Digital Footprints of International Migration on Twitter

Jisu Kim[1(✉)], Alina Sîrbu[2(✉)], Fosca Giannotti[3(✉)], and Lorenzo Gabrielli[3(✉)]

[1] Scuola Normale Superiore, Pisa, Italy
jisu.kim@sns.it
[2] University of Pisa, Pisa, Italy
alina.sirbu@unipi.it
[3] Istituto di Scienza e Tecnologie dell'Informazione, National Research Council of Italy, Pisa, Italy
{fosca.giannotti,lorenzo.gabrielli}@isti.cnr.it

Abstract. Studying migration using traditional data has some limitations. To date, there have been several studies proposing innovative methodologies to measure migration stocks and flows from social big data. Nevertheless, a uniform definition of a migrant is difficult to find as it varies from one work to another depending on the purpose of the study and nature of the dataset used. In this work, a generic methodology is developed to identify migrants within the Twitter population. This describes a migrant as a person who has the current residence different from the nationality. The residence is defined as the location where a user spends most of his/her time in a certain year. The nationality is inferred from linguistic and social connections to a migrant's country of origin. This methodology is validated first with an internal gold standard dataset and second with two official statistics, and shows strong performance scores and correlation coefficients. Our method has the advantage that it can identify both immigrants and emigrants, regardless of the origin/destination countries. The new methodology can be used to study various aspects of migration, including opinions, integration, attachment, stocks and flows, motivations for migration, etc. Here, we exemplify how trending topics across and throughout different migrant communities can be observed.

Keywords: International migration · Emigration · Big data · Twitter

This work was supported by the European Commission through the Horizon2020 European project "SoBigData Research Infrastructure—Big Data and Social Mining Ecosystem" (grant agreement no 654024) and partially by the Horizon2020 European project "HumMingBird – Enhanced migration measures from a multidimensional perspective" (grant agreement no 870661).

M. R. Berthold et al. (Eds.): IDA 2020, LNCS 12080, pp. 274–286, 2020.
https://doi.org/10.1007/978-3-030-44584-3_22

1 Introduction

Understanding where migrants are is an important topic because it touches upon multidimensional aspects of the sending and receiving countries' society. It is not only the demographic fabric of countries but also labour market conditions, as well as economic conditions that may alter due to demographic adjustment. Understanding their allocation is essential for both policy makers and researchers to bring the best of its effects.

Official data such as census, survey and administrative data have been traditionally the main data source to study migration. However, these data have some limitations [12]. They are inconsistent across different nations because countries employ different definitions of a migrant. Moreover, collecting traditional data is costly and time consuming, thus tracking instantaneous stocks of migrants becomes difficult. This becomes even harder when tracking emigrants because of the lack of motivation from citizens to declare their departure.

In recent years, however, we are provided with other alternative data sources for migration. The availability of social big data allows us to study social behaviours both at large scale and at a granular level, and to peek into real-world phenomena. Although known to suffer from other types of issues, such as selection bias, these data could bring complementary value to standard statistics.

Here, we propose a method to identify migrants based on Twitter data, to be used in further analyses. According to the official definition, a migrant[1] is "a person who moves to a country other than that of his or her usual residence for a period of at least a year". In the context of Twitter, we define a migrant as "*a person who has the current residence different from the nationality*".

Following this definition, we performed a two step analysis. First, we estimated the current residence for users by examining location information from tweets. The residence is defined as the country where the user spends most of the time in a year. Second, we estimated nationality, by considering the social network of users. In the international literature, nationality is defined as a relationship between a state and an individual, with rights and duties on both sides [1,6]. Related concepts are ethnicity - in terms of cultural features - and citizenship - in terms of political life. In this paper, we employ the term nationality to define the ensemble of features that make a person feel like they belong to a certain country [2,5]. This could be the country where a person was born, raised and/or lived most of their lives. By comparing labels of residence and nationality of a user, we were able to understand whether the person has moved from their home country to a host country, and thus if they are a migrant. We validated our estimation internally, from the data itself, and externally, with two official datasets (Italian register and Eurostat data).

One of the advantages of our methodology is that it is generic enough to allow for identification of both immigrants and emigrants. We also overcome one of the limitations of traditional data by setting up a uniform definition of

[1] Recommendations on Statistics of International Migration, Revision 1 (p. 113). United Nations, 1998.

a migrant across different countries. Furthermore, our definition of a migrant is very close to the official definition. We establish the fact that a person has spent a significant period at the current location. Also, we eliminate visitors or short-term stays that do not follow the definition of a migrant. This is also validated by the comparison with official datasets. Another advantage of our method is the fact that it uses only very basic features from the Twitter data: location, language and network information. This is useful since the settings of the freely available Twitter API change constantly. Some of the user attributes that the existing literature use to estimate nationality are no longer available. In addition, we make use of unknown locations of tweets by examining whether they intersect with identified locations. By doing so, we do not neglect any information provided by the tweets from unknown locations which later provide useful information on trending topics of Italian emigrants overseas.

One of the issues with our method is that the migrants that we observed are selected from the Twitter population, and not from the general world population, and it is known that some demographic groups are missing. Nevertheless, we believe that studying the Twitter migrant population can provide important insight into migration phenomena, even if some findings may not apply to the other demographic groups that are not represented in the data.

It is important to note that tracking individual migrants is not the objective of our study, but it is only an intermediate stage to enable further analyses. We simply perform user classification to identify migrants among users in our data, and then aggregate the findings. Further studies we envision are aimed at devising new population-level indices useful to evaluate and improve the quality of life of migrants, through targeted evidence-based policy making. No individual personal information nor migration status is released at any stage during the current analysis, nor in any population-level analysis, which is performed following the highest ethical and privacy standards.

The rest of the paper is organised as follows. In the next section we describe related work that studies migration using big data. In Sect. 3, we provide details of the experimental setting for data collection as well as data pre-processing. We then explain our identification strategy for both residence and nationality in Sect. 4. In Sect. 5, we evaluate our estimation using both internal and external data. Section 6 covers a possible application of our method on studying trending topics among Italian emigrants, while Sect. 7 concludes the paper.

2 Related Work

In the past few years, there have been several works on migration studies using social big data. Most of these employed Twitter data but Facebook, Skype, Email as well as Call Detail Record (CDR) data have also been used to study both international and internal migration [3,9,10,14,16]. Here, we focus on studies that have employed freely available data. The definition of a migrant varied from one work to another depending on the purpose of the study and the nature of the dataset. Thus, the definitions provided fit under different types of migration such as refugees, internal migrants, seasonal migrants or even visitors.

One example of using Twitter to observe migration flows is [15]. They defined residence as the country where the tweets were most frequently sent out for periods of four months. If one's residence changed in the following four months period, it was considered that the person has moved. In a more recent work, [11] measure migration flows from Venezuela to neighbouring countries between 2015 and 2019. They look at the bounding boxes and country labels provided by the tweets and identified the most common country of tweets posted monthly. Their definition of a migrant was "any individual leaving Venezuela during the time window of observation" which was observed when an identified Venezuelan resident appeared for the first time in a different country. Our definition of residence is somewhat similar to these works. However, unlike them, we are measuring stocks of migrants, and not flows. Thus, we take into account the aspect of duration of stay. This naturally eliminates short-term trips and visits.

Apart from geo-tagged tweets, there is other information provided by the Twitter API that can help us infer whether a person is a migrant or not. Although [8] did not directly study migrants, but looked at foreigners present in Qatar, it provides important insights to which of the features provided by Twitter is useful in identifying nationality of users. They gathered features from both profile and tweets of users. For features providing information on profile pictures and name, they performed facial recognition and name ethnicity detection. Their final results showed that ethnicity of name, race, language of tweet, language of mention, location of followers and friends are the first six features that are useful. In this paper, we purely employ data provided by Twitter for the analysis and therefore, we do not have name, ethnicity and race features. Nevertheless, our work also shows that locations of users and friends are the useful features. The difference here is that we propose to use the social network of users as one of the main features in identifying nationality, which is more flexible than having to perform ethnicity detection on names and profile pictures.

3 Experimental Setting for Data Collection

We began with a Twitter dataset collected by the SoBigData.eu Laboratory [4]. We started from a three months period of geo-tagged tweets from August to October 2015. Due to our focus on Italy, we selected from these data the users that tweeted from Italy, obtaining thus 34,160 users. We then crawled the network of geo-enabled friends of these 34,160 users, using the Twitter API. Friends are people that the individual users are following. We focused on friends because we believe that for a user, the information on whom they follow is more informative when it comes to nationality, than who they are followed by. We concentrated on geo-enabled friends because geo-location is necessary for our analysis. By collecting friends, the list of users crossed our initial geographic boundary, i.e., Italy. At this stage, the number of unique users grew to over 250,000. For all users we also scraped the profile information and the 200 most recent tweets using the Twitter API. During this process, we were able to collect all 200 recent tweets for 97% of users and at least 55 tweets for 99% of users. Our final user

network consists of 258,455 nodes and 1,205,133 edges which includes both our initial 34,160 users and their geo-tagged friends.

For the process of identifying migration status, we focus on the core users, i.e., 34,160 users. We assign a residence and a nationality to each user, based on the geo-locations included in the data, the language of tweets and profile information. The final dataset includes 237 unique countries from where individuals have sent out their tweets, including 'undefined' location. Even if a user enables geo-tags on their tweets, not all tweets are geo-tagged. As a result, 21% of our tweets are 'undefined'. As for the languages, there are 66 unique languages and 12% of our tweets are in English.

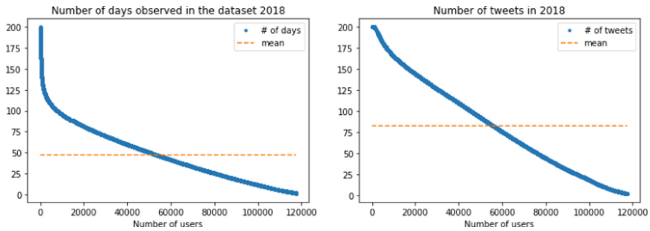

Fig. 1. Distribution of the number of days (left) and the number of tweets (right) observed in the data per user: on average, our users have tweeted 47 days and 82 tweets in 2018.

As for the profile features, we observe that 40% of the users have filled out location description. In addition, most of users have set their profile language to English. The number of unique profile languages detected in our data is 58 which is smaller than the languages used, indicating that some users are using languages different from their profile language when tweeting.

In order to assign a place of residence to users, we needed to restrict the observation time period. We have chosen to look at one year length of tweets from 2018, in order to assign the residence label for the 2018 solar year. We selected users that have tweeted in 2018, identifying 128,305 users. To remove bots, we looked at whether a user is tweeting too many times a day. We considered that tweeting more than 50 tweets on average in a single day was excessive and we have eliminated in this way 39 users. In addition, we removed users that were not very active in 2018. If the number of tweets was less than 20, we checked whether the tweeted days were spread out during the year. If the days were not well spread out, we filtered out the user. On the other hand, if it was well spread out, it meant that the user was regularly tweeting, so the user was kept. During this process, we removed 10,764 users. After removing bots and inactive users, we have 117,502 users. For these, we show the distribution of the number of tweets and number of days in which they tweeted in Fig. 1. On average we see 47 days and 82 tweets.

In addition to the Twitter data, we also collected a list of official and spoken languages for countries identified in our data[2].

4 Identifying Migrants

A migrant is a person that has the residence different from the nationality. We thus consider our core 34,160 Twitter users and assign a residence and nationality based on the information included in our dataset. The difference between the two labels will allow us to detect individuals who have migrated and are currently living in a place different from their home country. The methodology we propose is based on a series of hypotheses: a person that has moved away from their home country stays in contact with their friends back in the home country and may keep using their mother tongue.

4.1 Assigning Residence

In order for a place to be called residence, a person has to spend a considerable amount of time at the location. Our definition of residence is based on the amount of time in which a Twitter user is observed in a country for a given solar year. More precisely, a residence for each user is the country with the longest length of stay which is calculated by taking into account both the number of days in which a user tweets from a country but also the period between consecutive tweets in the same country. In this work we compute residences based on 2018 data.

To compute the residence, we first compute the number of days in which we see tweets for each country for each user. If the top location is not 'undefined', then that is the location chosen as residence. Otherwise, we check whether any tweet sent from 'undefined' country was sent on a same day as tweets sent from the second top country. In case at least one date matched between the two locations, we substitute second country as the user's place of residence. On average, 5 dates matched. This is done under the assumption that a user cannot tweet from two different countries in a day. Although this is not always the case if a user travels, in most of the days of the year this should be true. This approach allowed us to assign a residence in 2018 to 57,180 users.

For the remaining 60,322 users, a slightly different approach was implemented. We computed the length of stay in days by adding together the duration between consecutive tweets in the same country. We selected the country with the largest length of stay. In case the top country was 'undefined', we checked whether 'undefined' locations were in between segments of the second top country, in which case the second country was chosen. In this way, an additional 11,046 users were assigned a place of residence. The remaining 49,276 users were neglected because we considered that we did not have enough information to assign a residence.

[2] Retrieved from http://www.geonames.org and https://www.worlddata.info.

4.2 Assigning Nationality

In order to estimate nationalities for Twitter users, we took into account two types of information included in our Twitter data. The first type relates to the users themselves, and includes the countries from which tweets are sent and the languages in which users tweet. For each user u we define two dictionaries loc^u and $lang^u$ where we include, for each country and language the proportion of user tweets in that country/language.

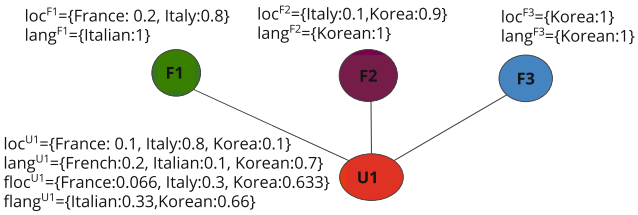

Fig. 2. Example of calculation of the $floc$ and $flang$ values for a user. The calculation of $floc^{U1}$ and $flang^{U1}$ is based of the $floc$ and $flang$ values for the three friends, showing the distribution of tweets in various countries/languages for each.

The second type of information used is related to the user's friends. Again, we look at the languages spoken by friends, and locations from which friends tweet. Specifically, starting from the loc and $lang$ dictionaries of all friends of a user, we define two further dictionaries $floc$ and $flang$. The first stores all countries from where friends tweet, together with the average fraction of tweets in that country, computed over all friends:

$$floc^u[C] = \frac{1}{|F(u)|} \sum_{f \in F(u)} loc^f[C] \tag{1}$$

where $F(u)$ is the set of friends of user u. Similarly, the $flang$ dictionary stores all languages spoken by friends, with the average fraction of tweets in each language l:

$$flang^u[l] = \frac{1}{|F(u)|} \sum_{f \in F(u)} lang^f[l] \tag{2}$$

Figure 2 shows an example of a (fictitious) user with their friends, and the four resulting dictionaries.

The four dictionaries defined above are then used to assign a nationality score to each country C for each user u:

$$N_C^u = w_{loc} loc^u[C] + w_{lang} \sum_{l \in languages(C)} lang^u[l] + \tag{3}$$

$$w_{floc} floc^u[C] + w_{flang} \sum_{l \in languages(C)} flang^u[l] \tag{4}$$

where $languages(C)$ are the set of languages spoken in country C, while w_{loc}, w_{lang} , w_{floc} and w_{flang} are parameters of our model which need to be estimated from the data (one global value estimated for all users). Each of the w value gives a weight to the corresponding user attribute in the calculation of the nationality. To select the nationality for each user we simply select the country C with maximum N_C: $N^u = \text{argmax}_C N_C^u$.

5 Evaluation

To evaluate our strategy for identifying migrants we first propose an internal validation procedure. This defines gold standard datasets for residence and nationality and computes the classification performance of our two strategies to identify the two user attributes. The gold standard datasets are produced using profile information as they are provided by the users themselves. We then perform an external validation where we compare the migrant percentages obtained in our data with those from official statistics.

5.1 Internal Validation: Gold Standards Derived from Our Data

Residence. To devise a gold standard dataset for residence we consider profile locations set by users. We assume that if users declare a location in their profile, then that is most probably their residence. Very few users actually declare a location, and not all of them provide a valid one, thus we only selected profile locations that were identifiable to country level. Among the user accounts for which we could estimate the residence, 3,065 accounts had a valid country in their profile location. Using these accounts as our validation data, we computed the F1 score to measure the performance of our residence calculation. Table 1 shows overall results, and also scores for the most common countries individually. The weighted average of the F1 score is 86%, with individual countries reaching up to 94%, demonstrating the validity of our residence estimation procedure.

Nationality. In order to build a gold standard for nationality, we take into account the profile language declared by the users. The assumption is that profile languages can provide a hint of one's nationality [13]. However, many users might not set their profile language, but use the default English setting. For this reason, we do not include into the gold standard users that have English as their profile language.

Table 1. Average precision, recall and F1 scores, together with scores for the top 7 residences in terms of support size.

	Weighted Avg	Macro avg	Micro avg	IT	KW	US	ID	SG	AU
F1-score	0.858	0.716	0.856	0.928	0.839	0.703	0.945	0.83	0.891
Precision	0.879	0.745	0.856	0.935	0.989	0.572	0.949	0.946	0.883
Recall	0.856	0.727	0.856	0.921	0.728	0.91	0.941	0.739	0.899
Support	3065	3065	3065	343	125	122	119	119	109

Table 2. Average precision, recall and F1 scores for top 8 nationalities in terms of support numbers

	Weighted avg	Macro avg	Micro avg	IT	ES	TR	RU	FR	BR	DE	AR
F1-score	0.99	0.98	0.72	0.99	0.96	0.98	0.95	0.94	0.95	0.92	0.97
Precision	0.99	0.98	0.73	1	0.94	0.98	0.98	0.9	0.96	0.91	0.98
Recall	0.98	0.98	0.75	0.99	0.97	0.99	0.93	0.98	0.94	0.93	0.95
Support	12223	12223	12223	10781	302	173	146	118	113	86	59

The profile language, however, does not immediately translate into nationality. While for some languages the correspondence to a country is immediate, for many others it is not. For instance, Spanish is spoken in Spain and most American countries, so one needs to select the correct one. For this, we look at tweet locations. We consider all countries that match with the profile language and, among these, we select the one with the largest number of tweets, but only if the number of tweets from that country is at least 10% of the total number of tweets of that user. This allows to select the most probable country, also for users who reside outside their native country. If no location satisfies this criterion the user is not included in the gold standard. We were able to identify nationalities of 12,223 users. Due to the fact that during data collection we focused on geo-tags in Italy, the dataset contains a significant number of Italians.

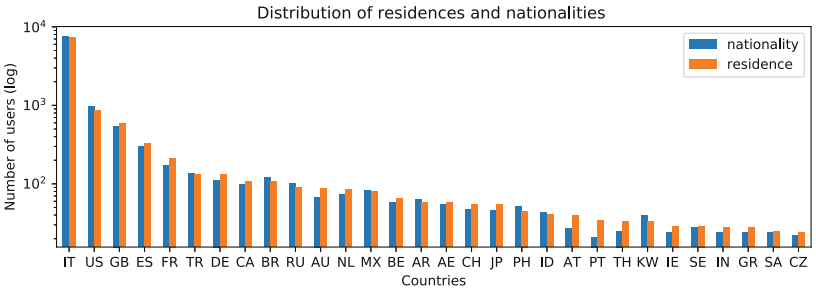

Fig. 3. Distribution of residences and nationalities of top 30 countries, for all users that possess both residence and nationality labels.

We employed this gold standard dataset in two ways. First, we needed to select suitable values for the w weights from Eqs. 3–4. These show the importance of the four components used for nationality computation: own language and location, friends' language and location. We performed a simple grid search and obtained the best accuracy on the gold standard using values 0 for languages and 2 and 1.5 for own and friends' location, respectively. Thus we can conclude that it is the locations that are most important in defining nationality for twitter users, with a slightly stronger weight on the individual's location rather than the friends. The final F1-score, both overall and for top individual nationalities, are included in Table 2, showing a very good performance in all cases.

To assign final residences and nationalities to our core users, we combined the predictions with the gold standards (we predicted only if the gold standard was not present). Figure 3 shows the final distribution of residences and nationalities of top 30 countries for all users that have both the residence and nationality labels. The difference in the residence and nationality can be interpreted as either immigrants or emigrants.

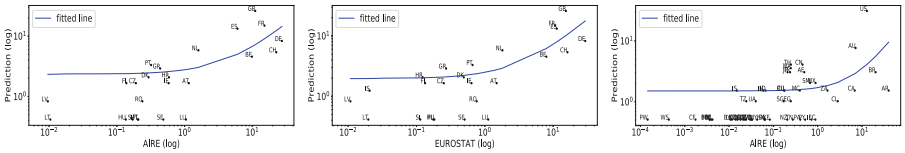

Fig. 4. Comparison between the true and predicted data; the first two plots show predicted versus AIRE/EUROSTAT data on European countries. The last plot shows predicted versus AIRE data on non-European countries.

5.2 External Validations: Validation with Ground Truth Data

In order to validate our results with ground truth data, we study users labelled with Italian nationality and non-Italian residence, i.e. Italian emigrants. We computed the normalised percentage of Italian emigrants resulting from our data for all countries, and compared with two official datasets: AIRE (Anagrafe Italiani residenti all'estero), containing Italian register data, and Eurostat, the European Union statistical office. For comparison we use Spearman correlation coefficients, which allow for quantifying the monotonic relationship between the ground truth data and our estimation by taking ranks of variables into consideration.

Figure 4 displays the various values obtained, compared with official data. A first interesting remark is that even between the official datasets themselves, the numbers do not match completely. The correlation between the two datasets is 0.91. Secondly we observed good agreement between our predictions and the official data for European countries. The correlation with AIRE is 0.753, while with Eurostat it is 0.711 when considering Europe. For non-European countries, however the correlation with AIRE data drops to 0.626. We believe the lower performance is due to several factors related to sampling bias and data quality in the various datasets. This includes bias on Twitter and in our methods, but also errors in the official data, which could be larger in non-EU countries due to less efficient connections in sharing information.

All in all, we believe our method shows good performance and can be successfully used to build population level indices for studying migration. We do not aim to perform nowcasting of immigrant stocks, but rather to identify a population that can be representative enough for further analyses.

6 Case Study: Topics on Twitter

In this section we show that our methodology can be employed to study how trending topics in Italy are also being discussed among Italian emigrants. As an example, we selected one hashtag that has been very popular in the last years: #Salvini. This refers to the Italian politician Matteo Salvini who served as Deputy Prime Minister and Minister of internal affairs in Italy until recently. To this, we added the top nine hashtags that appear frequently with #Salvini in our data: Berlusconi, Conti, Diciott, DiMaio, Facciamorete, Legga, M5S, Migrant, Ottoemezzo. Indeed, they all represent people that are often mentioned together or political parties or other issues that are associated with the hashtag #Salvini.

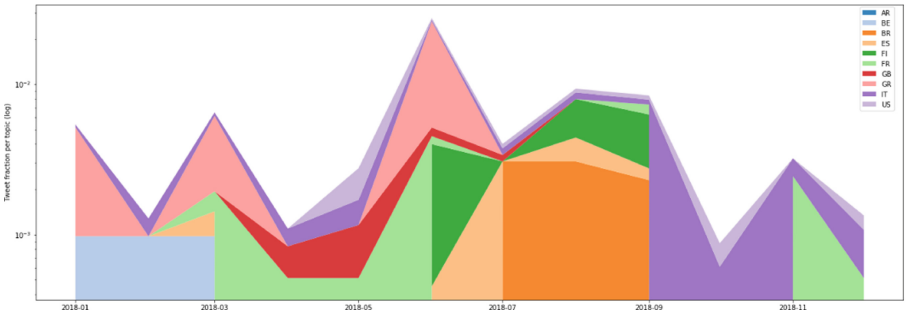

Fig. 5. Stream graph: appearance of hashtags related to #Salvini from Italians across 10 selected residence countries in 2018. The discussion continuously appeared in Italy throughout the year and it became more lively employed by Italians overseas as Salvini gained more political attention.

Figure 5 shows an evolution of the usage of the 10 above mentioned hashtags across different Italian communities both within and abroad Italy. The values shown are the number of tweets from Italian nationals residing in each country that include one of the 10 hashtags, divided by the total number of tweets from Italian nationals from that country. Values are computed monthly. Thus, we show the monthly popularity of the topics in each country. In this way, even the tweets from less represented countries are well shown. As the figure shows, the hashtag was continuously used by Italians in Italy. We observed that the hashtag gradually spread over other residence countries as Salvini received more and more attention. We also observe that most of the attention comes from Italians residing in Europe, with non-European countries less represented.

7 Conclusion and Future Work

We have developed a new methodology to provide a snapshot of migrants within the Twitter population. We considered the length of stay in a country as the

key factor to define a user's residence. As for the nationality, connections which migrants maintain with their country of origin provided us with a good indication. In particular, the location of friends seemed to be a strong feature in determining nationality, together with the location of the users themselves. Tweet language, on the other hand, was not considered relevant by our model. This is probably due to the fact that English is the dominating language on Twitter, since a language that is widely understood has to be spoken to get more attention from other users. We have validated our results both with internal and external data. The results show good classification performance scores and good correlation coefficients with official datasets.

The constructed dataset can be applied in different scenarios. We have shown how it can be used to study trending topics on Twitter, and how attention is divided between emigrants and non-migrants of a certain nationality. In the future, we plan to analyse social ties, integration and assimilation of migrants [7]. At the same time, one can investigate the strength of the ties with the community of origin.

References

1. Castillo petruzzi case (1999)
2. Assal, M.A.: Nationality and citizenship questions in Sudan after the Southern Sudan referendum vote. Sudan Report (2011)
3. Blumenstock, J.E.: Inferring patterns of internal migration from mobile phone call records: evidence from Rwanda. Inf. Technol. Dev. **18**(2), 107–125 (2012)
4. Coletto, M., et al.: Perception of social phenomena through the multidimensional analysis of online social networks. Online Soc. Netw. Media **1**, 14–32 (2017)
5. Donner, R.: The Regulation of Nationality in International Law, 2d edn, p. 289. Leiden, Brill Nijhoff (1994). https://brill.com/view/title/14000, ISBN 978-09-41-32077-1
6. Hailbronner, K.: Nationality in public international law and European law. JSTOR (2006)
7. Herdağdelen, et al.: The social ties of immigrant communities in the united states. In: Proceedings of the 8th ACM Conference on Web Science, pp. 78–84. ACM (2016)
8. Huang, W., et al.: Inferring nationalities of Twitter users and studying international linking. In: Proceedings of the 25th ACM Conference on Hypertext and Social Media, pp. 237–242. ACM (2014)
9. Kikas, R., et al.: Explaining international migration in the Skype network: the role of social network features. In: Proceedings of the 1st ACM Workshop on Social Media World Sensors, pp. 17–22. ACM (2015)
10. Lamanna, F., et al.: Immigrant community integration in world cities. PLoS One **13**(3), e0191612 (2018)
11. Mazzoli, M., et al.: Migrant mobility flows characterized with digital data. arXiv preprint arXiv:1908.02540 (2019)
12. Sîrbu, A., et al.: Human migration: the big data perspective. Int. J. Data Sci. Anal. (2020, under review)
13. Stokes, B.: Language: the cornerstone of national identity. Pew Research Center's Global Attitudes Project (2017)

14. Zagheni, E., et al.: Combining social media data and traditional surveys to nowcast migration stocks. In: Annual Meeting of the Population Association of America (2018)
15. Zagheni, E., et al.: Inferring international and internal migration patterns from Twitter data. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 439–444. ACM (2014)
16. Zagheni, E., Weber, I.: You are where you e-mail: using e-mail data to estimate international migration rates. In: Proceedings of the 4th Annual ACM Web Science Conference, pp. 348–351. ACM (2012)