Big Data Analytics

Lecture 3/A

Unsupervised classification methods:
the Structural Topic Model

UNIVERSITÄT LUZERN

Reference

- ✓ Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Luca, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, David G. Rand. 2014. Structural Topic Models for Open-Ended Survey Response, *American Journal of Political Science*, 58(4), 1064-1082
- ✓ Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley(2014). STM: R Package for Structural Topic Models, *Journal of Statistical Software*, https://cran.r-project.org/web/packages/stm/vignettes/stmVignette.pdf

Classification methods

Structural Topic Model (STM) innovates on Topic models in two different ways:

First: topic proportions (θ) are allowed to be **correlated**: this is a reasonable assumption given that in documents topics discussed are correlated!

For example, if a party manifesto contains discussion of Topic X (e.g. administrative reform), the probabilities that it will also contain discussion of Topics Y (e.g. curbing public works) and Z (e.g. reducing the number of Lower House members), are not independent of each other, but correlated

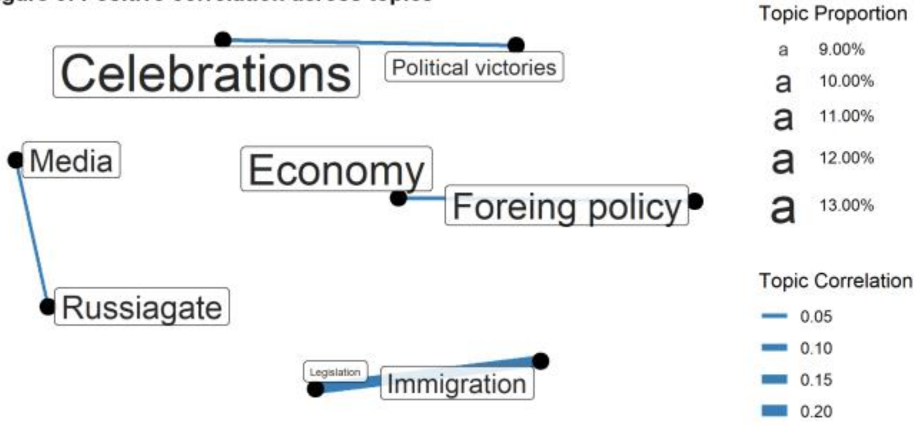
In this sense, STM fits a Correlated Topic Model

Classification methods

Graphical depictions of the (*positive*) correlation between topics provide insight into the organizational structure at the corpus level

In essence, the model identifies when two topics are likely to co-occur (by focusing on positive correlation) within a document

Figure 3: Positive correlation across topics



Source: Results from a Structural Topic Model on @realDonaldTrump Twitter account

Classification methods

Second: as we already discussed, topic models allow the analyst to estimate for each document the proportion of words attributable to each topic, providing a measure of *topic prevalence*

These models also calculate the words most likely to be generated by each topic, which provides a measure of *topical content*

Classification methods

However, in standard LDA, the document collection is assumed to be **unstructured**; that is, each document is assumed to arise from the same data-generating process irrespective of additional information (about the corpus) the analyst might possess

But that shouldn’t be always the case…

Classification methods

Suppose for example that you have reasons to believe that the **age** of a text’s author affects the *probability to discuss about a given topic* rather than some other alternatives

Or that author’s age affects the *probability of using some words when discussing about a given topic* rather than others

Classification methods

Given an unstructured Topic Model…

…you can still try to control if there is there any relationship between the **age** of the author of a document and the emphasis/salience she devotes in her document(s) towards a particular topic (for example, a topic related to migrants)

How?

Only *ex-pos*t, i.e., after you have ended to run the Topic Model and obtained the θs

Classification methods

Such θs have been **however** generated by assuming that they arise from the same data-generating process - irrespective of any document-level variables of interest

On the contrary, STM incorporates the information about **age** (i.e., that belongs to the *structure of texts* in your corpus) directly in the analysis

As a result, it is more efficient as an approach (as far as the document level variables you include in the model makes sense)!

Classification methods

How to reach this goal?

Rather than assuming that **topic prevalence** (i.e., thetas) and **topical content** (i.e., betas) prior distribution **are constant** across all documents (as a topic model does), i.e., you draw θ_i from one single Dirichlet distribution and similarly you draw β_k from one single Dirichlet distribution

…in a STM each document can have (according to the researcher's decision) its own prior distribution over topics, its own prior distribution over betas, or both, according to the values of the document-level variable(s) you decide to include in the fitted topic model

Classification methods

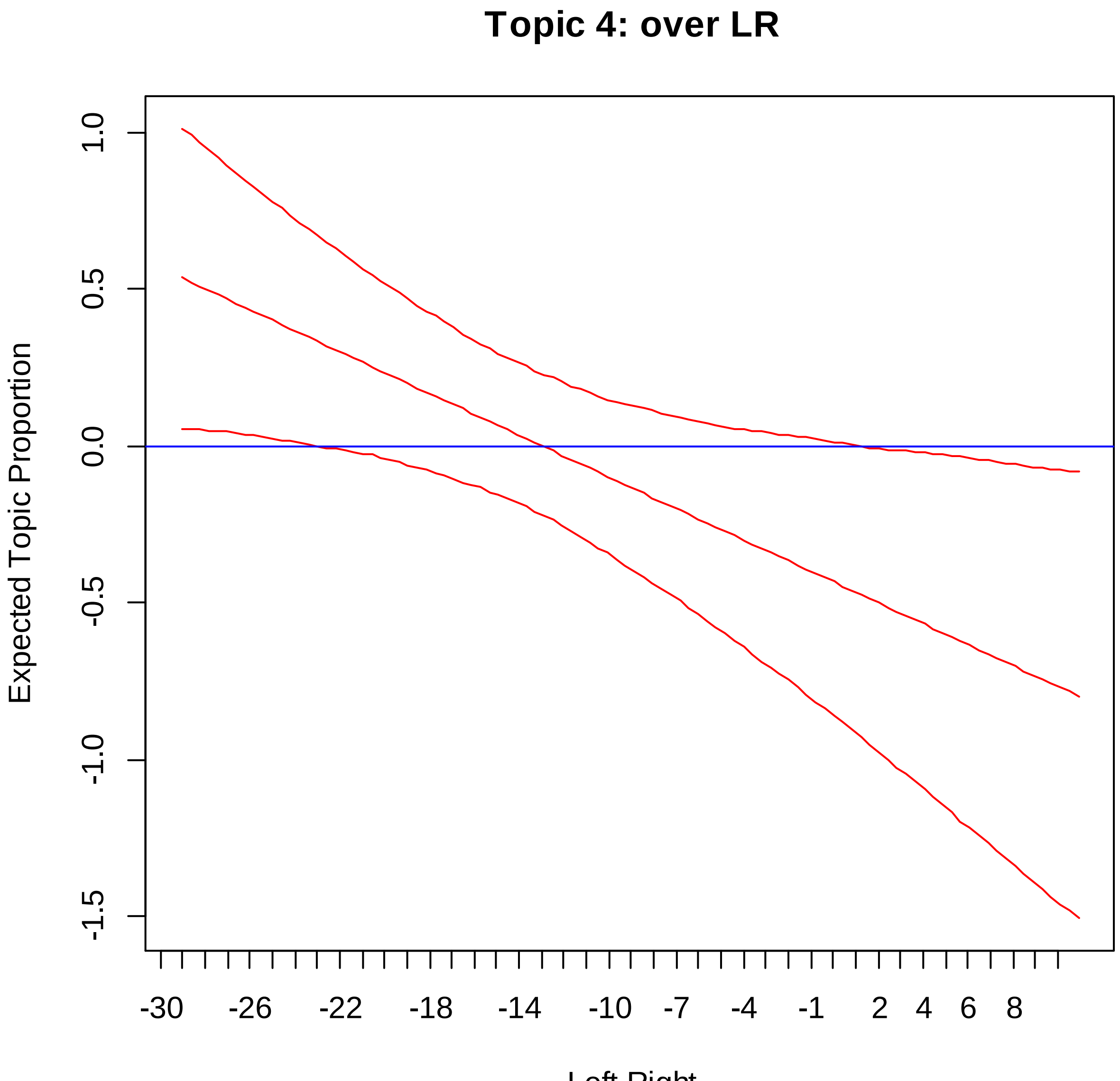
Through that, **topical prevalence** – the *thetas* – can be affected by the covariates you include in the topic model, i.e., we can obtain measures of how our treatment condition systematically affects how often a topic is discussed (prevalence) while **simultaneously estimating** the θs !

Same things can happen for **topical content**, i.e., the *betas* of your fitted topic model

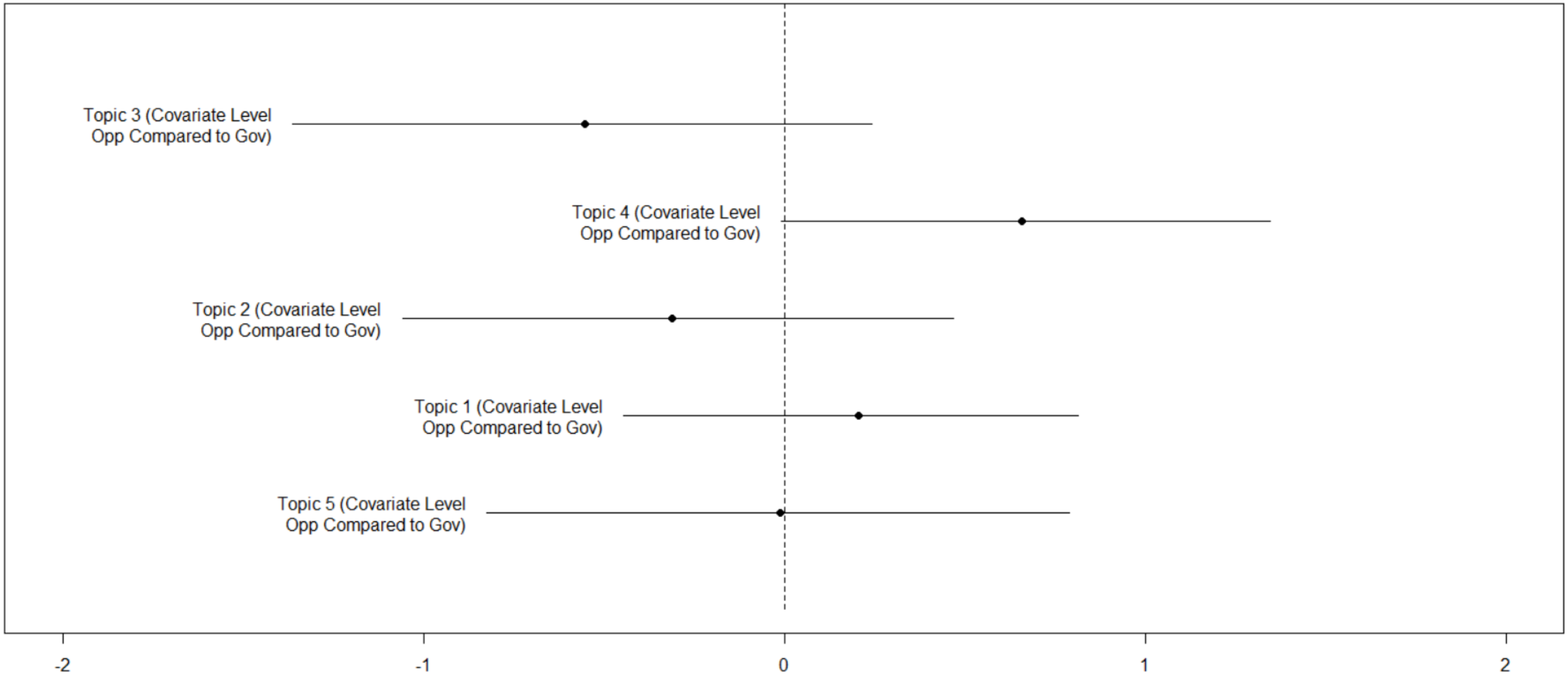
Classification methods

- for example, do documents of **left parties (or opposition politicians)** discuss more about a given topic than documents of right parties (or cabinet politicians)?

Classification methods



Classification methods



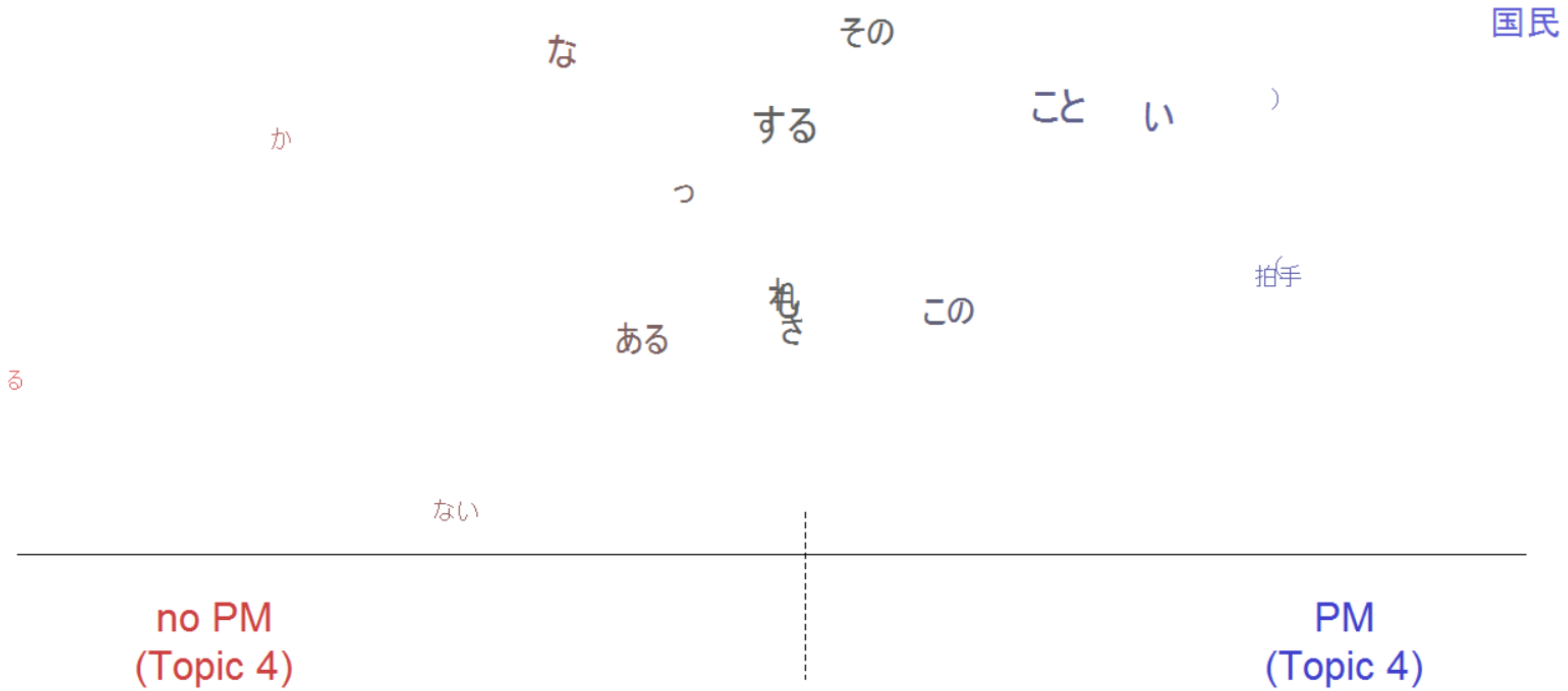
Reported coefficient:
«opposition – government»

Classification methods

Similarly, we can obtain measures of how the **language** used to discuss the same topic (content)

- for example, when **men** politicians discuss about a particular topic do they use the same words than **female** politicians?

Classification methods



Classification methods

In the STM framework, the researcher has therefore the option to choose covariates to incorporate in the model

These covariates inform either the **topic prevalence** or the **topical content** latent variables with observed information about the respondent

Classification methods

The analyst will want to include a covariate in the **topical prevalence portion of the model** when she believes that the observed covariate will affect *how much* the respondent is to discuss a particular topic

The analyst also has the option to include a covariate in the **topical content portion of the mod**el when she believes that the observed covariate will affect *the words which a respondent uses* to discuss a particular topic

These two sets of covariates can overlap, suggesting that the topic proportion and the way the topic is discussed change with particular covariate values

Twitter geolocations

- ✓ Kruspe, Anna et al. (2021). Changes in Twitter geolocations: Insights and suggestions for future usage. *arXiv:2108.12251v1*

Geolocation data

Terminology

“*geolocated*”: tweets containing explicit metadata about a geographic location they were posted from or are referring to

“*geotagging*”: user action that causes this metadata to be attached

Since mid-2019 Twitter’s policy radically changed with respect to geolocation availability

Motivations? Privacy

Geolocation data

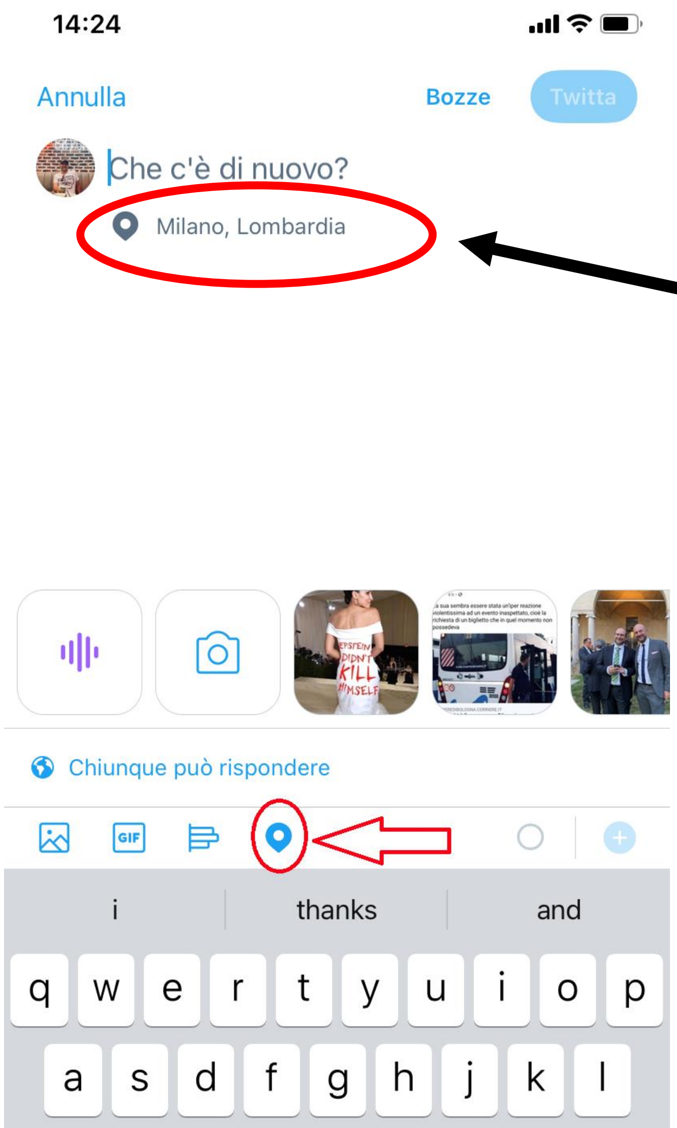
Which geo-data are then available (at least using `rtweet` 0.7.0 – using the 1.0.2. version, the geographical metadata are far less…)?

`place` attributes: the place attribute serves to assign a pre-defined geographic entity to a post

Twitter offers **users** the option to select this entity from a list of those found nearby (within a radius of roughly 200m) when sending a tweet

These entities may be countries, cities, neighborhoods, points of interest (POI), etc.

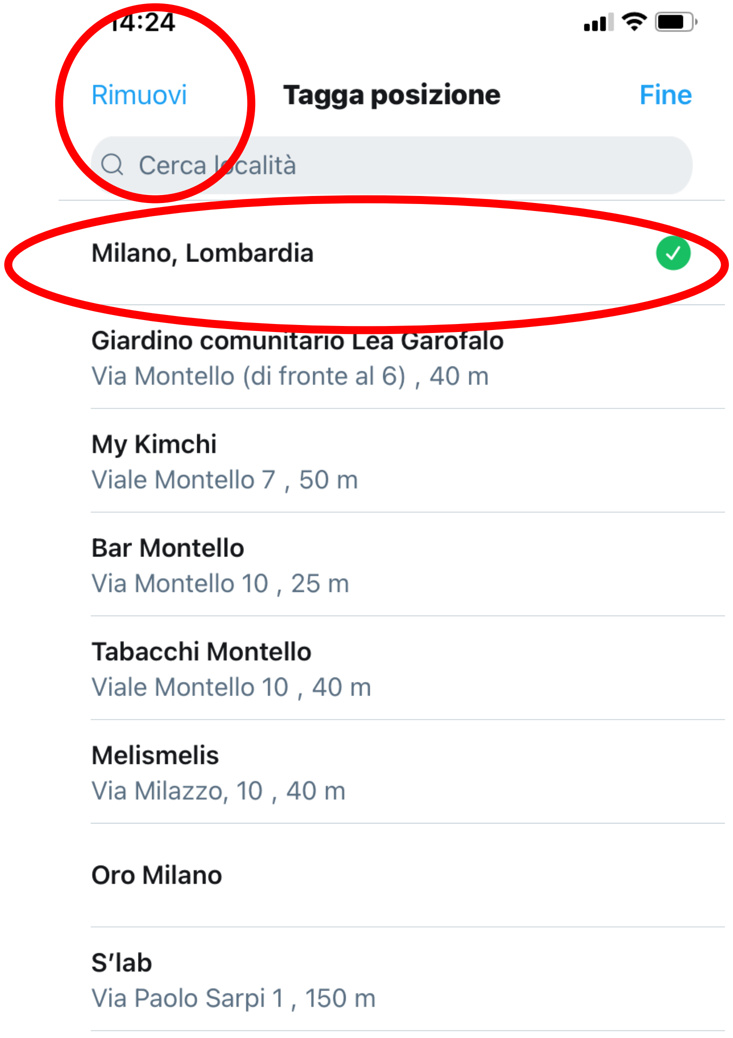
Geolocation data


14:24

Che c'è di nuovo?

Milano, Lombardia

Chiunque può rispondere


14:24

Tagga posizione

Cerca località

Milano, Lombardia ✓

Giardino comunitario Lea Garofalo
Via Montello (di fronte al 6) , 40 m

My Kimchi
Viale Montello 7 , 50 m

Bar Montello
Via Montello 10 , 25 m

Tabacchi Montello
Viale Montello 10 , 40 m

Melismelis
Via Milazzo, 10 , 40 m

Oro Milano

S'lab
Via Paolo Sarpi 1 , 150 m

Geolocation data

`place`’s sub-fields are then automatically filled using information from geolocation services

Among those subfields you have `bbox_coords` that contains a set of coordinates spanning a polygon

Geolocation data

`coords_coords` and `geo_coords` attributes: originally (pre-2019) they were containing the longitude-latitude values of the tweet (provided the users allowed the geotagging option on her smartphone)

Nowadays refer basically to two possibilities: a) a user is employing a very old version of Twitter software on her smartphone; b) the tweet is a cross-post from third-party sources (typically a post on Instagram), and the coordinates reported on Twitter are those picked on Instagram

Geolocation data

However note one important point in the latter case: in this case the coordinates are not anymore representative of the user’s geolocation from which the post was sent, but of some pre-defined location selected by the user, which may be very different from their physical location

In the case of native Twitter posts, these locations (via *`bbox_coords`*) will at least be somewhere close to the GPS location of the device (around 200m radius), whereas in Instagram, they may be anywhere in the world (as selected by the Instagram user)

Geolocation data

In general, the percentage of geolocated tweets out of all tweets is low at 1-2%

How to increase it? We can take advantage of the text either included in the tweets or in users’ profiles (30/40% of profiles contain some form of geolocations) via for example a Named Entity Recognition approach (or by paying the Enterprise API…)