September 2021

Asking the right questions

Assessing the value and futures of social media analysis

A report by Ipsos MORI, King's Centre for Strategic Communications, and the Social Intelligence Lab

Steven Ginnis
Michael Clemence
Faith Jones
Emily Mason
Tara Beard-Knowland
Thomas Colley
Jillian Ney

Ipsos MORI

Contents

1 Executive summary

1.1 Introduction

The Defence Science and Technology Laboratory (DSTL) commissioned Ipsos MORI, the King's Centre for Strategic Communications (part of King's College London) and the Social Intelligence Lab to conduct research assessing the utility of social media data for the purpose of research. Online sources provide a vast body of data, which can help leverage insight into public beliefs, attitudes and behaviours. However, there is still much debate over their reliability and utility compared to other forms of data and research.

The overarching objective of the project was to ask: "How robust, representative and reliable is content posted online in providing insight into the behaviours, motivations and attitudes of wider populations?"

To answer this question, the project combined four main strands of work:

1. A literature review to assess the latest research on the representativeness and usefulness of social media data. The review was conducted between January-February 2020.
2. 30 in-depth interviews with subject matter experts from across a range of disciplines including academia, politics, market research and defence and strategic communications. Interviews were conducted between March-May 2020.
3. A live case study to capture public opinion and experience during the COVID-19 pandemic in the UK. The case study compared the quality and value of findings from survey and social media research. Two waves of fieldwork were conducted in April 2020.
4. Twenty-three experts from the interviews participated in a Delphi panel survey to develop future scenarios for the social media sector in 2025. This survey explored their attitudes towards the key drivers identified though analysis of the interviews. The survey took place in May-June 2020.

1.2 Defining and assessing high quality

For the purpose of this project, social media data is defined in the broadest possible terms as the passive (e.g. location data) and active (e.g. a comment or 'like') data generated by users of social media platforms. The following working definitions were used in order to answer the primary research question:

- **Representativeness**: Whether the sample of data accurately reflects the broader phenomenon or population that is being studied.
- **Robustness**: How well a test performs when variables, assumptions or the environment are altered.
- **Reliability**: Whether repeating the same method would lead to the same result.

The Literature Review found no consensus on how representative, reliable or robust social media data can be. There appear to be three main, and conflicting, bodies of literature:

1. Data Science research supporting social media data's representativeness and predictive ability. These papers claim that social media data can be sufficiently representative to predict various behaviours, including elections, purchasing habits, epidemics and protest participation.

2. Critical papers highlighting a large range of possible errors that impede representativeness and prediction using social media data. These papers find that social media data is highly localised,

prone to bias, quickly outdated, neither reliable nor robust, and unrepresentative of broader online or offline populations.

3. Commercial literature claiming that social media data can be highly representative and predictive – particularly when analysed by the given companies' analytics software. These claims are hard to assess because companies lack commercial incentives to reveal their methods to competitors. Their software is typically proprietary and 'black-boxed'.

Each body of literature is a direct product of the underlying values and imperatives of the communities in which they have been produced. The critical perspective reflects the tension between the qualitative and quantitative research. Although the critiques are valid, many could easily apply to a wide range of different research methods and types of data.

It is therefore necessary to start any assessment of the value of social data by accepting that there is no one perfect data source for research. In reality, all analysts have is a range of different, albeit imperfect, sources of data, each providing them with a slightly different way of understanding the world. However, to date, there has been little public discussion of a robust framework to help assess the quality of, and potential bias within, social media data used for research.

1.3 Overall conclusion and key considerations

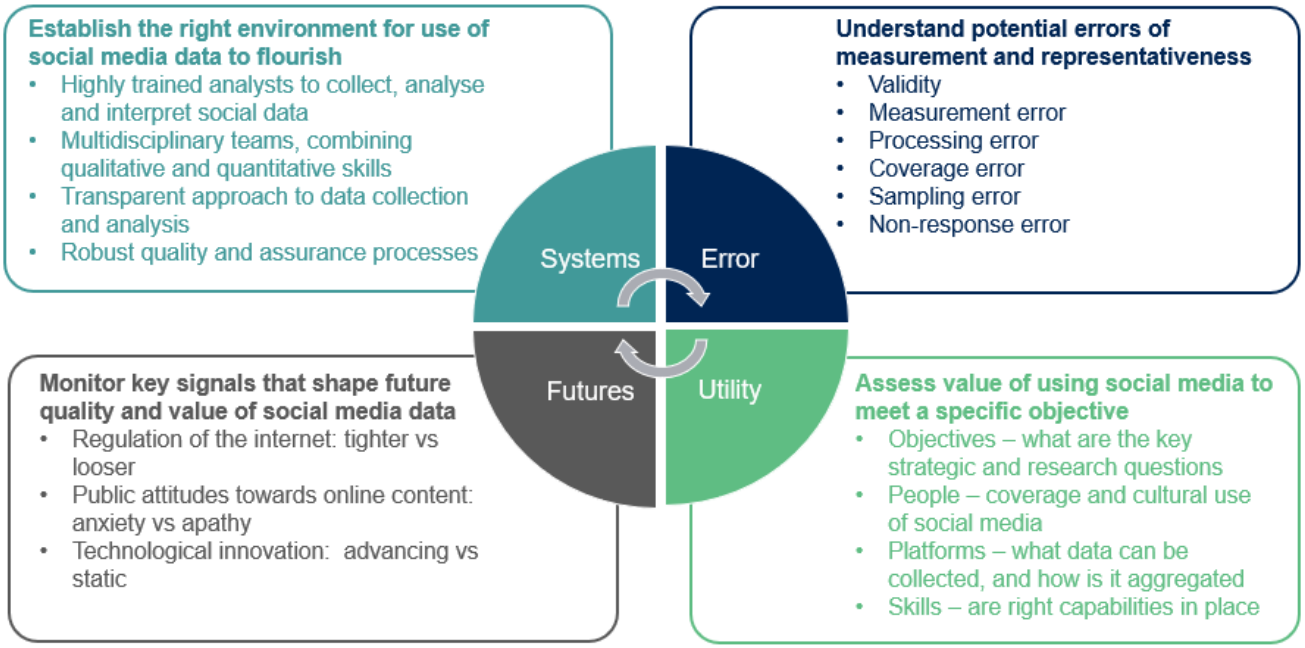
Based on the evidence from the literature review, expert interviews, and primary case study research, **we conclude that any form of social media data is unlikely to be truly representative of wider populations of interest; and that the collection and analysis of social media data is not yet as reliable and robust as other more established methods**. **However, we also conclude that this does not discount social media data from being a valuable tool for research**.

There is no universal truth as to the representativeness, reliability, or robustness of all forms of social media research. All three variables are context specific, and are dependent on the objectives of the research, the population of interest, the platforms analysed, and perhaps most importantly the skills of the analyst conducting the research.

The skills and the environment in which analysts work is a critical multiplier of the quality, and ultimate value to social media data for research. As demonstrated throughout this project, analysts make numerous decisions that shape aspects of data collection and analysis; these decisions can enhance or degrade the representativeness, robustness and reliability of results. Yet despite analysts' best efforts, there are inherent measurement biases within the construct of social media. Some aspects of quality remain outside researchers' control (for example use of platforms, and access to data defined by platforms). An analyst's ability to understand these limitations is therefore even more critical in assessing the value of social media data in any given context.

Drawing on the findings from this project, Figure 1.1 provides an overview of the key considerations for the use of social media data in research and policy making. There are many moving parts. The use of social media, access to social media, cultural-political context, and technological advancement will all continue to shape the content and value of social media data. With this in mind, the framework should be reviewed regularly.

Figure 1.1: Key considerations for the use of social media data in research and policy making



1.4 Understanding errors of measurement and representativeness

In judging the quality of social media analysis, it is important to first conceptualise potential errors within the research process. Inspired by the Total Survey Error Framework, we have identified six key sources of error that are relevant to social media analysis; these provide a useful way of thinking about the biases that may affect social media data. To date, there has been considerable focus on errors of representativeness within social media analysis, but less so on measurement. Both deserve equal scrutiny:

1. **Validity:** are the inferences made within the data valid (e.g. what is signified by a 're-tweet', or being 'friends' on social media)?
2. **Measurement error:** what is the deviation between true public opinion or behaviour, and those shared on social media (e.g. are some issues more taboo, or polarised, or open to social desirability bias[1])?
3. **Processing error:** how accurate are methods used to categorise, analyse and interpret data (e.g. automated sentiment analysis or appended demographics[2])?
4. **Coverage error**: to what extent does the target population match the population accessed and sampled through social media (e.g. restrictions to public data made available by platforms, or risk of editorial censorship by governments and platforms)?
5. **Sampling error:** to what extent does sampling of data accurately reflect the conversation of interest (e.g. through the development of a social media search query, or the variable terms of access to data)?

[1] For example, social media users expressing views that they feel others want to hear, or feeling compelled to present a certain image of themselves online that is deemed more 'acceptable' to society in order attract positive attention.
[2] For example classifying a post as being 'positive' or 'negative' or estimating age based on the contents of a post.

6. **Non-response error:** which groups cannot be contacted or persuaded to provide data (e.g. which groups are less likely to create accounts, and post; extent to which data is further skewed by prolific users)?

These six sources of error often interlink within assessments of representativeness, reliability and robustness.

Representativeness

In traditional market research, representativeness – or more accurately, drawing a representative sample – is important because it allows researchers to make observations and inferences about a broader population. The expert group felt that social media data is inherently different to survey research data and therefore that the traditional measures of quality, including representativeness, have limited read-across. Our research with experts found a broad consensus that social media data was not 'representative' in the traditional sense of the term (i.e. did not represent all of society); some argued that the use of representativeness as a benchmark for judging the quality or utility of social media was a false problem.

Attempts to derive representativeness from social media samples face several challenges. Firstly, the data can only reflect social media users and what they are willing and able to share online in the context of the design and structure of a given platform. Additionally, in a volatile online environment, there are many commercial and political actors attempting to influence public opinion and behaviour.

Both the literature review and the expert group suggested that the complex nature of social media data required a far more nuanced and context specific assessment of representativeness. It is difficult to make any absolute claims about whether social media is representative of online, or offline, populations. Analysts instead need to consider four factors to help assess the representativeness of any given social media dataset:

1. Platforms used – including data accessibility, platform design, the profile of users and culture of use.
2. Topic of conversation – whether the topic is present on social media, and/or at risk of social desirability bias.
3. Users – who uses social media, how they use it,, and what is the ratio between users and posts
4. Socio-political factors – including cultural difference in use of platforms, and extent to which there is freedom of speech or censorship.

The more analysts understand about the platform, topic, users and socio-political climate, the better they will be able to make their own judgement about limitations of representativeness, and thus the appropriate weight to attach to the findings.

Robustness

Robustness is defined as how well a test performs when variables, assumptions or the environment are altered. In the context of social data, it is important to consider robustness in terms of the quality of data collection, aggregation, analysis and interpretation.

Social media data was generally seen by the experts as not being inherently robust, due to significant design and researcher effects on data quality. It was viewed as highly vulnerable to changes across a range of variables. These included changes to social and analytical platforms' algorithms or data access

policies, and researcher decisions throughout the research process from query development to data cleaning. The issue was further compounded by the 'black-boxed' nature of social media analytical platforms, which experts criticised for lacking transparency in their approaches to collecting and cleaning social data.

However, experts suggested opportunities to improve data quality: it was deemed key for researchers to be aware of the methodological considerations and context they are working in when using social data. Important steps included building a relevant and accurate search query in the first place, identifying the differences between sources of data and how people use platforms differently, deciding how to deal with bots, excluding irrelevant data and checking for bias in the data and processing systems used.

Reliability and prediction

The reliability of social media and its ability to predict phenomena are some of the areas of greatest debate; both the literature and expert opinion is mostly divided into contrasting camps. For some, social data is reliable enough to make tactical predictions in specific contexts. Others believe a lack of transparency over methodology and analysis tools, enhanced feedback loops (that create false trends) and a constantly changing corpus of data makes this impossible.

From a data collection perspective, experts felt that reliability was also variable, and dependent on the topic of discussion, the volume of data available and the platform on which the conversation was being held. Yet the expert group disagreed on the details. Some said that higher salience topics such as politics would be more reliable due to the increased likelihood of people raising their personal opinions; others felt that these topics were more open to interference from bots and trolls, reducing data reliability. The platform hosting the data was also an issue, as the way different providers allow data access or screen their APIs was unclear.

There was greater agreement on the role of data analysis in reliability – the general view was that opacity from 'black box' analytics of existing research platforms hampered the ability to produce replicable research. Both experts and the literature review highlighted that researchers can also be opaque about their data collection methods from platforms. Issues related to low transparency in data collection and analysis could begin to be remedied by greater transparency of methods.

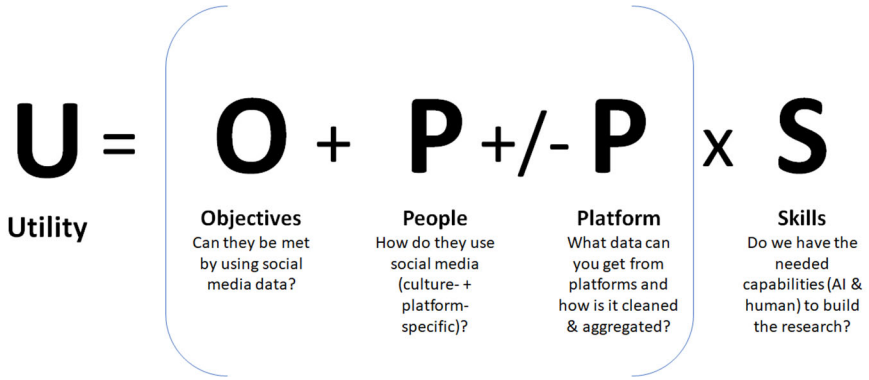
The expert group was not yet convinced that social media data could yield accurate predictions; the literature review found no clear conclusion on predictiveness either. Existing evidence suggests that social data is less accurate in predicting electoral outcomes than polling. There is also less understanding of what makes a good (or bad) prediction from social data. Echoing the point on data analytics above, there was a perception that those cases of 'successful' prediction had been cherry-picked from a wider range of failures. Many were not genuinely predictive, and instead were reverse-engineered to show that one could have predicted a past event if one knew which variables to look for. Again, a lack of full transparency makes it difficult to assess the claims and counter-claims.

1.5 Defining value

On balance, the available literature and expert testimony suggest that the representativeness, reliability and robustness of social media data is weak. Yet it would be wrong to assume that the utility of social media research is driven purely by the extent to which analysis is representative, reliable and robust.

We propose that the overall utility of social media data can be summarised as a formula, which considers representativeness, robustness and reliability alongside specific objectives and contexts; and where risks can be mitigated by analysts.

Figure 1.2: Key considerations for assessing the utility of social media data



Within this model there are 4 overarching factors:

- **O** is for **Objectives**. The first consideration is to identify the strategic objectives of the project. What do we want to do (broader strategic objectives)? And what do we need to know to do it (Research objectives)? The next stage is to review which methods are available to gather evidence relevant to these objectives; and to consider whether objectives can be met using social media data. Social media analysis should not be conducting just because it is technically possible.
- The first **P** is for **People**. This should seek to go beyond traditional debates of representativeness as defined by who uses social media. Further consideration is needed of the broader cultural context that shapes who uses social media and how. Political context and the role of the state is also key, in shaping how freely people post online.
- The second **P** is for **Platforms**. These can be a help or hindrance to the overall equation depending on their performance. Key considerations in how data can be obtained; what data and metrics can be collected from the social media platforms, and how do social listening platforms and aggregators collate and analyse the data?
- **S** is for **Skills**. Skills are a multiplier because they can both mitigate against risks and enhance the benefits. Analysts are required to make large numbers of justifiable decisions during data collection and analysis. The quality and transparency of these decisions directly impacts on the robustness and reliability of the findings. If an analyst has limited skills, whatever they do is likely to have limited utility. Skills partly reflects the AI capabilities available to conduct analysis, but it mainly concerns researchers' skills in building research and analysis that meets the objectives.

We propose that the combination of these factors is more of an art than a science. Each organisation and analyst must trade off the strengths and limitations of using social media for any given purpose. Despite its limitations, social media research may still yield significant value. The framework presented here provides a means to assess the quality of findings from social media research and should be used to guide how much weight should be given to results in decision-making.

1.6 The future

The world of social media moves fast. New platforms emerge, internet access continues to grow, and governments continue to grapple with the politics of online harms, privacy, and freedom of speech. These dynamic factors all affect the quality and value of social media data. What, then, does the social media sector look like in 2025?

Based on a detailed analysis of current drivers and trends derived from primary and secondary research, we have created three plausible scenarios for the near future. Our analysis has led to three plausible

directions of travel for the sector, each of which has implications for those individuals, businesses and governments which seek to engage with social media or to use its data as a basis for decision-making. These are not definitive predictions of how the social media sector will unfold and no single scenario has to be "correct". Their success is causing reflection on the present, as it may yet be possible to influence how the future unfolds.

Each scenario has specific implications for research, and key signals to monitor that will help indicate the future value of social media data.

Octopus Corporations

A world where existing social platforms prove their worth to citizens and governments, strengthening their position further and allowing them to extend their operations into new areas and services. The number and breadth of citizens on these platforms will be increasing. In many countries they will become increasingly important to governments as a way of communicating with citizens and measuring their needs and interests.

Key implications: Increasing user volumes on existing networks supports current research models but debate will continue over the representativeness of this data. The biggest challenges to research will come from keeping pace with expected technological developments in online content and increasing attempts to influence online opinion by state and non-state actors.

Signals of this future: Key social network penetration hits near-universal levels in many countries; networks widen their service range and work more closely with government.

Digital Fortresses

A world where attempts at regulating what goes on online have failed. As a result, a lack of trust among citizens and governments drives a retreat from open spaces into locked and private forums – and potentially offline completely. Geopolitical ructions also fragment the regulatory landscape further, making it harder for the same firms to operate across countries. Instead, nationally-aligned and smaller-scale social networks rise in their place.

Key implications: The capacity for representative, reliable and robust research using social data would decline in this scenario as no single network will cover a wide range of people and an anxious public becomes more mindful of what they express in remaining public online spaces.

Signals of this future: New government regulations on fake news and online harms are abandoned, watered down or deprioritised, while public trust in a wide range of communications channels falls with trust in online information.

The Curated Internet

A world where stronger regulations on social networks make running universal networks more challenging, while the public are becoming more aware of the value of tailored web experiences and their own data. As a result, social networking will start to move to a model that is based on private networks and small communities united by common interests, including subscriptions.

Key implications: In a more commercialised internet the cost of online research and data access will rise which will influence the scope and types of research conducted. Tighter regulation on speech online will also shift more online discussion into paywalled and encrypted platforms.

Signals of this future: Headline legislation defining social networks as publishers in several countries, with people gravitating to more controlled and subject-specific social networks and causing social networks to move towards subscription and other models to monetise their data. This specialisation would also drive specialisation of social data research and the people carrying it out.

1.7 Implications for research and policy making – establishing the right environment for social media data to flourish

As noted above, there is no data source that perfectly reflects public opinion and behaviour – all have margins of error. Social media data can offer significant value, yet its utility is variable. It is also impossible to judge the value of social media data outside of a given context. For example, in some situations it may be the only source of insight into populations available; in others its data may not be relevant to the target audience of interest. Whilst the high volume and low cost of social media data makes investment in collecting and analysing it worthwhile, it is important not to set unrealistic expectations of what it can achieve, and what the applications are within research and policy making.

Social media data should be used as part of a flexible and context specific research programme, which draws upon as broad a range of online and offline data sources as possible. It should be further assumed that quality will be variable between given contexts. It is also important to consider the many other ways social media can be useful beyond accurate measurement of popular opinions, behaviours and motivations. It is a useful qualitative insight tool, to help provide depth of insight into specific audiences and topics of interest. It can also inform question design, gather real time knowledge of new situations and identify gaps in need of further research.

The successful application of social media analysis relies on empowering analysts to marry the strengths of social media data to the right strategic communication objectives. It also requires informed judgement about how much weight to place on the findings based on the known limitations in any given context.

The use of social media data can flourish in the right environment, or flounder without due care and attention. Key principles for consideration include:

- being clear about research objectives and the key audience(s) of interest,
- cross-referencing the political as well as technical implications of social data analysis to ensure its use is appropriate,
- seeking to assess all aspects of data quality,
- considering the specific cultural context in which data is created and analysed,
- being open and transparent about any limitations,
- developing robust theoretical frameworks on which data can be assessed,
- using highly trained analysts with the appropriate skills to collect, analyse and apply social media data
- implementing a robust quality and assurance process to mitigate risks of processing and sampling error

2 Introduction

2.1 Project overview

Social media platforms such as Facebook and Twitter provide a vast body of data which can produce insight into public beliefs, attitudes and behaviours. However, there is still much debate over their reliability and utility compared to other forms of research.

The Defence Science and Technology Laboratory (DSTL) commissioned Ipsos MORI[3], the King's Centre for Strategic Communications (part of King's College London)[4] and the Social Intelligence Lab[5] to conduct research assessing the utility of social media data. Online sources provide a vast body of data, which can help leverage insight into public beliefs, attitudes and behaviours. However, there is still much debate over their reliability and utility compared to other forms of research.

This project uses primary and secondary research to build understanding of the value of using social media data to inform online audience analysis and the factors that should be taken into account to compare its utility in different contexts. The project also seeks to contribute to the wider body of academic knowledge and act as a reference document to researchers and analysts undertaking behavioural and attitudinal research more broadly.

The core question it seeks to answer is:

"How robust, representative and reliable is content posted online in providing insight into the behaviours, motivations and attitudes of wider populations?"

Within this question there are a number of areas of interest, including two key sub-questions:

1. How closely do the opinions and stated behaviours posted by people online reflect their actual opinions and behaviours?

2. How representative of wider populations (i.e. those posting no online content) are the views and behaviours of those posting online content?

Wider reflections include:

- How representative and reliable are other sources of data (such as opinion polling) against which we can compare social media data?
- How do these comparisons vary by topic of conversation?
- How do these comparisons vary by online platform?
- To what extent is access to data an issue?
- How do these comparisons vary by country?

[3] Ipsos MORI is a market leading global research company. As a full-service research agency, Ipsos MORI has multiple research specialisms, these include Research Methods, Innovation and Trends and Futures.
[4] The King's Centre for Strategic Communications (KCSC) a leading global centre of expertise on strategic communication. It is led by internationally renowned experts from the Department of War Studies and partners from the policy and practitioner communities
[5] The Social Intelligence Lab is the leading source of news and insight for social intelligence professionals and a professional membership association.

2.2 Objective and method

This project used a variety of methods to meet the research objectives of providing a clearer understanding of the robustness, representativeness and reliability of content posted. A summary of the four workstreams is included below:

	Objectives and outputs
Literature Review	A team at the King's Centre for Strategic Communications (KCSC), led by Dr Thomas Colley, conducted a review of academic literature into the quality of social media data and analysis. The review will be published as a standalone academic paper to summarise the state of the art at present.[6] The review of literature also helped identify key experts and laid the groundwork for the discussion guide to be used in the subject matter expert interview stage.
Subject matter expert interviews	A team from Ipsos MORI, KCSC and the Social Intelligence Lab conducted 30 hour-long interviews with global experts from across a broad range of disciplines including academia, politics, market research and defence and strategic communications. These interviews built on the literature review to provide a wide perspective on the quality and utility of social data and gather examples of the practical challenges and experiences faced by those working in the field currently.
Case study	During the project, Ipsos MORI conducted a live case study comparing the data collected online though social media with data gathered through a traditional survey approach to track public attitudes around the COVID-19 pandemic. The case study draws out the strengths and challenges faced by both.
Delphi survey	Twenty-three experts from the interviews participated in a Delphi panel survey to develop future scenarios for the social media sector in 2025. This survey explored their attitudes towards the key drivers identified though analysis of the interviews, resulting in three potential futures for the sector.

2.3 Notes on interpretation

The following notes should be considered when drawing conclusions from the workstreams reported in this Final Report.

Workstream 1:

The literature review can only cover a sample of the tens of thousands of papers employing social media analysis. The review focuses specifically on issues relating to representativeness, robustness and reliability. Its main focus is on literature published in English from the last five years, though it does include seminal texts and review articles from the last decade. Most of literature stems from academic research, but the review also draws on relevant research from think tanks, government organisations and private companies engaging in social media research. The review examines case studies from around the world, though it is worth noting that most come from Western liberal democracies. Greater diversity of cases is a key area for future research, since only limited generalisations can be made about social media platform usage and behaviour across cultures.

Workstream 2:

Subject matter expert interviews are intended to be illustrative rather than statistically reliable. Given their qualitative nature, the data collected from the in-depth interviews aims to provide detailed and

[6] Thomas Colley, Harris Kuemmerle, Yeseul Woo and Neville Bolt (2020) Social Media Data's Representativeness, Robustness and Reliability: A Review. King's Centre for Strategic Communications, King's College London, 2020

exploratory insights into the opinions, attitudes and judgement of subject matter experts. Although the sample was designed to ensure that a range of different experts were interviewed, the sample itself is not intended to be representative. It is not possible for qualitative research to provide a precise or meaningful indication of the broader prevalence of a certain beliefs and opinions due to the relatively small number of participants involved. The aim instead is to capture the range of opinion and identify areas of consensus and disagreement.

Workstream 3:

The initial stage of the case study comprised of an online survey of those aged 18-75 in Great Britain. The online survey ran over two separate weeks between the 10th and 27th of April 2020, with a total of 2,149 responses. The survey sample was weighted by age, gender, region and working status to ensure that it was broadly representative of those online aged 18-75 in Great Britain. However, it was a survey and as a result it is important to note that significant associations, and not causal effects, are reported.[7]

In parallel to the online survey, the case study involved social media data collection through the social media analytics platform Synthesio[8]. This focused on three topic areas, aligned with the questions covered in the survey:

- Concern about COVID-19 - for the individual and for the country as a whole
- Timing of UK government lockdown measures
- UK armed forces relating to COVID-19

Data was collected on these topics through user defined search queries developed by Ipsos MORI. The queries worked as a search formula, using a combination of keywords (which are not case sensitive) and Boolean operators (AND, OR, NOT, NEAR) to isolate information. Full queries can be found in the appendix.

Data was collected for the period 1st January – 5th June 2020, but a shorter period is used in parts of the report to align with survey fieldwork. Only public content that was still available at the time of data capture (i.e. had not been deleted from Twitter) was included and then anonymised.

The focus of this case study was to test different approaches to data collection, cleaning and analysis. As such, details of the various approaches used to filter and analyse the data are detailed throughout the report.

Workstream 4:

While the Delphi survey used for the future scenario element of this research was conducted using a survey methodology, it is a fundamentally qualitative exercise seeking to build on the reflections of the expert panel to create potential futures. Twenty-three of 30 experts participated in this stage of the research.

[7] Furthermore, there are natural limits to the representativeness that is achievable for the general population when doing online surveys, especially for older and less well-off sample groups. It is worth noting that the representativeness of the questions around internet use may be reduced because the survey was carried out online. For this reason. the survey was planned to be carried out using face-to-face methods, however this was not possible due to the impact of COVID-19 on face-to-face fieldwork.
[8] Synthesio is a social listening platform and part of the Ipsos Group. The platform sources social data from a range of social media platforms, and enriches content with metadata (such as age, gender and geo-location). Further information can be accessed at: https://www.synthesio.com/.

The aim of foresight research such as this is not to predict what will happen next, but rather to produce plausible future scenarios based on the potential trajectories of current key issues. They are not definitive predictions and the metric for success is not in providing a single scenario which proves to be correct. Instead, the future that emerges is likely to contain elements of all futures listed, plus additional factors that cannot be foreseen. The success of scenario planning is that it provokes greater thought about the future now, leading to better and more informed decision-making in the future.

2.4 Acknowledgements

The project team at Ipsos MORI would like to thank all the experts who took part in the subject matter interviews, and respondents to the online survey. We would also like to thank everyone who provided help and support in the design and delivery of this research, in particular: Thomas Colley, Harris Kuemmerle, Yeseul Woo and Neville Bolt and King's College London for their initial scoping and literature review, and Jillian Ney of the Social Intelligence Lab for her expert knowledge on the wider social media sector.

3 Defining 'high quality' in social media research

In this chapter we examine the different types of social data and outline the ways in which it has been defined in academic literature. We introduce the research constructs of representativeness, robustness and reliability and explore their application within traditional research methods. We then use the Total Survey Error framework to illustrate how we might assess the strengths and limitations of social data in the context of representativeness, robustness and reliability.

3.1 What is social data?

Broadly speaking, social data is anything user generated and shared publicly. Social data requires people interacting in social contexts, often social media platforms. Social media data is generally divided into two categories:

- **Passive data:** This is created as a by-product of a user's interaction with social media, such as location data, friends and networks, search history etc. These data are not explicitly created by the user and, indeed, are often collected without their knowledge.
- **Active data:** Active data is intentionally created by the user, such as messages, comments, status updates, tweets, audio and video content, or product reviews.

Social data, or 'Big Data' as it is sometimes called, is inherently hard to define. The literature warns against thinking about Big Data solely in terms of 'volume'. There is an important distinction to be made between 'Big' Data, which is a 'by-product of digital activity', and 'small' data, which is generated through a more formal research process. The distinction between how 'big' social data and 'small' data are generated is important, yet the literature indicates that it is regularly overlooked.

3.2 The 'Six Vs'

The literature highlights a number of attempts to define social data. Whilst definitions varied, there was appeared to be a broad consensus on Big Data possessing six key qualities:

1. Volume – at least measured in terabytes.
2. Velocity – being created in near to real time.
3. Variety – contains a mixture of structured and unstructured data both temporally and geographically.
4. Veracity – the data contains noise and bias making it hard to produce valid findings.
5. Volatility – changing technology and regulation makes it hard to produce reliable findings over time.
6. Value – something of value is derived from the data[9].

These six qualities speak to the fundamental 'messiness' that is social data. It is this natural messiness that make the methodological endeavour for representativeness, robustness and reliability difficult to achieve. The 6th V, value, is perhaps where there is scope to re-imagine the way researchers could

[9] Hammer et al., 'Big Data'; MacFeely, 'Big Data', p.27

make use of social data. It is clear that there is scope for researchers to derive valuable insights from social data. Although it is important to understand the strengths and limitations of representativeness, robustness and reliability of social data, it is imperative that the researcher's role is not overlooked.

3.3 Defining the 'Three Rs'

In order to answer the primary research question "How robust, representative and reliable is content posted online in providing insight into the behaviours, motivations and attitudes of wider populations?" we need to interrogate the meaning and function of representativeness, robustness and reliability.

Our working definitions are as follows:

- **Representativeness**: whether the sample of data accurately reflects the broader phenomenon or population that is being studied.
- **Robustness**: How well a test performs when variables, assumptions or the environment are altered.
- **Reliability:** Whether repeating the same method would lead to the same result.

Representativeness, robustness and reliability originate from a quantitative research tradition. In a very basic sense, quantitative research is predicated on the belief that phenomena, or constructs can be empirically measured. Within traditional social research, quantitative data can be collected though a measurement instrument (e.g. a survey). Phenomena such as behaviours, motivations and attitudes are harder to measure, and require a high-quality measurement instrument. The construct of quality here would be determined by, among other things, the degree to which the data collected was representative, robust and reliable by design.

3.4 Total Survey Error Framework

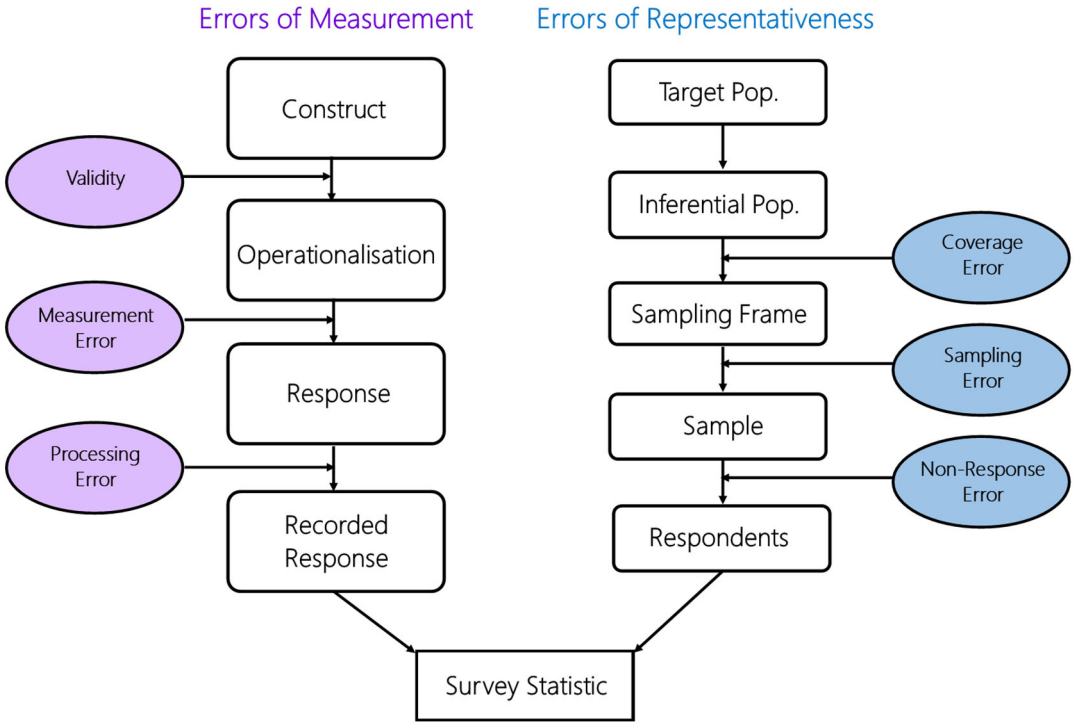
When it comes to representativeness, random probability surveys are widely thought as being the 'gold standard' of research. Within survey research there is a useful framework that allows methodologists to assess the quality of the survey data by identifying all the potential sources or error that arise at each stage of the survey design and implementation. The Total Survey Error framework[10] divides the survey process into two main strands, one concerning the representativeness of the survey sample and one concerning the accuracy of measurements made. Total survey error (TSE) refers to all sources of bias, or systematic error, and variance, random error, which might affect the validity and accuracy of data[11]. In theory, the lower the error the lower the sources of bias. The lower the bias, the more accurate the measurement of population characteristics and, the more representative the data. As such, recognising bias is crucial for determining the utility of the findings.

Survey methodologists use the Total Survey Error (TSE) framework to help conceptualise sources of error in traditional survey research. The diagram below has been adapted to highlight sources of error that are particularly relevant to the assessment of the representativeness, robustness and reliability of social media data.

[10] Groves, R. M., Fowler F. J. Couper, M. P., Lepkowski, J. L., Singer, E. and Tourangeau, R. (2009). Survey Methodology (2nd ed.). Wiley

[11] Lavrakas, P. J. (2008). Encyclopedia of survey research methods (Vols. 1-0). Thousand Oaks, CA: Sage Publications, Inc. doi: 10.4135/9781412963947

Figure 3.1: Total Survey Error Framework, adapted for social media research



We have identified six key sources or error that are relevant to social media research – a more detailed account of these can be found in the appendix. This framework provides a useful way of thinking about the potential areas of bias that should be examined when using social media data. To date, there has been considerable focus on errors of representativeness within social media research, but less so on measurement. We argue that both deserve equal scrutiny.

- **Validity:** are the inferences made within the data valid (e.g. what is signified by a 're-tweet', or being 'friends' on social media)?
- **Measurement error:** what is the deviation between true public opinion or behaviour, and those shared on social media (e.g. are some issues more taboo, or polarised, or open to social desirability bias[12])?
- **Processing error:** how accurate are methods used to categorise, analyse and interpret data (e.g. automated sentiment analysis or appended demographics[13])?
- **Coverage error**: to what extent does the target population match the population accessed and sampled through social media (e.g. restrictions to public data made available by platforms, or risk of editorial censorship by governments and platforms)?
- **Sampling error:** to what extent does sampling of data accurately reflect the conversation of interest (e.g. through the development of a social media search query, or the variable terms of access to data)?

[12] For example, social media users feeling compelled to present a certain image of themselves online that is more 'acceptable' to society in order attract positive attention.
[13] For example, classifying a post as being 'positive' or 'negative' or estimating age based on the contents of a post.

- **Non-response error:** which groups cannot be contacted or persuaded to provide data (e.g. which groups are less likely to create accounts, and post; extent to which data is further skewed by prolific users)?

It is important to reiterate the fundamental differences between social, or 'big' data, and traditional, or 'small' data. Although the three R's may provide a useful framework upon which to understand and assess the utility of social data in different context, we must be mindful that they are not going to be universally applicable. As discussed in chapter 6, we should be cautious when using them as a way of determining the overall 'value' of social data.

3.5 Social Data: The three bodies of literature

It is perhaps unsurprising that, due to the inherent complexity of social data, there is a lack of consensus on whether social data can represent the opinions and behaviours of populations. Despite this, three dominant 'bodies of literature' were identified in the literature review conducted by King's College London.

1. The Data Science perspective:

Data science research tends to support the notion that social media data has both representative and predictive abilities. This literature mostly assumes that social media data can be sufficiently representative of both behaviours and attitudes, and seeks to provide evidence of it. However, pressure to produce positive findings means that much research underplays significant issues of data quality, obscures or minimises outliers and sources of error, and findings are rarely reproduced in other studies. Negative findings are rarely published.

2. The Commercial perspective:

Commercial literature is extremely positive about the representative and predictive quality of social media data. Unsurprisingly, this optimism is greatest where the commercial literature discusses a given companies own analytics software programme. As is typical with commercial work, it is difficult to comprehensively assess the veracity of the claims being made. There is little commercial incentive for greater transparency and doing so would risk revealing methods to competitors. Therefore, commercially generated software for social media analysis is typically 'black boxed'.

3. The Critical perspective:

There is a significant body of research that highlights the range of potential issues that may hinder the representativeness and predictive ability of social media data. Critical literature indicates that social media data is highlight localised, vulnerable to bias, rapidly outdated, neither robust nor reliable and fundamentally unrepresentative of populations both online and offline. A common criticism raised is that data science research illustrating positive effects either lacks transparency about the methods used, or relatedly, that it is unclear whether researchers tested multiple variables until one achieved effect – which may be due to chance. These papers typically find dozens of sources of error, but the effect of each error is rarely clear, since more research is needed.

T. Colley et al (2020)[14]

[14] T. Colley et al (2020) 19-21

3.6 Conclusion – what does this mean for social data?

Each body of literature is a direct product of the underlying values and imperatives of the communities in which they have been produced. The critical perspective is underpinned by the broader epistemic tension between the qualitative and quantitative research. Although the critiques are valid, many of them bear similarity to a wide range of different research methods and types of data.

It is therefore necessary to start any assessment of the value of social data by accepting that there is no one perfect data source. All analysts have is a range of different and imperfect data sources, each providing them with a slightly different way of understanding the world. However, to date, there has been little public discussion of a robust framework to help assess the quality of, and potential bias within, social media research data. It is hoped that the findings presented in this report may help lead to a common set of standards.

4 Representativeness

In this chapter we examine the extent to which social media data is representative. In the context of this research, representativeness is defined by the extent to which a sample of data accurately reflects the broader phenomenon or population that is being studied. The chapter outlines the important contextual factors which impact upon representativeness; platform, topic of conversation, social media users and socio-political factors. We also discuss the important role the researcher plays in identifying and interpreting these contextual factors in relation to representativeness.

4.1 Representativeness and social media

Social media generates a large amount of data. For this reason, it tempting to assume that the sheer volume of content posted online will be representative. In the same way that people may consider a 10,000-person survey sample to be ten times as good as a 1,000 person sample (when the actual improvement is significantly less than this), the experts felt some people made an automatic link between the enormous volumes of data obtained through social networks and representativeness. There was concern among some experts that people had become blinded by large numbers.

"Culturally, our understanding of what social media is, is distorted by the way of reporting. Policy makers, analysts and journalists claim [social media data] must be representative of everyone because numbers are large"

Public sector analyst

The findings of the live case study highlight these difficulties interpreting volume. Comparing the topline results of the survey with the social media analysis reveals a large difference (up to 17 percentage points) in the overall proportions who said that that the UK government's COVID-19 response measures were too late. The sample taken from social media data was more likely to register the opinion that that the measures were implemented too late[15]. This discrepancy is explored further in chapter six.

Social media is a response-based medium. Unlike survey data, in which each person is counted only once, a social media dataset is made up of posts, not people. This means that multiple posts might come from a relatively small number of users. For example, in the social media dataset used in the live case study one topic was present in 26,183 posts, but these posts came from 20,632 individual accounts. The ratio of posts to users is an important consideration for analysing representativeness, as a small number of super users, bots or fake accounts reduces the potential for representativeness. Although volume is an interesting metric, it should not be used in isolation.

"Twitter has the volume - it is the loudest, but database of users is small, and not growing. It is not remotely representative."

Private sector analyst

While the size and complexity of social media is what makes it a lucrative data source, it also poses a key challenge to representativeness. Many of the experts suggested that the complexity, and uniquely unknowable conditions in which it is generated means it is almost too far removed from traditional

[15] Wave 1: 57% of survey respondents felt that the government measures were brought in too late, compared to 74% of social data sample.
Wave 2: 66% of survey respondents felt that the government measures were brought in too late, compared to 88% of social data sample.

methods to be assessed using the same criteria. Among this group it was felt that social media data was not 'representative' in the traditional sense of the term (i.e. did not represent all of society); some argued that the use of representativeness as a benchmark for judging the quality or utility of social media was a false problem.

The experts echoed a key point of difference between the two data sources that was also highlighted in the literature review: data generated by survey research is representative by design. For example, the online survey used in the live case study was purposefully designed to be representative of the national demographics. By contrast, social media is not designed to generate representative data, nor is it designed to be used for research. This does not mean it has no research value, rather that it should be assessed, analysed and used in a different way to more traditional research data.

"I've moved away from idea that we should try to mitigate representivity. It's not like a poll, less like generic views, more a report on what is happening to individuals. You don't have to try and generalise."

Social media academic

4.2 Context matters

One of the key findings that came out of the literature review and expert group was the importance of context. While it is difficult to make any absolute claim to the representativeness of social media data, it is clear that representativeness, and the importance of being representative, is highly context-dependent. As discussed elsewhere in this report, representativeness may not be the most important or relevant quality of a given data source. Traditional social research generally adopts the most suitable method depending on the purpose of the research. Social media research should be no different. No approach is perfect, and all methods have their limitations. However, understanding the context in which the data has been generated will help researchers to use data more effectively.

"It all depends on the context, and who uses social media and how in a given cultural context. Different platforms are used for different purposes in different places. What social media is representative of will vary over time, too, depending on the situation as well."

Defence/Strategic Communications analyst

4.3 Assessing the Representativeness of Content Posted Online

However, the complexity of the social media makes identifying these 'contextual' factors challenging. To help assess the representativeness of social media data we have identified four key factors: platform, topic of conversation, users and socio-political context. These are discussed in more detail below.

1. Platform

Accessibility of data

Platforms differ significantly in terms of privacy policies, access to data and number of users. Approaches range from Twitter, where users have very little 'private' data, to secure messaging services like Telegram and WhatsApp, which are completely end-to-end encrypted. The availability and accessibility of data has significant implications for representativeness.

Even on relatively open platforms there may be obstacles to accessing social media data. These obstacles range from private accounts to Application Programming Interface (API) limits on the volume of data that can be accessed for research. For example, Twitter's streaming API only allows researchers

to directly access one per cent of the data on their chosen topic.[16] Although this proportion may contains millions of tweets, it is difficult to know whether the tweets included in the one per cent sampled are representative of all social media users, let alone offline populations.

Platform Design

By their design, different social media platforms promote and report different attitudes and behaviours. This has two aspects: firstly, people use social media for a variety of different reasons. These reasons will determine the types of content they post on different platforms. Secondly, the platforms have complex feedback loops[17] and collaborative filtering algorithms which function to amplify and filter the content that users are exposed to. This can make it difficult to accurately interpret the gravity or legitimacy of the content being posted. For example, viral or 'trending' content is determined by an algorithmic feedback loop within a particular social media platform. It is difficult to determine whether a certain topic is 'trending' because it is of particular interest to a large proportion of social media users, or whether it has been artificially amplified by the feedback loop. This is particularly hard to establish when the content in question is controversial or polarising.

Platform Culture and Structure

The design and structure of different social media platforms creates different online cultures. The culture on a given platform might determine the type of content posted and the behaviours exhibited by users. This might have a profound impact on the representativeness of the content posted. For example, platforms such as 4chan and 8chan illicit a self-consciously offensive and ironic 'meme' culture. These platforms are now increasingly associated with far-right political groups and individual acts of terrorism. The platforms are designed as anonymous conversation threads, which means that assessing the representativeness of opinion or behavioural intention is essentially impossible. There is no filtering algorithm. Content only remains visible if someone replies to it, whereupon it is 'bumped' back to the top of a thread. Otherwise, comments can disappear from a thread in seconds. This incentivises users to post provocative content, which drives the culture on the site towards extremes.

2. Topic of conversation

Social desirability bias

The public nature of conversation on most social media means that 'social desirability' bias – people expressing views that they feel others want to hear – is a significant factor, as it is for traditional surveys and qualitative research. Social media adds another level, however, because unlike a survey, social media posts are usually publicly accessible. Views expressed online may therefore not be representative of an individual's 'offline' or 'real' views. Some users may moderate and self-censor, fearing criticism. Others may engage in a more exaggerated and exhibitionist way.

"You don't find the silent minorities, who tend to be under-represented more generally"

Social media academic

The expert group noted that these differences vary greatly depending on cultural norms. There are many social, cultural and political factors that influence online behaviour. Understanding the cultural context in

[16] Though social media monitoring platforms have negotiated access to greater levels of data, the level of access differs by provider.
[17] Feedback loops can create false trends that feed themselves – e.g. a user click on something that is trending because it is trending, not because they were looking for it.

which specific topics are being discussed on social media is extremely important, particularly when considering representativeness across multiple countries.

Some topics can be more representative than others

As with all research, it is easier to capture opinion and behaviour on some topics than it is for others. People are more likely to be honest about topics which are relatively uncontroversial, such as food or fashion. This is less likely with controversial topics.

"Representivity is when people can talk honestly about the subject. If the topic is explosive, or people have a personal or commercial relationship to the topic, there is less honesty"

Private sector analyst

However, it is generally the most controversial, compelling and provocative topics that gain the most traction. Experts tended to agree that social media analysis performs better in capturing views at the extreme ends of public opinion but tends to under-represent the middle ground. However, users' opinions on more controversial topics are considerably less representative of their 'offline' views.

"Social media is a game of extremes. It's very good at showing opinions of those at either end of the opinion bell curve. But it's not good at the middle bit"

Private sector analyst

3. Users

Who uses social media?

Another important consideration is coverage: the expert group acknowledged that social media users are generally unrepresentative of the demographics (and other characteristics) of populations at a national level. The literature review emphasises the importance of understanding the differences in demographic representativeness across platforms. For example, research indicates that globally 56% of Facebook users are male, with 62% of users aged under 35. Instagram is even more skewed in a different direction: in its user base, 32% are male and 90% are less than 35 years old[18]. The demographic representation of social media varies across countries, platforms, and active users.

There are a range of structural, cultural and political factors that affect who is (and who isn't) represented on social media. For example, in Somalia, social media users tend to be from more urban areas and have a higher level of education that the broader population. Traditionally most social media users in Somalia were men, who would go to internet cafes and use social media there. However, the proliferation of smartphones, mobile network coverage and internet dongles has made the internet and social media more accessible to a broader population. Similarly, ethnographic research conducted in Iraq indicated that when women use social media they often do so through male personas to avoid online harassment.

Although these structural and cultural limitations pose a challenge to population-level representativeness, social media may still be representative of certain groups. Understanding who does (and who does not) use social media in a given context is key to assessing its representativeness.

[18] London School of Economics, 'Social Media Platforms and Demographics', https://info.lse.ac.uk/staff/divisions/communications-division/digital-communications-team/assets/documents/guides/A-Guide-To-Social-Media-Platforms-and-Demographics.pdf, accessed 5th August 2020

"We have to build country specific models based on cultural behaviour. For example, in Africa, social media is only representative of those who have a smartphone. Access to electricity is difficult and data costs are high, and so infrastructure plays a role in how much they consume and where they comment."

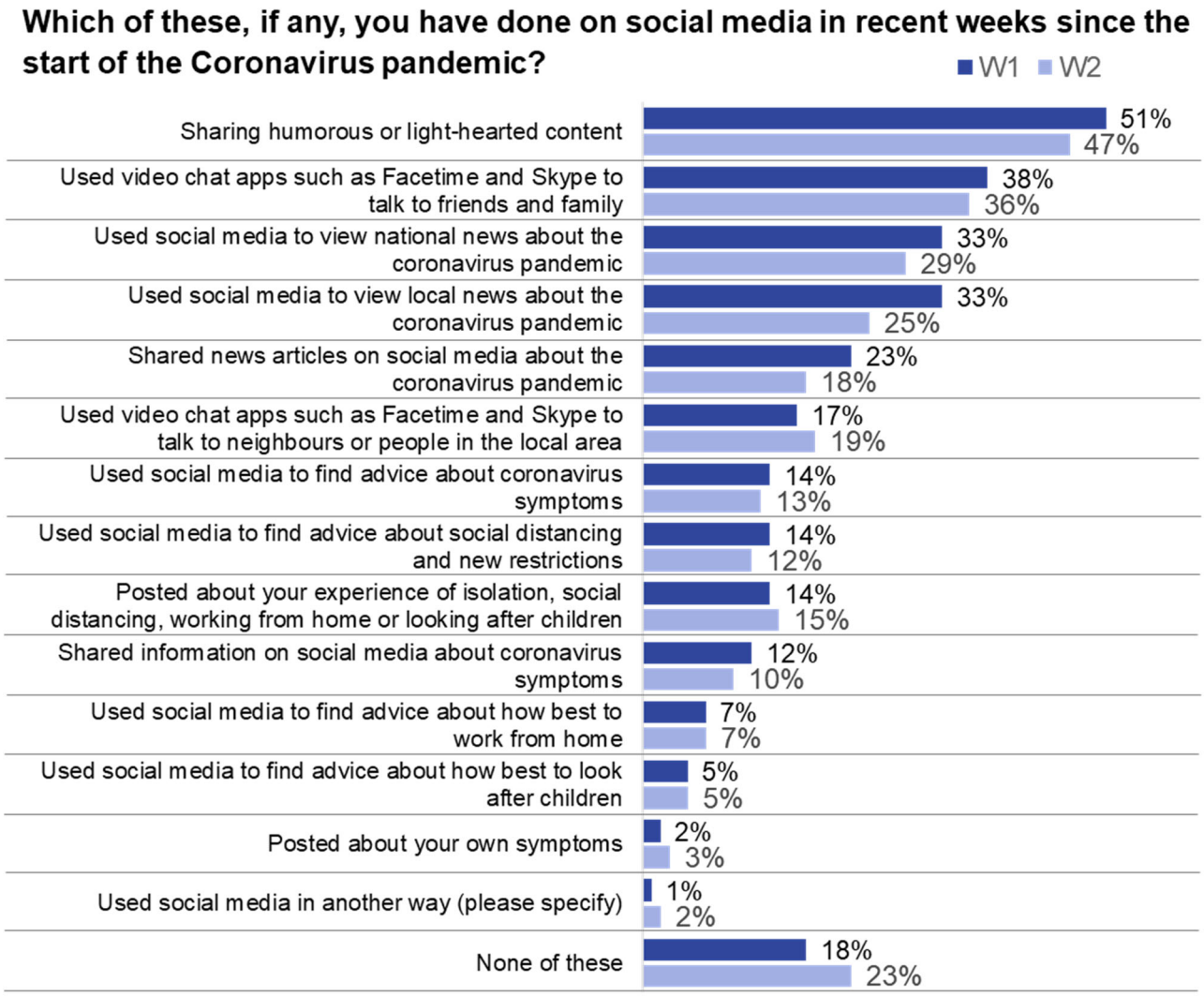
Public sector analyst

How do people use social media?

In addition to understanding who is using social media, it is important to reflect on how different social networks are used in different cultural contexts. Representativeness for any given research query is not limited to the percentage of people who have a social media account, it will also depend on who is an active user, who is posting and what they are posting about within that specific conversation. Not all social media users will share comments on that topic online, and of those that do, many of these opinions will not be relevant to researchers. As the findings of the survey highlight, social media users use social media in different ways.

For example, within the context of our case study on COVID-19, the research team was interested in comparing public opinion on social media to that collected through a survey. The case study found that only a small proportion of social media users used platforms to post about their specific experience. In contrast, around 50% of social media users said that they were relevant 'sharing humorous or light-hearted content' online, and other popular uses were to follow news, share and discover information, and seek advice. By the second wave of the survey, nearly a quarter (23%) of social media users had not used social media in any way related to COVID-19 in recent weeks. It is therefore reasonable to conclude that content that is representative of public opinion may be quite a small proportion of the total volume of user posts.

Figure 4.1: Common activities on social media (survey data)



4. Socio-political factors

Cultural differences in social media use:

Different cultures use social media in different ways. These differences mean that social media may be more or less representative depending on the cultural context. It is important to understand how behavioural, linguistic and cultural differences play out on social media. For example, culturally Turkish users are seen as being very sarcastic on Twitter, Germans use hashtags more than other countries, while Koreans tend to reply more often to tweets. As one expert stated:

"the specific country context is likely to be more important than people [researchers] think. Social media research projects need to consider the ways in which platforms are used culturally. For example, WhatsApp in India is used fundamentally differently to in Brazil; in the Philippines, Facebook is used in a more commercial context. If you don't understand these differences, you won't understand the value of your data."

Social media academic

Freedom of speech and state censorship:

In addition to cultural differences it is necessary to understand the unique political context of social media within a given country. The attitude of the state toward social media, freedom of speech and censorship can impact representativeness. It will also affect the types of social media platform used. For example, citizens in less democratic countries increasingly use encrypted platforms such as WhatsApp to share news with a wide range of contacts, many of whom they may not know personally.

4.4 The Role of the Researcher

"Researchers are the most important tools… they need to understand the context of the internet"

Private sector analyst

Identifying relevant contextual factors and understanding how they might impact upon the representativeness of social media data requires researchers to have a deeper 'offline' understanding of platforms, topics, users and socio-political climate. One expert suggested that the most important attribute to do social media analysis properly is "humanness". Another expert spoke about using ethnographers to gain a deeper 'offline' perspective of everyday social media use in Iraq. This insight is critical to informing their approach to 'online' social media research. Researchers need to understand the unique interaction between the platform, topic, users and socio-political climate in order to determine the representativeness of social media data in a given context.

"Researchers must set the correct expectations, knowing specifics about what is possible and what is not possible. They need to advise that the research will be qualitative in nature and be flexible with the analysis."

Private sector analyst

4.5 Conclusion

Attempts to derive representativeness from social media samples face several challenges. Firstly, the data can only reflect what users are willing and able to share online in the context of the specific design and structure of a given platform. Additionally, in this volatile social media environment there are a range of different commercial and political actors attempting to influence and undermine public option and behaviour.

The expert group were generally of the opinion that social media data is inherently different to survey research data and therefore the traditional measures of quality, including representativeness, have limited read-across. As a result, it is extremely difficult to claim with certainty how representative social data is, or could be, and the construct itself is of limited usefulness in assessing the quality of social media data.

The interviews also suggested new metrics and considerations that are more suitable to ascribing value and utility to social media data. From a representativeness perspective, a key takeaway is the importance of context. The more researchers understand about the platform, topic, users and socio-political climate, the better they will be able to use the data to answer their specific question.

5 Robustness

In this chapter we explore the robustness of social media data and analysis. In the context of this research, robustness is defined as how well a test performs when variables, assumptions or the environment are altered. In the context of social data, we consider robustness in terms of the quality of data collection and aggregation by exploring two elements: data collection and data analysis.

5.1 Data Collection

The initial data collection process

Social listening tools, or social media monitoring platforms, collect, collate, and store data in different ways. Some have access to more sources than others or may only collect a sample of the data from the available sources. They may have different approaches to storing historical data, or differing approaches to removing data when requested by social media users, producing a diluted pool of social data which differs between analytical platforms.

The expert group highlighted the importance of the quality of social data for the success of any research project, but there was no consensus. Some felt that very similar data was provided across platforms, while others acknowledged that social listening tools have their own strengths and weaknesses. For example, some tools were more difficult to use when entering queries in Arabic script. They felt it was important that researchers understand the strengths and weaknesses of the platforms they access and the tools they use, as well as the bias existing in sources and social data itself. This would enable them to be able to make appropriate decisions throughout the research process, at the query design, data cleaning and analysis stages.

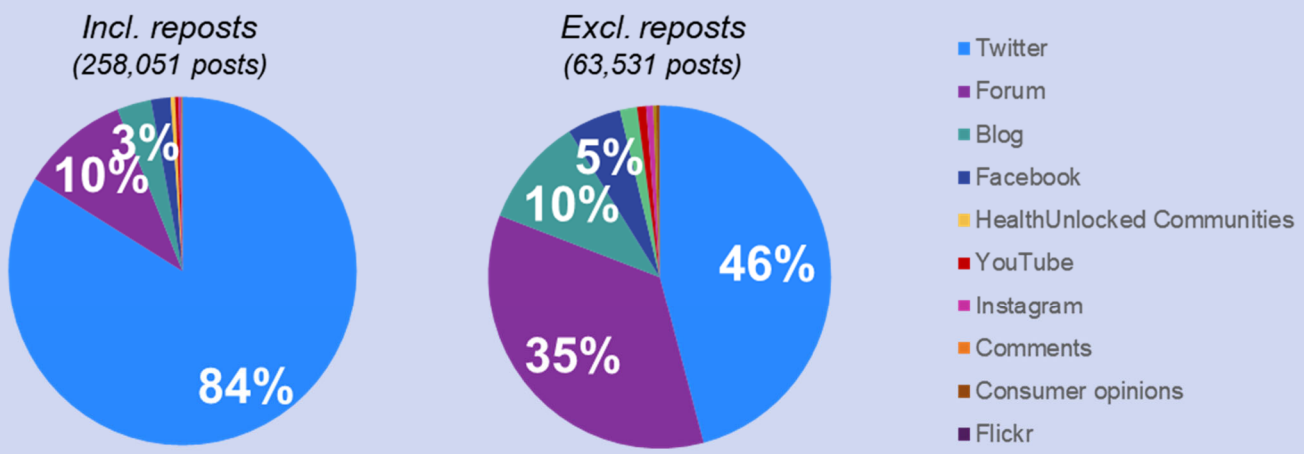
Figure 5.1: Lessons on data selection from within the live case study

'Live' case study: Social listening tools and data sources

Our 'live' case study examined the data sources collected by Synthesio for the topic of personal concern about coronavirus. The findings highlighted the disparity between the volume of content from certain sources in our social media data and the sources people are most likely to report using. For example, only 5% of the social data collected came from Facebook, despite being the most commonly used type of social media in the survey (used by 73%). And whilst less than a third (29%) of our survey respondents said that they use Twitter, the majority (84%) of the social data collected was on Twitter.

The case study also illustrates the challenges of defining a high quality dataset. For example, the chart below shows how the volume of data and distribution of sources changes significantly depending on whether reposts are included or excluded. Social listening tools may have one approach set as the default, but there are arguments for both: including them can skew the data towards those points that get lots of reposts, but sometimes the overall volume of conversation that people will be seeing around the topic is what you are interested in. Researchers must be aware of the factors that influence the quality of their dataset in relation to their specific research questions.

Social data: Concern about coronavirus for yourself



Survey base: Adults aged 18-75 in Great Britain (1,069); Fieldwork 10-13th April; Social data collected: 31st March - 27th April.

Dealing with bots

A key concern around data quality for many experts was how to deal with bots – automated fake accounts on social media. In part, this is related to different approaches to automated accounts between social networks. It was felt that some networks make more effort to remove bots than others: Twitter was mentioned as an example of a network that doesn't do as much as it could, particularly due to their lack of a "real name" policy[19].

[19] Some platforms ask for users' real name at sign up, whilst others such as Twitter ask for any username.

Although there was agreement on bots' ubiquity, the expert group tended to be relaxed about the presence of bots in social data. For most this was because they are easily identifiable, for instance by looking at duplication of posts, multiple posts in quick succession, and the age of the account. As a result, the experts were confident in the ability of analysis tools to identify bots and filter them out of datasets. However, there was some criticism about the opacity of this process and disagreement on who should be responsible for cleaning data. Some experts felt it was the social networks' responsibility to produce clean datasets, while others felt that data science teams or analysts should consider cleansing the data as part of their role.

Others in the group saw bots differently. Those conducting more qualitative analyses online considered bots to make up an important part of the online information environment, so thought it was necessary to include these posts in analysis to understand fully the context people are interacting in. If bots have a communication effect by making people think something is more popular than it is, that effect is genuine, regardless of whether one later cleans them from a dataset.

"They are playing a role in the ecosystem so you can't ignore them. Bots are still part of the conversation."

Private sector analyst

The analytical 'black box'

The opaque nature of many social listening tools was highlighted as an important obstacle to assessing the robustness of social data. It was noted that social data emerged from technology companies, meaning that the sector's incentives are more commercial and there has been correspondingly little interest or pressure to improve sampling quality.

Platform algorithms are rarely shared with analysts and this lack of clarity means that there is no way to understand why different platforms can yield different data from the same search query:

"I struggle with the black boxed nature of tools. We typically try a number of different platforms but sometimes there is no way of knowing how it got that result. If I can't reverse engineer it, we don't report it."

Social media academic

This was a particular issue with bots: while experts were confident in their own steps taken to clear bots from data, the lack of clarity meant they were unsure about the quality of bot removal conducted by the analytical platforms.

This concern was common on other elements too. 'Sentiment analysis' was especially distrusted as it was considered to be too simplistic an interpretation of data and there was no information to allow the experts to verify or check how it was being calculated. This bred distrust of the metric. The role of AI in reporting tools on platforms was seen as a further development of the black box mentality.

5.2 Data Analysis

Analyst decisions

Experts considered all stages of research important in building robust data. They highlighted the importance of building a good query in the first place to ensure data quality, with some seeing building a relevant query as the greatest challenge. Those operating in the commercial sector felt their clients were often disinterested in this issue and were prone to using simple keyword-based approaches, assuming a query is correct because it brings back relevant data, without thinking about cleaning the data first.

Building the query was seen by a number of experts as having the potential to be the weakest part of the project, and as such that it is really important to have a robust systematic and considered approach.

"Every method of research has a 'dark' side. [Cleaning social media data queries is] equivalent to weighting schemes for surveys."

Private sector analyst

Experts mentioned challenges due to the variability of social data depending on factors such as the topic, platform, context:

"There is no generic "how people use social media". It is used by different countries and contexts differently. One size does not fit all."

Defence/Strategic Communications analyst

The literature review highlighted that the evolving nature of platforms and language used means the robustness and reliability of social media data can decline markedly over time. Experts highlighted further challenges such as whether data collected through different queries can be compared when queries are designed in different ways. If, in a query about computers, you tailor your search for 'Apple' using lots of exclusions to remove irrelevant data, but you do not do the same for less popular brand such as 'Asus', these datasets may not be comparable.

Key steps in improving quality included determining the relevance of a query and data to the research question, thinking about sources and how people use them differently, adding in exclusions, looking at people, sites, and checking the bias in the data. Considerations for query development and the effect on the completeness vs. usefulness of the sample of data were illustrated in practice in the live case study.

Figure 5.2 below illustrates the compromises analysts make in attempting to identify relevant social media posts. As our live case study illustrates, a key decision is whether to risk more false positives or more falsefalse negatives in developing a query that is either too broad or too specific.

Figure 5.2: Lessons from query design and development within the live case study

'Live' case study: Query design and development

There is a balance in social data collection between **usefulness and completeness** (although with a good query, we can be more certain that we are using specific terms most relevant to the dataset). Key questions to consider in query design include:

- **How broad or narrow to go?** The 'broader' option will provide a more complete view, but can be very comprehensive and difficult to work with, as there is a huge volume of data that needs to be filtered down. A 'narrow' approach will mean that more of the dataset is relevant, but may be missing parts of the conversation.

  For example, to identify conversation about opinions on the timing of lockdown, one approach would be to use a broad query that captures any mention of lockdown measures, and then look for comments about the timing of lockdown within that dataset. This would produce a dataset with lots of irrelevant content, so cleaning and some way of identifying posts about the timing being too early/too late/ at the right time (such as using a sentiment algorithm) would be needed. Our approach was to build narrower queries that are likely to miss some conversation, but where we know the majority of the data collected will be relevant. This involved 3 queries with terms about lockdown measures as well as specific terms about timing such as "too late" or "not soon enough".

- **How are people talking about the topic?** Are names of specific entities, organisations or key public figures used when people are talking about the topics? Are they using slang, acronyms, key hashtags, or emojis? Desk research to identify the different ways people may be talking about a topic is key to creating a more comprehensive query.

  For example, in our 'military' query we included key public-facing members of the UK armed forces, who were often mentioned in posts reacting to announcements about military intervention in the coronavirus pandemic.

  Desk research is also key for identifying terms that have multiple meanings in different contexts. This may mean that whilst a term is relevant to a topic, it generates more irrelevant content than relevant content. For example, including the word 'Navy' in our military example would collect lots of irrelevant content about the colour, rather that only content about the naval force. Specifying co-occurring terms - e.g. ((Royal OR UK) AND Navy) - in a query can avoid this issue, but is also likely to exclude some relevant conversation.

- **How is your query language reflecting the goal?** For example, reflexive language may be needed in a query if the subject is something more personal. This will help to minimise irrelevant content.

  The 'concern' topics aimed to identify people feeling personally concerned (for themselves or for the country as a whole). To focus the data on individuals' posts, both queries used language such as "I'm worried about" or "my concern". The 'concern for yourself' query then used possessive pronouns (e.g. "my", "our") whilst the 'concern for the country' query used terms about the UK (e.g. British, "our country", "NHS").

- **How are you handling different languages in a multi-lingual market?** Different languages can be included in one query, which will give a better 'total' view. However, this might not be practical (if the query is longer or more complex).

5.3 Analysts

There were mixed views on the relevance of the skills of the analyst in ensuring that data is good quality. Some felt that as there is no one set rule for defining quality in social media, and as such intuition was sometimes needed to iteratively 'feel' or test and learn:

"If you were looking for social media data on care homes and there were clearly fewer older voices in there, then intuitively, that is not right."

Private sector analyst

Others argued that the analysts' skills became less important if you have good data in the first place. Moreover, they suggested that it will become easier for non-experts to do robust analysis end-to-end in the future, with social listening platforms helping them to ensure quality. Although again, the opacity of social listening tools becomes an increasingly important issue under this scenario.

There was broader agreement that rigorous analysis can take place teams have built in quality assurance procedures, such as having multiple people working on projects with review processes in place:

"Internally we always have two people working on a report. A draft is then reviewed by a third, more senior colleague. We ensure that we are clear in our sources, methods and [the] basis for our claims. Externally we approach people for peer review… to challenge our assumptions and conclusions and consider a counterfactual. They ask: 'can we evidence this?'"

Social media academic

In Figure 5.3, the 'live' case study looked to demonstrate some of the considerations involved in cleaning a social dataset and the significant impact of the researcher decisions made on the resulting dataset for analysis. As shown below, filters can dramatically affect the sample of social data, so it is important to consider whether certain filters are biasing the data in an unintended way (e.g. if a platform does not tag a poster's location, filtering by country could filter out all posts from that platform). Robustness is developed where analysts record and justify their decisions and establish consistency in decision making across multiple datasets and projects.

5.4 Conclusion

Social media data was generally seen by the experts as not being inherently robust, due to significant design and researcher effects on data quality. This type of data was viewed as highly vulnerable to changes across a range of variables, including any change to social and analytical platforms' algorithms or data access policies, and researcher decisions throughout the research process from query development to data cleaning. The issue was further compounded by the 'black box' nature of analytical platforms, which were criticised for lacking transparency in their approaches to collecting and cleaning social data.

However, there were some suggestions for improving data quality: it was deemed key for researchers to be aware of the methodological considerations and context they are working in when using social data. Important steps included building a relevant query in the first place, thinking about sources and how people use them differently, thinking about how to deal with bots, excluding irrelevant data, and checking for bias.

Figure 5.3: Illustration of the impact of analyst decisions on data collected during live case study

The 'concern for yourself' query produced an dataset of more than 2 million posts between 1st Jan and 5th June. As part of the set up, Synthesio offers a noise reducer filter, which excludes mentions that are considered noise based on a machine learning model (e.g. ads, job offers, competitions, vulgar content). This was applied to our case study dataset.

As our aim was to capture online conversation from individuals (rather than news/organisations) over a similar timeframe to our online survey fieldwork, we applied a number of additional filters to remove content that we deemed to be irrelevant.

2,208,502

1. Time period
To compare the social data to our survey fieldwork periods, we decided that when answering survey questions participants could be thinking back on the experiences/events of the previous week or two rather than only the day they were asked. The social data we looked at therefore consisted of two two-week periods to line up with our survey fieldwork (W1: 31st March – 13th April; W2: 14th April – 27th April). This reduced the dataset to 682,358 posts; 371,914 posts for W1 and 310,454 posts for W2.

682,358

2. Language
As the query used English-specific terms, it was not surprising that the vast majority of posts were categorised as English. The most common language after English was Japanese (130 posts). A small number of posts (79) could not be categorised as a language. Posts that appeared in the dataset but were categorised as a language other than English mainly contained English website links only, posts of one word only, account names only, or contained a mix of languages within a post. Non-English posts were excluded from the dataset, as we did not have the language knowledge to be able to analyse posts in other languages.

681,220

3. Location
Countries to include in the data are selected at the data collection set up stage. We were interested in UK posts only as the survey data covered Great Britain, but to look at the impact of the country filter we also included posts that were not tagged with location information. Filtering these unidentified country posts out had a large impact on the data, with 463,347 posts unidentified. Based on previous experience often these posts come from the places with the biggest populations – for example in the US in relation to our queries. Looking at a sample of these posts, some did appear to mention the US but for others it was unclear. It is also important to note that filtering out non-identified posts changed the platform profile of the dataset significantly. For example, all Reddit posts (217,375) were non-identified so were filtered out of the data.

217,873

4. Media types and platforms
As we were interested in posts from individuals about their personal concerns, we excluded news posts and posts from organisations. This filter removed a large amount of data, which were categorised as 'general news, magazines' (95,986), 'regional newspaper' (54,491), 'national newspaper' (20,112), 'TV, radio' (13,160), 'professional news, industry news' (7,723), 'news agency' (7,049), 'institution, government' (4,627), 'company' (369), 'Dailymotion' (32), 'institution, professional organisation, syndicate, association, NGO' (4).

64,409

5. Individual authors / accounts
To further refine our data to focus on individuals, we also excluded specific news and organisational accounts. This involved manually filtering out accounts (with at least 2 posts in the dataset) based on the name of the account. This had a relatively small impact on the data as no individual account had more than 50 posts in the dataset.

63,525

6. Reposts
The initial dataset excluded identical reposts (quoted retweets, which involve a comment on the reposted content, were included). Reposts can add a huge volume of repeated posts to a dataset, making it more difficult to clean. Once the dataset was filtered and the majority of posts appeared relevant from a qualitative review, we added the reposts (73,427) back in to be able to size the conversation about each topic.

136,645

As a result of these decisions, the initial sample of more than 2 million posts became a more focused, filtered sample of 136,645 posts. Another set of researchers might make different decisions at each stage that may also be defensible, which often explains why social media analysis is hard to replicate.

6 Reliability and prediction

In this chapter we explore the extent to which data from social media sources can be considered reliable. In the context of this research, reliability is measured by the extent to which repeating the same methods produces the same result, which is an important building block in scientific research. We examine two key facets – replication of data collection and replication of data analysis.

The chapter also considers the power of social media to predict attitudes and behaviours in the wider world, using case studies including elections and transmissible diseases.

6.1 Reliability

Replicating data collection

The nature of social data poses the first issue for reliable research and some experts felt this made it impossible for social media data to be reliable. Social data is highly mutable; posts can be withdrawn, edited or updated by authors. This means that between any two points in time there is the potential for reported attitudes or behaviours to shift, making it a challenge to return similar results from the same method. This is also true of opinion and attitudes in qualitative research and surveys, yet as these methods are person-based (rather than based on volume of posts, regardless of the number of people involved) the scope for change is smaller.

"You cannot enter the same data source twice - each time you do you will get a different snapshot. You cannot tell the "truth" with social media data, but can open lines of enquiry. It is not accurate but it is useful"

Private sector analyst

Of course, where the data from a query is downloaded and fixed there is the potential to design queries iteratively, which is a positive development for the utility of social data research. Some experts held a more nuanced view on the capability for social data to be reliable, with a view that the topic under consideration, the culture of the platform, and the volume of data were important factors in determining the reliability of social data:

- **Topic**: Broadly, social data on uncontroversial topics, where the author did not have a personal stake in what they were saying, was thought more likely to be reliable – examples provided included product reviews and discussion of roads in the UK.

- **Volume**: Reflecting a common view about social media explored elsewhere, some suggested that higher volume conversation topics would be more reliable as a wider range of perspectives were likely to be included. This edged towards a data science perspective – the higher the volume and the wider the range of sources, the more confident a researcher can be in the reliability of their data. However, as discussed elsewhere in this report, it is not analytically sound to assume a correlation between higher volume and a wider range of perspectives.

- **Platform**: Platform effects were also felt strongly. In part this was related to the volume point above; larger platforms are more likely to offer a level of reliability. But it also related to the nature of the platform, for instance data on an open platform such as Twitter might be more reliable than that on an anonymous social network, such as Whisper. Yet, this suggestion is highly dependent on the social context of the platform and the topic under analysis.

There are contradictions in expert opinion here – for instance, politics is a high salience topic (with greater volume) but one that can have a high level of personal involvement and controversy. Open platforms might be more reliable as they have lower barriers to entry, but this also facilitates a higher level of bots and advertising which has significant impacts on the reliability of data. This suggests that the concept of reliability for social media research remains underdeveloped.

A broader issue raised by experts relates to the structure of the social media sector, specifically data ownership. The long term, reliable, public opinion research that can be provided by surveys is built on collecting and storing a dataset so it can be returned to in future. By contrast, as the networks retain ownership of the data, long-term reliability of social media data is a major uncertainty. On many sites it is not possible to look back at past social media data – it has to be collected at the time. As the field is new longer-term questions around data have not yet been answered. Experts expressed concerns that reliability could be impeded by: changes in algorithms by the platform or the analysis software (see below); the shift of target audiences to new platforms requiring incorporation into research, and changes in legal ownership of communication networks. What happens to social data if a social network goes bust?

Replicating data analysis

Replicating analysis is a significant challenge for existing social media analysis. Existing literature suggests a poor record, with some studies failing to replicate the results from more than half of the Twitter analyses they attempt.[20] The expert group agreed, highlighting that different analytical platforms might return different results from analysis of the same topic, using the same social networks.

Again, this issue finds its counterpart in traditional research methods – we expect to see slight differences in the results returned by different polling companies covering the same topic, such as elections. However, these survey 'house effects' are well studied,[21] while in social data there remains little analysis or understanding of the difference in results returned by different analysis tools. The expectation was that running the same query on the same networks, but using different analysis software, would return different results due to internal cleaning steps each takes with raw social data – a factor identified in the literature review.[22]

> "[A 1% Twitter API sample] may or may not be representative of the rest of activity on Twitter, and there is no easy way for researchers to tell without conducting studies multiple times simultaneously.
>
> Studies have shown up to 96% similarity between the content two different researchers might receive with the same query to Twitter's streaming API… However, other studies have found that the API produces significant error compared to Twitter's firehose API (providing 100% of tweets) when used to correlate with real world events such as flu incidence."
>
> T. Colley et al. (2020)

Even where the differences between the same search query run twice are minimal, experts noted a lack of clarity in research methods: the steps taken between harvesting raw data and the publication of results are often not recorded in detail, which impedes peer review. The literature review reaches a similar conclusion: that given a specific dataset and research question, two different research teams can

[20] See reference 228 in the project literature review
[21] https://sotonpolitics.org/2019/09/10/house-effects-and-how-to-read-the-polling-tea-leaves/
[22] T. Colley et al (2020). Pp 54

make a series of defensible methodological choices of what to analyse and how, and they can come to completely different outcomes. Experts suggested that consistency and transparency in platform usage, API settings, finalised search strings, and filters, could improve replicability– even before the choice of analytical tools was considered.

Experts felt that much research lacks a robust theoretical framework that links what happens online with offline behaviour which can be tested with the data. In many examples, automated platforms are used to run through large volumes of data to find any positive result – meaning that any result may be just due to chance. An equivalent in survey research might be reporting all differences that are statistically significant regardless of the context of the research question.

"The reliability of social media analysis is made worse because we have a tendency to collect high volumes of data and then run lots of analysis to see what sticks… [we are] doing this because we can, not because it is based on a research question or hypothesis"

Public sector analyst

A comparison can be made with qualitative research: an ethnography is not expected to return the same result twice even if it uses the same discussion guide, moderator and participant. The conversation would be changing constantly and often the researcher is expected to deal with data collection issues as they emerge. Nevertheless, having an established framework, especially one tied to external data sources – helps to ground the research and shape the discussion.

Experts also identified a lack of transparency in tools and methods as a significant impediment to measuring the reliability of social data research. As the automated processes these analytical tools use tend to be 'black boxed' by the software provider, users typically cannot validate the steps used to produce through their results. This fed a common demand among experts for greater transparency from analytical platforms and the adoption of standards across social media data. Allowing third parties to review the steps taken in analysis could build greater replicability in results.

As discussed elsewhere, the approaches analysts take also affect reliability:

- The **type of research question** that is being answered will have its own impact, as it does in all research. Qualitative studies such as ethnography or observation of forums have different standards of reliability than large-scale quantitative analysis of Twitter.
- The **structure of the sector** imposes additional constraints. Experts acknowledged that analytical platform charges and data access costs can shape the choice of network for study, rather than it being a decision based on research quality. This again echoes offline research. Polling is increasingly done using cheaper online platforms when face-to-face or telephone might be preferable.
- **Analyst understanding** of the data and the analysis platform is also key. Ideally the researcher would understand the strengths and limitations of the data they are collecting, the network they are working with, and the platform being used to analyse it.

6.2 Prediction

These factors also feature in discussions of prediction. Predicting key events is a well-established method of promoting research methods, especially political polling. Elections are the highest profile example for prediction which stretches across both methods. Other commonly cited examples for social data include predicting the popularity of new products and films and the spread of seasonal diseases such as flu and norovirus.[23]

A lack of consensus

Prediction is one of the busiest and most adversarial fields in social research. Our literature review identifies claims of predictiveness from social media for elections stretching back to 2009.[24] A similarly large corpus of work refutes claims of prediction, alongside a high volume of commercial publications claiming that their tools are predictive but with limited transparency about their methods. As elsewhere, this lack of transparency is a significant stumbling block to establishing a clear answer on the power of social media analysis.

> "There is little clear and convincing evidence for social media's ability to predict outcomes. Every study claiming that it can predict elections, for example, is contradicted by others saying it cannot come close... Looking back at data retrospectively, at best these studies show that it might have been possible to predict outcomes if we knew what to look for in advance.
>
> Some studies may exaggerate their findings, and academic pressures may be leading to the overreporting of positive findings. Less is known about tests that showed no effects, and few studies' findings are replicated because replication studies look less original."
>
> T. Colley et al. (2020)

The expert group reflected the state of division found in the literature review. We found experts held a wide range of views on the predictive ability of social data. Those working in consumer trends and private sector research were the most positive. Examples of success were closer to what is called 'market sizing' in market research – understanding the likely popularity of new products, or box office success for new releases. Even here some disagreed, though, saying that social data should not be used for quantitative prediction but for understanding the culture of what makes products appealing:

"It is not good for telling you how many people buy Kinder Eggs, but the raw data will get you closer to the consumer to understand their thoughts"

Private sector analyst

Others said that prediction using social media was still highly inexact and that those success cases had been retrospectively cherry-picked, ignoring a far larger body of negative results (an issue for academic study more widely):

"It almost never works. Because most of the time, you are over-fitting to a particular type of answer from your dataset"

Social media academic

[23] https://www.bbc.co.uk/news/blogs-trending-38227094
[24] T. Colley et al. (2020). Pp 42

Examples of prediction

Google's flu modelling[25] is a good case study as some experts cited it as an example of the power of social media to predict, while others used it as an example of its challenges. Initially, the use of Google search data to predict the incidence of winter flu in different areas of the US worked reasonably well. However, during this experiment, and without the knowledge of the researchers, Google changed its search algorithm so that searches returned a larger number of health-related sites. This led to the model making huge overstatements of the prevalence of winter flu, removing its predictive power. Readings from this less accurate model could still be correlated with flu outbreaks, but as the bar for success was simply a flu outbreak emerging, any signal from the flu trends, however weak, can in hindsight be interpreted as proof of a successful prediction.

Again, experts suggested that social media is more useful in tightly constrained circumstances where researchers are looking for tactical insights, rather than where the goal is to predict broader, strategic shifts. Again, experts suggested that social media is more useful in tightly constrained circumstances where researchers are looking for tactical insights, rather than where the goal is to predict broader, strategic shifts. Predicting shifts in flu incidence during an outbreak will be easier than predicting the COVID-19 pandemic. And as research has shown, predicting the outset of revolution is rarely possible, but forecasting how protests may develop once they have started may be more doable.

Elections are another area where social data is more commonly used for prediction. They have many of the characteristics required that make a social data prediction feasible – a tightly bounded contest that contains limited unpredictability. In many countries' elections there are frequently just two feasible outcomes: the government is either deposed or retains its position. However, as explored in the literature review, there are few clear-cut examples of social media correctly predicting electoral outcomes.[26]

> "Studies into the cases of the 2008 and 2010 US elections found that there was no correlation between analysis results and electoral outcomes. Any predictions were found to be 'no better than chance', and sometimes worse… Election prediction articles have claimed that they can rank party vote share effectively, but this is hardly a substantive finding, particularly in countries with one dominant party."
>
> T. Colley et al. (2020)

Where there are examples, experts noted that social media was never the most predictive method. In most elections, polling is more accurate and as readily available. Perhaps most importantly, polling predictions are issued ahead of an election while most examples of social media 'predictions' are issued after the fact, with researchers retrofitting the data to a known outcome. Polling has appeared to fare differently over the past decade, with the UK in 2015 seen as the year of a significant "miss" while the 2017 and 2019 elections are seen as "hits". Its accuracy has been relatively consistent over time, however,[27] compared with social media-based election predictions which often produce errors of 40 to 50 per cent, even when retrofitting findings to the eventual result. In countries where polling and survey infrastructure is well established, these have almost always proved more predictive than social media data.

[25] https://www.google.org/flutrends/about/
[26] T. Colley et al. (2020)
[27] https://eprints.soton.ac.uk/421260/1/JenningsWlezien_PollingErrors.pdf

6.3 Conclusion

The expert view on the reliability of social media data was split, with no clear consensus. From a data collection perspective, it was felt that reliability varied by the topic of discussion, the volume of data available and the platform on which the conversation was being held. Yet the expert group disagreed on the details. Some said that higher salience topics such as politics would be more reliable due to the increased likelihood of people raising their personal opinions. Others felt that these topics were more open to interference from bots and trolls, reducing the reliability of the data. These may both be the case. Either way it is hard to generalise, because whether people discuss politics publicly depends on which culture and platform one is studying. The platform hosting the data also affects reliability if data access is difficult or how they screen their APIs is unclear.

There was greater agreement on the role of data analysis in reliability. The general view was that opacity from 'black box' analytics of existing research platforms hampered the ability to produce replicable research. Both experts and the literature review cited concerns that researchers working in the area were also being opaque about how they collected, sorted, and analysed their data.

On the possibility of using social media data for prediction, the expert group tended to feel that the case was yet to be proved. Existing evidence suggests social data is less accurate in predicting electoral outcomes than polling. There is also less understanding of what makes a good (or bad) prediction from social data. Experts perceived that those cases of 'successful' prediction had been cherry-picked from a wider range of failures. Many were not genuinely predictive, and instead were reverse-engineered to show that one could have predicted a past event if one knew which variables to look for.

7 Defining the value of social media for online audience analysis

Having established the credentials (and limitations) of social media data more broadly, this chapter considers the specific value social media data brings to the research community, and to evidenced based policy more broadly.

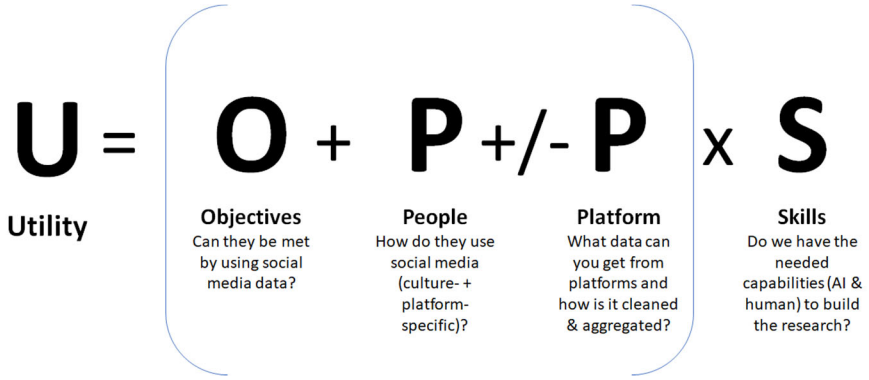
7.1 Assessing 'utility' in social media data

It would be wrong to assume that the utility of social media data is driven purely by the extent to which analysis is representative, reliable and robust. Furthermore, it is important to remember that no single data source can claim to be perfectly accurate, and that all sources contain errors. The evidence is that the use of social media data for online audience analysis can have practical utility within specific contexts, and that this utility can be enhanced through the quality of project execution.

Understanding the dynamic in which social media analysis operates, and the limitations of different approaches, is key to assessing the value of online audience analysis. Our recommendation based on the evidence is that it is wise to initially assume that raw social media data is likely to be unrepresentative of the target population's opinions and behaviours, but that the usefulness of social media data can improve the more skilled human analysts are in addressing the large number of factors that affect data quality.

For ease of reference, we propose that the overall utility of social media data can be summarised as a formula, where judgements relating to the representativeness, robustness and reliability are considered alongside specific objectives and contexts; and where risks can be mitigated by analysts.

Figure 7.1: Key considerations for assessing the utility of social media data



Within this model there are 4 overarching factors:

- **O** is for **objectives**. The first consideration is to identify the strategic objectives of the project. What do we want to do? (Broader strategic objectives) and What do we need to know to do it? (Research objectives). The next stage is to review which methods are available to gather evidence relevant to these objectives; and to consider whether objectives can be met using social media data. Social media analysis should not be conducting just because it is technically possible.

- The first **P** is for **People**. This should seek to go beyond traditional debates of representativeness as defined by who uses social media. Further consideration is needed of the broader cultural

context that shapes who uses social media and how. Political context and the role of the state is also key, in shaping how freely people post online.

- The second **P** is for **Platforms**. These can be a help or hindrance to the overall equation depending on their performance. Key considerations in how data can be obtained; what data and metrics can be collected from the social media platforms, and how do social listening platforms and aggregators collate and analyse the data?

- **S** is for **Skills**. Skills are a multiplier because they can both mitigate against risks and enhance the benefits. Analysts are required to make large numbers of justifiable decisions during data collection and analysis. The quality and transparency of these decisions directly impacts on the robustness and reliability of the findings. If an analyst has limited skills, whatever they do is likely to have limited utility. Skills partly reflects the AI capabilities available to conduct analysis, but it mainly concerns researchers' skills in building research and analysis that meets the objectives.

We propose that the combination of these factors is more of an art than a science. Within each factor, there are signals that are likely to lead to a higher degree of representativeness, robustness and reliability. Table 7.1 below provides an overview of how these signals correspond to our overall formula of utility. However, it will be for each organisation and each analyst to trade off the strengths and limitations of using social media for any given purpose. It may be, that despite its limitations, social media research will still yield significant value. The framework presented here provides a useful reference point for making judgements about quality of findings from social media research, and should be used to guide how much weight should be given to results in decision-making.

It is important to further note that in projects that align more closer with 'signals of low quality' (for example, illiberal political systems, where internet penetration may be lower, where there is extensive online manipulation, where there may be limited other sources of data), these characteristics can be a relative strength of social media data. For example:

- The speed and agility of social media data is extremely attractive in a fast-paced environment where the use of high-quality traditional methods is not a practical option. In such circumstances, the lack of alternative viable methods or data sources will enhance makes social media data more valuable in relative terms.

- The exploration of taboo topic areas can also be, in the right context, an advantage for social media, for some topics social media displays opinions and behaviours that people would be unwilling to share publicly. This is particularly relevant for understanding extremist communication, even if the link between communication and behaviour is harder to determine.

- High manipulation of content through use of bots and hostile actors may not be issue where it is the very existence of the bots and hostile actions that is under investigation.

This further demonstrates the need to judge the utility of social media analysis in each given context, rather than making blanket assumptions about its value.

Table 7.1: Key considerations and signals in defining quality of social media data

	Key questions / considerations	Signals of high quality	Signals of low quality
Objectives	• What do we want to do? (Broader strategic objectives) • What do we need to know to do it? (Research objectives) • What other data is already available? • What other research methods are viable?	• Topic of interest that is not taboo, so that people are willing to share views freely • Multiple other data sources for triangulation and validation • Where the population of interest is those who are online (often in developing markets this aligns to those of higher social grade) • Short term situations less likely to invalidated by shifts in platform use or design • Where the group of interest is not expected to align with broader representative public opinion	• High levels of social desirability bias within topic of interest where online opinion is unlikely to be widely representative • Topic is not routinely discussed online • Limited alternative data sources for triangulation or verification • Where the aim seeks to quantify public opinion at large
People	• Demographics and other population characteristics which shape access to and use of social media • The broader cultural context in which people are using social media data, including extent of freedom of speech vs censorship • Presence of bots, fake accounts and other adversarial strategic actors • Dominance of prolific users, influences and institutions	• Widespread access to the internet among the target group, and high levels of use of social media • Liberal democracy in which freedom of expression is widely established • Well understood manipulation of content by adversaries	• Limited internet connectivity and low social media penetration • Illiberal political systems involving restrictions on internet use and censorship • Unknown extent of online manipulation by adversaries
Platforms	• Structure of platform and how this shapes content (e.g. anonymity, encryption, features) • Availability of data for research – how content is aggregated, analysed and served • Role of data aggregators in harvesting and enriching data	• Platform granting extensive access, in sufficient detail, to significant and representative sample of content • Quality of and access to data is reliable and transparent	• Limited data available • Unreliable quality of data – lack of transparency, changeable aggregation techniques, or incomplete data
Skills	• What are the software capabilities and limitations for analysis (including use of AI)? • What are the capabilities of the human analyst? • Understanding of how to improve reliability and robustness?	• High levels of transparency in steps of both data collection and analysis, • Quality assurance and peer review • Quantitative and qualitative analysis is conducted in tandem	• Linguistic barriers and/or poor cultural understanding • Low levels of transparency and quality assurance • Black box analytics which cannot be assessed or replicated

7.2 Illustrative case study: capturing public opinion during the COVID-19 pandemic

This case study demonstrates that although data collected on social media may not be precisely representative, it can still offer a number of benefits compared to other traditional methods.

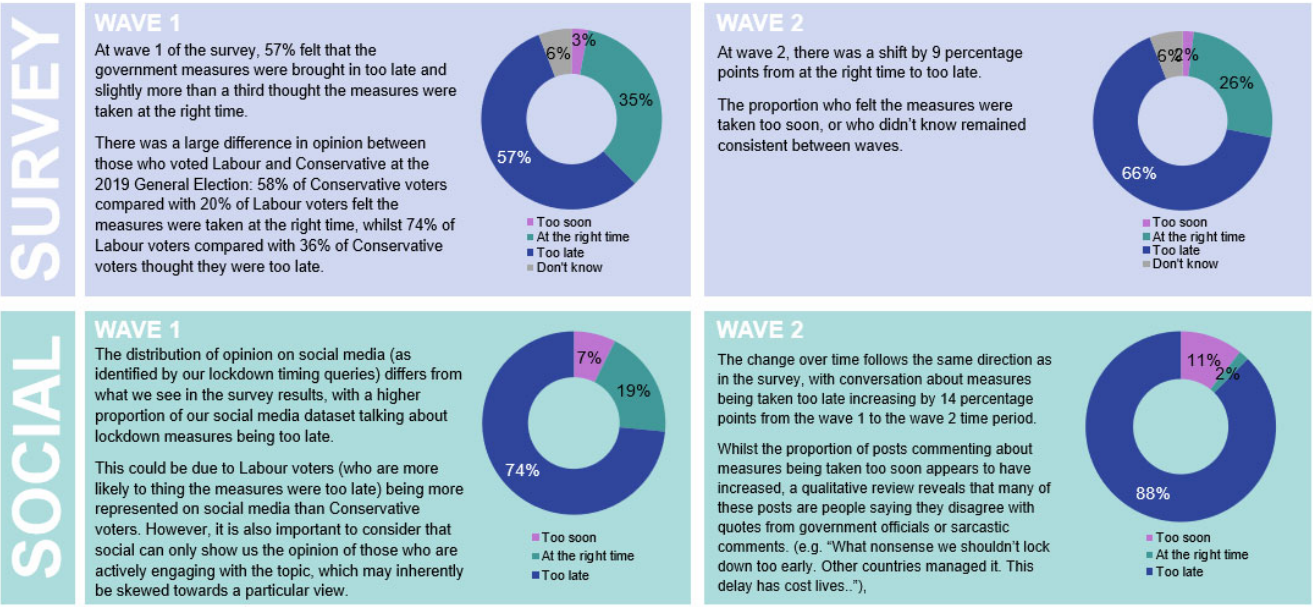
As part of this project, the authors conducted a live case study to examine the value of social media research to better understand public opinion surrounding the COVID-19 pandemic. The case study explored public opinion across three fronts: i) level of concern for yourself and the country; ii) views on whether government intervention was timely; and iii) views towards the use of the military to support the response to the pandemic. Alongside the social media analysis, two waves of survey research were conducted online – seeking to provide a comparison to a nationally representative sample of the online population in the UK. The findings demonstrate differences in results but also the value of social media analysis.

There are clear differences in comparisons of overall sentiment of the nation; however, trends within the data showed similar swings in public opinion over time. For example, as shown in figure 7.2 below, at wave 1, the proportion of believe that action to place the UK in lockdown took place too late is higher when estimated on social media than when estimated through the online survey (74% compared to 54%). This difference may be attributed in part to the political leaning measured on Twitter, the largest source within this data set; however, it should also be noted the survey research also found that those who are more active on social media[28] were more likely to hold the view that government acted too late in dealing with the COVID-19 pandemic. Either way, it is likely that the social media data represents a skew in the extent to which it truly represents public opinion of all in the UK. This is further evidenced when exploring support for the use of Armed Forces in the pandemic response, where support is weakest among younger adults who are most likely to be using social media. Yet, accounting for these initial differences, the shift in sentiment overtime is broadly in line and suggests that social media data may be a useful indicator of some shifts in public opinion.[29]

[28] through use of private messaging, sharing content or following companies they like

[29] Survey saw a rise of 11 percentage points (from 57% in wave 1 to 66% in wave 2) in proportion saying government measures were too late. Social media data raw a rise of 14 percentage points on the same measure (from 74% to 88%).

Figure 7.2: Comparison of social media and survey findings of COVID-19 case study



*It is important to note that the **social media analysis counts posts, rather than individuals:** whilst survey data counts each respondent only once, a dataset of social media posts could be made up of a large number of posts from a small number of posters. For example, in our dataset a publisher name was present for 26,183 Twitter and social network posts, so we could see that these posts came from 20,632 different individual accounts. The most prolific publisher accounted for 58 posts in the data. However, not all social media platforms tag posts with a name, and some users choose to post anonymously.*

Base: Adults aged 18-75 in Great Britain (1,069), Fieldwork dates April 2020

To this end, a key strength of survey data is subgroup analysis. Surveys are able to accurately capture a wide range of demographics directly from respondents; in contrast, social media has to rely on best estimates, which often have poor levels of accuracy and vary by platform. The exception to this is location. Most surveys have too few sampling points to generate robust regional analysis; in contrast some social media platforms can, in theory, provide detailed geographical analysis. For example, the case study was able to uncover levels of personal concern about the virus across major cities of the UK. However, location data on social media should be treated with caution, as not every platform provides this level of detail and research indicates that the majority of users do not turn their location on.

In its favour, the social media analysis provided rich insight not just to what proportion of people were concerned for themselves and their country, but more valuably why, and in what context. Working with the text analytics tool Insius, human analysts identified five key concerns relating to family, work, health, children and the future, and the relative importance of these. This included real in-the moment, unprompted, testimony of emotions and experiences, rather than relying on survey recall. However, it proved to be difficult to assign value to some opinion where this was not expressed frequently or consistently. For example, automated analysis was confused and contradictory when within the same posts users cited Matt Hancock's claim that the UK had 'lockdown was at the right time' alongside their own criticism of his position. Only manual analysis could correctly assign this opinion. In a further example, when asked whether they support using Armed Forces to help the pandemic response, 76% of the public gave a positive response within the survey. However, as many as 78% of the posts were assigned to 'neutral sentiment' on this issue within the social media data, 14% negative and 8% positive. This is clearly an inadequate assessment of overall public sentiment, on a topic that yielded comparatively little discussion on social media. Such high measures of 'neutral' sentiment are typical of sentiment analysis currently, showing how this method needs to become far more refined before being consistently useful analytically.

Finally, the dynamic of the social media data was clearly driven by events and media coverage. This can be a positive, providing precise evidence as to the exact moments of public reaction and the impact of key events on public opinion – something which is notoriously difficult to do within survey research. However, media interest can at times cloud underlying opinion. For example, on the day after the UK Chief of Defence Staff participated in the daily coronavirus briefing, a bigger topic of conversation seems to be reports of the army using insect repellent as protection for coronavirus.

In summary, it would be wise to conclude that the social media data does not reflect the same proportions of public opinion, and should not be used to make concrete conclusions about the exact number of the population that hold a particular view. The utility of this data would be further reduced if analysis purely relied on automation and received little engagement or oversight from analysts. However, if the objectives were to track changes over time, or to develop a deeper understanding of the range of experience or opinion at a given moment, or to seek a better understanding of how events interact with public opinion, then the value of the data becomes greater. This utility is further enhanced if analysts are able to correctly identify limitations, and develop transparent collection and analytical processes that stand up to review and scrutiny.

7.3 Conclusion

It is impossible to judge the value of social media outside of a given context. Social media data has the potential to offer significant value, yet its utility is variable. For example, in some situations it may be the only source of insight into populations available; in others its data may not be relevant to the target audience of interest. Whilst the high volume and low cost of social media data makes investment in collecting and analysing it worthwhile, it is important not to set unrealistic expectations of what it can achieve, and what the applications are for making evidence based policy.

Social media data should be used as part of a flexible and context specific research programme, which draws upon as broad a range of online and offline data sources as possible. However, it should not be assumed that every social media project will yield the same standard of data and analysis.

It is also important to consider the many other ways social media can be useful beyond seeking to directly estimate the exact proportion of a population that hold set opinions, behaviours and motivations. For example, it should also be considered as a qualitative insight tool, to help provide depth of insight into specific audiences and topics of interest. It can also be used to inform question design, gather real time knowledge of new situations and identify gaps in need of further research.

The successful application of social media data relies on empowering analysts to marry the strengths of social media data to the right strategic communication objectives. It also requires informed judgement about how much weight to place on the findings based on the known limitations in any given context.

8 The futures of social media research

This chapter presents the full detail of the approach used to generate scenarios of the future of the social media sector in 2025. It details the key drivers uncovered during analysis of the expert interviews and the results of a 'Delphi' panel survey that identified high impact shifts that may occur between now and the middle of the decade. The implications of the three potential futures – Octopus Corporations, Digital Fortresses and the Curated Internet – for social media and online audience analysis are discussed in depth, followed by a conclusion covering the risks and opportunities each poses for analysts.

8.1 The context for the scenarios

The scenarios have been derived from analysis that identified the key drivers and trends acting on the sector currently. Drivers are the larger forces acting in society; as they act on the medium term, understanding their current trajectory can give an understanding of how they will develop into the future. The seven drivers were derived from analysis of the 30 expert interviews conducted in this research project. They fall within three broad themes which appear to define the shape of the sector currently: how controlled or open cyberspace will become; how far and how quickly new technological innovations reach mainstream adoption; and the extent to which there will be international co-operation – or fragmentation – in managing citizens' access to the internet.

Table 8.1: Key drivers for the social media sector

Driver themes	Driver	Main oppositions	Key questions
Controlled or Open cyberspace?	**Network access models**	Walled VS open	User data is the principal source of income for social networks. Will a more open model continue, or will all social networks protect their data behind paywalls?
	Public attitudes towards online content	Anxiety VS apathy	Public opinion towards how their personal data is used by govt/businesses varies between anxiety and apathy (acceptance) worldwide. Will this balance continue, or will we see lesser/greater demands for privacy from users?
	Regulation of the internet	Tight VS loose	Many governments have signalled an intention to control online harms/fake news through legislation. Will these regulations continue and be enforced, or will they be unenforceable?
Tech revolution or evolution?	**Technology: innovation and adoption**	Advancing VS static	Will smartphones remain the main device used to access social networks, and will the adoption of 'Internet of things' technology be widespread or remain centred on a small elite?
	Research method development	Traditional qual/quant VS new computational methods	Will new computational methods and algorithms change how researchers deal with social media or will it continue to be viewed through a traditional "qual vs quant" research lens?
Fragmented or convergent?	**The future of big networks**	Monopoly VS hollowed out	Will existing big networks continue to be the networks people use for social interactions, or will this fragment to other smaller and specialised networks?

	International policy co-ordination	Global VS fragmented	Regulation of tech is breaking into competitive global blocs – EU vs US vs China. Will one model win out or will multipolarity increase?

Having developed these seven drivers as a framework for constructing potential futures a second wave of primary research was conducted with social media experts to test their validity, impact and likelihood, in the form of an online survey. During this survey, 28 statements were tested which mapped the ends of the oppositions within each driver (the full list is provided in the Appendices).

The survey asked participants three main questions about each statement:

- How **likely** they felt it was to occur by 2025.
- How **impactful** it would be for the social media sector were it to occur.
- The level of **uncertainty** they had about the impact this future would have.

They were also asked which of the seven drivers were likely to be the most influential in shaping the future of social media over the next five years.

The results of the survey outline how expert consensus differs across the seven drivers; these are detailed below.

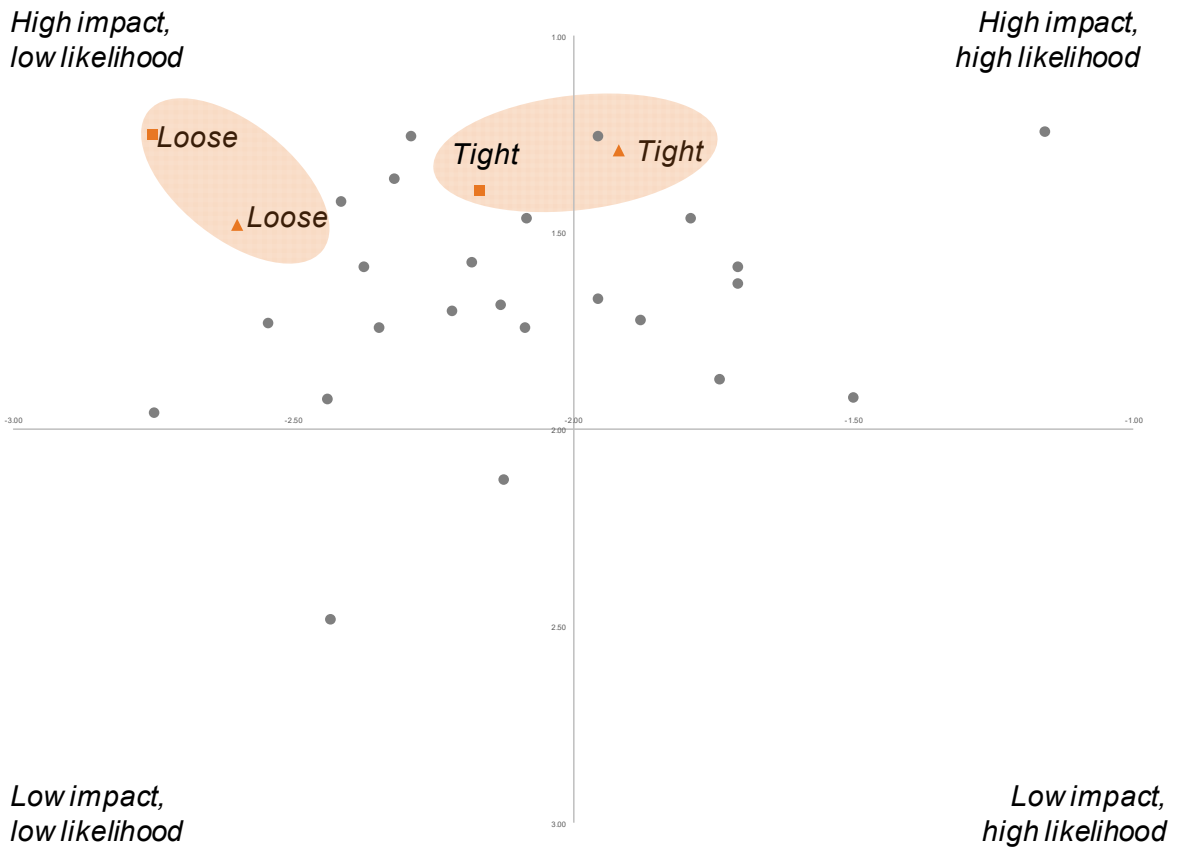
Regulation of the internet

How governments seek to regulate topics such as fake news, disinformation and harmful content on social media was rated as the most important axis for determining the immediate future of the sector. This likely reflects the current focus on regulation across a range of countries, which may result in diverging rules in countries across Europe and North America – but also the potential for countries to shift towards models based on the very different regulatory regimes in place in autocratic regimes.

In the survey, statements associated with a higher level of regulation were considered more likely than lower levels of regulation: specifically, the prospect of social media networks being treated as publishers rather than platforms (echoing discussion in the US about Section 230 of the Communications Decency Act) was rated as both higher likelihood and higher impact.

However, the prospects of looser and tighter regulation of social networks were both considered to be high impact; specifically, governments moving away from trying to regulate online life. This suggests that regardless of the type of regulatory change we see, the extent and impacts on the current picture will be similarly large.

Figure 8.1: Regulatory driver – impact and likelihood grid



Looser regulation statements	Tighter regulation statements
Square: Governments will have abandoned attempts to regulate disinformation and fake news online	Square: Those wishing to sign up to any social network will be required to provide proof of identity
Triangle: The public will only trust in information which has come through offline sources, especially personal connections	Triangle: Social media networks will be treated as publishers, with responsibility for all content on their sites

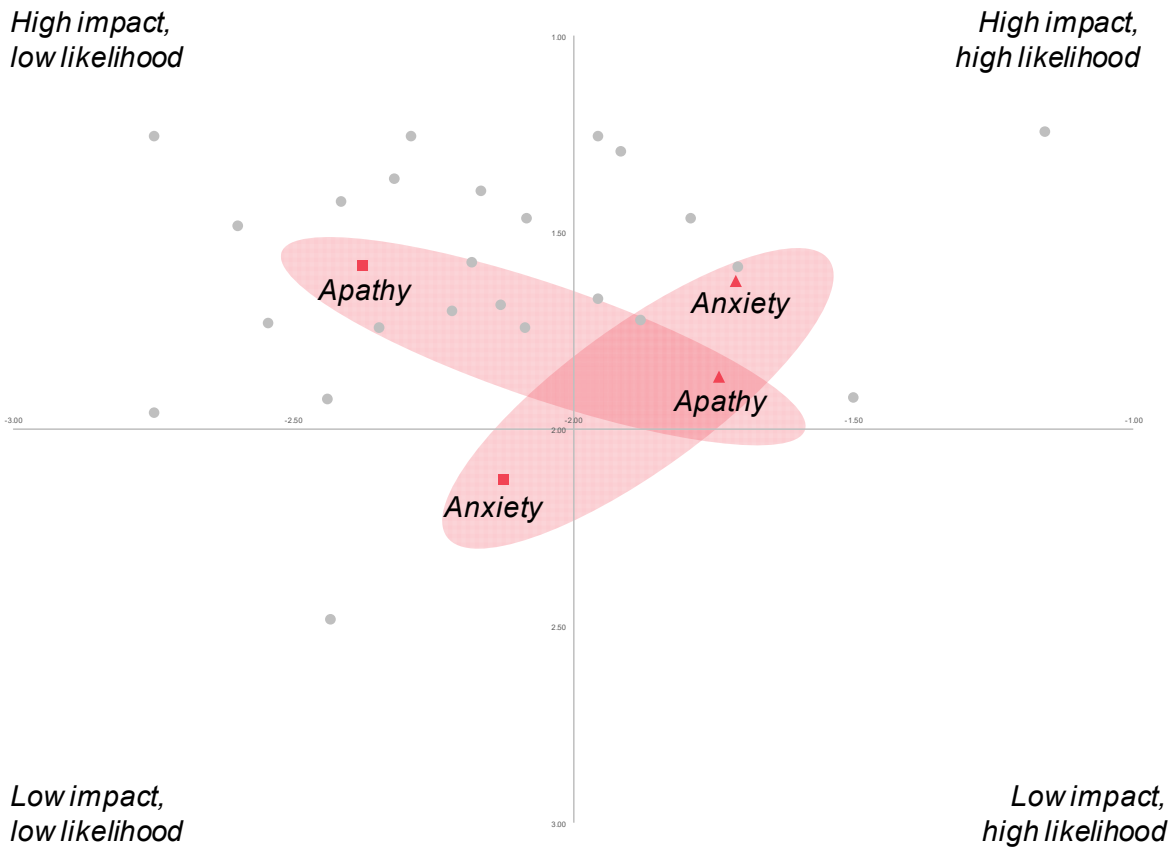
Public attitudes towards online content

Public attitudes to the internet – measured by trust in online information, the use of encryption and the representativeness of social networks – were seen as the second most important driver for the future of the sector.

Attitudes in this space were split between two poles seen in other research into this area. The first is anxiety – concern about who is using or viewing personal data and an attempt to control or limit access. The other is apathy – a level of rational ignorance around who holds personal data and how it is used, as well as a fatalistic attitude around how far personal data can be protected in any case.

The statements tested reveal a mixed picture among experts: widespread use of encryption by the public and the potential of fake news to degrade public trust in other information sources are both considered higher impact and likelihood but cover opposite ends of the spectrum. This suggests a lack of consensus among experts on how this driver might play out, meaning that it will remain a contested area into the near future. All scenarios for this time scale will contain conflicting elements of both drivers

Figure 8.2: Public attitudes driver – impact and likelihood grid



Data Apathy statements	Data Anxiety statements
Square: The public will expect social networks to pay for access to their own data Triangle: The increasingly low credibility of information available on social media will have reduced public trust in other information sources	Square: Publicly available social media conversations will be less representative of the opinions in a given population than it is now Triangle: Most social media interactions will occur through encrypted or secure channels

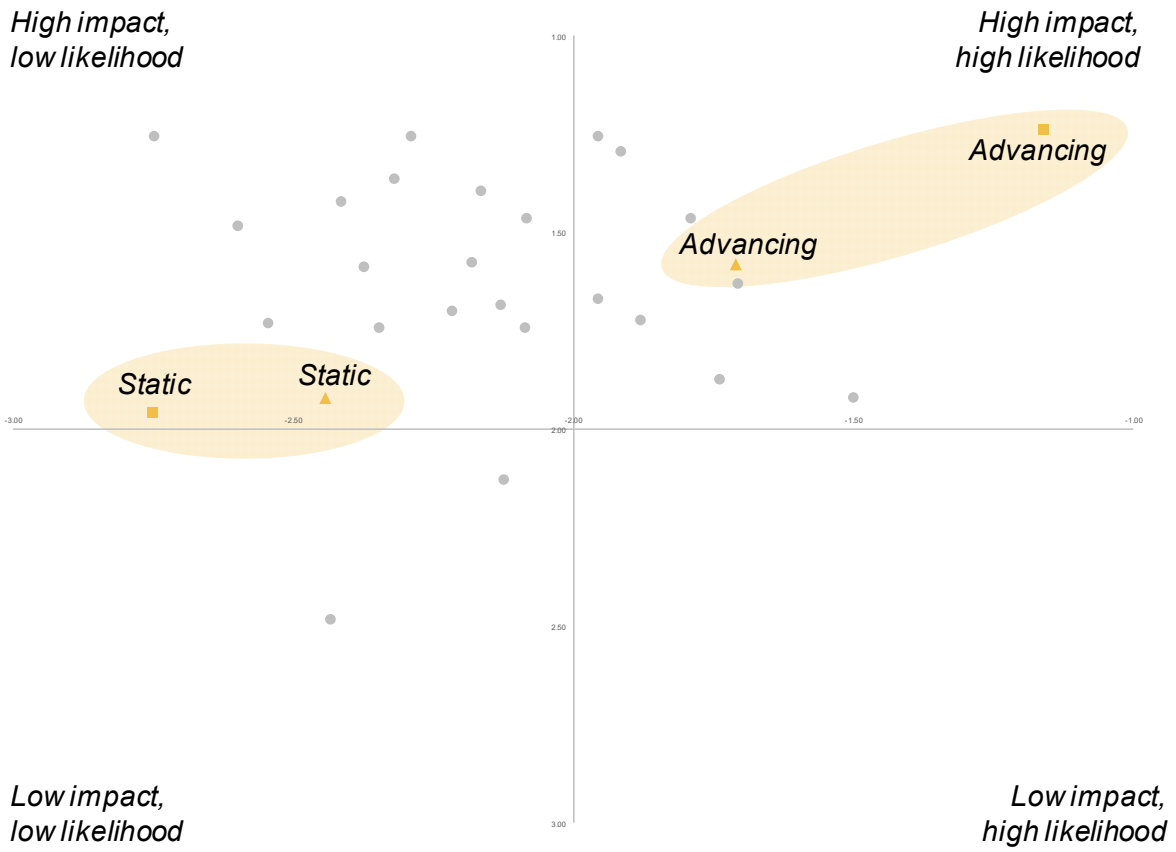
Technology adoption and innovation

The third most powerful driver for the sector was the level of technological innovation we will see and the extent to which these innovations enter mainstream usage.

Here the experts told a clear story – continued advances in technology such as a move away from text- and image-based social networking to video and the continued spread of social networks into new services are not only highly likely but will also be high impact. A world where the conveyor belt of technological progress slows or stalls would also be high impact but are considered among the least likely of all 28 statements.

As a result, all scenarios should include an assumption about the continued pace of innovation, with a slowing in advances becoming a ‘wildcard’ – a highly unlikely event which would have a systemic impact on all aspects should it occur.

Figure 8.3: Technology driver – impact and likelihood grid



Static technology statements	Advancing technology statements
Square: Social media application optimisation will focus on ensuring battery and device preservation rather than best interface	Square: Social media networks will increasingly offer other services to their users, including payments, healthcare and transport
Triangle: Social media networks will keep updating their interfaces, but there will be no major changes to their core offer	Triangle: The most popular social networks will be based around video sharing rather than words and images

The expert view on the remaining four drivers are listed below

International policy co-ordination

The statements in this driver were all considered high impact but low likelihood. This suggests that there are unlikely to be any significant developments in the direction this driver might take over the life of the scenarios considered in this analysis.

Network access models

More open models were considered marginally more likely and impactful, but the relationship was not clear, implying that in all scenarios to 2025 a mixed system would remain.

The future of big networks

The future of big social networks was considered by experts to be one of the least impactful drivers for the sector to 2025 – itself a notable finding that suggests they are no longer seen as being at the forefront of social media. Statements associated with a more monopolistic future were seen as having a

bigger potential impact on the sector but views on their likelihood were more varied, suggesting that this is another axis unlikely to have a clear direction by 2025.

Research methods development

This driver was seen as having the least impact on the sector overall, but a consensus emerged that the emergence of new social data research methods was both high likelihood and high impact, suggesting this as an area for significant innovation and turnover in tools and methods. However, one element of the traditional axis – that those performing social analysis will increasingly be generalists rather than specialists – was also considered likely, suggesting that this area will remain closely tied to existing market and opinion research.

The impact of COVID-19

Fieldwork for the Delphi study was conducted over March and April 2020, a period covering the most stringent lockdown period of the first wave for many countries in Europe and Asia. While it is unclear which of the changes wrought by the pandemic will last into the medium and long term, the experts were asked to reflect on what this period might mean for the seven drivers as this large external shock has implications for all futures of the sector.

Expert feedback on this question highlighted public attitudes as the driver most likely to be significantly altered by the experience of the pandemic. Under isolation and quarantine policies many have experienced a greater reliance on social networks and digital platforms to keep in touch with friends and family as well as to work and carry out everyday tasks, which may lead to greater public acceptance of these uses and a more positive view of the companies providing the service. Further, the use of smartphone tracking apps in test, track and isolate services to combat the disease has wide public acceptance and may make people receptive to sharing their data with other organisations more widely.

Other potentially significant shifts included an acceleration of tighter regulation in response to false information about the disease circulating on social media and even fast adoption of new technologies as more of life is pushed online. A table summarising experts' reflections on each driver is provided below.

Table 8.2: Anticipated impacts of Coronavirus on key drivers

Driver	Covid impact	
Regulation of the internet – tighter versus looser	Tighter regulation	Governments around the world will seek to tighten regulation on social networks as the pandemic eases – tackling the spread of disinformation online will be the primary driving factor
Public attitudes towards online content – anxiety versus apathy	**Key point of acceleration**	**Increased reliance on social networks during lockdowns may have a positive effect on public views of technology and the firms themselves. Many people appear willing to be tracked closely using smartphone apps for health reasons**
Technology innovation and adoption – advancing versus static	More of life online	A ratchet effect will push more of life online and not all will return to offline once the pandemic has passed.
International policy co-ordination –	More fragmentation	Policy towards covid has been fragmented across countries, including appetite for the use of tech

global versus fragmented		
Network access models – walled versus open	Public more platform-agnostic	Some are interacting with a wider range of platforms during lockdown and may become more agnostic around brands as they experiment. But many are underpinned by the same (Google or Facebook) login
The future of big social networks – monopoly versus hollowed out	Potential to go either way	Established firms have been able to trial new products with captive audiences; governments are coming to rely on many of them for essential analytics or tech delivery
Research methods development – traditional approaches versus computational innovation	Accelerate – adoption and mixed methods	A Large number of projects have switched to a mixed-methods approach over COVID-19 and this may normalise the use of online and mixed methodology polling more generally, even once face to face research resumes

8.2 The scenarios in detail

This section will review each of the three scenarios in detail, providing background on the key statements and drivers associated with each. It will outline important signals we could expect to see should the world begin to head towards the scenarios and the potential implications for research. We will also consider the possible “wildcards” – extremely high impact and uncertainty events which would immediately reshape the sector were they to occur.

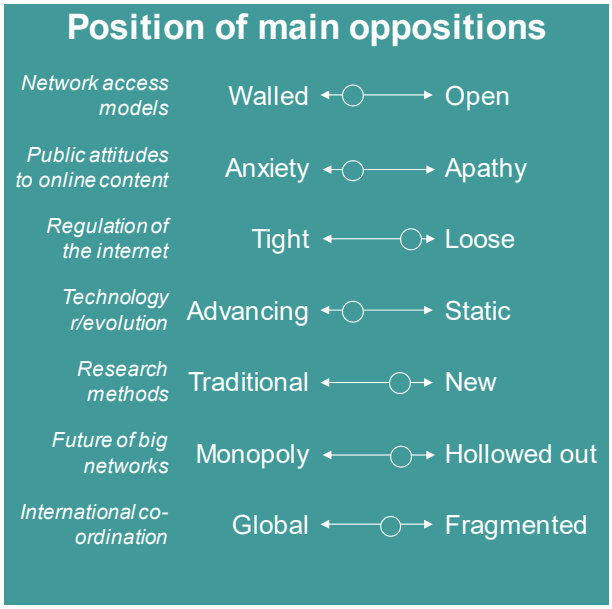
Generating the scenarios

The three scenarios were created using an inductive scenario process, a creative method which relies on the researcher exploring the emerging drivers and issues and understanding how they might fit together into cohesive scenarios. An alternative approach is known as the deductive method, where key drivers are used to structure a grid or framework and scenarios are created at in the quadrants formed by the axes.

An inductive method is more suitable for this project as the expert survey revealed a high level of uncertainty around the likely development of trends – this meant that several of the drivers had near-equal importance. This freer method resulted in three potential futures for 2025, which are detailed below.

Digital Fortresses

In this world, attempts by governments to regulate social media on fake news and online harms, initially given greater impetus following the COVID-19 pandemic, have proved to be impractical, with several new scandals emerging in quick succession during the 2020 US Presidential election and other major events. The online environment is becoming more hostile, driving a fall in public views of the quality and veracity of information available online. It also means that the proportion of the public who are actively concerned about sharing their information online is rising.



At the geopolitical level an increasingly adversarial environment, primarily between China and the US but also within and between other nations, is driving up the level of state-sponsored misinformation and disinformation campaigns. This leads to a new arms race and pushes countries to develop their own resources to build national resilience, rather than work with others.

These shifts are beginning to change the economics for existing, more open platforms, which suffer from a lack of public and regulatory trust. This means there is strong potential for new networks to emerge which play to public demand for security by offering a different, more closed experience which is also more nationally-focussed and adheres to local norms and rules – closer to current social media experiences in China and Russia.

Implications for research

The capacity for representative, reliable and robust research using social data would decline in this scenario, as it would become likely that no single platform can offer a wide enough level of coverage or generate access to a large enough dataset of publicly available data. Heightened public anxiety about the safety of online platforms would have a strong influence on what people post about, especially on public platforms. This would further reduce the potential for collecting views on social media which are reflective of the opinions of the general public or sub-groups within society. In response, research would need to focus on platforms used by specific communities, which could differ dramatically between countries, language groups and regions in terms of the level of data available and the extent of access. Some might be entirely encrypted and inaccessible for any type of research.

Signals to monitor

Some key categories of events that suggest this future might be emerging are listed below:

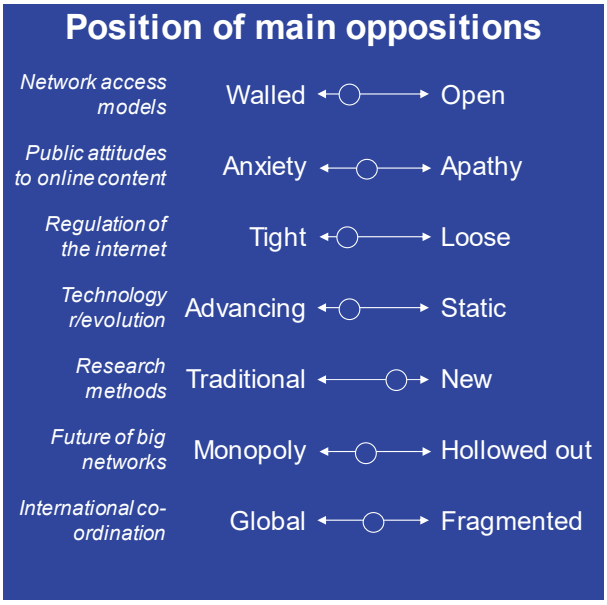
- Government regulations on fake news and online harms are abandoned, watered down or deprioritised.
- Public discourse shifts to hidden and private channels, evidenced by rising popularity of new apps and services that put encryption and privacy at the heart of their advertising and brand.

- Public trust in the information available online falls markedly and the trustworthiness of other media and communications channels drops too.

The Curated Internet

In this scenario, concerted public and government action on tightening regulation of social networks would make the business of running a universal social network more challenging. National requirements and regulation would be mounting up, even if the international space picture is not becoming more fragmented than now.

At the same time, the public appreciation of the benefits social networks can bring will be maintained and begin to grow, with an increasing minority also becoming cognisant of the value of their data.



The response is the rise of 'curated networks' – private communities united by common interests with a more explicitly commercial outlook. Subscription models will be increasingly common and for some powerful users, networks will be paying them to maintain a presence, interact and to give access their data and followers. The wider public will also begin to move between networks, taking their data with them when they leave.

Implications for research

In this more commercial world, network data access requirements and costs will be increasing for researchers. This means there will be greater consideration of the choices of platform used for research projects, with a general impact that research using social data will become more specialised. Quantitative scraping studies that seek a level of representativeness will become uneconomical and replaced by cheap online polling or more in-depth online studies.

Tighter regulations on online harms, fake news and other aspects of speech would drive a greater proportion of political groupings and views into more walled and encrypted forums, reducing the variety of opinions expressed in more open internet settings, reducing the potential for online research to be representative of the range of opinions in a given society.

Signals to monitor

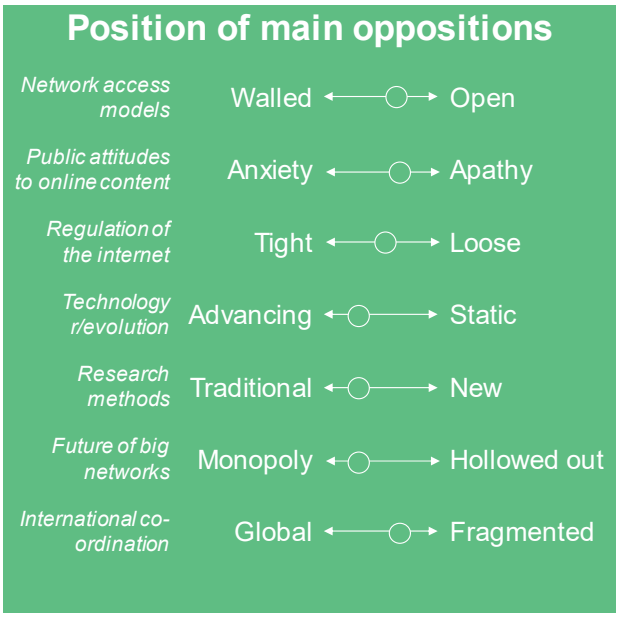
Some key categories of events that suggest this future might be emerging are listed below:

- Legislative developments in several countries will mean countries begin to treat social media networks as publishers, with responsibility for content that is available on their sites.
- A growth in the popularity of networks which explicitly value users' data, either through paying people to exist on networks, or offering service discounts in exchange for joining them.
- User numbers on social networks becoming more dynamic as people leave networks and move to others.
- The development of new guidelines and standards for social data research as the discipline becomes distinct from market research and existing secondary research.

Octopus Corporations

In this world, the principal impact of the COVID-19 pandemic for social networks is the same as it has been for other sectors – reinforcing the position of existing players and accelerating existing trends.

The current big networks become more entrenched and their role in connecting people during lockdowns means that public attitudes towards platforms become more positive. Their increasing coverage and utility also makes them valuable to governments. Buoyed by public and government trust, their coverage of national populations expands further as more of life moves online.



From this strong position these networks can broaden their offering into adjacent and new industries including payments mobility and healthcare, utilising new technologies such as the Internet of Things. As they branch out, they become stronger still through the network effects brought by rising user numbers and deepening interaction with existing users. Interacting with friends will ultimately become just one of a number of potential motivations for using a social network.

Implications for research

Increasing user volumes on existing networks would support current research models. “Scraping” research approaches that access free data from networks will be able to continue and rising user numbers will mean that despite an unresolved debate over representativeness, a wider range of users will be brought into social networks, which will make their data more useful and less open to errors of coverage. The current landscape of mixed methods in social data research would continue in this scenario.

Keeping pace with advancing technology as existing platforms innovate will be one of the key challenges for research in this scenario: for instance, analysis will need to track the existing shift from text to image and video-based communication in social media closely. Another challenge will be addressing concerns over the robustness and reliability of social data. This more open model of social networks will mean a more level playing field for new entrants trying to influence public opinion – whether they are individuals, companies or states.

Signals to monitor

Some key categories of events that suggest this future might be emerging are listed below:

- Penetration of key social networks will climb to near-universal levels in many countries: beyond 90% or more of the public will be on a single network.
- Social networks widening the range of services they offer, successfully branching into banking/finance, health and transport.
- Social networks begin working with local and national governments to provide state services such as voter registration, tax payments and census measurements.

- Widening access to social data research: an increase in non-specialists using data from social networks to make claims about society.

Wildcards

Scenarios are plausible projections of the direction the future might take – there are potentially millions of possible futures which will likely share elements of the three identified here. In the course of this project we have identified some potential events and drivers of the future which are highly unlikely, but the impact they would have on social media were they to occur is such that they should be reflected in our report. These 'wildcards' are listed below:

- **A technological slowdown**: One area of expert consensus in this research is that the pace of innovation in technology is unlikely to slow between now and 2025. If this is the predominant assumption, then any slackening in the development of new technology has the potential to disrupt the sector significantly. "When will Moore's Law be broken?" is a perennial question in the technology sector – this wildcard assumes that it will be answered in the next few years.

- **Facebook crashes**: Many commentators and experts say that Facebook will suffer from falling user numbers and lower relevance in the coming years. This wildcard envisages a sudden implosion rather than a slow death – were Facebook to disappear overnight the implications on the technology and research sectors would be profound.

- **Payment for data**: A sudden shift in the value assigned to personal data would have a huge impact on the social network sector. If a critical mass of the public shift their view in some countries this would invert the business models of existing networks and reshape the entire sector.

- **Infrastructure sabotage**: A successful attack on the physical underpinnings of the internet – for instance cutting the undersea cables which still form the backbone of the World Wide Web – would have the immediate impact of damaging connectivity (especially for the UK) and longer-term impacts on the importance of resilience on the national agenda.

8.3 Conclusions

This chapter has presented three potential scenarios for the social media sector up to 2025. The wide range covered by these potential futures is a clear indication of the rate of change in this area, and the wildcards point to the possibility of far greater disruption.

In a present this uncertain, monitoring over the medium term is the surest way to understand how far our world will come to resemble one of the three futures. In particular, it will help to think about the key risks and opportunities each world presents.

Octopus corporations

Opportunities

In this world where increasing domains of everyday life are held on linked databases, there is powerful potential for the value of social media analysis to increase – the dichotomy between online and offline worlds will fuse and understanding what is going on online will become more important for governments and companies.

Risks

The dominating position of these companies would also be key risk in this world – for instance, established social media research could be invalidated by a small change in algorithms, decided by an unrepresentative company board based in the United States. Relations between these firms and national governments could become strained.

The Curated Internet

Opportunities

This world also presents opportunities for social media research, although this scenario is more fragmented. Finding specific groups of interest would require significant analyst understanding of the relevant culture and the potential for cost through access fees and subscriptions. However, this approach would promote qualitative methods which would give a fuller understanding of the specific groups in question.

Risks

The risks in this world are that the remaining open forums become even more strongly dominated by extreme voices and misinformation. People would become less likely to express opinions in open cyberspace, leaving space on remaining forums to those invested in promoting extreme opinions and it is likely that these would still be picked up by the public and journalists. The reported gap between public opinion and stated opinions online would grow wider.

Digital Fortresses

Opportunities

This world is marked by social media research having lower utility or value for most purposes. The opportunities are instead in the offline world – traditional methods of communication may be viewed as more trustworthy by contrast to an online world typified by viruses, misinformation and bots.

Risks

The risk in this scenario is that it becomes harder to understand public opinion through any single medium- response to traditional methods of surveying continue to decline and post-Covid restrictions limit face-to-face qualitative research. A triangulation approach using skilled analysts becomes necessary to provide even basic information on population-level attitudes.

Appendices

A. Expert Interviews Overview

Interviews with a group of 30 global experts formed the basis of the analysis in this report. Participants were sourced by Ipsos MORI, King's Strategic Communications Centre, and the Social Intelligence Lab to have a global view of how social media analysis is developing across research areas including consumer goods, academia and defence and counter-terrorism.

The breakdown of experts across areas of expertise is included below

Area of expertise	Number of interviews
Defence/strategic communications	6
Social data analysts (private sector)	13
Academics and public sector analysts	7
Online ethnographic specialists	2
Social media networks	2

B. Table 8.3: Relevance of Total Survey Error Framework

	Definition	Application to social media
Validity	Validity is a concern in all research; it concerns the approximate truth of an inference. In social research validity is the degree to which the question measures the 'construct' it is intended to measure. It's central to establishing the overall validity of a method. A construct refers to a concept or characteristic that can't be directly observed.	Need to consider the extent to which the inferences you make about social media are legitimate. For example, it is necessary to consider how technical and social processes have contributed to the creation of the data. Researchers may associate 'friends' as evidence of a social connection or re-tweets as endorsements, but without knowing how and why people use these functions it would be difficult to use them as a valid measure of say, their social circle or opinion on a topic.
Measurement Error	Occurs when there is a deviation between the value measured and the true value. There are numerous sources of measurement error. In surveys they might be caused by researchers, interviewers, respondents and data processors. For example, when asking questions about socially 'undesirable' behaviours or attitudes you are likely to see a pattern of under-reporting which would distort the data.	Social-desirability bias is likely to play a significant role on social media. People are aware of how they are being perceived. It is necessary to consider whether certain topics, particularly those that are polarising or socially undesirable, could be accurately measured using social media. it is difficult to establish whether the opinions people post online are reflective of what they really think; however, it will depend on the topic.
Processing Error	Includes all post-collection operations. Processing errors may include: errors of transcription, errors of coding, errors of data entry, errors in the assignment of weights or and errors of arithmetic in tabulation.	Need to consider the extent to which text analytics and statistical analysis software correctly classifies data. E.g. sentiment analysis is often carried out using natural language processing and machine learning, processing errors may occur if the identification and categorisation of the sentiment of user's posts is incorrect.
Coverage Error	Occurs when specific members of a population are excluded from the sample. This may arise from inadequate sampling frame, or flaws in the data collection method. Coverage error results because of under or over-coverage of a certain group.	There may be certain reasons which make it impossible to sample a specific group of people. For example, on Facebook you can only collect 'public' posts meaning private pages and closed forums cannot be accessed. Even if you increased the sample size or repeated the study it would still be impossible to sample those posts.
Sampling Error	Occurs when the selected sample does not represent the entire population. This may be due to biased sampling procedures or it may happen by chance.	In the context of social media, it is important to consider whether the sample of data that you access (e.g. using the Twitter API) reflects the wider population. This is likely to be improved with a larger sample or repeating the same analysis on a newly drawn sample.
Non-Response error	When a proportion of the approached eligible sample members cannot be contacted or persuaded to provide data. This generates non-response bias if the survey responders and non-responders differ in important ways.	There may be specific groups who are not 'active' on social media, or who do not engage with the topic of interest. Within this, there might be important demographic differences in those who are active. For example, if older demographics do not actively post on social media this could create a non-response bias.

C. Delphi Futures Statements

The delphi survey was conducted with 23 of the 30 experts interviewed for the main stage of this project using an Ipsos survey platform. Fieldwork was between the 5th and 27th May 2020.

All experts were sent an anonymised link to complete the survey, which contained 28 statements about the future of social media, aligned with the seven drivers identified by our analysis. The full list of statements is included below:

1. Social media sources will only sell their data directly, rather than via third party vendors (e.g. web listening platforms).
2. Most social discussion online will occur on locked forums, or within online game universes.
3. Most social media interactions will occur through encrypted or secure channels.
4. Publicly available social media conversations will be less representative of the opinions in a given population than it is now.
5. Social media networks will be treated as publishers, with responsibility for all content on their sites.
6. Those wishing to sign up to any social network will be required to provide proof of identity.
7. The most popular social networks will be based around video sharing rather than words and images.
8. Social media networks will increasingly offer other services to their users, including payments, healthcare and transport.
9. Advances in analytical algorithms will have kept pace with the growth in non-text data.
10. New standards of quality for research using social data, distinct from the standards governing surveys, will be universally developed and applied.
11. Active user numbers for dominant social networks such as Facebook and VKontakte will fall by more than half.
12. An anti-trust action will have been launched in the US, with the aim of forcing one of Google, Amazon or Facebook to demerge.
13. The global internet will be completely divided into a series of regional or sometimes national internets.
14. Social media networks will be required to store all data about a country's citizens within the borders of that nation.
15. Widespread public use of data portability rights will mean social media sites struggle to monetise users' data.
16. Social networks will be portals to access other services rather than destinations in themselves.

17. The increasingly low credibility of information available on social media will have reduced public trust in other information sources.

18. The public will expect social networks to pay for access to their own data.

19. Governments will have abandoned attempts to regulate disinformation and fake news online.

20. The public will only trust in information which has come through offline sources, especially personal connections.

21. Social media networks will keep updating their interfaces, but there will be no major changes to their core offer.

22. Social media application optimisation will focus on ensuring battery or device preservation (rather than best interface).

23. Social media data analysis will become an approach used by all types of researchers, rather than by specialists.

24. An existing social data metric (such as ‘sentiment’, or ‘reach’) will remain the most-used analytical approach.

25. In most countries there will be at least one social media network which reaches more than 90% of the online population.

26. Some national governments will be designating dominant social networks as utilities, like electricity or broadband.

27. Global internet governance standards will have been established, based on current standards in the EU, US or China.

28. The internet will be a two-tier system with global and national levels, each with different user audiences.

For more information

3 Thomas More Square
London
E1W 1YW

t: +44 (0)20 3059 5000

www.ipsos-mori.com
http://twitter.com/IpsosMORI

About Ipsos MORI Public Affairs

Ipsos MORI Public Affairs works closely with national governments, local public services and the not-for-profit sector. Its c.200 research staff focus on public service and policy issues. Each has expertise in a particular part of the public sector, ensuring we have a detailed understanding of specific sectors and policy challenges. Combined with our methods and communications expertise, this helps ensure that our research makes a difference for decision makers and communities.

Ipsos MORI