Archiving twitter data linked to UKHLS

Paulo Serôdio
ISER, University of Essex

Linking Survey and Digital Trace data, Natcen 2024

Acknowledgements

- ESRC grant: “Understanding [Offline/Online] Society: Linking Surveys with Twitter Data”

- Luke Sloan (Cardiff, PI)
- Matthew Williams (Cardiff, Co-I)
- Shujun Liu (Cardiff)
- Curtis Jessop (NatCen, Co-I)
- Tarek Al Baghal (Essex, Co-I)

Background

Motivation

- UKHLS Innovation Panel (wave 10) asked respondents for consent to link their Twitter data to the survey responses.
  - Annual probability panel, fielded in 2017
  - N = 1945
  - Consented to linkage = 171
  - Active, public accounts throughout data collection = 127
- Twitter consent also asked in IP15, but only IP10 data was used for deposit.

What?

- Link UKHLS survey data to participants' Twitter accounts

+

- Deposit linked data under a EUL

Why?

- Continuous, real-time data collection
- New behavioural metrics
- Adjustments to non-response, recall, social desirability bias, errors in self-report
- Survey Augmentation/Replacement
- Validation survey measurements
- Crosses disciplinary boundaries (sociology, psychology, data science, survey methodology)

Contribution

- Augmenting & sharing social media data:

  1. Unconstrained by Twitter's ToS requirement that content is published "unaltered and with attribution", we can deposit *user* and *tweets* metadata (not just tweet IDs): implications for access, replicability and verifiability in post-API age.

  2. Detailed longitudinal survey data on Twitter users;

- Overcoming practical, legal & ethical challenges;

- Creation of principled framework that inform the different stages of the archiving process

Past Research

- Acquiring consent [Al Baghal et al 2019; Stier et al. 2020]
- Quality of data linkage [Al Baghal et al. 2021]
- Security measures around storage [Sloan et al. 2020]
- Producing study-level metadata [Breuer et al. 2020]

… yet, little guidance on the hurdles of producing **usable** linked data which maintains respondents **anonymity**.

Data linkage approach

Steps

1. Data collection protocol
    - How do collect the data (API, screen scraping, third-party purchase?);
    - Determine **query** and **frequency** of requests: consider velocity of social media data production and how to capture it.
2. Data management workflow
    - License (EULAs, Special License, Secure Data Access?)
    - Derived metrics and raw data
    - De-identification procedures
    - Volume
    - Data Organisation
3. Security Assessment
4. Documenting (study & variable with metadata for archiving)
5. Re-hydrating Tweets + Batch compliance*

Steps

1. Data collection protocol
    - How do collect the data (API, screen scraping, third-party purchase?);
    - Determine **query** and **frequency** of requests: consider velocity of social media data and how to translate this into data.
2. Data management workflow
    - License (EULAs, Special License, Secure Data Access?)
    - **Derived metrics and raw data**
    - **De-identification procedures**
    - Volume
    - Data Organisation
3. **Security Assessment**
4. Documenting (study & variable metadata for archiving)
5. Re-hydrating Tweets + Batch compliance*

Main Hurdle

Main Hurdle

Privacy

WIRED LONG READS BUSINESS CULTURE GEAR SCIENCE SECURITY VIDEO SUBSCRIBE

CHRIS STOKEL-WALKER BUSINESS 09.07.2018 07:00 AM

Twitter's vast metadata haul is a privacy nightmare for users

Working with publicly available metadata from Twitter, a machine learning algorithm was able to identify users with 96.7 per cent accuracy

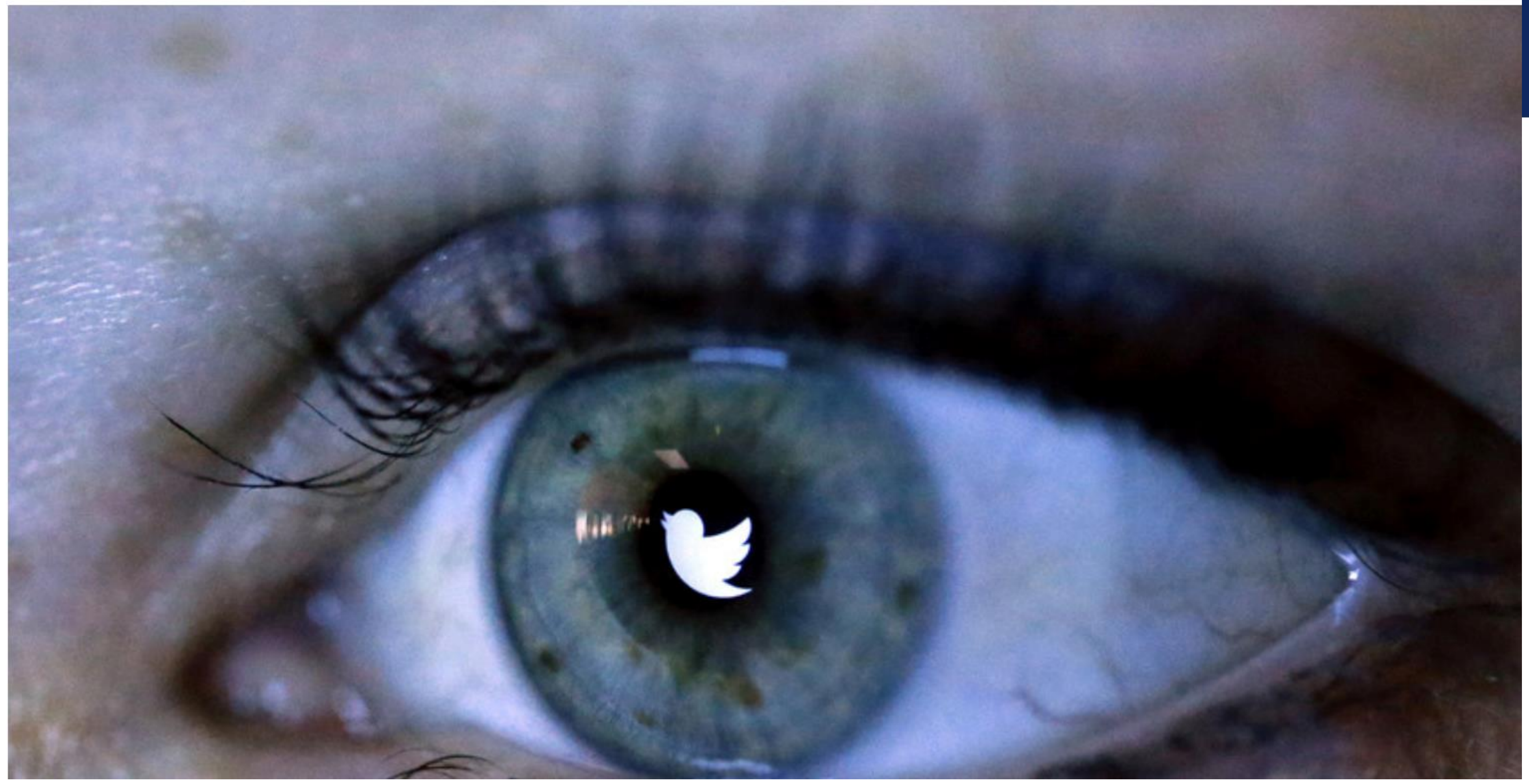
.com NEWS IN YOUR AREA ELECTIONS 2018 BUSINESS SPORT SHOWBIZ LIFE & STYLE EVENTS GUIDE

No anonymity on Twitter, UK researchers find

Researchers could correctly identify a Twitter user from a group of 10,000 with 96.7 percent accuracy

Share Tweet SHARE

BY RUSSIAN TELEVISION
21:18, 20 JUL 2018



MOST POPULAR

Every *Black Mirror* Episode, From

Print subscriptions | Sign in | Search jobs | Search | International edition

Support the Guardian

Fund independent journalism with €5 per month

Support us

The Guardian

News Opinion Sport Culture Lifestyle More

World UK Coronavirus Climate crisis Environment Science Global development Football **Tech** Business Obituaries

Data protection

This article is more than **3 years old**

'Anonymised' data can never be totally anonymous, says study

Findings say it is impossible for researchers to fully protect real identities in datasets

Alex Hern
@alexhern
Tue 23 Jul 2019 16.40 BST

Most viewed

'Soul-crushing': converted Bay Area office apartment fail goes viral

Family members join condemnation of Robert Kennedy Jr's Covid remarks

The Sixth Commandment review - as immaculate a piece of TV as you will ever see

Australia Commonwealth Games 2026: Victoria cancels event after costs blow out to $7bn

White House condemns

Two contradictory objectives

Protection of individual privacy

Publication of detailed individual records

Solutions

- Access restriction
- Statistical Disclosure Control (data obfuscation)
    1. Non-perturbative, deterministic methods: top/bottom coding; banding/grouping
    2. Perturbative, probabilistic methods (noise addition): differential privacy algorithms
- De-identification

Solutions

- Access restriction
- Statistical Disclosure Control (data obfuscation)

undesirable statistical biases

ineffective

1. Non-perturbative, deterministic methods: top/bottom coding; banding/grouping
2. Perturbative, probabilistic methods (noise addition): differential privacy algorithms
3. Differential Privacy algorithms (noise addition)

- De-identification

Query-based

Query-based de-identification

- Social media metadata can be used to unmask anonymised respondents (Perez, Mussolini & Stringhini 2018), but only if the following conditions are met:

    1. Sample properties are known AND

    2. Archived metadata can be used to query the API;

Solution:

- Remove any metadata that can be used to query the API;

- Assess whether API has capacity to recreate the sample & whether it is feasible.

Risks

1. Respondent is re-identified using combination of metadata by agents with access to millions of tweets collected over a period.

    - How was data collected? API v1 very limiting;
    - Probability sample was recreated 100%? Geolocation not active for many users

2. Respondent is re-identified by friend/acquaintance:

    - Chances are the main survey is faster and more effective for re-identification purposes
    - Rocher et al. (Nature 2019) used copula functions to demonstrate 99.98% of Americans could be re-identified in any dataset using 15 demographic attributes.

Data Processing

Our guiding principles [CURTIS]

- **Consistency** in deriving metrics
- **Utility** of the data for research purposes across disciplines
- **Reproducibility** of analytical metrics
- **Transparency** of analytical decisions
- **Integrity** with respect to the raw data
- **Security** of de-identified survey participants

Two datasets

- **Platform-based behaviour** (raw and derived metrics from user-level metadata) [**30 variables**]

- **Tweet metadata** (raw and derived metrics from tweet-level metadata) [**135 variables**]:
  - Tweet raw metadata
  - Sentiment Analysis
  - Syntactic and Lexical Features
  - Readability
  - Lexical Diversity
  - Complex content: Part-of-Speech tagging

Platform-based Behaviour

Variable Name	Description	Type	API Endpoint	Software Dependency (R package)
following	Count of the number of accounts the user was following (at the time of the last API request, in the first quarter of 2023).	integer	User	-
followers	The most recent count of the number of followers of the user’s account.	Integer	User	-
count_reply	The most recent count of the number of tweets posted by the user’s account in reply to a tweet by another user.	Integer	User	
count_quote	The most recent count of quote of tweets posted by the user.	Integer	User	
count_original	The most recent count of original content tweets posted by the user (excludes quoted tweets).	Integer	User	
prop_unique_tweets	Proportion of unique (non-repeated) tweets posted by the respondent. Calculated by dividing the count of distinct tweets by the total number of tweets posted by the respondent.	Numeric	Derived	
own_tweets	Count of the total number of original tweets posted by the respondent excluding simple retweets and liked tweets. This variable includes tweets in which the respondent posts original text and quoted retweets.	Integer	Derived	
hashtoken_ratio	The ratio of the total number of hashtags to the total number of tokens in all the tweets posted by the respondent. It's calculated by pre-processing the tweets using the function described at the beginning of this section, concatenating the text of all	Numeric	Derived	quanteda::ntoken

Tweet-level metadata

Variable Name	Description	Type	Software Dependency (R package)
Sentiment Analysis Tweets were subject to the following pre-processing steps: remove “RT”, remove irregular whitespace, remove URLs, remove emojis, remove hash symbol, separate camel case hashtags into separate words, remove @ symbol from mentions, offset punctuation, create endmarker punctuation for tweets when absent. Sentiment analysis was run at the sentence level and averaged for each tweet			
sentimentr_jockers_rinker_b	Average sentiment score for sentences in the tweet using the combined and augmented version of Jockers (2017) & Rinker’saugmented Hu & Liu (2004) positive/negative word list as sentiment lookup values, ie dictionary of positive/negative word list.	Numeric	sentimentr::sentiment; lexicon::hash_sentiment_jockers_rinker
sentimentr_jockers_b	Average sentiment score for sentences in the tweet using a modified version of Jockers (2017) sentiment lookup table used in szuhet R package. Sentiment values ranging between -1 and 1.	Numeric	sentimentr::sentiment;
sentimentr_huliu_b	Average sentiment score for sentences in the tweet using an augmented version of Hu & Liu’s (2004) positive/negative wordlist as sentiment lookup values. Sentiment values ranging between -2 and +1.	Numeric	sentimentr::sentiment; lexicon::hash_sentiment_huliu

Syntactic and Lexical Features Tweets were subject to the following pre-processing steps: remove “RT”, remove irregular whitespace, remove URLs, remove emojis, remove hash symbol, separate camel case hashtags into separate words, remove @ symbol from mentions, offset punctuation, create endmarker punctuation for tweets when absent.			
chars	Count of characters per tweet.	Integer	quanteda_textstats
sents	Count of sentences in the tweet.	Integer	quanteda_textstats
tokens	Count of tokens (words) per tweet.	Integer	quanteda_textstats

Tweet-level metadata

Variable Name	Description	Type	Software Dependency (R package)
Readability Tweets were subject to the following pre-processing steps: remove “RT”, remove irregular whitespace, remove URLs, remove emojis, remove hash symbol, separate camel case hashtags into separate words, remove @ symbol from mentions, offset punctuation, create endmarket punctuation for tweets when absent.			
Flesch.Kincaid	Flesch-Kincaid Readability Score (Flesch and Kincaid 1975)	Numeric	quanteda_textstats::textstat_readability
Flesch	Flesch’s Reading Ease Score (Flesch 1948)	Numeric	quanteda_textstats::textstat_readability
ARI	Automated Readability Index (Senter and Smith 1967)	Numeric	quanteda_textstats::textstat_readability
Lexical Diversity Tweets were subject to the following pre-processing steps: remove “RT”, remove irregular whitespace, remove URLs, remove emojis, remove hash symbol, separate camel case hashtags into separate words, remove @ symbol from mentions, offset punctuation, create endmarket punctuation for tweets when absent.			
C	Herdan's C (Herdan, 1960, as cited in Tweedie & Baayen, 1998; sometimes referred to as LogTTR)	Numeric	quanteda.textstats::textstat_readability
R	Guiraud's Root TTR (Guiraud, 1954, as cited in Tweedie & Baayen, 1998)	Numeric	quanteda.textstats::textstat_readability
TTR	The ordinary Type-Token Ratio	Numeric	quanteda.textstats::textstat_readability
Complex Content: part-of-speech tagging Tweets were subject to the following pre-processing steps: remove “RT”, remove irregular whitespace, remove URLs, remove emojis, remove hash symbol, separate camel case hashtags into separate words, remove @ symbol from mentions, offset punctuation, create endmarket punctuation for tweets when absent.			
pr_noun	proportion of nouns in tweet	Numeric	sophistication:: covars_make_pos
pr_verb	proportion of verbs in tweet	Numeric	sophistication:: covars_make_pos
pr_adjective	proportion of adjectives in tweet	Numeric	sophistication:: covars_make_pos

Deposited with the UK Data Archive

UK Data Service

Search the site... Login

Find data | Deposit data | Learning Hub | Training and events | About | News | Impact | Help | Contact

Home > Data catalogue > Studies > Study

Studies

Series

https://doi.org/10.5255/UKDA-SN-9208-1

Copy study DOI

Understanding Society: Innovation Panel Twitter Study, 2007-2023

Details | Documentation | Resources | Access data

Details

Title:	Understanding Society: Innovation Panel Twitter Study, 2007-2023
Alternative title:	UKHLS
Study number (SN):	9208
Access:	These data are safeguarded
Persistent identifier (DOI):	10.5255/UKDA-SN-9208-1
Series:	Understanding Society
Data creator(s):	University of Essex, Institute for Social and Economic Research

Next steps

Auteregressive LLMs for de-identification

- Use **paraphrasing** models to rewrite tweets;
- Produce multiple versions of the original tweet that reflect key features:
  - Sentiment
  - Lexical diversity
  - Readability
  - Stance
  - Persuasiveness

I’m baffled & exasperated by the numbers doing statistical analysis who imagine that the patients in a clinical trial can be regarded as a random sample from the population defined by the inclusion criteria.
Some heinous crime was committed in educating them.

The sheer number of researchers performing flawed statistical analyses on clinical trials astonishes me. They mistakenly treat the participants as representatives of the entire population based on narrow eligibility criteria. Their education seems woefully inadequate to handle such complex data.

- Twitter post-Musk:
    - access to API;
    - new platform features can alter participants behaviour (exogenous change);
    - Construct measurement stability over time/changes;