Road to Representivity

Addressing quality in social media research

Josh Keith, Ipsos MORI

SRA Annual Conference

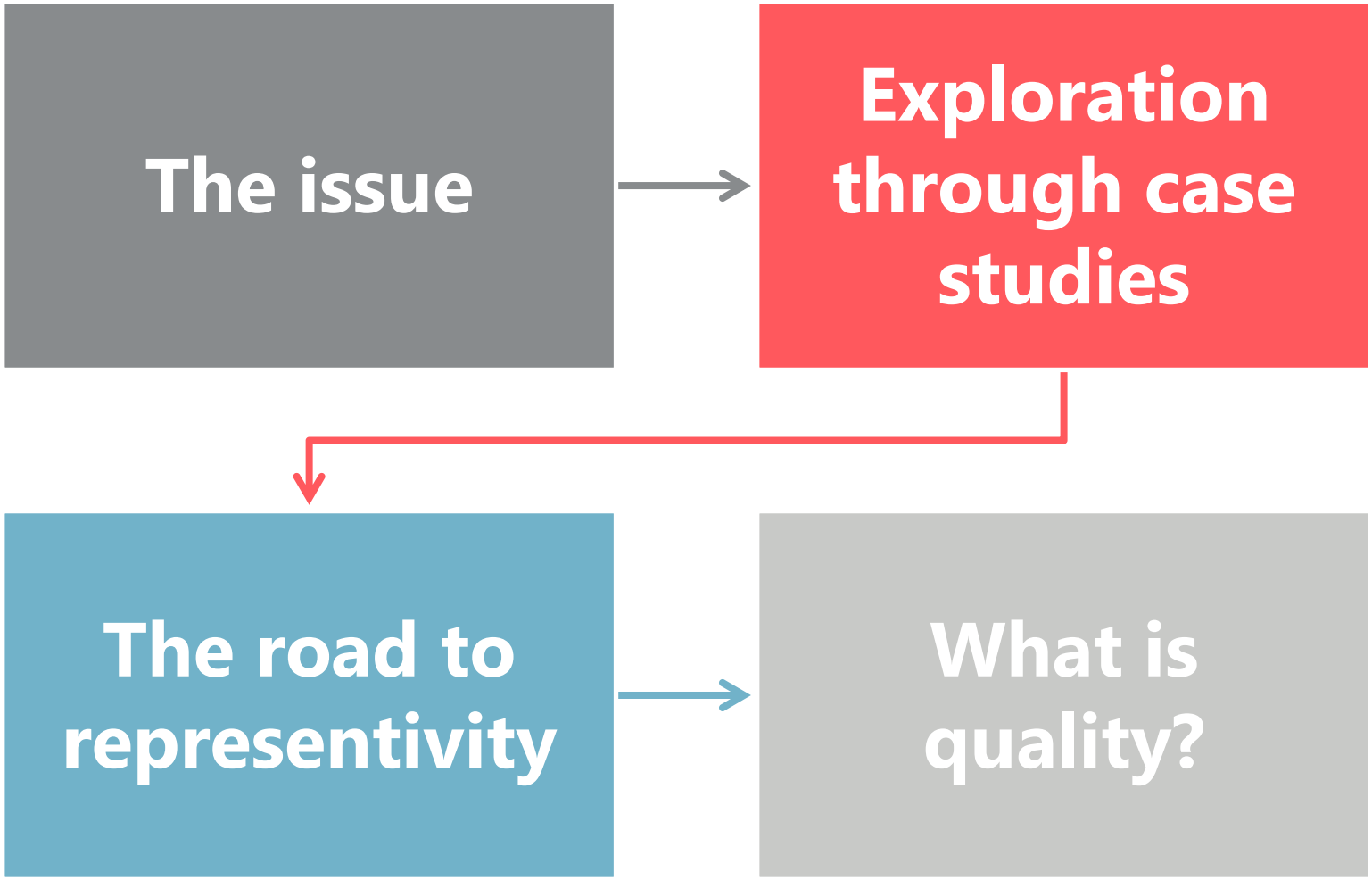
14th December 2015

Innovate UK
Technology Strategy Board

Ipsos MORI

#SRAevents

Overview


The issue
Exploration through case studies
The road to representivity
What is quality?

The issue

Opportunities

- The rise of social media usage
- Generating access to a huge volume of data – 6,000 Tweets a second, 500 million a day
- Improvements in tech allows for tracking/trending and analysis of metadata

Challenges

- Reflecting principles of social science
- Understanding large, complex datasets
- The 'two faces of representivity' – population AND conversation
- Challenging the technology

Exploration through case studies

- Collection of **online** and **offline** data across 3 themes.
- Interrogate **differences** between **online** and **offline.**
- Exploration of differences and how they **vary** across **themes**.

Themes

Politics: Miliband v Cameron in run-up to GE2015.

Brands: Popularity in online and offline discussions.

Issues: Importance of issues in online and offline conversations.

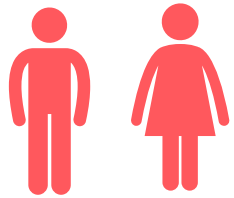
Exploration through case studies

Case study 1: Politics

- Twitter much **harsher** than offline research.
- Online and offline direction of travel **different**.

Why?

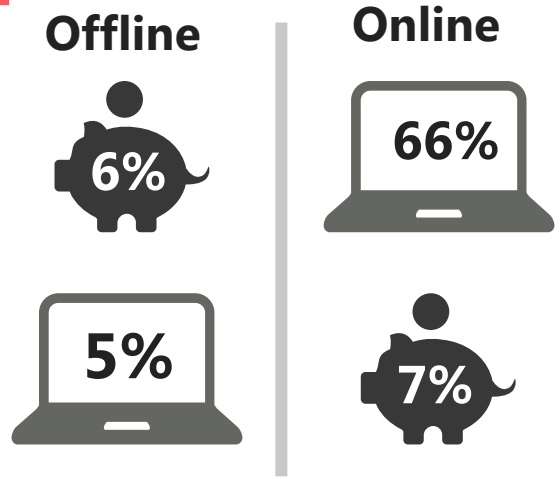
- ***Power-users?*** 33% of Tweets from 1% of users.
- ***Institutions?*** 10% of Tweets from institutional accounts.
- ***Location?*** London over-represented – 24% of Tweets.
- ***Socio-demographics?*** Twitter exaggerates the gender divide in favourability.



Exploration through case studies

Case study 2: Brands

- Tech companies dominate online conversation.
- Again, very little relationship online vs. offline.



Why?

- ***Power-users?*** 29% of Tweets from 1% of users.
- ***Institutions?*** 20% of Tweets from institutional accounts – but little impact.
- ***Location?*** Significant regional variation, and over-representation.

Exploration through case studies

Case study 3: Issues

- Immigration and the NHS dominate the **offline** conversation.
- **BUT** education and crime top the **online** discussion.

Why?

- **Data collection?** Large volumes of irrelevant Tweets.
- ***Location?*** Significant regional variation, and over-representation.

The road to representivity

Data collection

Challenges

- Not comprehensive
- Non-random
- Non-relevant

Solutions

- Over-sample
- Relevancy classification
- Remove irrelevant data

Prolific accounts

Challenges

- Small number of vocal accounts
- Not representative of Twitter users

Solutions

- Avoid focus on ‘n’ Tweets
- Analysis ‘n’ unique users
- ‘Tweets per user’ to assess issue

Bots

Challenges

- Automated ‘bots’
- Coordinated content sharing

Solutions

- Exclude accounts with very low follower numbers
- Work in progress

The road to representivity

Institutions	Location	Socio-demographics	Comparability
Challenges			
▪ Not just individuals ▪ Corporations and institutions	▪ Unrepresentative of a given country or regions within	▪ Nature of users ▪ Over-representation of subsets of society	▪ What is being measured?
Solutions			
▪ Algorithms to identify institutions – 87% accuracy ▪ Remove from data or account for in analysis	▪ Geo-tags for some Tweets ▪ Algorithmic analysis of metadata ▪ 80-90% accurate when possible	▪ Algorithms to assign gender to non-institutional accounts	▪ Contextualise Twitter research

What is *quality* in social media research?

1. Social media *analytics* → *science*.

2. Common methodological approach.

3. Clarity of approach.

4. Reflective of limitations.

5. A companion to offline research.

Thank you

For more information

Josh Keith

Research Manager, Social Research Institute

☎ 020 7347 3151

077 6618 4888

Josh.Keith@Ipsos.com

Wisdom of the crowd

US University of Sussex · DEMOS · CASM Consulting · Ipsos MORI

Innovate UK
Technology Strategy Board

Ipsos MORI

www.ipsos-mori.com/wisdomofthecrowd