagged. This means that applications are not utilizing the large volume of non-geotagged data. Document geolocation is, therefore, of a paramount importance to promote geospatial research. Only few researchers considered the evaluation of document geolocation [Jurgens et al. 2015a, Han et al. 2016] with some limitations which were discussed in §2.4.3 and §2.4.4.

In the previous chapter, we demonstrated that GPS-coordinates are a more effective output format for evaluating geolocation systems. As such, we develop an evaluation framework for comparing geolocation systems where reverse-geocoding only is involved. This tool aims to empower researchers and decision-makers to choose the correct geolocation system for the particular needs of their application.

Document geolocation has been an active research area over the last decade, resulting in hundreds of publications, geolocation systems and datasets [Melo and Martins 2017, Zheng et al. 2018]. Comparison of such systems share the same challenges of Twitter user

geolocation:

i. The earth can be represented differently, as coordinate points, a grid of cells, a list of administrative locations, or a list of points of interest.

ii. The distribution of documents over locations tends to be imbalanced, which has to be managed.

iii. A wide range of evaluations metrics have been used in previous research.

iv. Social media datasets are often not reproducible, especially Twitter, because of the social networks' policy of not sharing raw data.

We, therefore, share our evaluation framework with the research community, hoping researchers will employ it in their future geolocation research. We demonstrated the utility of this framework in evaluating Twitter user geolocation systems in Chapter 5.

6.2 Replicable Evaluation

Replicability — the process of reproducing scientific experiments to obtain a consistent result — is a fundamental pillar of the scientific process. For a researcher to demonstrate an advancement in their experiment, they have to replicate either the existing research methodologies or the evaluation design (including the datasets). Replicable evaluation can be categorized into three approaches.

Shared Evaluation Settings provide public datasets and annotations for evaluation to a particular research problem, known in the research community as shared tasks. This approach paid its dividends in many research areas with standing evaluation conferences such as TREC and CLEF in Information Retrieval. With the rise of social media and the availability of user-generated content, researchers adopted the same approach to promote research using social media datasets. W-NUT[1] (Workshop on Noisy User-generated Text) and SSM4h[2] (Social Media Mining for Health) are examples of shared tasks in Natural Languages Processing based on social media datasets. However, Twitter datasets are non-transferable because the terms of usage allows sharing tweet IDs only. Therefore, any reconstructed Twitter datasets will be subject to data decay (see §2.4.3).

[1] https://www.aclweb.org/anthology/venues/wnut/
[2] https://www.aclweb.org/anthology/W19-3203/

Remote Evaluation requires participants to submit a program to a hosting platform to be tested on unseen data that is never released. This approach is common in competitive programming platforms such as HackerRank[3] and CodeChef[4]. Each competition defines the input and output formats of the data.

Hosted Evaluation Environments provide access to non-transferable private data (such as Twitter datasets) hosted by a single operator. An experimenter submits a request to the host along with a code to be remotely executed and evaluated on their behalf. 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 – limits the utility of this approach.

Both Remote and Hosted Evaluation don't release their datasets. However, they differ in the type of tasks supported. The former focuses on computational tasks such as the Fibonacci problem, while the latter supports research experiments involving machine learning and natural language processing.

Standardized Evaluation Process follows the traditional model of researcher-owned datasets to overcome the challenge of non-transferable social media datasets. It provides a unified evaluation framework covering the factors that influence each task to ensure a fair evaluation against previous research. This approach, however, requires researchers to open-source their models for replicability purposes, which aligns with the growing trend in the research communities these days.

In this chapter, we adopt a standard evaluation process for the task of document geolocation. In the next section, we propose an evaluation tool that addresses the three main challenges introduced in §6.1 by:

i. Unifying the input and output location to a generic format which gives researchers the flexibility to employ the earth-representations that suit their models.

ii. Maintaining the wide range of evaluation metrics employed in existing research while providing more analysis about the real differences between these metrics.

iii. Promoting the usage of researcher-owned datasets will resolve the challenge of social

[3]https://www.hackerrank.com/
[4]https://www.codechef.com/

media data decay.

6.3 GeoLocEval

GeoLocEval is an open source python package[5] to evaluate the performance of a given set of geolocation systems.

6.3.1 Evaluation Process

We follow the process presented in § 5.1, by first comparing systems under different evaluations and geographic granularities, next examining rank correlations of systems to assess the effectiveness of the evaluation metrics, and finally studying the statistical significance to validate the differences observed between geolocation systems. All the generated results are exported to a text file.

GeoLocEval employs ten evaluation metrics categorized into three groups. Continuous evaluation is based on the estimated GPS coordinates of a document, and represented by median and mean error distances from the original location. Discrete evaluation is based on the resolved location of a document (e.g. city), 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.

GeoLocEval assesses the effectiveness of the ten evaluation metrics using rank correlations. Kendall's Tau-b is calculated for all combinations (45) of the employed metrics.

Statistical tests employed in GeoLocEval are categorized into two groups given the multi-class and imbalanced nature of the task discussed in §5.1.2. **Micro tests** are the micro sign test (s) and proportions z-test (p). **Macro tests** include macro sign test (S), macro t-test (T), and macro t-test after rank transformation (T', a.k.a Wilcoxon). We use two-sided versions of the tests to avoid any bias in the direction of the effect. Our implementation of the multi-class statistical significance tests are based on the binary classification statistical routines provided by Scipy[6] — a python library which provides

[5] https://bitbucket.org/amourad/geoloceval.git
[6] https://www.scipy.org/scipylib/index.html

an efficient routines for statistics.

6.3.2 Reverse-Geocoding

In Section 5.1.3, we established the need for a unified location format (GPS-coordinates) to ensure a fair comparison of geolocation models. While GPS-coordinates are accurate representations of locations, spatial applications require evaluation at the level of administrative regions (e.g. city, country) because they are human-readable. Reverse-geocoding, therefore, is employed to resolve a pair of coordinates to an address. GeoLocEval supports two of the most common reverse-geocoding APIs based on the limit of reverse-geocoding requests per day and the price. The choice of a reverse-geocoding service, however, is irrelevant to the evaluation process. The quality of the chosen reverse-geocoding service will influence the considered geolocation systems with the same degree.

The Nominatim reverse-geocoding service[7] is a search engine for OpenStreetMap data. Its API limits access to 2,500 requests per day. The Nominatim API is chosen because it provides a free service, which makes it affordable to everyone. The resolved location includes a breakdown of the address into eight administrative regions, namely: country, state, county, city, suburb, major streets, major and minor streets, and building. A location can also include a list of alternative names (e.g. language variants), if available. An example of the output is shown in Listing 6.1.

Google reverse-geocoding[8] is a search engine for Google maps. Its API allows up to 100,000 requests per day. Although Google API V3 is a commercial service, it is the most feasible option to reverse-geocode large datasets. It took two weeks to reverse-geocode our dataset of 1.5 million users for the cost of $1000 AUD. Nominatim would take 600 days to reverse-geocode the same dataset. The resolved location includes a breakdown of the address into eight administrative regions, namely: country, administrative_area_level_1 (state), administrative_area_level_2 (county), locality (city), neighbourhood, route, and street number. if available. An example of the output is shown in Listing 6.2.

GeoPy[9] is a python client for several popular geocoding web services, which unifies

[7] https://nominatim.org/release-docs/develop/api/Reverse/
[8] https://developers.google.com/maps/documentation/geocoding/start
[9] https://geopy.readthedocs.io/en/stable/

```
{
  "type": "FeatureCollection",
  "licence": "Data  OpenStreetMap contributors, ODbL 1.0. https://osm.org/copyright",
  "features": [
    {
      "type": "Feature",
      "properties": {
        "place_id": "18512203",
        ... Additional information truncated in this example[] ...
        "display_name": "71, Via Guglielmo Marconi, Saragozza-Porto, Bologna, BO,
        ↪  Emilia-Romagna, 40122, Italy",
        "address": {
          "house_number": "71",
          "road": "Via Guglielmo Marconi",
          "suburb": "Saragozza-Porto",
          "city": "Bologna",
          "county": "BO",
          "state": "Emilia-Romagna",
          "postcode": "40122",
          "country": "Italy",
          "country_code": "it"
        }
      },
      "bbox": [ 11.3397676, 44.5014307, 11.3399676, 44.5016307 ],
      "geometry": {
        "type": "Point",
        "coordinates": [ 11.3398676, 44.5015307 ]
      }
    }
  ]
}
```

Listing 6.1: Nominatim JSON Output Format

the requests and retrieving the results from different reverse-geocoding APIs. This adds to the flexibility of GeoLocEval to support any API required by the research community for any particular needs. To use Nominatim and Google V3, a user will have to register at the corresponding API service and provide the key access to GeoLocEval. In terms of operational modes, GeoLocEval maintains the batch mode provided by GeoPy (the ability to reverse-geocode records in batch), while maximizing efficiency by supporting multi-threading. A batch of records is processed in parallel over the cores of the deployment server.

While both APIs support eight administrative levels, we limit the analysis of GeoLocEval to the most common four administrative granularities in existing Twitter geolocation research, namely: city, county, state and country. GeoLocEval caches all the resolved GPS

```
{
   "plus_code" : {
      "compound_code" : "P27Q+MC New York, NY, USA",
      "global_code" : "87G8P27Q+MC"
   },
   "results" : [ {
         "address_components" : [ {
               "long_name" : "279", "short_name" : "279",
               "types" : [ "street_number" ]
            }, {
               "long_name" : "Bedford Avenue", "short_name" : "Bedford Ave",
               "types" : [ "route" ]
            }, {
               "long_name" : "Williamsburg", "short_name" : "Williamsburg",
               "types" : [ "neighborhood", "political" ]
            }, {
               "long_name" : "Brooklyn", "short_name" : "Brooklyn",
               "types" : [ "political", "locality", "locality_level_1" ]
            }, {
               "long_name" : "Kings County", "short_name" : "Kings County",
               "types" : [ "administrative_area_level_2", "political" ]
            }, {
               "long_name" : "New York", "short_name" : "NY",
               "types" : [ "administrative_area_level_1", "political" ]
            }, {
               "long_name" : "United States", "short_name" : "US",
               "types" : [ "country", "political" ]
            }, {
               "long_name" : "11211", "short_name" : "11211",
               "types" : [ "postal_code" ]
            }
         ],
         "formatted_address" : "279 Bedford Ave, Brooklyn, NY 11211, USA",
         ... Additional information truncated in this example[] ...
      },
   ... Additional results truncated in this example[] ...
   ],
   "status" : "OK"
}
```

Listing 6.2: Google JSON Output Format

coordinates by any of the APIs to reduce the number of duplicate requests and maximize the usage of the limited requests per day.

6.3.3 Input and Output Formats

The input is a list of JSON files, one for each system to be compared. Each file includes a list of documents. Each document has a unique identifier and a location. Locations are expressed in the most generic format: GPS coordinates, as in Listing 6.3. This format is

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

Listing 6.3: GeoLocEval JSON Input Format

```
{
    "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
        },
    }
}
```

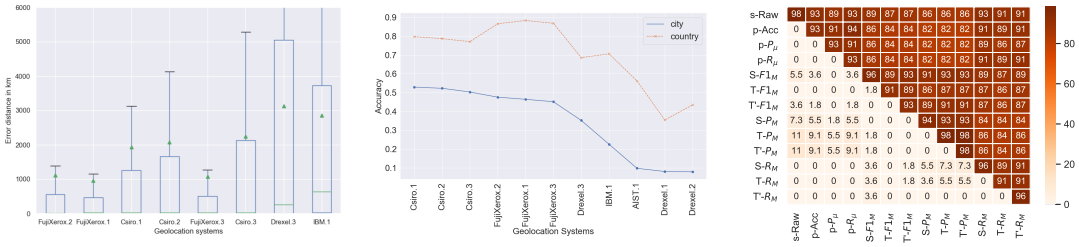
Listing 6.4: GeoLocEval JSON Output Format

compatible with the Twitter geolocation prediction shared task at the level of tweets and users [Han et al. 2016], known as WNUT.

The GPS coordinates are expanded using a single geocoding API, Nominatim by default. Results are exported to a JSON file, as in Listing 6.4. Each document has a unique identifier and a list of the result objects corresponding to each geolocation system, including the ground-truth. An object includes the coordinates estimated by the geolocation system, the reverse-geocoded administrative regions and error distance in km.

6.3.4 Visualization

GeoLocEval visualizes the results using three types of graphs which have been designed to convey the differences between geolocation models easily, see Figure 6.1. Boxplots are employed to evaluate the continuous range of error distances (Fig 6.1a). Line graphs are used for visualizing discrete metrics at multiple geographic granularities (Fig 6.1b). Heatmaps are employed to present the rank correlations and statistical agreements and disagreements on ranking geolocation systems based on the evaluation metrics (Fig 6.1c). For the statistical significance heatmaps, the diagonal values represent the percentage of



(a) Boxplots for continuous evaluation

(b) Linegraphs for discrete evaluation

(c) Heatmaps for correlations and statistical analysis

Figure 6.1: Overview of graphs generated by GeoLocEval

```
python geoloceval.py --online --api nominatim --geoloc <coords-cache> --outdir
↪  <output-directory> <ground-truth>.json <geolocation-system>*.json
```

Listing 6.5: GeoLocEval Online Mode

systems pairs that are significantly different based on a single metric, the values above the diagonal show the percentage of statistically significant agreements, the values below the diagonal show the percentage of statistically significant disagreements.

6.3.5 Operational Modes

GeoLocEval supports three operational modes of document geolocation evaluation, independent of the task, e.g. user, tweet or wikipedia-article geolocation:

Online Mode: This mode supports an end-to-end evaluation of document geolocation including online reverse-geocoding of locations. A user passes a list of files that represent the ground truth and predictions of each system in a file in the form of GPS-coordinates (see Listing 6.3). All coordinates, including the ground-truth, are reverse-geocoded using the same API to administrative regions and exported to a single file in the format shown in Listing 6.4. All results are then evaluated and statistically analyzed. Generated results and graphs are saved to an output directory. An example of a command line to run this mode is shown in Listing 6.5.

Offline Mode: This mode allows users to utilize an offline cache of previously reverse-geocoded coordinates for incremental development of their geolocation systems.

Statistical Significance Mode: Earlier in §5.1.2, we highlighted the importance of

```
python geoloceval.py sst --outdir <output-directory> <ground-truth>.json
↪  <geolocation-system>*.json
```

Listing 6.6: GeoLocEval Statistical Significance Mode

statistical significance (SS) analysis in the field of natural language processing (NLP) [Dror et al. 2018], and the set of required SS tests to deal with a skewed category distribution in classification tasks [Yang and Liu 1999]. Later in §6.3.1, we pointed out the limitation of existing statistical libraries to provide SS tests for binary classification only with no consideration to the frequency distribution of multi-class categorization tasks. GeoLocEval, therefore, provides a standalone operational mode for the NLP research community to utilize our implementation of five SS tests at micro and macro levels. An example of a command line to run this mode is shown in Listing 6.6.

6.4 Summary

In this chapter, we introduced GeoLocEval, an evaluation framework that allows geolocation systems to be compared fairly and accurately. It encompasses the best practices for an effective evaluation based on our analysis of Twitter user geolocation in previous chapters. It is domain and application independent, and a fair comparison between geolocation systems is guaranteed regardless of the underlying earth representation. Ten evaluation metrics are employed over four geographic granularities to cover the diverse needs of applications. GeoLocEval conducts comprehensive statistical analysis aligned with the evaluation metrics. Sharing social media datasets is not required as long as researchers allow retraining of their geolocation models. We encourage researchers to employ GeoLocEval in their future geolocation research based on our findings.

CHAPTER 7

Conclusion

This thesis aims to improve the utility of Twitter geolocation systems with downstream applications. We examined the influence of geographical biases in social media population and language use on the quality of geolocation prediction. In this chapter, we summarize our findings and contributions. More specifically, we raised four research questions: i) what is the influence of language bias, in comparison to population bias, on the quality of geolocation prediction?; ii) how important is urban and rural evaluation of geolocation systems from the point of view of applications?; iii) how can geolocation systems be evaluated effectively taking into account the challenges of social media data and the generic information needs of applications?; and iv) how does our proposed effective evaluation process of Twitter geolocation generalize to the broader context of document geolocation?

In Chapter 2 we reviewed previous work on Twitter geolocation finding that: i) most past work utilizes the influence of language bias, with little comparison to the influence of population bias; ii) English is the most prominent language in existing research; iii) geolocation applications might have different needs which might require evaluation from several perspectives; and iv) a wide range of evaluation metrics has been employed, yet these do not address the previous challenges.

In Chapter 3, we quantified the influence of social media biases among other features, finding that population bias has more influence on the accuracy of Twitter geolocation than language influence, and typical evaluation metrics are biased towards urban cities.

We proposed contrasting urban and rural evaluation metrics to address the population bias in social media datasets.

In Chapter 4, we quantitatively assessed the need for urban and rural evaluation in the context of applications, finding that evaluation metrics should treat rural and urban cities with the same degree of importance.

Chapter 5 focused on devising a standardized and statistically principled evaluation process to geolocation systems, finding that a unified output format and reverse-geocoding are vital for a fair evaluation.

Lastly, Chapter 6 demonstrated the utility of our Twitter geolocation evaluation framework to the the broader discipline of document geolocation, while describing its design principles and operational modes.

We now summarise the key contributions and findings of this research.

7.1 Language Influence

Language is a valuable source of location information that helps to enhance the accuracy of Twitter user geolocation. Existing research hypothesized that the *language bias* and *population bias* played an important role in determining geolocation accuracy. However, English is the most prominent language in the datasets of previous research, with only a few researchers having considered the language in which a tweet is written as a feature for Twitter geolocation. Another shortcoming is that the datasets used were of different sizes in comparison to English.

To address these shortcomings, we first demonstrated that *population* bias has more influence on the accuracy of Twitter user geolocation than *language* bias. The widely-used accuracy metric is biased towards the most populous locations. We proposed contrasting micro and macro averaging over precision, recall and f1-score for the generic evaluation of geolocation systems. The value of macro-averaging is that it treats rural and urban locations with the same degree of importance. The decision to favor one averaging technique over the other depends on the particular needs of an application.

7.2 Urban and Rural Evaluation

Geo-spatial applications require evaluating geolocation applications from several perspectives. Typical evaluation practices focused on a few measures introduced in early literature ignoring the needs of applications. To tackle this shortcoming, we used a global database of events as an indicator of the importance of rural and urban locations from a journalistic standpoint, and a Twitter user dataset to be employed for training geolocation systems. We demonstrated that both datasets follow a power-law distribution, similar to census data, where a few urban locations have the largest populations, but rural locations collectively have a substantial population which should be considered for evaluation. In this case, a combined micro and macro evaluation is required.

7.3 Evaluation Effectiveness

The evaluation of geolocation methods is affected by many factors, such as dataset availability, pre-processing, ground-truth construction, geographic coverage, population bias, and how the earth is represented. Moreover, evaluation at multiple levels of geographic granularity is not widely used despite it being required by some applications. To address these challenges, we developed a standard geolocation evaluation framework that helps decision makers with choosing a geolocation model suitable to their needs. We showed that previous research practices influence the fairness of evaluation, and some geolocation systems are not performing as previously thought. To enhance the credibility of our evaluation framework, we developed a statistically principled process for evaluation given the multi-class categorization nature of geolocation.

7.4 Geolocation Evaluation Framework

Twitter user geolocation challenges introduced in this thesis are common in the broader field of document geolocation. To tackle this shortcoming, we open-sourced our geolocation evaluation framework hoping that the research community will employ it in future document geolocation research. This framework follows the design principles defined in

previous research for usability, efficiency and the ability to extend the current features to support more evaluation metrics as needed.

7.5 Future Work

This thesis was originally motivated by studying the lexical variations of languages and their impact on geolocating Twitter users. A simple feature represented by the number of location indicative words per language, due to the lack of enough resources other than English, was found to have no impact. It was hard to assess the richness of the vocabulary associated with the different languages (English is the pivot), or even dialects within the same language (no definitive list of dialects per language). Gonçalves and Sánchez [2014] showed that major local varieties of Spanish can be recognized in Twitter and categorized into regions covering urban cities versus rural areas and small towns. They used a hybrid approach which combined linguistic resources (concepts and utterances) with computational methods. Recently, there is a growing interest in studying the lexical variations of languages at a finer granularity (i.e. county) than before (i.e. cities), providing lexicons for such variations [Huang et al. 2016]. These linguistic resources for English and Spanish can be utilized to measure how the difficulty of a language will influence the quality of Twitter geolocation.

Our language analysis was limited to a single geolocation system. Future work might consider evaluating other geolocation inference techniques from a language perspective, making use of a wide range of open source frameworks. For instance, Wing and Baldridge [2014] demonstrated that probabilistic language models and hierarchical logistic regression outperform location indicative words and text-categorisation for English, but on a different representation of location (i.e. not cities). Jurgens et al. [2015b] released a framework for nine different network-based geolocation systems. Recently, Rahimi et al. [2018] explored using a joint text and network based approach.

Twitter geolocation research started with limited geographic coverage such as few cities or states, mainly in the United States. As research advanced and the usage of social media spread in the entire world, the global geographic coverage of geolocation systems

has become the norm. We demonstrated that the distribution of Twitter users over cities follows a power-law. However, not all cities fit this model. Based on this observation, the estimated parameters of a power-law model (x_{min}) can be utilized to define a threshold on the number of Twitter users required for a city to be considered while training geolocation systems. This could reduce the influence of noise, locations with insufficient data.

The 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. Wing and Baldridge [2014] evaluated their models on six datasets from different domains, namely Twitter, Wikipedia and Flickr. In Chapter 5, the statistical significance of the differences between geolocation models were evaluated on a single dataset. A future direction could be to extend our geolocation evaluation guide to include the replicability analysis for statistical significance analysis over multiple datasets [Dror et al. 2017]. This should further enhance the credibility of the superiority of one algorithm over another across multiple datasets.

Gao and Sebastiani [2015] 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 should therefore 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 followed a quantitative approach to anticipate the information needs of geolocation applications, promoting for the idea of urban and rural evaluation. While this approach is a vital step to close the gap between theoretical research on social media geolocation and the application, we believe future work should start investigating the evaluation of geolocation models qualitatively. Instead of anticipating the needs of an applications, researchers should collaborate closely with domain experts, such as journalists, linguists, or humanitarians to develop the needs and evaluation metrics in the context of a specific task.

7.6 Summary

The importance of considering social media geographic biases in evaluating geolocation systems is highlighted in this thesis. In contrast to previous work, we demonstrated that population bias has more influence on the quality of Twitter geolocation than language bias. We proposed a standardized evaluation guide for Twitter geolocation that guarantees a fair, statistically principled and effective evaluation taking into account the challenges highlighted throughout the thesis. We also provided a geolocation evaluation framework to the research community hoping to advance the research on the broader field of document geolocation.

Bibliography

A. Ahmed, L. Hong, and A. J. Smola. Hierarchical geographical modeling of user locations from social media posts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 25–36, 2013.

O. Ajao, J. Hong, and W. Liu. A survey of location inference techniques on Twitter. *Journal of Information Science*, 41:855–864, 2015.

J. Alstott, E. Bullmore, and D. Plenz. Powerlaw: a python package for analysis of heavy-tailed distributions. *PloS one*, 9:e85777, 2014.

L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World Wide Web*, pages 61–70, 2010.

J. Bakerman, K. Pazdernik, A. Wilson, G. Fairchild, and R. Bahran. Twitter geolocation: A hybrid approach. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 12:34, 2018.

E. Baykan, M. Henzinger, and I. Weber. Web page language identification based on URLs. *Proceedings of the VLDB Endowment*, 1:176–187, 2008.

C. Biemann, F. Bildhauer, S. Evert, D. Goldhahn, U. Quasthoff, R. Schäfer, J. Simon, L. Swiezinski, and T. Zesch. Scalable construction of high-quality web corpora. *JLCL*, 28:23–59, 2013.

D. A. Broniatowski, M. J. Paul, and M. Dredze. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. *PloS one*, 8:e83672, 2013.

S. H. Burton, K. W. Tanner, C. G. Giraud-Carrier, J. H. West, and M. D. Barnes. "Right time, right place" health communication on Twitter: value and accuracy of location information. *Journal of medical Internet research*, 14:e156, 2012.

Z. Cheng, J. Caverlee, and K. Lee. 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*, pages 759–768, 2010.

L. Chi, K. H. Lim, N. Alam, and C. J. Butler. Geolocation prediction in Twitter using location indicative words and textual features. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 227–234, 2016.

A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM Review*, 51:661–703, 2009.

A. Culotta. Reducing sampling bias in social media data for county health inference. In *Joint Statistical Meetings Proceedings*, pages 1–12, 2014.

K. Darwish, W. Magdy, and A. Mourad. Language processing for arabic microblog retrieval. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2427–2430, 2012.

C. A. Davis Jr, G. L. Pappa, D. R. R. de Oliveira, and F. de L. Arcanjo. Inferring the location of Twitter messages based on user relationships. *Transactions in GIS*, 15: 735–751, 2011.

N. Diakopoulos, M. De Choudhury, and M. Naaman. Finding and assessing social media information sources in the context of journalism. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 2451–2460, 2012.

T. H. Do, D. M. Nguyen, E. Tsiligianni, B. Cornelis, and N. Deligiannis. Multiview deep learning for predicting Twitter users' location. *arXiv preprint arXiv:1712.08091*, 2017.

M. Dredze, M. J. Paul, S. Bergsma, and H. Tran. 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*, volume 23, pages 20–24, 2013.

R. Dror, G. Baumer, M. Bogomolov, and R. Reichart. Replicability analysis for natural language processing: Testing significance with multiple datasets. *Transactions of the Association of Computational Linguistics*, 5:471–486, 2017.

R. Dror, G. Baumer, S. Shlomov, and R. Reichart. 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*, pages 1383–1392, 2018.

M. Ebrahimi, E. ShafieiBavani, R. Wong, and F. Chen. Twitter user geolocation by filtering of highly mentioned users. *Journal of the Association for Information Science and Technology*, 69:879–889, 2018.

J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, 2010.

R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of machine learning research*, 9:1871–1874, 2008.

W. Gao and F. Sebastiani. Tweet sentiment: From classification to quantification. In *Proceedings of the 2015 IEEE/ACM international conference on Advances in Social Networks Analysis and Mining 2015*, pages 97–104, 2015.

D. W. Goldberg, M. Ballard, J. H. Boyd, N. Mullan, C. Garfield, D. Rosman, A. M. Ferrante, and J. B. Semmens. An evaluation framework for comparing geocoding systems. *International journal of health geographics*, 12:50, 2013.

B. Gonçalves and D. Sánchez. Crowdsourcing dialect characterization through Twitter. *PloS one*, 9:e112074, 2014.

B. Han and T. Baldwin. Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 368–378, 2011.

B. Han, P. Cook, and T. Baldwin. Text-based Twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49:451–500, 2014.

B. Han, A. Rahimi, L. Derczynski, and T. Baldwin. 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*, pages 213–217, 2016.

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 *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 237–246, 2011.

B. J. Hecht and M. Stephens. A tale of cities: Urban biases in volunteered geographic information. In *Proceedings of the eighth international AAAI Conference on Web and Social Media*, pages 197–205, 2014.

Y. Huang, D. Guo, A. Kasakoff, and J. Grieve. Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, 59: 244–255, 2016.

D. D. Ingram and S. J. Franco. NCHS urban-rural classification scheme for counties. *Vital and health statistics. Series 2, Data evaluation and methods research*, pages 1–65, 2012.

G. Jayasinghe, B. Jin, J. Mchugh, B. Robinson, and S. Wan. CSIRO Data61 at the WNUT geo shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 218–226, 2016.

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 *Proceedings of the 2017 CHI conference on Human Factors in Computing Systems*, pages 1167–1178, 2017.

D. Jurgens. 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*, pages 273–282, 2013.

D. Jurgens, T. Finethy, C. Armstrong, and D. Ruths. 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*, pages 8–17, 2015a.

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 *Proceedings of the ninth international AAAI Conference on Web and Social Media*, pages 188–197, 2015b.

S. Kinsella, V. Murdock, and N. O'Hare. I'm eating a sandwich in Glasgow: modeling locations with tweets. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 61–68, 2011.

Y. Kryvasheyeu, H. Chen, E. Moro, P. Van Hentenryck, and M. Cebrian. Performance of social network sensors during Hurricane Sandy. *PLoS one*, 10:e0117288, 2015.

W. Labov, S. Ash, and C. Boberg. *The Atlas of North American English: Phonetics, phonology and sound change*. Walter de Gruyter, 2008.

V. Landeiro and A. Culotta. Robust text classification in the presence of confounding bias. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, 37:145–151, 1991.

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 *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 207–216, 2016.

P. A. Longley, M. Adnan, and G. Lansley. The geotemporal demographics of Twitter usage. *Environment and Planning A*, 47:465–484, 2015.

M. Lui and T. Baldwin. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL-2012 System Demonstrations*, pages 25–30, 2012.

M. M. Malik, H. Lamba, C. Nakos, and J. Pfeffer. Population bias in geotagged tweets. In *Ninth international AAAI conference on web and social media*, pages 18–27, 2015.

F. Melo and B. Martins. Automated geocoding of textual documents: A survey of current approaches. *Transactions in GIS*, 21:3–38, 2017.

A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist. Understanding the demographics of Twitter users. In *Proceedings of the fifth international AAAI Conference on Weblogs and Social Media*, pages 554–557, 2011.

Y. Miura, M. Taniguchi, T. Taniguchi, and T. Ohkuma. A simple scalable neural networks based model for geolocation prediction in Twitter. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 235–239, 2016.

Y. Miura, M. Taniguchi, T. Taniguchi, and T. Ohkuma. 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*, pages 1260–1272, 2017.

D. Mocanu, A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, and A. Vespignani. The Twitter of Babel: Mapping world languages through microblogging platforms. *PloS one*, 8:e61981, 2013.

A. Moffat, F. Scholer, and P. Thomas. Models and metrics: IR evaluation as a user process. In *Proceedings of the seventeenth Australasian Document Computing Symposium*, pages 47–54, 2012.

F. Morstatter, N. Lubold, H. Pon-Barry, J. Pfeffer, and H. Liu. Finding eyewitness tweets during crises. *ACL 2014*, page 23, 2014.

A. Mourad, F. Scholer, and M. Sanderson. Language influences on tweeter geolocation. In *Proceedings of the European Conference on Information Retrieval*, pages 331–342, 2017.

A. Mourad, F. Scholer, M. Sanderson, and W. Magdy. 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*, pages 437–440, 2018.

A. Mourad, F. Scholer, W. Magdy, and M. Sanderson. A practical guide for the effective evaluation of Twitter user geolocation. *Transactions of the Association for Computational Linguistics*, 0:1–23, 2019.

U. Pavalanathan and J. Eisenstein. Confounds and consequences in geotagged Twitter data. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2138–2148, 2015.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

A. Poulston, M. Stevenson, and K. Bontcheva. Hyperlocal home location identification of Twitter profiles. In *Proceedings of the 28th ACM conference on Hypertext and Social Media*, pages 45–54, 2017.

R. Priedhorsky, A. Culotta, and S. Y. Del Valle. Inferring the origin locations of tweets with quantitative confidence. In *Proceedings of the 17th ACM conference on Computer Supported Cooperative Work & Social Computing*, pages 1523–1536, 2014.

A. Rahimi, T. Cohn, and T. Baldwin. Pigeo: A python geotagging tool. In *Proceedings of the ACL-2016 System Demonstrations*, pages 127–132, 2016.

A. Rahimi, T. Baldwin, and T. Cohn. 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*, pages 167–176, 2017.

A. Rahimi, T. Cohn, and T. Baldwin. Semi-supervised user geolocation via graph convolutional networks. In *Proceedings of the 56th annual meeting of the Association for Computational Linguistics*, pages 2009–2019, 2018.

E. Rodrigues, R. Assunção, G. L. Pappa, D. Renno, and W. Meira Jr. Exploring multiple evidence to infer users' location in Twitter. *Neurocomputing*, 171:30–38, 2016.

S. Roller, M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 joint conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1500–1510, 2012.

D. Ruths and J. Pfeffer. Social media for large studies of behavior. *Science*, 346:1063–1064, 2014.

T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860, 2010.

S. Schifferes, N. Newman, N. Thurman, D. Corney, A. Göker, and C. Martin. Identifying and verifying news through social media: Developing a user-centred tool for professional journalists. *Digital journalism*, 2:406–418, 2014.

A. Schulz and P. Ristoski. The car that hit the burning house: Understanding small scale incident related information in microblogs. In *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

A. Schulz, A. Hadjakos, H. Paulheim, J. Nachtwey, and M. Mühlhäuser. A multi-indicator approach for geolocalization of tweets. In *Seventh International AAAI Conference on Weblogs and Social Media*, pages 573–582, 2013.

A. Schulz, B. Schmidt, and T. Strufe. Small-scale incident detection based on microposts. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 3–12, 2015.

A. Schulz, E. L. Mencía, and B. Schmidt. A rapid-prototyping framework for extracting small-scale incident-related information in microblogs: Application of multi-label classification on tweets. *Information Systems*, 57:88–110, 2016.

R. Schwartz, M. Naaman, and R. Teodoro. Editorial algorithms: Using social media to discover and report local news. In *Proceedings of the ninth international AAAI Conference on Web and Social Media*, pages 407–415, 2015.

F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34:1–47, 2002.

L. Sloan and J. Morgan. Who tweets with their location? understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. *PloS one*, 10:e0142209, 2015.

M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45:427–437, 2009.

K. Starbird, G. Muzny, and L. Palen. 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*, pages 1–10, 2012.

A. R. Tinoco and H. Ueda. The VARILEX project-spanish lexical variation. *Linguistica Atlantica*, 27:117–121, 2007.

B. Wing and J. Baldridge. Hierarchical discriminative classification for text-based geolocation. In *Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing*, pages 336–348, 2014.

B. P. Wing and J. Baldridge. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 955–964, 2011.

Y. Yang. An evaluation of statistical approaches to text categorization. *Information retrieval*, 1:69–90, 1999.

Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 42–49, 1999.

X. Zheng, J. Han, and A. Sun. A survey of location prediction on Twitter. *IEEE Transactions on Knowledge and Data Engineering*, pages 1652–1671, 2018.

A. Zubiaga, H. Ji, and K. Knight. Curating and contextualizing Twitter stories to assist with social newsgathering. In *Proceedings of the 2013 international conference on Intelligent User Interfaces*, pages 213–224, 2013.