Social Network Analysis and Mining

Using Geolocated Tweets for Characterization of Twitter in Portugal and the Portuguese Administrative Regions

--Manuscript Draft--

Manuscript Number:	
Full Title:	Using Geolocated Tweets for Characterization of Twitter in Portugal and the Portuguese Administrative Regions
Article Type:	Original Article
Corresponding Author:	Joao Paulo Carvalho, Ph.D. INESC-ID / Instituto Superior Tecnico, Universidade de Lisboa Lisboa, PORTUGAL
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	INESC-ID / Instituto Superior Tecnico, Universidade de Lisboa
Corresponding Author's Secondary Institution:	
First Author:	Gaspar Brogueira, MsC
First Author Secondary Information:	
Order of Authors:	Gaspar Brogueira, MsC
	Fernando Batista, PhD
	Joao Paulo Carvalho, Ph.D.
Order of Authors Secondary Information:	
Funding Information:	Fundação para a Ciência e Tecnologia (PTDC/IVC-ESCT/4919/2012) — Not applicable
Abstract:	The information published by the millions of public social networks users is an important source of knowledge that can be used in academic, socio-economic or demographic studies (distribution of male and female population, age, marital status, birth), lifestyle analysis (interests, hobbies, social habits) or be used to study online behaviour (time spent online, interaction with friends or discussion about brands, products or politics). This work uses a database of about 27 Million Portuguese geolocated tweets, produced in Portugal by 97.8K users during a 1 year period, to extract information about Portugal, the distinct Portuguese regions, and respective habitants. We analyze the behavior of an important part of the Portuguese Twitter community and show that with this information it is possible to extract indicators such as: the daily periods of increased activity per region; prediction of regions where the concentration of the population is higher or lower in certain periods of the year; how do regional habitants feel about life; or what is talked about in each region. To the best of our knowledge, this work is the first study that shows Portuguese users' behaviour in Twitter focusing on geographic regional use.
Suggested Reviewers:	Zhiyuan Cheng Texas A&M University zcheng@cse.tamu.edu He is the author of many relevant papers within this subject: - Who is the barbecue king of texas?: a geo-spatial approach to finding local experts on twitter - Finding local experts on twitter - Location prediction in social media based on tie strength - Spatio-temporal dynamics of online memes: a study of geo-tagged tweets
	Aron Culotta Illinois Institute of Technology College of Science aculotta@iit.edu

	Several recent papers related with this topic: (2015) - A demographic and sentiment analysis of e-cigarette messages on Twitter (2015) - Predicting the Demographics of Twitter Users from Website Traffic Data (2014) - Using county demographics to infer attributes of Twitter users
	Michael J. Widener University of Toronto michael.widener@uc.edu He is an active researcher in this field, having authored papers such as: - Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US - Visualizing Demographics Based on Micro-blogger Tracking: Twitter Case Study

Using Geolocated Tweets for Characterization of Twitter in Portugal and the 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

The information published by the millions of public social networks users is an important source of knowledge that can be used in academic, socio-economic or demographic studies (distribution of male and female population, age, marital status, birth), lifestyle analysis (interests, hobbies, social habits) or be used to study online behaviour (time spent online, interaction with friends or discussion about brands, products or politics). This work uses a database of about 27 Million Portuguese geolocated tweets, produced in Portugal by 97.8K users during a 1 year period, to extract information about Portugal, the distinct Portuguese regions, and respective habitants. We analyze the behavior of an important part of the Portuguese Twitter community and show that with this information it is possible to extract indicators such as: the daily periods of increased activity per region; prediction of regions where the concentration of the population is higher or lower in certain periods of the year; how do regional habitants feel about life; or what is talked about in each region. To the best of our knowledge, this work is the first study that shows Portuguese users' behaviour in Twitter focusing on geographic regional use.

Keywords: Twitter; Geolocated Tweets; Portuguese tweets; Portuguese Districts; Twitter Data Analysis.

1 Introduction

The change of the social interaction 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 improves the interval time between sampling (Housley *et al*, 2014). Channels for expressing opinions have been increasing on a regular basis, and when these opinions are relevant to a company, they are important sources of business insight, whether they represent critical intelligence about customers, the impact of an influential reviewer on others purchase decisions, or early feedback on product releases, company news or competitors. Capturing and analysing these opinions is a necessity for proactive product planning, marketing and customer service, and it is also critical in maintaining brand integrity. The importance of harnessing opinion is growing as consumers use technologies such as Twitter to express their views directly to other consumers (Kalarikkal and Remya, 2015). Recently several studies have been developed that aim to characterize the communities formed in social networks by analysing the use of certain symbols or own conventions of writing small messages on Twitter to disseminate information, to express emotions or to make a topic a major issue in social network. The study presented by (Honeycutt and Herring, 2009) focused on the importance of using the @ symbol in British tweets as a form of direct interaction

with other users. (Hong *et al*, 2011) compared the behaviour and use of Twitter by various communities identified in a corpus made up of 62M tweets. This work identified the top 10 most used languages on Twitter, extended Honeycutt's work (Honeycutt and Herring, 2009), looked for patterns of sharing URLs, hashtags, mentions, replies and retweets. (Boyd *et al*, 2010) examined the dissemination of information on Twitter by retweeting action, and (Burnap *et al*, 2014) presented some models to predict information flow size and survival using data derived from Twitter by defining information flows as the propagation over time of information posted to Twitter via the retweeting action. (Bora *et al*, 2015) analysed the ability to predict the emergence of virality of a hashtag. (Java *et at*, 2006) identified several categories of intention to use Twitter including: i) daily chatter where users discuss events in their lives or their current thoughts; ii) sharing information or URLs and iii) reporting news which includes commenting on current events or automated news agents posting weather or news stories.

In addition to the knowledge and interpretation of behavior in social networks, many such applications could benefit from information about the location of users, but only less than 1% of tweets are geotagged and information available from the location fields in users' profiles is scarce. (Mahmud *et al*, 2014) present some algorithms to predict the home location at the city-level of Twitter users from the content of their tweets and their tweeting behavior. They also examined the possibility of predicting at other larger levels of granularity, such as state, time zone and geographic region. Other authors built geographic topic models to predict the location of Twitter users in terms of regions and states (Eisenstein *et al*, 2010). (Hecht *et al*, 2011) built Bayesian probabilistic models from words in tweets for estimating the country and state-level location of Twitter users. They used location information submitted by users in their Twitter profiles, resolved via the Google Geolocation API, to form the ground-truth of a statistical model for location estimation. (Cheng *et al*, 2010) describe a city-level location estimation algorithm, which is based on identifying *local words* from tweets and building statistical predictive models from them, but their method requires a manual selection of such *local words* for training a supervised classification model. (Chandra *et al*, 2011) described location estimation using the conversation relationship of Twitter users in addition to the text content used in the conversation. (Chang *et al*, 2011) described yet another content based location detection method using Gaussian Mixture Model (GMM) and the Maximum Likelihood Estimation (MLE). Their method also eliminates noisy data from tweet content using the notion of non-localness and geometric-localness.

Capturing human movement patterns across political borders is difficult. (Blanford *et al*, 2015) analyzed 10 months of geo-located tweets for Kenya and studied movement of people at different temporal (daily to periodic) and spatial (local, national to international) scales. The use of geolocated tweets is also reported in (Widenera and Li, 2014) in order to present a study about the consumption of healthy and unhealthy foods by the US population. Tweets with known location are also used by (Saravanan *et al*, 2013) and (Kim *et al*, 2013) 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 is possible to discover the occurrence of a relevant event in a certain place, by collecting and analyzing geo-located tweets is proposed by (Kim *et al*, 2011). Another recent and interesting study uses two years of geo-located data from Twitter to track trends in migration patterns (Zagheni *et al*, 2014). The paper shows that publicly available geo-located tweets can, without additional

information, help to understand the relationships between internal and external migration. Other related work includes a method presented by (Culotta *et al*, 2015) to predict the particular user location, based on the user's followers. An analysis on how geo-located 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 (Manfredini *et al*, 2011).

Most of the above-referred studies base their analysis on tweets published in English (Hong *et al*, 2011). To the best of our knowledge, our work is the first to systematically study how Portuguese users behave in Twitter, more specifically in each of the Portuguese districts. This work uses a database of geo-located tweets, produced in Portugal during a one-year period, and written in European Portuguese. The database was created using several strategies for overcoming some of the Twitter API limits (Brogueira *et al*, 2015). We use information about a tweet's date and time to analyze the distinct Portuguese regions in terms of the number of tweets produced at a given period of time (day, season, etc.) and provide insights about the Portuguese population, such as rate of Internet in using new technologies, population distribution, main interests, or territorial mobility.

Our research is motivated by the attempt to characterize the use of Twitter in Portugal and results in the following main contributions:

- An algorithm for predicting the Portuguese districts where each tweet was published;
- Pattern identification about Twitter usage during work periods and holiday periods;
- Characterization of Twitter usage by the Portuguese community (based on hashtag, mention, retweet, replies and URL use);
- Characterization of district sentiment based on the analysis of the frequency of emoticons.

The remainder of this paper is organized as follows: Section 2 describes Twitter's conventions. Section 3 presents the methodology for data acquisition and processing. Section 4 presents the temporal and geographical analysis of collected data. Section 5 addresses topics and results concerning the characterization of Twitter usage in Portugal. Finally, Section 6 presents the major conclusions and prospects for future work.

2 Background

2.1 Twitter

Twitter is a microblogging service based on short messages limited to 140 characters called tweets. Twitter has currently around 255 million active users that publish about 500 million messages per day. Users can indicate whether they wish their tweets to be public (the messages appear in reverse chronological order on the "public timeline" on Twitter home page) or private (only those who have subscribed to the user's feed, "followers", are able to see the messages). Tweets can be posted via Twitter.com, text messaging, instant messaging or from third party clients (Facebook, Instagram, etc.) (Honeycutt and Herring, 2009). The freedom to share thoughts and opinions about different aspects of daily life, 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 (Rill *et al*, 2014), tourism, marketing or health (Culotta, 2014; Santos and Matos, 2013). Much of that data is public and available for

mining and it has fueled the rapid growth of consumer-generated content such as consumer satisfaction, opinion extraction, ratings and sentiment analysis (Kalarikkal and Remya, 2015). Further, research suggests that the online purchase intent is significantly impacted by negative/positive sentiments found online. For example, Mittal *et al*. used twitter data to understand public mood and use the predicted mood values to infer the stock market movements (Mittal and Goel, 2011).

The Twitter API provides a limited access to the total volume of produced tweets. For example, the Streaming API accesses in real time a continuous stream of tweets that, depending on the level of used permissions authentication (Kumar *et al*, 2014), corresponds to 1% of the total tweets produced at a given time. Alternative APIs limit the access in other ways.

2.2 Categories of Users and Intention to use Twitter

Twitter is used for purposes as diverse as: i) sharing ideas and thoughts; ii) information dissemination and news; iii) communication or conversation with friends. (Java *et al*, 2006) identified three main categories of Twitter users: information sources, friends and information seekers. Information sources post news and tend to have a large base of "followers". These sources may be individuals or automated services. Friends are a broad category that encompasses most users, including family, co-workers and strangers. Information seekers tend to be users who may post rarely but who follow others regularly. As previously mentioned, (Java *et al*, 2006) also identified several categories of intention to use Twitter, including: (1) daily chatter, where users discuss events in their lives or their current thoughts; (2) sharing information or URLs; (3) reporting news, which includes commenting on current events or automated news agents posting weather or news stories. Other category of user intention is conversation (Java *et al*, 2006), a quite popular use within the Twitter Portuguese community, as we reveal later in this article.

2.3 Twitter conventions

Each category of users or type of Twitter usage is characterized by the use of certain Twitter symbols. Twitter users can refer to a specific user by including a mention anywhere in their tweets, done in the form of *@username*, where *username* is the mentioned user's *screen_name*. The information about all mentions contained in a single tweet is presented in the field *entities.mentions*. A reply to a previous tweet is a specific form of mention, with the *@username* appearing at the beginning of the reply tweet. Retweeting is typically used to spread information received from followers to followers (Boyd *et al*, 2010). It is the equivalent of forwarding an email, and is an action to information sharing and a social act, recognizing and promoting someone's message. A common form of retweeting is "RT @username message", where "message" is a tweet created by "@username". Users have also adopted a variety of other syntactical markers such as "RT:@", "retweeting @", "retweet @", "(via @)", "RT (via @)", "thx @", "HT @", and "r @" (Boyd *et al*, 2010).

Hashtags are keywords included in tweets, in the form of *#keyword*. Including a hashtag creates a tag for a social bookmarking system and specifies a mechanism useful to collect tweets related to the given topic suggested by the keyword. The field *entities.hashtags* contains the information about all hashtags present within the message.

In order to share information, Twitter users can include URLs or links in their tweets. The information about any used URLs is presented in the field *entities.urls.*

3 Methodology

3.1 Data Acquisition and Processing

The dataset used in the scope of this paper was collected between September 15th, 2014 and September 14th, 2015, using the Streaming API *filter/status* and considering only the collection of tweets produced in Portugal and written in European Portuguese. The data was collected by a Python script that directly accesses the Twitter API and was restricted to the geographic limits corresponding to the Portuguese mainland and 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". The existing collection contains about 27.8M tweets produced by 97584 users, and is stored in a large MongoDB database.

The information about each published tweet contains not only the message, but also author information and location at the time of the post. One of the goals of this work concerns the analysis of each of the 20 administrative districts within Portugal mainland. Therefore, all results depend on how well we 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. The remainder of this section describes the approach used in assigning the district to the location where the tweet was produced.

3.2 District Inference by Locality Name

The information contained in each tweet has a flexible schema, and the data about the author and the location where it was produced is included in documents imbedded in the tweet structure. All geographically localized tweets contain the embedded document "*place*", which assembles a number of fields that 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 country to which the location belongs. For instance, with the value of the field "Lisboa, Portugal", the reference to "Lisboa" is the name of the city Lisbon and the reference to "Portugal" is the country name to which Lisbon belongs. Using the information found on the *place.full_name* field, we developed a method to obtain the district name based on the locality name (Fig. 1).

This method involves several steps and is based on the list of postal codes[1] provided by *CTT - Correios de Portugal SA* (Portugal's postal service). 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; locality name. Table 1 shows an example of such 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.

[1] http://www.ctt.pt/feapl_2/app/restricted/postalCodeSearch/postal CodeDownloadFiles.jspx (visited in 06-11-2015)

Table 1: Match of district and municipality codes.

```
01;04;69893;Picoto;;;;;;;;;;4540;205;AROUCA
04;05;23585;Argana;;;;;;;;;;5340;171;LAMALONGA
```

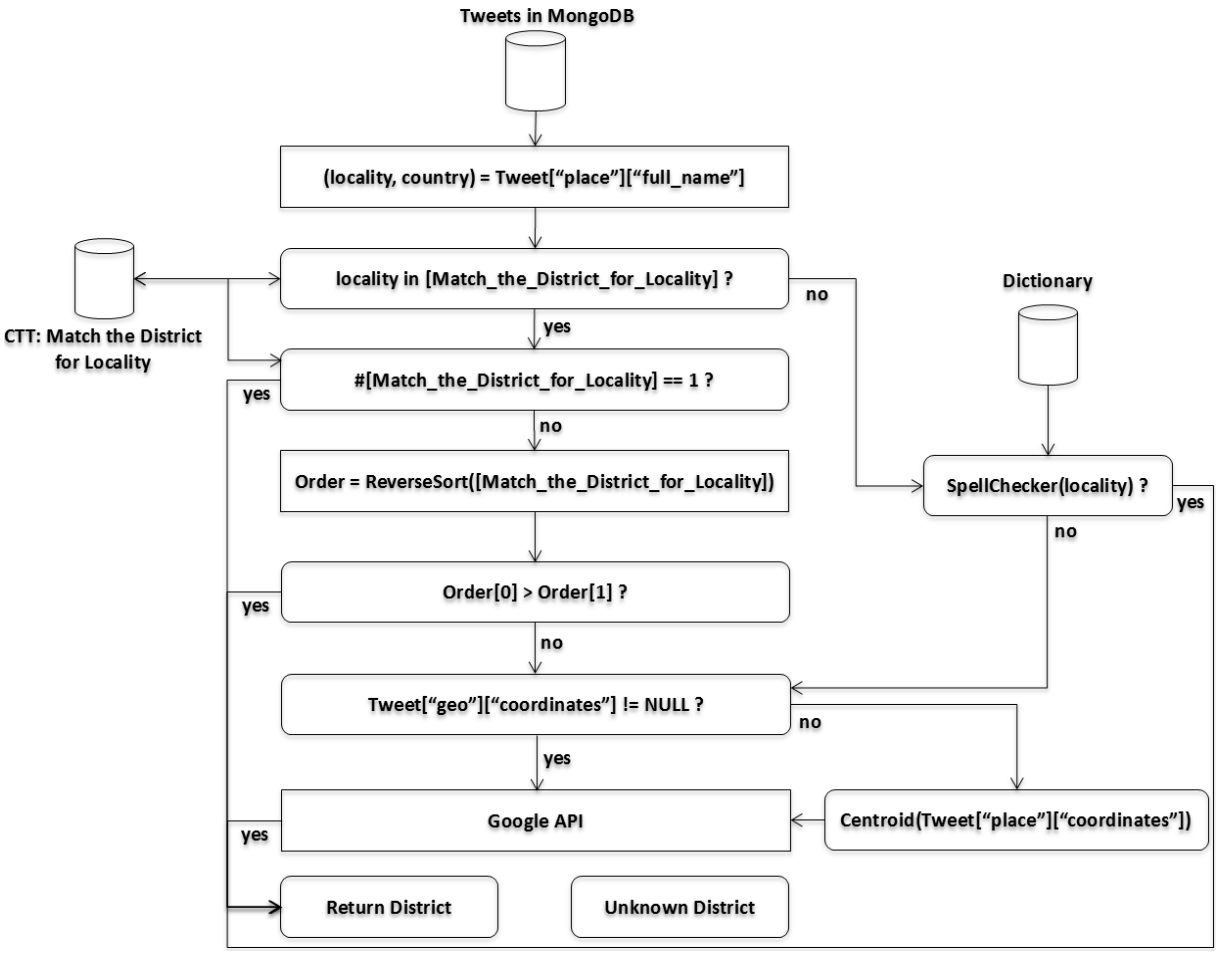


Fig. 1 Algorithm to infer the district of each locality.

The process begins by checking whether the name of the settlement indicated in the first value of the *place.full_name* field is included in the list of locations of *CTT - Correios de Portugal SA*. If not, we consider the possibility that the location name contains a typographical error, and we use a dictionary containing the most frequent errors in order find the probable matching district. For example, if the location shows the values of "Lisbon" or "lisbonne", we consider that the respective tweets are deemed to have been published in the town of Lisbon (and as such, in the district of Lisbon). Location name typographical errors occurred in around 2.93% of the database tweets. When the location name is correctly written and is present in the list of locations of *CTT - Correios de Portugal SA*, finding the corresponding district is a direct task unless the name occurs in different districts. However, 47% of the tweets refer locations names that exist in more than a single district. One example is

"Covilhã", a well-known (in Portugal) city in the Castelo Branco district, but whose name is also associated to two smaller localities in Porto and Braga districts. Another example is the city "Seixal", which is both a city in Setúbal district, and a much smaller locality in Aveiro district. These examples are shown in Table 2.

Table 2: 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
```

In 43% of the tweets, the relative size of the locations is substantially different, and in such cases we assume that the tweet is originated from the locality with the much larger area. To calculate the area of each place it was considered the number of matching streets, being considered that a particular locality has a much greater area if has a much larger number of streets listed in the database. If the location exists in more than 2 districts, one locality must be much larger than all others.

Table 3: Examples of localities sharing the same name.

- In districts Aveiro (01), Coimbra (06), Faro (08), Guarda (09), Leiria (10), Lisboa (11), Setubal (15) and Santarém (14), Viseu (18) and Madeira (31) exists one locality with name "Seixal":

```
01;04;60744;Seixal;;;;;;;;;;4540;497;ROSSAS ARC
06;05;18064;Seixal;;;;;;;;;;3090;651;FIGUEIRA DA FOZ
08;09;28536;Seixal;;;;;;;;;;8550;376;MONCHIQUE
09;06;59047;Seixal;;;;;;;;;;6290;310;GOUVEIA
10;02;2603;Seixal;;;;;;;;;;3250;168;ALVAIÁZERE
11;08;22369;Seixal;1023300000;Travessa;da;;;Igreja;;;;;2530;254;LOURINHÃ
15;10;43887;Seixal;200101015;Rua;;;;Silvana Alves Cunha;;;;;2840;471;SEIXAL
18;22;51785;Seixal;;;;;;;;;;3650;079;TOURO
31;06;60857;Seixal;;;;;;;;;;9270;133;SEIXAL PMZ
```

- Area of the locality "Seixal" in one of each district (District, Area) is:

  [('Setúbal', 90), ('Lisboa', 62), ('Aveiro', 16), ('Leiria', 2), ('Coimbra', 1), ('Faro', 1), ('Guarda', 1), ('Madeira', 1), ('Viseu', 1)]

- So the district of locality "Seixal" is "Setúbal"!

When the locations of the two largest values are not too dissimilar, we solve the ambiguity by taking into account the geographical coordinates present in the field *geo.coordinates* (if different from null). This is accomplished by invoking a service of the Google Maps API that given a pair of coordinates (latitude, longitude) lets you know which town and the district corresponds to the coordinate pair. Also for cases where the

place.full_name field contains the value of "Portugal" invoking the Google Maps API service allows you to get the name of the district, which otherwise would be impossible to determine. Table 4 shows an example of this invocation and the respective Google Maps API return.

Table 4: Using Google Maps API.

- Tweet with field *place.full_name* = Portugal

```
{
 "geo" :  { "coordinates" : [ 38.697843, -9.173279 ] },
 "place" : { "full_name" : "Portugal", … }
}
```

- Request Google Maps API

```
response ← http://maps.googleapis.com/maps/api/geocode/json?latlng=38.697843,-9.173279
response = {
            "status": "OK",
            "locality": Lisboa,
            "lat": 38.697843,
            "lng": -9.173279,
            "country": "Portugal",
            "district": "Lisboa"
            }
```

The reason why the disambiguation process starts by comparing locality area size instead of using the geographical coordinates lies in the fact that the Google Maps API imposes a maximum limit to the number of daily invocations from a given IP address. Due to the size of the used corpus, it would be impossible to use the API to determine the district for all ambiguous cases. Using our method, the Google Maps API is only invoked in 0.03% of the tweets (corresponding to the disambiguation of cases where the *place.full_name* field is not mentioned, or the tweet was posted in a location that can belong to different districts and have a similar areas).

The previous steps allowed for resolving the district in 99.32% of tweets. In the remaining 0.68% of the corpus, the *place.full_name* field did not contain any locality name and instead of precise coordinates it is presented a geographic. In such cases we used the center of the referred area and the Google Maps API to find the corresponding district.

4 Temporal and Geographical Data Analysis

The Portuguese mainland is divided into 18 districts: Aveiro, Beja, Braga, Bragança, Castelo Branco, Coimbra, Évora, Faro, Guarda, Leiria, Lisboa, Portalegre, Porto, Santarém, Setúbal, Viana do Castelo, Vila Real, Viseu. The Madeira archipelago is composed by Madeira and Porto Santo islands. The Azores archipelago consists of nine

islands: Santa Maria, São Miguel, Terceira, Graciosa, São Jorge, Pico, Faial, Corvo e Flores. For representation clarity and due to the population size, we grouped the islands into two districts corresponding to each archipelago: Madeira and Azores.

The Portuguese population is not equally distributed in the Portuguese territory (Instituto Nacional de Estatística, 2011). 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. Census 2011 (Instituto Nacional de Estatística, 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. As such, it is common to divide the territory into coastal and interior regions when performing data analysis (Fig. 3a).

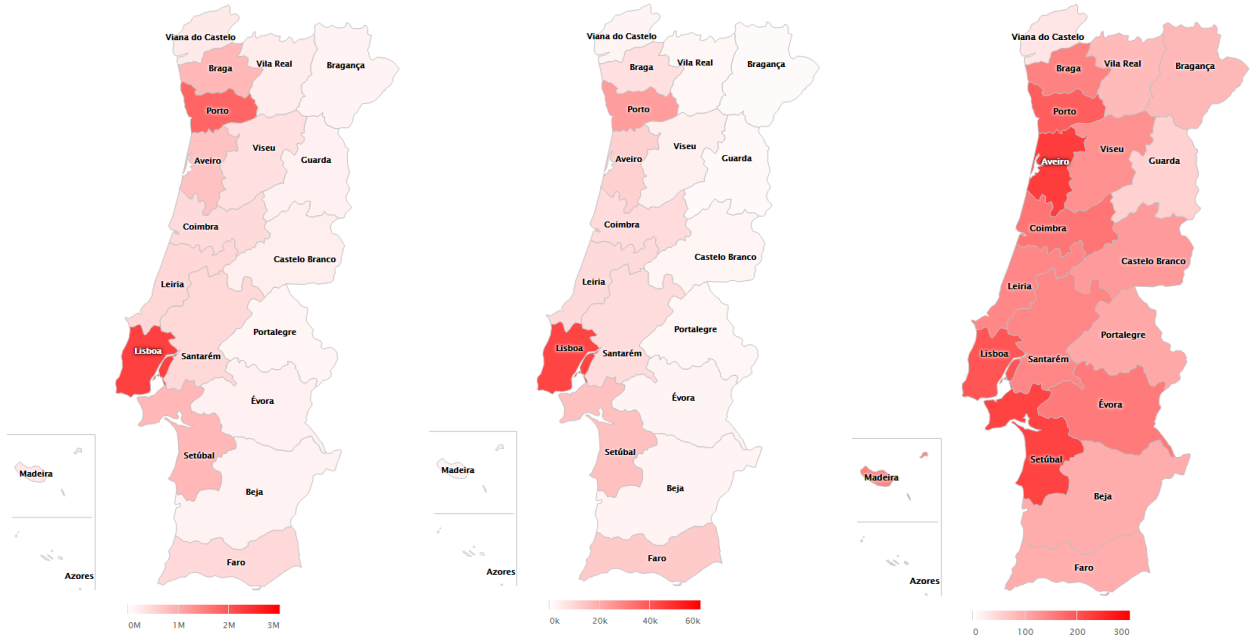
Fig. 2 shows the Portuguese population per district and the distribution of Twitter users together with their activity for the Portuguese territory, based on our database of tweets. It is clear a high correlation between population and Twitter use. Most users are active in the coastal districts of Portugal, particularly in *Lisboa* (~44K), *Porto* (~23K), *Setúbal* (~15K) and *Faro* (~12k). *Faro* district corresponds to the Algarve region, and has such, its number of users is inflated by influx of population during the summer holiday period. Map c) shows that the most active users are in the districts of *Aveiro*, *Setúbal*, and *Lisboa*.

The Portuguese Twitter community is essentially composed of teenagers or young adults (Brogueira *et al*, 2014), which given the highest percentage of young people on the coast of Portugal, partly explains the higher volume of tweets collected in coastal districts as well as the higher user activity. The largest tweet production occurs in *Lisboa* (~8.8M), followed by *Porto* (~4.3M) and *Setúbal* (~3.2M), which is consistent with the distribution of the Portuguese (Instituto Nacional de Estatística, 2011).

In addition to the coast/interior differences, it is also usual to look for regional differences between the North, the Center and the 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* (Fig. 3b).

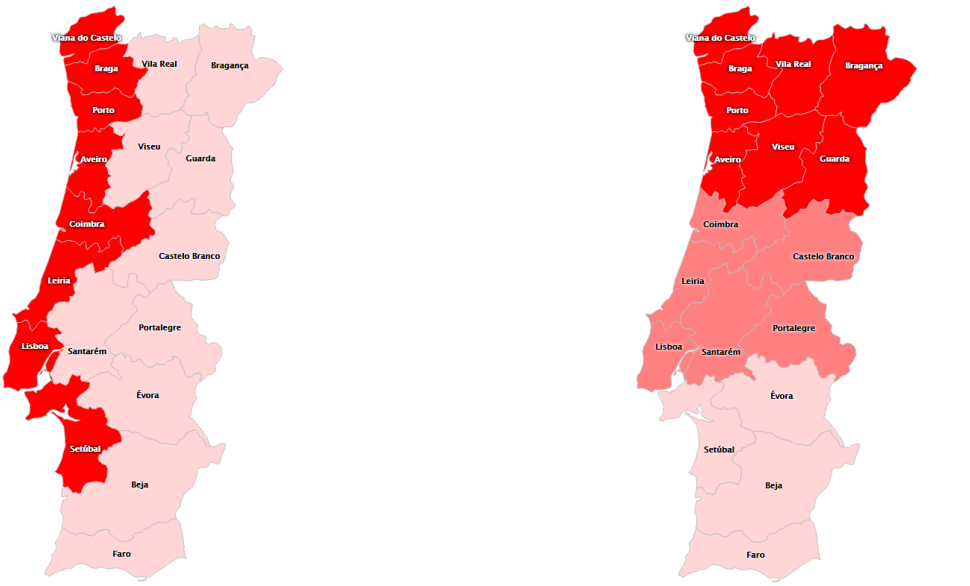
A significant part of the population usually moves away from the major urban areas to their homelands or to the leisure areas (mostly to Algarve). In order to account for any seasonal regional population flows we also decided to divide our analysis into eight working and holiday periods according to the 2014/2015 school calendar as defined by the Ministry of Education and Science of the Portugal Government[2]. The working periods consisted of: September 15th to December 16th, 2014; January 6th to February 15th, 2015; February 19th to March 20th, 2015; and April 7th to June 12th, 2015. The Christmas holiday period and New Year's Eve was considered from December 17th, 2014 to January 5th, 2015; the Carnival holiday period from February 16th to 18th, 2015; the Easter holiday period from March 21st to April 6th, 2015; and the Summer holiday period was considered from June 13th to September 14th, 2015.

[2] https://dre.pt/application/dir/pdf2sdip/2014/07/126000000/1728617289.pdf



a) Population in Census 2011[3]. b) Twitter users per district. c) Average tweets per user.

Fig. 2 Distribution per district of the Portuguese population, the Twitter users, and also the average number of tweets per user per day

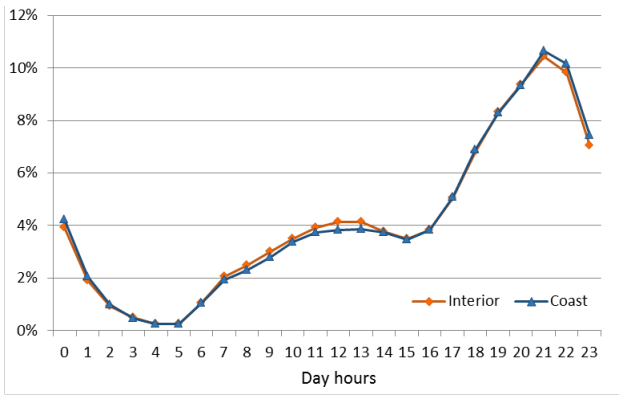


a) Coast (dark) and Interior (light) districts b) North, Center and South districts

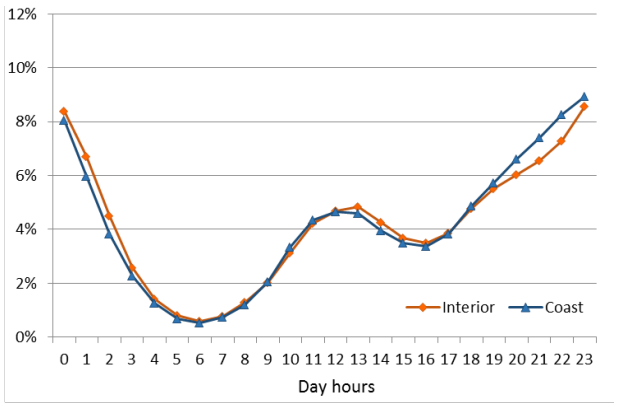
Fig. 3 Common divisions used for analysis of Portuguese regional data

[3] https://pt.wikipedia.org/wiki/XV_Recenseamento_Geral_da_População_de_Portugal

An analysis of daily Twitter usage between the different areas (Coast/Center; North/Centre/South) does not show any significant differences especially during working periods. However the activity profile is significantly different between work periods and holidays. Fig. 4 and Fig. 5 show respectively the Twitter activity throughout the day during work and holiday periods for Coast/Centre and North/Centre/South.



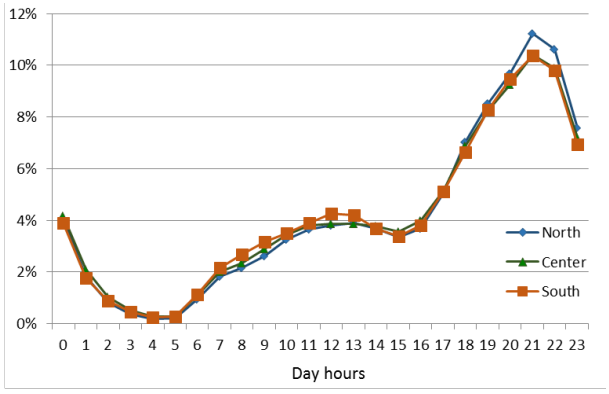
a) During a Work Period.



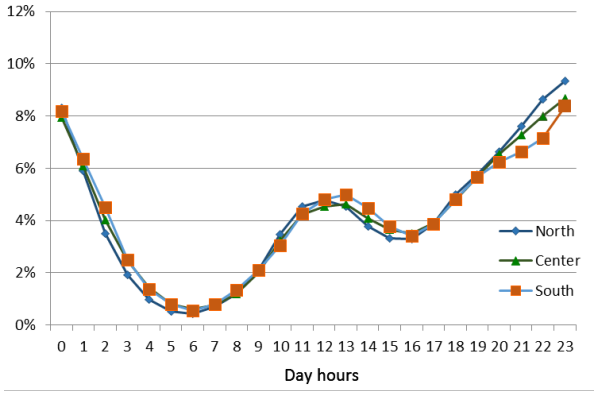
b) During Holidays.

Fig. 4 Daily activity in interior regions and coastal regions

During the work periods, Twitter activity distribution is quite similar throughout the country, without significant regional differences. The activity reaches a minimum between 3:00 and 5:00, has a constant growth rate between 7:00 and 12:00 (lunch break start), has a slight decay during and after lunch hours, and grows at a maximum rate between 16:00 and 21:00, when it reaches peak hour usage. Activity then decreases more or less constantly until the mentioned minimum activity hour.



a) During a Work Period.



b) During Holidays.

Fig. 5 Daily activity for the North, Center and South

Daily activity during holiday periods is quite different from the work periods, even if it is still possible to observe the lunch hour peak and the minimum and maximum activities occur roughly around the same time of day. One of the most notorious differences is the extended activity during the night period. Instead of peaking around 21:00-22:00 and rapidly decreasing, holiday activity extends throughout the night: activity until 2:00 in the morning is higher than lunch time activity; and the minimum activity period is between 6:00 and 7:00, i.e, one to two hours later than during working periods. The lunch time peak also occurs one hour later than during the work periods.

Regional differences are also more significant during holiday periods than during work periods, which indicate that, as expected, certain areas are more vacation oriented than others. Fig. 6 summarizes the activity in all working and holiday periods per district (data is not normalized by period duration). Fig. 7 shows the same information but aggregating all work periods and normalizing each period by its duration. The first conclusion to obtain from these charts is that twitter activity is lower during work periods than during holiday periods, and that the most prolific season for tweet production in Portugal is Carnival, a period of days largely associated with partying and celebration.

Regional differences are visible during the summer period. Regions that are common vacation destinations (such as Algarve, district of Faro) exhibit an expected activity increase. But the most interesting seasonal differences are seen in regions that receive a large influx of emigrants visiting their families (Guarda, Viana do Castelo, Azores). Such regions have an aging (and non-technological) population that is transformed with the seasonal arriving of the emigrants and their child. This results in a very noticeable Twitter activity increase. Note that the summer period lasts from mid June to mid September, but most vacationing emigrants stay only for a 1 month period. Their impact would be even more visible in that period.

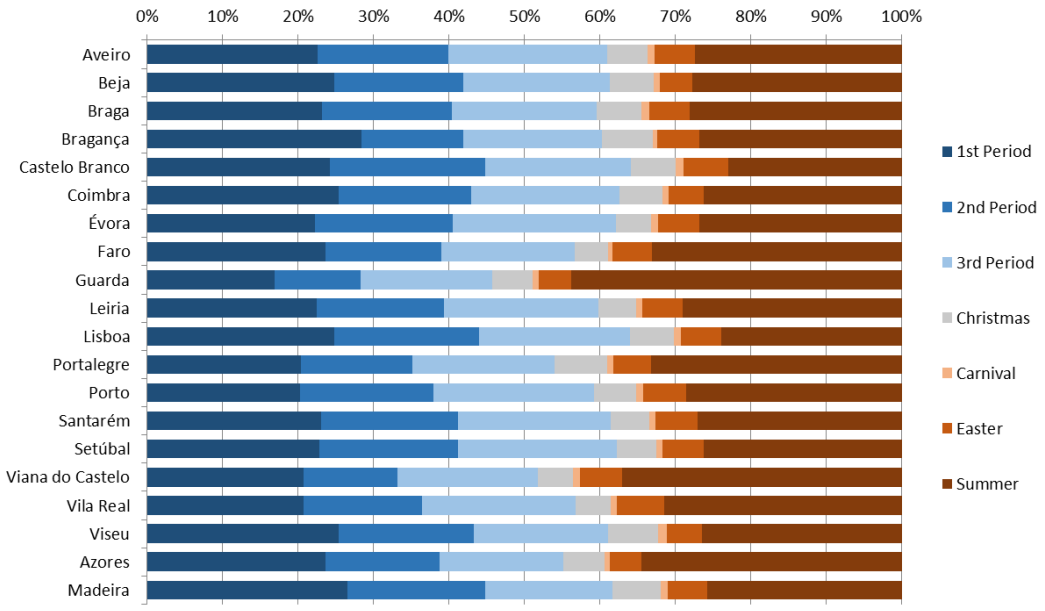


Fig. 6 Twitter activity during work and holiday periods.

The information presented in the seasonal charts can prove to be useful when deciding the timing of launching advertising campaigns, or for a more efficient dissemination of news targeted to the interests of the population in each region. The same is valid in what concerns the knowledge of 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.

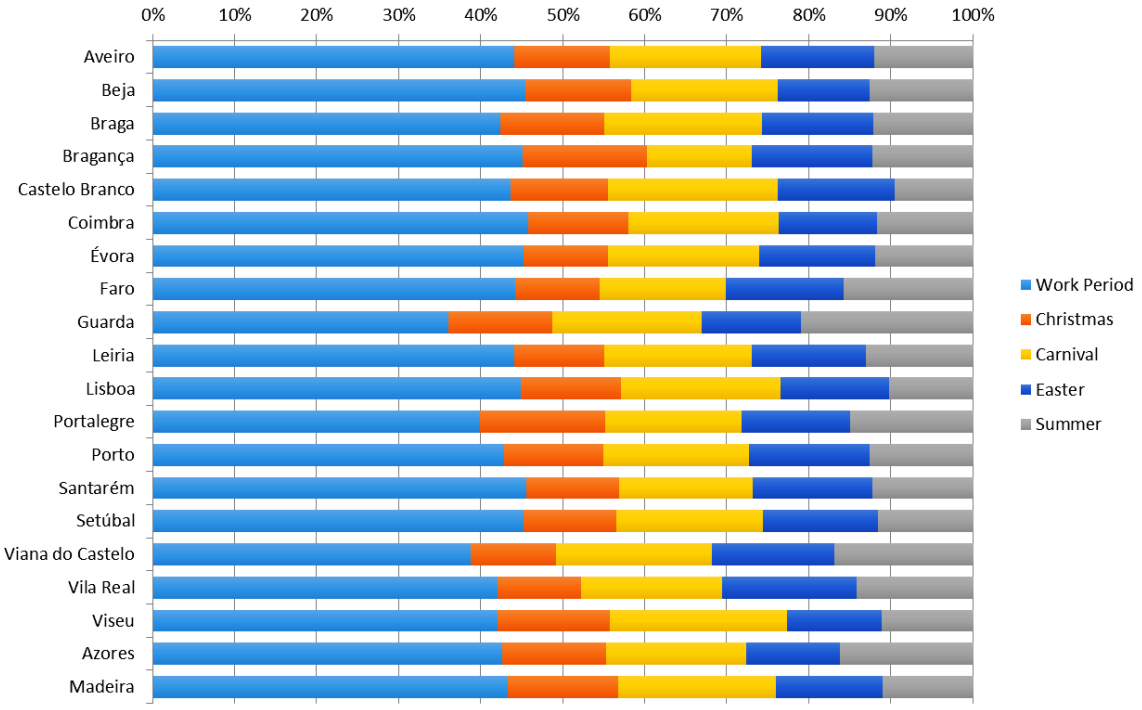


Fig. 7 Average number of tweets per day (as a percentage) for each considered period. Twitter production is lowest during work periods and reaches its highest during Carnival.

5 Data Characterization

5.1 Usage of Twitter conventions

The proportion of occurrence of Twitter conventions, i.e., URLs, hashtags, mentions, replies and retweets, allows the characterization of Twitter use on a particular community. A study by (Hong *et al*, 2011), performed such analysis for different languages. It included the study of 6M tweets in Portuguese language (mostly Brazilian Portuguese) collected between April 18th and May 16th, 2010. The study found that 13% of tweets contained URLs, 12% contained hashtags, 50 % of tweets made reference to other users via mentions, 32% of tweets were replies to other tweets, and 12% were retweets. It should be emphasized that the tweets used in the study did not consider geolocalization and were not related to specific countries, but to specific languages.

A similar analysis was performed on our geolocated corpus of 27.8M Portuguese tweets. Table 5 shows the results we obtained for the whole country and per district. The results we found are noticeably different, showing a much lower usage for each category (in some cases one order of magnitude less!) These large differences are surprising but can be explained (or at least attenuated) by some factors, the most important of which is the fact that all the tweets in our database are geolocalized. This is a very relevant issue because most Portuguese mass tweet producers, such as news agencies, newspapers, TV channels or TV shows, do not publish geolocalized tweets, and as such are absent from our database. The only notorious exception is the newspaper "Jornal de Notícias" (Table 6). Tweets by such users are characterized by including conventions: they incentivize the use of #hastags, often include

URLs to sites where the readers can find more details on the message they are commenting or transmitting, and often retweet related messages (with or without recurring to mentions). The weight of the mass tweet producers within all the produced tweets is enough to explain the much lower Twitter convention values we found. Less important reasons include the fact that in (Hong *et al*, 2011) most of the tweets are from Brazilian users (which at the time outnumbered the Portuguese Twitter community by a large factor: 30M users vs. less than 600K), and Twitter Brazillian and Portuguese communities do not necessarily behave similarly.

Analyzing the data presented in Table 5 while taking into consideration the categories of Twitter use suggested by (Java *et al*, 2006), it is possible to state that the Portuguese community on Twitter uses this social network essentially to chat, exchange thoughts and opinions. This is supported by the fact that the conventions with a significant usage percentage are mentions (17.88%) and replies (16.97%), while the conventions associated with the dissemination of information, URLs (2.13%) and retweets (0.02%) have a residual usage.

Differences between districts exist but are not very significant: mentions and replies are less used in the interior (which might or not indicate a lesser tendency for conversational use in the interior); URLs are strangely more frequently in Viana do Castelo; Lisbon users seem to be more adept at hashtagging than habitants from other regions; etc.

Table 5: Percentage of tweets using various conventions.

	URLs	**Hashtags**	**Mentions**	**Replies**	**Retweets**
Portuguese language tweets, 2010 (Hong et al)	**13%**	**12%**	**50%**	**32%**	**12%**
Portugal, geolocated tweets	**2.13%**	**3.35%**	**17.88%**	**16.97%**	**0.02%**
1 - Aveiro	1.68%	2.64%	19.50%	18.36%	0.02%
2 - Beja	1.73%	2.86%	12.31%	11.56%	0.03%
3 - Braga	2.15%	3.53%	18.57%	17.63%	0.02%
4 - Bragança	2.74%	3.14%	14.15%	12.57%	0.23%
5 - Castelo Branco	2.11%	3.95%	17.26%	16.18%	0.05%
6 - Coimbra	1.80%	3.02%	21.13%	20.07%	0.02%
7 - Évora	1.13%	2.61%	17.73%	16.85%	0.01%
8 - Faro	2.32%	2.64%	16.50%	15.55%	0.01%
9 - Guarda	2.21%	2.88%	13.65%	12.71%	0.09%
10 - Leiria	1.71%	2.90%	19.92%	18.84%	0.03%
11 - Lisboa	2.65%	4.15%	17.43%	16.34%	0.02%
12 - Portalegre	1.76%	3.00%	12.85%	11.91%	0.02%
13 - Porto	2.24%	3.23%	18.81%	17.84%	0.03%
14 - Santarém	1.64%	2.80%	18.08%	16.99%	0.02%
15 - Setúbal	1.45%	2.70%	16.76%	15.85%	0.01%
16 - Viana do Castelo	5.13%	5.03%	15.44%	13.44%	0.01%
17 - Vila Real	2.18%	3.70%	16.73%	15.77%	0.01%
18 - Viseu	1.50%	2.93%	17.79%	16.90%	0.02%
19 - Madeira	1.89%	3.89%	13.04%	19.32%	0.01%
20 - Azores	3.30%	3.51%	15.21%	13.97%	0.00%

5.1.1 Topics of Conversation in Portugal: Hashtags

In terms of the content of messages posted in each of the districts during the period of September 2014 to September 2015, it was found that the most used hashtags in each district involve football (soccer) and entertainment prime time TV shows. The most used #hashtag is common to all districts, #carregabenfica, which is related to "Sport Lisboa e Benfica", the Portuguese football champion in 2014/2015. The hashtags #ss5, #idolospt and #unicamulher refer to prime time entertainment TV programs and are also top hashtags in all districts. Fig. 8 shows the frequencies of the top 4 hashtags common to all districts. The discussion and exchange of views on football and television programs cuts across all districts of Portugal, but the theme of football has prevalence, since among the top-k hashtags for each district are #diadesporting, #sportingcp, #somosporto, #fcporto, #rumoao34, all related to the three largest Portuguese football clubs.

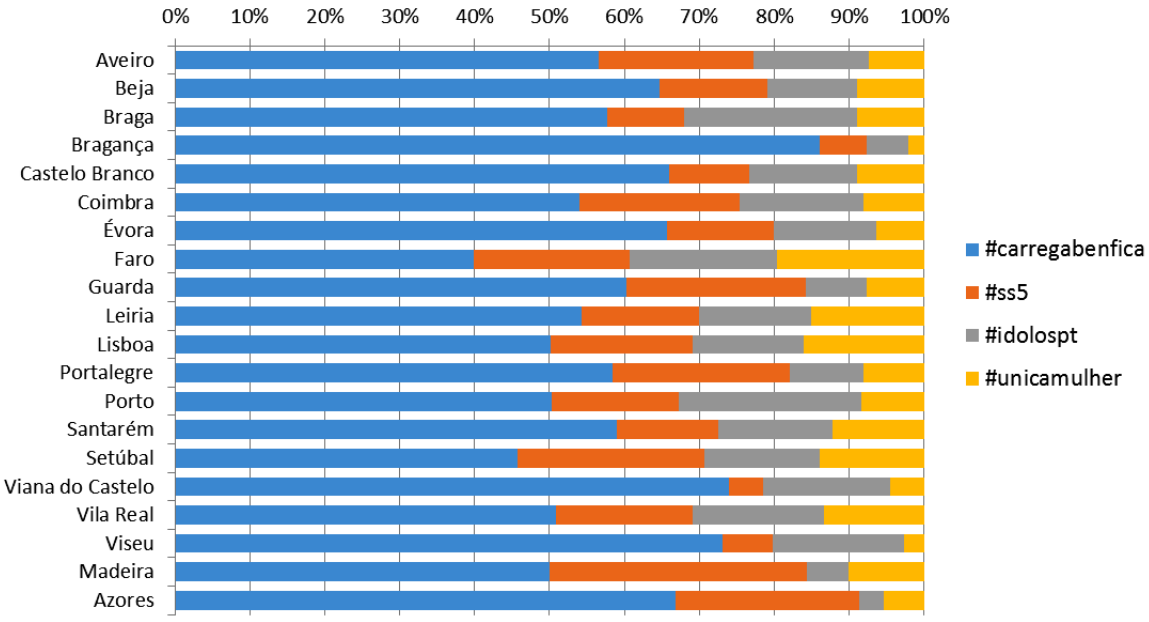


Fig. 8 Occurrence of the Top 4 hashtags in all districts

5.1.2 Information Share: URLs

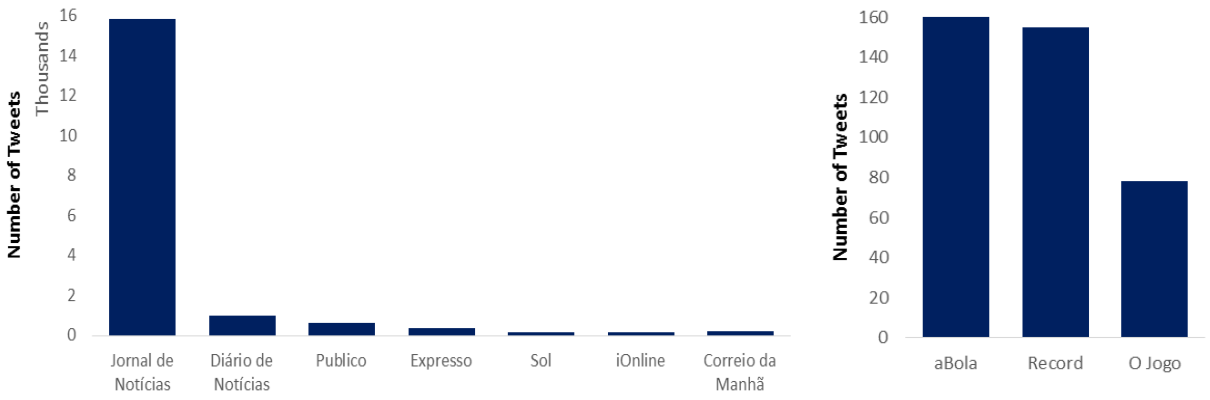
Twitter users can include URLs or links in their tweets with intention to share, or elaborate, on the information published on a tweet. The top 5 domains shared in the analyzed corpus of tweets are: 1 – www.instagram.com; 2 – www.trendinalia.com; 3 – www.dlvr.it; 4 – www.swarmapp.com; 5 – www.youtube.com.

Compared to the tweets in English analyzed in (Hong *et al*, 2011), the only top common link is the sharing of videos from YouTube, even though Instagram also appears in many common lists. No significant regional differences were observed in the Top 5 positions, but going lower it is possible to find differences, such as, for example, a local car dealership in the district of Braga that uses Twitter to promote its products.

Sharing URLs is directly related to information sharing. Fig. analyzes the frequency of occurrence of URLs related with each of the main Portuguese newspapers. As mentioned before, “Jornal de Notícias” is the only major newspaper that publishes on Twitter using geolocalized information (Table 6). As such, it is no wonder that “Jornal de Notícias” is by far the most frequent URL in our database (Fig. 9a). Note that the “Correio da Manhã” newspaper's website got in July 2015 about five times more visits than the site “Jornal de Notícias[4], and as such it should be much more popular in Twitter if one excludes the geolocation bias. The same source states that in terms of sports newspapers, the website of newspaper "aBola" is the leader in both visits and previews, followed by the websites from “Record” and “O Jogo”. This trend is confirmed in Fig. 9b. In this case there is no bias since none of the 3 sports newspapers publishes using geolocalized information.

Table 6: Jornal de Notícias, the only major Portuguese newspaper that tweets using geolocalized information.

```
{
   "id": 15391813,
   "id_str": "15391813",
   "name": "Jornal de Notícias",
   "screen_name": "JornalNoticias",
   "location": "Porto - Portugal",
   …
}
```



a) Generalist newspapers b) Sports newspapers

Fig. 9 Newspapers URLs shared by geolocated Twitter users.

[4] http://www.jn.pt/PaginaInicial/Nacional/Media/Interior.aspx?content_id=4730582. Accessed 12 November 2015.

5.2 Emoticons

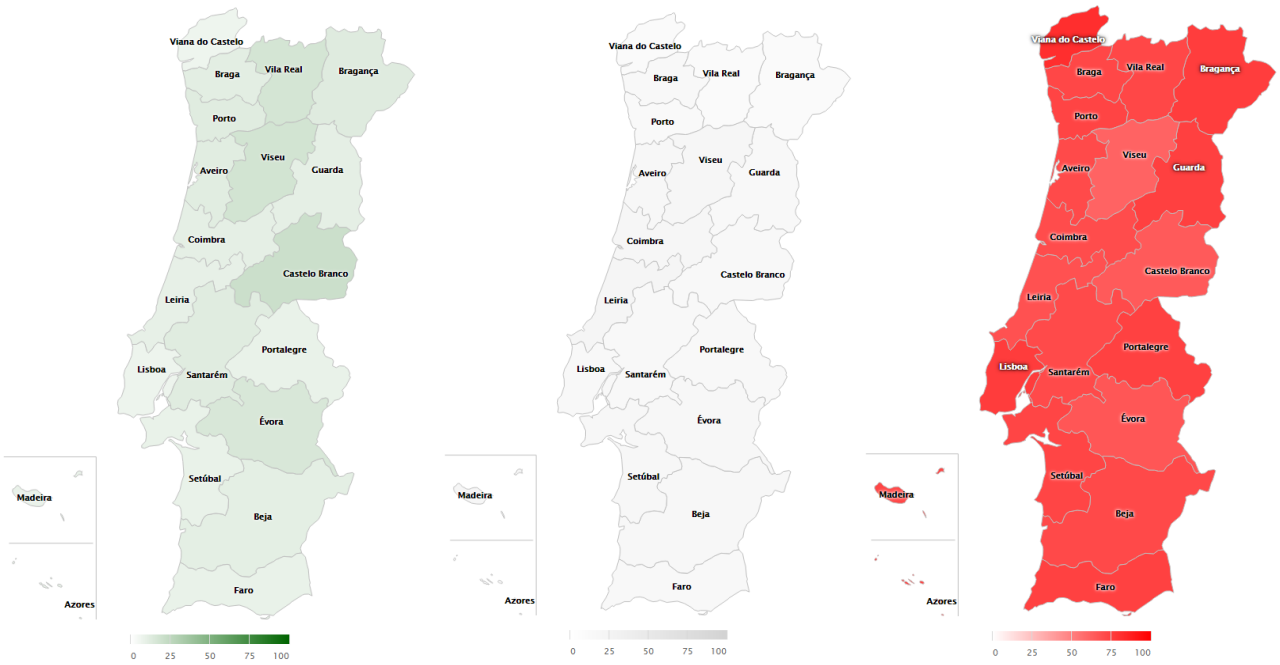
Individual happiness is a fundamental societal metric (Dodds *et al*, 2011), and as such one factor worth of analysis using any available data. Emoticons allow expressing a large and diverse set of emotions using a compact representation that uses just a couple of characters, and the Twitter 140 character message size limit incentives a their use. Therefore emoticons are *a priori* a means to characterize user happiness in Twitter.

Table 7: Top 10 emoticons used for representing Positive, Neutral and Negative feelings and respective frequency.

Positive (freq)		**Neutral** (freq)		**Negative** (freq)	
xD	183,320	??	449,922	:/	**1,355,699**
:3	88,843	-.-	73,915	o/	**1,294,110**
:)	83,386	%	57,543	:(	59,447
;)	37,624	:0	56,605	:c	50,195
:D	23,188	@x	34,623	:$	26,755
:p	19,876	*-*	34,501	:'	25,827
:-)	15,833	:o	34,273	((	18,655
XD	15,368	OOOOOO	23,501	:'(	14,458
^^	8,021	-_-	4,899	:s	12,211
:*	6,963	L.	4,895	:\	3,890

We looked for emoticons within our geolocalized database and categorized them into “positive”, “neutral” and “negative” feelings. Table 7 presents the top 10 emoticons for each category and respective occurrence frequency in the database. The results were rather surprising: the most frequent emoticons (by a large margin), “:/” and “0/”, are used to express “confusion” and “frustration”. The occurrence of staples such as “:)” or “:(“ is one order of magnitude lower than “:/” or “0/”. This can be seen as a huge indicator of the young age of most Twitter users in Portugal, and shows how they are lost about their future and their lack of perspectives under the crisis and austerity affecting the country during the analyzed period. The results also reveal the lack of anger and revolt usually associated to youngsters that are more politically oriented and use Twitter for dissemination of ideas, reinforcing the notion that Twitter is mostly used in Portugal for more soft and/or recreational purposes.

Fig. 10 shows how (a) positive, (b) neutral and (c) negative districts are. The darker the tone, the more intense is the feeling. Intensities are comparable between the figures and show the prevalence of negative feelings and a generalized degree of frustration, dissatisfaction and unhappiness (tone intensity is higher for negative feelings across all districts), which, once again is in line with the overall sentiment associated with the austerity imposed to the country during the period in question.



a) Positive feelings b) Neutral feelings c) Negative feelings

Fig. 10 Emoticon frequency per district.

5.3 Tweet Publishing Source

Each tweet contains a "source" field that is used to indicate the type of device used for the tweet publication. In addition it is possible to know the device operating system (e.g. Android, Windows Phone, iPhone, etc.). An analysis by district shows that Android is the most prevalent source for geolocalized tweet publishing in all Portuguese districts. Fig. 11 shows the top publishing sources used in each district.

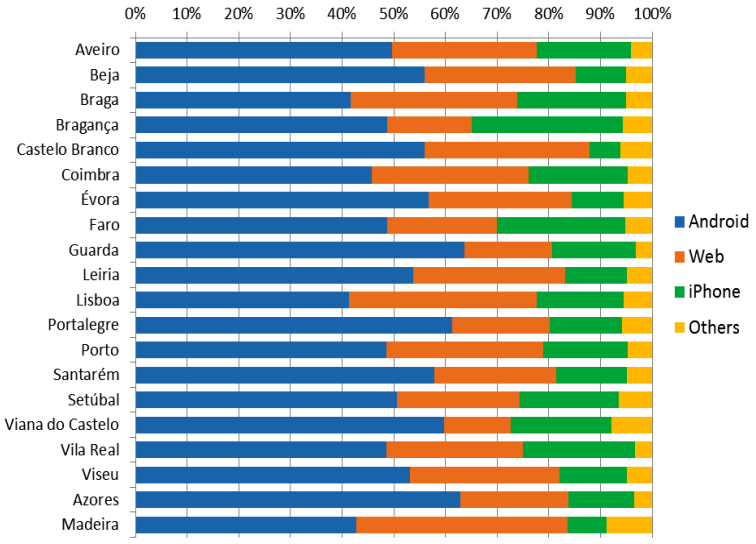
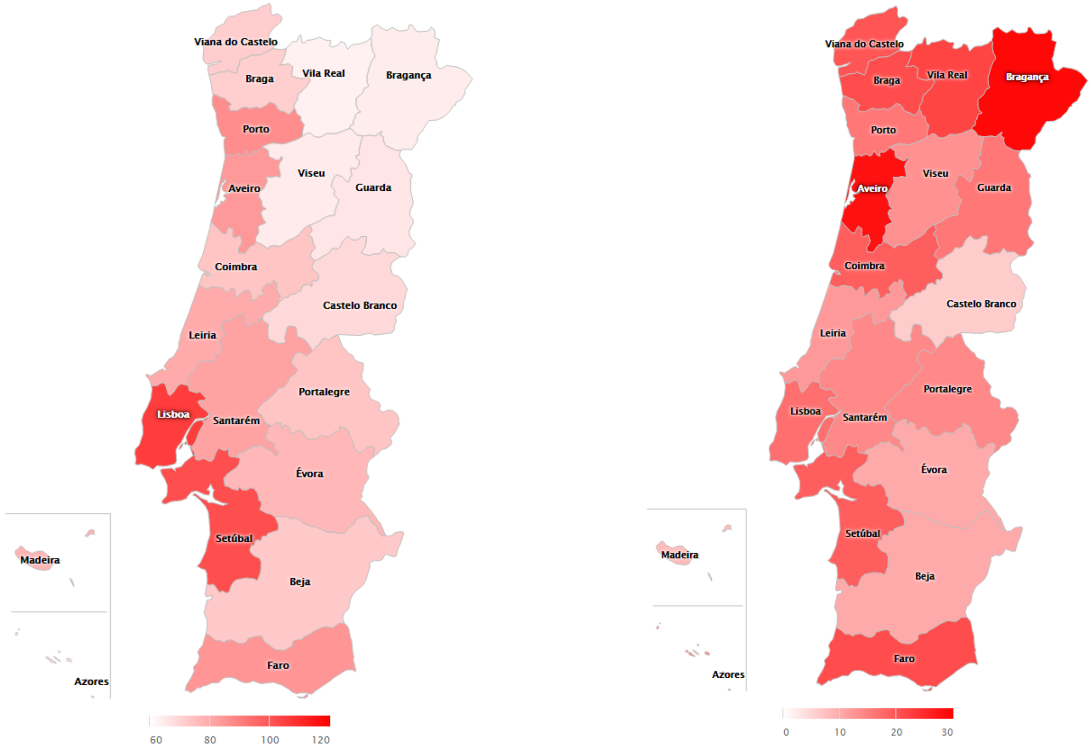


Fig. 11 Top Twitter publishing sources per district.

A possible interesting analysis on this data arises from (Edwards, 2014) statement in Business Insider: "the rich, it seems, use iPhones while the poor tweet from Androids". If one accepts such statement as true – which is obviously very debatable –, it would be possible to infer the overall wealth level of the country (more "poor" than "rich"), and an analysis based on iPhone usage for tweeting could indicate which are the "wealthiest" Portuguese regions. As can be seen in Fig. 12, this statement does not seem to hold in the case of Portugal, since districts such as Bragança, and Vila Real, which are some of the poorest Portuguese regions, have the highest Twitter iPhone usage.



a) Wealth per capita (100 is the EU average)

b) iPhone Tweeter usage

Fig. 12 Is iPhone usage a true wealth indicator?

6 Conclusions and Future Work

This paper presents an analysis over a database of about 27.8 Million Portuguese geolocated tweets, produced in Portugal by 97.8K users during a one-year 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; the prediction of regions where the concentration of the population will be higher or lower in certain periods of the year; what are the most satisfied and dissatisfied regions; what Portuguese use Twitter for; what do Portuguese tweet about; etc. Such information could prove useful for areas as different as marketing, tourism, sociological studies or even public health.

The decision to base the study solely on geolocated tweets has the advantage of allowing us to remove the influence of most tweet mass producing users, which tend to distort statistics due to their weight within the twitter community.

Among the most interesting conclusions is the fact that the Portuguese community on Twitter, which is in large part constituted by youngsters, uses this social network essentially to chat, and exchange thoughts to friends, instead of news dissemination. Moreover, the most discussed topics involve sports and tv shows, instead of "more serious subjects". It was also possible to denote the negativism and frustration usually associated with the Portuguese people, and the notorious absence of anger and revolt.

This paper is a first step in understanding the idiosyncrasies of Portugal and the Portuguese regions in terms of contents in daily-based or yearly-based periods. The presented analysis shows just a few examples of what can be done with the available data. This study will be further extended in order to better characterize each of the regions in terms of daily habits, user profiles, and also in order to better understand the way people travel across 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.

7 References

Blanford J, Huang Z, Savelyev A, MacEachren A (2015) *Geo-Located Tweets. Enhancing Mobility Maps and Capturing CrossBorder Movement*, PLoS ONE, doi:10.1371/journal.pone.0129202.

Bora S, Singh H, Sen A, Bagchi A, Singla P (2015) *On the role of conductance, geography and topology in predicting hashtag virality*, In Journal of Social Network Analysis and Mining, Springer, pages 1-15.

Boyd D, Golder S, Lotan G (2010) *Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter.* Proceedings in System Sciences (HICSS), pages 1-10.

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

Brogueira G, Batista F, Carvalho JP (2015) *Sistema Inteligente de Recolha, Armazenamento e Visualização de Informação proveniente do Twitter*, 15th Conferência da Associação Portuguesa de Sistemas de Informação, CAPSI'2015, Lisboa.

Burnap P, Williams M, Sloan L, Rana O, Housley W, Edwards A, Knight V, Procter R, Voss A (2014) *Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack*, In Journal of Social Network Analysis and Mining, Springer.

Cheng Z, Caverlee J, Lee K (2010) *You are where you tweet: a content-based approach to geo-locating twitter users*. In Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM '10). ACM, pages 759-768.

Chandra S, Khan L, Muhaya FB (2011) *Estimating twitter user location using social interactions - A content based approach*. In Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011, pages 838-843.

Chang J, Sun E (2011) *Location3: How Users Share and Respond to Location-Based Data on Social Networking Sites*, In Proceedings of the 5th. International Conference on Weblogs and Social Media (ICWSM'11), AAAI Press, pages 74 – 80.

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

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

Diário da República Portuguesa (1989) *Decreto-Lei n.º 46/89*, pages 590 – 594, 15 February 1989.

Dodds P, Harris K, Kloumann I, Bliss C, Danforth C (2011) *Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter*. PLoS ONE, doi:10.1371/journal.pone.0026752.

Edwards J (2014) *These Maps Show That Android Is For People With Less Money*.

http://www.businessinsider.com/android-is-for-poor-people-maps-2014-4. Accessed 12 November 2015.

Eisenstein J, O'Connor B, Smith N, Xing E (2010) *A latent variable model for geographic lexical variation*. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP'10). Association for Computational Linguistics, pages 1277-1287.

Hecht B, Hong L, Suh B, Chi E (2011) *Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles*. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11). ACM, pages 237-246.

Honeycutt C, Herring S (2009) *Beyond Microblogging: Conversation and Collaboration via Twitter*. Proceedings of the Forty-Second Hawaii International Conference on System Sciences (HICSS-42). Los Alamitos, CA: IEEE Press, pages 1-10.

Hong L, Convertino G, Chi EH (2011) *Language Matters In Twitter: A Large Scale Study*. Fifth International AAAI Conference on Weblogs and Social Media.

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

Instituto Nacional de Estatística (2011) *Censos 2011 Resultados Definitivos – Portuga*l.

Java A, Song X, Finn T, Tseng B (2006) *Why we Twitter: Understanding microblogging usage and communities*. Joint 9th WEBKDD and 1st SNA-KDD Workshop '07, San Jose, CA.

Kalarikkal S, Remya PC (2015) *Sentiment analysis and dataset collection: A comparitive study*. Advance Computing Conference (IACC), IEEE International, pages: 519 – 524.

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

Kim T, Huerta-Canepa G, Park J, Hyun SJ, Lee D (2011) *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.

Kumar S, Morstatter F, Liu H (2014) *Twitter Data Analytics*. Springer.

Mahmud J, Nichols J, Drews C (2014) *Home Location Identification of Twitter Users*. ACM Trans. Intell. System Technology 5, volume 3, Article 47, 21 pages.

Manfredini F, Tagliolato P, Rosa CD (2011) *Monitoring Temporary Populations through Cellular Core Network Data*. In Computational Science and Its Applications - ICCSA 2011. Lecture Notes in Computer Science, Springer Berlin Heidelberg, volume 6783, pages 151—161.

Mittal A, Goel A (2011) *Stock prediction using Twitter sentiment analysis, Standford University*, CS229 (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.375.4517&rep=rep1&type=pdf). Accessed 16 November 2015.

Rill S, Reinel D, Scheidt J, Zicari RV (2014) *PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis.* Knowledge-Based Systems, vol. 69, pages 24-33.

Santos JC, Matos S (2013) *Predicting flu incidence from* Portuguese *tweets*. In International Work-Conference on Bioinformatics and Biomedical Engineering - IWBBIO, pages 11–18.

Saravanan M, Sundar D, Kumaresh VS (2013) *Probing of geospatial stream data to report disorientation*. IEEE Recent Advances in Intelligent Computational Systems (RAICS).

Widener MJ, Li W (2014) *Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US.* Applied Geography, vol. 54, pages 189 – 197.

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