Linking Digital Traces and Survey Data: Politus Project

M. Fuat Kına
Koç University, Sociology
PhD candidate

Şükrü Atsızelti
Koç University, Sociology
PhD student

Outline

- ❑The problem
- ❑Recent literature
- ❑Politus project
- ❑Why are we linking digital trace and survey data?
- ❑Online survey design
- ❑Outputs

Problem

The Problem

Classical surveys have problems and limitations to understand human behavior in a fast-moving world.
Social media monitoring / listening enables catching up with the high pace, yet bringing data based problems in itself.

CLASSICAL SURVEYS

- ➢ No trend tracking
- ➢ Geographical limitations
- ➢ Not fast, not real time
- ➢ Low response rate
- ➢ Response bias

(when respondents' answers do not reflect their true beliefs)

SOCIAL MEDIA MONITORING / LISTENING

- ➢ Biased (unrepresentative) population of social media
- ➢ Noisy (not clean) data
- ➢ Non-organized data
- ➢ Vocal users
- ➢ Too much data: difficulty in context interpretation

Recent literature

- Ex ante vs ex post linking
- Aggregate vs individual-level linking

Table 1. Linking types with examples from the literature.

Ex Ante Linking	Ex Post Linking
(A) Aggregate level • Analysis of audience overlaps (e.g., Mukerjee et al., 2018; Nelson & Webster, 2017) • Analysis of aggregate audience statistics (e.g., political ideology, Nelson & Webster, 2017)	(B) Aggregate level Linking survey responses to digital trace data . . . • Temporally: both are generated during the same time period (e.g., Mellon, 2014; O'Connor et al., 2010; Stier et al., 2018) • Topically: both focus on the same topic (e.g., Pasek et al., 2019) • Geographically: both can be located within same geographic area (e.g., Beauchamp, 2017)
	(C) Public actors Link publicly available digital trace data of public actors (e.g., politicians or organizations) to their survey responses (e.g., Karlsen & Enjolras, 2016; Quinlan et al., 2017)
(D) Individual level Ask individuals in surveys for informed consent to record in real time: • Website visits (e.g., Guess, 2015; Jürgens et al., 2019; Möller et al., 2019; Vraga & Tully, 2018) • Smartphone data (e.g., Boase & Ling, 2013; Jürgens et al., 2019; Kreuter et al., 2019) • Sensor data (e.g., Génois, Zens, Lechner, Rammstedt, & Strohmaier, 2019)	(E) Individual level Ask individuals in surveys for informed consent to collect their historical digital trace data . . . • From social media APIs (e.g., Al Baghal et al., 2019; Haenschen, 2019; Hofstra, Corten, van Tubergen, & Ellison, 2017; Hopp, Vargo, Dixon, & Thain, 2018; Vaccari et al., 2015; Wells & Thorson, 2015) • via data donation, for example, personal Google or Facebook histories (e.g., Thorson et al., 2018)

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review*, *38*(5), 503-516

Recent literature

- Social media, geospatial data, sensor data

	Data type	Ex ante	Ex post
Aggregate level	**Social media**	• Collecting tweets for the same region and period of time as the survey data using the Stream API	• Linking survey data with counts of posts (about a certain topic) or aggregate sentiment scores for posts from existing social media data collections for specific regions or time periods
	Geospatial data	• Simultaneous recording of data for surveyed area (e.g., weather, pollution or noise data collected via sensors)	• Linking aggregated survey data for specific geographic areas to available geospatial data (e.g., on access to certain amenities, pollution, noise, etc.)
	Sensor data	• Simultaneous recording of health data of a surveyed group (e.g., a sports team)	• Linking aggregated medical data for specific populations (e.g., blood oxygen levels in previous studies)

Individual level	**Social media data**	• Ask survey respondents for consent to collect their current/latest social media data (e.g., via an API or a browser plugin) for a specified period of time (during + maybe also after the survey field time)	• Ask individuals in surveys for informed consent to collect their historical digital trace data...
			  • via social media APIs
			  • via data donation (e.g., personal Google, Twitter or Facebook archives)

	Geospatial data (note: these are usually not generated/available on the individual level)	• Record location data from respondents (self-report, e.g., via experience sampling or tracked GPS data from devices)	• Linking survey data to existing geospatial data (e.g., on access to certain amenities, pollution, noise, etc.) on the level of the location/address of individual participants
	Sensor data	• Equiping respondents with fitness trackers for the time of the study	• Accessing fitness tracker data stored on respondents devices (e.g., via data donation)

Beuthner, Christoph, Johannes Breuer, and Jünger, Stefan (2021). Data Linking - Linking survey data with geospatial, social media, and sensor data. Mannheim, GESIS Survey Guidelines.

Politus Project: Using Digital Traces to Predict Political and Social Trends

Politus Project

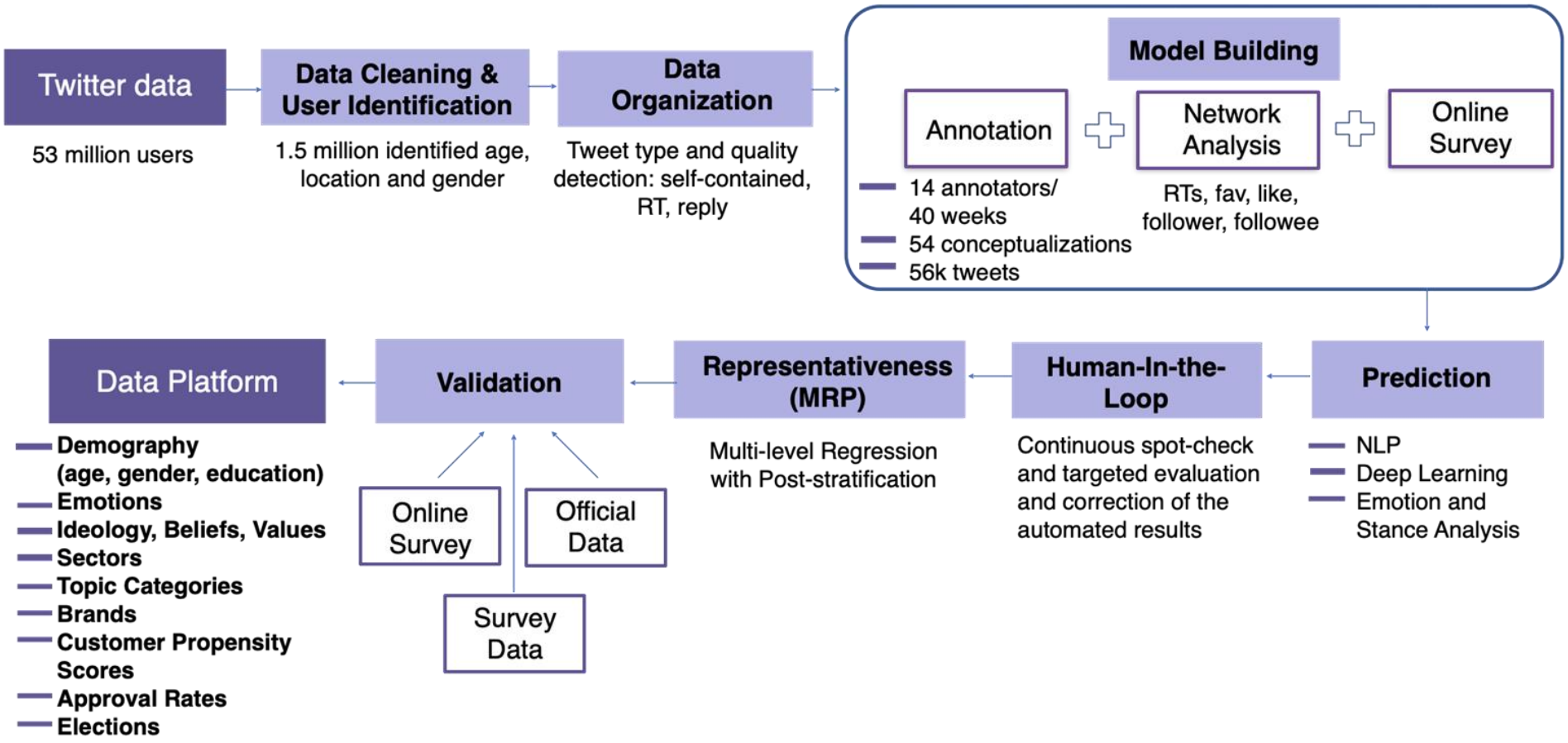
Politus Project

- Politus briefly develops an AI-based innovation that combines quantitative and computational methods, to create a data platform.
- It aims to deliver representative, valid, instant, real-time panel data on key political and social trends.

Politus Project

Methodology

Large team composed of social scientists, computer scientists, mathematicians and economists

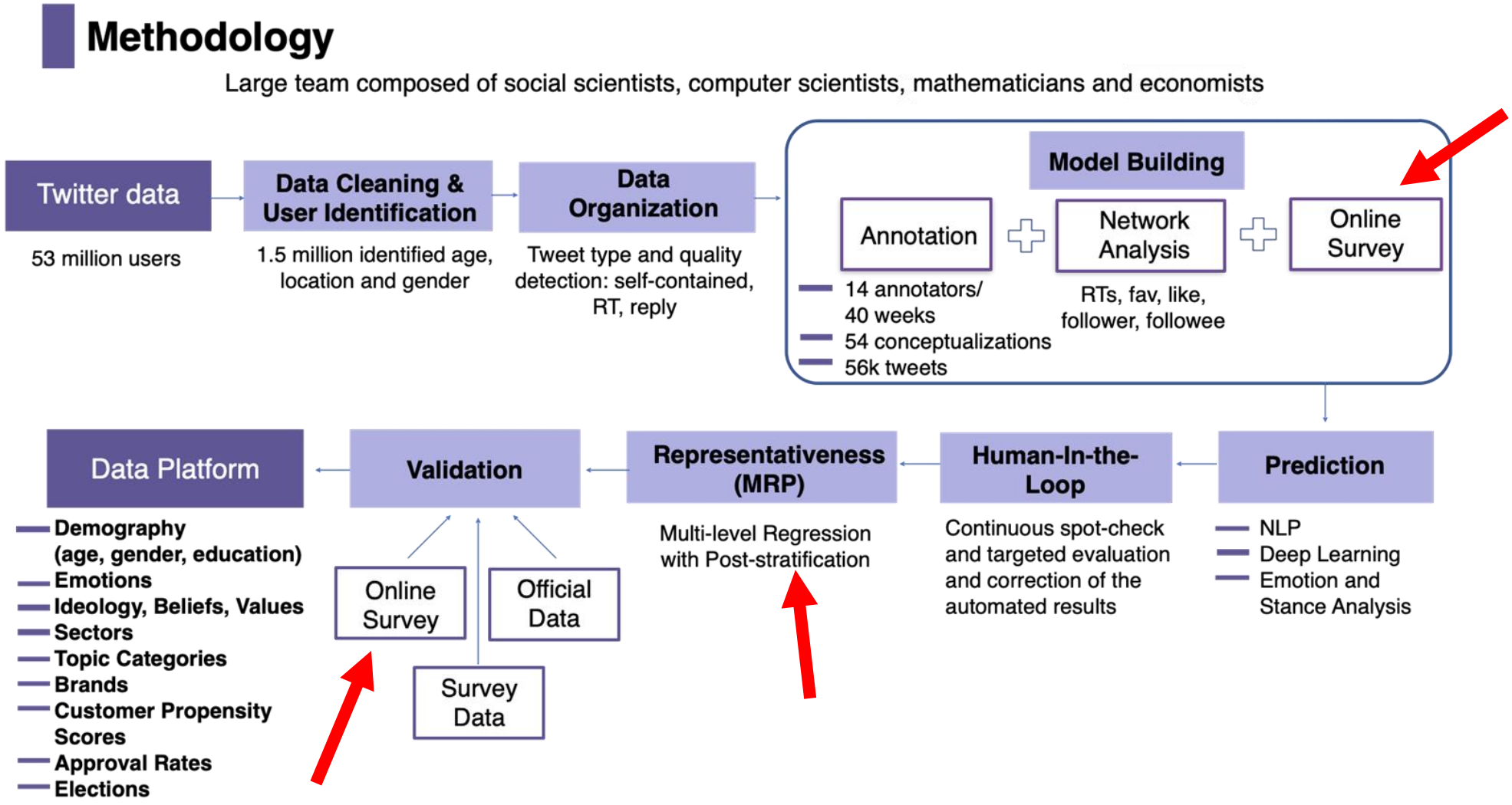


Linking

Why are we linking digital traces and survey data?

1. Validation
2. Measurement
3. Predicting education level

Why are we linking digital traces and survey data?



1. Validation

Lack of validity

Tweet-to-user transformation

- How can we decide whether this user is socialist, or happy; what is the relevant time range; what can be the thresholds?

User A: 3 positive among 100 tweets
User B: 2 positive among 2 tweets

2. Measurement

- Measuring the error
  - Self-censorship bias
  - Vocal user problem

- Model building for user-level information
  - Connecting text-based tweet-level estimations with user-level network and image data

3. Predicting Education Level

- Why is education variable vitally important?
  - Multilevel regression and poststratification: age, gender, location

- Lack of ground truth data for twitter users
- Existing research develop rule-based transformation
  - Twitter bio/description field
  - Facebook profiles

- However, survey data provide self-reported education levels of users.

Online Survey

Politus Project

Platform

- Facebook vs. Twitter

Twitter	Facebook
Target platform	Intermediary platform
Underemployed	Established practices for surveys
Broadly targeting/Uploading list option	Detailed targeting
Needs an intermediary company	Needs identity confirmation
Min. 10.000 TL for the first time	Low cost trials

Informed Consent

- GDPR Requirements

“Processing personal data is generally prohibited…” 😱😱😱

“unless it is expressly allowed by law, or the data subject has consented to the processing.”

Informed Consent

- Clear information about the project, PI, the data processing steps, data confidentiality and data storage
- Clear information about why we demand their handles
- Clear statement on the right of withdrawal
- Contact info if participant wants to withdraw his/her data

Incentive

- Bad for representative sample list
- Good for unattractive tasks

We had an unattractive task for a representative sample

Questionnaire Design

- Handle
- Topics covered in Politus annotation manual
- Variables predicted from Twitter data through inference tools (like gender, age and location)
- Variables that haven’t been inferred but are of interest like education, occupation and ethnicity.
- Vote preference and job approval
- Tweeting behavior
- Self-censorship

Advertisement Process

- Step 1

Creating a representative sample

- Users divided into 5 groups based on activity levels over various periods.
- Algorithm calculates necessary person count per category according to Tüik
- Users drawn from the most active users.

Advertisement Process

	gender	age_group	location	user_level_edited	count
0	female	19-29	1	1	90
1	female	19-29	1	2	126
2	female	19-29	1	3	342
3	female	19-29	1	4	123
4	female	19-29	1	5	238

Advertisement Process

	gender	age_group	count	location
0	male	<=18	365921	1
1	male	<=18	119395	2
2	male	<=18	108447	3
3	male	<=18	112799	4
4	male	<=18	69901	68

Advertisement Process

```python
#Function for dividing
def divide_counts(df, desired_num):
    total_count = df['count'].sum()
    ratio = desired_num / total_count
    df['target'] = df['count'] * ratio
    df['target'] = df['target'].round()
    return df
```

- Determining needed user counts for each specific category

- Target numbers for different categories calculated for a sample of 500.000 participants

	gender	age_group	count	location	target
0	male	<=18	365921	1	2145.0
1	male	<=18	119395	2	700.0
2	male	<=18	108447	3	636.0
3	male	<=18	112799	4	661.0
4	male	<=18	69901	68	410.0

Advertisement Process

- Checking most active users for each category

if user count is greater than expected:
then the code randomly takes all
if user count is equal to the expected:
then the code takes all
if else (if the user count smaller than expected):
then the code takes all and check the next most active category

- Checking the missing values and determining the sample size

Advertisement Process

- Why Step I failed?

- Sample is too small!
- Accessing the Twitter help desk

Advertisement Process

- Step II

Accessing Twitter users via Facebook

- Why it has failed?

Almost nobody shared their handle.

Advertisement Process

- Step III

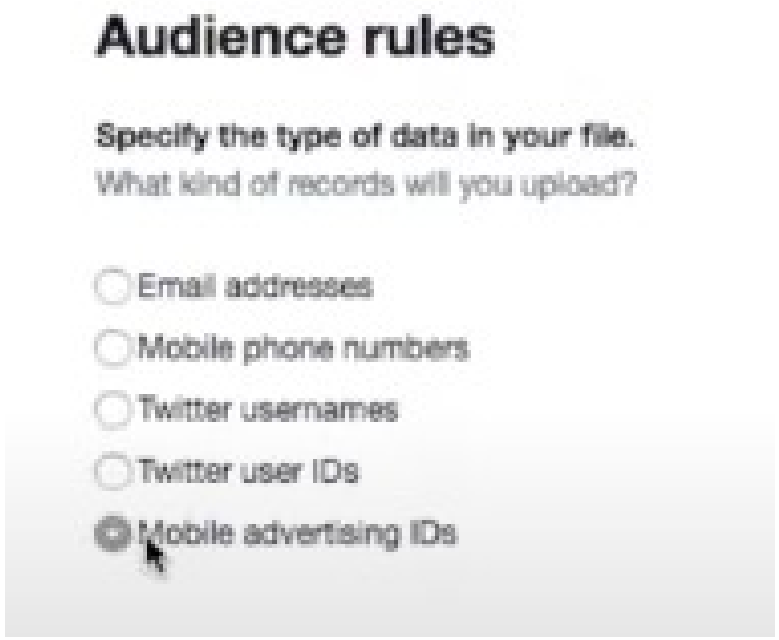
Two-step advertisement from Twitter

- First advertisement toward general population
- Second advertisement toward the underrepresented groups
- Problems with platform affordance, blue tick

Advertisement Process

Create your List Custom Audiences

You might gather these records from your mailing list, past purchasers, or potential customers who have shown interest. You can upload lists of email addresses, Mobile Advertising IDs (iOS Advertising Identifiers and Google Advertising IDs, or when not available, Android IDs), Twitter @handles, or Twitter user IDs.



Advertisement Process

- Surprisingly, many participants are already in our dataset

- Need for Twitter data collection

New APIs, ethicality of web scraping

- Last catastrophes

Twitter’s (almost) unique place among others

Capitalism at the end?

Future Tasks

- Getting the remaining Twitter user data from Twitter 😢
- Validation of models
- Preparation of artificial neural network model for predicting the education
- …

Thank you

Politus Project