



Twitter and migration research

Jisu Kim¹

¹Max Planck Institute for Demographic Research

Outline



- ▶ Twitter data and migration research
- ▶ How migrants are identified on Twitter
- ▶ Validation
- ▶ What can we do beyond identifying migrants on Twitter?
- ▶ Pros and Cons of Twitter data

Twitter data and migration research

1. We should identify migrants:

Information from Twitter data that can be used in migration research:

- ▶ Geo-tagged tweets
- ▶ Self-declared location information from profile
- ▶ Social networks: followers and friends
- ▶ Name
- ▶ Language

Definitions of migrants on Twitter in the literature

- ▶ “Any individual leaving Venezuela during the time window of observation.” (Mazzoli et al. 2020)
- ▶ “Anyone who tweeted exclusively from Venezuela in the time period between Feb. 1 and April 30 2017.” (Hausmann et al. 2018)
- ▶ “A Twitter user has the nationality that others believe you have.” (Huang et al. 2014)
- ▶ “Migrants are users that are identified as people who moved to a different country for at least one of the 4-month periods.” (Zagheni et al. 2014)
- ▶ “A migrant is a person that has the residence different from the nationality.” (Kim et al. 2020)

Identifying mobility

“Any individual leaving Venezuela” (Mazzoli et al. 2020),
“Anyone who tweeted from Venezuela” (Hausmann et al. 2018),
“people who moved to a different country” (Zagheni et al. 2014)

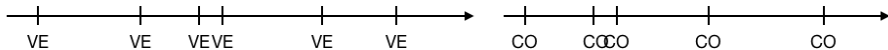
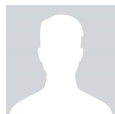


Figure: Mock example of geo-tagged tweet timeline

Inferring nationalities of Twitter users in Qatar (Huang et al. 2014)



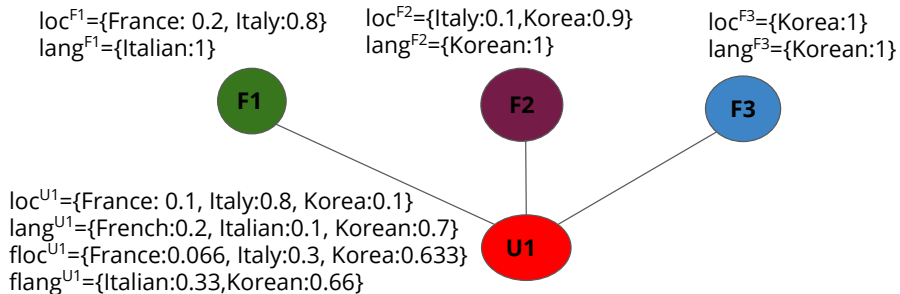
- ▶ Twitter page: xx
- ▶ Homepage: XX
- ▶ Location: Doha, Qatar
- ▶ Time zone: UTC
- ▶ Interface language: English
- ▶ Tweets languages: (1) en: 97.81% (2) de: 2.13% (3) ru: 0.02%
- ▶ Follower locations: (1) US: 151 (2) DE: 72 (3) QA: 26
- ▶ Following locations: (1) US: 128 (2) DE: 89 (3) ES: 57
- ▶ Tweets from (1) US: 23 (2) QA:19 (3) TR: 4

Identifying migrants on Twitter (Kim et al. 2020)

“A migrant is a person that has the residence different from the nationality.”

- ▶ Country of residence: “the country with the longest length of stay”
- ▶ Country of nationality: “the ensemble of features that make a person feel like they belong to a certain country”

Identifying migrants on Twitter (Kim et al. 2020) 2



Potential issues

- ▶ Tourists
- ▶ Travel blogs
- ▶ Company or Organisation accounts
- ▶ Bots
- ▶ ...

We need to validate that migrants have correctly been identified!

Method validation

We need to validate the identification method to prove that the method is reliable and reproducible.

- ▶ Compare with ground truth data (Kim et al. 2020)
- ▶ Compare with golden standard data¹ (Huang et al. 2014; Kim et al. 2020)
- ▶ Diff-in-Diff (Zagheni et al. 2014) (Correction of bias)
- ▶ Manually checking subsample of data

¹A gold standard data is a high quality data that is the closest to the ground truth data that you can get.

Validation-Official statistics (Kim et al. 2020)

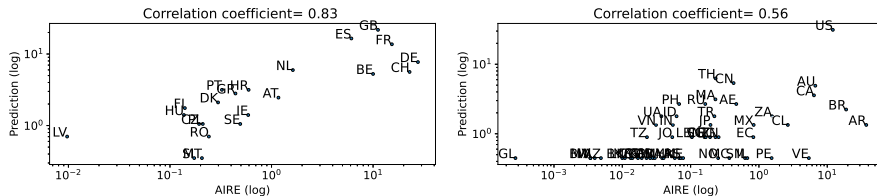


Figure: Correlation between predicted data and official statistics (Kim et al. 2020)

Validation-Gold standard data (Huang et al. 2014)

	Pre.	Rec.	F1
QA	86.67%	95.37%	90.81%
ARA	82.96%	71.16%	76.56%
WES	70.86%	70.62%	70.64%
SA	93.35%	90.48%	91.89%
IN	82.19%	71.13%	76.00%
OTH	78.76%	40.72%	53.54%
UN	30.78%	15.13%	20.16%

Table: The average Precision, Recall, and F1 scores for each nationality group (Huang et al. 2014)

Inferring international and internal migration patterns from Twitter data (Zagheni et al. 2014)

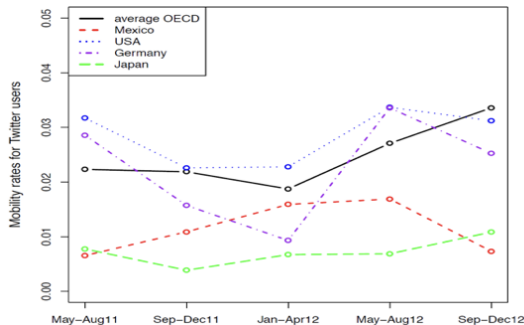


Figure: Mobility rates for Twitter users (Zagheni et al. 2014)

Difference-in-Differences

- ▶ Out-migration rates clearly an overestimate
- ▶ Non-representative user set
- ▶ Selection bias is changing over time
- ▶ Focus on between-country differences

$$\hat{\delta}_c^t = (m_c^t - m_{oeed}^t) - (m_c^{t-\Delta} - m_{oeed}^{t-\Delta})$$

→ Diff-in-diff estimator to evaluate relative changes in trends (Zagheni et al. 2014)

Cont.

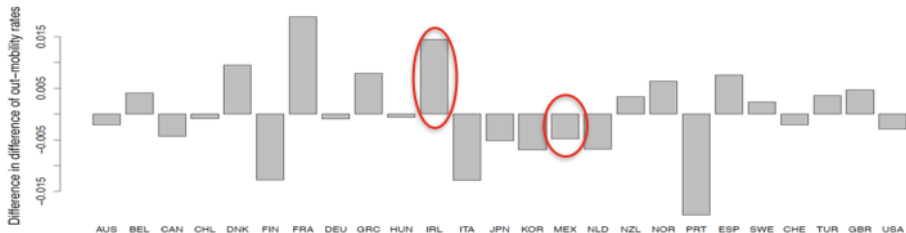


Figure: (Soft) Validation: Ireland out-migration rate grew by 2.2 percent 2011 - 2012, more than most countries (Irish Central Statistics Office) Mexico also sees a reduction in out-migration (Pew Research Center)

Other migration related researches

- ▶ Cultural integration (Kim et al. 2021b)
- ▶ Spatial integration (Mazzoli et al. 2020; Lamanna et al. 2018)
- ▶ Sentiment analysis (Öztürk et al. 2018; Arcila-Calderón et al. 2021)
- ▶ Social network analysis (Kim et al. 2021a)

Cultural integration (Kim et al. 2021b)

*Q. How much do migrants absorb the culture of their destination society?
Do they lose connection with their home country?*

Traditional studies focus on elements such as:

- ▶ Language proficiency
- ▶ Marital status
- ▶ Role of media ...

	Low HA	High HA
Low DA	Marginalisation	Separation
High DA	Assimilation	Integration

Table: Theories of integration and their relation to HA and DA (Kim et al. 2021b)

Home and Destination attachment indexes (Kim et al. 2021b)

$$H(h) = \frac{-\sum_c P_h(c) \log P_h(c)}{\log(|P_h(c)|)} \quad (1)$$

$$HA(u) = \frac{\# C_n(u) \text{ hashtags}}{\# \text{ total hashtags}} = \frac{HT(u, C_n(u))}{HT(u)} \quad (2)$$

$$DA(u) = \frac{\# C_r(u) \text{ hashtags}}{\# \text{ total hashtags}} = \frac{HT(u, C_r(u))}{HT(u)} \quad (3)$$

```
'Salvini':  
  {'CH': 1,  
   'CL': 1,  
   'CZ': 2,  
   'DE': 2,  
   'ES': 3,  
   'FR': 2,  
   'GB': 1,  
   'IT': 544,  
   'LU': 1,  
   'NL': 2,  
   'PL': 1,  
   'TH': 2,  
   'TR': 1,  
   'US': 6}  
    /569  
Entropy score=0.11  
∴ Italian specific
```

HA & DA indexes

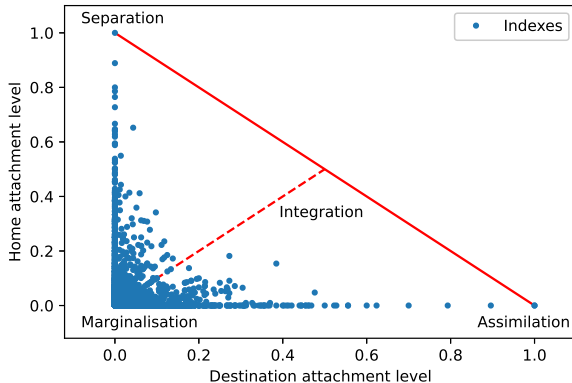


Figure: Scatter plot between home and destination attachment indices for all the migrants in the data (Kim et al. 2021b)

Sentiment analysis (Arcila-Calderón et al. 2021)

Hate speech towards migrants/refugees? What are the main topics behind the messages in Twitter in Spain with hate speech against refugees?

Topic	Most Common Words and Their Frequency within the Topic (the Underlined Words are the Most Determinant for the Labelling of the Topic)
Pull effect and consequences	"immigrants" (0.008) + "spain" (0.007) + "boat" (0.005) + <u>"effect"</u> (0.005) + <u>"pull"</u> (0.004) + "goes" (0.004) + "valencia" (0.004) + "consequences" (0.004) + "europe" (0.004) + <u>"concentration"</u> (0.004)
Pull effect and not welcoming "illegals"	"immigrants" (0.013) + "people" (0.008) + "spain" (0.008) + "pull" (0.007) + <u>"effect"</u> (0.007) + <u>"illegal"</u> (0.007) + <u>"protectyourborders"</u> (0.007) + "spaniards" (0.006) + "harbor" (0.006) + <u>"aquariusnotwelcome"</u> (0.006)
Not welcoming and terrorism	"go" (0.008) + "spain" (0.008) + "people" (0.006) + <u>"aquariusnotwelcome"</u> (0.005) + "country" (0.005) + <u>"boko"</u> (0.005) + <u>"haram"</u> (0.005) + "immigrants" (0.005) + "boat" (0.004) + "have" (0.004)
Smugglers and NGOs	"immigrants" (0.023) + "spain" (0.010) + "spaniards" + <u>"come"</u> (0.008) + <u>"mafias"</u> (0.008) + "people" (0.007) + "go" (0.006) + <u>"illegal"</u> (0.006) + <u>"ngos"</u> (0.006) + "boat" (0.006)
Money and jobs	"refugees" (0.011) + "immigrants" (0.009) + <u>"pay"</u> (0.007) + "spain" (0.007) + "countries" (0.005) + "boat" (0.005) + "spaniards" (0.005) + "people" (0.005) + <u>"solution"</u> (0.005) + <u>"work"</u> (0.005)
Entrance to Europe	"spain" (0.018) + "immigrants" (0.009) + "europe" (0.006) + "boat" (0.006) + "valencia" (0.005) + "spaniards" (0.005) + "government" (0.005) + "immigration" (0.005) + <u>"north"</u> (0.004) + <u>"millions"</u> (0.005)

Social network analysis (Kim et al. 2021a)

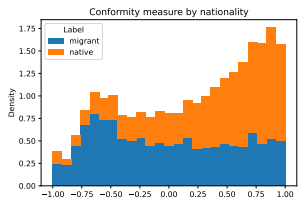


Figure: Nationality label

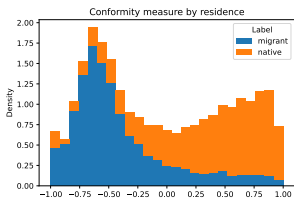


Figure: Residence label

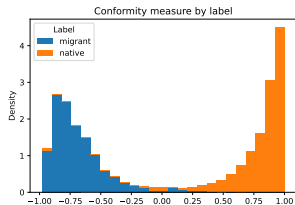


Figure: Migrant label

Figure: Stacked histogram of conformity measures

Pros and cons

Pros:

- ▶ Access
- ▶ Open box
- ▶ Longitudinal: access to historical data

Cons:

- ▶ Bias
- ▶ Sample size

Issue:

- ▶ Ethics and Privacy

What can we do to mitigate the issues related to Ethics and Privacy?

To be continued...

Supplement materials

F1-score measures the accuracy;

$$F1score = 2 * \frac{precision * recall}{precision + recall}$$






- Precision: How many selected items are relevant?





$$\frac{Truepositive}{Truepositive + Falsepositive}$$

- Recall: How many relevant items are selected?

$$\frac{Truepositive}{Truepositive + Falsenegative}$$

The score ranges from its worst score of 0 to its best score 1.

-  Arcila-Calderón, Carlos et al. (2021). “Refugees Welcome? Online Hate Speech and Sentiments in Twitter in Spain during the Reception of the Boat Aquarius”. In: *Sustainability* 13.5, p. 2728.
-  Hausmann, Ricardo et al. (2018). *Measuring venezuelan emigration with twitter*. Tech. rep. Kiel Working Paper.
-  Huang, Wenyi et al. (2014). “Inferring nationalities of twitter users and studying inter-national linking”. In: *Proceedings of the 25th ACM conference on Hypertext and social media*, pp. 237–242.
-  Kim, Jisu et al. (2020). “Digital footprints of international migration on twitter”. In: *International Symposium on Intelligent Data Analysis*. Springer, pp. 274–286.
-  Kim, Jisu et al. (2021a). “Characterising different communities of Twitter users: Migrants and natives”. In: *International Conference on Complex Networks and Their Applications*. Springer, pp. 130–141.
-  Kim, Jisu et al. (2021b). “Home and destination attachment: study of cultural integration on Twitter”. In: *arXiv preprint arXiv:2102.11398-Under review*.

-  Lamanna, Fabio et al. (2018). “Immigrant community integration in world cities”. In: *PloS one* 13.3, e0191612.
-  Mazzoli, Mattia et al. (2020). “Migrant mobility flows characterized with digital data”. In: *PloS one* 15.3, e0230264.
-  Öztürk, Nazan et al. (2018). “Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis”. In: *Telematics and Informatics* 35.1, pp. 136–147.
-  Zagheni, Emilio et al. (2014). “Inferring international and internal migration patterns from twitter data”. In: *Proceedings of the 23rd International Conference on World Wide Web*, pp. 439–444.