Online Social Networks and Media 1 (2017) 14–32

Contents lists available at ScienceDirect

Online Social Networks and Media

journal homepage: www.elsevier.com/locate/osnem

Perception of social phenomena through the multidimensional analysis of online social networks

Mauro Coletto[a], Andrea Esuli[b], Claudio Lucchese[b], Cristina Ioana Muntean[b,*], Franco Maria Nardini[b], Raffaele Perego[b], Chiara Renso[b]

[a] Department of Computer Science, Ca' Foscari University of Venice, Italy
[b] ISTI-CNR, Pisa, Italy

ARTICLE INFO

Article history:
Received 23 December 2016
Revised 28 February 2017
Accepted 11 March 2017
Available online 14 April 2017

MSC:
00-01
99-00

Keywords:
Twitter
Multidimensional analysis of OSNs
User polarisation
Topic and sentiment tracking

ABSTRACT

We propose an analytical framework aimed at investigating different views of the discussions regarding polarized topics which occur in Online Social Networks (OSNs).

The framework supports the analysis along multiple dimensions, *i.e., time, space and sentiment* of the opposite views about a controversial topic emerging in an OSN.

To assess its usefulness in mining insights about social phenomena, we apply it to two different Twitter case studies: the discussions about the *refugee crisis* and the *United Kingdom European Union membership referendum*. These complex and contended topics are very important issues for EU citizens and stimulated a multitude of Twitter users to take side and actively participate in the discussions. Our framework allows to monitor in a scalable way the raw stream of relevant tweets and to automatically enrich them with location information (user and mentioned locations), and sentiment polarity (positive vs. negative). The analyses we conducted show how the framework captures the differences in positive and negative user sentiment over time and space. The resulting knowledge can support the understanding of complex dynamics by identifying variations in the perception of specific events and locations.

1. Introduction

During the last few years, major social and political events happened in the European Union in concomitance with a deep economic crisis that polarized the opinion of the citizens into opposite sides with respect to many important issues, giving a new vision of a "Divided Union". The existence itself of the European Union (EU) seems to be at stake since even the EU leaders fail to show unity.

One of the most controversial topics undermining the union is constituted by the continuous waves of migration flows reaching Europe from Arabic countries. Indeed, we are nowadays witnessing one of the largest movement of migrants and refugees from Asian, African and Middle-east countries towards Europe. The United Nations High Commissioner for Refugees (UNHCR) estimates that more than one million of refugees arrived to the Mediterranean coasts in 2015 mainly from Syria (49%), Afghanistan (21%) and Iraq (8%). The map shown in Fig. 1 reports the main routes followed by migrants and refugees to reach EU coasts and northern countries. Understanding how the debate is framed between governmental organizations, media and citizens may help to better handle this emergency.

Another thoroughly divisive topic has been the unexpected results of the *United Kingdom European Union membership referendum*, held in the United Kingdom countries on June 23, 2016, which shows a clear will of the majority of the UK citizens to leave EU. This has been informally called by media the *Brexit referendum*, name that we often use throughout the paper. The result of the vote reveals a geographically and politically divided United Kingdom, where Scotland, Northern Ireland and the city of London are crisply pro-Remain, while England and Wales are essentially pro-Leave[1]. The map depicted in Fig. 2 illustrates this sharp division by reporting the results of the vote in each region of the UK.

These two controversial topics were intensively debated in traditional media and on Online Social Networks (OSNs) as well. We believe that a proper analysis of the discussions contributed by EU citizens on popular OSNs, aimed at capturing the users polar-

* Corresponding author.
E-mail addresses: mauro.coletto@unive.it (M. Coletto), Andrea.Esuli@isti.cnr.it (A. Esuli), Claudio.Lucchese@isti.cnr.it (C. Lucchese), cristina.muntean@isti.cnr.it (C.I. Muntean), FrancoMaria.Nardini@isti.cnr.it (F.M. Nardini), Raffaele.Perego@isti.cnr.it (R. Perego), Chiara.Renso@isti.cnr.it (C. Renso).

[1] http://www.bbc.com/news/politics/eu_referendum/results

http://dx.doi.org/10.1016/j.osnem.2017.03.001
2468-6964/

Fig. 1. The routes to European countries (source Business Insider – Europol, Reuters, Washington Post, AFP, ICMPD).

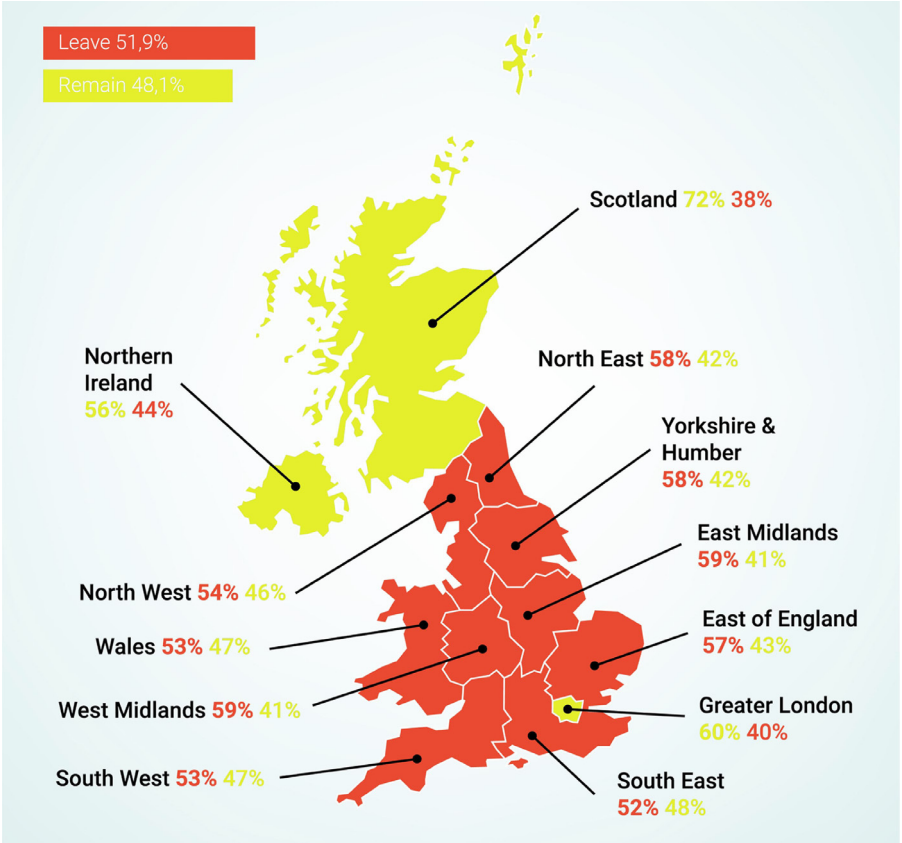


Fig. 2. United Kingdom European Union membership referendum results (source SHRM).

ity through time and space, may help to understand such complex phenomena.

Basic analyses of Twitter have been already performed by media in a simplistic way, mainly through manual analysis of content, statistics on news and hashtags for small samples of tweets. For instance, in the migration phenomenon it has been found that the *#welcome refugees* and *#germany* hashtags are used mainly from outside Germany, or in other articles the usage of terms *refugees* and *migrants* have been compared[2] in offline and online media: the former indicates someone forced to leave her country to avoid war or imprisonment, the latter is instead someone moving from his/her country searching for better living conditions.

At the same time the Brexit referendum has been well covered by social media and many web sites provide analyses of the "Pro Leave" or "Pro Remain" positions emerging from Twitter, Google and Facebook data[3].

However, the volume of messages exchanged on popular OSNs is massive and the efficient extraction of sentiment is challenging. Furthermore, a comprehensive work trying to systematically analyze how the debate and the perception of users about these events are evolving, shaped across the dimensions of places, time and sentiment is still missing. This work attempts to fill this gap: *i)* by proposing an analytical framework to capture and interpret users polarity about divisive topics from large collections of OSN data; *ii)* by discussing its application to the study of the refugees crisis and the Brexit referendum on Twitter.

The main characteristic of our approach is the extraction and analysis of relevant information along three dimensions: time, location and sentiment. Given an event, we identify the relevant messages in the OSN by choosing an initial set of seed hashtags. From the relevant posts, we first extract the spatial and temporal information. One characteristic of the spatial aspect is that we identify both the users location and the locations mentioned in the media posts. This allows us to perform analysis of the sentiment of users living in a given geographical area, while the mentions capture the polarity of users with respect to a different location. A second step is the tracking of the topic discussion over time and the computation of the sentiment of the messages, and therefore the polarity of the user regarding the topic of analysis. This is done by an iterative classification algorithm that progressively classifies both posts and users into a positive or negative class on the basis of hashtags co-occurrence. The algorithm has been proved to be efficient and to reach a good approximation in a few iterations. Third, the social media collection, enriched with spatial, temporal and polarity information, enables multidimensional analyses to be performed by querying the enriched data along these, possibly combined, orthogonal dimensions. In the final part of the paper we link together the analyses of the two social phenomena by showing common traits and correlations.

The approach proposed is general and can be easily adapted to any polarizing topic of interest involving multiple dimensions in OSNs. Moreover, it is efficient and scalable due to the automatic sentiment enrichment procedure. In this paper we experiment our framework using Twitter data related to our two case studies: the European perception of the 2015 refugee crisis and the Brexit referendum analysis before and after the vote. The analysis conducted shows how our framework allows us to easily identify the differences in positive and negative sentiment over time and space. The resulting knowledge can thus support the understanding of complex opinion dynamics by matching variations in perception with specific events and locations. Particularly, we highlight the sentiment of Twitter users of European countries towards the refugee crisis and how this sentiment evolved through time, space and in relation to some major events. Symmetrically, we follow the evolution of the Brexit referendum sentiment and the perception in other EU countries. Finally we try to correlate the two phenomena trying to understand if and how the migrant crisis sentiment has a correlation with the Brexit sentiment and vote outcome.

This paper is an extended version of a previous conference paper [1]. The original contributions presented in this paper include: a more detailed description of the proposed framework, the multidimensional analytical possibilities, and the study of the perception of the Brexit referendum, conducted by analyzing the English tweets posted during the weeks preceding and following the day of the vote. Finally, we report about the possibility enabled by our framework to exploit the common space dimension of the largest UK cities to correlate the user polarization on the two different analyzed phenomena.

The rest of the paper is organized as follows. Section 2 discusses the related work, while Section 3 describes the methods used to build the analytical framework. We describe our findings in applying the framework to the two case studies in Sections 4 for refugees and in Section 5 for Brexit. Finally, Section 6 concludes the paper and details future work.

2. Related work

Using Twitter for opinion mining and user polarization is a vast subject [2]. The existence of polarization in Social Media was first studied by Adamic et al. [3] who identified a clear separation in the hyperlink structure of political blogs. Conover et al. [4] studied afterwards the same phenomenon on Twitter, evaluating the polarization based on the retweets. Most of the studies on polarization are still based on sentiment analysis of the content. The sentiment analysis methods proposed are numerous and they are mainly based on dictionaries and on learning techniques through unsupervised [5] and supervised methods (lexicon-based method [6]) and combinations [7]. Opinion mining techniques are widely used in particular in the political context [3] and in particular on Twitter [8]. Recently new approaches based on polarization, controversy and topic tracking in time have been proposed [9,10]. The idea of these approaches is to polarize users of a social network in groups based on their opinion on a particular topic and tracking their behavior over time. These approaches are based on network measures and clustering [9] or hashtag classification through probabilistic models [10] with no use of dictionary-based techniques.

Twitter is also exploited to better understand how the communication flows during political movements and events. Donover et al. study Twitter data covering the birth and maturation of the American anti-capitalist movement *Occupy Wall Street* [11]. The authors analyze the geo-spatial dimension of tweets in combination with the communication dimension building a geographic profile for the communication activity of the movement. An extensive analysis of these data produced many interesting results. For example, it appears that proximity to events plays a major role in determining which content receives the most attention in contrast to the stream of domestic political communication.

Moreover, there is a significant increase of interest in collecting and analyzing geo-located data from online social networks. Several works study different aspects of the geographical dimension of OSNs, a broad study on this argument is reported in [12]. Here, the authors propose a framework to compare social networks based on two new measures: one captures the geographical closeness of a node with its network neighborhood and a clustering coefficient weighted on the geographical distance between nodes. Twitter geo-located posts are studied to understand how Twitter social ties are affected by distance [13]. Linked users are identified as "egos" and "alters" and the distance between them is

[2] https://www.storify.com/ImagineEurope/what-is-associated-with-europe
[3] http://www.pol.ed.ac.uk/neuropoliticsresearch/sections/remote_content

Table 1
Notation and datasets statistics summary. The total number of Brexit-related tweets include those collected on the day of the vote, which have been excluded from the analysis.

Symbol	Description	Refugees	Brexit referendum		
		Total	Before	After	Total
$\mathcal{G}$	Collected English tweets	97,693,321	–	–	157,064,081
$\mathcal{T}$	Set of relevant tweets	1,238,921	724,399	3,331,058	4,343,548
$\mathcal{T}_{c+}$	Pro refugees/Brexit tweets	459,544	140,186	649,775	–
$\mathcal{T}_{c-}$	Against refugees/Brexit tweets	387,374	95,584	291,917	–
$\mathcal{T}_{ML}$	Tweets with mentioned location	421,512	–	–	–
$\mathcal{T}_{UL}$	Tweets with user location	101,765	231,873	969,071	1,302,981
$\mathcal{U}$	Users	480,660	368,732	1,467,600	1,725,523
$\mathcal{U}_{c+}$	Users with pro sentiment	213,920	55,470	380,956	–
$\mathcal{U}_{c-}$	Users with against sentiment	104,126	59,185	153,543	–
$\mathcal{U}_L$	Users with country location	47,824	42,075	151,569	453,375
t	Period of analysis (initial day - final day)	05-Aug-15 17-Sept-15	18-Jun-16 22-Jun-16	24-Jun-16 04-Jul-16	18-Jun-16 04-Jul-16

analyzed by considering the correlation with the air travel connection distance and with national borders and languages. An analogous objective is the focus of [14] where the authors infer the location of 12 million Twitter users in a world-wide dataset. Differently from the previous paper, they studied the correlation between the Twitter population and the socio-economic status of a country, suggesting that high developed countries are characterized by a larger Twitter usage. The geographical properties of Twitter are also useful to study the movements of people and migration phenomena. A study of mobility using geo-located Twitter messages is presented in [15]. The authors introduce a detailed study aimed at estimating international travelers based on the country of residence. They identify a number of characteristics including radius of gyration and mobility rate to describe the traveling phenomena thorough the Twitter lens. Zagheni et al. show how the analysis of 500,000 geo-located Twitter users may help to predict the migration turning points and to better understand migration in OECD countries [16]. The authors estimate the migration rate of users moving from one "home" country to another country. The reported results depict some interesting trends such as the decrease of migration from Mexico to US, consistent with official estimations.

The novelty of our proposal compared to the state-of-the-art approaches is mainly the fact that we propose an analytical framework to study a mass event from Twitter messages as a combination of three dimensions: time, space and sentiment. The sentiment analysis method adopted is efficient in tracking polarization over Twitter compared to other methods. Concerning other approaches for studying social phenomena, we do not base our analyses on the change of location of Twitter users to measure the flow of individuals through space, but rather we aim at understanding the impact on the EU citizens perception of migrants' movements and their resulting decision to vote for Brexit. Compared to other Twitter Brexit analysis available on the Web[4] our approach is general and reusable to assess the opinion of users in different contexts. Moreover we not base our analysis on a set of manually extracted keywords but we rely on a method which allows to rigorously extract the most informative concepts to monitor user opinion. As a further contribution our paper highlights a correlation between the migrants and the Brexit polarity in different UK cities, finding how the pro-Brexit negatively correlates with the pro-refugees.

3. The analytical framework

In this section we introduce the analytical framework by detailing how data collection and enrichment steps are performed in order to extract, for a given polarizing topic, a dataset of relevant tweets. During this process we capture the analytical dimensions we are interested in, namely the spatial, temporal and sentiment dimensions. The multi-dimensional dataset obtained at the end of the collection and enrichment steps can be analyzed and queried along the spatial, temporal and sentiment axes and, more interestingly, on combinations among them. The characteristics of the datasets collected and the notation used in the following are summarized in Table 1. The first dataset refers to the *Refugees crisis* and contains about 1.2 M tweets, while the second one refers to the *Brexit referendum* and contains about 4.3 M tweets.

We used the Twitter Streaming API under the *Gardenhose* agreement (granting access to 10% of all tweets) to collect the English tweets posted in two periods: from mid August to mid Sept 2015 for the refugees dataset, and from mid June to the beginning of July 2016 for the Brexit dataset, respectively. We first filtered out the tweets not related to the specific events analyzed. To this end, we simply chose two subsets of frequently-used hashtags and keywords specifically related to the refugees (208 hashtags) and the Brexit (111 hashtags and keywords) case studies.

We selected as *relevant* all the English tweets containing at least one of these hashtags. The filtered collections of relevant tweets are denoted in the following as $\mathcal{T}$. Then we enriched $\mathcal{T}$ by associating with tweets, when possible, information about their spatial, temporal, and sentiment dimensions. The characteristics in terms of number of users and polarity of tweets in the resulting datasets are detailed in Table 1. The methodologies used for the enrichment steps are discussed in the following two subsections.

3.1. Spatial and temporal dimensions

For each tweet we extract (when present) the *user location* of the person posting the message. The *user location* is structured in two levels: the city (when present) and the country. The user city is identified from the *GPS coordinates* or *Place* field, when available. Since GPS and *Place* data are quite rare (less than 2% of the relevant tweets) we also used the free-text *user location* field to enrich location metadata. We identified the locations in the user generated field based on data from the Geonames[5] dictionary which fed a "parsing and matching" heuristic procedure. This technique provides high-resolution, high-quality geo-location in presence of meaningful user location data [17]. The user country is collected in a similar way. When the country is not explicitly present we infer the country from the city. In the case of the refugees dataset we also extracted the *mentioned locations* within the tweet text. We used the mentioned locations to correlate the phenomenon of

[4] http://www.pol.ed.ac.uk/neuropoliticsresearch/sections/remote_content

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

Table 2
Seed hashtags used in *PTR*.

Refugees Dataset		Brexit Dataset	
$H^0_{c_+}$ Pro-Refugees	$H^0_{c_-}$ Againts-Refugess	$H^0_{c_+}$ Pro-Brexit	$H^0_{c_-}$ Against-Brexit
`#refugeeswelcome`	`#refugeesnotwelcome`	`#voteleave`	`#voteremain`
`#refugeesnotmigrants`	`#migrantsnotwelcome`		
`#welcomerefugees`	`#norefugees`		

the refugees with their locations (origins, destinations or significant cities in their routes or countries involved in the migration crisis). Also the *mentioned locations* are represented both at the city and the country level and they are extracted from tweets' text using the same heuristic procedure adopted for user location.

For both datasets we limit our analysis to the perception and sentiment of European citizens. The volume of tweets with user location $\mathcal{T}_{UL}$ and mentioned locations $\mathcal{T}_{ML}$ is reported in Table 1.

Finally, we extract the publishing time from each tweet and the period of time when each user was active. This information is necessary to study the temporal dimensions of social phenomena.

3.2. Sentiment dimension

We are interested in understanding if the user has a positive feeling towards the analyzed social event or if he/she mainly expresses negative feelings, antagonist and opposing ideas. In particular, for the *Refugee crisis* case study, we look at opinions of the users about the migrants. For the *Brexit referendum* case study, we split the dataset into two parts: posts made in the period before the Brexit referendum and posts made in the days after the vote. The sentiment analysis procedure is applied independently on the two parts of the dataset. It is worth noticing that the polarization registered *before* the referendum is connected to the vote intentions, while the discussion afterwards is more focused on the results: it reflects the impact of the pro-Brexit results (sentiment +) or the reasons why the *remain* supporters lost (sentiment -). Therefore, the pre and post datasets are enriched with information about the sentiment for both tweets and users ($\mathcal{T}_{c+}$, $\mathcal{U}_{c+}$, $\mathcal{T}_{c-}$ and $\mathcal{U}_{c-}$).

In both case studies, we consider two polarized classes $c \in \mathcal{C}$: *pro-refugees* / *pro-Brexit* (c_+) and *against-refugees* / *againts-Brexit* (c_-). We used our *PTR* (Polarization Tracker) algorithm [10] to assign a class to each polarized tweet and to each polarized user in an iterative way by considering his/her tweets and the hashtags contained. The approach proposed in [10] is suitable to track polarized users according to a specific topic.

Algorithm 1 *PTR* Algorithm.

Require: a set of users $\mathcal{U}$, their tweets $\mathcal{T}$with hashtags $\mathcal{H}$, a set of hashtags H^0_c for each class $c \in \mathcal{C}$
Ensure: Classification of users U_c and hashtags H_c

```
procedure PTR({H_c^0}_{c∈C})
    τ ← 0
    repeat
        {T_c^τ}_{c∈C} ← TwClass({H_c^τ}_{c∈C}, T)   ▷ Classify tweets on the basis of hashtags
        {U_c^τ}_{c∈C} ← UsClass({T_c^τ}_{c∈C}, U)   ▷ Classify users on the basis of tweets
        {H_c^{τ+1}}_{c∈C} ← HtClass({U_c^τ}_{c∈C}) ▷ Find better hashtags on the basis of U_c^τ
        τ ← τ + 1
    until convergence
    return {U_c^τ}_{c∈C}, {T_c^τ}_{c∈C}, {H_c^τ}_{c∈C}
end procedure
```

The pseudo-code reported in Algorithm 1 illustrates the procedure. The algorithm receives as input an initial set of polarized hashtags $\{H^0_c\}$ (initial *seed*) for each class c and the collection of relevant tweets $\mathcal{T}$. The initial seeds have been selected by analyzing the most frequent among the relevant hashtags used in the datasets. Among them, we selected two set of polarized hashtags which indicate the sentiment of the two opponent parts. The initial seeds used in the two datasets are reported in Table 2.

In the *refugees* dataset the hashtags in the seed set $H^0_{c_+}$ occur in 36K tweets, whereas $H^0_{c_-}$ hashtags are used in only 2K tweets. In the *Brexit* dataset, the initial seeds $H^0_{c_+}$ are used in 60K tweets before the referendum and in 12K tweets after, whereas $H^0_{c_-}$ in 21K tweets before the referendum and in 4K tweets after. One of the benefits of *PTR* is that after only a few iterations the results are less dependent on the size of the original seed, correcting the unbalanced number of occurrences per class. The final polarized tweets ($\mathcal{T}_{c+}$,$\mathcal{T}_{c-}$) reported in Table 1 are indeed more balanced than the seeds.

The internal functions of Algorithm 1 are defined as follows:

- TwClass: a tweet is polarized to one class c only if it contains only hashtags of one class $\{H_c\}_{c\in\mathcal{C}}$.
- UsClass: a user is polarized to one class c only if his polarized tweets of class c are at least twice the number of his polarized tweets of any other class.
- HtClass: a hashtag h is assigned to one class c if $S_c(h) > S_{c'}(h) \quad \forall c' \neq c$. The candidate hashtags h are chosen among all the hashtags used by at least two polarized users in their tweets, which have been filtered in order to be relevant to the topic of the analysis as it has been discussed in the data section.

The score is defined as follows:

$$S_c(h) = \frac{|\mathcal{T}_{\ h} \cap \mathcal{T}_{\ U_c}|}{|\mathcal{T}_{\ U_c}|} \cdot \prod_{c'\in\mathcal{C}\ ,c'\neq c} \left(1 - \frac{|\mathcal{T}_{\ h} \cap \mathcal{T}_{\ U_{c'}}|}{|\mathcal{T}_{\ U_{c'}}|}\right)$$

The score indicates the property of seeing the hashtag h among the tweets of the polarized users $\mathcal{T}_{\ U_c}$ predominantly for one class c. For our experiments we considered only hashtags with a score > 0.005: the threshold for the score $S_c(h)$ has been chosen after an empirical evaluation.

The procedure adds information about the polarization of users through the extension of the polarized hashtags by the analysis of all the tweets written by an already polarized user. First, the polarized tweets are identified through polarized hashtags. Then polarized users are labeled by means of polarized tweets and after the new polarized hashtags are extracted. This can be done by checking if they are representative of a sentiment class from all the relevant tweets written by polarized users. The procedure includes the relevant and not polarized tweets that have not been used to classify the users. It means that, among all the tweets written by a polarized user related to the topic, it may happen that only a few tweets are labeled as polarized. The remaining of tweets are used together with the polarized ones to extract the new hashtags that might be representative of a sentiment class. The iterative procedure has been run four times until the convergence was

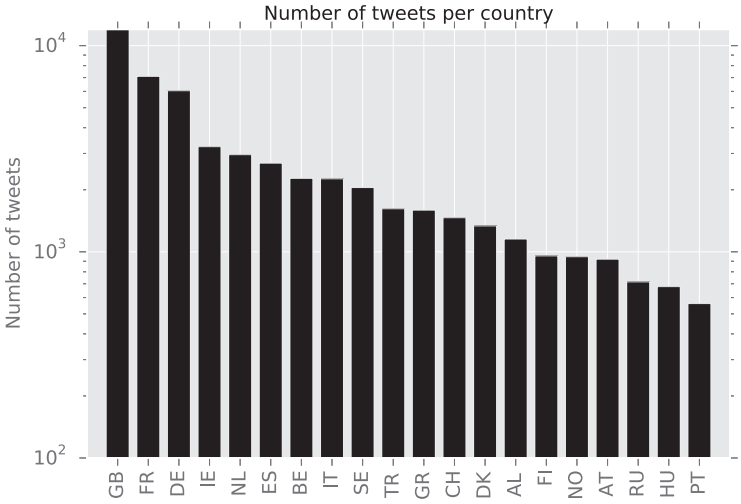


Fig. 3. $\mathcal{T}_{UL}$ per top-20 countries in log scale (refugees).

reached: i.e. the procedure does not find new polarized hashtags in the further iteration of the algorithm. We excluded from the hashtags retrieved by *PTR* all the hashtags which directly mention a city or a country. This is done to keep the sentiment independent of the location in the computation of the polarization. From empirical evaluation the presence of locations in the polarized tweets affects negatively the analysis since it assigns a specific sentiment to all the tweets regarding a place.

4. Perception of the refugee crisis in Europe

Our study is driven by the following analytical questions:

Refugees-AQ1: What is the evolution of the discussions about refugees migration in Twitter?

Refugees-AQ2: What is the sentiment of users across Europe in relation to the refugee crisis? What is the evolution of the perception in the countries affected by the phenomenon?

Refugees-AQ3: Are users more polarized in the countries that are most impacted by the migration flow?

4.1. *Applying* PTR *to data*

Note that the algorithm may classify both users and tweets as non polarized, thus favoring accuracy of truly polarized content. The polarization algorithm was able to assign the sentiment to 68% of the tweets and to 66% of the users in our dataset. Regarding EU-geolocated tweets and users, the algorithm assigned the sentiment to 73% of tweets and to 71% of the users. The sentiment analysis has been performed through PTR [10] since this method does not need external dictionaries or supervision, and provides a classification of polarized users in a flexible way by looking at terms used by members of different opinions. The method is fast and scalable and in [10] it is proved to be accurate, providing an improvement over the baseline from 7% to 71% on different datasets.

4.2. *Spatial and temporal analysis*

We navigate our multidimensional dataset by first analyzing the spatial and temporal dimensions to answer Refugees-AQ1. These analyses can quantify the volumes of relevant Twitter messages based on the countries of the users and the country mentions, since these volumes are strong indicators of real-world events [18]. Note that this spatial and temporal analysis is not considering the user polarization and it is therefore conducted on the full set of tweets available. Fig. 3 depicts the total number of tweets for the 20 most active countries. Since the dataset is in English most of the tweets (56.1%) come from users located in United Kingdom (UK), therefore in Section 4.4.3 we focus our analysis on UK. Nevertheless, a significant fraction of the data comes from other countries, e.g., France (FR) accounts for 6.9% of the tweets and Germany (DE) accounts for 5.9%. Without loss of generality, our methodology can be extended to other languages by simply extending the seed hashtags used in the sentiment dimension construction. As far as the mention location is concerned, we see that users from 51 countries mention 154 countries in Europe, Asia and Africa. This is analyzed in further detail in Section 4.4.3.

Fig. 4 illustrates tweets volumes along the temporal dimension, and relates volumes to events, as summarized in Table 3. The volume includes all the tweets in $\mathcal{T}$ and not only the geo-located ones. We observed significant volume peaks in days August 19, September 4 and September 19. As we can see from Table 3, these days match the major events like the UK and France security deal signed on the 18th of August regarding Calais, the drowned Syrian boy found on the beach in Greece, Hungary taking refugees to Austrian border by bus during September 2nd to September 4th, and migrants breaking through Hungarian border on September 16.

Next, we analyze location mentions to countries related to the refugee migration. Fig. 5 reports the volumes of tweets mentioning the EU countries most impacted by the refugees route, namely Austria, Germany, Croatia, Macedonia, Hungary, Serbia, Greece. We see that there is an interesting correspondence between the peaks of mentions and the events timeline. An evident peak for Germany, Austria and Hungary is the first week of September, probably related to the news of borders being opened to refugees. We also notice a peak of mentions of Croatia corresponding to the closing of borders with Serbia. Macedonia also sees an important increase of mentions around the 20th of August, probably in relation to the Macedonian Police using tear gas on refugees.

Similarly, Fig. 6 focuses on the mentions of relevant non European countries. The number of tweets mentioning Syria increases dramatically after the aforementioned facts of September 4. Turkey also has a peak on the 4th of September, probably due to the Alan Kurdi news. We also observe how the mentions to other countries remain more or less stable along this period to witness the fact that they were not directly related to the events reported by the media in that period involving mainly the Syrian refugees topic.

4.3. *Content-based analysis*

In this section we show a study related to the tweet content in terms of hashtags. In a first analysis we correlate the hashtags

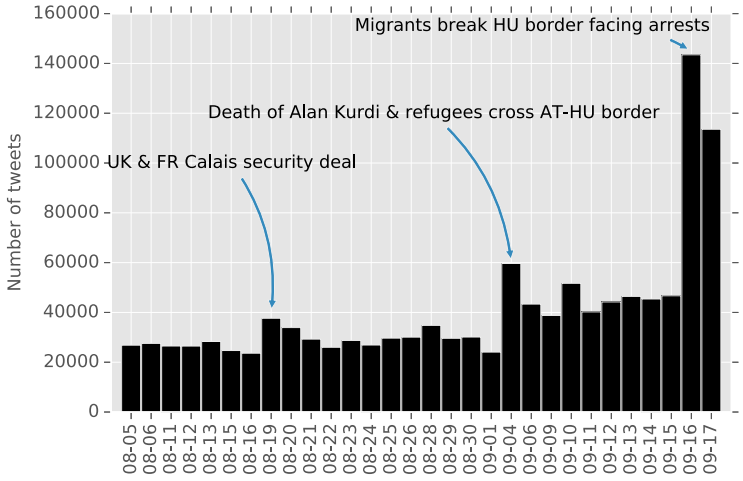


Figure 4: $\mathcal{T}$ per day and news headlines.

Fig. 4. $\mathcal{T}$ per day and news headlines.

Table 3
Major events during the observation period as reported by UK Newspapers.

18.08	UK and France sign the Calais security deal.
20–21.08	Macedonian police have used tear gas with thousands of refugees crossing from Greece and declares state of emergency.
27–28.08	71 dead refugees found dead in truck in Austria.
31.08	Angela Merkel: Europe as a whole must help with refugees.
1.09	Hungary closes main Budapest station to refugees.
2.09	Alan Kurdi drowned off the shores of Turkey.
4–6.09	Migrants are allowed to cross the Austro-Hungarian border; Refugees welcomed warmly in Germany.
8.09	Hungarian Journalist appears to kick and trip fleeing refugees.
14.09	Austria followed Germany's suit and instituted border controls; Refugee boat sinking
15.09	Croatia starts to experience the first major waves of refugees; Hungary announces it will start arresting people crossing the border illegally.
16.09	Refugee crisis escalates as people break through Hungarian border;
17.09	Croatia decides to close its border with Serbia.

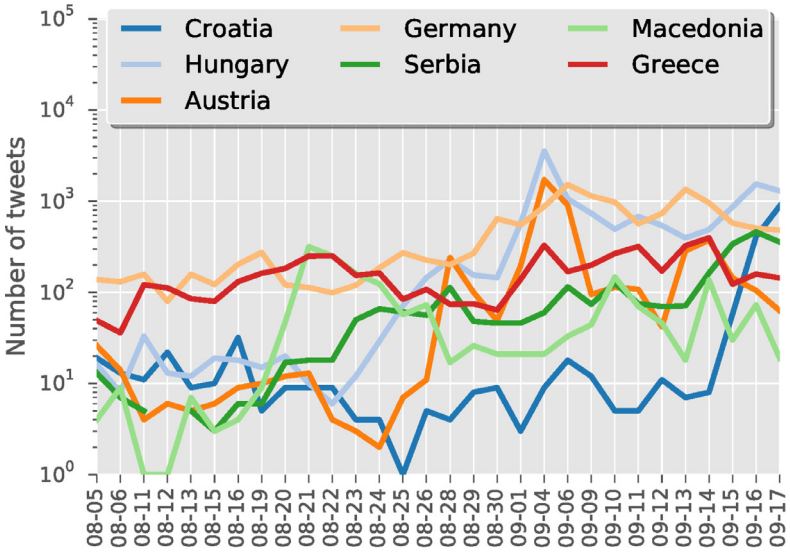


Fig. 5. EU country mentions per day in log scale.

usage across the whole Refugees dataset to the events occurring within the observed time frame. We measured the frequency of each hashtag per day, then performed a two-pass normalization. First we normalized the frequencies of hashtags on each day so as to avoid that days with lower recorded traffic are given less importance. Then, we normalized each hashtag over the observed period, so that the values are comparable among different hashtags. We then measured the variance of the normalized frequencies, considering that hashtags with higher variance are those with a more unbalanced distribution among days. The hypothesis is that the unbalanced distribution is due to a close relation of the hashtag with a specific temporal event (usually one or two days). Fig. 7 shows the resulting twenty highest-variance hashtags. The plot shows how this simple method allowed us to quickly spot hot topics in the observed stream of tweets and to correctly place them in time. The top ranked hashtag mentions a 2013 event that is

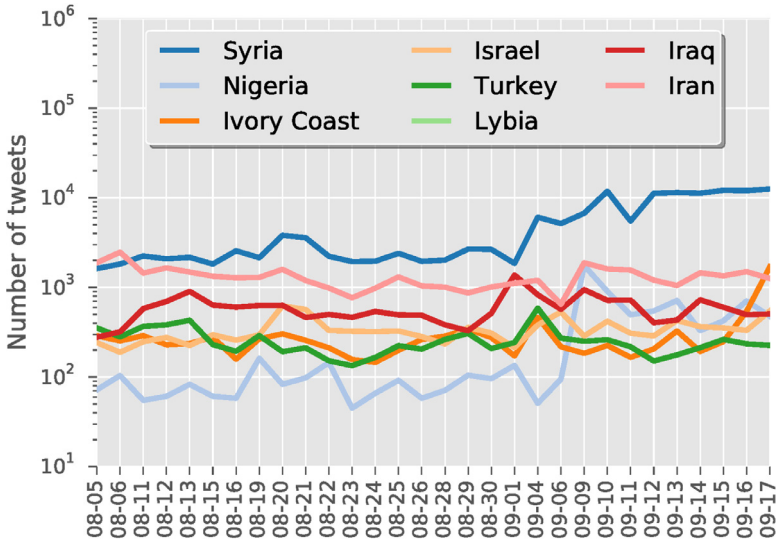


Fig. 6. Non-EU country mentions per day in log scale.

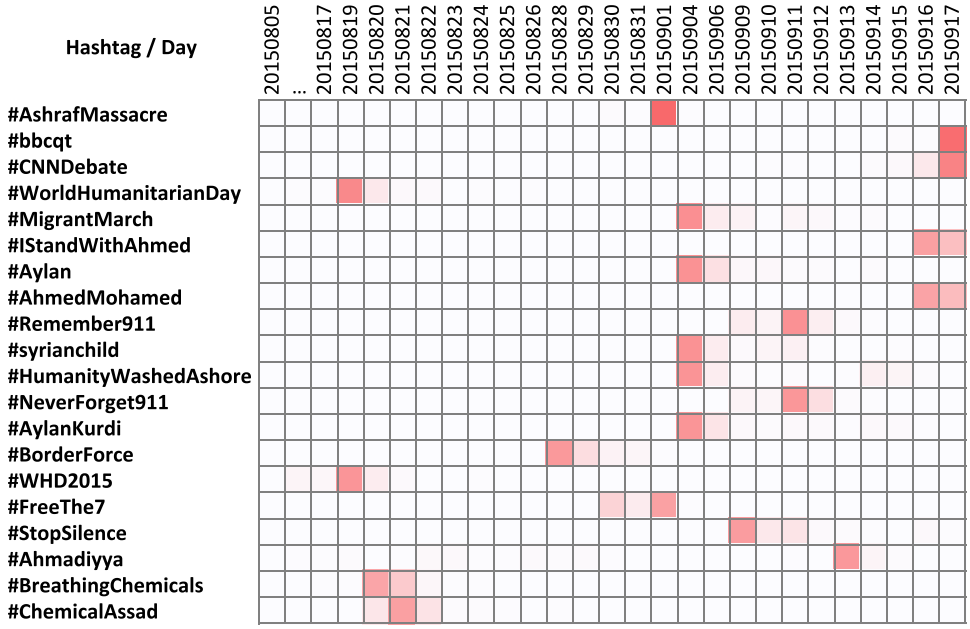


Fig. 7. Highest-variance hashtags per day; intense red represents higher relative frequency.

related to refugees, yet completely unrelated to the Syrian crisis. On 9/11 there are hashtags about the Twin Towers attack of 2001. The second and third hashtags in the ranking mention debate events related to the refugee crisis. A large group of hashtags are related to the Alan Kurdi death. A couple of hashtags mention Ahmed Mohamed, a 14 years old boy of Muslim faith arrested at school in the USA, when a teacher confused Ahmed's do-it-yourself clock with a bomb. Although some of the events are not directly connected to the refugees crisis, all of them are relevant to the debate.

We then focused on the usage of polarized hashtags as extracted by the *PTR* algorithm. We report in Table 4 the most relevant retrieved hashtags in addition to seed ones after the final iteration of *PTR*. From the analysis of the extracted hashtags we can see that people with a positive sentiment prefer to use the term *refugees*, while people with a negative sentiment refer to them as *migrants*, thus minimizing the fact that they are escaping war and persecution. Users with a negative sentiment frequently use *refugees* and the Islamic religion together, somehow correlating, in a prejudicial way, refugees with Islam and terrorism. Finally, we observe that individuals with negative sentiment are often patriotic and not pro Europe.

4.4. Sentiment analysis

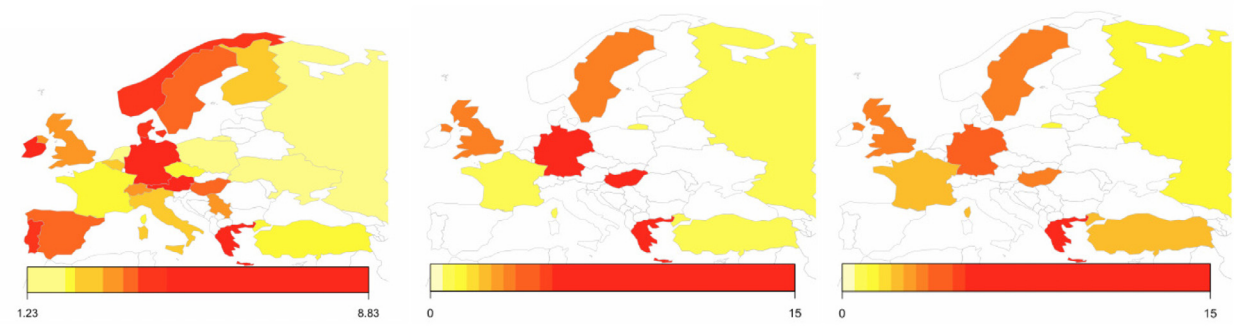
To answer the analytical question AQ2, we analyze the perception of the refugee crisis phenomenon by the European countries by exploiting the sentiment and location dimensions of the Twitter users in our dataset. To simplify the notation in the following we refer to $\mathcal{U}_L$ simply by $\mathcal{U}$. Let us define the ratio ρ between polarized users, namely the number of *pro refugees* users and the number of *against refugees* users:

$$\rho = \frac{|\mathcal{U}_{c_+}|}{|\mathcal{U}_{c_-}|}$$

The index ρ gives a compact indication of the sentiment of a group of users. In the following, we first analyze the sentiment

Table 4
Polarized hashtags discovered by the *PTR* algorithm.

Classes	Polarized Hashtags
Pro-refugees $H_{c_+}^{\tau=final}$	#campliberty #health #humanrights #marchofhope #migrantmarch #refugee #refugeecrisis #refugeemarch #refug
Against-refugees $H_{c_-}^{\tau=final}$	#alqaeda #guns #illegalimmigration #illegals #invasion #isis #islamicstate #justice #migrant #migrantcris



(a) Global perception (b) Internal perception (c) External perception

Fig. 8. Index ρ across European countries: red corresponds to a higher predominance of positive sentiment, yellow indicates lower ρ. (a) Refers to the whole dataset. (b) Is limited to users when mentioning locations in the their own country. (c) Is limited to users otherwise.

across countries, and then we differentiate discussions referring to internal versus external locations. The analysis was conducted by exploiting the polarized tweets and users data produced by the *PTR* algorithm from the whole Refugees dataset.

4.4.1. Sentiment by country

Fig. 8(a) shows the value of ρ for users belonging to the different European countries. We observe that Eastern countries in general are less positive than Western countries. In particular, Russia and Turkey have a low sentiment index probably because they are highly affected by the flow of arrivals. On the contrary, countries like Germany and Austria are more positive and this can be confirmed by the news reporting their decision of opening borders to migrants. Among western countries, France, UK, Italy and Netherlands have a low ρ index. In Italy the large amount of refugees arrived mainly through the sea directly from Lybia or Tunisia and the tone of the discussion is often characterized by negative notes. In France the sentiment expresses all worries about the situation of the "Calais jungle". The situation in Greece appears very different. The sentiment is positive even though this country remains by far the largest single entry point for new sea arrivals in the Mediterranean, followed by Italy. Greece captured the attention of humanitarian organizations. Countries like Ireland, Norway or Portugal are less interested by the phenomenon and therefore their perception might result more positive. Even for Spain ρ is not particularly low since the number of refugees coming from Western Mediterranean was low compared to central and eastern countries.

4.4.2. Internal and external country perception of the refugees crisis

In the following we study the perceived sentiment in relation to the country of the user. We denote as *internal perception* the sentiment of a user when mentioning his/her own country (or a city in his/her country). *External perception* on the contrary refers to polarized tweets with no references to his/her own country.

Fig. 8(b) shows the sentiment ratio ρ by country considering the internal perception, thus tweets mentioning the country itself. The ρ computation refers to the users of a country who mentioned in their tweets the country itself (or indirectly a city in the country). We report countries for which we have a sufficient amount of data. We can see that Russia, France and Turkey have a really low ρ index. We conjecture that the sentiment of a person, when the problem involves directly his/her own country, could be more negative since we are generally more critical when issues are closer to ourselves.

The external perception ratio is depicted in Fig. 8(c). Comparing the two maps, we see that internal and external perception is stable for UK and Sweden. Other countries have a much lower internal sentiment ρ than external, and this is the case of France, Russia and Turkey. All these countries were indeed facing many critical problems due to the arrival of refugees at their borders. The case of *Calais* is one of the most significant examples which could explain the case of the low ratio in France. Germany, Hungary and Greece, on the contrary, have a better internal perception which might be due to the decision of Germany to open borders to allow many people to transit from Hungary to Germany, releasing the extremely difficult situation at the national borders.

4.4.3. Sentiment analysis: the UK case

In this section we focus on the sentiment analysis of UK citizens, as UK is the most represented country in our dataset and therefore a more detailed sentiment analysis can be done. Indeed, for UK we detail the results at the granularity of largest cities by filtering the outcome of the *PTR* algorithm produced from the full Refugees dataset.

Fig. 9(left) shows the polarization in the most represented cities of the country, where at least 100 polarized users are present in the dataset. We can see from the heat map that there is a difference in the sentiment from north to south. This could be due to the fact that the cities in the south were more involved in the welcoming process of refugees and this might have generated more discontent. On the other hand Scotland shows a more positive perception of the refugees migration. From the time series of ρ for UK users we see an increase in the general sentiment ratio of the country after September 4. We found news[6] regarding that period from BBC and we think that the increase in the sentiment polar-

[6] Sept, 04: http://www.bbc.com/news/uk-34148913 – Sept, 16: http://www.bbc.com/news/uk-34268604

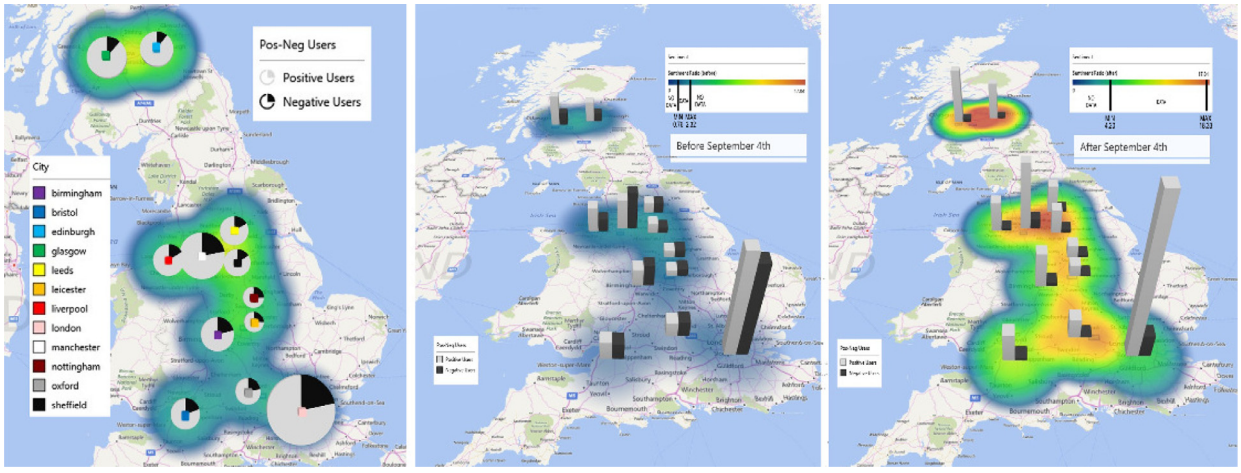


Fig. 9. Positive and negative users for different cities in UK in all period (left) before (center) and after (right) September 4. In the infographic the pies/bars show the number of polarized positive and negative users by city and the heat map in background indicates the value of ρ for the cities considered in the legend. For some cities the tweets are not sufficient to compute polarization, therefore when the heat degrades to 0 it indicates no data.

ization could be due mainly to the decision of the Prime Minister Cameron of acting with "head and heart" to help refugees. He allocated substantial amounts of money to humanitarian aid, becoming, at that time, the second largest bilateral donor of aid to the Syrian conflict (after the US). Fig. 9 (middle) and (right), shows the comparison of the opinion in UK before and after September 4, respectively. We highlight again a gradient of polarization from north to south in both cases even though the sentiment ratio ρ before and after that day is completely overturned. After September 4 the spreading of positive news in UK increases the sentiment and the volume of relevant tweets in all the country and probably government position reflects the sentiment of a vast majority of users which show support to refugees in their social media statements.

4.5. Mentioned location analysis

The last analysis we conducted aims at exploring AQ3 by studying the sentiment of the tweets when mentioning specific countries. We show how events impact differently the volume of tweets with positive or negative sentiment. Furthermore, we relate the sentiment changes to events.

Fig. 10(a–c) shows the sentiment of tweets when mentioning three of the countries most impacted by the refugees routes: Hungary, Austria and Croatia. We highlight an overall low number of mentions of these countries until the beginning of September. In the case of Hungary and Austria there is a sudden increase in the beginning of September in the overall number of mentions, predominantly for c_+ with a relative increase in c_-. This is mostly due to the overall positive sentiment towards the events from the previous days (the Alan Kurdi story), but also due to positive news about migrants being allowed to cross the Austro-Hungarian border. The negative sentiment appears, and continues to grow, until the middle of September when c_- tends to increase more than c_+, due to tweets expressing negative feelings towards border controls in Austria (13–15 Sept) and Hungary arresting refugees crossing the border illegally (15–17 Sept). Croatia comes into play towards the end of our observation period, when on September 16th it becomes a valid alternative to Hungary which closed its borders with Serbia. A similar analysis has been done for Greece, Macedonia and Serbia, but due to lack of space we are omitting here.

In Fig. 10 (d–f) we look at the sentiment in relation to mentions of UK, France and Germany. Both UK and Germany are rather balanced between positive and negative tweets. Germany presents exceptions on certain days when a positive feeling arises in support to the sad incidents related to refugees. We notice that the official media news at the end of August reported that Germany was welcoming refugees, while UK started showing a positive sentiment after the dramatic facts of Alan Kurdi and the announcement of welcoming 20,000 refugees by 2020. France seems to have more negative feelings, probably due to the difficult situation in Calais and the news about victims trying to cross the country, while a positive peak appears in correspondence to the young Alan Kurdi news.

5. The Brexit referendum

Similarly to the refugee related analysis, our study about the Brexit referendum is driven by the analytical questions below:

Brexit-AQ4: What is the evolution of the discussions in Twitter regarding the Brexit referendum?

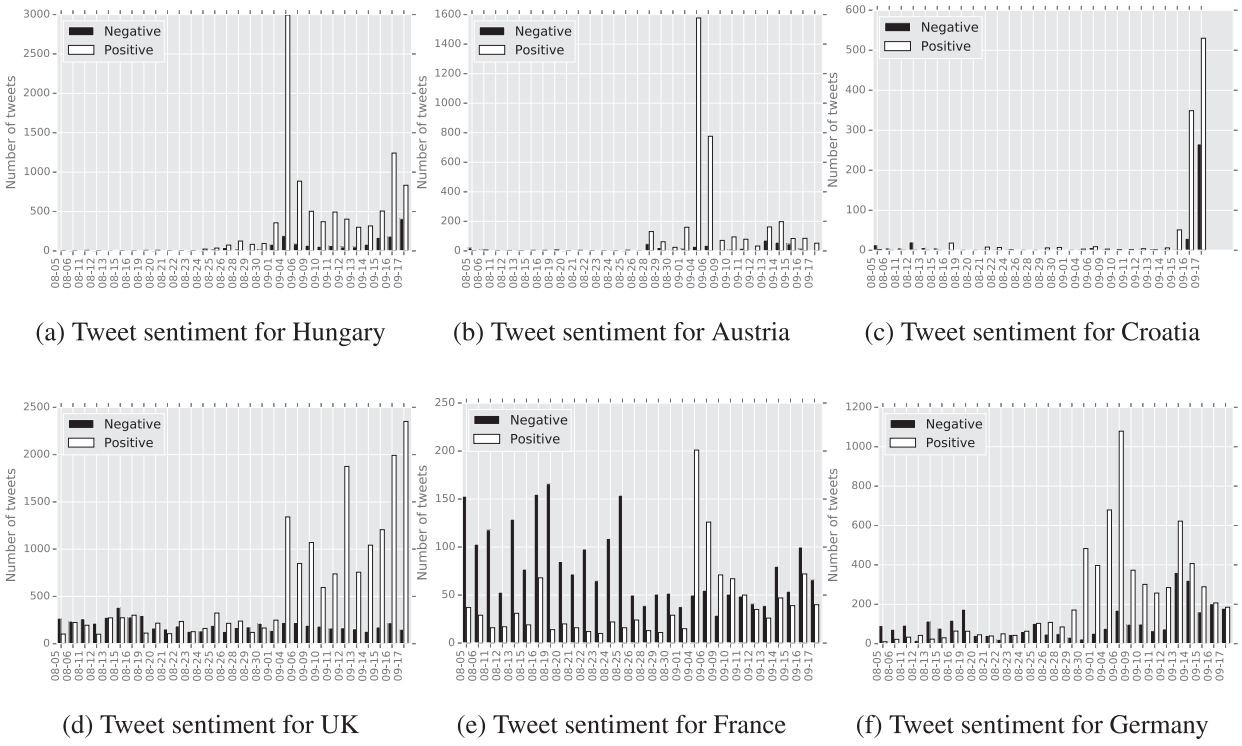
Brexit-AQ5: What is the sentiment of Twitter UK users on the Brexit referendum topic, before and after the vote? What is the perception in other European countries?

Refugees-Brexit-AQ6: Is the polarization of the users about refugees and the Brexit referendum somehow correlated? If so, how are the two topics correlated?

By applying our PTR algorithm (Algorithm 1) to the Brexit dataset we discover new hashtags starting from the selected seeds, namely: `#voteleave` and `#voteremain`. In Table 5 we report the most relevant hashtags found for each class c. The polarization algorithm was able to assign the sentiment to 31% of the users in the tweets before the referendum and 36% after. Most of the users are not polarized indicating that the dictionary used in their tweets does not allow the algorithm to decide one polarization class. Compared to the previous case study the polarization for a referendum is a more difficult task since the hashtags used are not very well separated.

The PRT algorithm however succeeds in finding relevant hashtag for both sides, pro and against Brexit. It is worth noticing that the hashtag `#brexit` is classified as part of the pro-Brexit (leave) hashtag set. However the meaning of `#brexit` changes after the referendum day, since the mentions of Brexit tend to refer to the referendum results instead of being a polarization in favor of leaving the EU.

The Before and After Brexit hashtag lists present some differences. While before the referendum date the discussion is centered



(a) Tweet sentiment for Hungary (b) Tweet sentiment for Austria (c) Tweet sentiment for Croatia

(d) Tweet sentiment for UK (e) Tweet sentiment for France (f) Tweet sentiment for Germany

Fig. 10. Tweet sentiment for country mentions per day.

Table 5
Polarized hashtags discovered by the *PTR* algorithm.

Classes	Polarized Hashtags
Pro-Brexit BEFORE $H_{c_+}^{final}$	`#brexit #bbcdebate #eu #uk #leave #bbcqt #leaveeu #takecontrol #inorout #projecthope #news #marr #bbc #r`
Against-Brexit BEFORE $H_{c_-}^{final}$	`#euref #voteremain #remain #strongerin #eureferendum #labourinforbritain #c4debate #votein #catsagainstb`
Pro-Brexit AFTER $H_{c_+}^{final}$	`#brexit #eu #uk #leave #euro2016 #leave #voteleave #indyref2 #engice #eng #toryleadership #ukip`
Against-Brexit AFTER $H_{c_-}^{final}$	`#euref #eurefresults #keepcorbyn #eureferendum #corbyn #remain #labourcoup #notmyvote #voteremain #corby`

around how to vote, after the results have been published we can see new emerging topics, highly related to Brexit: `#indyref2`, for example, refers to the fact that Scotland discusses a possible second referendum for independence from UK, as a consequence to the referendum results (we recall that Scottish people mainly choose to remain in EU). Another example is related to J. Corbyn (`#corbyn, #keepcorbyn, #corbynstays`) and the Labour party (`#labourcoup, #labour`). The politician threatened to block Brexit, but later had to clarify the Labour Party's position on Brexit, namely respecting the referendum results. We can see that before the vote day there are encouragement hashtags like `#strongerin, #greenerin, #intogether`, while after the vote the discussion reflects how part of the population feels unrepresented `#notmyvote`.

5.1. Spatial and temporal analysis

To answer Brexit-AQ4, we start by analyzing the spatial and temporal dimensions. Differently from the refugee crisis, the Brexit referendum involved directly only one country, the UK. However, the result of the UK vote impacted on the opinion of citizens of every European country thus bringing the discussion to all EU. Fig. 11 depicts, in a logarithmic scale, the total number of related tweets posted in the period observed for the 20 most active European countries.

Due to the clear UK focus, but also because we collected only tweets in English language, there is a huge difference between the number of tweets coming from the UK respect to other countries. After UK, we have Republic of Ireland (IE), a neighbor and English-speaking country.

We further look at the distribution of relevant tweets across the referendum period. In Fig. 12 we detect an increasing volume of tweets the days before the referendum, slowly declining the days after. The highest tweet volume is on June 24 when the results became official. The result stirred an influx of tweets from both UK and the other European countries. The following days we see a peak around June 27 when a EU summit took place. In this summit EU decided to refuse any informal agreement until the UK does not trigger article 50 of the Treaty, by issuing formal notification of its intention to leave.

An overview of the main events during our observation period are presented in Table 6. The analyses conducted in the following sections will bring more light towards the key dates and topics involved.

Another interesting view is to see what are the most popular hashtags related to the referendum during the selected period. We applied the normalized variance measure, as described

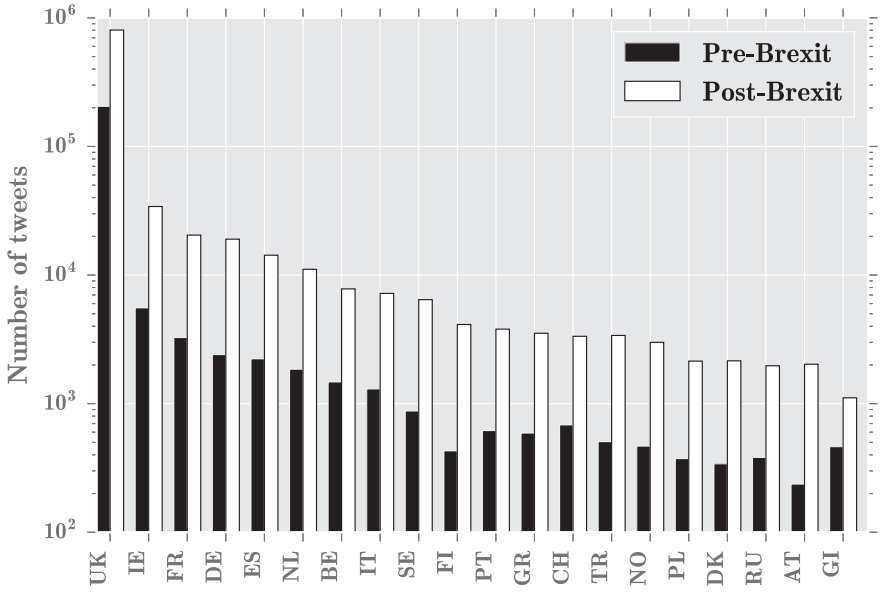


Fig. 11. $\mathcal{T}_{UL}$ per top-20 countries in log scale.

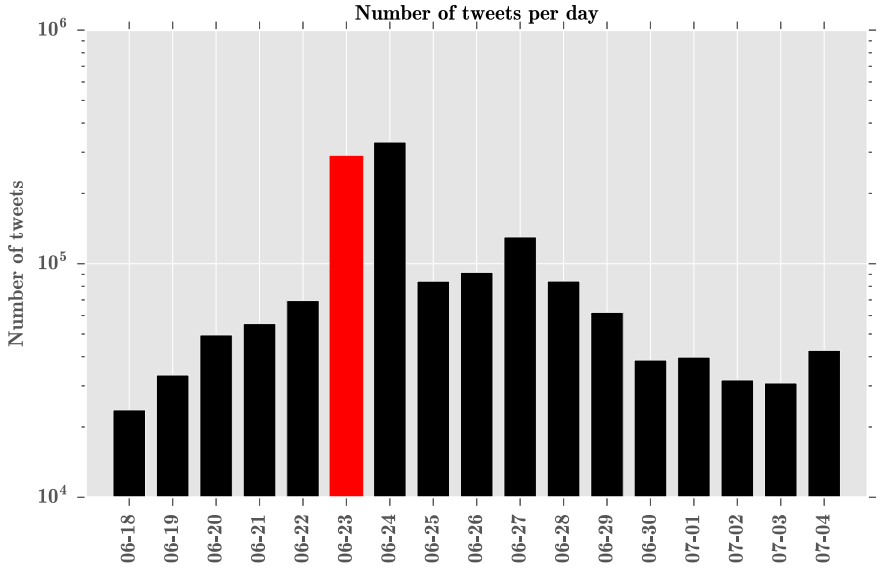


Fig. 12. Number of tweets per day.

Table 6
Major events during the observation period as reported by UK Newspapers.

06–17	The Times comes out in support of staying in the European Union
06–19	Andrew Marr Show BBCs final televised debate of the referendum campaign
06–21	EU referendum live: Khan accuses Boris Johnson of leading 'Project Hate' in BBC's Great Debate
06–23	**EU Referendum**
06–24	UK votes to leave EU after dramatic night divides nation
06–26	Nicola Sturgeon: Scottish parliament could block Brexit
06–27	EU may refuse informal Brexit talks until UK triggers article 50
	Boris Johnson holds a press conference in which it fails to provide a post-Brexit plan
07-02	Brexit live: thousands 'march for Europe' in post-referendum protest
07-04	Farage resigns
07-05	Brexit: May wins first round of voting

in Section 4.2, to identify the hashtags that are most related to events that have a precise location in time. In this case in Fig. 13, we separately ranked hashtags in the days before the vote and in the days after the vote. Before the vote most of the top-ranked hashtag are related to pre-vote debate events. Another relevant group of hashtags is related to the Euro 2016 soccer tournament, in which a comparison about leaving or remaining in the tournament and leaving of remaining in EU is often made. The top hashtags in the post-vote analysis are in fact related to the defeat that caused the England team to be knocked out of the Euro 2016 tournament. Other after-vote tags are those related to immediate reactions the

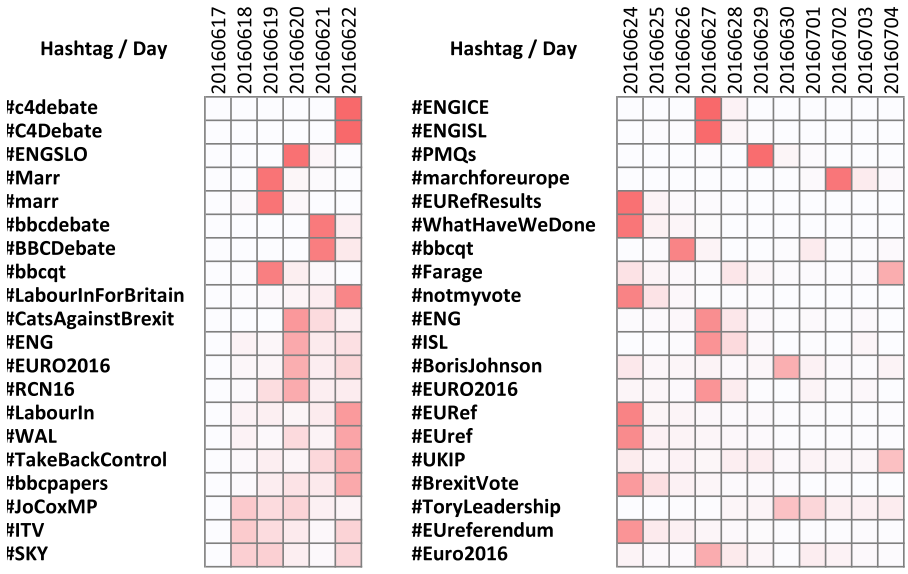


Fig. 13. Highest-variance hashtags per day. Intense red represents higher relative freq.

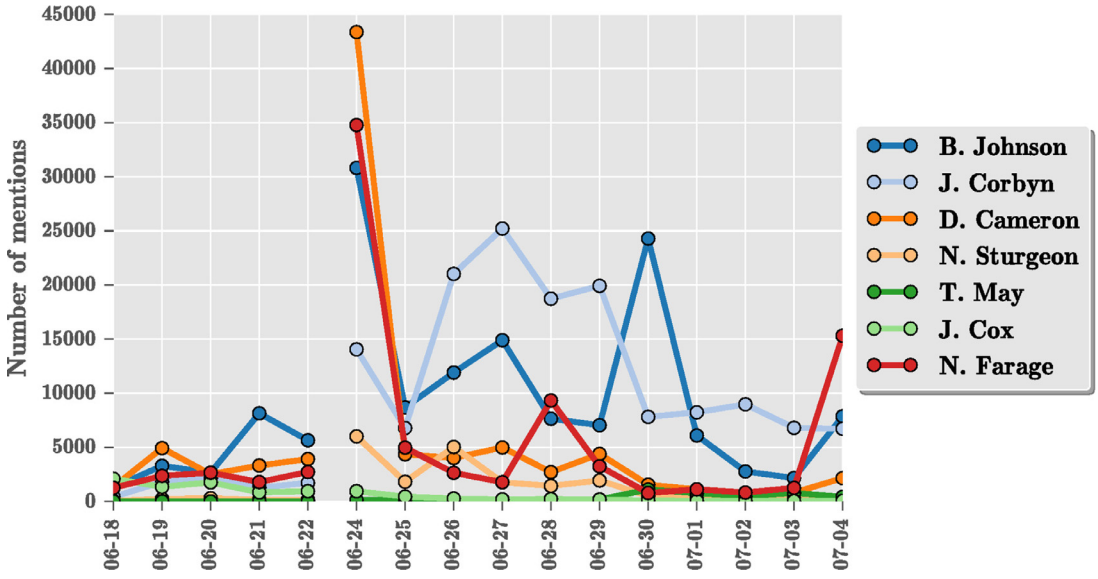


Fig. 14. Politician mentions before and after the Brexit referendum.

day after the vote and those related to the leadership changes in the Conservative Party.

Besides identifying the most popular hashtags, we also want to see who are the politicians and parties most frequently mentioned in our dataset.

As we can see in Fig. 14 before the referendum the politicians being mentioned are a few. However, some politicians are discussed, such as B. Johnson and D. Cameron, due to specific events. On June 19, D. Cameron participates in BBC's Question Time show, a special edition dedicated to the referendum, while on June 30, B. Johnson rules himself out of Conservative leader race, despite his active role in the leave campaign. After the vote, the mentions of politicians increase and show several peaks. For example we measure a peak of mentions for B. Johnson in concomitance with the press conference he held on June 26, after D. Cameron announcement of resignation as Prime Minister. The highest peak of mentions for the former mayor of London occurs however on June 30 when he announces that he will not join the race to be the next leader of the Conservative Party and prime minister. Mentions for J. Corbyn shows a similar peak on June 27 in concomitance with his possible resignation as leader of the Labour party.

We apply the same kind of analysis to the main political parties in UK (Fig. 15). We notice that the two most discussed are the Labour and the Conservative parties, followed by a peak on UKIP on the 4th of July, when its leader, N. Farage, resigned. This increase in the number of mentions to N. Farage is reported also in Fig. 14. Overall, we can observe that parties are not much discussed before the referendum, while after the referendum the mentions increase as leaders and parties started discussing their stand on the Brexit referendum results and possible consequences.

5.2. Sentiment analysis

To answer the analytical question Brexit-AQ5 about the polarization of Twitter users towards the leave and remain option of UK, we analyze the sentiment and spatial dimensions in the collected Brexit dataset. To be consistent with the previous case study on refugees, we refer to $\mathcal{U}_L$ simply by $\mathcal{U}$ and we consider *the ratio* $\boldsymbol{\rho}$

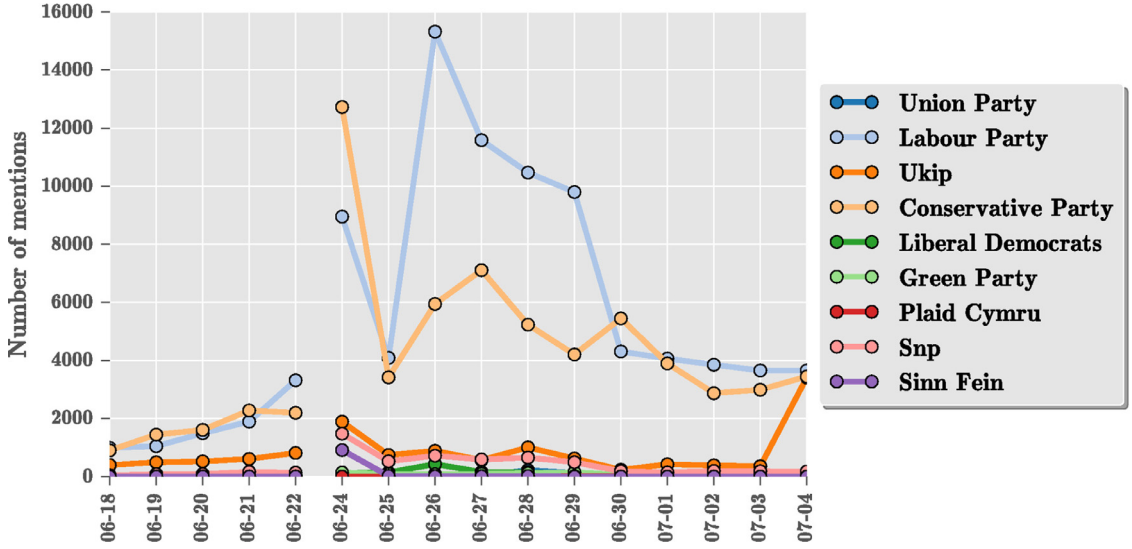
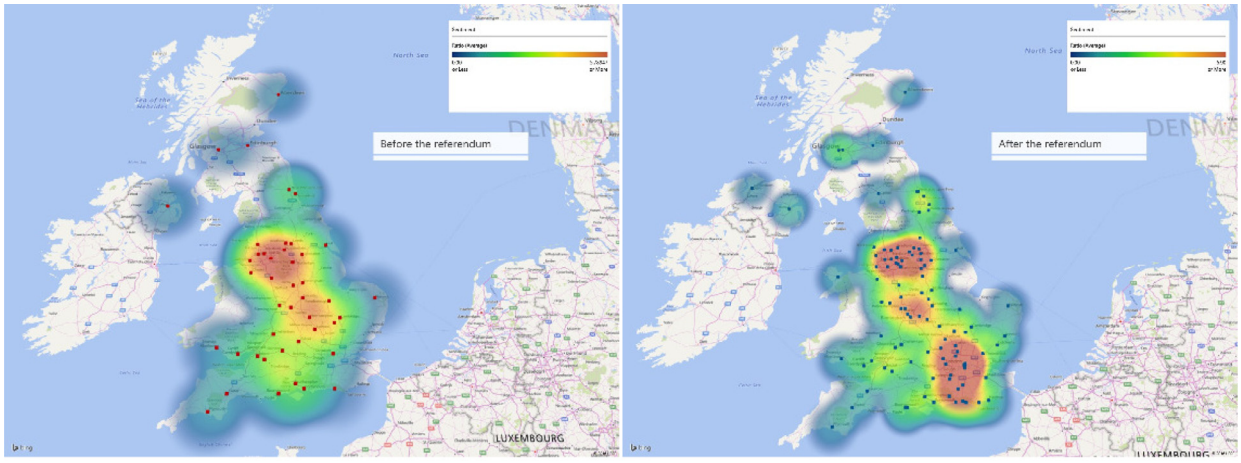


Fig. 15. Parties mentions before and after Brexit.



(a) Before the vote (b) After the vote

Fig. 16. Polarization index ρ across UK cities before and after the vote.

between ***pro Brexit users****over* ***against Brexit users*** as a measure for evaluating the polarity of a set of users. Here we split the dataset into two independent periods: before and after the vote day.

5.2.1. Sentiment analysis in UK

Since the referendum took place in UK we initially restrict the analysis only to UK users. In Fig. 16 the sentiment ratio ρ is reported as a heat map for the largest cities of the country. When we consider on the period before the vote (Fig. 16(a)) the ratio can be seen as an indicator of the vote intention of the users. This value is particularly high (pro Brexit) in cities in the central part of the UK (England), while cities in Scotland (e.g. Edinburgh) register the lower pro Brexit values.

By looking at the figure we can observe how the London area and its surroundings are not really polarized pro Brexit, thus confirming a slight predictive behavior towards the actual referendum outcome. What is most interesting is the comparison with the situation after the referendum analyzed in Fig. 16(b). It is clear that the polarization of the users towards pro Brexit moves from the center of the country to the London area. However, we have to keep in mind that the meaning of the polarization after the referendum is less related to the vote intention, but it is more related to the discussion about the vote outcome. In particular we can observe that the discussion about Brexit became heated among users around the capital, probably debating the results, not representative of the intention of the majority of the UK citizens in that area. The value ρ before the referendum can be in fact considered a weak indicator of the vote outcome - with all the limitations of using a social network to predict political events (biased towards young people and altered by propaganda and media). After the referendum, instead, the ρ measure indicates the places where the discussion and comments about Brexit became more intense.

5.2.2. Users through time: before and after the Brexit referendum.

Fig. 17 shows how users are polarized before and after the Brexit referendum. The pie chart refers to only 3.5% of the users who are polarized according to our sentiment enrichment procedure both before and after the vote. Unfortunately, the intersection between users present in our dataset before and after the vote is not large enough to perform a detailed analysis of users who may have changed polarity. However, we observe that 63% of the users polarized in both time frames keep their polarization (in particular those pro Brexit which are the 37%). The remaining 37% consists of users changing polarization from pro to against (11%) and from

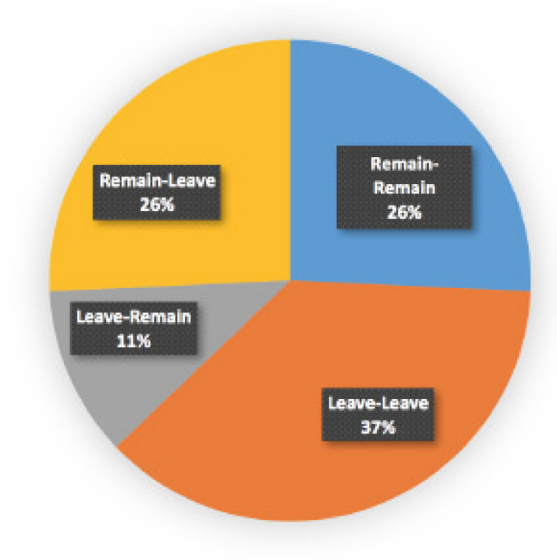


Fig. 17. User polarization before and after the Brexit referendum for those users who are polarized in both time frames for leave or for remain.

against to pro (26%). It is important to recall again that the polarization after the referendum is more related to the topic of discussion more than vote intention. For this reason, this change in the polarization means that after the vote users mainly discussed the result of leaving the EU rather than a change of vote intention. The actual intention to change the vote as regretting the actual result is probably present in reality, but hard to detect in our dataset.

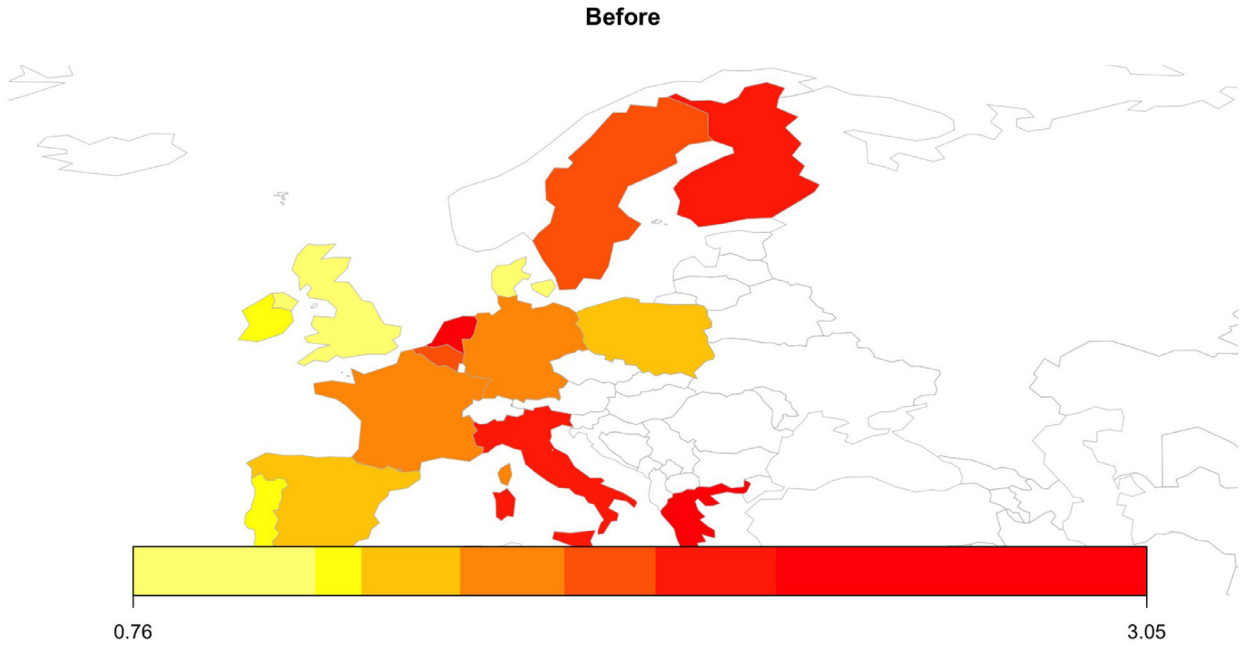
5.2.3. Sentiment of the European countries

In this section we discuss the opinion of the other European users about the Brexit result. Fig. 18 shows the value of the ρ indicator for users located in different European countries. In particular, we focus of the situation before Brexit to capture the perception of users outside UK. From the map we see that the sentiment in UK is quite against Brexit (in yellow) compared to other European countries more pro Brexit. In particular Italy, Netherlands, Finland and Greece have a very high ρ meaning that the users we have analyzed are more pro Brexit (in red). The interpretation of this value can be twofold: these countries are pervaded by a sentiment pro Brexit or simply they use a vocabulary which is anti-Europe and for this reason it is similar to the one used by pro Brexit people. Fig. 19 shows a comparison between the sentiment before and after the referendum for the European countries on a common color scale. The discussion about Brexit increased after the referendum also in countries where there were no particular discussions before the vote (e.g. Russia).

5.3. Mentioned concepts analysis

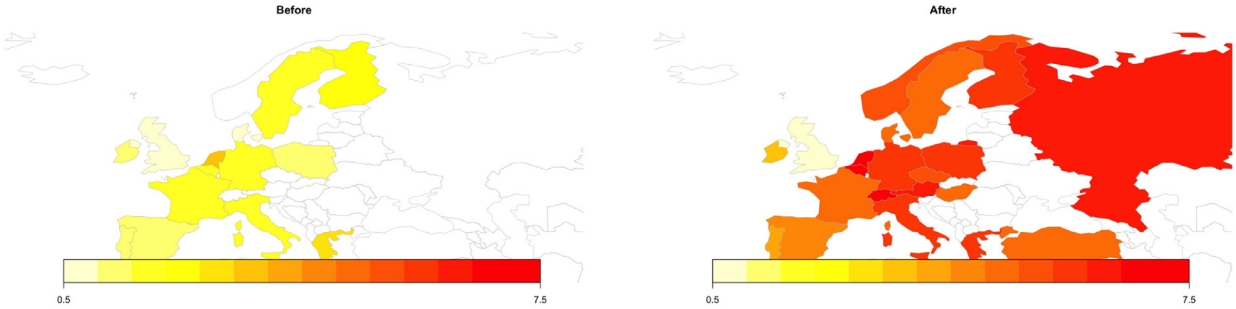
In this section we analyze how users talk about the Brexit referendum by considering geo-located tweets. In order to get a glimpse into the discourse of pro and against Brexit, we look at the hashtags selected as seeds and the final hashtags, extracted by the PTR algorithm for the two classes of users.

In Fig. 20 we plot the mentions of seed hashtags τ_0 (`#voteleave, #voteremain`) before and after Brexit, with continuous lines. We can observe an increased frequency the days before the referendum. However after the referendum the volume decreases significantly. By using the derived hashtags of the PTR algorithm, we are able to follow this topic through the following days. We cluster the final hashtags τ_{final} for each of the two classes, pro and against, and count the mentions. As we can see, although we follow the same topic, there is a much higher volume of mentions when we look at the set of final hashtags. The peak



Pre referendum situation by user location

Fig. 18. Index ρ across European countries: red is related to Brexit, yellow indicates the remain choice.



(a) Before the vote (b) After the vote

Fig. 19. Index ρ across European countries: red is related to Brexit, yellow indicates the remain. (a) refers to the data collected before the vote. (b) refers to the data after the vote.

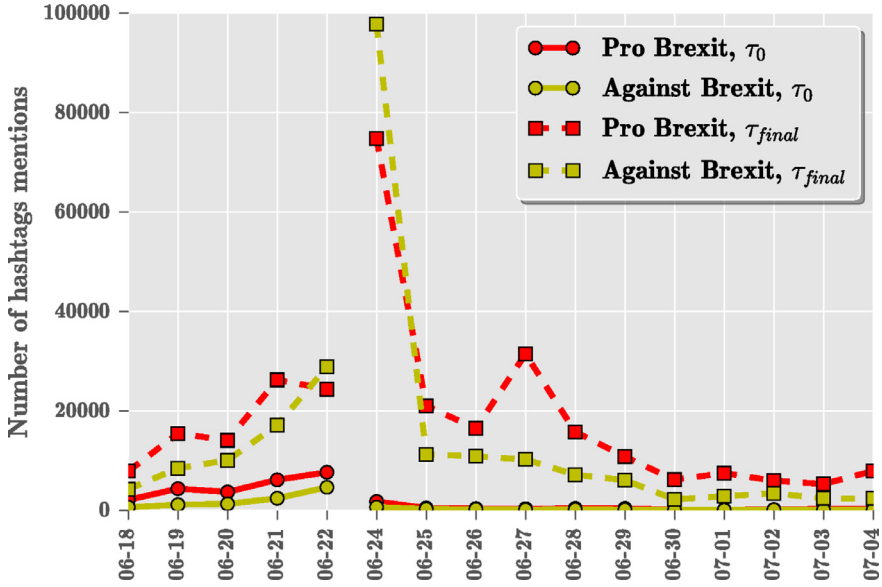


Fig. 20. The evolution of seed and final hashtags of PTR in time.

of discussion about Brexit happens on June 24, when the official result came out, and decreasing the following days.

Another interesting element to be mentioned is the inversion of popularity between pro and against Brexit on June 22, the day before the referendum. On that day, due to the worry of a possible Brexit result, there was an activation of the against Brexit side, pleading for voting remain, shift also present on June 24. The activity of the pro Brexit side remains persistent throughout the time. This is also due to the fact that after the referendum, users tend to refer to Brexit much more often, regardless of their sentiment.

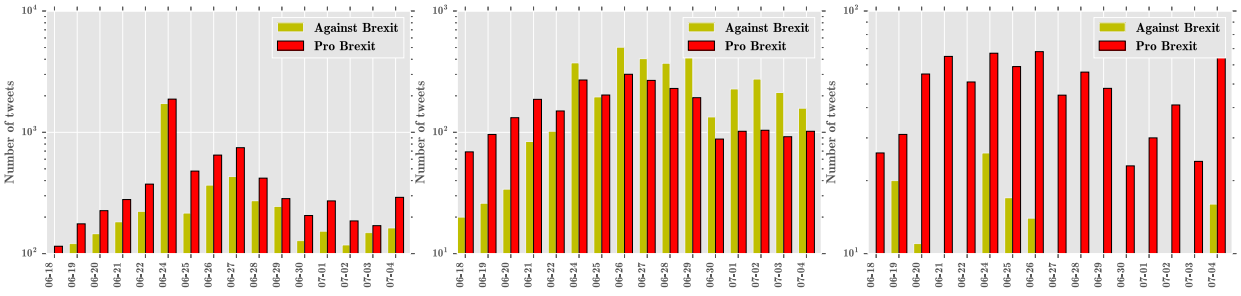
Fig. 21 reports the topics of discussion of the classified users. We look at the mentions of the most discussed political parties before and after the referendum. We plot the mentions to the Conservative, Labour and UKIP parties of pro and against Brexit users. We can see that there is a certain balance between the pro and against users when mentioning the Conservative party, with a slight dominance for the pro users. Alternatively, the Labour party is mentioned more by against users, especially after Brexit, probably due to the news regarding J. Corbyn's no confidence vote, and many users supporting him (`#keepcorbyn` etc.). On the other hand, UKIP is mostly mentioned by pro Brexit users.

We also investigated how users mention the representatives of the political parties. The results of the analysis are plotted in Fig. 22. We see that D. Cameron maintains a certain balance throughout the whole period, while J. Corbyn, similarly to the Labour Party in Fig. 21, seems to be mentioned mostly by against Brexit users. An interesting view regards UKIP leader, N. Farage, who is mentioned more frequently by the against users rather than the pro ones, most probably as a controversial figure. After the referendum both N. Farage and B. Johnson are mentioned mostly by pro Brexit users. The volume of mentions regarding these two leaders are however quite low respect to J. Corbyn.

5.4. Correlations in Refugees and Brexit discussions

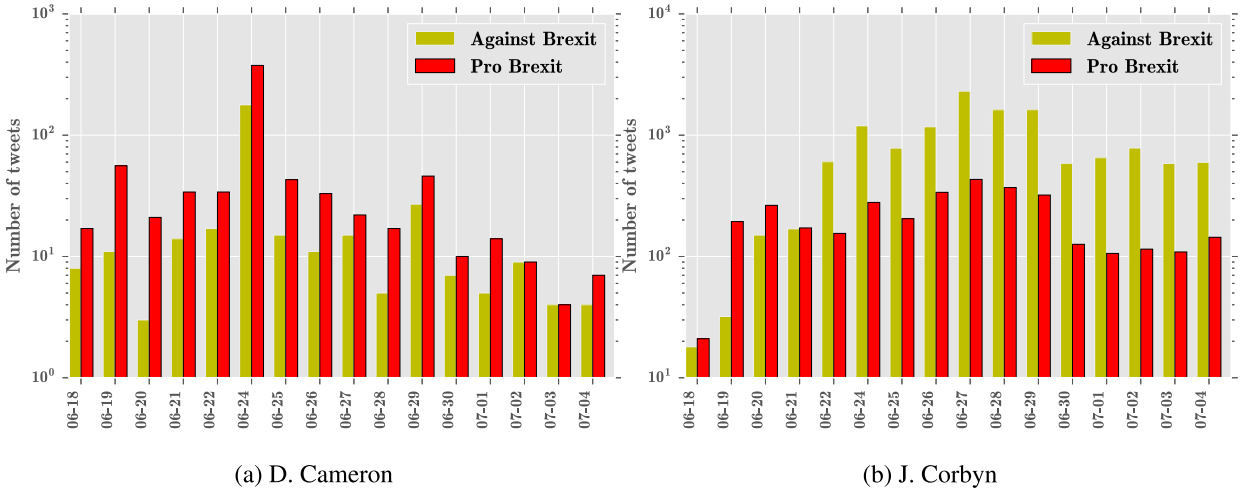
The analytical question Refugees-Brexit-AQ6 proposes a cross analysis of the two topics to see if and how they are correlated in Twitter in terms of polarization of the users. Indeed, we know that one of the main points of the pro Brexit campaign was the migrants crisis. How do these two topics relate in our Twitter datasets?

We selected in both datasets the polarized geolocated users in UK and we used the geolocation to extract the city information. We obtained a list of 16 major cities in UK for which we calculated ρ for the geolocated users, both for the refugees dataset and for the Brexit dataset. In particular, for each city we obtained ρ_R in relation to the polarized users in the refugees dataset, $\rho_{\overline{B}}$ in relation to the polarized users in the Brexit dataset before the vote

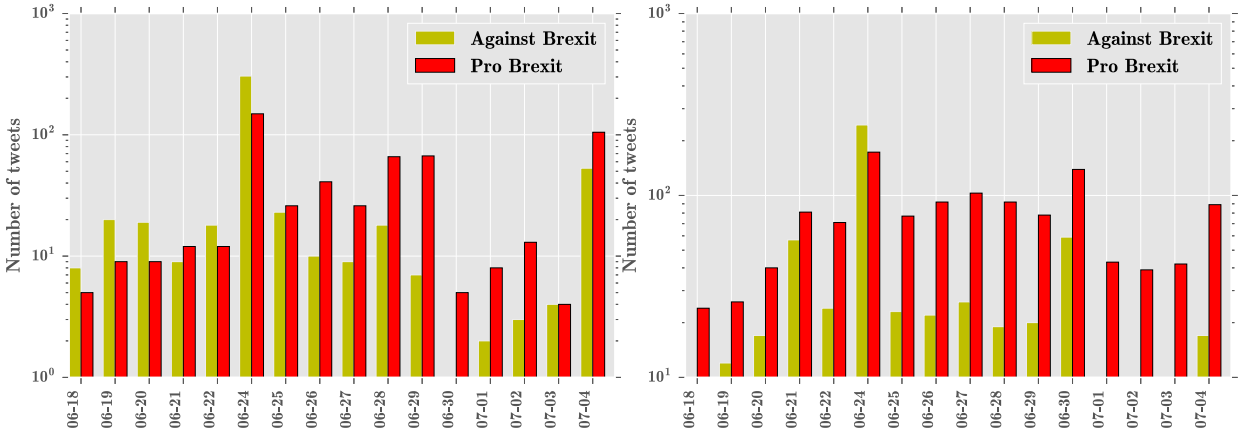


(a) Conservative party (b) Labour party (c) UKIP

Fig. 21. Sentiment for users speaking about political parties.



(a) D. Cameron (b) J. Corbyn



(c) N. Farage (d) B. Johnson

Fig. 22. Sentiment for users speaking about politicians.

and $\rho_{\overrightarrow{B}}$ in relation to the polarized users in the Brexit dataset after the vote. Moreover, for each city we report ρ_V which is the ratio between actual pro-Brexit citizens and against-Brexit citizens according to the official referendum results for that region.

In Table 7 we report six representative cities in UK, three with high ρ_V and three with low ρ_V index. To make the comparison easier we reported the $log_2(\rho)$ values. Positive values of $\rho_{\overleftarrow{B}}$, $\rho_{\overrightarrow{B}}$ and ρ_V indicate pro-Brexit attitude and negative values against-Brexit attitude; for ρ_R positive values mean pro-Refugees attitude, while negative ones against-Refugees attitude. This table highlights some interesting relations: cities in Scotland (e.g., Edinburgh or Glasgow) were against Brexit in the referendum (negative ρ_V) and, consistently, the scores related to the Brexit dataset of the polarized users before the election are negative and low, while the score ρ_R in both cities is positive and greater than for other cities, indicating a polarization pro-Refugees in both Edinburgh and Glasgow. On the other hand, cities like Nottingham which were pro-Brexit in the referendum, show a lower ρ_R and a higher $\rho_{\overleftarrow{B}}$. The polarization after the referendum in the dataset is more difficult to interpret since it is more related to the discussion of the vote

Table 7
Polarization of UK main cities according to our datasets (refugees, before Brexit, after Brexit). In the last column we report the official results of the referendum for the region of each city calculating the ratio of the positive votes over negative votes.

City	Region	Refugees $\log(\rho_R)$	Before Brexit $\log(\rho_{\overleftarrow{B}})$	After Brexit $\log(\rho_{\overrightarrow{B}})$	Official Brexit results $\log(\rho_V)$
Birmingham	West Midlands	1.92	−0.94	−0.35	0.52
Edinburgh	Scotland	3.07	−1.51	0.07	−0.91
Glasgow	Scotland	2.96	−1.18	0.00	−0.91
Leicester	East Midlands	1.62	−0.62	−0.51	0.52
London	Greater London	1.85	−0.55	0.28	−0.57
Nottingham	East Midlands	1.56	−0.37	−0.34	0.52

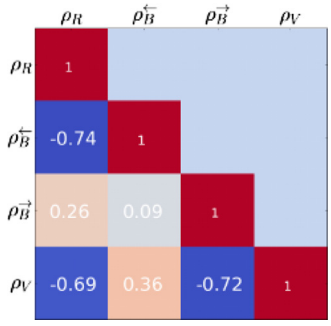


Fig. 23. Pearson correlation among the ρ values.

results more than an indication of the user sentiment or opinion. In Table 7 we already see that $\rho_{\overrightarrow{B}}$ seems anti correlated to ρ_V: a possible explanation of that could be that after the vote users started discussing about the results of the referendum to criticize it.

To better understand the correlation between these two topics we computed the Pearson correlation matrix between the ρ values of the major UK cities whose results are depicted in Fig. 23. We considered the following UK cities: Belfast, Birmingham, Brighton, Bristol, Cambridge, Cardiff, Edinburgh, Glasgow, Leeds, Leicester, Liverpool, London, Manchester, Nottingham, Oxford and Sheffield. From the matrix we see that ρ_R is highly anti-correlated with both ρ_V and $\rho_{\overleftarrow{B}}$: this suggests that the sentiment against-Refugees could be one of the reason for the Brexit result. The anti-correlation is stronger with $\rho_{\overleftarrow{B}}$ in agreement with the presence of a bias in Twitter users in relation to the voters population. The sentiment before and after the referendum is not correlated at all, confirming that the use of similar hashtags before and after the referendum has a completely different meaning. Surprisingly, the referendum outcome is highly anti-correlated with the sentiment after the vote, confirming our assumption about the use of opponent part vocabulary to criticize the outcome of the referendum.

We conclude that investigating polarization on one topic may help in understanding polarization in related topics. We found that the migrant phenomena was a relevant topic for the Brexit referendum discussion. Indeed, there is a correlation between the user polarization we detected in the two datasets. We believe that the proposed multi-dimensional analysis can thus be a useful tool to break down polarized discussions into relevant topics and to investigate users sentiment around those topics.

6. Conclusions and future work

We proposed an adaptive and scalable multidimensional framework to analyze the spatial, temporal and sentiment aspects of a polarized topic discussed in an online social network. Besides enriching tweets with spatial and temporal information, a contribution of this paper is the sentiment enrichment algorithm, capable of identifying the polarity of users and tweets. We experimented this methodology analyzing Twitter data for two case studies: the perception of the Mediterranean refugee crisis in Europe and the discussion about the EU Referendum in UK. The combination of the sentiment aspects with the temporal and spatial dimension is an added value that allows us to infer interesting insights. For example, our analysis revealed that European users are sensitive to major events and mostly express positive sentiments for the refugees. However, in some cases this attitude suddenly changes when countries are exposed more closely to the migration flow. As for the Brexit referendum, we observed how the discussion evolved in the pre and post vote, we identified who are the UK leaders more discussed and the sentiment inside and outside UK. Finally we converged the two analyses in investigating the correlations between the refugee crisis sentiment of UK citizens with the Brexit sentiment and the final vote outcome. As future work we intend to adapt the framework to a real-time streaming scenario and to add more dimensions such as the type of user and the network relationships in the Twitter user graph.

Acknowledgments

This work was supported by the EU H2020 Program INFRAIA-1-2014-2015 *SoBigData: Social Mining & Big Data Ecosystem* (654024).

References

[1] M. Coletto, A. Esuli, C. Lucchese, C.I. Muntean, F.M. Nardini, R. Perego, C. Renso, Sentiment-enhanced multidimensional analysis of online social networks: Perception of the mediterranean refugees crisis, in: Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016, San Francisco, CA, USA, August 18–21, 2016, 2016, pp. 1270–1277.
[2] B. Pang, L. Lee, Opinion mining and sentiment analysis, Found. Trends Inform. Retrieval 2 (1–2) (2008) 1–135.
[3] L.A. Adamic, N. Glance, The political blogosphere and the 2004 us election: divided they blog, in: Proceedings of the 3rd international workshop on Link discovery, ACM, 2005, pp. 36–43.
[4] M. Conover, J. Ratkiewicz, M.R. Francisco, B. Gonçalves, F. Menczer, A. Flammini, Political polarization on twitter., ICWSM 133 (2011) 89–96.
[5] A. Pak, P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining., in: Proceedings of the LREc, 10, 2010, pp. 1320–1326.
[6] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, Comput. Linguist. 37 (2) (2011) 267–307.
[7] O. Kolchyna, T.T. Souza, P. Treleaven, T. Aste, Twitter sentiment analysis: Lexicon method, machine learning method and their combination, in: Handbook of Sentiment Analysis in Finance(2015).
[8] M. Coletto, C. Lucchese, S. Orlando, R. Perego, Electoral predictions with twitter: a machine-learning approach, in: Proceedings of the IIR 2015, Cagliari, Italy, 2015.
[9] K. Garimella, G. De Francisci Morales, A. Gionis, M. Mathioudakis, Quantifying controversy in social media, in: Proceedings of the ACM International Conference on Web Search and Data Mining, in: WSDM '16, 2016.
[10] M. Coletto, C. Lucchese, S. Orlando, R. Perego, Polarized user and topic tracking in twitter, in: Proceedings of the SIGIR 2016, Pisa, Italy, 2016.
[11] M.D. Conover, C. Davis, E. Ferrara, K. McKelvey, F. Menczer, A. Flammini, The geospatial characteristics of a social movement communication network, PloS One 8 (3) (2013).

[12] S. Scellato, C. Mascolo, M. Musolesi, V. Latora, Distance matters: Geo-social metrics for online social networks, in: Proceedings of the Conference on Online Social Networks, in: WOSN'10, 2010.
[13] Y. Takhteyev, A. Gruzd, B. Wellman, Geography of twitter networks, Social Netw. 34 (1) (2012) 73–81. http://dx.doi.org/10.1016/j.socnet.2011.05.006.
[14] J. Kulshrestha, F. Kooti, A. Nikravesh, K.P. Gummadi, Geographic dissection of the Twitter network, in: Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, 2012.
[15] B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, C. Ratti, Geolocated Twitter as proxy for global mobility patterns, Cartography Geogr. Inform. Sci. 41 (3) (2014) 260–271, doi:10.1080/15230406.2014.890072.
[16] E. Zagheni, V.R.K. Garimella, I. Weber, B. State, Inferring international and internal migration patterns from Twitter data, in: Proceedings of the WWW Conference, in: WWW'14 Companion, 2014.
[17] J.-P. Onnela, S. Arbesman, M.C. González, A.-L. Barabási, N.A. Christakis, Geographic constraints on social network groups, PLoS one 6 (4) (2011) e16939.
[18] J. Weng, B.-S. Lee, Event detection in twitter., ICWSM 11 (2011) 401–408.

Mauro Coletto is a research fellow at DAIS, Ca' Foscari University, and he received his Ph.D. degree from IMT - Institute for Advanced Studies (Lucca) in the track "Computer, Decision, and Systems Science". He graduated in Information Management Engineering at the University of Udine in 2012. During his doctoral studies at IMT, he has worked in collaboration with CNR (ISTI-HPC) on the following research topics: Web Mining, Online Social Networks, and Social Media Analysis. His Ph.D. thesis is concerned with the development of a framework for automatically extracting knowledge about topics, opinions of users, and dynamics of polarised communities in Online Social Networks.

Andrea Esuli has master in Computer Science and a Ph.D. in Information Engineering. Since 2011 is a researcher at Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", which is part of the National Research Council (CNR). His research is in Human Language Technologies area, with a focus on Text Classification, Information Extraction and Sentiment Analysis. His main topic of interest are statistical machine learning and information retrieval. In 2010 he won the European "Cor Baayen" award as a "promising young researcher in computer science and applied mathematics".

Claudio Lucchese (http://hpc.isti.cnr.it/~claudio) is a researcher with the Italian National Research Council (CNR). He received his MS.c. and Ph.D. from the Univ. Ca' Foscari di Venezia in 2003 and 2008, respectively. His main research activities are in the areas of data mining techniques for information retrieval, large-scale data processing and cloud computing. He has published more than 60 papers on these topics in peer reviewed international journals and conferences.

Cristina Ioana Muntean (http://pomino.isti.cnr.it/~muntean) is a research fellow at ISTI CNR, Pisa, Italy. She received her Ph.D. from the Babes-Bolyai University, Cluj-Napoca, Romania in 2012, and ever since has been working in Pisa in the High Performance Computing lab. Her main research interests are Data Mining and Machine Learning applied to text and mobility, publishing in international conferences and journals. She also gained experience in working on regional and European projects.

Franco Maria Nardini (http://hpc.isti.cnr.it/~nardini) is a researcher at Italian National Research Council. He received the Ph.D. in Information Engineering from the University of Pisa in 2011. His research interests focus on Web Information Retrieval (IR), Data Mining (DM), and Machine Learning. He served as program committee member of several top-level conferences of IR and DM. He authored more than 40 papers in peer reviewed international journal, conferences and other venues.

Raffaele Perego (http://hpc.isti.cnr.it/~raffaele) is a senior researcher at ISTI-CNR, where he leads the HPC Lab. His main research interests include large-scale information systems, information retrieval, data mining, and machine learning. He co-authored more than 140 papers on these topics published in journals and in proceedings of international conferences. In 2016 he co-chaired the annual ACM SIGIR Conference. He serves as program committee member in the top-tier conferences of his research field.

Dr. Chiara Renso holds a PhD and M.Sc. degree in Computer Science from University of Pisa (1992, 1997). She is permanent researcher at ISTI-CNR, Italy. Her research interests are related to spatio-temporal data mining, reasoning, data mining query languages, semantic data mining, trajectory data mining. She has been involved in several EU projects about mobility data mining. She has been the coordinator of an FP7 Marie-Curie project on semantic trajectories knowledge discovery called SEEK (www.seek-project.eu) with 8 partners. She was also coordinator of a bilateral CNR-CNPQ Italy-Brazil project on mobility data mining. She is author of more than 100 peer-reviewed publications. H-Index on Google Scholar is 21. She is co-editor of the book "Mobility Data: Modelling, Management, and Understanding" edited by Cambridge Press in 2013. She has been co-chair of three editions of the Workshop on Semantic Aspects of Data Mining in conjunction with IEEE ICDM conference. She has been local co-chair for the international conference SIGIR 2016. She is regular reviewer of top level journals and conferences on data mining, knowledge management, spatio/temporal data, semantics.