Towards an unbiased approach for the evaluation of social data geolocation

Carlo Bernaschina, Ilio Catallo, Eleonora Ciceri, Roman Fedorov, Piero Fraternali
Politecnico di Milano, Dipartimento di Elettronica, Informazione e Bioingegneria
Piazza Leonardo da Vinci, 32 - 20133 Milan - Italy
name.surname@polimi.it

ABSTRACT

We present a study that reveals a significant statistical bias in the distributions of geolocated and non-geolocated social data. We state that this bias affects the real performance of social geolocation algorithms and can impair the results of these algorithms, which are commonly trained and tested on datasets consisting of crawled geolocated data. At last, we propose the construction of an a-posteriori geolocated dataset for an unbiased estimation of new and state-of-the-art algorithms alike.

1. INTRODUCTION

With the tremendous success of social media, an increasing body of research has started addressing the problem of analyzing user-generated content with the aim of extracting the latent information it contains. Indeed, social media provide researchers with an unprecedented means for inexpensively collecting and analyzing data: every action a user makes on a social platform carries a certain amount of information, an uploaded photo shows a snapshot of reality in a given place at a given time, a new status reflects an opinion, an added friend reveals a social and potentially geographical relationship, etc. In this respect, the advent of handheld devices as preferred tools for accessing social media has the potential to provide significant insights on the geo-spatial dimension of social data. Thanks to the widespread of such GPS-equipped devices, the number of geolocated social data has been steadily increasing for the past few years, which allowed the application of social media analysis to novel contexts, such as event detection and disaster management [1].

Despite these promising results, in many application contexts the fraction of geolocated social data is still too scarce to enable a fine-grained geographical analysis. In this regard, experiments conducted on Twitter reveal that a small percentage of data (usually around 0.3%) is geolocated [2]. Moreover, API restrictions on tweet access only increase this problem (as also confirmed by our experimental evaluation). Because of this, many works in the literature propose techniques for estimating the position of non-geolocated social data. Clearly, the assessment of each such technique depends on the availability of a groundtruth dataset, which is commonly constructed by collecting from the social media of choice only those data carrying the required geolocation information. The performance of a given technique is therefore assessed on the groundtruth dataset by contrasting, for each social datum, its actual location with its estimate. We argue that this evaluation approach, although allowing a precise performance assessment, is subject to an important bias. In particular, we claim that any performance so obtained should be intended as the quantification of the ability to *locate geolocated data with removed geolocation*, as opposed to the more desirable ability to *locate non-geolocated data*. Often in the literature, it is assumed the former ability to entail the latter. That is, it is assumed that geolocated and non-geolocated data *only* differ in the presence of the geolocation information. However, at the best of our knowledge, there is no evidence supporting the validity of such an assumption, and if it had to fail, it would imply that the performance obtained on groundtruth dataset might not relate to the ability to perform well in real-case scenarios.

In this work, we study the validity of the afore-mentioned assumption when applied to the Twitter social platform. Namely, we perform a statistical analysis of two populations - tweets with geolocation and tweets without geolocation - in terms of features commonly used by state-of-the-art geolocation algorithms. Our preliminary results show significant statistical difference between the two populations. We deem that this work highlights a bias in the traditional approach to geolocation estimation. The next sections describe the details of the experiments and propose a future direction of work aiming at quantitatively evaluating the impact of this bias on the performance of state-of-the-art algorithms.

2. PROBLEM STATEMENT

The key tenet of traditional geolocation algorithms is that geolocated and non-geolocated data only differ in the presence of the geolocation information. In this work, we question such an assumption. That is, we seek for an answer to the following question: *Do populations of geolocated and non-geolocated tweets significantly differ w.r.t. features other than the presence of a geolocation?* Since many state-of-the-art techniques extract features from tweets in order to esti-

This work is supported by the POR-FESR 2007-2013 PROACTIVE Project, http://www.proactiveproject.eu

GIR '15, November 26-27, 2015, Paris, France

DOI: http://dx.doi.org/10.1145/2837689.2837697

mate their location, any difference in the distribution of such features among the two populations might hinder the predictive capabilities on non-geolocated data. Hence, in the next section we first identify the most commonly used features in state-of-the-art techniques, and consequently verify their distribution among geolocated and non-geolocated tweets.

3. EXPERIMENTAL STUDY

In this section we illustrate the experiments that were conducted to compare the non-geolocated and the geolocated tweet distributions.

Dataset. We collected more than 24 million tweets by randomly sampling the Twitter real time feed for a week, so as to obtain an unbiased sample of what users publish on the platform. Then, we divided the sample into two subsets:

- *Geolocated dataset:* tweets whose geographical coordinates (in the form of latitude and longitude) are specified;

- *Non-geolocated dataset:* all the other tweets.

Overall, the geolocated dataset contains $105K$ tweets (0.4% of the total amount of collected tweets).

Features. We identified several features that proved useful for the geolocation of tweets, and analyze their variation between geolocated and non-geolocated datasets. For some fatures we measure their binary variation (e.g., *Has User Location*), whereas some others are precisely quantified (e.g., *N. of Geonames*). We hereby list the precise set of measured features and how variations are computed:

- *Has User Location*: percentage of tweets that are produced by users that specified their home location on their profile. Some approaches (e.g., [3]) use this feature to geolocate the tweets a user produces.

- *Has Hashtags*, *N. of Hashtags* and *N. of Hashtags (>0)*: respectively, percentage of tweets containing at least one hashtag, number of hashtags per tweet and number of hashtags per tweet considering only tweets with at least one hashtag. Some approaches (e.g., [4]) look for known hashtags associated with a specific place to locate tweets.

- *Has Geonames* and *N. of Geonames*: respectively, percentage of tweets containing at least one geoname (i.e., geographical name, such as a city or a country) and number of geonames per tweet. Some approaches (e.g., [2, 4]) locate tweets based on the geonames they contain. In this work, we define a geoname as a sequence of one or more consecutive words associated with a city listed in the GeoNames database[1] and with at least 100000 inhabitants.

- *N. of Words*: number of words per tweet. Some approaches (e.g., [2, 3, 4]) classify tweets as geolocated/non-geolocated according to their word distribution.

Results. For each of the above features, we computed the sample mean $\bar{x}_G$ and $\bar{x}_{NG}$ for, respectively, the geolocated and non-geolocated dataset, as well as the gap $\rho = \frac{\bar{x}_G - \bar{x}_{NG}}{\bar{x}_{NG}}$ between $\bar{x}_G$ and $\bar{x}_{NG}$ (Table 1). As shown, the two samples exhibit different values of sample mean for all the considered features. While one might expect the gap ρ to be especially notably for geographical-related features (e.g., $\rho = 400\%$ for the *N. of Geonames* feature), we registered important differences also for less obvious features, such as *N. of Hashtags* ($\rho = 126\%$). In order to test the statistical significance of our findings, we conducted two statistical analyses, namely: *i)* unpooled unpaired t-test for continuous features (e.g., *N. of Geonames*), and *ii)* χ^2-test for categorical features (e.g., *Has Hashtags*).

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

Table 1: Results of the feature statistical study

Feature	$\bar{x}_{NG}$	$\bar{x}_G$	ρ	Test Stat.
Has User Location	53.6%	71.8%	34%	$\chi^2 = 13.8k$
Has Hashtags	20.0%	24.7%	24%	$\chi^2 = 1.4k$
N. of Hashtags	0.34	0.77	126%	$t = 81$
N. of Hashtags (>0)	1.68	3.13	86%	$t = 106$
Has Geonames	2.03%	9.23%	345%	$\chi^2 = 26.7k$
N. of Geonames	0.02	0.10	400%	$t = 77$
N. of Words	19.9	22.2	12%	$t = 63$

All tests were performed with a statistical significance level $\alpha = 0.01$. We report the test statistic values in the last column of Table 1. The tests confirmed a statistical significant difference between the sample means of the two datasets we collected. That is, there exists a significant bias between the two populations for all the features we tested. Furthermore, it is important to note that state-of-the-art geolocation algorithms, which rely on the reported features for performing their estimation, most likely benefit from the existence of such a bias. This is indeed evident if we notice that the sample mean $\bar{x}_G$ dominates $\bar{x}_{NG}$ for all the considered features.

4. CONCLUSION AND FUTURE WORK

In this work we experimentally evaluated the existence of a difference between geolocated and non-geolocated tweets. To this end, we collected a random sample of geolocated and non-geolocated data from Twitter and verified their statistical difference w.r.t. the most significant features in the literature for geolocation estimation. We showed that the two populations are in fact characterized by a significant statistical difference. For this reason, we argue the existence of a bias in the usage of geolocated data for the assesment of geolocation algorithms. As future work, we propose the creation of an unbiased groundtruth dataset of tweets. Such a dataset will be collected through the application of volunteer crowdsourcing and user engagement techniques. Hopefully, the availability of an unbiased groundtruth dataset will ultimately allow not only a better comparison of state-of-the-art geolocation algorithms, but also an efficient training of new algorithms capable of better performing in real-world scenarios.

5. REFERENCES

[1] M. A. Cameron, R. Power, B. Robinson, and J. Yin. Emergency situation awareness from twitter for crisis management. In *Proceedings of the 21st international conference companion on World Wide Web*.

[2] R. Gonzalez, G. Figueroa, and Y.-S. Chen. Tweolocator: a non-intrusive geographical locator system for twitter. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Location-Based Social Networks*. ACM, 2012.

[3] B. Han, P. Cook, and T. Baldwin. A stacking-based approach to twitter user geolocation prediction. In *ACL (Conference System Demonstrations)*, pages 7–12, 2013.

[4] J. Mahmud, J. Nichols, and C. Drews. Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2014.