Tracking Language Mobility in the Twitter Landscape

Izabela Moise*, Edward Gaere*, Ruben Merz†, Stefan Koch† and Evangelos Pournaras*
*Computational Social Science, ETH Zurich, Switzerland
†Swisscom, Switzerland,
Email: {imoise, egaere, epournaras}@ethz.ch, {merzru,stefkoch}@student.ethz.ch

***Abstract*—The unprecedented data explosion has drastically changed the data science landscape. At the same time, Big Data analytics have reshaped the design and implementation of the applications that analyse the data. In this paper, we explore the use of Big Data tools for extracting value from Twitter data. We acquire a large set of Twitter data (10TB in size) and process it by relying on Spark DataFrame. The purpose of our analytics pipeline is to study the mobility of languages as captured by the Twitter signal. We study the evolution of languages from both a temporal and a spatial perspective, by applying density-based clustering and Self-Organising Maps techniques. The analysis enabled the detection of tourism trends and real-world events, as perceived through the Twitter lens.**

I. Introduction

Big Data is drastically changing the data science landscape. The most obvious transformation originates from the unprecedented explosion in the sheer amount of data. Multiple dimensions of our social life have Big Data proxies: our opinions and sentiments leave traces in blogs and tweets; our movements are recorded by GPS tracks and mobile phone logs; our social links are reflected in social media or phone networks. This increasing abundance of data represents a valuable resource for building data-driven systems that study complex social phenomena, in particular social behaviour. With the size and complexity of Big Data, new computational challenges emerge: Big Data analytics [1], [2] have reshaped the design of applications that analyse the data.

Although Big Data tools have experienced significant progress in terms of maximising efficiency and scalability, applying them into practical deployment settings is still a challenging task that is highly dependent on the particularities of the data. In this light, we deploy an analytical pipeline that aims at extracting value from Twitter data. The goal of our study is to explore language mobility in the Twitter landscape, by tracking temporal and spatial evolution of languages, as captured by the Twitter signal.

The Twitter dataset we consider for analysis consists of the 1% of public tweets for the period 2013-2014. Although Twitter has rapidly become a prolific data source for researchers, harvesting information from Twitter data is a challenging task. The analytical pipeline begins with data preprocessing, with the main purpose of extracting coupled information on location and language, at the tweet-level. The sequence of preprocessing steps retains only geo-localised tweets and filters out social bots. We processed an initial dataset of 10 TB of Twitter data with Spark DataFrame on a cluster of 1000 cores, in 1 hour and 6 minutes.

Our quantitative analysis proceeds in two steps. First, we explore the temporal evolution of languages, as expressed in the Twitter-sphere. At this level, we zoom into the temporal language composition of the Twitter signal, in three specific case studies. By performing linear regressions, we find evidence that Twitter adoption over time is highly heterogeneous and partially reflects language distribution in countries with multi-linguistic landscape (e.g. Switzerland). We are also able to detect migration patterns. In a second step, we shift the perspective through which we explore the data, from a temporal dimension to a spatial one. The analysis builds monthly snapshots illustrating the spread of languages, by applying a density-based clustering technique. Building on the insights provided by the temporal and the spatial analyses, we perform a study on both dimensions, by investigating how language mobility over time is reflected in Twitter data. The mobility of a language is emulated through the time-lapse of the centres of the language mass which are constructed by applying a neural network model on the data, more specifically Self-Organising Maps. Finally, we show that exploring shifts in language composition, based on the Twitter signal, represents a promising approach to discovering tourism trends and fluxes.

The rest of this paper is structured as follows: Section II provides an overview of related work. Section III-B presents the first phase of the analytical process, data preprocessing. Sections IV and V focus on the two analyses we perform on our dataset, while Section VI reports on our findings regarding language mobility. Section VII provides the final remarks.

II. Context

A. Twitter

Twitter is a massive social networking site designed for fast communication and relying on population engagement as the main pillar. Official numbers indicate that in June 2016, the number of monthly active users reached 313 millions [3].

Twitter's predominant nature as a real-time information network (and less as a social network), along with Twitter's speed and ease of information sharing have led to the broad adoption of Twitter as a valuable data source for research in various fields. A few examples include the prediction of earthquake occurrences [4], and detecting and tracking epidemics [5].

Harvesting geographic and language information from tweets is far from being a trivial task. A comparative survey of the key methods used to infer language and location in Twitter can be found in [6]. The study in [7] characterises the Twitter population of the U.S. along three axes (geography, gender and race) concluding that the Twitter users represent a highly non-uniform sample of the U.S. population. Previous studies also explored worldwide linguistic indicators mined from large-scale datasets of microblogging posts, as a useful tool for explaining various social phenomena. In [8], the authors explore the potential of geographical and language levels: first, at a coarser granularity by looking at correlations between the adoption of Twitter and the GDP of a country and also discovering touristic seasonal patterns within countries; second, at a fine granularity level, exploring the spatial distribution of languages in cities and neighbourhoods.

Modeling and predicting disease spread heavily rely on information about human mobility. In [9] and [10], the authors investigate the feasibility of using Twitter as a proxy for human mobility, and as a potential replacement of more time-consuming methods such as census data or mobile phone logs. The adoption and usage of social media across different countries and societies are analysed in [11], in which the authors report differences and similarities in terms of activity, sentiment, use of language and network structure. Previous research found that geographic distances, national boundaries, and languages hold considerable influence on the formation of social ties on Twitter [12], and also on the usage patterns of Twitter features between different language communities [13]. In [14], the authors explore the socio-spatial relations extracted from geo-tagged tweets with the purpose of studying the process of segregation in multi-cultural neighbourhoods.

B. Big Data tools

Apache Spark [2] is a high-performant cluster computing platform designed to be general-purpose. The main feature that gears Spark toward high-speed computing is the ability to run computations entirely in memory. Spark provides high-level built-in libraries for supporting the development of Data Science applications: machine learning, graph processing, structured data processing, stream processing. Spark SQL is a Spark module for processing structured data. The core component of Spark SQL is a DataFrame, a distributed collection of data, organized into named columns. DataFrame supports reading data from popular formats, including JSON files, and provides a domain-specific language for distributed data manipulation.

III. Data acquisition and preprocessing

A. Data acquisition

Twitter usage statistics report a rate of 6,000 tweets generated every second, corresponding to a daily rate of 500 million tweets. Twitter has made available to developers a sampled view of its data through two APIs: the *REST APIs* (granting access to historical tweets) and the *Streaming APIs* (providing a continuous stream of public tweets).

The Twitter dataset we consider for our analysis is a subset of a large data collection, put together by the Archive team [15]. The Archive data was acquired through the Twitter Streaming API, the Twitter protocol which grants access to 1% of public tweets. The streaming mechanism provides useful metadata information for each tweet, such as posting time and date, author information, geographical information (if available), retweet counts and network indicators (number of friends and followers). The stream of tweets is returned in the JavaScript Object Notation (JSON) format. The Archive Twitter dataset spans over a period of 4 years (2012 - 2015), consisting of monthly compressed archives. Each archive is structured as a multi-layer hierarchy of folders with the following levels: `year` $\rightarrow$ `month` $\rightarrow$ `day` $\rightarrow$ `hour` $\rightarrow$ `minute`. At the last level (i.e. the minute level) a JSON file contains the tweets streamed in a minute's time. The dataset we consider here, covers the period from January 2013 to December 2014, reaching 10 TB in size.

B. Data preprocessing

In the following, we describe the preprocessing pipeline applied on the raw JSON files with the goal of extracting the fields relevant for further analysis. We first discuss the Twitter metadata that specify language and location information.

Twitter allows users to share their location in two ways: (i) attaching geographical coordinates to tweets and/or (ii) setting a home town in their public profile. Although the former supplies a reliable location indicator, it requires deliberate effort and is consequently hardly used. Whenever the user's device enables location information (GPS coordinates) for the Twitter account, each tweet posted from this device will have associated GPS coordinates. However, a large audience of users will not use this GPS enabled service for social media.

In contrast, profile-level location sharing is far more common. Users can supply as part of their profile, a short non-structured string of text representing their location. In practice, very few users provide a valid string for these fields [6]. Additionally, Twitter incorporates a mechanism that automatically determines the location of the tweet based on a combination of geographic coordinates or based on the text of the message. However, the matching of a location mentioned in a tweet and the real location, is often not entirely accurate. For example, a tweet that mentions "London" in the text field will have the location field set to London, even if the message is sent from a different location. As these fields are not automatically updated, the information they provide might be stale or incorrect.

Language indicators for a given tweet are also supplied at two levels: user profile, where the language field is set by the user, and automatically determined by Twitter, based on the text of the message.

a) Filtering geo-tagged tweets: As a result of the above discussion, we retain for further analysis the following fields: `id`, `created_at`, `user.id`, `coordinates`, `lang`. In other words, we discard all the tweets that are not geo-localized (the `coordinates` fields are empty).

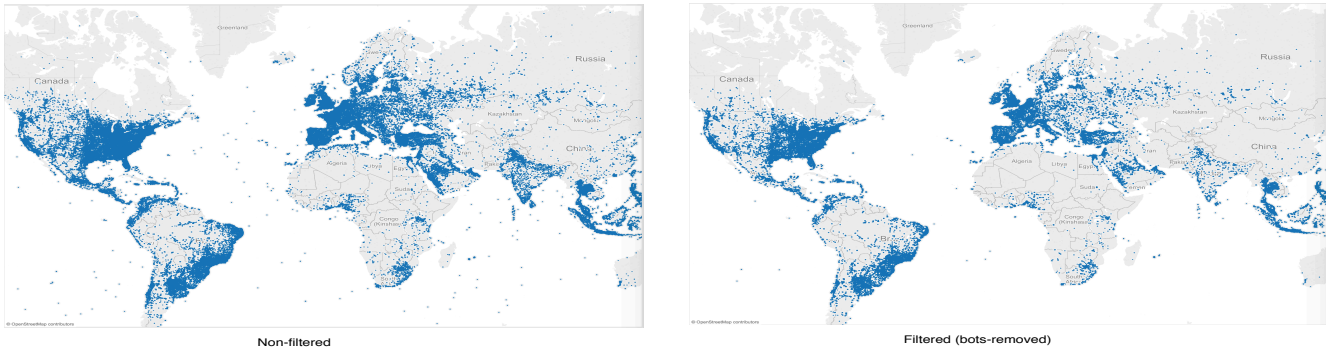


Fig. 1: Geolocated Twitter dataset, October 1st 2014. Each dot on the map represents a tweet.

b) Removing social bots: Social platforms such as Twitter are vulnerable to automated accounts that mimic real users, the so-called *social bots*, seeking to spread misinformation and to temper statistics by posting messages generated automatically and interacting with legitimate users. These social bots pose a high risk as they could bias or even invalidate many existing services, by infiltrating the social networks and acquiring trust of real users [16], [17]. A first indicator of a social bot activity is an unusually high number of posted messages. Starting from this valid assumption, we compile a list of the most active users in our geolocated dataset, at a monthly level. We observed a significant difference in the number of tweets posted by the most active users and the majority of users. While 61% of users tweet only once per month (in the geolocated dataset), the most active users post up to 439 times per month.

As we are interested in exploring patterns in users mobility, we add a motion analysis to the social bot removal method: we consider the distances traveled by users in the geolocated dataset. Messages from bots appear to have at least one of the following characteristics: high similarity in the structure of the text, extreme stationarity (social bots located in a data centre or in a weather station, exhibit no movement over time) or extreme velocity (geographical coordinates indicate locations very distant from each other, between two consecutive tweets). As an example, we mention a weather station in United Kingdom, broadcasting meteorological data and radio stations indicating names of songs currently played. At the monthly level, we discarded users that have not moved at least 100 meters and users that travel at speeds greater than 1000 km/h between two consecutive tweets. Figure 1 shows a snapshot of the geolocated tweets posted on October 1st 2014, in the initial dataset and in the filtered dataset, after bot removal.

c) Mapping geo-coordinates to cities: The goal of this third step is to enhance the dataset by applying a reverse geocoding operation. By taking as input a pair of geographic coordinates, this mechanism provides the `country` and `city` (city corresponds to the nearest city) fields that match the geo-coordinates. As a reverse geocoding tool, we relied on a Python-based library available at [18] and GeoNames [19].

We further proceed with a partial ranking of countries, based on the number of daily tweets. The ranking is presented in Figure 2, in terms of the average number of tweets per day. We only include the countries with a daily average number of tweets greater than 100 (out of the total of 200 countries identified in the dataset).

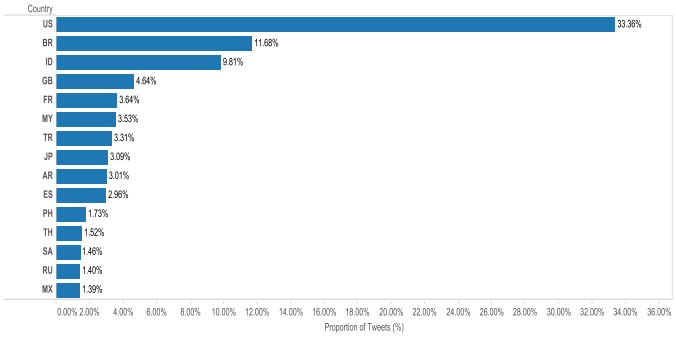


Fig. 2: Countries ranked by total number of tweets, in the filtered dataset. Only countries with at least 1% of all tweets are shown.

d) Converting to time-series: We further transform the geo-localised, bots-filtered dataset into time-series counting the daily number of tweets per language and per country.

e) Technical implementation: We have implemented this preprocessing pipeline by relying on Spark DataFrame, a Spark module for processing structured data. The experiments were carried out on a YARN [20] cluster deployed on 250 dedicated nodes, part of a high-performance computing infrastructure, available at ETH Zurich. The YARN cluster is deployed on a set of 250 nodes, each disposing of 4 cores and 16 GB of RAM. Apache Hadoop YARN version 2.7.1 is deployed as a resource-management platform. Spark application properties were set to 250 executors, each configured with 4 cores and 7 G of memory. Table I presents the completion time for processing 10 TB of Twitter data.

TABLE I: Data preprocessing with Spark

data size	completion time	post-processed data size	post-processed # tweets
10 TB	1h 6min 22 sec	4 GB	55 million

The data size reduces drastically: as an example for the month of January 2013, only 1.54% of messages contain

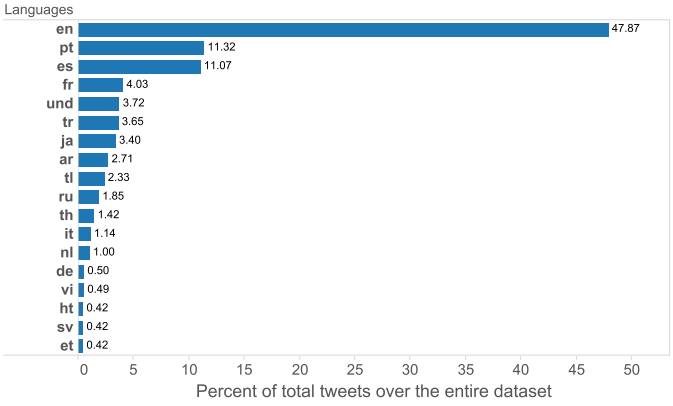


Fig. 3: Languages ranked by total number of tweets in the localised and filtered data-set.

geographic coordinates. As a result, the rest of the analysis was built on Python-based machine-learning modules: statsModels, scikit-learn and Kohonen.

IV. Temporal evolution of languages

A first analytical goal is to observe the temporal dynamics of languages, as reflected in the Twitter-sphere. We consider the time-series of the daily number of tweets per language and per country. This analysis heavily relies on linear regressions for detecting increases and declines on the time dimension.

We begin with a quantitative description of the distribution of languages contained in the localised dataset. Figure 3 provides the ranking of all languages extracted from the geo-localised, filtered dataset. According to expectations, English is the clear leading language on Twitter, followed by Portuguese, Spanish and French. However, this ranking does not reflect the official estimates of language speakers in the world: Chinese leads the ranking, followed by English, Spanish, and Hindi. This divergence serves as a reminder that we are observing the linguistic landscape through the lenses of a social platform that, for example, is not available in China.

In the following, we zoom into the temporal language composition of the Twitter signal, in three specific case studies.

A. Case study: Switzerland

We begin with a study of temporal evolution of languages in Switzerland. We consider the period April 2013 to December 2014. For this purpose, we have aggregated the time-series at a weekly level and we performed linear regressions on the data in order to capture the temporal evolution of languages in Switzerland. Figure 4 shows, on the left, the total number of tweets per week along with the number of languages detected (also at the weekly level) in Switzerland. A first observation is that the number of detected languages is rather stable over time, with an upward trend confirmed by the linear regression (represented by the dashed blue line). A further examination of the figure reveals that the number of tweets exhibits strong seasonal variations and a downward sloping trend, as shown by the linear regression (represented by the dashed green line). Figure 4, on the right, depicts the weekly evolution of all languages in Switzerland with a median proportion above 2.5%. Until July 2014, the three most prevalent languages are clearly French (`fr`), English (`en`) and German (`de`). There is a strong decrease of French and increase of English over time. The linear regression on the time-series confirms these changes over time. In addition, French and English exhibit the strongest slope coefficients (with an absolute value of 0.025 and 0.017 respectively), with French heavily more present than German. Other notable languages with a positive regression slope are Portuguese, German and Arabic (note that Arabic represents only 1.1% of the dataset). While the top three languages in Switzerland, German, French and Italian, are well represented in the dataset, their proportion does not correspond to the distribution in the Swiss population, with German more used than French [21]. The arabic language exhibits a fairly seasonal trend, reaching more than 6% and 9% in September 2013 and 2014, respectively.

B. Analysis of southern European languages

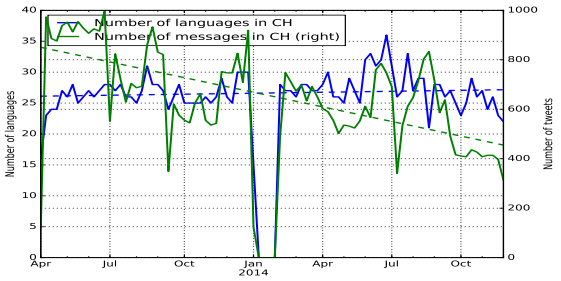
A second case study we consider focuses on four southern languages: Greek (`el`), Portuguese (`pt`), Italian (`it`) and Spanish (`es`). Similarly to the first case study, the time evolution of these languages is indicated by the changes over time in the density of Twitter conversations in each of these languages. We follow a similar methodology of conducting linear regressions on the time-series corresponding to each of the considered languages: more precisely, we consider the time-series of the number of countries covered weekly by each language and also the weekly number of tweets per language.

Figure 5 summarises our results. As a first observation, the conversation on Twitter in Greek and Italian is less dense than the communication in Portuguese and Spanish. Although the number of Italian tweets is one order of magnitude smaller than the number of tweets (at a weekly level) in Portuguese or Spanish, the number of countries covered by Italian is comparable to the countries covered by the other languages. Spanish is the language that covers the most countries.

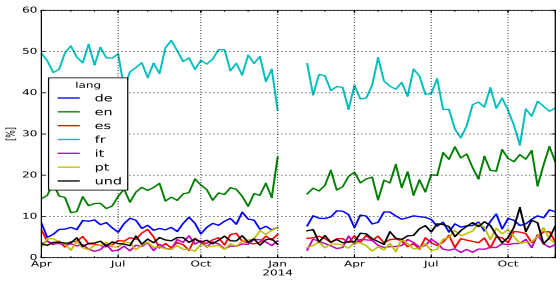
Another notable remark refers to the Greek language: it exhibits a high stability over time, Greece being its main and largest location. For the considered time period, our analysis did not detect any significant trend in other countries.

The Portuguese language is strongly localised with only three countries with a median weekly proportion above 1.0%. Brazil represents the largest share of the Portuguese language. Figure 6 shows a graphical representation of the evolution of Portuguese in Portugal. The number of tweets grows by a factor of 7 over the time-period.

Spanish exhibits an impressive growth in South America as shown in Figure 7 for Argentina, Uruguay and Brazil. Argentina surpasses Spain in terms of absolute proportion in the course of 2014. A study on migration routes between Latin America and the European Union [22] supports an indication of a heavy migration from Spain to Latin America. The study reports that for the first time in 14 years, the migration flux from the EU (with Spain the main source) to Latin America is more dense than the reverse.



(a) Number of languages detected in Switzerland and total number of tweets per week. Dashed lines correspond to linear regressions of the respective time-series.



(b) Languages in Switzerland with a weekly median proportion of at least 2.5%.

Fig. 4

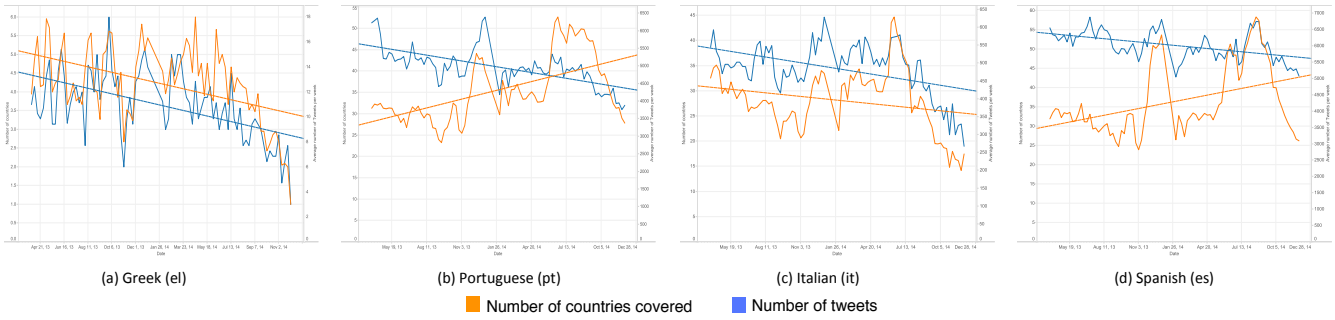


Fig. 5: Number of countries with tweets in Greek, Portuguese, Italian, and Spanish and total number of tweets per week. Dashed lines correspond to linear regressions of the respective time-series. Note that the respective scales can be different for each language.

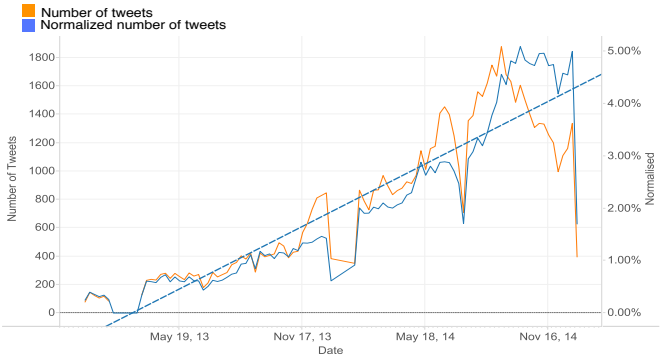


Fig. 6: Evolution of Portuguese in Portugal, weekly. Localized dataset, filtered

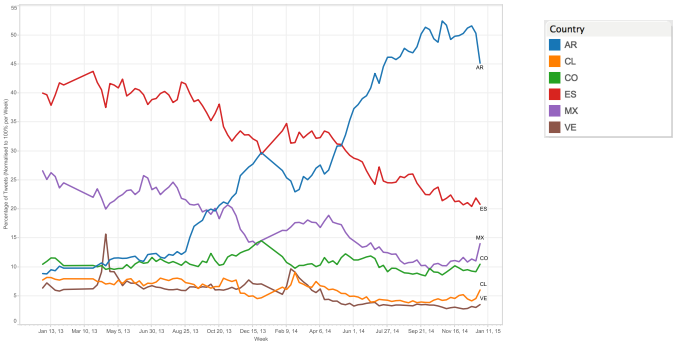


Fig. 7: Countries with tweets in Spanish where the median daily rate of tweets in Spanish is at least 4%.

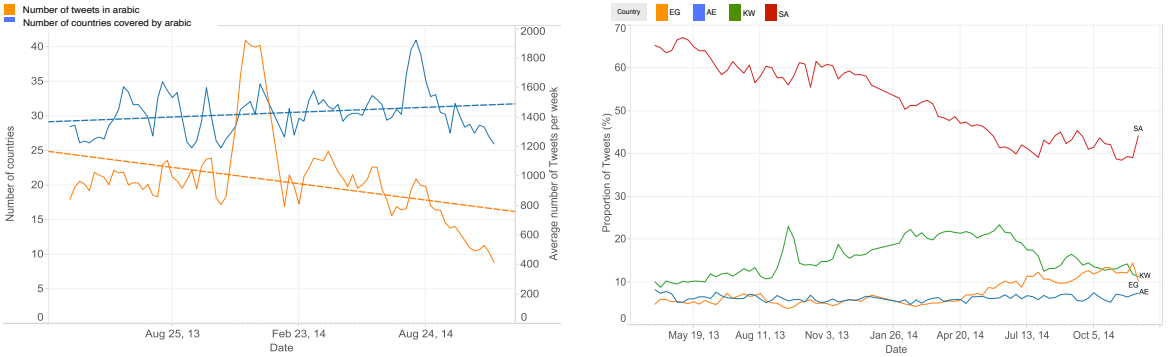
C. The Arabic language

Figure 8 illustrates our observations for the Arabic language: the left figure shows an increase in the number of countries where tweets in Arabic are posted. The regression analysis confirms that the most significant growth happened in Egypt, Kuwait, Jordan and Oman (depicted in the right figure). The large spike taking place in Kuwait in September 2013. The regression analysis shows that the Arabic language exhibits a generally growing trend outside of Saudi Arabia (the majority of slope coefficients are positive).

Twitter adoption over time is highly heterogeneous and partially reflects language distribution in countries with multi-linguistic landscape (e.g. Switzerland). The temporal thread in evolution of languages in Twitter may represent a reliable proxy for migration indicators. The southern European and Arabic language study brings evidence in this direction.

V. Spatial Evolution of Languages

In this section, we shift the perspective through which we explore the data, from a temporal dimension to a spatial one. More precisely, for each language in the dataset, we aggregate the number of tweets at a monthly level. We examine these monthly snapshots of language distributions over countries, as captured by the Twitter signal. As our purpose is to examine the spread of languages in spatial dimensions, we rely on a



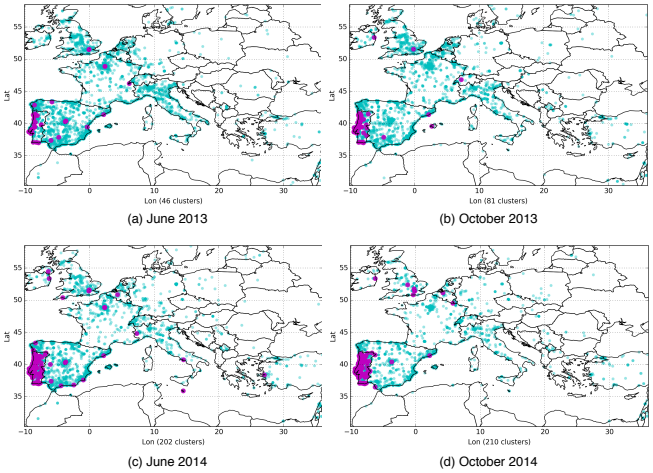
(a) Number of countries with tweets in Arabic

(b) Countries with tweets in Arabic with a weekly median $> 2.5\%$

Fig. 8: Evolution of Arabic languages, weekly. Localized dataset, filtered

cluster analysis for discovering dense concentration of tweets that share the same language. We selected a density-based clustering technique, namely *DBSCAN* [23]. DBSCAN is a density-based clustering algorithm that looks for dense neighbourhoods of traces defined by two parameters: the radius r of the neighbourhood (set to 0.01) and the number of points contained by the neighbourhood, which must be greater than a lower bound of $MinPts$ (set to 10). The rationale of the algorithm is that clusters correspond to areas with a higher density of points than areas outside the clusters.

We present here some of our findings for the months of June 2013, October 2013, June 2014 and October 2014, for the Portuguese and the Arabic languages. These timeframes and languages exhibited interesting patterns: as noticed from Figure 9, the number of cluster centres increases in 2014, as compared to 2013, indicating a higher density of Portuguese tweets in Portugal. Also, centres appear in United Kingdom and Ireland in 2014. Figure 10 shows a noticeable increasing trend in the spread of the Arabic language, in particular in the southern Mediterranean Sea. This is indicated by the increase in the number of detected clusters from one year to the other.



(a) June 2013 (b) October 2013

(c) June 2014 (d) October 2014

Fig. 9: Outcome of DBSCAN for the Portuguese languages. Cyan points represent original data and magenta points represent cluster centres.

VI. Language Mobility

Building on the insights provided by the temporal and the spatial analyses, we conduct a study on both dimensions, by investigating how language mobility over time is reflected in Twitter data. In order to track the mobility of a certain language, first we need to identify where the language is located, at a given time. A valid indicator is provided by computing the centres of mass associated with a language, for each month of Twitter data between January 2013 and December 2014. Therefore, we are able to observe the locations where a language is mostly used on Twitter. The centres of mass for each language are then plotted on the world's map in order to visualise possible movements over time.

A. Self-Organising Maps

To compute the centres of mass for each pair (language, month), we rely on Self Organising Maps (SOM) [24], a popular neural network model which performs a mapping from high-dimensional space to map units while preserving the initial topology. Usually, the map units form a two-dimensional lattice, and thus the mapping is done from high dimensional space onto a plane. For this reason, SOMs are well-suited for our two-dimensional geolocated dataset.

A two-dimensional matrix with 4 centres of mass may not yield sufficient granularity to capture dynamics of languages in less frequent locations. A 4×4 matrix would be too granular, with the risk of leading to an overfit. Thus, Twitter messages for each pair (language, month) are projected onto a 3×3 matrix, yielding 9 centres of mass. During the training process, tweets that are close to each other in space, are clustered together around a centre of mass.

B. Results

Figure 11 illustrates the centres of mass obtained for the Spanish language. The time-lapse of the centres of mass brings further evidence to support the migration trend from Spain to Latin America, also discussed in Section IV-B.

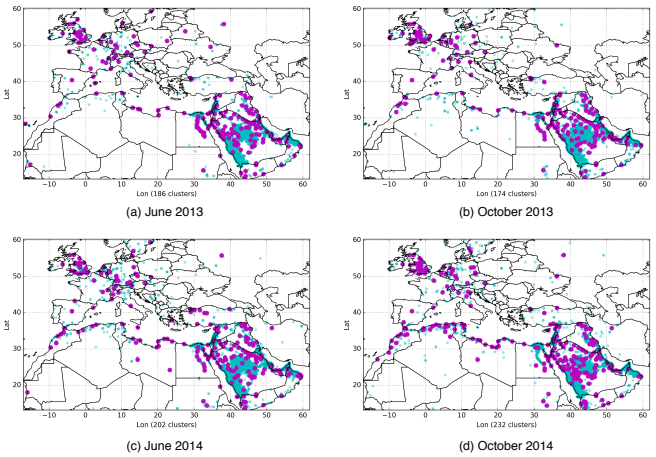


Fig. 10: Outcome of DBSCAN for the Arabic language. Cyan points represent original data and magenta points represent cluster centres.

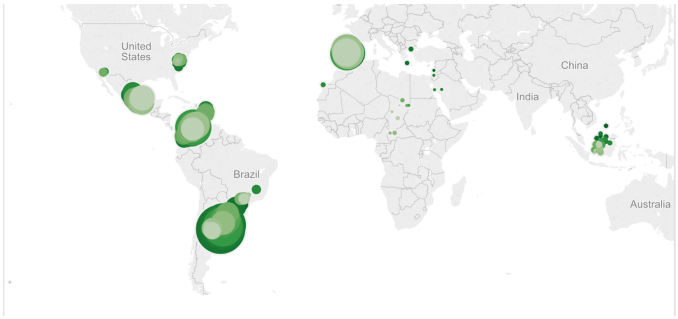


Fig. 11: Time-lapse of 9 centres of mass per month for Spanish. Large circles represent a higher number of tweets. Light green corresponds to January 2013 and dark green to December 2014.

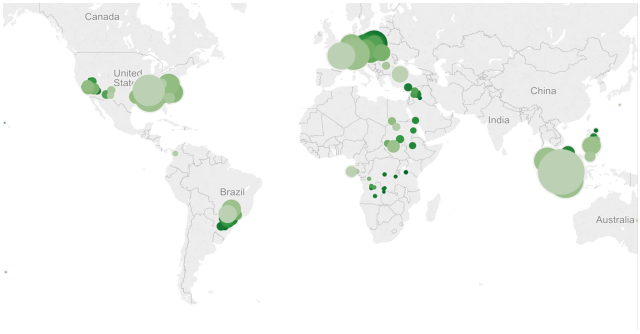


Fig. 12: Time-lapse of 9 centres of mass per month for Polish. Large circles represent a higher proportion of data. Light green is January 2013, dark green is December 2014.

Results for Polish and Portuguese are presented in Figures 12 and 13. These two languages exhibit a significant trend in motion, with an increase of density in Europe which may suggest a migration trend in Europe.

C. Discovering tourism patterns using language mobility

The language mobility analysis previously described is a good starting point for exploring tourism trends, by studying the variation of the language composition of a given country.

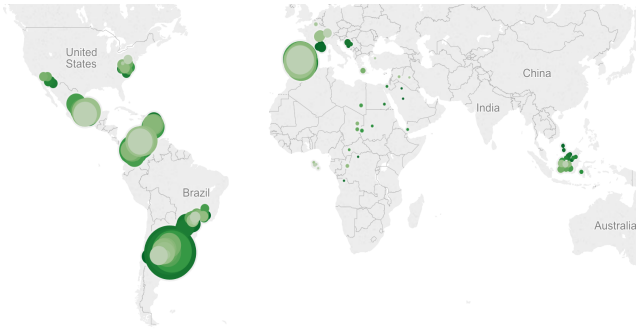


Fig. 13: Time-lapse of 9 centres of mass per month for Portuguese. Large circles represent a higher proportion of data. Light green is January 2013, dark green is December 2014

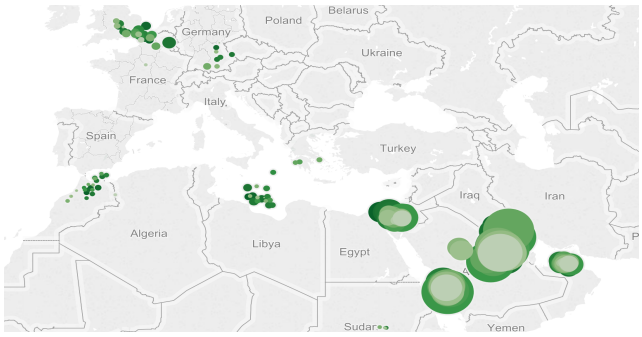


Fig. 14: Centres of mass for Arabic. Large circles represent a higher proportion of data. Light green is January 2013, dark green is December 2014

In a nutshell, the approach consists in differentiating between the country of origin and the current location of a person. We proxy the former though the users languages, as extracted from the Twitter dataset. Tweets geo-information provides a good indicator of the current location of the author of the tweets. We assume that on a macro level, we are able to observe trends based on country-language coupling, which may indicate tourism patterns.

In a first step, we group all the tweets by country, language, year and month and compute the total number of tweets and the monthly percentage for each pair (country, language), for 2013 and 2014. We normalised the monthly count of tweets by the total number of tweets for a given country-language pair for the two full years of data. This resulted into 1424 time-series of country-language pair. We compute the differenced time-series for each language, in terms of number of tweets compared to the previous month. For Croatia and Greece, we illustrate these numbers in a heat map depicted in Figure 15. A seasonal trend is indicated by the rapid increase followed by a decline in Russian, Spanish, English, French and German tweets over the summer months. As these languages are not representative for Greece and Croatia, we infer that these patterns show the inflow and outflow of tourists.

This mechanism also enables us to identify world events, mirrored in Twitter, such as the world cup in 2014, hosted by Brazil. Figure 16 plots the distribution of languages in Brazil over the period 2013-2014. The heatmap reveals sudden and large increases in the amount of tweets posted in certain

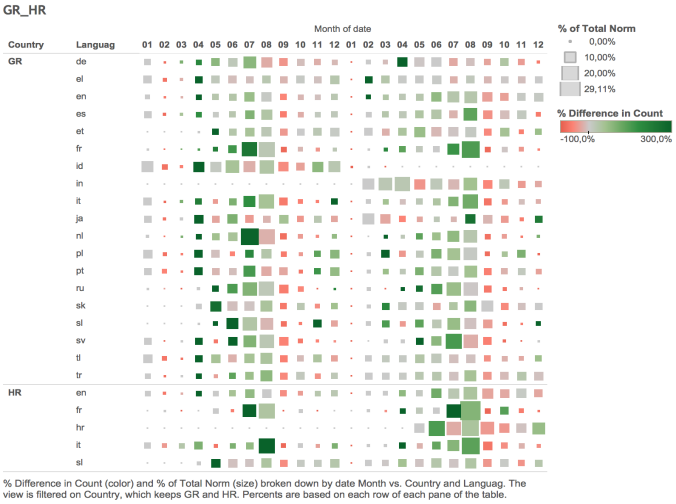


Fig. 15: Normalized tweet frequency in Croatia and Greece grouped by languages. Coloured by difference in number of tweets compared to the previous month.

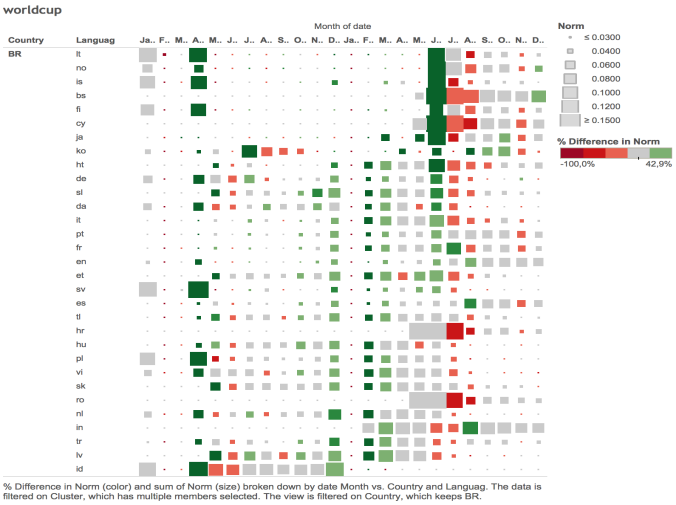


Fig. 16: Language distribution in Brazil

languages, during the month of June 2014, followed by a rapid decline in July 2014. We argue that the time-frame is too short to indicate a normal tourism trend, and instead it indicates a language mobility triggered by the world cup 2014.

Thus, exploring the shifts in the language composition at a country-level, based on the Twitter signal, represents a promising approach to observing tourism fluxes and trends.

VII. Conclusion

In this paper, we explored the use of Big Data analytics for tracking language mobility in Twitter data. We performed an extensive analysis on a 10TB Twitter dataset by relying on Spark DataFrame. The analytical pipeline heavily relies on Data Science techniques, such as density-based clustering and Self-Organising Maps. By tracking language and location in the Twitter signal, we studied the evolution of languages from a spatial and temporal perspectives. We restricted our analysis to geo-located tweets, in order to ensure the reliability of location information. Future work includes the exploration of methods that extract location from the text of the tweets, as well as applying the same analytical steps on other countries and languages.

Acknowledgments

This work is supported by the European Community's H2020 Program under the scheme 'INFRAIA-1-2014-2015: Research Infrastructures', grant agreement #654024 'SoBigData: Social Mining & Big Data Ecosystem' http://www.sobigdata.eu.

References

[1] "The Apache Hadoop Project," http://www.hadoop.org.
[2] "Apache Spak," http://spark.apache.org/.
[3] "Twitter Usage," https://about.twitter.com/company.
[4] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in *Procs. of the 19th International Conference on World Wide Web*. ACM, 2010.
[5] V. Lampos, T. De Bie, and N. Cristianini, "Flu detector: Tracking epidemics on twitter," in *Procs. of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part III*. Springer-Verlag, 2010.
[6] M. Graham, S. A. Hale, and D. Gaffney, "Where in the world are you? geolocation and language identification in twitter." *CoRR*, 2013.
[7] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist, "Understanding the Demographics of Twitter Users," in *Procs. of the 5th Intl. AAAI Conf. on Weblogs and Social Media (ICWSM'11)*, 2011.
[8] D. Mocanu, A. Baronchelli, N. Perra, B. Gonalves, Q. Zhang, and A. Vespignani, "The twitter of babel: Mapping world languages through microblogging platforms," *PLoS ONE*, 2013.
[9] J. Liu, K. Zhao, S. Khan, M. A. Cameron, and R. Jurdak, "Multi-scale population and mobility estimation with geo-tagged tweets," *CoRR*, 2014.
[10] M. Dredze, M. García-Herranz, A. Rutherford, and G. Mann, "Twitter as a source of global mobility patterns for social good," *CoRR*, 2016.
[11] B. Poblete, R. Garcia, M. Mendoza, and A. Jaimes, "Do all birds tweet the same?: Characterizing twitter around the world," in *Procs. of the 20th ACM Intl. Conf. on Information and Knowledge Management*, 2011.
[12] J. Kulshrestha, F. Kooti, A. Nikravesh, and K. Gummadi, "Geographic Dissection of the Twitter Network," in *Proc. 6th Intl. AAAI Conf. on Weblogs and Social Media (ICWSM'12)*. AAAI, 2012.
[13] L. Hong, G. Convertino, and E. H. Chi, "Language matters in twitter: A large scale study." 2011.
[14] T. Shelton, A. Poorthuis, and M. Zook, "Social Media and the City: Rethinking Urban Socio-Spatial Inequality Using User-Generated Geographic Information," in *Landscape and Urban Planning*, 2015.
[15] "Archive Twitter dataset," https://archive.org/details/twitterstream.
[16] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Who is tweeting on twitter: Human, bot, or cyborg?" in *Procs. of the 26th Annual Computer Security Applications Conf.* ACM, 2010.
[17] C. Freitas, F. Benevenuto, S. Ghosh, and A. Veloso, "Reverse engineering socialbot infiltration strategies in twitter," in *Procs. of the 2015 Intl. Conf. on Advances in Social Networks Analysis and Mining*.
[18] "Python reverse-geocoder," https://github.com/thampiman/reverse-geocoder.
[19] "Geonames," http://www.geonames.org/.
[20] "Apache Hadoop YARN," http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html.
[21] "Distribution of languages in Switzerland," http://official-swiss-national-languages.all-about-switzerland.info/.
[22] "International Organization for Migration: Migratory Routes and Dynamics between Latin American and Caribbean Countries (LAC) and between LAC and the European Union."
[23] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise." AAAI Press, 1996.
[24] T. Kohonen, "Neurocomputing: Foundations of research." MIT Press, 1988, ch. Self-organized Formation of Topologically Correct Feature Maps.