UNIVERSITÀ DEGLI STUDI DI BERGAMO | Dipartimento di Scienze Aziendali, Economiche e Metodi Quantitativi

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

ITSEW2019:

INTERNATIONAL TOTAL SURVEY ERROR WORKSHOP

Social media data for social indicators

Assessing the quality through case studies

Silvia Biffignandi [1]
Annamaria Bianchi [1]
Camilla Salvatore [2]

[1]University of Bergamo
[2]University of Milan-Bicocca

LOCATION

University of Bergamo

DATE

11th June 2019

INDEX

Social media data for social indicators

Assessing the quality through case studies

Introduction

Social Media: Opportunities and Challenges

OPPORTUNITIES

- Answer to new questions;
- Provide an insight on people's preferences, behaviors and political movements;
- Provide complementary, faster and specific information about a topic;
- Help to assess unmeasured or partially measured socioeconomic phenomena.

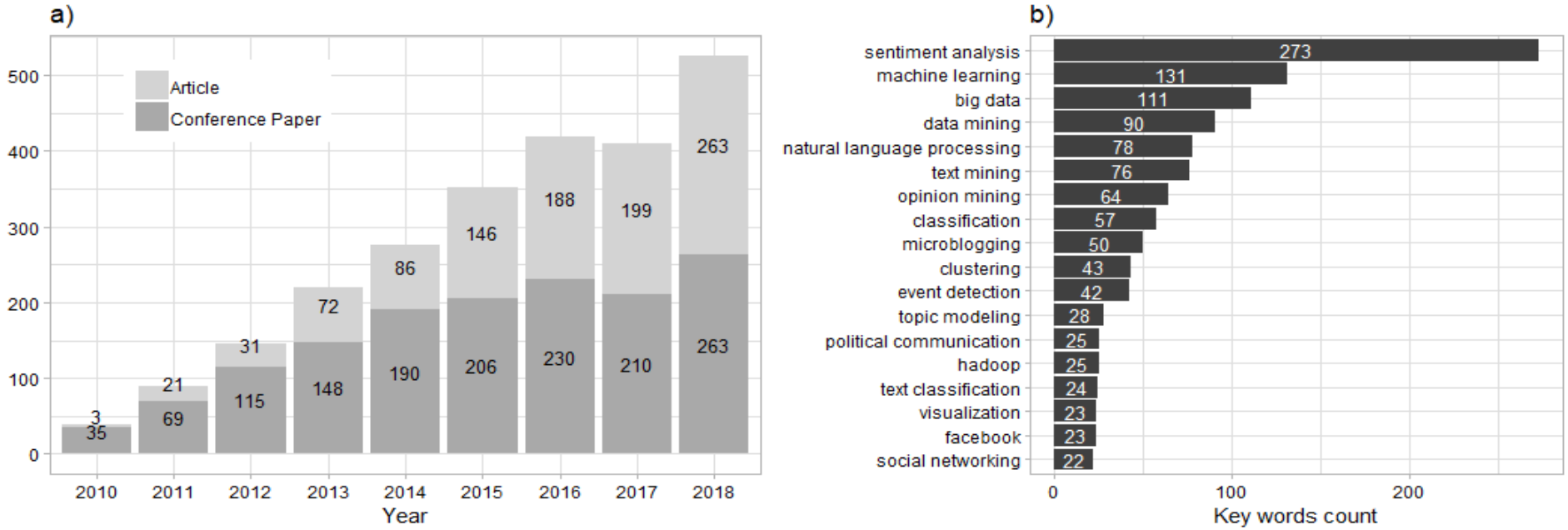
CHALLENGES

- Data, Process and Management challenges;
- Privacy;
- Quality → low quality data can lead to wrong conclusion.

Introduction

Social Media data for social indicators: examples

- Papers that analyse Twitter Data:



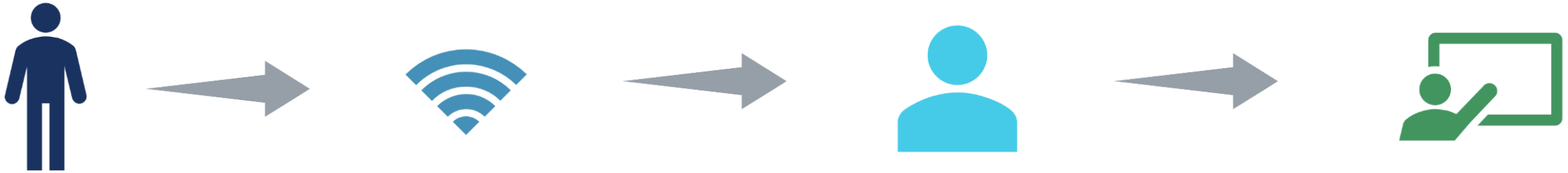
- Research Area: Economics; Politics; Well-Being;

- In Official statistics:
  - VM (security) survey + STI (social tension indicator based on social media);
  - CCI (consumer confidence index) survey + SMS (social media sentiment).

Sources: Author's own elaboration

Statistical Considerations

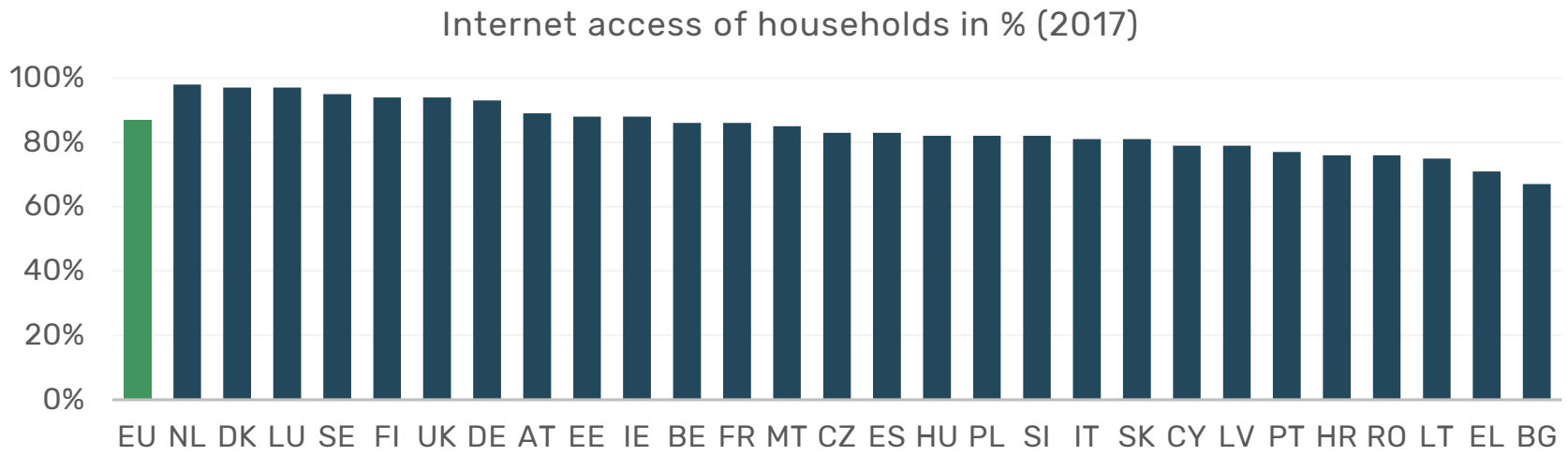
Self-selection process



1: Have an Internet connection

2: Decision to join a specific social-media platform

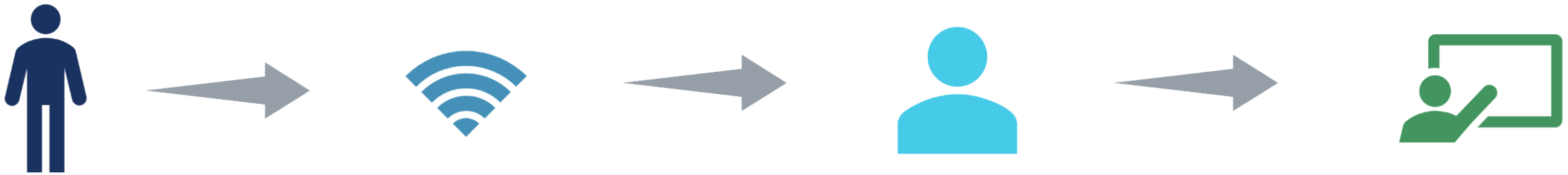
3: Decision to publish contents



Sources: Eurostat.

Statistical Considerations

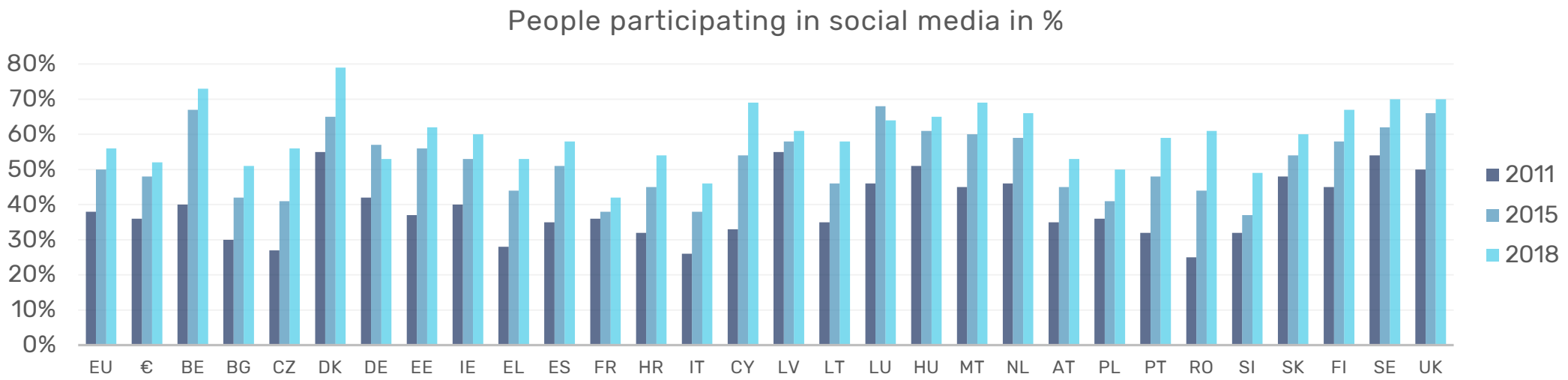
Self-selection process



1: Have an Internet connection

2: Decision to join a specific social-media platform

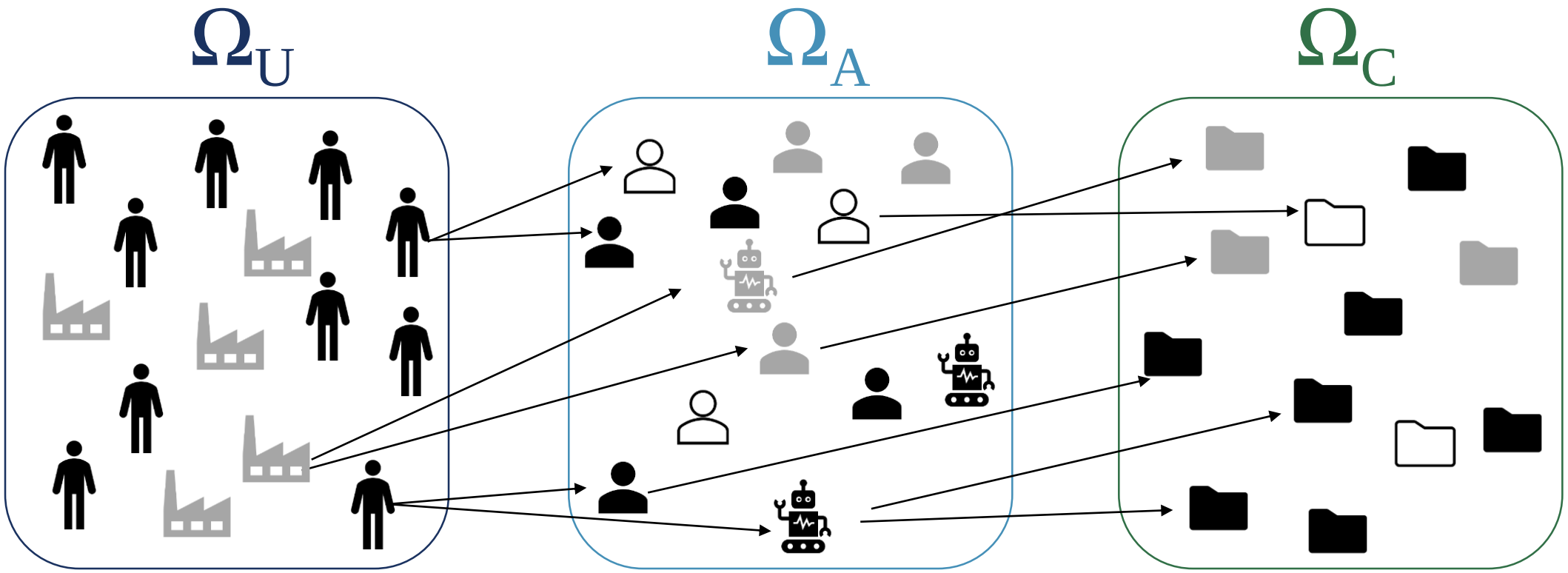
3: Decision to publish contents



Sources: Eurostat.

Statistical Considerations

Populations in social media



Sources: Author's own elaboration

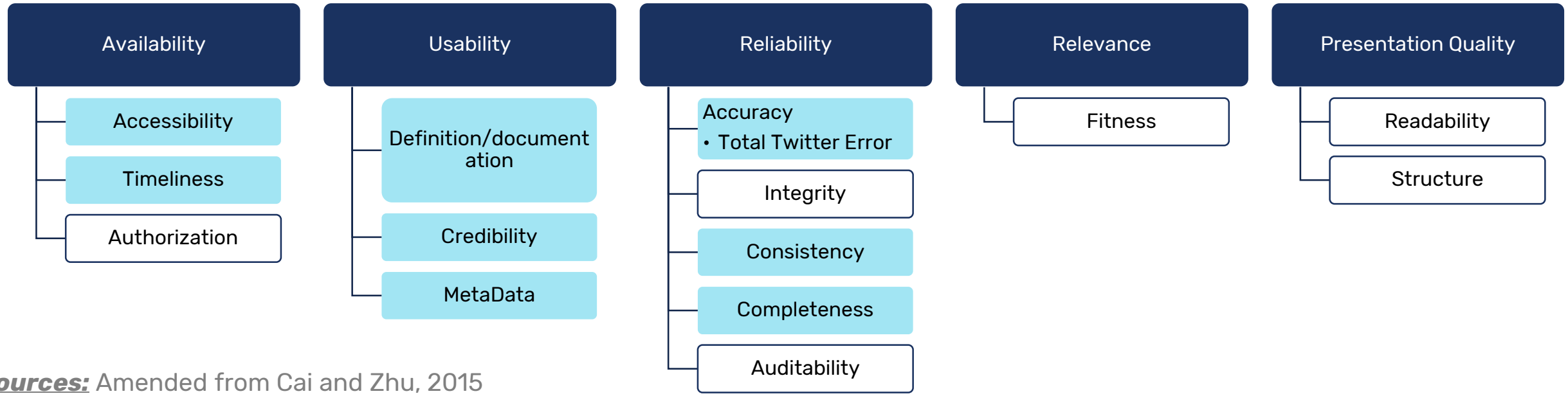
Statistical Considerations

Summary

- We do not observe directly the characteristics of Ω_U.
- Ω_A includes also malicious accounts.
- The link between the statistical phenomena of interest and the data collected is indirect.
- Nature of the data: Twitter message ≠ survey answer.
- Other considerations related to Big Data in general:
  - Data deluge;
  - Methodological issues
  - Volatility
  - Consent to the use of data;
  - Privacy and other issues.

Total Quality Twitter Framework

- Quality is a multidimensional concept;
- Any Survey Quality framework contains at least nine dimensions: accuracy (TSE), credibility, comparability; usability/interpretability, relevance, accessibility, timeliness/punctuality, completeness and coherence;
- These dimensions are general enough to be adapted also to big data with some adjustment;
- Cai and Zhu (2015) proposed a hierarchical definition of quality and its indicators considering similar dimensions:



Sources: Amended from Cai and Zhu, 2015

Total Quality Twitter Framework

Availability

It refers to the ease and the conditions under which the data and the related information can be obtained. We can consider two sub-dimensions, the *accessibility* and the *timeliness*.

Accessibility

Twitter data are accessible with few restrictions. Twitter provides several APIs to access data according to the different use cases;

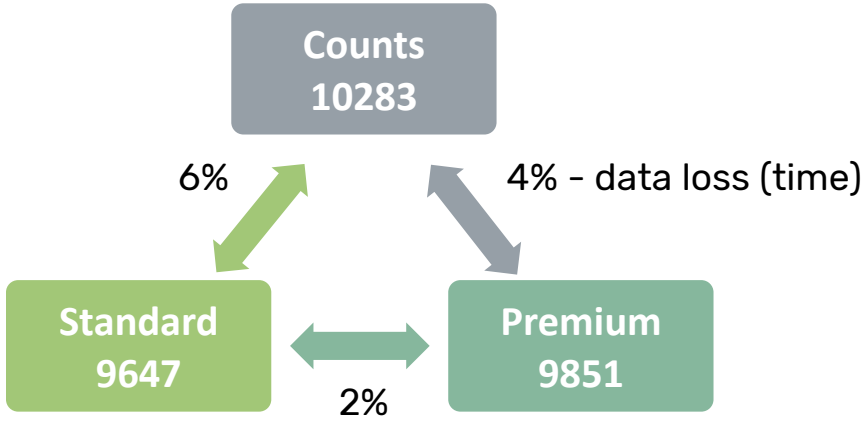
Access Type	Description	Free/Paid	Completeness
Search API: old tweets	Standard: 7 days	Free	NO
	Premium: 30 days or Full-Archive	Free (Sandbox) or Paid	YES
	Enterprise:30 days or Full-Archive	Paid	YES
Filter real-time tweets: Streaming API	Standard: statuses/filter	Free	NO
	Enterprise: PowerTrack API (Firehose)	Paid	YES
Sample: all public tweets	Standard: statuses/sample	Free	NO
	Enterprise: Decahose	Paid	10% random sample
Batch: Historical tweet	Enterprise: Historical Power Track API	Paid	YES

Total Quality Twitter Framework

Accessibility

The type of access affects the analysis results:

- Real-time Streaming (free) vs Firehose (paid) APIs (Morstatter et al.,2013):
  - They found that that the results of using the Streaming API depend strongly on the coverage and the type of analysis that the researcher wishes to perform;
  - They used Firehose data to get additional samples to better understand the results from the Streaming API and they found that the Streaming API performs worse than randomly sampled data, especially at low coverage.
- Standard (free) vs Premium (paid) Search APIs:
  - We retrieved tweets with query "#BrexitShambles" the 16th of January relative to the 15th January. The results of counts and data endpoints are:



Total Quality Twitter Framework

Timeliness

There are different time-dimensions to consider:

- The first one is the time between the data request and the data delivery which varies according to the access type.
- Tweets of non-protected accounts are available 30 seconds after the publication but they are not stored forever.
- An indicator of the data loss due to the time lag between the data generation and the retrieval can be the difference between the estimates obtained through the counts endpoint and the quantity of data retrieved through the data endpoint.

Total Quality Twitter Framework

Usability

It refers to the ease with which data can be used.

- Twitter is committed in providing documentation, in enriching and regularly updating Metadata.
- Of course, with upgraded access the usability is improved since premium search operator and extra support services are provided and Metadata are enriched.
- Data are provided in JSON format (JavaScript Object Notation) – semi structured form.

Total Quality Twitter Framework

Reliability

The key issue is whether we can trust data. We analyse the following sub-dimensions: accuracy, consistency and completeness.

Accuracy

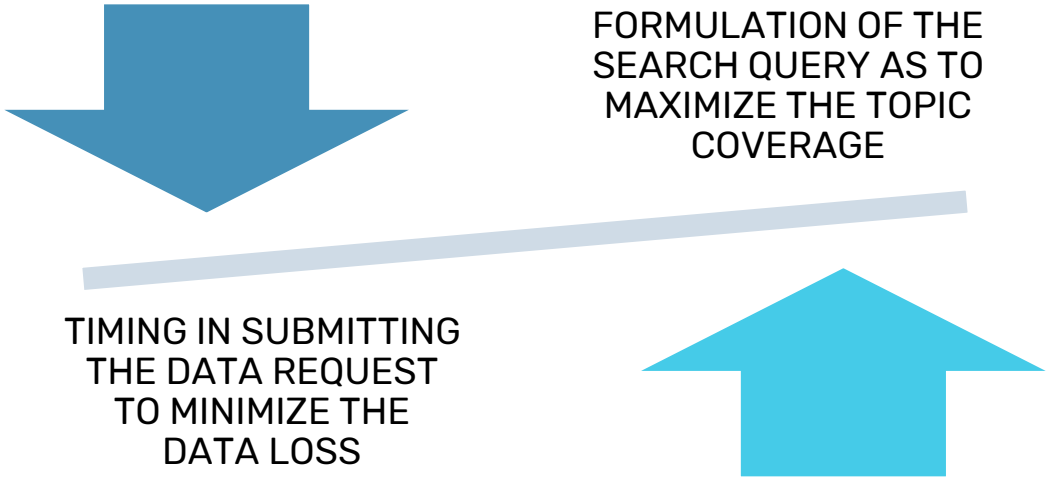
It is linked to the concept of “errors”

- *Textual errors:*
  - Typos: Misspelled words cannot be recognized and elaborated by algorithms and this affects the results of the analysis.
  - We can consider the percentage of misspelled words as an indicator of the accuracy of tweets at the origin.
  - Also abbreviations and *slang* are difficult to evaluate by machines. In this context, text mining techniques represent a fundamental tool to identify and correct errors before the implementation of any analysis.
- *Total Twitter Error Framework (TTE).* Hsieh and Murphy (2017) adapted the TSE paradigm to Twitter and developed the Total Twitter Error framework. They identify three exhaustive and mutually exclusive sources of errors:
  - query error
  - coverage error
  - interpretation error.

Total Quality Twitter Framework

Total Twitter Error : Query Error

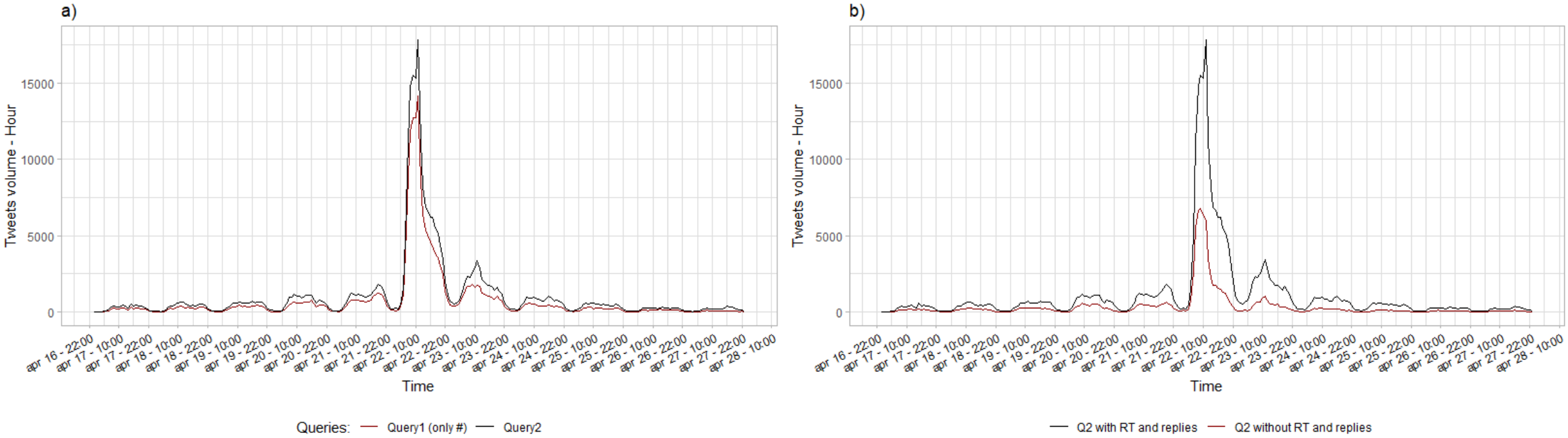
- Researchers formulate the query as to maximize the topic coverage.
- Sources of error:
  - Misspecification of the search string.
  - Inclusion or exclusion of retweets and replies.
  - To other search constraints (ex. Geolocalization).
- TRADE-OFF between:



Total Quality Twitter Framework

Total Twitter Error : Query Error

- Example:
  - Query 1: "`#londonmarathon OR #londonmarathon18 OR #londonmarathon2018`"
  - Query 2: "`#londonmarathonOR #londonmarathon18 OR #londonmarathon2018 OR (london +marathon)`"

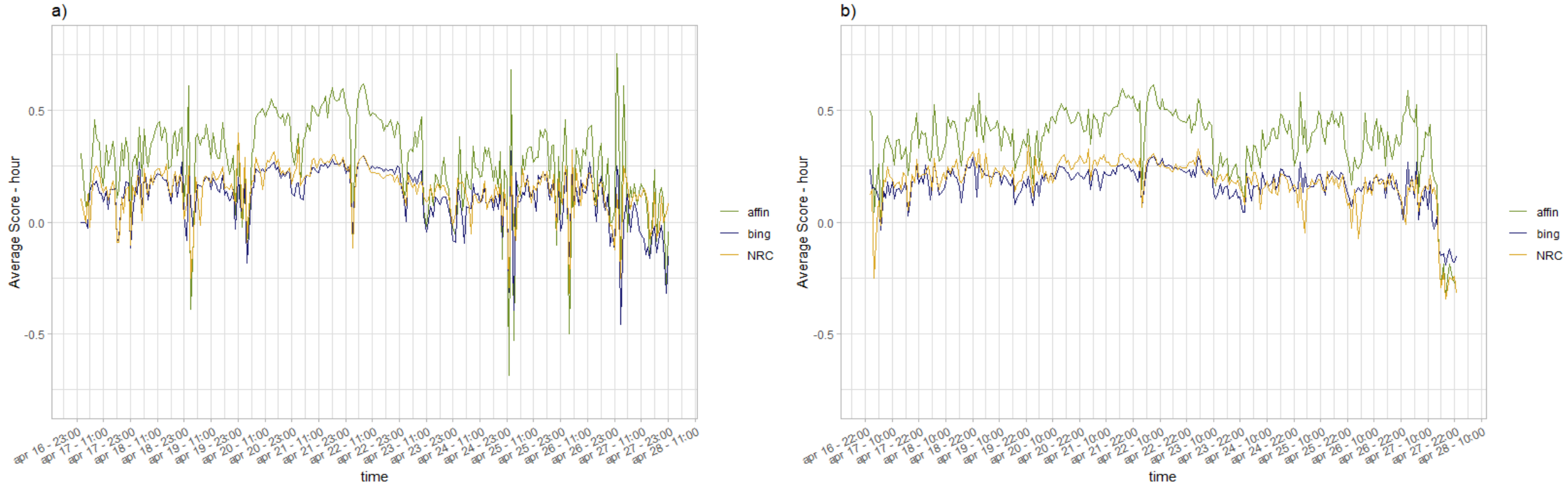


Sources: Author's own elaboration

Total Quality Twitter Framework

Total Twitter Error : Query Error

- How the query formulation affects the analysis:



Sources: Author's own elaboration

Total Quality Twitter Framework

Total Twitter Error : Interpretation Error

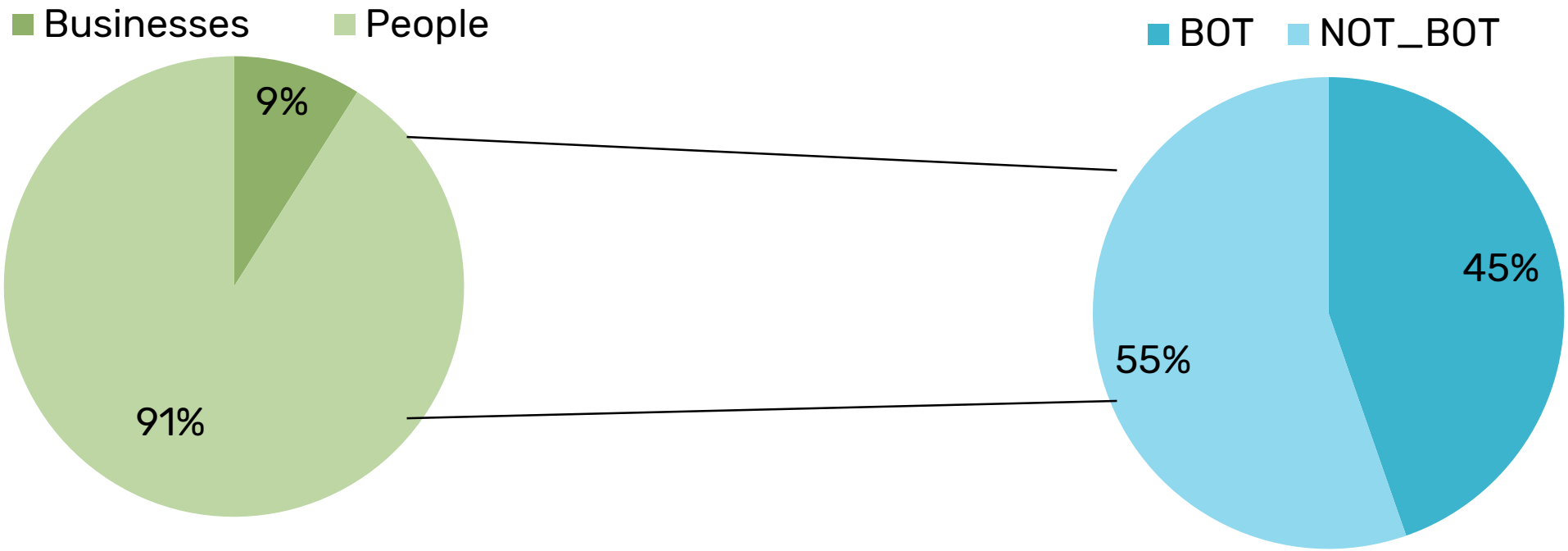
- It is due to the process of extracting insight from the text or to the process of inferring users missing characteristics.
- Kiefer suggests that for automatically sentiment classifier an indicator of the similarity between the input data and the training data can be measured using the Cosine Similarity or the Greedy String Tiling (Kiefer, 2016).
- For dictionary-based approaches, we should consider the characteristics of the lexicons:
  - Lexicons that accounts for the “shade” of the opinion words can give more accurate results;
  - It useful to evaluate the ratio between positive and negative words for each lexicon to obtain an indicator of the negative or positive propensity of the lexicon;
  - Context-specific lexicons should be preferred.

Total Quality Twitter Framework

Total Twitter Error : Coverage Error

Sources of error:

- Under-coverage: the observed sample is not representative of the target population.
- Over-coverage: the observed sample is composed by accounts that are associated to people, businesses as well as BOT.



Sources: Author's own elaboration

Total Quality Twitter Framework

Consistency

It refers whether the data remain consistent and verifiable over time. To show the data loss over time, we decided to investigate whether our London Marathon's tweets are still available.

Day	No. Tweets (count endpoint)	LM tweets Apr 2018	Available Apr. 2019	Loss	% of data loss
April 17th	3,803	3,731	2,342	1,389	37.22%
April 18th	5,055	4,814	2,940	1,874	38.92%
April 19th	6,236	6,153	3,782	2,371	38.53%
April 20th	9,833	9,645	5,999	3,646	37.80%
April 21st	14,968	14,854	9,068	5,786	38.95%
April 22nd	116,185	115,494	72,580	42,914	37.15%
April 23rd	24,954	24,176	14,777	9,399	38.87%
April 24th	8,257	7,870	4,845	3,025	38.43%
April 25th	4,443	4,428	2,494	1,934	43.67%
April 26th	2,309	2,307	1,438	869	37.66%
Total	**196,043**	**193,457**	**120,265**	**73,207**	**38%**

Sources: Author's own elaboration

Total Quality Twitter Framework

Completeness

- The completeness of data and Metadata depends on the data access.
- In the Standard Search API data returned are based on the relevance and not on the completeness. Completeness is assured with the Premium and Enterprise access.
- An indicator of the completeness can be the percentage of missing values.

Conclusions

- Big Data does not mean Big Information → “imperfect, yet timely, indicator of phenomena in society” (Braaksma and Zeelenberg, 2015).

- To trust data we must assess the Quality and reduce the Error.

- Our study presents same experimental analysis to build up quality indicators on Twitter data and a framework for the Total Twitter error.

- It is fundamental to use a mixed method based on quantitative as well as on qualitative analysis to built quality and errors indicators.

Questions

- Which other dimensions of quality could be considered?

- Do you have examples of similar analyses on Twitter data quality? What was the conclusion in such cases?

References

- BRAAKSMA, B., ZEELENBERG, K. (2015). "Re-make/Re-model": Should big data change the modelling paradigm in official statistics?. *Statistical Journal of the IAOS*, *31*(2), 193-202.
- CAI, L., & ZHU, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 14.
- HSIEH, Y. P., MURPHY, J. (2017). Total Twitter Error. *Total Survey Error in Practice*, 23-46.
- KIEFER, C. (2016). Assessing the Quality of Unstructured Data: An Initial Overview. In *LWDA* (pp. 62-73).
- MORSTATTER, F., PFEFFER, J., LIU, H., & CARLEY, K. M. (2013, July). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *ICWSM*.
- TANUPABRUNGSUN, S., & HEMSLEY, J. (2018). Studying celebrity practices on Twitter using a framework for measuring media richness. *Social Media+ Society*, 4(1), 2056305118763365.

UNIVERSITÀ DEGLI STUDI DI BERGAMO | Dipartimento di Scienze Aziendali, Economiche e Metodi Quantitativi

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Thank you for your attention!

Social media data for social indicators

Assessing the quality through case studies

Questions?

Contacts:

c.salvatore4@campus.unimib.it

silvia.biffignandi@unibg.it

annamaria.bianchi@unibg.it