Using Geolocated Tweets for Characterization of Portuguese Administrative Regions

Gaspar Brogueira
INESC-ID, Lisboa, Portugal
ISCTE-IUL – Instituto Universitário de Lisboa, Portugal
gmrba@iscte.pt

Fernando Batista
INESC-ID, Lisboa, Portugal
ISCTE-IUL – Instituto Universitário de Lisboa, Portugal
fernando.batista@inesc-id.pt

Joao Paulo Carvalho
INESC-ID, Lisboa, Portugal
Instituto Superior Técnico - Universidade de Lisboa, Portugal
joao.carvalho@inesc-id.pt

Abstract

This paper uses a database of about 18 Million Portuguese geolocated tweets, produced in Portugal during a ten-month period to extract indicators about the distinct Portuguese regions and their population. This paper analyses the part of the Portuguese Twitter community and reveals that it is possible to extract relevant indicators, namely: the daily periods of increased activity per region; and predict regions where the concentration of the population is higher or lower in certain periods of the year. Such information can be useful for distinct studies in distinct areas, such as: marketing studies for the planning of launch schedules for advertising campaigns in times of the day when there is a bigger chance of a broader audience reach in social networks; and public health issues, for instance in the prediction of disease outbreaks, by observing population clusters during certain periods of the year in certain regions.

Keywords: Portuguese Tweets, Geolocated Tweets, Twitter API, MongoDB

1 Introduction

The information published by millions users in social networks such as Twitter, is an important source of knowledge, that could lead to academic, socio-economic or demographic studies (distribution of male and female population, age, marital status, birth), lifestyle analysis (interests, hobbies, social habits) or study online behavior (time spent online, interaction with friends or discussion about brands, products or politics).

The change of social interactions paradigm, due to spread of social networks, allows access to more data and additional information than traditional methods such as surveys, interviews or researches, as well as a greater interval time between the taking of samples [7].

Twitter is a microblogging service, with about 255 million active users that publish about 500 million messages per day, limited to 140 characters (tweets). The freedom to share thoughts, opinions, feelings or news about various subjects, makes the volume of information present on Twitter, potentially interesting for several studies in diverse areas such as policy [3], tourism, marketing or health [11], [12].

The Twitter API provides a limited access to the total volume of produced tweets. For example, the Streaming API accesses in real time to a continuous stream of tweets that, depending on the level of used permissions authentication [1], corresponds to 1% to 10% of the total tweets produced at a given time. Alternative APIs limit the access in other ways.

This paper uses a database of geolocated tweets, produced in Portugal and written in European Portuguese, and previously created using several strategies for overcoming some of the Twitter API limits. Such database covers a period of about 10 months and can provide insights about the Portuguese population, such as rate of Internet in using new technologies, population distribution, relationship between people, level of culture, literacy, and territorial mobility. The paper used information about a tweet's date and time and analyzes the distinct Portuguese regions terms of the number of tweets produced at a given period of the day or at a different period of the year.

The use of geolocated tweets is also reported by [2], who presents a study about the consumption of healthy and unhealthy foods by the US population. Tweets with known location are also used by [4] and [5] for real-time information on the most relevant topics covered by users, by conducting a review of feelings indicating if a discussed topic is positive or negative. A methodology by which it's possible to discover the occurrence of a relevant event in a certain place, by collecting and analyzing geolocated tweets is proposed by [6]. Another recent and interesting study uses two years of geolocated data from Twitter to track trends in migration patterns [13]. The paper shows that publicly available geolocated tweets by themselves can help to understand the relationships between internal and external migration. Other related work includes a method presented by [14] to predict the particular user location, based on the user's followers. An analysis on how geolocated information coming from cellular data can help monitoring and mapping spatial and temporal variability of population in a specific region can also be found in [15].

The remainder of this paper is organized as follows: Section 2 describes the process of collecting and processing geolocated tweets and describes the dataset. Section 3 presents the analysis of the data. Section 4 presents major conclusions and prospects for future work.

2 Data Acquisition and Processing

The dataset used in the scope of this paper was collected between February 20 and December 31, 2014, using the Streaming API *filter/status*, and considering only the collection of tweets produced in Portugal and written in Portuguese. The data was restricted to the geographic limits corresponding to the Portuguese mainland and also the Autonomous Regions of Azores and Madeira. Additionally, the tweets were also restricted to those in which the language field *lang*, automatically assigned by Twitter, contains the

value 'pt' and the *place.country* field contains the value "Portugal." Considering the above restrictions, the existing collection contains about 18.4M tweets and is stored in a large MongoDB database.

The information about each published tweet contains, not only the message, but also the author's information and location at the time of the post. The analysis performed in the scope of this paper aims at distinguishing between the 18 districts in which the Portuguese mainland is divided. Therefore, all results depend on how well one can assign the location where a given tweet was produced to the corresponding district. However, most of the times such information cannot be easily retrieved from the tweet. For that reason, the remainder of this section describes the approach used in assigning the district to the location where the tweet has been produced.

All geographically localized tweets contain the "place" field, which assembles a number of other fields and provides, as a whole, information about the geographical location where the tweet was produced. Such information can be found in the *place.name* and *place.full name* fields. In some cases, *place.full_name* contains, not only the location, but also the district to which the location belongs. For instance, with the value of the field "Lisboa, Lisboa", the first reference to "Lisboa" is the name of the city Lisbon and the second reference to "Lisboa" is the district name to which Lisbon belongs to, that is, in this case, the district of Lisbon.

The information was not kept consistent during the time the tweets were collected. For that reason, only about 8.37% of the tweets contain information about the district where the tweet was produced directly on the *place.full_name* field. To work around this problem, the district name can be obtained based on the name of the locality that can be found on the *place.name* field, by consulting the list of postal codes[1] provided by *CTT - Correios de Portugal SA*. The list contains, among other information, the association between the locality and the district, for all locations of the Portuguese mainland, Azores and Madeira. The information is stored as CSV (comma separated value) files, where each line contains 16 data fields separated by semicolons, including the following information: district code, county code, locality code, and locality name. The following example shows an example of an entry, where "01" corresponds to *Aveiro* district, "04" is the code of *Arouca* municipality, and 69893 is the code of *Picoto*, the corresponding location. The district and municipality codes are also available as separated files.

```
01;04;69893;Picoto;;;;;;;;;;;4540;205;AROUCA
```

In order to successfully apply the previously mentioned list, some of the tweets must be normalized. For example, sometimes the *place.name* field contains the value "Lisbona" or "Oporto" instead of "Lisboa" because names are written in different languages, other than Portuguese. Additionally, some cases of spelling mistakes in the location name can also be found, such as "*Setubal*" or "Guimaraes" that correspond to *Setúbal* and *Guimarães*, respectively.

[1] http://www.ctt.pt/feapl_2/app/restricted/postalCodeSearch/postalCodeDownloadFiles.jspx (visited in 21-01-2015)

The above-described method works well for about 99.69% of the tweets. However, we could not infer the district of about 0.31% of the tweets, mostly because the field *place.name* in these tweets contain a generic value, such as "Portugal". An almost negligible amount of tweets contain values that don't match a name of a Portuguese locality. For example, names of restaurants: "Restaurante Zé Pinto" or "Casa dos Presuntos". Finally, we have also found several cases of Portuguese localities with identical names in different districts. One example is "*Covilhã*", a city in the *Castelo Branco* district, whose name is also associated with a locality in *Porto* district and another one in *Braga*. Another example is the city "*Seixal*", which is also a city in *Setúbal* district, has a namesake locality in *Aveiro* district. These examples are shown in Table 1.

Table 1: Examples of localities sharing the same name.

```
03;13;53346;Covilhã;;;;;;;;;;;4730;490;
  SANTIAGO CARREIRAS
05;03;14718;Covilhã;1000305;Rua;dos;;;
  Barreiros;Vila do Carvalho;;;;6200;224;
  COVILHÃ
13;01;4000;Covilhã;;;;;;;;;;;4600;757;
  TELÕES AMT

01;04;60744;Seixal;;;;;;;;;;;4540;497;
  ROSSAS ARC
15;10;43887;Seixal;200101015;Rua;;;;
  Silvana Alves Cunha;;;;;2840;471;SEIXAL
```

We have disambiguated these cases by considering the largest area of each one of the possible localities, i.e., by choosing the locality with the largest area.

3 Data Analysis

This paper considers the Portuguese mainland divided in 18 districts: Aveiro, Beja, Braga, Bragança, Castelo Branco, Coimbra, Évora, Faro, Guarda, Leiria, Lisboa, Portalegre, Porto, Santarém, Setúbal, Viana do Castelo, Vila Real, Viseu. Azores and Madeira were not considered in this study for simplification. Alternatively, we could have considered dividing the country in administrative units, known as Nomenclature of Territorial Units for Statistics (NUTS), also used in the literature [8]. This paper is a first attempt to use tweets for characterizing the Portuguese regions. Future extensions of this work will consider this alternative division.

According to "Censos 2011" [9], the Portuguese population is not equally distributed in the Portuguese territory. In fact, a sharp desertification is noticed in large areas of interior and a high population density can be found on the coast and metropolitan areas, in particular Lisbon and Oporto. Censos 2011 also refers to the distribution of young and elderly population: the coastline contains a superior percentage of young people. The situation is reversed in relation to the elderly population, and the percentage of elderly population in the interior is higher than the percentage of elderly population on the coast of Portugal.

Figure 1: Distribution of collected tweets and tweets authors.

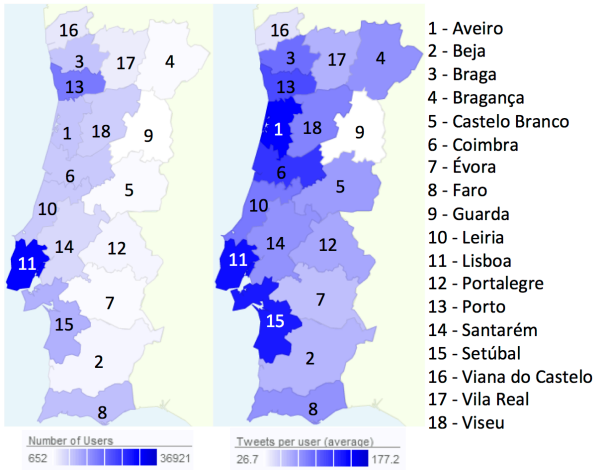


Figure 1 represents the distribution of Twitter users together with their activity for the Portuguese territory, based on our database of tweets. The left map shows the distribution of the users, verifying a larger number of users mainly in coastal districts of Portugal, particularly in *Lisboa* (~37K), *Porto* (~20K) and *Faro* (~10k). *Faro* has a high number of users, due the influx of population for this district in the summer holiday period. The map on the right shows that the most active users concentrate in the districts of *Aveiro*, *Lisboa,* and *Setúbal*. The Portuguese Twitter community is essentially composed of teenagers or young adults [10], which, given the highest percentage of young people on the coast of Portugal, partly explains the increased volume of tweets collected in coastal districts as well as increased user activity, also located in the coastal districts. A large production of tweets can be found in *Lisboa* (~6.1M), followed by *Porto* (~2.5M) and *Setúbal* (~1.9M), which is consistent with the distribution of population density of Portugal [9]. The study presented by [9] states that the coast contains a greater proportion of workers with working hours exceeding 45 weekly hours, particularly in the metropolitan areas of *Lisboa* and *Porto*, which may be related to the lower activity on Twitter halfway through day, that is, during the lunch hour (12h-14h) and the increased production of tweets during the night (18h-24h) in the coastal districts.

The economic activity carried out on the interior of country, mainly related to the primary sector, associated with the lower development of this region of the country, provides the populations to earlier sleep and there is less tweets production activity during the night, as occurs in the graph of Figure 2. Another possible explanation concerns the difference of habits in large and smaller cities. For example, people from big cities may be more prone to engage in night activities, and because big cities are mostly located along the coast, it turns out that Twitter activity during the night becomes more prominent along the coast.

Another division of the districts of Portugal that is common to observe, is the division in the North, Center and South. The North region includes the districts of *Aveiro*, *Braga*, *Bragança*, *Guarda*, *Porto*, *Viana do Castelo*, *Vila Real*, and *Viseu*; the Center region contains the districts of *Castelo Branco*, *Coimbra*, *Leiria*, *Lisbon*, *Portalegre* and *Santarém*; and the South contains districts of *Beja*, *Évora*, *Faro*, and *Setúbal*. Figure 3 shows the usage of Twitter throughout the hours of a day in these three regions, revealing that the production of tweets during the night period (6pm to 12am) is higher at the north region, whilst the southern region has lower activity during the same period, by contrast with noon, in which the southern region has a higher tweet activity.

Figure 2: Activity during a day for the *interior* and *coast.*

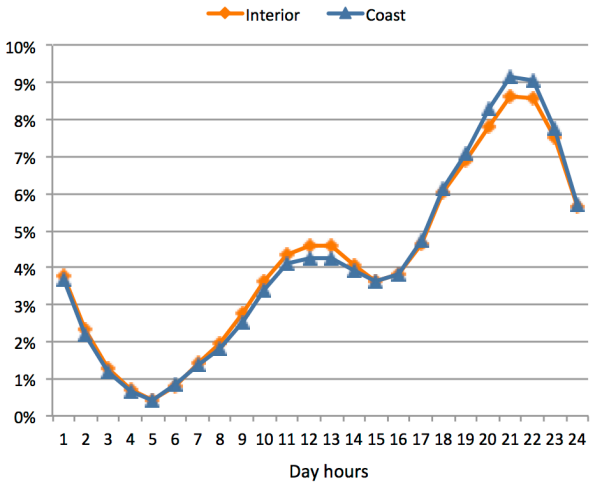
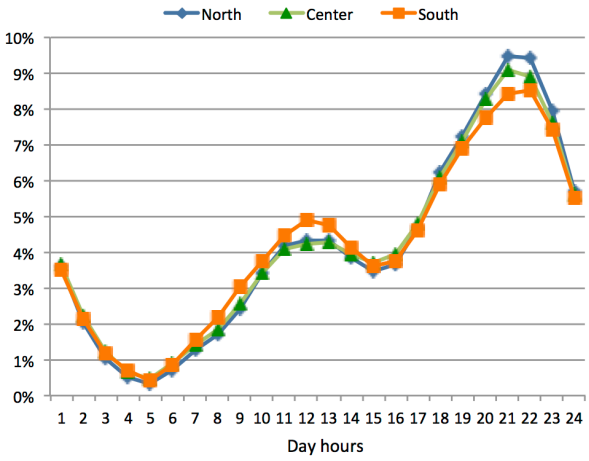


Figure 3: Daily activity for the *North, Center* and *South.*

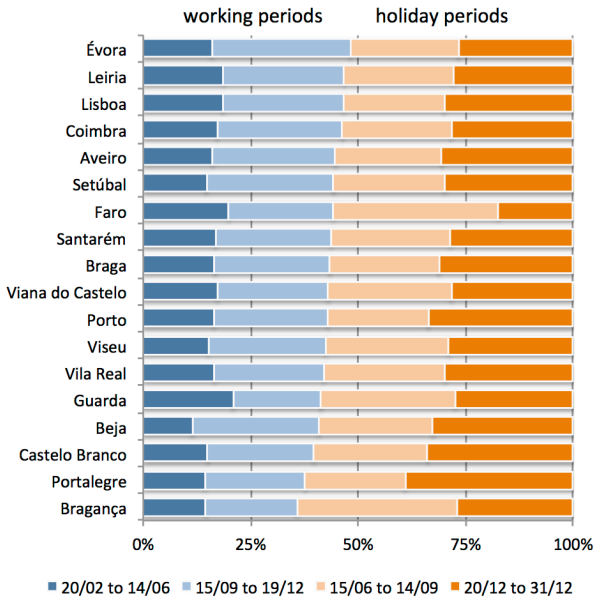


This type of information could prove to be quite useful, for example, to launch advertising campaigns or for the propagation of news, targeted to the interests of the population in each region. By knowing the time of the day when the target audience in a certain region is more active on Twitter, the information can be propagated more effectively and viewed by a larger number of potential customers.

The following set analyses consider the period of the year when the tweets were produced, and relate the user activity with working vs. holiday periods. The data was divided into four periods of work and holidays. Were considered as working periods the intervals between February 20th and June 14th, and September 15th to December 19th. The summer holiday period was considered from June 15th to September

14th and the Christmas holiday period and New Year's Eve from December 20th to 31st. Figure 4 summarizes the activity in those periods, revealing and generalized increased activity during the holidays.

Figure 4: Activity per day in work and holiday periods.



The summer holiday period is especially active in *Faro*, location elected by a considerable number of Portuguese people for holidays, due beaches and good weather. The districts of *Bragança* and *Guarda* are also particularly active during this period mostly because of the many immigrants who return to Portugal in this time of year to come together with their families. The Christmas holiday period is the most active in most regions, being particularly active in *Portalegre*, *Castelo Branco*, *Porto*, and *Beja*. The only exception concerns to *Faro* that, in contrast to the summer holiday period, is the one with lower production of tweets during the Christmas holiday. Finally, Figure 4 shows that the most active working period goes from September to December, which corresponds to the beginning of the school. This behavior is certainly correlated with the fact that most of the users in our database are young people that use Twitter as a way to send messages to the school colleagues.

Our final analysis relates the daily activity with the different working and holiday periods, and is illustrated in Figures 5 and 6. Figure 5 shows the daily activity corresponding to working periods, whereas Figure 6 shows the corresponding activity for holiday periods.

During working periods the activity distribution is quite similar for the two periods considered, following a similar shape. The activity peaks are around 21h and between 11h to 13h. The lowest activity corresponds to 4h and 5h.

In what concerns to the holiday periods, they also follow a similar shape too, but differences are more notorious. During the day, the activity peak is before noon during the summer and after noon during Christmas time. The activity goes over the night and achieves its lower bound around 6 a.m.

The two figures reveal that working periods and holiday periods are in fact quite distinct in terms of user activity. One of the most notorious differences is the extended activity during the night in holiday periods, in opposition to the activity peek that can be found around 21h during working periods.

Figure 5: Activity per day in working periods.

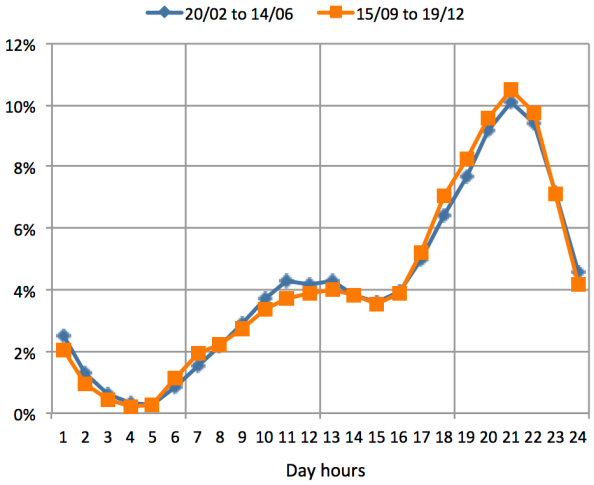
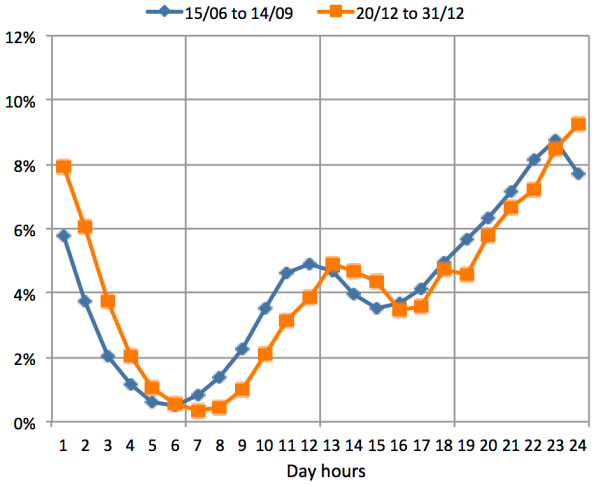


Figure 6: Activity per day in holiday periods.



4 Conclusions and Future Work

This paper presents an analysis over a database of about 18 Million Portuguese geolocated tweets, produced in Portugal during a ten-month period. By observing the Twitter usage by the Portuguese community this paper reveals that it is possible to extract relevant indicators, such as: the daily periods of increased activity, predict regions where the concentration of the population will be higher or lower in certain periods of the year. Such information could prove useful for distinct areas, such as: marketing studies for the planning of launch schedules for advertising campaigns in times of the day when there is a bigger chance of a broader audience reach in social

networks; and public health issues, for instance in the prediction of disease outbreaks, by observing population clusters during certain periods of the year in certain regions.

This paper is a first step in understanding the idiosyncrasies of each one of the Portuguese regions in daily-based or yearly-based periods. This study will be further extended in order to better characterize each one of the regions in terms of daily habits, user profiles, and also in order to better understand the way people travel between regions.

Acknowledgements

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) under project PTDC/IVC-ESCT/4919/2012 (MISNIS) and funds with reference UID/CEC/50021/2013.

References

[1] S. Kumar, F. Morstatter and H. Liu. *Twitter Data Analytics*. Springer. 2014.

[2] M. J. Widenera and W. Li. *Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US*. Applied Geography, vol. 54, pages 189 – 197, 2014.

[3] S. Rill, D. Reinel, J. Scheidt and R. V. Zicari. *PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis.* Knowledge-Based Systems, vol. 69, pages24-33, 2014.

[4] Saravanan M, D. Sundar and Kumaresh V. S. *Probing of geospatial stream data to report disorientation*. IEEE Recent Advances in Intelligent Computational Systems (RAICS), 2013.

[5] H. Kim, S. Lee and S. Kyeong. *Discovering Hot Topics using Twitter Streaming Data*. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2013.

[6] T. Kim, G. Huerta-Canepa, J. Park, S. J. Hyun and D. Lee. *What's Happening: Finding Spontaneous User Clusters Nearby Using Twitter*. IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing 2011

[7] W. Housley, R. Procter, A. Edwards, P. Burnap, M. Williams, L. Sloan, O. Rana, J. Morgan, A. Voss and A. Greenhill. *Big and broad social data and the sociological imagination: A collaborative response.* Big Data & Society, pages 1–15, 2014.

[8] Diário da República Portuguesa. *Decreto-Lei n.º 46/89*, pages 590 – 594, 15 Fevereiro 1989.

[9] Instituto Nacional de Estatística. *Censos 2011 Resultados Definitivos – Portuga*l, 2011.

[10] G. Brogueira, F. Batista, J. P. Carvalho and H. Moniz. *Portuguese geolocated tweets: An overview*. In Proceedings of the International Conference on Information Systems and Design of Communication, ISDOC, pages 178-179. ACM, 2014.

[11] A. Culotta. *Estimating County Health Statistics with Twitter.* In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Pages 1335-1344, ACM, 2014.

[12] J. C. Santos and S. Matos. *Predicting flu incidence from* Portuguese *tweets*. In International Work-Conference on Bioinformatics and Biomedical Engineering - IWBBIO, pages 11–18, 2013.

[13] E. Zagheni, K. Garimella, B. State and I. Weber. *Inferring international and internal migration patterns from Twitter data.* Proceedings of the 23rd International Conference on WWW '14 Companion, Seoul, Korea, 2014

[14] A. Culotta, N. Ravi, and J. Cutler. *Predicting the demographics of Twitter users from social evidence using website traffic data*. 29th AAAI Conference on Artificial Intelligence (AAAI-15), 2015

[15] F. Manfredini, P. Tagliolato and C. Di Rosa. *Monitoring Temporary Populations through Cellular Core Network Data*. In Computational Science and Its Applications - ICCSA 2011. Lecture Notes in Computer Science. Volume 6783, pp. 151—161. Springer Berlin Heidelberg. 2011.