“Researching Twitter can transform how we understand trust…”

A QUESTION OF TRUST

Carl Miller

DEMOS

Demos is Britain's leading cross-party think tank. We produce original research, publish innovative thinkers and host thought-provoking events. We have spent 20 years at the centre of the policy debate, with an overarching mission to bring politics closer to people.

Demos is now exploring some of the most persistent frictions within modern politics, especially in those areas where there is a significant gap between the intuitions of the ordinary voter and political leaders. Can a liberal politics also be a popular politics? How can policy address widespread anxieties over social issues such as welfare, diversity and family life? How can a dynamic and open economy also produce good jobs, empower consumers and connect companies to the communities in which they operate?

Our worldview is reflected in the methods we employ: we recognise that the public often have insights that the experts do not. We pride ourselves in working together with the people who are the focus of our research. Alongside quantitative research, Demos pioneers new forms of deliberative work, from citizens' juries and ethnography to social media analysis.

First published in 2014

Magdalen House, 136 Tooley Street,
London, SE1 2TU, UK

ISBN 978 1 909037 75 5
Series design by modernactivity
Typeset by Chat Noir Design, Charente

Set in Gotham Rounded
and Baskerville 10
Cover paper: Flora Gardenia
Text paper: Munken Premium White

A QUESTION OF TRUST

Carl Miller

DEMOS

Contents

Acknowledgements

First, my sincere thanks to Daniela and Ralph at Demos for their help in checking, formatting and producing the paper. Also to Thomas Plant, whose feedback and comments on earlier drafts greatly improved the final product.

The paper would not have been possible without the help of a fantastic group of interns and research assistants, including Chris Waller, Louis Reynolds, Charlie Beirouti, Sofia Patel and Alex Krasodomski-Jones. Thanks to all of them for the varied and constant contributions they made to this paper, from analysis and writing to visualisation and presentation.

Nor would it have possible without the talented group of technologists at the University of Sussex who form the other half of CASM: David Weir, Simon Wibberley, Jeremy Reffin and Andrew Robertson. Their patience, support and innovation remain very appreciated indeed.

But, in this, my first solo report for Demos, my deepest thanks are reserved for Jamie Bartlett and Jonathan Birdwell, without whose help and guidance I never would have reached this point.

All errors and omissions, of course, remain completely my own, as do the views expressed in this report.

Carl Miller
November 2014

1 Introduction

Trust

Many things we hold to be true, we can neither discover nor prove ourselves. Claims about the world often rely on specialised or esoteric knowledge, on information that we cannot access, or on experiences that we have not had. We often rely on other people and institutions – experts, governments, journalists, our friends and family – to decide what is true and false, to make decisions and take actions.[1]

We believe secondhand knowledge because of who provides it. When we believe people and institutions are capable and able sources, and also ones acting with credibility and integrity, we 'trust' them. Trust is one of the most important concepts for explaining each of our intellectual and moral worlds, and our relationship with the people, institutions, technologies and processes that surround and shape our lives.

When well placed, trust is a social good, foundational to a healthy democracy, and vitally necessary for human beings to work confidently with one another. Holding trust is an important asset for governments, organisations and individuals. Reasonable scepticism and criticism of institutions and individuals is important: to hold the powerful to account, to challenge conventions, to produce new solutions, and to enable genuine choice. However, when mistrust is high and generalised, it is harmful. It increases friction in society, makes interactions between people more difficult, and undermines the capacity of government to benefit the people it serves.[2]

Measuring trust is important to understand society, to know how messages are understood, how organisations and processes are interacted with, and why individually and collectively we make the choices that we do. It is an important part of sociology, and a vital requirement of informed public policy.

The measurement of trust

Trust is an elusive concept as well as an important one. Like other concepts that we use to explain the social world, it is abstract and intangible. It cannot be measured directly, but only – indirectly – through social behaviours that imply or indicate it. These are called 'trust indicators'.[3]

A range of disciplines – psychology, psephology, anthropology, medicine, computer science and sociology – measure trust using a variety of indicators. Some ways of researching trust are highly individual: extended interviews, diaries and logs. Others are based on group discussions or observing key parts of natural life as they unfold. Large society-wide attempts to measure trust use structured techniques to ask people questions and record their responses in the form of polls. These are conducted regularly in an attempt to understand and measure the extent of our trust in governments, organisations and individuals around us, and how it changes. Polling companies, for instance, are especially interested gathering a representative national picture of trust, especially in politics, politicians, important institutions and the major professions. The polling company Ipsos MORI has systematically measured general public trust in professions since 1983, as has YouGov since 2003.[4] Eurobarometer, the biannual survey undertaken for the European Commission, measures trust in national governments, parliaments and the European Union.[5]

While polls are the most widely used method to measure trust, they have important weaknesses. In her 2002 Reith lectures on trust, the philosopher Onora O'Neill argued that trust is highly dependent on context and situation, and that we trust some kinds of people or institutions in some matters, some of the time, but not all of the time.[6] Polls struggle to grasp this context. They are expensive, difficult to conduct constantly, and often necessarily involve – in the act of polling – lifting people out of the day-to-day context where trust really exists. O'Neill complained that when asking generic questions about trust, polls smooth out the important everyday distinctions that sit at the heart of what trust really is.

Social media

Over the last five years, the way we communicate with each other has changed dramatically. Around the world, 1.2 billion people use an app or website to generate as well as consume information.[7] Social media is now the most popular activity on the internet.[8] In the UK, 48 per cent of British adults use a social networking site, and this number is growing rapidly.[9] The explosion of social media has radically changed where and how we live our lives, and how we talk about the experiences we have had and the attitudes that we hold.

The growth of social media has opened up new opportunities to study the social world. The recent Demos paper *Vox Digitas* argued that a new research technique – 'digital observation' – could be used to understand attitudes from Twitter, a social media platform that allows users to create accounts and post 'tweets' to its site.[10] Tweets are small micro-blogs, and can only contain a few short sentences, pictures or links. Twitter has been operating since 2006 and its 200 million active users have posted over 170 billion tweets since the platform was first created. Around 10 million of these users are British.[11] Twitter makes many of these tweets available for researchers. Taken together – *Vox Digitas* argued – Twitter created and released data that were large, constantly refreshing, and sociologically rich. This new space could be systematically researched to uncover more information about our attitudes than ever before.[12]

Research Questions

This paper examines whether it is possible to learn more about trust through researching Twitter. It addresses three central challenges: data, technology and theory.

Data: are indicators of trust produced by Twitter?

Twitter's users now produce 500 million tweets per day. These Tweets are produced for a number of reasons, have a wide variety of functions, and are on a broad range of topics. First, do these tweets contain indicators of trust? If so, what kinds of indicators

exist, in what contexts, quantities and qualities, and in relevance to what issues?

Technology: can indicators of trust be reliably measured?

Second, if indicators of trust exist, could they be dependably collected, counted, measured and differentiated? This is a question of determining the reliability of social media research methodology.

Twitter often produces datasets that are too large to be read manually. They overwhelm conventional social scientific methods for analysing them. This paper employs a new method – digital observation – which uses computer-based techniques to handle very large bodies of data. However, these techniques are unfamiliar in social research, and carry with them new implications and consequences. This paper will therefore examine whether it is possible to collect and measure indicators of trust dependably, using this technology. For a fuller description of the methodology see the section 'Method' below, and the methodology annex.

Theory: does researching Twitter advance our understanding of trust?

The third challenge is the most fundamental and important. Given the datasets that Twitter produces, and the ability of researchers to collect and analyse them, does studying Twitter improve our overall understanding of trust? This question opens up a number of issues about the broader place of social media research within social science. They cannot be conclusively addressed within the scope of this – or indeed any current – research, but are important to raise here to contextualise what else is found.

The structure of the report

The report addresses the three research questions set out above through three case studies. Each introduces a different context wherein trust is important.

Case study 1 looks at political trust in the claims and statements of individuals who hold political office. Political trust is especially necessary and important, but conventional polls suggest it is currently deficient.[13]

Case study 2 considers official trust. It is important to understand trust in the large, impersonal processes and institutions that shape people's lives. This case study looks at trust directed towards selected official organs of the British Government.

Case study 3 addresses event-based trust. This sort of trust is not related to a particular individual or institution but instead to a specific event. Trust between countries is vitally important, especially for maintaining international norms, agreements and treaty regimes. The event selected in the case study focuses on a trust-based controversy – the disputed violation of a treaty by Russia, a country in which trust is deeply divided.[14]

Method: digital observation

This section describes how digital observation was used to collect, analyse and interpret tweets related to trust.

Data collection – the application programming interface

Researchers collected one set of tweets for each case study via Twitter's 'stream' and 'search' application programming interface (API), which allows researchers to collect publicly-available tweets. The 'search API' returns a collection of relevant tweets from an index that extends up to roughly a week in the past. The 'stream API' continually produces tweets that contain one of a number of keywords for the researcher, in real time as they are made.[15] For each case study, a set of words was created based on a manual review of Twitter conversations before data collection.

Data analysis – the 'classifier'

Twitter produces datasets too large to be analysed manually or understood in their totality. Digital observation was developed as a method capable of handling datasets of this kind. It uses natural language processing (NLP) 'classifiers', which are algorithms trained by analysts to recognise the linguistic difference between different categories of language. This training is conducted using a technology developed by the Centre for the Analysis of Social Media called Method51, which allows non-technical analysts to train and use classifiers.[16] For a full description of Method51 and the training of classifiers, see the methodology annex.

The assessment of classifiers

The accuracy of each classifier used for this paper was measured by comparing the classifications given to 100 randomly selected tweets by a computer classifier and human analyst. These comparisons are recorded to show the number of times the algorithm made the same decisions as a human (and therefore was correct), and the number of times it did not (and was therefore wrong). For a fuller description of this process, see the methodology annex.

There are three outcomes of this test: recall, precision and overall. Each measures the ability of the classifier to make the same decisions as a human – and thus its overall performance – in a different way.

'Recall' describes the number of correct selections that the classifier makes as a proportion of the total correct selections it could have made. If there were ten relevant tweets in a dataset, and a relevancy classifier successfully picks eight of them, it has a recall score of 80 per cent

'Precision' describes the number of correct selections the classifiers makes as a proportion of all the selections it has made. If a relevancy classifier selects ten tweets as relevant, and eight of them are indeed relevant, it has a precision score of 80 per cent.

The overall outcomes combine recall and precision, as all classifiers are a trade-off between recall and precision. Classifiers with a high recall score tend to be less precise, and vice versa.

The overall score reconciles precision and recall to create one overall measurement of performance for the classifier.

Ethics

Conducting research using Twitter data presents new ethical challenges over how researchers should collect, store, analyse and present publicly posted tweets. Because it is a new field of research, there are no widely accepted protocols and approaches for ethical social media research. Some useful recent guidance has been issued by the 'New Social Media, New Social Science' network of social media researchers which recognises that a number of outstanding ethical questions for research of this kind remain.[17]

The Economic and Social Research Council has six principles of ethical research.[18] After reviewing these principles, we judged two to be important to consider in our research: (1) whether informed consent is necessary, and (2) whether there are possible harms to individual participants entailed.

Informed consent

Informed consent is widely understood to be required when personal data are used and research subjects have an expectation of privacy. Determining whether research subjects can reasonably expect privacy is important in both offline and online research contexts. How to do this is not simple. The individual must expect the action – in this case a Tweet – to be private and this expectation must be societally accepted as objectively reasonable.

Within this frame, an important determinant of an individual's expectation of privacy on social media is by reference to whether individuals have made any explicit effort or decision in order to ensure that third parties cannot access the information they provide there.

Applying these two tests to Twitter for our work, I believe that there is, in general, a low level of expectation of privacy among those who tweet publicly available messages. (This is not true of all social networks.) Twitter's terms of service and privacy

policy state: 'What you say on Twitter may be viewed all around the world instantly',[19] and the terms of service explicitly encourage re-use: 'We encourage and permit broad re-use of content. The Twitter API exists to enable this.'[20] Societal expectation of privacy on Twitter, I believe, is also relatively low given recent court cases that have determined that tweets are closely analogous to acts of publishing, and can thus also be prosecuted under laws governing public communications, including libel.[21]

Possible harms to individual participants

The chief burden on researchers is to make sure they are not causing any likely harm to users, if those users have not given a clear, informed, express consent that they might be identified. It is difficult to measure harm in social media research. It was judged that individual harm to participants was possible through quoting individual tweets – especially those that contained a message that was critical, offensive, obscene or represented a behaviour or attitude considered to be socially deviant. While the tweet was public, and the user could not reasonably expect it to be private (see above), it could be traced back to the individual, possibly with negative consequences. Other users might find simply having their details published distressing or upsetting, especially if used in a context they had not consented to.

There is material value to the research of directly quoting social media. As a general principle, it is considered good practice where possible to quote research subjects directly and faithfully. This is because it is more accurate as a research method and it allows other researchers to scrutinise and potentially replicate your research work more closely. On weighing these two concerns together, we determined to 'cloak' direct quotes, given the sensitive subject matter and the fact that precise, identifiable data were not materially important for the rigour of the research work. Thus we retained the essence of the data, but changed small parts so that no one could be easily identified.[22]

We did not cloak institutional accounts and those of public figures with a large following, for two reasons. First, the accurate description of how these kinds of accounts behave is particularly important for the understanding of trust; second, because the possibility of causing individual harm and the reasonable expectation of privacy for public figures and institutions is likely to be lower than for individuals.

2 Case studies

Case study 1 Political trust

As its first case study, the report investigated the 30-minute Twitter question-and-answer session hosted by Deputy Prime Minister Nick Clegg at 12pm on Monday 26 June 2014. Called a 'tweetchat', the event was advertised as a chance for British voters to ask the deputy prime minister about education policy. It involved a number of questions being sent to Nick Clegg's Twitter account (@nick_clegg). Clegg then typically retweeted a question, and then tweeted a response.

During the 30-minute duration of the tweetchat 2,946 tweets were posted containing '@nick_clegg', which were directed towards Nick Clegg's Twitter account. Two classifiers were built to analyse the dataset.

Tweets containing explicit statements of trust

The first classifier was trained to identify tweets that contained explicit expressions of trust. To qualify for this category, the tweet needed to contain an obvious, direct and clear statement of trust or mistrust in Nick Clegg – the statements he made, assurances he gave, and the facts, figures and claims that he used. Examples of tweets in the explicit trust category include:

The difference is we have a liar as PM & another liar as DPM.

Do you regret breaking your promises, raising university fees, and generally selling out to Tories?

How do u stop being seen as someone who's [sic] promises are worthless? #lameduck #liar

Any tweet that did not meet this criterion was categorised as 'not trust'.

The classifier judged 247 tweets to contain explicit statements of trust, and 2,699 not to contain explicit statements of trust.

Tweets containing explicit and implicit statements of trust

A second classifier was created to identify tweets that contained not only explicit statements of trust or mistrust, but also statements that implied or suggested the presence or absence of trust. This wider category could include statements that expressed quiet confidence or subtle doubt about the policies, transparency or reliability of Nick Clegg, the Liberal Democrats, and/or the government. Examples of tweets that fell into the implied trust category included:

Watching @nick_clegg actually answering questions in civilised manner on Twitter & @AlexSalmond full of rudeness & bluster at #fmq

Its [sic] the principle of the matter libdems wanted to abolish tution [sic] fees yet they put it up.

Mr Clegg, you supported the Conservatives [sic] oppression of people with disabilities. Flying a flag means nothing.

The classifier identified 710 tweets that contained statements of explicit or implied trust, and 2,236 that did not.

Case study 2 Official trust

The second case study aimed to identify indicators of trust directed towards an institution, department, process or official of the British Government. Government departments now maintain a number of official Twitter accounts in order to publicise and explain government policy, and to engage with citizens, answer their questions, and provide clarification and additional information.

Tweets were collected that contained the account name of one of 11 official British Government Twitter accounts. Each collected tweet either 'mentioned' the official account, usually in order to comment about it, or was addressed to the official account, often in order to pose a question or to address the institution directly.

The official Twitter accounts selected for study were:

- @ForeignOffice: the official and main Twitter account of the Foreign & Commonwealth Office (FCO)
- @DECCgovuk: the account of the Department of Energy & Climate Change (DECC)
- @FCOTravel: an account that provides consular assistance for British nationals overseas
- @FCOHumanRights: the account of the Human Rights and Communications Team at the FCO
- @SimonFraserFCO: the permanent under-secretary of the FCO and head of the Diplomatic Service
- @end_svc: the account of the campaign Time to Act and the Global Summit to End Sexual Violence in Conflict hosted by the FCO in June 2014
- @UKUrdu: the FCO's official account in Urdu
- @FMUnit: the account of the Forced Marriage Unit

@FCOClimate: the account of the Science, Innovation and Climate Department within the FCO

@SudanUnit: the account of the Sudan Unit, based at the FCO and the Department for International Development (DFID)

@LondonCyber: the account of the International Cyber Policy Unit based at the FCO

Between 26 June and 18 July 2014 we collected 22,568 tweets that mentioned at least one of these accounts.

Identifying tweets related to trust from this sample was a complex analytical task. A multi-stage method was used, and is explained below. Overall, this method aimed to build NLP classifiers to peel away successive layers of irrelevant data, eventually producing a kernel of trust indicators.

Step 1 Filter language

Some tweets sent to one of the official accounts were not in English and fell outside the scope of the report. We therefore used a classifier that was trained to recognise tweets that contained English language from all other tweets. This reduced the initial sample from 22,568 to 21,408.

Step 2 Remove institutions

Inspection of the 21,408 tweets now in the dataset revealed that many were sent from large institutional accounts – embassies, large companies, schools and government departments. We judged that these were unlikely to contain meaningful trust indicators.

A classifier was created to filter out tweets sent from institutions rather than individuals. Every tweet contains the tweeter's profile description as a piece of metadata – a short paragraph where people often describe who they are. A classifier was built to recognise the difference between profiles from institutions, and those written by individuals.

An example of institutional accounts profile description is:

The Chartered Institute of Public Relations is the professional body for PR professionals with 10,000+ members in the UK. You'll catch us on weekdays 8am–6pm.

An example of a non-institutional profile description is:

British. Teacher. May contain traces of Irishness. In Beirut.

The profile of each tweet was then classified on the basis of whether its creator was likely to be an individual. The algorithm classified 10,902 tweets as produced from a non-institutional account.

Step 3 Identify tweets containing an opinion

On inspection the dataset was divided between tweets that contained an opinion or commentary – including judgements, leading questions and non-neutral reportage – and those that did not. Statements that are indicators of trust are necessarily opinions. A classifier was therefore built to separate tweets that contained an opinion from those that did not.

An example of a judgemental tweet is:

I'm actually disgusted at my government right now. What a disgrace. @foreignoffice

An example of a non-judgemental tweet is:

@foreignoffice international monitors say parts of the wreckage were changed and cut into since they first saw them.

There were 6,383 tweets classified as containing a judgement.

Step 4 Identify tweets directed towards the UK Government

Inspection of this dataset revealed that the opinions contained in these tweets were directed at a large number of different objects. A classifier was trained to identify tweets that were directed towards a department or institution of the British Government.

An example of a tweet directed towards a government institution is:

Shameless complicity @foreignoffice UK in the Gaza massacre. Hollow words, craven kowtowing to Israeli state murder. Not in my name!

An example of a tweet not directed towards a government institution is:

@TomsonSwarb @LBC @foreignoffice with military planes being shot down in that area, it was suicidal for M Airlines to fly in that area.

There were 2,985 tweets classified as being directed towards or that were about a department or institution of the Government.

Step 5 Identify tweets that contain an indicator of trust

At the final stage, the dataset consisted of 2,985 tweets that were classified as in English, from an individual, containing an opinion and directed towards the UK Government.

A final classifier was built to identify tweets that contained a statement qualifying as either an explicit or implicit indicator of trust in the government. This could include statements that implied a level of confidence, or a lack thereof, in the policies, transparency or reliability of the FCO, DECC or their representatives.

Examples of 'trust' tweets include:

@foreignoffice You are acting outrageously. Im [sic] a UK citizen & I demand FCO puts human rights FIRST & denounces#Bahrain regime's HR abuses.

@foreignoffice I am PROUD to be from a country that flies a rainbow flag on one of it's [sic] main government buildings @LondonLGBTPride

@foreignoffice Doubtful you care really, come on? Your government is supporting fascist coup by Maidan nazis ther [sic].

There were 2,985 tweets identified as relevant to trust, 722 as not relevant to trust. At the end, therefore, approximately 10 per cent (2,263) of tweets from the original 22,568 tweets were judged to be indicators of trust.

A random sample of 150 trust-indicator tweets were qualitatively analysed to understand better the nature and kind of trust that they contained (table 1).

Case study 3 Event-specific trust

The third case study focused on expressions of trust related to a specific event. On 28 July 2014, the US delivered a letter written

Table 1 **The categories a random sample of 150 tweets fell into, with examples**

Category	Description	Count	Example
'Abusive'	Contained insulting language. Targeted the FCO and William Hague; many related to Gaza.	50	*@foreignoffice William Hague is a <expletive>!*
'William Hague'	Focused on the news that Hague had been reshuffled, including congratulations and wishes of luck.	27	*@WilliamJHague is a true statesman. A real shame that the country is losing his service in @foreignoffice.*
'Double standards'	Alluded to the double standards of British foreign policy. Majority focused on Gaza.	24	*@foreignoffice never see your sympathies expressed to the families of Palestinians murdered by the IDF. Why pray tell?*
'Gaza/Israel'	On FCO policy towards the Israel–Palestine conflict. Especially angry reaction to FCO tweet that appeared to accept the Israeli version of events.	18	*I am disgusted beyond belief @foreignoffice latest tweet blaming Palestinians for Israel's relentless bombings!*
'Bahrain'	On FCO policy towards Bahrain – including criticisms towards British support for Bahrain, and requests for FCO to maintain peace proactively.	15	*@FCOHumanRights You do NOT condemn torture in #Bahrain. Instead, you support the Bahrain regime + allow alleged torturers into UK. Why?*
'Miscellaneous trust'	Trust-relevant tweets that do not fall into the other categories.	8	*@DECCgovuk I didn't say you have. I said the report you commissioned and published suggested that. Have you read your own report?*
'Praise'	These tweets praised the FCO.	6	*@DLidington @foreignoffice Glad you are there to support the birth of a real democracy #ukraine*
'Foreign policy'	These tweets related to general issues of foreign policy, and touched on Iraq, Saudi Arabia and the general behaviour of the FCO.	2	*@foreignoffice Stop covering up British connivance in CIA torture programme at Diego Garcia.*

by Barack Obama to the Russian Embassy containing the allegation that Russia was violating the 1987 Intermediate Nuclear Forces (INF) Treaty. This treaty mutually prohibits the US and the USSR from developing, testing, possessing or deploying intermediate-range (500–5,500km) ground-launched ballistic and cruise missiles. A US State Department report released shortly after the letter was delivered concluded that Russia was violating the INF treaty by testing banned intermediate-range, ground-launched missiles.[23] The Russian government denied the allegation. Subsequently, a series of factually competing claims about the INF Treaty were made by the US and Russian Governments, specialist NGOs, academic experts and private citizens.

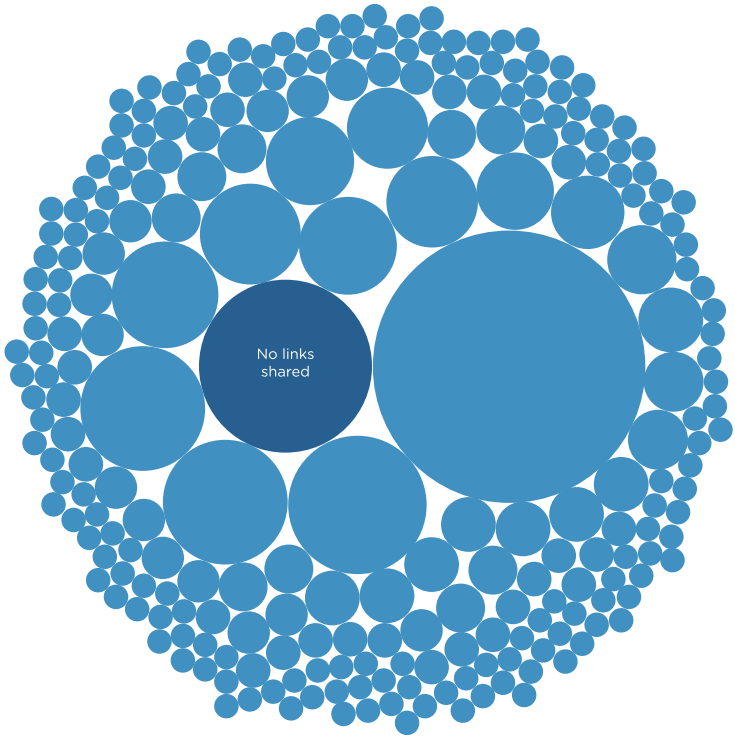
Six phrases were selected to identify tweets about the INF treaty violation. These were: 'INF treaty', 'nuclear missile treaty', '1987 treaty', 'intermediate range nuclear forces treaty', 'missile treaty' and 'nuclear arms treaty'. Between 6 and 19 August 2014, 743 tweets were collected that contained at least one of six phrases. Almost all (99 per cent) of the tweets identified by these phrases were relevant.

Very few tweets in this sample contained statements that could be construed as relevant to trust. However, 693 (93 per cent) of the collected tweets contained a link. This is a significantly higher proportion of tweets with links than were found in the other datasets collected for this report – typically 50–70 per cent.[24] It is consistent with a key research finding of *Vox Digitas:* Twitter is often used to share information rather than express opinions. Datasets of tweets often contain a substantial number of links to media stories, often with no additional comment by the tweeter.[25]

Figure 1 illustrates the link-sharing behaviour within the dataset. Each blue ball denotes a link that was shared, and the size relates to the number of tweets sharing it. It shows that a number of links were shared, many only a small number of times, but a few at very great volume.

The links were analysed to study whether the content of the shared article was relevant to the understanding of trust. Just under two-thirds (469 or 63 per cent) of tweets shared links to

Figure 1 **Link-sharing within the dataset on event-specific trust**

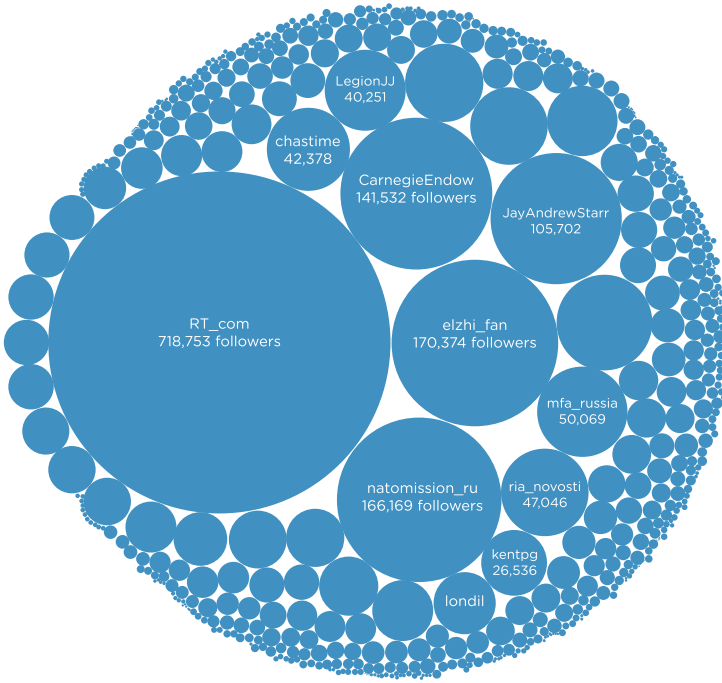


articles that themselves disputed, questioned, supported or confirmed the factual claims and statements made by either the US or Russia. The most commonly shared news item, shared by 150 of the 743 tweets (under two URLs), was an article in *Russia Today* entitled 'US nuke accusations "part of infowar set to discredit Russia" over Ukraine crisis'.[26] That article contained an interview with Russian Deputy Defence Minister Anatoly Antonov, and alleged that the US accusation was a deliberate act of information warfare to undermine Russia's reputation in the context of the Ukraine crisis. The second most widely shared item, with 32 shares, was an article on the consequences for the

Table 2 **The most frequently shared articles in the dataset on event-specific trust**

URL shared	Title of news article	Source	Number of shares
http://on.rt.com/5ku8ke	'US nuke accusations "part of infowar set to discredit Russia" over Ukraine crisis'	Russia Today	124
http://www.pism.pl/publications/bulletin/no-107-702	'Russia's violation of the INF treaty: consequences for NATO'	Polish Institute of International Affairs	32
http://news.investors.com/ibd-editorials/072914-710915-russia-violates-1987-inf-missile-treaty.htm	'Obama invited Russia missile treaty violations'	Investors.com	26
http://rt.com/politics/official-word/180136-inf-treaty-antonov-nuclear/	'US nuke accusations "part of infowar set to discredit Russia" over Ukraine crisis'	Russia Today	26
http://nationalinterest.org/feature/how-respond-russia%E2%80%99s-inf-treaty-violation-11024	'How to respond to Russia's INF treaty violation'	National Interest	19
http://bit.ly/1sQRCSj	'Intermediate-range nuclear forces treaty: setting the record straight'	Russia Today	17
http://freebeacon.com/national-security/destabilizing-threat/	'Destabilizing threat: Russian cruise missile violation of arms treaty a "serious threat"'	Washington Free Beacon	16
http://on.rt.com/ur180e	'Intermediate-range nuclear forces treaty: setting the record straight'	Russia Today	14
http://ceip.org/1zTTUST	'How to respond to Russia's INF treaty violation'	Carnegie Endowment for International Peace	13
http://bit.ly/1emAxZu	'US says Russia conducted missile test banned by 1987 treaty'	Last Great Stand	11

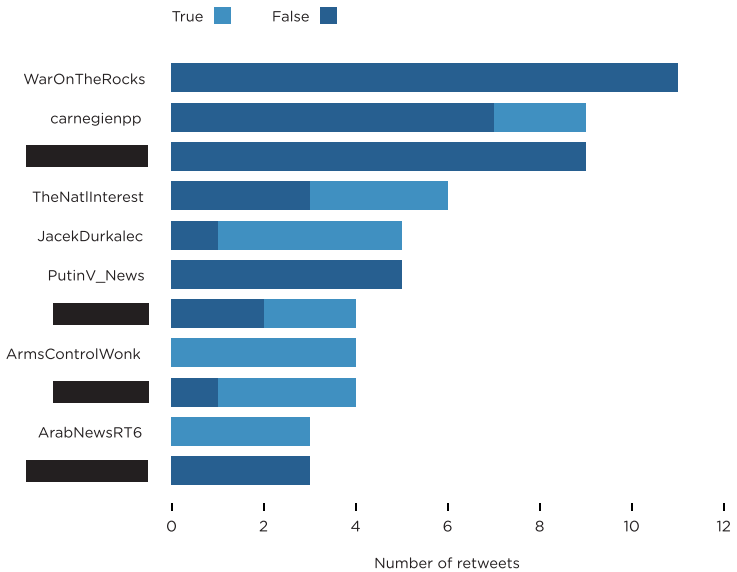
Figure 2 **The most followed Twitter accounts within the dataset on event-specific trust**



North Atlantic Treaty Organization (NATO) of Russian violations written for the Polish Institute of International Affairs. The third most shared article was from Investors.com, written from a US perspective, which was critical of what the author perceived to be Obama's weak reaction to the violation, shared 26 times (table 2).

The Russia Today account was, by a considerable margin, the most followed Twitter account that posted a tweet within the dataset on event-specific trust. Other accounts with a large number of followers included Russia's mission to NATO, and the Carnegie Endowment for International Peace (figure 2).

Figure 3 **The most prolific commentators within the dataset on event-specific trust, and whether their retweets were true or false**



While Russian-based accounts constituted a majority of the followers within the dataset, the most prolific commentators within the dataset were predominantly Western-based institutions: 'War on the Rocks', a group of international security commentators, the Carnegie Nuclear Policy Program, a (cloaked) American individual explicitly affiliated with conservative politics, and the *National Interest*, a US international affairs magazine (figure 3). Jacket Durkalec is an analyst at the Polish Institute of International Affairs. PutinV_news is an unofficial feed of news related to Vladimir Putin. Arms Control Wonk (a US-based non-proliferation academic) and Arab News contributed a large number of retweets, but did not post original tweets themselves.

3 Conclusion

Data: do indicators of trust exist on social media?

The report considers the quantity, quality and kind of data on Twitter that were identified to be related to trust. Throughout three case studies, 26,257 tweets were collected and analysed, and 5,937 indicators of trust were found, of three different types: explicit indicators of trust, implicit indicators of trust, and sharing and following behaviour on Twitter relevant to trust (table 3).

Explicit indicators of trust

These are the strongest, most definitive indications of the trust of the tweeter. These were tweets that contained statements that directly indicated a tweeter's belief in the reliability, credibility or ability of a person or institution. Tweets in this category typically contained a small number of highly relevant terms – such as 'liar' or 'trustworthy'. Only a small proportion of tweets – 247 – were

Table 3 **Indicators of trust found in the tweets collected in the three case studies**

Case study	Total tweets collected	Explicit indicators of trust (% of total)	Implicit indicators of trust (% of total)	Shares and links indicating trust (% of total)
Political trust	2,946	247 (8.38%)	2,236 (75.9%)	N/A
Official trust	22,568	N/A	2,985 (9.09%)	N/A
Event-based trust	743	N/A	N/A	469 (63.12%)
Total	26,257	247 (0.94%)	5,221 (19.88%)	469 (1.79%)

identified to contain statements so directly and explicitly related to trust.

Implicit indicators of trust

A much larger number of tweets – 5,221 – contained statements that more broadly implied or suggested a level of trust. These tweets represent weaker indicators of trust, where an inference or interpretation has to be made by the analyst in order to associate the statement with the tweeter's beliefs about trust.

Sharing and following behaviour on Twitter relevant to trust

The findings from case study 3 implied that two important behaviours on Twitter can be seen as relevant to trust: the act of sharing tweets (through retweets) and the act of following an account. Neither is a clear indication of trust. It is still unclear why people are motivated to share the information that they do on social media. People who use Twitter are likely to follow accounts for a number of reasons, some related to trust and some not. Some research suggests that we are most likely to share information that elicits 'high-arousal' emotions, negative and positive.[27]

There is some association between an individual's sharing behaviour, and the beliefs that they hold. Recent research from the University of Western Ontario concluded that 'content or news items being shared must have value to the person or the audience to whom it is directed', and that the sharing of information 'is part of online identity construction, since sharing is in a large part determined by how the content reflects on the sharer'.[28] However, both sharing and following behaviour on Twitter are possibly valuable. They reflect the political dimension of trust, and how it plays out on Twitter. Studying 'information diffusion', how messages and information spread online, is a useful future avenue for the study of trust. Especially in controversial circumstances where differing accounts exist, trust is especially important.

Features of the data

These were the main features of the data:

- *Large amounts*: 5,937 indicators of trust were found. Within the limited context of a short report, this is a significant quantity of relevant data – and larger than what could be gathered using conventional sociological methods under the same time and resource constraints.
- *Real or near-real time*: Relevant tweets were collected almost immediately after they were posted. Automated techniques – like digital observation – are capable of analysing tweets almost immediately after they are collected. This opens the possibility of analysing trust expressed in tweets in real time.[29]
- *Reactive, event-specific and context-specific*: It broadly emerged throughout the case studies that tweets did not in general express generic sentiment on Twitter about trust. A tweet is overwhelmingly a reaction to an event that the tweeter has otherwise encountered – either online or offline – in their lives. Therefore expressions of trust were most often made in the context of a specific event, at a specific time. This is notably different from the generic and broad measurements of trust often made by polls.

However:

- *The relationship of this data with trust is inferential*: The indicators of trust found in the report were either statements that people had made in tweets, or articles that people had shared. Both of these kinds of data need to be interpreted before they can link to trust.
- *The indicators are of varying degrees of quality*: Only a relatively small amount of data explicitly related to trust. A much wider body of data implicitly did so. A greater amount of interpretation is required to relate this wider body of data to trust, and – like any interpretation – the link is contestable.
- *Many tweets were adversarial*: Most contained expressions demonstrating lack of trust rather than trust. While it is possible that this reflects genuine underlying attitudes, it is also possibly a product of a 'platform-specific effect': an attribute of Twitter itself (see below).

Technology

In the course of any research it is important to be sure that what is being measured is what is described as being measured; this is called measurement veracity. In this context, it is important to know whether it is possible to identify and measure indicators of trust accurately and at great scale. The measurement of social media, especially in very large quantities, is a young and rapidly developing field of research that cannot have recourse to the depth of literature and experience of other academic methodologies, use cases or research contexts. Consequently, two issues have been identified relevant to the accuracy of measuring trust using digital observation: classifier performance and probabilistic outcomes.

Classifier performance

A key consideration is whether the automated classifiers could accurately identify the meaning of each tweet. On the basis of the empirical tests carried out for this research, the average accuracy of all the classifiers built and used for this paper is 84.8 per cent.

However, the accuracy of the classifiers was uneven. A classifier could only reliably identify explicit indicators of trust 25 per cent of the time, and implied trust 63 per cent of the time (see table 4), and was more accurate identifying irrelevant material. This suggests that classifiers can usefully identify and sort away irrelevant material, but where a high degree of accuracy is required, automated analysis should not be the final stage of the analysis. Manual verification would also be required.

Probabilistic outcomes

The products of digital observation produce inherently probabilistic judgements. Any given indicator of trust has a certain, measured chance of being present. This unavoidably introduces uncertainty into the results that are produced. Uncertainty increases when multiple classifiers are used sequentially. This shift towards results with a measured uncertainty is general in big data analysis.[30] It demands new

Table 4 **The accuracy of classifiers used in this study**

Classifier	Decision	Precision (%)	Recall (%)	F-score (%)
Tweet contains explicit trust?	Explicit trust	46.7	17.1	25.0
	No explicit trust	91.2	97.8	94.4
	Overall accuracy: 89.5%			
Tweet contains explicit or implied trust?	Implied trust	64.1	63.1	63.6
	No trust	89.8	90.2	90.0
	Overall accuracy: 84.3%			
Tweet is from an institution?	Institutional	78.6	86.8	82.5
	Non-institutional	90.2	83.6	86.8
	Overall accuracy: 84.9%			
Tweet contains an opinion?	Judgemental	83.7	91.1	87.2
	Non-judgemental	90.2	82.2	86.0
	Overall accuracy: 86.7%			
Tweet contains an opinion about government?	Towards government	76.4	79.7	73.3
	Not towards government	76.2	82.1	79.0
	Overall accuracy: 77.8%			
Tweet contains an indicator of trust?	Trust	84.1	96.7	89.9
	Not trust	90.5	63.3	74.5
	Overall accuracy: 85.6%			

ways of caveating and communicating uncertainty that is clear to the reader or research end-user.

Theory: does this research contribute to our understanding of trust?

Overall, the study of trust on social media is a promising new avenue for research. It allows the collection of large amounts of data that can be used to understand trust in the context of people's lives and real time. However, it raises two broad issues related to sociological practice and the study of attitudes. These

are who the expressions of trust are produced by, and whether they genuinely reflect an underlying attitude.

Trust of whom? Representivity

The beliefs that we hold, including trust, tend to reflect our background and experiences; how old we are, what we do for a job, where we live, and how much we earn. In conventional studies of attitudes or beliefs, research is almost always based on a sample of people that is carefully composed to represent a broader group: a particular area, a certain age group, a social or economic bracket, or the whole of the UK. Representivity is crucial to allow results from a smaller group to be generalised onto a larger one. It is a central facet of social science, and an important part of modern polling and survey methodologies. The data gathered for this report – and usually on Twitter generally – are not representative of any identifiable group. There are three reasons for this: Twitter does not represent society, data gathered from Twitter may not represent the whole of Twitter, and tweets may not represent Twitter users.

Twitter does not represent UK society, or global societies

Some kinds of people are much morel likely to use Twitter than others. Research by Ipsos MORI has found that Twitter users tend to be younger (especially between 15–34) and wealthier. The oldest and poorest demographic groups are almost entirely unrepresented on Twitter.[31] This means that data gathered from Twitter is likely to reflect the views of younger and richer demographic groups, but not older and poorer ones.

The data gathered from Twitter may not represent the whole of Twitter

As the description of the methodology above explains, the report relies on datasets of tweets that match certain keywords. There is no way to ensure that these keywords identify all tweets that relate to a particular individual, issue or event. Indeed, relevant tweets that are not collected may differ in a relevant and important way from those that do – both in the views they

contain and in the backgrounds of the people that hold those views. This is called systemic non-random bias. Data sampling strategies that have it, such as those used in this paper, can introduce skews into the analysis.

Tweets may not represent Twitter users

A small number of users are responsible for most Twitter activity. Recent research by Nielsen found that 80 per cent of the time spent on the site is accounted for by 7 per cent of 'power' Twitter users.[32] Gathered datasets of tweets reflect this: a disproportionately high number of Tweets are likely to have been sent by a small number of dedicated commentators or campaigners on a related issue.

Genuine trust?

Crucial to any research method is the quality of the indicators that are measured. The things that people share or say in tweets must relate to genuine attitudes that they hold related to trust. A number of issues are associated with this crucial link: deception, platform effects and lack of context.

Deception

Twitter as a medium is subject to deception, deliberate and otherwise. Twitter accounts and tweets can be generated automatically. Twitter estimates that 5 per cent of its monthly active users – 10.75 million – are fake.[33] Independent researchers have produced estimates that are twice as large.[34] These fake accounts are often linked to produce centrally controlled 'bot' networks, to produce large amounts of content without genuine human involvement.[35] This can be done for a number of different reasons, as hoaxes and campaigns or for the purpose of crime, spam, fraud and or as a deliberate attempt to misrepresent the popularity of an issue online. It is difficult to provenance, authenticate and therefore detect, track and predict deceptive practice. Attempts to detect fake accounts dynamically interact with counter-attempts to evade such detection. This sort of 'arms race' continues to be an important area of research.[36]

Platform effects

The intent, motivation, social signification, denotation and connotation – the meaning – of any behaviour depends on the situation and culture that it occurred within. Social media platforms are new social contexts, with distinct norms, ways of transacting and speaking. These exert 'platform effects' on the message, and change how what is said should be interpreted. An online 'disinhibition effect' is a well-evidenced influence on how we act online. We are often ruder, more critical and less inhibited in online forums than we are offline.[37] Previous research conducted by Demos demonstrates that Twitter is often used as a way to critique or complain about those in positions of power. In analysis we conducted for the Nigel Farage and Nick Clegg debates of spring 2014 we found almost 90 per cent were negative (irrespective of which candidate). We term this the 'protest-platform' phenomenon.[38]

Lack of context

Twitter is used to hold conversations with one or more other discussants, as well as to broadcast to a wider viewership. A tweet therefore may be a standalone statement, a response, or a contribution within a longer conversation. The methodology used in this report takes each tweet from its context, and treats it as an isolated statement. This is likely to lead to misinterpretations and errors in understanding what each tweet means and intends.

Moving forward

The report concludes with recommendations for how digital observation should be deployed to better understand trust.

Use digital observation to study trust iteratively and in situ

The largest advantage of this technique is that it allows trust to be measured constantly and in specific contexts. With all the caveats and reservations that this report recognises, this is nevertheless an important step forward from the occasional polling of generic levels of trust. It presents the opportunity to

study trust at scale and in the context of people's everyday lives, with all the nuances and detail that this entails.

Use digital observation to study trust related to events and in real time

Real-time research reflects the advantages of the method and the way that Twitter is most often used. The ability to discern reactions to events as soon as they occur is a powerful tool for institutions to have. It allows them to be agile, and react quickly to groundswells of anger, support or criticism quickly enough to influence the underlying developments and events that drive these attitudes.

Identify failures and crises of trust

The case studies suggested that people are more likely to use Twitter as a protest platform – manifesting behaviours that indicate a lack of trust rather than the presence of it. Digital observation is therefore less useful to understand balances of trust overall, and more useful to identify contexts and situations where trust is most absent, or where it has suddenly and seriously declined.

Use digital observation to augment polls on trust, not replace them

This report has identified that digital observation has both important advantages and significant disadvantages in comparison with conventional methods of studying trust. It should be used alongside conventional methods to augment and widen how we understand trust, in real-time, in specific contexts, and in reaction to events. However, in the face of a number of stiff challenges to credibility, digital observation should be cross-referenced and compared with results produced by more methodologically mature forms of offline research. The comparisons – whether as overlays, correlations or simply reporting that can be read side-by-side – can be used to contextualise the robustness of findings from digital observation alone.

Consider digital observation to be a long-term prospect

Digital observation is a new and rapidly developing research method. Its capability, accuracy and importance are likely to change rapidly in the future, given the growth of social media, developments in the technologies used to study social media, and improvements in social media research methods.

The growth of social media

Driven by increasing affordability and ease of access, Internet and social media use is increasing rapidly. [39] Twitter specifically has been the fastest growing social media platform in the world.[40] The annual growth of Twitter is expected to decrease over time, by 24.4 per cent in 2014, 15.1 per cent in 2016, and 10.7 per cent in 2018, when it is expected that Twitter will have 400 million users (compared with around 200 million active users in 2014).[41]

In addition, the fastest growing demographics on Twitter are the ones currently least represented. The number of users aged between 55 and 64 grew 79% between 2012 and 2013 – proportionately more than any other.[42] This growth is likely, over time, to make Twitter increasingly representative of both UK society, and of the global population. However, the future user bases of specific social media platforms are volatile and difficult to predict accurately. MySpace was once the largest social media site in the world with 75.9 million regular users.[43] It now has only 36 million, and has been far surpassed by other platforms, most notably Facebook, which has over 1 billion users.[44]

Rapid developments of natural language processing

The ability of automated, computer-based technologies to understand natural language – such as Tweets – is increasing rapidly. Improvements in hardware and data storage are allowing unprecedented amounts of data to be leveraged to train computers to establish the meaning of language accurately. A sub-field of NLP – distributional semantics – is now producing 'deep learning' techniques, modelled on the human brain. Multi-modal techniques are beginning to be able to understand meaning across different formats, including film and pictures as

well as text. The result is that the technological aspects of digital observation are likely to significantly increase in accuracy within the next five years.

Improvements in social media research method

There are significant attempts to improve social media research frameworks to make them better reflect the principles of sociological good practice. The UK's academic research councils are making significant investments in the development of methods to understand social media rigorously.[45] As an example, Demos, the University of Sussex and Ipsos MORI have begun a major project – 'In the Hands of the Analyst' – to make develop methodologies capable of studying attitudes on social media in a representative way.[46]

Methodology annex

Data collection

All data from Twitter were collected from its application programming interfaces (APIs). Twitter has three different APIs that are available to researchers.

- *search API*: return a collection of relevant tweets matching a specified query (word match) from an index that extends up to roughly a week in the past
- *stream API*: continually produce tweets that contain one of a number of keywords to the researcher, in real time as they are made
- *sample API*: return a small number (approximately 1 per cent) of all public tweets in real time

Each of these APIs (consistent with the vast majority of all social media platform APIs) is constrained – or 'rate-limited' – by the amount of data they will return. This limit was not exceeded by any collection used in this report.

Data analysis

Natural language processing

The amount of Twitter data collected was too large to be analysed manually or understood in its totality. Natural language that occurs on social media can be automatically understood at great scale and speed using NLP. A long-established sub-field of artificial intelligence research, NLP combines approaches developed in the fields of computer science, applied mathematics and linguistics. It is increasingly used as an analytical 'window' into big datasets, such as the ones collected for this report.

The value of NLP in the context of this work is its ability to create classifiers, which are algorithms that automatically place tweets in one of a number of pre-defined categories of meaning. To build classifiers, the study uses a web-hosted software platform, developed by the project team, called Method51. This uses NLP technology to allow researchers to construct bespoke classifiers rapidly to sort defined bodies of tweets into categories (defined by the analyst). The process to create each classifier was to go through the following phases. Each phase is undertaken via a user interface within Method51.

Phase 1 Define categories

The formal criteria explaining how tweets should be annotated is developed by defining between two and five categories within which the classifier will try to place each (and every) tweet. The exact definition of the categories develops throughout the early interaction of the data. The categories are not arrived at a priori, but only through an iterative interaction with the data – wherein the definition of each category can be challenged by the actual data. This is to ensure that the categories reflect the evidence rather than the preconceptions or expectations of the analyst. This is consistent with a well-known sociological method called grounded theory.[47]

Phase 2 Create a gold-standard test dataset

This phase provides a baseline of truth against which the classifier performance is tested. A number of tweets (usually 100, but more are selected if the dataset is very large) are randomly chosen to form a gold-standard test set. These are manually coded into the categories defined during phase 1. These tweets are then removed from the main dataset, and are not used to train the classifier in phase 3.

Phase 3 Train

This is the process wherein training data are introduced into the statistical model, called 'mark up'. Through a process called 'active learning', each unlabelled tweet in the dataset is assessed

by the classifier for the level of confidence it has that the tweet is in the correct category. The classifier selects the tweets with the lowest confidence score, and these are presented to the human analyst via a user interface of Method51. The analyst reads each tweet, and decides which of the pre-assigned categories (see phase 1) that it should belong to. When ten have been selected, they are submitted as training data, and the NLP model is recalculated. The NLP algorithm looks for statistical correlations between the language used and the meaning expressed to arrive at a series of rules-based criteria.

Phase 4 Review and modify performance

The updated classifier is then used to classify each tweet within the gold-standard test set. The decisions made by the classifier are compared with the decisions made (in phase 2) by the human analyst. On the basis of this comparison, classifier performance statistics – 'recall', 'precision' and 'overall' (see 'The assessment of classifiers', in chapter 1) – are created and appraised by a human analyst.

Phase 5 Retrain

Phases 3 and 4 are iteratively repeated until classifier performance ceases to increase. This state is called 'plateau', and, when reached, is considered the practical optimum performance that a classifier can reasonably reach. Plateau typically occurs within 200–300 of annotated tweets, although it depends on the scenario: the more complex the task, the more training data that are required.

Phase 6 Process

When the classifier performance has reached a plateau, the NLP model is used to process all the remaining tweets in the dataset into the categories defined during phase 1 along the same, inferred, lines as the examples it has been given. Processing creates a series of new databases – one for each category of meaning – each containing the tweets considered by the model to be most likely to fall within that category.

Phase 7 Create a new classifier (phase 1), or post-processing analysis (phase 9)

Practically, classifiers are built to work together. Each is able to perform a fairly simple task at a very large scale: to filter relevant tweets from irrelevant ones, to sort tweets into broad category of meanings, or to separate tweets containing one kind of key message with those containing another. When classifiers work together, they are called a 'cascade'. Cascades of classifiers were used for case studies two and three. After phase 7 is completed, a decision is made about whether to return to phase 1 to construct the next classifier within the cascade, or, if the cascade if complete, to move to the final phase – phase 8, post-processing analysis.

Phase 8 Carry out post-processing analysis

After tweets have been processed, the new datasets are often analysed and assessed using a variety of other techniques. These are:

- *Metadata analysis*: There are around 150 pieces of metadata attached to every tweet. They include information about the tweeter, such as their public profile, the number of followers they have, and their screen name; about the tweet's context, such as whether it was a retweet, or a reply; possible geographic information about where the tweet was sent from, or where the tweeter has stated they are from; and whether the tweet contains objects like links, hashtags or media content. The metadata of processed datasets are often analysed to understand better their nature and meaning, such as the most retweeted tweets, the users with the most followers, and geographic distributions of tweets.
- *Time series analysis*: The datasets are often graphed over time in order to understand their relationship to offline events, and to identify significant moments when volume sharply increased or decreased.
- *Qualitative analysis*: a random sample of tweets is often drawn from processed datasets and analysed using qualitative sociological coding methodologies. These techniques attempt to draw out the detail, nuances and subtleties of meaning contained within the dataset, which automated analysis is not able to identify.

Notes

1 J Hardwig, 'The role of trust in knowledge', *Journal of Philosophy* 88, no 22, 1991.

2 S Parker et al, *State of Trust: How to build better relationships between councils and the public*, London: Demos, 2008.

3 A Bryman, *Social Research Methods*, 3rd edn, Oxford: OUP, 2008.

4 Ipsos MORI, *Trust in Professions*, 2013, www.ipsos-mori.com/researchpublications/researcharchive/15/Trust-in-Professions.aspx?view=wide (accessed 18 Nov 2014); YouGov, 'How much do you trust the following to tell the truth?', 2012, http://cdn.yougov.com/cumulus_uploads/document/smmygtcodn/YG-Archives-Pol-Trackers-TrustTrend-Dec.pdf (accessed 18 Nov 2014).

5 European Commission, Public Opinion in the European Union, Eurobarometer 81, 2014, http://ec.europa.eu/public_opinion/archives/eb/eb81/eb81_first_en.pdf (accessed 18 Nov 2014).

6 O O'Neill, Reith lectures, 2002, www.bbc.co.uk/radio4/reith2002/lecture1.shtml (accessed 18 Nov 2014).

7 Emarketer, 'Where in the world are the hottest social networking countries?', 29 Feb 2012, www.emarketer.com/Article/Where-World-Hottest-Social-Networking-Countries/1008870 (accessed 18 Nov 2014).

8 Comscore, 'It's a social world: social networking leads as top online activity globally, accounting for 1 in every 5 online minutes', 21 Dec 2011, www.comscore.com/Insights/Press-Releases/2011/12/Social-Networking-Leads-as-Top-Online-Activity-Globally (accessed 18 Nov 2014).

9 ONS, 'Internet access – households and individuals, 2012 part 2', release, Office for National Statistics, 28 Feb 2013, www.ons.gov.uk/ons/rel/rdit2/internet-access—-households-and-individuals/2012-part-2/index.html (accessed 18 Nov 2014).

10 J Bartlett et al, *Vox Digitas*, London: Demos, 2014

11 M McGrail, 'Twitter reveals latest UK usage statistics', blog, The Social Penguin, 2012, www.thesocialpenguinblog.com/2012/05/15/twitter-reveals-latest-uk-usage-statistics/ (accessed 18 Nov 2014).

12 Bartlett et al, *Vox Digitas*.

13 'Trust, politics and institutions', British Social Attitudes 30, 2013, http://bsa-30.natcen.ac.uk/read-the-report/key-findings/trust,-politics-and-institutions.aspx (accessed 18 Nov 2014).

14 'Russia's global image negative amid crisis in Ukraine: Americans' and Europeans' views sour dramatically', Pew Research Global Attitudes Project, 9 Jul 2014, www.pewglobal.org/2014/07/09/russias-global-image-negative-amid-crisis-in-ukraine/ (accessed 18 Nov 2014).

15 Twitter, 'Current performance and availability status', 18 Nov 2014, https://dev.twitter.com/status (accessed 18 Nov 2014).

16 Method51 is a software suite developed by the project team over the last 18 months. It is based on an open source project called DUALIST (B Settles, 'Closing the loop: fast, interactive semi-supervised annotation with queries on features and instances', *Proceedings of the Conference on Empirical Methods in Natural*

Language Processing, 2011, pp 1467–78.) It enables non-technical analysts to build machine-learning classifiers. The most important feature of it is the speed wherein accurate classifiers can be built. Classically, an NLP algorithm would require roughly at least 10,000 examples of 'marked-up' examples to achieve 70 per cent of accuracy. This is expensive, and takes days to complete. However, DUALIST innovatively uses 'active learning', an application of information theory that can identify pieces of text that the NLP algorithm would learn most from. This radically reduces the number of marked-up examples from 10,000 to a few hundred. Overall, in allowing social scientists to build and evaluate classifiers quickly, and therefore to engage directly with big social media datasets, the Method51 system makes possible the digital observation methodology used in this project.

17 J Salmons, 'New social media, new social science... and new ethical issues!', 2014, http://vision2lead.com/2014/02/new-social-media-new-social-science-and-new-ethical-issues/ (accessed 18 Nov 2014).

18 ESRC, 'Framework for research ethics', Economic and Social Research Council, 2012, www.esrc.ac.uk/about-esrc/information/research-ethics.aspx (accessed 18 Nov 2014).

19 Twitter, 'Terms of service', 2014, www.twitter.com/tos (accessed 27 Mar 2013); Twitter, 'Privacy policy', 2014, www.twitter.com/privacy (accessed 27 Mar 2013).

20 Twitter, 'Terms of service'.

21 For instance, A. Sherwin, 'Twitter libel: Sally Bercow says she 'learned the hard way' as she settles with Tory Peer Lord McAlpine over libelous Tweet', *Independent*, 24 May 2013, http://www.independent.co.uk/news/uk/crime/twitter-libel-sally-bercow-says-she-has-learned-the-hard-way-as-she-settles-with-tory-peer-lord-mcalpine-over-libellous-tweet-8630653.html.

22 Salmons, 'New social media, new social science... and new ethical issues!'.

23 US Department of State, 'Adherence to and compliance with arms control, non-proliferation and disarmament agreements and commitments', Jul 2014.

24 Bartlett et al, *Vox Digitas*.

25 Ibid.

26 'US nuke accusations "part of infowar set to discredit Russia" over Ukraine crisis', *Russia Today*, 13 Aug 2014, http://rt.com/politics/official-word/180136-inf-treaty-antonov-nuclear/ (accessed 19 Nov 2014). Under different URLs, this was the both the first and the fourth most-shared news article.

27 L Shifman, *Memes in Digital Culture*, Cambridge MA: MIT Press, 2014.

28 L Wong, 'CAIS poster: information diffusion on social media: why people share and "re-share" online', *Proceedings of the Annual Conference of the Canadian Association for Information Science*, 2014

29 C Miller and B Duffy, 'The birth of real-time research', *Demos Quarterly* 25 Apr 2014, http://quarterly.demos.co.uk/article/issue-2/the-birth-of-real-time-research/ (accessed 18 Nov 2014).

30 V Mayer-Schonberger and K Cukier, *Big Data: A revolution that will transform how we live, work, and think*, New York: Houghton Mifflin, 2013.

31 Ipsos MORI, MediaCT Tech Tracker Q2 2014: Full report, 12 June 2014, https://www.ipsos-mori.com/Assets/Docs/Publications/IpsosMediaCT_Techtracker_Q2_2014_Final.pdf (accessed 18 Nov 2014).

32 Nielsen, 'Social Norms Twitter Users Follow the 797 rule in the UK', 23 Feb 2010, www.nielsen.com/us/en/insights/news/2010/social-norms-twitter-users-follow-the-797-rule-in-the-u-k.html (accessed 18 Nov 2014).

33 J d'Onfro, 'Twitter Admits 5% of its "users" are fake', *Business Insider*, 3 Oct 2014, www.businessinsider.com/5-of-twitter-monthly-active-users-are-fake-2013-10 (accessed 18 Nov 2014).

34 J Elder, 'Inside a Twitter robot factory: fake activity, often bought for publicity purposes, influences trending topics', *Wall Street Journal*, 24 Nov 2013, http://online.wsj.com/news/articles/SB10001424052702304607104579212122084821400 (accessed 18 Nov 2014); K Wagstaff, '1 in 10 Twitter accounts is fake, say researchers', NBC News, 26 Nov 2013, www.nbcnews.com/tech/internet/1-10-twitter-accounts-fake-say-researchers-f2D11655362 (accessed 18 Nov 2014).

35 M Austin, 'Self-deception and social media', *Psychology Today*, 6 May 2013, www.psychologytoday.com/blog/ethics-everyone/201305/self-deception-and-social-media (accessed 18 Nov 2014).

36 E Santor Jr and G Johnson Jr, 'Towards detecting deception in intelligence systems', Occasional Paper, University of Connecticut, www.dartmouth.edu/~humanterrain/papers/article.pdf (accessed 18 Nov 2014).

37 J Suler, 'The online disinhibition effect', *Cyberpsychology and Behavior* 7, no 3, 2004.

38 C Miller, 'Nick versus Nigel: live analysis', Mar 2014, www.demos.co.uk/blog/nickvnigellive (accessed 18 Nov 2014).

39 'Social networking reaches nearly one in four around the world', eMarketer, 18 Jun 2013, www.emarketer.com/Article/Social-Networking-Reaches-Nearly-One-Four-Around-World/1009976 (accessed 18 Nov 2014).

40 TJ McCue, 'Twitter ranked fastest growing social media platform in the world', Forbes, http://www.forbes.com/sites/tjmccue/2013/01/29/twitter-ranked-fastest-growing-social-platform-in-the-world/ (accessed 18 Nov 2014).

41 'Emerging markets drive Twitter user growth worldwide: more than 40% of Twitter users worldwide will be in Asia-Pacific by 2018', 27 May 2014, www.emarketer.com/Article/Emerging-Markets-Drive-Twitter-User-Growth-Worldwide/1010874 (accessed 18 Nov 2014).

42 B Cooper, '10 surprising social media statistics', Fast Company, http://www.fastcompany.com/3021749/work-smart/10-surprising-social-media-statistics-that-will-make-you-rethink-your-social-stra (accessed 18 Nov 2014).

43 F Gillette, 'The rise and inglorious fall of MySpace', *Business Week*, 22 Jun 2011, www.businessweek.com/magazine/content/11_27/b4235053917570.htm (accessed 18 Nov 2014).

44 M McHugh, 'Myspace now boasts 36m users and a 340 percent increase in artists using the network', Digital Trends, 1 Oct 2013, www.digitaltrends.com/social-media/myspace-releases-new-user-numbers/#!bNVy2a(accessed 18 Nov 2014); 'How Facebook has grown: number of active users at Facebook over the years', Yahoo! News, 1 May 2013, http://news.yahoo.com/number-active-users-facebook-over-230449748.html (accessed 18 Nov 2014).

45 For instance, Phase 3 of the Economic and Social Research Council's Big Data Network is 'social media and third sector data', http://www.esrc.ac.uk/_images/Communication-on-ESRC-Big-Data-Network-Phase-3_tcm8-29496.pdf (accessed 18 Nov 2014).

46 Demos, Ipsos MORI, University of Sussex, CASM LLP, 'In the Hands of the Analyst'.

47 BG Glaser and AL Strauss, *The Discovery of Grounded Theory: Strategies for qualitative research*, New Brunswick NJ: AldineTransaction, 1967.

48 Twitter, 'GET search/tweets', 2014, https://dev.twitter.com/rest/reference/get/search/tweets (accessed 18 Nov 2014).

References

Austin M, 'Self-deception and social media', *Psychology Today*, 6 May 2013, www.psychologytoday.com/blog/ethics-everyone/201305/self-deception-and-social-media (accessed 18 Nov 2014).

Bartlett J et al, *Vox Digitas*, London: Demos, 2014

Bryman A, *Social Research Methods*, 3rd edn, Oxford: OUP, 2008.

Comscore, 'It's a social world: social networking leads as top online activity globally, accounting for 1 in every 5 online minutes', 21 Dec 2011, www.comscore.com/Insights/Press-Releases/2011/12/Social-Networking-Leads-as-Top-Online-Activity-Globally (accessed 18 Nov 2014).

d'Onfro J, 'Twitter Admits 5% of its "users" are fake', *Business Insider*, 3 Oct 2014, www.businessinsider.com/5-of-twitter-monthly-active-users-are-fake-2013-10 (accessed 18 Nov 2014).

Elder J, 'Inside a Twitter robot factory: fake activity, often bought for publicity purposes, influences trending topics', *Wall Street Journal*, 24 Nov 2013, http://online.wsj.com/news/articles/SB10001424052702304607104579212122084821400 (accessed 18 Nov 2014).

Emarketer, 'Where in the world are the hottest social networking countries?', 29 Feb 2012, www.emarketer.com/Article/Where-World-Hottest-Social-Networking-Countries/1008870 (accessed 18 Nov 2014).

'Emerging markets drive Twitter user growth worldwide: more than 40% of Twitter users worldwide will be in Asia-Pacific by 2018', 27 May 2014, www.emarketer.com/Article/Emerging-Markets-Drive-Twitter-User-Growth-Worldwide/1010874 (accessed 18 Nov 2014).

ESRC, 'Framework for research ethics', Economic and Social Research Council,2012, www.esrc.ac.uk/about-esrc/information/research-ethics.aspx (accessed 18 Nov 2014).

European Commission, Public Opinion in the European Union, Eurobarometer 81, 2014, http://ec.europa.eu/public_ opinion/archives/eb/eb81/eb81_first_en.pdf (accessed 18 Nov 2014).

Gillette F, 'The rise and inglorious fall of MySpace', *Business Week*, 22 Jun 2011, www.businessweek.com/magazine/content/11_27/b4235053917570.htm (accessed 18 Nov 2014).

Glaser BG and Strauss AL, *The Discovery of Grounded Theory: Strategies for qualitative research*, New Brunswick NJ: AldineTransaction, 1967.

Hardwig J, 'The role of trust in knowledge', *Journal of Philosophy* 88, no 22, 1991.

'How Facebook has grown: number of active users at Facebook over the years', Yahoo! News, 1 May 2013, http://news.yahoo.com/number-active-users-facebook-over-230449748.html (accessed 18 Nov 2014).

Ipsos MORI, *Trust in Professions*, 2013, www.ipsos-mori.com/researchpublications/researcharchive/15/Trust-in-Professions.aspx?view=wide (accessed 18 Nov 2014).

Mayer-Schonberger V and Cukier K, *Big Data: A revolution that will transform how we live, work, and think*, New York: Houghton Mifflin, 2013.

McGrail M, 'Twitter reveals latest UK usage statistics', blog, The Social Penguin, 2012, www.thesocialpenguinblog.com/2012/05/15/twitter-reveals-latest-uk-usage-statistics/ (accessed 18 Nov 2014).

McHugh M, 'Myspace now boasts 36m users and a 340 percent increase in artists using the network', Digital Trends, 1 Oct 2013, www.digitaltrends.com/social-media/myspace-releases-new-user-numbers/#!bNVy2a (accessed 18 Nov 2014).

Miller C, 'Nick versus Nigel: live analysis', Mar 2014, www.demos.co.uk/blog/nickvnigellive (accessed 18 Nov 2014).

Miller C and Duffy B, 'The birth of real-time research', *Demos Quarterly* 25 Apr 2014, http://quarterly.demos.co.uk/article/issue-2/the-birth-of-real-time-research/ (accessed 18 Nov 2014).

Nielsen, 'Social Norms Twitter Users Follow the 797 rule in the UK', 23 Feb 2010, www.nielsen.com/us/en/insights/news/2010/social-norms-twitter-users-follow-the-797-rule-in-the-u-k.html (accessed 18 Nov 2014).

O'Neill O, Reith lectures, 2002, www.bbc.co.uk/radio4/reith2002/lecture1.shtml (accessed 18 Nov 2014).

ONS, 'Internet access – households and individuals, 2012 part 2', release, Office for National Statistics, 28 Feb 2013, www.ons.gov.uk/ons/rel/rdit2/internet-access—-households-and-individuals/2012-part-2/index.html (accessed 18 Nov 2014).

Parker S et al, *State of Trust: How to build better relationships between councils and the public*, London: Demos, 2008.

'Russia's global image negative amid crisis in Ukraine: Americans' and Europeans' views sour dramatically', Pew Research Global Attitudes Project, 9 Jul 2014, www.pewglobal.org/2014/07/09/russias-global-image-negative-amid-crisis-in-ukraine/ (accessed 18 Nov 2014).

Salmons J, 'New social media, new social science... and new ethical issues!', 2014, http://vision2lead.com/2014/02/new-social-media-new-social-science-and-new-ethical-issues/ (accessed 18 Nov 2014).

Santor E Jr and Johnson G Jr, 'Towards detecting deception in intelligence systems', Occasional Paper, University of Connecticut, www.dartmouth.edu/~humanterrain/papers/article.pdf (accessed 18 Nov 2014).

Settles B, 'Closing the loop: fast, interactive semi-supervised annotation with queries on features and instances', *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011, pp 1467–78.

Shifman L, *Memes in Digital Culture*, Cambridge MA: MIT Press, 2014.

'Social networking reaches nearly one in four around the world', eMarketer, 18 Jun 2013, www.emarketer.com/Article/Social-Networking-Reaches-Nearly-One-Four-Around-World/1009976 (accessed 18 Nov 2014).

Suler J, 'The online disinhibition effect', *Cyberpsychology and Behavior* 7, no 3, 2004.

'Trust, politics and institutions', British Social Attitudes 30, 2013, http://bsa-30.natcen.ac.uk/read-the-report/key-findings/trust,-politics-and-institutions.aspx (accessed 18 Nov 2014).

Twitter, 'Current performance and availability status', 18 Nov 2014, https://dev.twitter.com/status (accessed 18 Nov 2014).

Twitter, 'GET search/tweets', 2014, https://dev.twitter.com/rest/reference/get/search/tweets (accessed 18 Nov 2014).

Twitter, 'Privacy policy', 2014, www.twitter.com/privacy (accessed 27 Mar 2013).

Twitter, 'Terms of service', 2014, www.twitter.com/tos (accessed 27 Mar 2013).

US Department of State, 'Adherence to and compliance with arms control, non-proliferation and disarmament agreements and commitments', Jul 2014.

'US nuke accusations "part of infowar set to discredit Russia" over Ukraine crisis', *Russia Today*, 13 Aug 2014, http://rt.com/politics/official-word/180136-inf-treaty-antonov-nuclear/ (accessed 19 Nov 2014). Under different URLs, this was the both the first and the fourth most-shared news article.

Wagstaff K, '1 in 10 Twitter accounts is fake, say researchers', NBC News, 26 Nov 2013, www.nbcnews.com/tech/internet/1-10-twitter-accounts-fake-say-researchers-f2D11655362 (accessed 18 Nov 2014).

Wong L, 'CAIS poster: information diffusion on social media: why people share and "re-share" online', *Proceedings of the Annual Conference of the Canadian Association for Information Science*, 2014

YouGov, 'How much do you trust the following to tell the truth?', 2012, http://cdn.yougov.com/cumulus_uploads/document/smmygtcodn/YG-Archives-Pol-Trackers-TrustTrend-Dec.pdf (accessed 18 Nov 2014).

Demos – Licence to Publish

The work (as defined below) is provided under the terms of this licence ('licence'). The work is protected by copyright and/or other applicable law. Any use of the work other than as authorised under this licence is prohibited. By exercising any rights to the work provided here, you accept and agree to be bound by the terms of this licence. Demos grants you the rights contained here in consideration of your acceptance of such terms and conditions.

1 Definitions

A **'Collective Work'** means a work, such as a periodical issue, anthology or encyclopedia, in which the Work in its entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole. A work that constitutes a Collective Work will not be considered a Derivative Work (as defined below) for the purposes of this Licence.

B **'Derivative Work'** means a work based upon the Work or upon the Work and other pre-existing works, such as a musical arrangement, dramatisation, fictionalisation, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Work may be recast, transformed, or adapted, except that a work that constitutes a Collective Work or a translation from English into another language will not be considered a Derivative Work for the purpose of this Licence.

C **'Licensor'** means the individual or entity that offers the Work under the terms of this Licence.

D **'Original Author'** means the individual or entity who created the Work.

E **'Work'** means the copyrightable work of authorship offered under the terms of this Licence.

F **'You'** means an individual or entity exercising rights under this Licence who has not previously violated the terms of this Licence with respect to the Work, or who has received express permission from Demos to exercise rights under this Licence despite a previous violation.

2 Fair Use Rights

Nothing in this licence is intended to reduce, limit, or restrict any rights arising from fair use, first sale or other limitations on the exclusive rights of the copyright owner under copyright law or other applicable laws.

3 Licence Grant

Subject to the terms and conditions of this Licence, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) licence to exercise the rights in the Work as stated below:

A to reproduce the Work, to incorporate the Work into one or more Collective Works, and to reproduce the Work as incorporated in the Collective Works;

B to distribute copies or phonorecords of, display publicly, perform publicly, and perform publicly by means of a digital audio transmission the Work including as incorporated in Collective Works; The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. All rights not expressly granted by Licensor are hereby reserved.

4 Restrictions

The licence granted in Section 3 above is expressly made subject to and limited by the following restrictions:

A You may distribute, publicly display, publicly perform, or publicly digitally perform the Work only under the terms of this Licence, and You must include a copy of, or the Uniform Resource Identifier for, this Licence with every copy or phonorecord of the Work You distribute, publicly display, publicly perform, or publicly digitally perform. You may not offer or impose any terms on the Work that alter or restrict the terms of this Licence or the recipients' exercise of the rights granted here under. You may not sublicence the Work. You must keep intact all notices that refer to this Licence and to the disclaimer of warranties. You may not distribute, publicly display, publicly perform, or publicly digitally perform the Work with any technological measures that control access or use of the Work in a manner inconsistent with the terms of this Licence Agreement. The above applies to the Work as incorporated in a Collective Work, but this does not require the Collective Work apart from the Work itself to be made subject to the terms of this Licence. If You create a Collective Work, upon notice from any Licensor You must, to the extent practicable, remove from the Collective Work any reference to such Licensor or the Original Author, as requested.

B You may not exercise any of the rights granted to You in Section 3 above in any manner that is primarily intended for or directed towards commercial advantage or private monetary

compensation. The exchange of the Work for other copyrighted works by means of digital filesharing or otherwise shall not be considered to be intended for or directed towards commercial advantage or private monetary compensation, provided there is no payment of any monetary compensation in connection with the exchange of copyrighted works.

C If you distribute, publicly display, publicly perform, or publicly digitally perform the Work or any Collective Works, You must keep intact all copyright notices for the Work and give the Original Author credit reasonable to the medium or means You are utilising by conveying the name (or pseudonym if applicable) of the Original Author if supplied; the title of the Work if supplied. Such credit may be implemented in any reasonable manner; provided, however, that in the case of a Collective Work, at a minimum such credit will appear where any other comparable authorship credit appears and in a manner at least as prominent as such other comparable authorship credit.

5 Representations, Warranties and Disclaimer

A By offering the Work for public release under this Licence, Licensor represents and warrants that, to the best of Licensor's knowledge after reasonable inquiry:

- **i** Licensor has secured all rights in the Work necessary to grant the licence rights hereunder and to permit the lawful exercise of the rights granted hereunder without You having any obligation to pay any royalties, compulsory licence fees, residuals or any other payments;
- **ii** The Work does not infringe the copyright, trademark, publicity rights, common law rights or any other right of any third party or constitute defamation, invasion of privacy or other tortious injury to any third party.

B except as expressly stated in this licence or otherwise agreed in writing or required by applicable law, the work is licenced on an 'as is' basis, without warranties of any kind, either express or implied including, without limitation, any warranties regarding the contents or accuracy of the work.

6 Limitation on Liability

Except to the extent required by applicable law, and except for damages arising from liability to a third party resulting from breach of the warranties in section 5, in no event will Licensor be liable to you on any legal theory for any special, incidental, consequential, punitive or exemplary damages arising out of this licence or the use of the work, even if Licensor has been advised of the possibility of such damages.

7 Termination

A This Licence and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this Licence. Individuals or entities who have received Collective Works from You under this Licence, however, will not have their licences terminated provided such individuals or entities remain in full compliance with those licences. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this Licence.

B Subject to the above terms and conditions, the licence granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different licence terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this Licence (or any other licence that has been, or is required to be, granted under the terms of this Licence), and this Licence will continue in full force and effect unless terminated as stated above.

8 Miscellaneous

A Each time You distribute or publicly digitally perform the Work or a Collective Work, Demos offers to the recipient a licence to the Work on the same terms and conditions as the licence granted to You under this Licence.

B If any provision of this Licence is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this Licence, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.

C No term or provision of this Licence shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.

D This Licence constitutes the entire agreement between the parties with respect to the Work licenced here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This Licence may not be modified without the mutual written agreement of Demos and You.

Well-placed trust is a vital social good. It is foundational to a healthy democracy, and necessary for human beings to work confidently with one another. A lack of trust increases social frictions and collective endeavour. Understanding trust is vital to understand society: to know how messages are received, how organisations and processes are interacted with, and why – individually and collectively – we make the choices and live the lives that we do. It is an important part of social science, and a vital requirement of informed public policy.

We have long tried to measure and understand trust: major national polls have measured trust in politicians, organisations and the professions for decades. But these techniques are criticised for only measuring trust as an abstract and general thing: trust is often highly contextual and dependent on situation. The rise of social media offers an opportunity to study trust in a new way. As we use new digital platforms, we create large new bodies of information about what we think, what we experience and who we are. A burgeoning new discipline – social media science – attempts to rigorously and ethically research social media to understand society.

This report uses new technologies and research methods to ask whether the study of Twitter can allow us to understand trust more contextually, constantly and ethically. It scopes the quality and quantity of trust-relevant data on Twitter, the ability of emerging technologies to accurately measure them, and, overall, whether this new form of research can add to our understanding of trust, and how it relates to the standard principles of good research and sound evidence.

Carl Miller is Research Director at the Centre for the Analysis of Social Media at Demos.

ISBN 978-1-909037-75-5 £10