Using Face Recognition with Twitter Data for the Study of International Migration[1]

Alexandru FLOREA[1,2], Monica ROMAN[1,3,4]
[1]The Bucharest University of Economic Studies
[2]Electronic Arts, Bucharest, [3] Institute of Labor Economics, IZA Bonn, [4]CELSI, Bratislava
alflorea@ea.com, monica.roman@csie.ase.ro

The level of widely used technology has tremendously increased in recent years, and the internet has genuinely re-shaped the way we learn, communicate and live. As a result, social data availability, complexity, and diversity steadily grow. In the same time, the computational machine power constantly unlocks opportunities to provide innovative techniques. By leveraging that power in statistics, powerful algorithms, such as neural networks, started to be applied in various fields. Image Processing has been a subject of broad interest and face recognition has been an essential part of this field in recent years. This paper aims to leverage all these resources to provide an overview of how social media data can be collected and analyzed using R. The result of this paper is represented by an innovative algorithm able to retrieve and analyze Twitter information. Moreover, this paper also provides a snapshot of the Romanian Twitter users' demographics and mobility.

Keywords: *Internet technology, Image Processing, Face recognition, Twitter, Social media Data Collection, International migration*

Introduction

Over the last 40 years, the Information and telecommunications revolution has reshaped individuals' lifestyle so dramatically that life without mobile devices, telecommunication, internet or social media platforms has become hard to imagine.

In the same time, international mobility has become an essential driver in the process of defining economies and societies. In Romania 's case, high skilled migrants are the ones with high mobility, and they also seem to be more connected and active in the social media networks [1]. Due to the high increase of migration, it has been a real challenge to track mobile population and migrants in recent years accurately. In the age of globalization, the traditional census re-searches cannot merely cover the actual needs of having accurate, up-to-date data. As a result, statisticians have started to explore alternative data sources in order to improve the official international mobility figures.

In this paper, we aim to prove the usefulness of studying mobility characteristics with data retrieved from social media, with a focus on using the visual identification of individuals in this context. Social media became one of the most comprehensive data sources available used as an alternative to official data. Moreover, it steadily improves its accuracy, completeness, and reliability.

Machine face recognition from images is the result of various intersected researches in multiple fields, such as computer vision, image processing, pattern recognition, or neural networks. Face recognition technology (FRT) has excellent applicability in real life and might help researchers draw more accurate views of the international population mobility. Using FRT on Twitter data, this paper proves that social media data can provide critical demographic characteristics and increase the accuracy of the official data by leveraging the power of big data. Also, the developed application is exemplified for the Romanian case, being novel for this specific population and having a high potential for being adapted for other geographical areas.

[1] Initial results of this paper were presented at the EGE Conference (Constanta, 2018) and at the International Conference on Economic Informatics (Iasi, 2018). The authors thank participants for their valuable suggestions.

DOI: 10.12948/issn14531305/22.4.2018.03

In the first part of the paper, it will be provided an overview of international migration, and social media data, while the second part will review a few of the available technologies used in migration research. In the third part, we aim to leverage the power of face recognition to estimate the demographic details for a sample of Twitter users active within the territory of Romania, by developing an R algorithm for this purpose. The last sections will provide a snapshot of the most recent mobility figure, based on the sample of Twitter users.

1. International Mobility Researches and Social Media data

In this section of the paper, we aim to briefly describe the Romanian mobility flows from two angles. In the first part of the section, we will analyze the dimension and the "push & pull" factors that contributed to the Romanian migration after the fall of the socialist state. The second part of the section will provide valuable information about how social media could capture the migration flows.

1.1 Romanian international migration: magnitude and recent trends

Romania is one of the most important European players regarding migration figures. However, Romanian migration went through various regimes: controlled migration until 1990, limited migration in the 1990s, irregular migration between 2002 and 2007, and unrestricted migration within the European Union after 2007. Although after 1989 Romanian state no longer prohibited the mobility of Romanians, some European states restricted the East-West migration by implementing migration policies [2].

Romanian migration (re)started in the early 90's with the group of asylum seekers, ethnic Germans, and the type of migrants seeking new opportunities in Western Europe. In some Romanian cities (e.g. Sighişoara) almost 50% of the German population married Romanians [3] [4]. Those networks facilitated the migration of Romanians to Germany [5].

Table 1: Number of Romanian applications for political asylum in Germany, 1990-1999

1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
11,191	27,089	57,464	146,738	21,424	10,274	3,168	1,672	917	537

Source: [2]

Nearly half of asylum applicants from Romania in Germany were Roma who expressed their concern about the growing discrimination in Romania [6]. However, at the end of 1993, Romania was recognized as a secure country and the applications that came after this year were denied. Moreover, Romania and Germany agreed on the repatriation of 60,000 Roma and 40,000 Romanians [7].

Another "pull" factor was the fact that after 2002, Romanians no longer needed a Visa to travel to European Countries. Moreover, after Romania joined EU, in 2007, Romanians were no longer irregular emigrants, and the number of Romanian emigrants increased dramatically [2].

Between 1990 to 1993, there was a time with strong ethnic, and asylum seekers migration [8]. During the same period, the migration of Romanians was not that intense [9]. Soon after 1993 these migration flows to Germany almost stopped, and new migration destinations emerged.

While during the socialist regime, the main push factors were politically related, after the fall of state socialism, economic factors prevailed [10].

After the fall of the communist regime, new forms of international mobility rose, one example was the petty trade to Turkey, former Yugoslavia, Poland, and Hungary. Petty trades practices extended afterward and many of them became pioneers of migration [11].

Religion remained one of the most influential factors in the Romanian migration during the 1990s. Examining the movement from Orthodox and Catholic communities in Romania, studies have shown significant differences between the two religious groups [12]. The inclusion of the Catholic emigrants seemed to

DOI: 10.12948/issn14531305/22.4.2018.03

be higher than the one of the Orthodox groups [2].

Romanian migration reached the first phase of maturity after 1997. The decisive "push" factor was the de-industrialization of the country which led to a severe impoverishment of the population [13]. This process affected an important number of Romanian families. Similar to other post-socialist societies, among the first Romanians who lost their job after the fall of the communist regime, were commuters [14] [15].

This process put a heavy pressure on rural households which transformed them from an active labor force to a class of potential emigrants [16].

This massive shift can partially explain why the migration developed so fast after 1997. Impoverishment and unemployment faced a dramatic increase, therefore in only ten years (1990-1999) active labor force decreased from 9.5 million to 4.5 million people [17].

As expected, the migration phenomenon can have a significant negative impact on sending countries, especially for the states with an emerging economy like Romania.

In terms of the "brain drain" of physicians, Romania is one of the most prominent European contributors. Studies show that Romania invested plenty of resources in developing a robust medical labor force, but unfortunately, a worrying part of it left the country or expressed its intention to migrate and small intention to return [18] [1].

While in the '90s, Romanian medical doctors preferred to emigrate to the US or Canada, after Romania's admission to the European Union, this class has mainly migrated to UK, Germany, France, and Belgium [19] [2]. Various studies show that, in 2010, Romanians physicians represented the main group of foreign medical experts in France [19] Moreover, in the next two years, Romanian physicians represented a third of foreign doctors registered in France [20]. The importance of this phenomenon is reinforced by studies that present a similar situation in Belgium, a country with an impressive number of Romanian medical specialists [19].

To understand the Romanian emigration magnitude, we can mention that there are more Romanian emigrants (approx. 4 million) than US emigrants (3 million), or French (2.2 million), emigrants.

As the high skilled Romanian migrants are the ones with high mobility and also those more connected and active in the social media networks [1], we aim to prove the usefulness of studying mobility trajectories with data retrieved from social media, with a focus on visual identification of individuals.

The next section presents a summary of the initiatives taken to improve the understanding of migration by using social media data.

1.2 Social media data leveraged in international migration researches

Social media platforms are websites and applications that help users to share content and build social networks/relations with other people who appear to have similar visions. Social media environment develops with a tremendously high pace and reshapes the way people live and make decisions.

There are various studies on the migration phenomenon that use social media data. For instance, some of them took advantage of the geolocated tweets. The geographic coordinates can be measured either by using the GPS location of the mobile device or by using the nearest address computed based on the IP location [21].

The current Twitter terms and conditions represent a challenge for working with social media data. The underlying API from twitter used for Academic purposes returns only 1% of the total Twitter feed. However, thanks to the growing penetration of smart devices, the amount of geolocated tweets constantly grows and becomes a more and more valuable register of human migration with every day.

In 2012, which is a long time ago concerning social media evolution, the absolute amount of geolocated tweets was estimated to 3.5M tweets per day [21]. Nowadays, the official sources estimate that there are over 500 million tweets per day. It can be stated that in just a few years, the activity on twitter registered an enormous increase.

DOI: 10.12948/issn14531305/22.4.2018.03

Twitter is probably one of the best social media platforms in terms of academic researches because of its flexibility, consistency and because of its terms and conditions [22] [23] [24]. Tweets were used for assessing a mood to society, by examining the content of Twitter posts [25] [26] [27] [28]. Geo-located tweets were used in many fields, such as analyzing the urban activity, public health, global distribution of languages, and others. [21].

A solid foundation was provided by Cheng's study (add source) in 2011, which tackled different mobility aspects based on Twitter check-ins, dominated in that period by Foursquare (another location sharing service). Other exploratory studies were made by Cho [29] who built a model for the influence of human mobility in respect with social ties, or by Noulas [30], who analyzed the intra-urban mobility using Foursquare check-ins.

In specific papers, it is emphasized the importance of cleaning the dataset obtained from social media, before any analysis. For instance, one approach is to analyze all consecutive locations of a Twitter user, and if he/she appears to travel from one place to another with >2000 km/h, it can be safely stated that the user is most probably a "bot" and it must be filtered out [21].

Preferably, non-personal Twitter activity (e.g. #tweetmyjob) must be filtered out as well because it does not reflect any kind of human (individual) physical activity or geographical mobility.

Some migration analyses on Twitter recommend excluding regions with a penetration rate smaller than 0.5% or countries with less than 10,000 resident users [21].

It has been proved that "despite the unequal distribution over the different parts of the world and possible bias toward a certain part of the population, in many cases, geo-located Twitter can and should be considered as a valuable proxy for human mobility, especially at the level of country-to-country flows [21]. In the study "*Geo-located Twitter as a proxy for global mobility pattern"* [21], the suggested methodology tries to capture human mobility by assigning the nationality of travelers based on the country of residence. This approach allowed the researchers to quickly and yet effectively compare the flows of travelers between countries. The study revealed that people from more developed countries as the West-European ones, are more likely to travel and more than this, the diversity of West-European travelers' destinations is usually higher compared to one of the East-European travelers. Moreover, in the same study, the scientists were able to identify specific patterns driven by cultural or special events occurring in a specific region. Even if in most of the cases Twitter analysis the reasonable expectations, it is essential to understand that Twitter data can be treated as a global register of human mobility and geolocated tweets can be used as a proxy of global mobility behavior.

Other researches have studied migration patterns using only the text, the date and the geographic coordinates of the tweets. Some demographic information could be estimated using different tools, for instance, the gender and age could be estimated using special tools like Face++ which computes those characteristics using computer vision and data mining processes to analyze the Twitter profile picture of a given user [31]. Using this technique, the "Inferring International and Internal Migration Patterns from Twitter Data" study managed to provide a comprehensive picture of the recent migration trends in OECD countries. Moreover, the study claims that there were significant reductions of out-migration flows from countries like Mexico. On the other hand, it also illustrates how in the countries affected by the economic crisis like Greece and Ireland, those flows dramatically increased just after 2008. However, even if this study presented the migration trends only for the short term, "this type of information becomes available well before official statistics and thus can be seen as a barometer of mobility patterns that is useful to nowcast recent trends" [31].

More niched studies captured the trends in the migration of the class represented by highly skilled people using LinkedIn data. LinkedIn is a social networking platform for professionals, and it counts over 450 million users from

DOI: 10.12948/issn14531305/22.4.2018.03

all over the world. People generally use LinkedIn to have more visibility on the latest job opportunities, by making public their resumes. As a result, LinkedIn data is probably one of the most "powerful" and up to date dataset available of the highly skilled migrants [32] [33].

Although the available studies stressed the fact that LinkedIn members are not a representative sample for the population of highly skilled migrants, the sample of LinkedIn users is convenient because of its complexity and size. Moreover. Those studies showed that LinkedIn information could provide a significant amount of insights regarding the recent migration flows of high-skilled people [32].

To conclude, social media is probably one of the best social data sources available because of the large number of potential statistical units that can be provided. Moreover, it constantly improves in terms of accuracy, completeness, and reliability. If it keeps the same pace in terms of development, it is expected that social media data will represent a real opportunity for future migration analyses.

2. New informatic technologies for migration research

In the last 40 years, the power of computers has helped us to explore the universe, find medical treatments, and even redefine the way life is created. Powerful machine learning algorithms like the neural network help researchers to solve complex problems and find creative solutions for the current challenges. For the last years, image processing has been a topic of significant interest. Thanks to these algorithms that leverage more and more powerful machines, we were able to build an algorithm which leverages multiple resources to draw a picture of how the Twitter users might look like in a specific city, country, region.

2.1. R: a useful tool for treating social media data

R is a programming language and an excellent environment for statistical computing or graphics generating. "It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S".[2]

R can provide a wide variety of graphical and statistical techniques. Moreover, R is highly extensible and combines both, statistical and programming skills in one single robust tool.

Being an Open Source tool facilitated its spectacular popularity growth and made R one of the most used statistical tool in the world. R has a lot of great predefined routines and packages, but it can also be connected with most of the current great tools. As a result, its effectiveness attracts more and more researchers every day. R can run on platforms like Windows, Linux, and MacOS.

Among the great resources designed for R, one of the "must have" resource that needs to be installed and used from the first day of using R is RStudio.

"RStudio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management." [3]

RStudio has an open source version as well as a commercial edition and is available for Windows, Mac, and Linux. It can run locally, on a desktop or in a browser connected to RStudio Server.

2.2. Social Media: a potential primary data provider

Social media platforms are now so well established, that the top 5 most popular social network platform did not change recently. However, the usage of those internet services significantly varies in different regions. When comparing social media platforms, it is imperative to take into consideration not only the

[2] R Documentation: https://www.r-project.org/other-docs.html

[3] RStudio Documentation: https://www.rstudio.com/products/rstudio/

DOI: 10.12948/issn14531305/22.4.2018.03

number of users but the number of active accounts as well [34]

It is already common sense to say that social media can positively or negatively impact our lives in equal measure. While each user/person is essential, from a scientific perspective, it is more important to study the networks of users and understand how social media drives changes in the behavior of these networks. Researches proved that social media contributes to migration by reinforcing the bonds between friends/relatives [35].

The daily billions social media interactions provide an enormous amount of information about the global population. Therefore the analytical potential of the social media data is undoubtful high [36].

Twitter is an online social networking and news service available to anyone on which users post messages known as "tweets".

There is no doubt that as a communication platform, Twitter continues to be a game changer by succeeding to provide a clean and effective environment to its millions of monthly active users.

Twitter did not only reshape the way people communicate but also unlocked massive opportunities and gave researches new ways to look at the social behavior of the global population. Even if Twitter can provide a vast amount of information, it is still not representative of the population, but it can be used as a proxy for global migration patterns [21] [31].

3. Face recognition in social media data using R

Machine detection and recognition of faces from images is the result of the emergence of multiple fields, such as computer vision, image processing, pattern recognition, and neural networks. Face recognition technology (FRT) has excellent applicability in real life. For example, it can be used in the process of matching the photo from a passport or an ID card with the real person that is using it. Another case could be represented by the increasing need of having up to date social researches like migration researches, in which FRT might be combined with other tools to estimate the structure of a defined population.

Even if human beings do not generally have difficulties to recognize faces in cluttered scenes, machine face recognition is a significantly higher challenge. Various researches touched this topic, but in 1995, Rama Chellappa, Charles Wilson, and Saad Sirohey concluded that "the most important step in face recognition is the ability to evaluate existing methods and provide new directions on the basis of these evaluations." [37].

In terms of the rapid development of science, the year 1995 seems like a very long time ago and since then, the power of computers enormously increased, and the creativity and knowledge served the needs of having robust face recognition algorithms available for the extensive usage. In 2015, "on the widely used Labeled Faces in the Wild (LFW) dataset, a system achieved a new record accuracy of 99.63%" by using deep convolutional networks [38].

"Kairos is an artificial intelligence company, founded in 2012 and specialized in face recognition. Through computer vision and machine learning, Kairos can recognize faces in videos, photos, and the real-world - making it easier than ever to transform the way your business interacts with people."[4].

Kairos handle the complexity of image processing by using neural networks. Their servers are leveraged to provide accurate face analyses that cannot be run on ordinary machines.

Considering the expertise and the processing power possessed by Kairos, algorithms provided by this company are going to be leveraged to determine the Twitter user's demographics.

[4] Kairos Documentation: https://www.kairos.com/about

DOI: 10.12948/issn14531305/22.4.2018.03

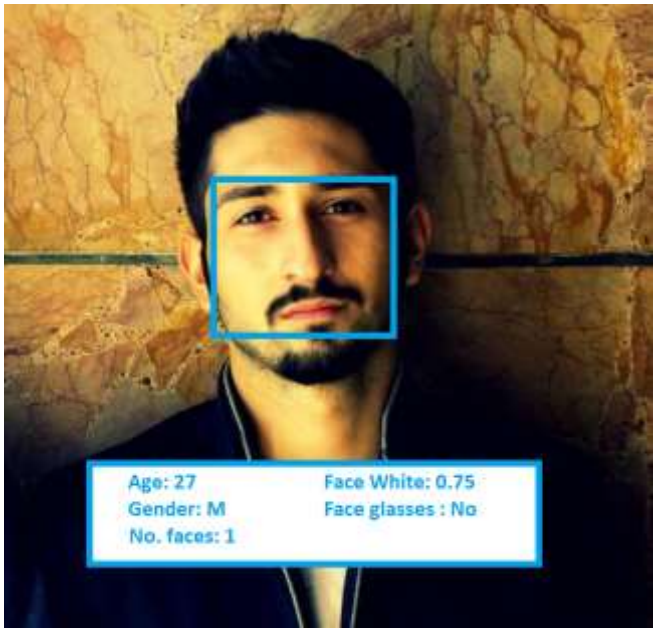

Fig. 1. Face recognition using Kairos algorithm
Source: Generated by Authors

In figure 1, there were estimated demographic characteristics like age, gender, ethnicity, and skin color by leveraging Kairos algorithm. The subject in the figure is a 26 years old, white-skinned Romanian man. The subject gave his consent to include his profile picture in this paper, therefore no rights were violated by this action.

Using powerful face recognition algorithms, like the one provided by Kairos, might significantly reduce the time of building comprehensive social researches.

In the following, we build an innovative algorithm by combining the power of face recognition with the power of social media and using R programming language, for identifying demographic characteristics of the individuals.

This case study on Romanian Twitter users described below uses R programming language to access the power of social media platforms (Twitter) and face recognition algorithms (Kairos) by using their API. In the computer programming field, API stands for "application programming interface" and is a set of communication protocols and subroutine definitions. In simple words, an API is a communication bridge among various components of a system.

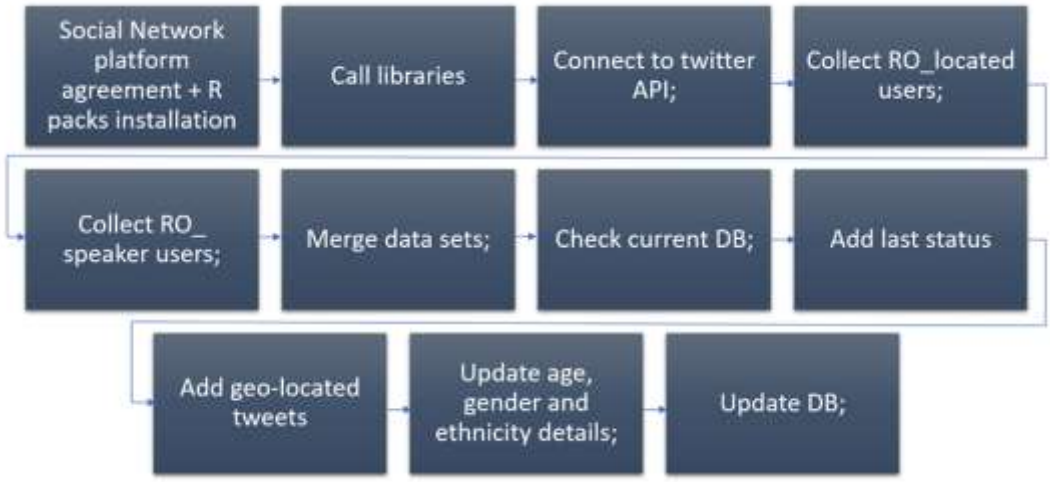

Fig. 2. Algorithm Architecture used in the paper

DOI: 10.12948/issn14531305/22.4.2018.03

Source: Generated by Authors

As presented in figure 2, the proposed algorithm is structured in 11 steps. The first step requires the gain of an agreement with the social media platform which is going to provide social, behavioral information and the installation of proper R libraries. Settings preparations represent the second and third steps while the fourth and fifth steps represent the raw collection of the information. The next four steps facilitate the increase of usage effectiveness while in the 10th step face recognition algorithm is leveraged to estimate demographic details of the social media users. The last step of the algorithm written in R is just about updating the existing database.

In order to be able to "communicate" with Twitter and Kairos in an efficient manner, different collections of precompiled routines (known as libraries) are needed; therefore, the following packages were installed:

```
################################    Call libraries    ################################
library(twitteR)
library(facerec)
library(data.table)
################################    Call libraries    ################################
```

The above script is calling library "twitteR", which will facilitate the communication with Twitter, library "facerec"- which will facilitate the communication with Kairos resources, and library "data.table"- which will provide useful predefined functions.

After setting up connections, creating stable tokens and collecting the coordinates of the biggest 150 cities in Romania, we collected the activity of users who used Twitter within the territory of Romania in 2018.

```
#########################    Collect Ro Activity    ##############################
coordinates<-read.csv('path\\Cities_GeoCoordinates.csv', header=T)
coordinates<-coordinates[!is.na(coordinates$lat),]
ro_located_data = twitteR::searchTwitter(' ', n = 100000,since = '2018-01-01',
                geocode =paste0(coordinates$lat[1],',',coordi-
nates$long[1],',','20km'),
                retryOnRateLimit = 3)
ro_located_data<-twitteR::twListToDF(ro_located_data)


for( i in 2:dim(coordinates)[1])
{
  ro_located = twitteR::searchTwitter(' ', n = 100000,since = '2018-01-01',
               geocode = paste0(coordinates$lat[i],',',coordi-
nates$long[i],',','20km'),
               retryOnRateLimit = 3)
  ro_located = twitteR::twListToDF(ro_located)
  ro_located_data<-merge(ro_located_data,ro_located, all=T)
Sys.sleep(100)
}
#########################    Collect Ro Activity
##############################
```

In the above script, Twitter users' activity was collected by using filters of 100,000 tweets/request in order to not exceed the limits imposed by Twitter terms and conditions. To be sure, the entire Romanian territory was covered, it was used a range of 20 km from the city center.

DOI: 10.12948/issn14531305/22.4.2018.03

name	created	protected	verified	screenName	location
[redacted]	2018-09-16 15:06:16	FALSE	FALSE	klauditzu	Botosani, România
[redacted]	2018-09-17 05:14:57	FALSE	FALSE	alexdimit2	Moldova
[redacted]	2017-12-24 10:47:44	FALSE	FALSE	mXmaraXm	Constanta, Romani
[redacted]	2018-09-20 16:26:04	FALSE	FALSE	KoczanCsaba	Satu Mare, Români
[redacted]1><U+044E><U+0448><U+0430>	2018-09-15 06:07:53	FALSE	FALSE	KCeeYPEfFCVgL0H	Moldova
[redacted]	2018-09-05 14:20:09	FALSE	FALSE	AlexTud04011744	Buzau, România

Fig. 3. Twitter Activity sample of data collected
Source: Generated by Authors

In figure 3 is displayed a sample of the data set obtained after running the above script. It still requires the usual process of data cleaning, but the potential unlocked by this feature might help researchers to get a proxy of the population structure in each region.

```
###################### AGE & GENDER & Ethnicity Estimates ######################
for (i in 1: dim(userDetails)[1])
{
  tryCatch(facerec<-as.data.frame(detect(userDetails$profileImageUrl[i])))
    if(names(facerec)[3] %like% 'error')
    {
      print(paste('No faces found at line:', i))
      userDetails$age[i]<- "No faces found"
      userDetails$gender[i]<-"No faces found"
      userDetails$face_hispanic[i]<-"No faces found"
      userDetails$face_black[i]<-"No faces found"
      userDetails$face_white[i]<-"No faces found"
      userDetails$face_asian[i]<-"No faces found"
    }else
    { #facerec
      userDetails$age[i]<-facerec$face_age
      userDetails$gender[i]<-facerec$face_gender_type
      userDetails$face_hispanic[i]<-facerec$face_hispanic
      userDetails$face_black[i]<-facerec$face_black
      userDetails$face_white[i]<-facerec$face_white
      userDetails$face_asian[i]<-facerec$face_asian
    }
  Sys.sleep(sample(1:30,1)
}
###################### AGE & GENDER & Ethnicity Estimates  ######################
```

The script above is calling Kairos algorithm to detect and analyze the profile images of the Twitter users collected by the previous section of the script. If the user did not provide a profile picture, a message indicating that no faces were found for user number "i" will be displayed in the log file. Moreover, all the demographic variables will take the value "No faces found" for this category of users.

DOI: 10.12948/issn14531305/22.4.2018.03

name	age	gender	face_hispanic	face_black	face_white	face_asian
[redacted]	29	F	0.00045	0.00035	0.28818	0.70777
[redacted]	34	M	0.05437	0.00104	0.93274	0.00984
[redacted]	27	F	0.02918	0.00789	0.16507	0.7806
[redacted]	33	F	1e-04	0.00018	0.99965	6e-05
[redacted]	25	F	0.00364	0.00035	0.99579	0.00011
[redacted]	25	F	0.00417	1e-04	0.97726	0.01266
[redacted]	22	F	0.01143	0.00021	0.98666	0.0015
[redacted]	35	M	2e-05	1e-05	0.99995	1e-05
[redacted]	30	F	0.00044	0.00011	0.99941	3e-05
[redacted]	37	M	0.00073	3e-05	1e-04	0.99902
[redacted]+021B>a	36	M	0.21232	0.00311	0.77841	0.00413
[redacted]	21	M	0.00207	1e-04	0.17771	0.81394

Fig. 4. Demographics obtained by using advanced face recognition algorithms
Source: Generated by Authors

Figure 4 represents a sample of the demographics obtained by using the advanced face recognition algorithm provided by Kairos. As it can be noticed, most of the users are young, with ages between 20 and 40 years old and white-skinned. As expected, there is a balanced split between genders. The names in the table were anonymized.

4. Using Face Recognition Technology to identify demographic characteristics

After running all the sections in the algorithm, we obtained valuable information about a large number of about 60,000 Twitter users from Romania.

However, only some of them have a relevant profile picture which might provide useful information related to their demographic characteristics: gender, but also age and race.

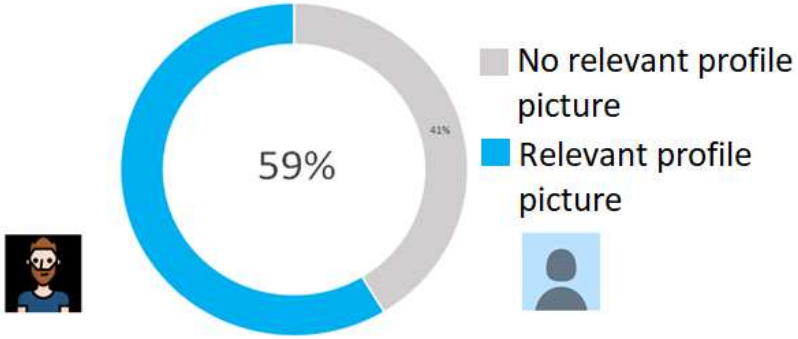

Fig. 5. Percentage of users with a relevant profile picture
Source: Generated by Authors

As seen in figure number 5, although the algorithm provided by Kairos Company was accurate and managed to handle cluttered images, only 59% of the users had at that moment a profile picture that could be analyzed the face recognition algorithm. Most of the other 41% percent of the users used the default Twitter profile picture, at that moment (2018-09-20).

DOI: 10.12948/issn14531305/22.4.2018.03

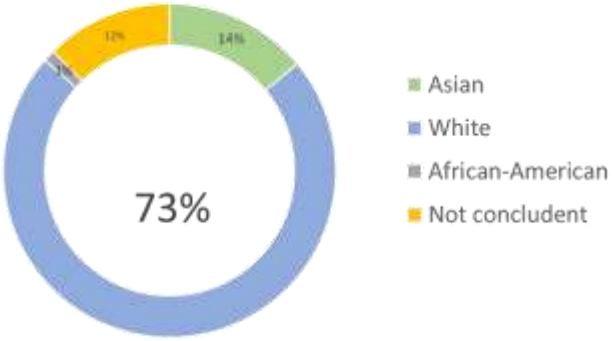

Fig. 6. Race split of Twitter users in the sample
Source: Generated by Authors

On the other hand, the race of the users with a relevant profile picture could be estimated accurately in almost 90% of the cases. It is easy to notice that seen in figure 6, 73% of collected users were in "White" ethnic category, 14% in "Asian" category and as expected, only 1% in the "African-American" group. Moreover, age and gender were estimated accurately in 99% of the cases.

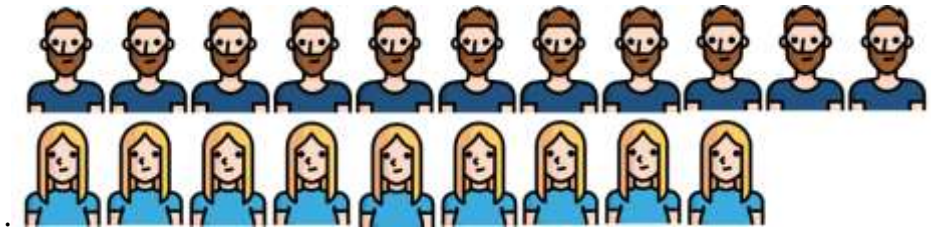

Fig. 7. Gender split of collected Twitter users
Source: Generated by Authors

In figure 7, can be noticed that the obtained dataset is balanced, regarding gender split. About 55% of the collected users were estimated to be men while 45% of users were women.

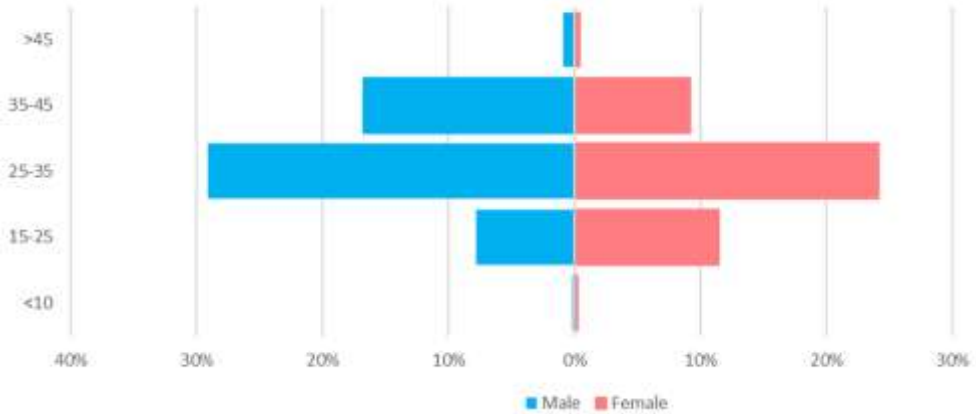

Fig. 8. Age pyramid for the individuals in the sample

DOI: 10.12948/issn14531305/22.4.2018.03

Source: Generated by Authors

Figure 8 provides more information on the demographic structure of the collected dataset. Although the overall gender split is relatively balanced, it is slightly skewed by the age, meaning that the percentage of men increases in the older groups. Even that, the overall percentage of men is slightly bigger than the percentage of women (55% vs. 45%), the number of women is bigger in the "under 25 years old" category. In the same time, most of the collected users have ages between 15 and 45 years old.

As reported from the current section, the potential of using Twitter data for migration research, but also for other purposes, is worth to be considered.

However, as our research is at one initial phase, there are some limitations. The number of characteristics we can identify is not very large at this stage of our research, and the quality and reliability of the information are limited to the one provided by the Twitter user. However, these limitations could be further addressed by considering a large number of potential cases that might be accessed through social media.

5. Romanian migration snapshot

Romanian migration increased dramatically in 2007, the same year Romania became an official European Union member. Unrestricted migration convinced almost 460.000 Romanians to leave Romania permanently [39]. Since 2007, the Romanian migration has reached worrying thresholds, therefore it is important to be able to track and monitor the very recent trends of migration. By understanding the recent trends, the current labor market, educational system, and other institutions can reshape their strategy to avoid mass migration of valuable people and knowledge.

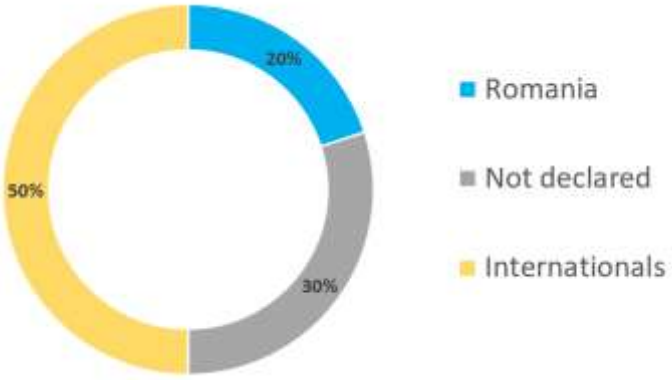

Fig. 9. Self-declared permanent location of the Twitter users in the sample
Source: Generated by Authors

A large share of 70% of the Twitter users in the sample have declared their address and this allows for tracking potential mobility of the individuals in the sample. A share of 20% of them declared that they have their permanent residence in Romania. This split does not necessarily mean that only 20% are Romanians. 2% of the users who do not have a permanent location in Romania use Twitter in Romanian language and are probably Romanians. Moreover, a lot of Romanians did not declare their permanent address at all.

However, 20% of our large data set of 60000 persons correspond to a subsample of 12000, and this might be enough to build a solid proxy for the real population with ages between 15-45 years old.

DOI: 10.12948/issn14531305/22.4.2018.03

Another aspect worth mentioning is that, in only one month, 0.2% of the users with a permanent location in Romania changed their country location. Starting from Twitter users' migration, proxies of the very recent real migration can be developed, and mobility trends can be identified long before any official data collection.

Migration is a vital economy and society driver, especially in emerging markets in which might have a negative impact on the long-term (i.e. brain-drain). Understanding the very recent trends of migration and being able to define the demographic structure of the migration groups, is essential for economies and societies.

Considering the above findings, the understanding of the very recent migration might be significantly improved by using multiple techniques, such as image processing, big data, computer programming, social media, and statistics.

Conclusion

Social media data increases in complexity and accuracy and might facilitate innovative and robust researches soon. Although collecting and analyzing social media data currently represent a big challenge, we believe it worth monitoring the evolution of these alternative data sources and be ready to integrate them in the official statistics. Nevertheless, powerful machine algorithms, such as image processing, neural networks, and pattern recognition will support this kind of analysis and help researchers draw more accurate pictures of international mobility.

Obstacles will be faced in any journey of reinventing or innovating a methodology, but it is essential to overcome them and provide solid foundations for future researches. One major obstacle in our research is that legal terms have recently changed and according to GDPR the storage of any personal information is not allowed anymore. However, it is allowed to work with aggregated figures and provide a proxy for real international mobility.

This paper proved that, even with limited access and restrictive legal terms, social media data could provide relevant information and draw the big picture of how social media users look like in Romania. We aim to go further and analyze social media data as much as the legal boundaries allow us. In the same time, we would like to emphasize that the collaboration between official entities and social media platforms would increase improve the official stats and user experience.

This paper presented a technique of collecting and analyzing data referring to Twitter users, who were active in Romania. Important demographics like age, gender, and race were estimated using advanced face recognition algorithms. The topic needs to be further explored, and there are clear directions that we plan to address in the future. The first direction for future research refers to including geographical coordinates for better identifying the geographical mobility trajectories of social media users. Also, there seems to be an enormous potential for developing new statistical and econometric models for analyzing a large amount of data available on social media, for specific research purposes. On the other hand, official institutions and social media platforms could collaborate to improve the quality of official statistics and ultimately, the quality of life. Therefore, the result of the present research could be relevant to the research community, but also for data scientist and public authorities.

Bibliography

[1] M. Roman and Z. Goschin, "Return migration in an economic crisis context. A survey on Romanian healthcare professionals," Romanian Journal of Economics, vol. 39, issue 2, pp 100-120, 2014.

[2] R. Anghel, A. Botezat, A. Cosciug, I. Manafi and M. Roman, International Migration, Return Migration and their effects. A comprehensive review on the Romanian Case, CELSI Discussion paper no.43, 2017.

[3] K. Verdery, "The unmaking of an ethnic collectivity: Transylvania`s Germans.," American Ethnologist, 1985.

DOI: 10.12948/issn14531305/22.4.2018.03

[4] R. I. Poledna, "Transformări sociale la saşii Ardeleni după 1945. Phd Thesis, Babeş-Bolyai University, Cluj Napoca.," 1998.

[5] B. Michalon, " Circuler entre Roumanie et Allemagne. Les Saxons de Transylvanie, de l`emigration ethnique au va-et-vient.," Balkanologie, 2003.

[6] A. Reyniers, "Migrations tsiganes de Roumanie.," In Visible mais peu nombreux. Les circulations migratoires roumaines, 2003.

[7] H. Kurthen, " Germany at the crossroads: National identity and the challenges of immigration.," International Migration Review, 1995.

[8] M. Baldwin-Edwards, "Migration policies for a Romanian within the European Union: Navigating between the Scylla and Charybdis.," Mediteranean Migration Observatory Working Paper No.7, 2005.

[9] D. Sandu, C. Radu, M. Constantinescu and O. Ciobanu, "A country report on Romanian migration abroad: Stocks and flows after 1989.," Multicultural center Prague, 2006.

[10] B. Dietz, "Ethnic German immigration from Eastern Europe and the former Soviet Union to Germany: The effects of migrant networks.," IZA discussion paper no. 68, 1999.

[11] D. Diminescu, R. Ohlinger and V. Rey, "Les circulations migratoires roumaines: une intégration européenne par le bas?," Cahiers de recherche de la MiRe, 2003.

[12] R. Monica and G. Zizi, "Does religion matter? Exploring economic performance differences among Romanian emigrants.," *Journal for the Study of Religions and Ideologies*, pp. 183-212, 2011.

[13] I. Horváth, "Focus migration," Country Profile, 2007.

[14] C. Hann, "The skeleton at the feast: Contributions to East European anthropology.," University of Kent. , Canterbury, 1995.

[15] C. Hann, " Postsocialism: Ideals, ideologies and practices in Eurasia.," Routledge, London, 2002.

[16] P. Cingolani, "Prin forţe proprii. Vieţi transnaţionale ale migranţilor români în Italia.," Sociologia migraţiei. Teorii şi studii de caz româneşti, 2008.

[17] I. Horváth and R. G. Anghel, ""“Migration and its consequences for Romania.," Südosteuropa. Zeitschrift für Politik und Gesellschaft, 2009.

[18] M. Wismar, C. B. Maier, I. A. Glinos, G. Dussault and J. Figueras, "Health professional mobility and health systems. Evidence from 17 European countries.," World Health Organization on behalf of the European Observatory on Health Systems and Policies,, 2011.

[19] J. Buchan, M. Wismar, I. A. Glinos and J. Bremner, "Health Professional Mobility in a Changing Europe," 2014.

[20] R. Séchet and D. Vasilcu, "Physicians' migration from Romania to France: a brain drain into Europe?," Cybergeo: European Journal of Geography, 2015.

[21] B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos and C. Ratti, "Geo-located Twitter as proxy for global mobility pattern," Global Networks- International Journal of Geographical Information Science Vol. 00, No. 00, 2014.

[22] A. Java, X. Song, T. Finin and B. Tseng, "Why We Twitter: Understanding Microblogging Usage and Communities.," In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007, 2007.

[23] H. Kwak, C. Lee, H. Park and S. Moon, "What Is Twitter, a Social Network or a News Media?," Proceedings of the 19th International Conference on World Wide Web, 2010.

[24] B. A. Huberman, D. M. Romero and F. Wu, "Social Networks That Matter:

DOI: 10.12948/issn14531305/22.4.2018.03

Twitter Under the Microscope," SSRN Scholarly, 2008.

[25] L. Mitchell, K. D. Harris, M. R. Frank, P. S. Dodds and C. M. Danforth, "The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place.," ArXiv e-print arXiv, 2013.

[26] A. Pak and a. P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining.," Proceedings of LREC, 2010.

[27] J. Bollen, A. Pepe and H. Mao, "Modeling Public Mood and Emotion: Twitter Sentiment and Socio-economic Phenomena.," Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 2011.

[28] S. A. Golder and M. W. Macy, "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures.," Science 333, 2011.

[29] E. Cho, S. A. Myers and J. Leskovec, "Friendship and Mobility: User Movement in Location-based Social Networks.," Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2011.

[30] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil and a. C. Mascolo, "A Tale of Many Cities: Universal Patterns in Human Urban Mobility.," PLoS ONE, 2012.

[31] E. Zagheni, V. R. K. Garimella, I. Weber and B. State, "Inferring International and Internal Migration Patterns from Twitter data," Stanford University, 2014.

[32] B. State, M. Rodriguez, D. Helbing and E. Zagheni, "Migration of Professionals to the U.S. - Evidence from LinkedIn data," Stanford University, 2014.

[33] S. O. and T. Elenurm, "Comparing Online Social Networks Ties as Tool for Entrepreneurial Learning Readiness in Small Economies," *Informatica Economica*, pp. 62-74, 2018.

[34] D. Chaffey, "Global Social Media Statistics Summary," 2016.

[35] R. Dekker and G. Engbersen, "How social media transform migrant networks and facilitate migration," Global Networks- A journal of transnational affairs, Volume 14, Issue 4, 2013.

[36] MacEachren, Robinson, Jaiswal, Pezanowski, Savelyev, Blanford and Mitra, "Geo-Twitter Analytics: Applications in Crisis Management," Proceedings, 25th International Cartographic Conference, Paris, France, 2011.

[37] R. Chelappa, C. L. Wilson and S. Sirohey, "Human and Machine Recognition of Faces: A Survey," in *IEEE*, 1995.

[38] F. Schroff, D. Kalenichenko and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in *IEEE Xplore*, 2015.

[39] INS, Migratia Internationala a Romaniei, 2014.

Alexandru FLOREA has graduated the Faculty of Cybernetics, Statistics and Economic Informatics in 2015. He holds a Master diploma in Statistics from 2017 and he is currently enrolled in a PhD program in Cybernetics and Statistics. He also holds a senior position in the Analytics department of Electronic Arts. His PhD work is mainly focused on finding innovative approaches to study to migration.

DOI: 10.12948/issn14531305/22.4.2018.03

Monica ROMAN is Professor at the Department of Statistics and Econometrics, the Bucharest University of Economic Studies, since 2007, where she teaches Business Statistics, Econometrics, and Quantitative Research Methods. Since 2011 she is affiliated as research fellow at the Institute for the Study of Labor IZA Bonn and at the Central European Labour Studies Institute in Bratislava. Her research interests are international mobility, demographic economics and regional studies. Prof. Roman has coordinated research grants as principal investigator on migration and labour issues, financed by the Romanian Government or the European Commission. She is the "MOVE" project responsible on behalf of the Bucharest University of Economic Studies. She has published two books on Romanian migration and articles on mobility and migration issues in peer reviewed journals such as Transnational Social Review, Journal of Comparative Economics, Panoeconomicus, Journal of Identity and Migration Studies, Journal for the Study of Religions and Ideologies.

DOI: 10.12948/issn14531305/22.4.2018.03