Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter

by Luke Sloan, Jeffrey Morgan, William Housley, Matthew Williams, Adam Edwards, Pete Burnap and Omer Rana
Cardiff University School of Social Sciences; Cardiff University School of Social Sciences; Cardiff University School of Social Sciences; Cardiff University School of Social Sciences; Cardiff University School of Social Sciences; Cardiff University; Cardiff University School of Computer Science and Infomatics

Sociological Research Online, 18 (3) 7
<http://www.socresonline.org.uk/18/3/7.html>
10.5153/sro.3001

Abstract

A perennial criticism regarding the use of social media in social science research is the lack of demographic information associated with naturally occurring mediated data such as that produced by Twitter. However the fact that demographics information is not explicit does not mean that it is not implicitly present. Utilising the Cardiff Online Social Media ObServatory (COSMOS) this paper suggests various techniques for establishing or estimating demographic data from a sample of more than 113 million Twitter users collected during July 2012. We discuss in detail the methods that can be used for identifying gender and language and illustrate that the proportion of males and females using Twitter in the UK reflects the gender balance observed in the 2011 Census. We also expand on the three types of geographical information that can be derived from Tweets either directly or by proxy and how spatial information can be used to link social media with official curated data. Whilst we make no grand claims about the representative nature of Twitter users in relation to the wider UK population, the derivation of demographic data demonstrates the potential of new social media (NSM) for the social sciences. We consider this paper a clarion call and hope that other researchers test the methods we suggest and develop them further.

Keywords: New Social Media, Demographics, Twitter, Social Media Analytics, Social Science, Sampling

Introduction

1.1 In their account of the 'coming crisis of empirical sociology', Savage and Burrows (2007, 2009) argue that, in the previous four decades, social scientists were able to claim a distinctive expertise in investigating social relations through such methodological innovations as the sample survey and the in-depth interview. However, with the development and proliferation of computing within late modern society and economy this claim has been compromised by the proliferation of 'social transactional data' generated, owned and increasingly analysed by large commercial organisations as well as government departments. The advent of 'big and broad data' on, for example, credit card transactions, telephone communications, use of store loyalty cards and insurance claims in addition to the digital production of official, 'curated', datasets such as the census of populations, general household surveys, police recorded crime, victim of crime surveys and labour market surveys, provokes an existential question for academic sociology: is it, 'becoming less of an "obligatory point of passage" for vast swathes of powerful agents … if so, how can the discipline best respond to this challenge?' (Savage & Burrows 2007: 886). One of the ways in which social science is responding is through the use of publically available digital data streams that are both 'big' and 'broad' in character. One such source of this type of data is social media. This data is not without problems or issues that are related to the clarity of generated datasets, storage, ethics, interpretation or how they relate to traditional 'terrestrial methods' (see Housley et al. 2013). However, the sheer volume of communications and its status as mass naturally occurring mediated data that potentially renders populations observable 'on the move' in ways that inform locomotive methods has resulted in considerable interest and engagement (Edwards et al. 2013).

1.2 The global adoption of social media over the past half a decade has seen the exponential expansion of 'digital publics' to an unprecedented level. Estimates put social media membership at approximately 2.5 billion non-unique users, with Facebook, Google+ and Twitter accounting for over half of these (Smith 2012)[1]. These online populations produce hundreds of petabytes of information daily, with Facebook users alone uploading 500 terabytes of data daily, including 2.7 billion 'likes', 300 million photos and just under half a billion status updates (Tam 2012)[2]. Corporations are already mining this wealth of transactional and social data to track brand popularity via topic detection and sentiment analysis (Scarfi 2012). Sentiment analysis can be used to understand large amounts of text based data and is a useful means of reducing the complexity of big text data such as archived tweets. Conventionally, sentiment analysis attaches sentiment scores to different texts or individual tweets and can therefore inform our understanding of the character of content and frequency of different opinions within social media data. Packages such as SentiStrength (see Thelwall et al. 2010) have proved popular within the social sciences. Sociologically speaking, sentiment analysis renders the emotive characteristics of large text available to further qualitative inspection and interpretation and enables cross-referencing and correlation with other variables of interest, e.g. geo-location, types of event and gender. Thus, despite its reductive characteristics it provides a further tool through which to characterize and interpret large data in ways that configure different population group perceptions and understandings of particular events in meaningful ways.

1.3 Similarly, researchers in computer and information science disciplines are beginning to explore the utility of social media data in predicting human behaviours. Tumasjan et al. (2010) identified that social media data were as accurate at predicting voting patterns as traditional polls, while Asur and Huberman (2010) demonstrated how 'tweets' were more accurate at predicting movie revenue compared to the Hollywood Stock Market. Bruns et al. (2012) discuss how social media, particularly Twitter, was adopted by the public to communicate information regarding major natural disasters including the South East Queensland floods in 2011 whilst Sakaki et al. (2010) have demonstrated how the online platform can be interpreted as an early warning system for earthquakes in Japan. Beyond social networks, Ginsberg et al. (2009) successfully traced the spread of flu in the United States by correlating flu based search terms in Google with visits to the local doctor. Arguably, social media streams are coming to be considered as new sources of data on the perceptions, opinions, feelings and actions expressed by digital publics (Bruns et al. 2011). Twitter in particular has become a favoured source for social media data collection and analysis. This is partly because Twitter provides three different level of data access (the lowest of which is free) and the data can be obtained using an online query via the Application Programming Interface (API). The free random 1% of the Twitter stream is dubbed the 'spritzer', the 'garden hose' provides access to a random 10% and the 'fire hose' affords 100% access. In average circumstances the 1% accounts for approximately 3.5 million tweets a day. Twitter can be conceptualised as a 'digital agora' (Housley et al. 2013) that provides an insight into mass user generated opinions, sentiments and reactions to social events. Twittermetrics has found a home in market research for the purposes of brand tracking (Mendes et al. 2010) natural disaster tracking (Sakaki et al. 2010) and more recently social scientific analyses of civil disturbance (Lewis et al. 2011).

1.4 However, a significant barrier to the integration of social media data into the social sciences is the apparent lack of demographic data (Mislove et al. 2011) and there is very little conceptualisation and re-purposing of social media data in ways that enable the interrogation of said data in terms of variables that are central to much social scientific explanation. Ironically, the social media data sources that are the easiest to access for the purposes of analysis, such as Twitter, are often the most 'data-light' and this has led some academics to dismiss NSM as a potentially rich source of social science data (Gayo-Avello 2012). In this paper we challenge this stance by demonstrating a series of techniques that we are using at the Cardiff Online Social Media ObServatory (COSMOS) (Burnap et al. 2013) to derive or estimate information on gender, language and various geographies from a corpus of over 113 million tweets collected over the 31 days of July 2012.

1.5 Through knowing the demographic characteristics of the tweeters we can formulate hypotheses about group differences and avoid the criticisms often levelled at 'big data' over the use of data mining (Burgess & Bruns 2012). With group comparisons in mind, the identification of gender, language and geography within NSM data represent some key variables of interest and a step towards dealing with issues surrounding the use of social media data in ways that can allow social scientific inference and explanation and take advantage of the potential use of global social media streams as 'big and broad' data. The key to operationalising this potential is to be found in sampling strategies and the development of programmes and rules that can identify 'proxies' for gender and geographical location (given the relatively small use of geo-location options by social media users at this point in time). This approach balances the low-fidelity of social media data with the opportunities provided by big and broad naturally occurring mediated data through a triangulation strategy realised through the orienting concept of 'signature science' (Housley et al. 2013) within which the development of proxies for key variables of interest is a central feature. A further consideration here is the framing of this analytic strategy within the wider methodological context of surrogacy, augmentation and re-orientation (Edwards et al. 2013). In the context of this paper, the ways in which we can derive key demographic information from Twitter represents a positioning of social media analysis within the frame of augmentation, i.e. not as a stand-alone analysis per se but as an approach that can be used to augment and enhance more traditional locomotive, punctiform, intensive and extensive research strategies including terrestrially curated datasets.

Sampling

2.1 The computational platform developed by COSMOS (<http://www.cs.cf.ac.uk/cosmos/>) collects and archives tweets from the 1% Twitter stream (the 'Spritzer'), which was accessed via an (API) free of charge. This means that with a daily total of over 350 million tweets authored worldwide, we capture and store approximately 3.5 million tweets per day. In this paper we illustrate our demographic-collection techniques by analysing a sample of 113,828,224 tweets collected during the 31 days of July 2012. This sample contains tweets authored by 13,260,819 different users and written in 53 different languages. Of the 48% of Twitter users for which we could identify gender, approximately 45% were male and approximately 47% were female.

2.2 One concern with collecting and analysing tweets from Twitter's one per cent sample of the full stream of tweets is how Twitter produces that sample. Although the method used by Twitter to whittle the full Twitter stream down to one per cent is unknown, our analyses of the one per cent stream give us three reasons to be confident that the one per cent stream is a random and representative sample of the full Twitter stream. First, the gender distribution of male and female Twitter users corresponds to the general population. In our sample of 13 million Twitter users, 30,273 were located in the UK. Of those 30,273 users, 4,133 were identified as either male or female. Of those 4,133 users, 2,017 (48.8%) were male, 2,116 (51.2%) were female. This percentage split between males and females in the Twitter sample is identical to that reported in the latest 2011 Census Statistics (ONS 2011) with 27.8 million males (48.8%) and 51.2 million females (51.2%), indicating that the 1% sample supplied by Twitter is representative of the population in this regard.

2.3 Second, all 53 languages identifiable by our language detection software are present in our sample. Third, when we plot the location of Twitter users on a map of the world, we see that the geographic distribution of users corresponds with the population distribution of each country, i.e. more users are located in areas known to have higher population densities and vice versa. We describe in detail our methods of identifying the gender, language and location of Twitter users below.

2.4 Whilst this represents a significant step forward in developing social media data for social science research there are still issues to be overcome which we do not address in this paper, such as building-in the ability to check that respondents are 'real' rather than 'bots'. Computational approaches for the classification of fake accounts (or 'bots') in Twitter have been developed, with encouraging accuracy. The aim of such approaches is to distinguish non-human users and fake accounts from legitimate human accounts, using features such as intervals between posts (i.e. looking for regular intervals), how often URLs are included in a tweet, and friends-to-followers ratio (Chu et al. 2012). A recent study demonstrated the ease with which between 20,000 and 70,000 followers could be purchased for fake Twitter accounts, with at least 20 sellers on eBay, selling at an average price of $18USD per 1,000 followers (Barracuda Labs 2012). The study also found that 75% of abusers have a URL in their profile, compared with 31% for all Twitter users. This further demonstrates the possibility to identify fake accounts through account behaviour and metadata features. Our study does not distinguish between real and fake accounts but in future we may develop the ability to identify fake accounts using COSMOS and further refine our demographic study based on real vs. fake accounts.

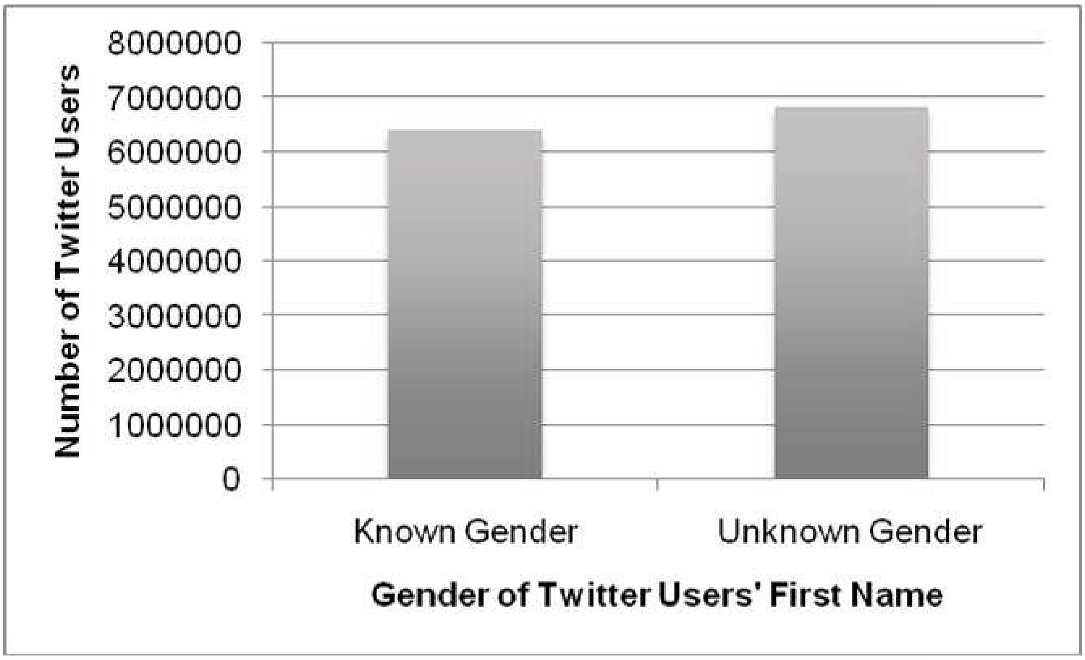
Gender

3.1 Following previous work by Mislov et al. (2011) we derived the gender of Twitter users from their first name as specified in the name field of their Twitter profile. Mislov et al. used 3,034 male and 3,643 female names from the US Social Security Administration (USSA) to identify the gender of Twitter users. The USSA records the 1,000 most popular names given to boys and girls born in the US each year dating back to the 1880s (accessed 2012). Although the population of the US is composed of a wide variety of ethnic backgrounds, English speakers of Anglo-Saxon decent dominate. As well as Anglicised names dominating the list of most popular names in the US each year, many parents follow prevailing trends when naming their babies and this fashion for baby names is clear when comparing the most popular baby names from year to year. The bias toward Anglicised names and the relatively small number of names available from the USSA prompted us to search for a larger list of names that is more representative of the names found across the world.

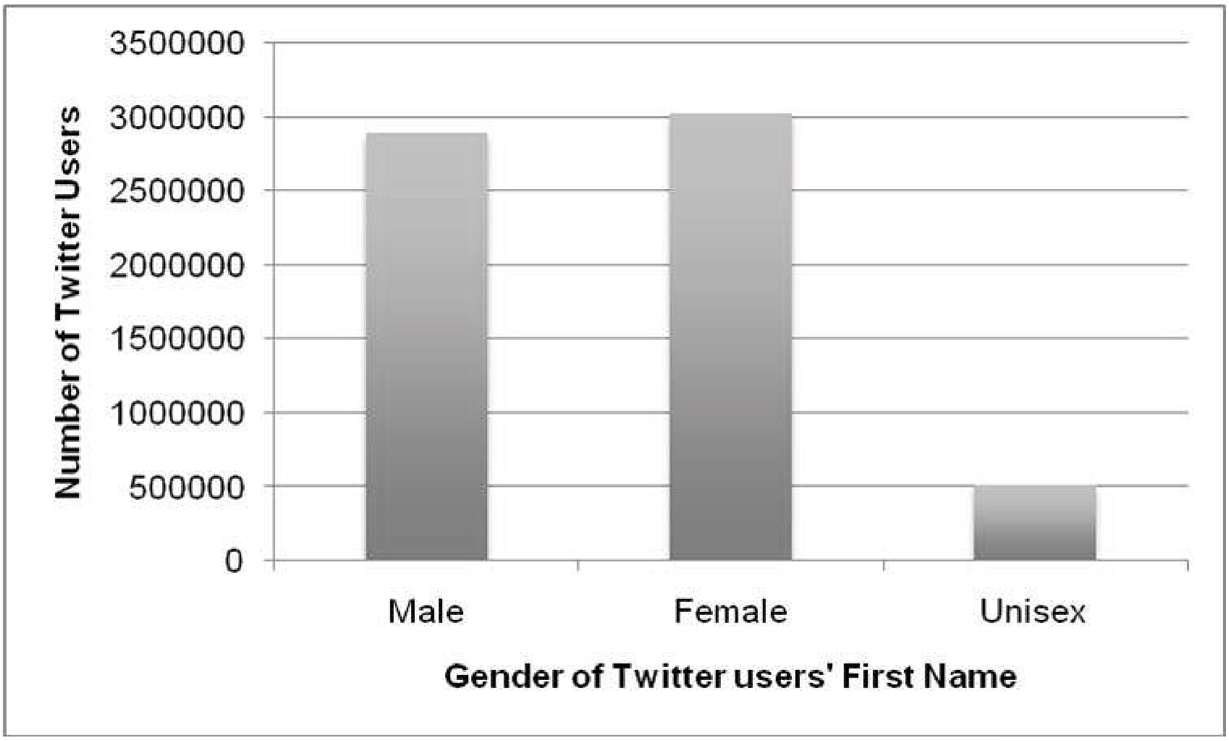
3.2 A better solution is to use the 40,000 Namen (40N) database to identify the gender of Twitter users (Michael 2007). 40N maps first names onto the gender of people commonly given those names. The 40N database was chosen because it is a comprehensive database of over 44,000 names (17,740 male, 16,921 female and 9,907 unisex) from 54 countries around the world, which is less likely to be influenced by anglo-centric naming trends and other biasing factors that affect data sources such as the USSA. Another compelling reason for using the 40N database is its recognition of name abbreviations. For example, since 40N knows that 'Jeff' is short for 'Jeffrey' and that 'Jeffrey' is a male name, 40N can correctly identify the gender of 'Jeff' as male. Since the USSA would record the full version of a baby's first name, gender-identification methods that use the USSA data may not be able to identify the gender of name abbreviations (unless, of course, the baby's name was actually the abbreviated form).

3.3 The 40N database classifies names into one of four categories: male, female, unisex and unknown[3]. The unisex category represents names often given to males and females, such as Leslie and Francis. The unknown category represents names that are not listed in the 40N database. Figure 1 shows the breakdown of the genders identified by 40N from our sample of 13,260,819 Twitter users. In this sample, 48% (6,428,491) of users had a first name known to 40N. Of those 6,428,491 users, 45% (2,895,371) were identified by 40N as having a male name, 47% (3,021,389) were identified as having a female name, and 8% (511,731) were identified as having a unisex name.

3.4 At first 48% may seem a relatively small percentage of Twitter users for which gender can be identified, particularly when compared to the much higher percentage of gender information collected from respondents to social science surveys. A compelling difference between the traditional social science survey method and social media systems like Twitter is reach. As of April 2012 Twitter had over 500 million user accounts meaning that our method has the potential to identify the gender of over 250 million Twitter users. The vast number of social media users available for demographic analysis contrasts sharply with the much smaller sample sizes available to traditional social science.



(a) The number of Twitter users for which 40N could identify first name gender.



(b) The number of Twitter users identified by 40N as male, female or unisex.

Figure 1. The breakdown of genders identified by the 40N database in our sample of 13,260,819 Twitter users.

3.5 Even with a list of 40,000 names it can be challenging to establish the gender of Twitter users. Some tweeters do not write anything at all in the name field whilst others write nonsensical sequences of letters, numbers, punctuation and other symbols. Additionally, many users make creative use of spelling, word boundaries, symbols and punctuation to customise their name, which makes name identification particularly difficult. Another reason why name identification is challenging is that Twitter users often choose pseudonyms. Without knowing each Twitter user's real name it is impossible to speculate on the proportion of Twitter users with pseudonyms. However we are certain that Twitter users do adopt pseudonyms because, unlike the pseudonyms chosen by authors of published written work, Twitter users regularly choose the names of characters from popular films, TV shows and computer games. When identifying gender from names, we are not actually concerned that Twitter users adopt pseudonyms, only that they choose pseudonyms of the same gender. For example, we would correctly identify the gender of a Twitter user whose real name was Sarah but who went under the pseudonym Jane, because the gender of the pseudonym, Jane, is the same as the gender of the real name, Sarah. Although there is limited and dated evidence that some females may mask their gender when interacting online not relating to Twitter (Jaffe et al. 1999), more recent studies have demonstrated that virtual representations of the self normally reflect reality (Huffaker & Calvert 2005).

3.6 In response to these challenges, we are continually improving our gender identification system and aim to classify the gender of as many Twitter users as possible. To accurately identify the gender of a name, the 40N database must be able to match the names we supply it with the names it knows. We use a number of text-cleaning techniques to give 40N the best chance possible of identifying a name and providing us with the gender assigned to that name. Our first technique deletes a wide variety of punctuation characters and other symbols often found in the name field of Twitter profiles, including , : ; ? ( ) { } [ ] @ # $ & * + ~ % and |. We can safely delete these characters because they contain no information about the name. At this stage, we frequently find that the name contains no characters at all. In this case, we move on to identify the gender of the next Twitter user.

3.7 Our second text-cleaning technique replaces with a space those characters commonly used in lieu of a space such as - and _. We replace rather than delete these connecting characters because at the end of our cleaning process we want to extract the first name, which we do by choosing all the characters up to the first space. If we delete connecting characters, we would miss an opportunity to identify a name. For example, cleaning the name variants 'luke-sloan' and 'luke_sloan' should produce 'luke sloan' and not 'lukesloan', which would happen if we deleted the connecting characters - and _.

3.8 The third text-cleaning technique addresses another common method of connecting names, which is to simply join them together, for example 'LukeSloan'. Although it is easy for human agents to pick out 'Luke' as the first name, the 40N database cannot. In response to this we have built a rule for breaking up joined names by looking for a particular letter pattern. When we examine a name and find that an upper-case letter immediately follows a lower-case letter, we treat the point in between those letters as a word boundary. For example, in the joined name 'LukeSloan' an upper-case 'S' follows a lower-case 'e' which meets the criteria for splitting a name. We now have two words ('Luke' and 'Sloan') and we take the first word as the first name. Although there are inevitably many cases where this letter pattern will occur in text that is not a name, the 40N database will fail to recognise the word we incorrectly identify as a first name (just as it would fail to identify both words as a name when joined together). Our rule implements our best guess and in the case where this pattern does occur in an actual name we will have correctly submitted a name to 40N for gender identification.

3.9 One limitation to gender identification through the profile name analysis approach is that more than 44,000 names (those stored in the 40k Namen database) are in use around the world. Human annotation is frequently used to derive a gold standard for validating machine annotated data (Yuen et al. 2011). In particular, Amazon Mechanical Turk has been used to 'crowdsource' human annotations on a large scale (Finin et al. 2010). In this study, human annotators were asked to classify named entities mentioned in tweets as to whether they were person, organisation or place names. To enhance our work this approach could be applied to male, female or unisex names extracted from user profiles to assist in classifying previously unclassified names and confirming machine classified names. Limitations that crowdsourcing cannot assist with include cases where users provide the name 'God', 'Darth Vader', or some other non-gendered name. And of course, we have no way to determine if the user is actually a male pretending to be a female, or vice versa. To assist with enhancing the accuracy of user location, it could also be possible to use human annotation to analyse the text of the user's last ***n*** tweets to determine if there are clues in the text to suggest their location (for example, place name recognition as in Finin et al. 2010). The language of the tweet is identified using the Language Detection Library for Java, which can identify 53 different written languages from a text sample. As with names, there will be languages, such as Welsh, that are not detected within the 53 known languages. This is another case where human annotation could be used to extend the number of languages COSMOS could automatically classify by crowdsourcing human annotations of unclassified languages.

3.10 An alternative approach to gender identification of Twitter users is to analyse the text of the tweets they write. Thomson and Murachaver (2001) revealed that people use gender-preferential language patterns in electronic discourse such as e-mail and that people trained to identify these language patterns can accurately identify the gender of authors. Argamon et al. (2006) also found significant differences between male and female language use in a large subset of the British National Corpus that spanned a variety of genres. Recent work has focussed on automating the gender-identification process. For example, Cheng at al. (2011) successfully used machine-learning techniques to automate gender-identification of Internet text authors. Automated identification of author gender through text analysis is becoming more popular as techniques improve and the computing power required to implement them increases.

3.11 Gender identification by tweet text analysis is an approach we may consider as a secondary method when our gender identification by name approach returns the unknown gender category. The main advantage of gender identification by name is simplicity: looking up the gender of a name in the 40N database is a single, fast operation well suited to both real-time and archived tweet processing. In contrast, gender identification by tweet text analysis requires several steps, including accumulating a sufficient number of tweets written by each user and then subjecting those tweets to language analysis. This type of multi-step operation is better suited to processing archived rather than real-time twitter data. However, as we are continually improving our computing infrastructure, real-time gender identification by tweet text analysis is a possible direction for future work.

Language

4.1 We can determine the language preferred by Twitter users with two methods. The first method takes the language that Twitter users set in their Twitter profile. This language setting specifies in which language Twitter users prefer to interact with the Twitter website. For example, the default language of the Twitter website is English, but French-speaking Twitter users may prefer to see a French-language version of the Twitter website. Similarly, Arabic-speaking users may prefer to see an Arabic-language version and so on. This language setting gives us a strong indication of each user's preference for and proficiency in a particular language.

4.2 The second method of identifying the language used by Twitter users is to analyse the text of the tweets written by each user. To identify the language in which a tweet was written we apply the Language Detection Library for Java (LDLJ 2012) software to the text of the tweet. The LDLJ software recognises 53 languages, as listed in Table 1. We believe this is a comprehensive subset of the languages in which the worldwide Twitter community writes their tweets. In our sample of 113 million tweets, 99.3% (113,000,453) were written in a language identifiable by the LDLJ software. Furthermore, all 53 languages were present in this sample. Table 1 details the number of tweets written in each of the 53 languages identified by the LDLJ software in our sample of 113 million tweets.

Table 1: The number of tweets in our sample of 113 million written in each of the 53 languages identifiable by the Language Detection Library for Java (LDLJ) software.

1	English	45,594,240	28	Hungarian	235,894
2	Japanese	12,738,687	29	Slovak	219,654
3	Spanish	10,136,337	30	Lithuanian	200,237
4	Indonesian	9,142,131	31	Albanian	178,708
5	Portuguese	6,991,330	32	Vietnamese	162,587
6	Arabic	3,172,589	33	Czech	108,080
7	Somali	2,553,774	34	Persian	105,789
8	Dutch	2,240,281	35	Latvian	95,477
9	Tagalog	1,899,788	36	Simplified Chinese	86,284
10	French	1,767,104	37	Bulgarian	80,332
11	Italian	1,705,202	38	Greek	78,015
12	Turkish	1,536,013	39	Traditional Chinese	52,063
13	German	1,446,948	40	Urdu	40,736
14	Korean	1,337,590	41	Macedonian	37,081
15	Afrikaans	1,307,274	42	Ukrainian	27,455
16	Estonian	1,223,220	43	Hebrew	12,827
17	Thai	836,832	44	Tamil	4,933
18	Finish	833,097	45	Hindi	1,942
19	Russian	728,551	46	Nepali	1,420
20	Swahili	721,658	47	Bengali	936
21	Norwegian	709,519	48	Malayalam	898
22	Slovene	547,050	49	Punjabi	688
23	Danish	521,356	50	Marathi	442
24	Swedish	477,127	51	Telugu	220
25	Polish	395,858	52	Kannada	183
26	Romanian	384,550	53	Gujarati	183
27	Croatian	319,283			

4.3 As well as providing an important demographic for Twitter users, language detection enables us to improve the efficiency of subsequent tweet analyses. For example, one of the techniques we apply to each tweet is sentiment analysis. Sentiment analysis is a text analysis technique that provides a numerical measure of the overall emotional content in a piece of text (for example see Dodds & Danforth 2010). Current sentiment-analysis tools are built to process English-language text so by detecting the language in which a tweet was written, we can efficiently skip non-English-language tweets when performing sentiment analysis and other English-language analyses.

4.4 Efficiency is also important because as well as analysing tweets that we collect and archive we also take tweets directly from the Twitter stream and process them in real time. The LDLJ software uses a fast language-detection algorithm that examines letter patterns that are known to occur in natural languages. Fast language detection enables us to perform a sequence of real-time analyses on each tweet without falling behind the incoming stream of tweets.

4.5 At first it may seem more efficient to determine the language preferred by each Twitter user from the language setting in their profile, rather than to perform the extra computation required to identify the language in which their tweets were written. However, in our sample of 113 million tweets, 33% (37,649,491) were written in a language different from the language setting in the tweet author's profile. One possible reason for this difference is that whereas Twitter users may prefer to interact with the Twitter website in their native language, the international nature of social media may encourage communication in other more widely spoken languages such as English. However, the 140 character limit on Tweets poses particular challenges for language detection as many detection algorithms are trained on longer documents (Hale et al. 2012) and there is further work to be done on fine-tuning such approaches to take account of the use of acronyms, slang and non-standard punctuation and grammar.

Location

5.1 Twitter provides three opportunities for collecting geographical information about Twitter users: from the user profile, from geo-tagged tweets, and from the content of tweets. The first opportunity comes from the location field in the Twitter user's profile. This is the location where Twitter users say they live. When we analyse user-supplied locations, we face the same challenges that we face when analysing user-supplied names, i.e. some users write nothing at all, some users write sequences of nonsensical letters, punctuation and other symbols, and some users write actual place names but customise them with creative use of spelling, word boundaries, symbols and punctuation. Another challenge is that some users lie about where they live, which is particularly obvious when they make humourous or wishful references to real or imaginary places such as 'on the beach' or 'in hell'.

5.2 Despite the challenges of deriving geographic information from the location field of Twitter user profiles, we have been successful at identifying the country for over 50% of the Twitter users we process. After cleaning the location field with techniques similar to those we use to clean the Twitter user's name field we use the Yahoo! PlaceFinder (2012) geographic database to extract location information. Yahoo! PlaceFinder returns rich, hierarchical location descriptions that include the city, county and country in which a location is found as well as the latitude and longitude co-ordinates of the location. Yahoo! PlaceFinder can return accurate location information from very small snippets of text, such as the snippets found in the location field of Twitter user profiles after text cleaning. For example, given the partial postcode SW19, Yahoo! PlaceFinder can identify this postcode as belonging to the Wimbledon district in London.

5.3 Although Yahoo! PlaceFinder is a rich source of geographical information its main drawback for large-scale user-location efforts is that it allows only 50,000 location identifications per day. To put this in context, it would take 260 days to process our sample of 13 million Twitter users at a rate of 50,000 per day. To speed the user-location process we selected a one per cent sub-sample from our sample of 113 million tweets by taking every 100th tweet which gave us a sub-sample of 1,138,266 tweets. Of those 1,138,266 tweets 73,025 were written by Twitter users already present in our sub-sample. We removed tweets written by Twitter users already present in our sub-sample to produce our final sub-sample of 1,065,257 tweets which were authored by 1,065,257 different Twitter users.

5.4 In our sub-sample of 1,065,257 Twitter users we were able to locate the country for 52% of users, the state[4] for 43% of users, the county for 36% of users, the city for 40% of users and the postcode for 10% of users (as visualised in Figure 2). Although these percentages may seem small we apply the same reasoning to location identification as we applied to gender identification above, i.e. that the number of social media users in general, and the number of Twitter users in particular, is so vast that even these small percentages produce sample sizes orders of magnitude larger than the samples sizes available to traditional social science.

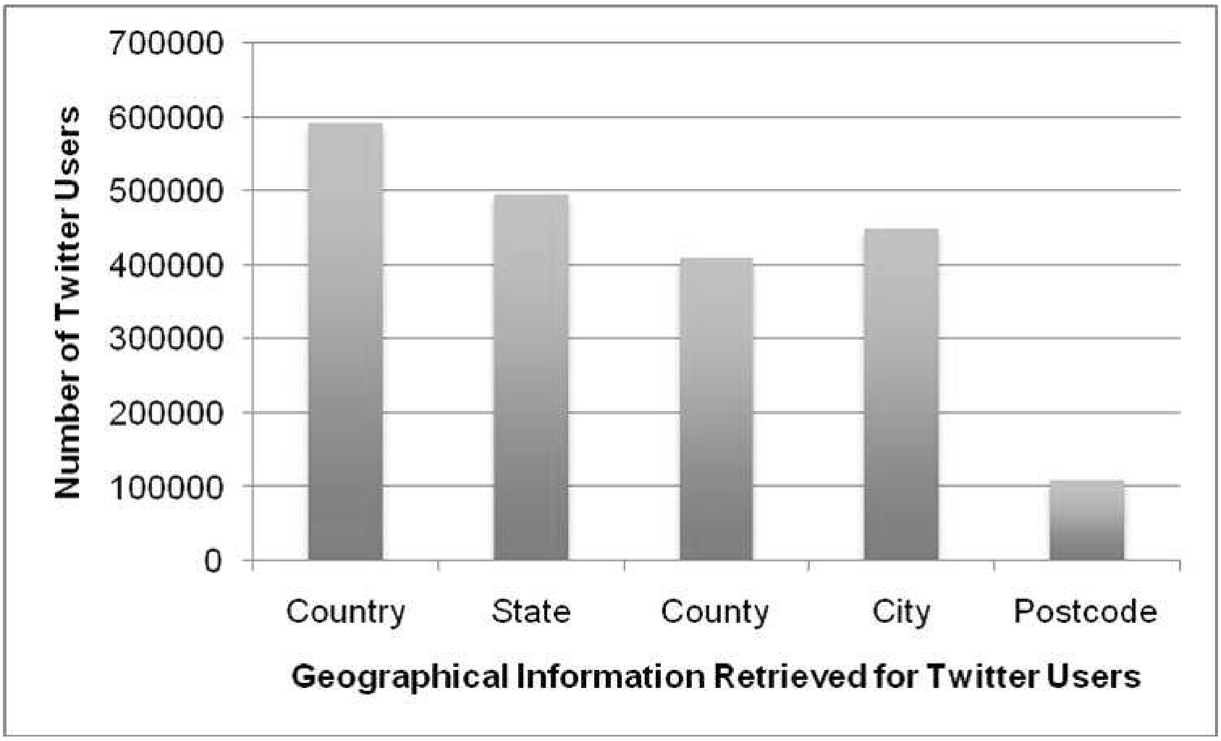


Figure 2. The number of Twitter users for which we were able to identify country, state, county, city and post code geographical information from a sub-sample of 1,065,257 million tweets.

5.5 Table 2 lists the number of tweets sent from the 50 most-tweeted-from countries, as determined by our analysis of each Twitter user's location entry in their profile. Deriving geographical information about Twitter user locations, such as the city, county and latitude and longitude co-ordinates is important for combining or 'mashing' traditional curated data sources with Twitter social media data. For example, when we know the district in which a Twitter user lives we can look up a wide range of statistics about their area from the Indices of Multiple Deprivation. Similarly, using the latitude and longitude co-ordinates of a location, we can use the Police Crime Database to look up the frequency of different types of crime perpetrated near that location.

Table 2: The number of tweets sent from the 50 most-tweeted-from countries, as determined by our analysis of each Twitter user's location entry in their profile.

1	United States	232,354	25	The Netherlands	3,143
2	Indonesia	54,516	27	Kuwait	2,869
3	Brazil	39,822	28	Norway	2,712
4	United Kingdom	30,347	29	Russia	2,614
5	Spain	16,689	30	Singapore	2,594
6	Mexico	14,012	31	South Africa	2,491
7	Japan	13,387	32	Portugal	2,125
8	Canada	12,035	33	Burkina Faso	2,090
9	France	9,428	34	Ireland	1,966
10	India	8,250	35	Nigeria	1,940
11	Argentina	7,884	36	United Arab Emirates	1,891
12	Malaysia	7,757	37	Ecuador	1,844
13	Italy	7,596	38	Sweden	1,842
14	Turkey	7,327	39	Dominican Republic	1,827
15	Philippines	6,394	40	Guatemala	1,821
16	Belgium	6,370	41	Egypt	1,731
17	Venezuela	6,353	42	Taiwan	1,533
18	Colombia	5,946	43	Switzerland	1,498
19	Saudi Arabia	4,851	44	Denmark	1,494
20	Australia	4,702	45	China	1,338
21	Netherlands	4,408	46	Austria	1,276
22	Thailand	4,391	47	Poland	1,073
23	Germany	4,214	48	Peru	1,071
24	South Korea	3,791	49	Paraguay	1,036
25	Chile	3,306	50	New Zealand	923

5.6 Mashing traditional and social media data in this way leads to new opportunities for analysing and visualising data about populations that were not possible before the rise of social media. For example, we could identify tweets that contain racist terms and by locating these within existing geographies we can show which areas can be characterised by proportionally high levels of racist sentiment. This can then be cross-referenced with Census 2011 data on the religious and ethnic population composition and the relationship between tweet sentiment and contextual area factors can be explored. However, using profile information to establish location is not without its problems. Where a location is explicitly artificial it is easy to identify (e.g. made-up place names such as 'Hell' or 'on the beach') but we cannot control for inaccurate or false content. It is also not clear whether the location field refers to where the tweeter is from, where they live or where they work and this applies to all geographies from country to postcode. Even when the location information is accurate it may not be useful for research purposes. A tweeter who has their location as 'London' may live and work in this city, but the sheer size of the area covered reduces the utility of this information and the same can be said for Manchester, Birmingham and other large cities.

5.7 The second opportunity for collecting geographical information about Twitter users comes from the tweets themselves and is a more reliable source than profile information. Each tweet can be geo-tagged with the latitude and longitude co-ordinates of the place where the tweet was authored. In most cases geo-tagging is performed when tweets are sent from mobile devices such as smart phones, tablets and laptops. Such devices do not need to be fitted with special GPS equipment because Wi-Fi and 3G network service providers calculate the latitude and longitude of mobile devices.

5.8 Despite the ubiquity of mobile devices the proportion of geo-tagged tweets is very small. In our sample of 113 million tweets only 0.85% (966,082) were geo-tagged. There are two main reasons for such a low geo-tagging rate. First, geo-tagging is turned off by default on most mobile devices and many people do not know how to activate geo-tagging or even that their mobile device is capable of geo-tagging their tweets. Second, there is increasing concern over privacy issues and leaving a digital trail (Lang 2007; Kapadia et al. 2007; Debatin et al. 2009; Coll et al. 2011). For example, it is straightforward to map the whereabouts of Twitter users that have geo-tagging enabled on their mobile device. Such a map could be annotated with the date and time Twitter users visit each location and could also show which other Twitter users were near those locations at the same time (as long as those other users tweeted on a mobile device with geo-tagging enabled).

5.9 Figure 3 plots each of the 966,082 tweets from our 113 million sample that were geo-tagged with latitude and longitude information. The dotted-areas in Figure 3 follow closely the population of the Earth whereas the dot-free areas follow the regions of the Earth known to have little or no population, such as Northern Canada and the Arctic, North-Eastern Russia and Central Australia. Figure 3 also reveals that people tweet aboard ships and other vessels, as shown by the many dots located over the seas and oceans of the world.

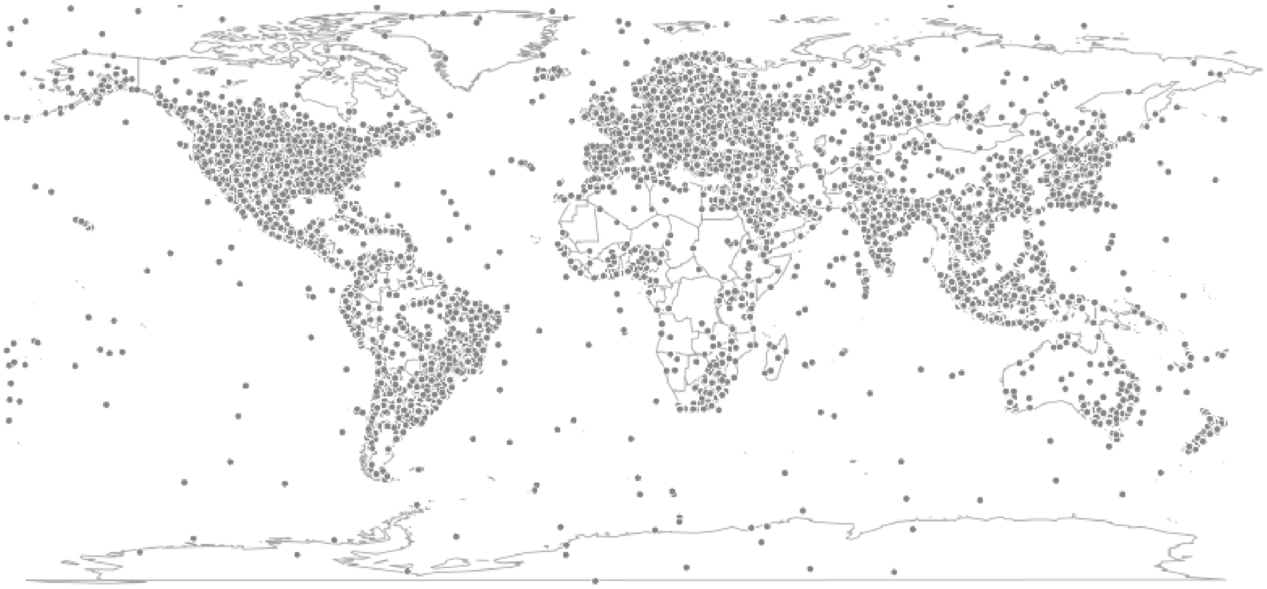


Figure 3. The geographic distribution of the 966,082 tweets from our 113 million sample that were geo-tagged with latitude and longitude information.

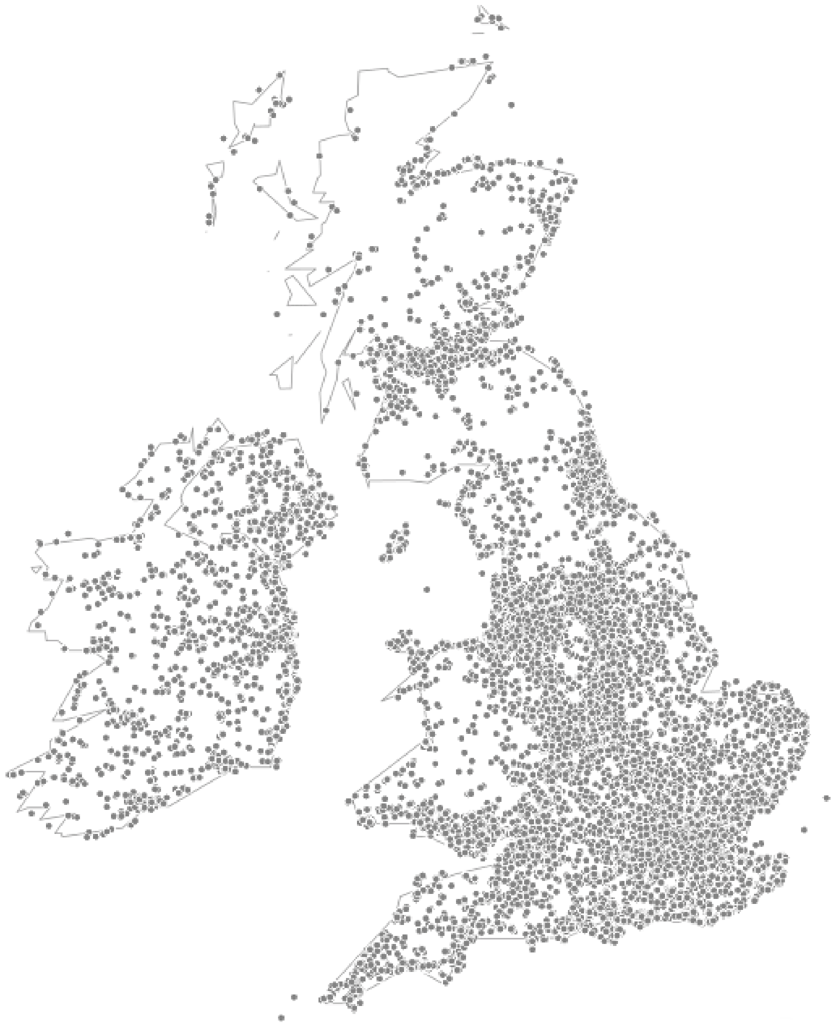


Figure 4. The geographic distribution of the geo-tagged tweets in our sample sent from the United Kingdom and the Republic of Ireland mirrors the population distribution.

5.10 The scale of the map in Figure 3 makes it difficult to see how closely the geographic distribution of tweets mirrors the geographic population distributions of individual countries. Zooming in to show just the geo-tagged tweets in our sample sent from the United Kingdom and the Republic of Ireland, Figure 4 shows clearly that the Twitter users in our sample sent tweets in proportion to the population densities of the United Kingdom and the Republic of Ireland. However it is unlikely that the small proportion of users with geocoding enabled are representative of the wider Twitter population. Indeed, 'the division between geocoding and non-geocoding users is almost certainly biased by factors such as social-economic status, location, education' (Hale et al. 2012: 2).

5.11 The third opportunity for collecting geographical information about Twitter users involves extracting the places mentioned in their tweets. However, extracting location information from the text of tweets requires much deeper analyses than the techniques described so far and we also need to consider the context in which a location is mentioned (i.e. is it where a user is from, going to, visiting or simply commenting on?). These issues aside, if we were to derive geographical information from the tweet content itself; then one approach would be to use natural language processing to parse the meaning of each tweet. Another approach would be to use probabilistic methods to estimate the likelihood that a given Twitter user lives in a particular place based on the locations they mention in their tweets. For example, Cheng et al. (2010) use a probabilistic model to estimate the likelihood of Twitter users living in a particular city. Using only the text of several hundred tweets written by a Twitter user, Cheng et al.'s technique can place that user within 100 miles of their city with 51% accuracy. We have not attempted to extract user locations from tweet texts for the same reason that we have not attempted to identify gender, i.e. accumulating sufficient tweets written by each user and subjecting them to a probabilistic analysis is a multi-step operation better suited to processing archived rather than real-time Twitter data. However, extracting location information from tweet texts is an interesting area for further work.

Ethics

6.1 A clear set of concerns for NSM analysis within the context of the social sciences include ethics. This involves the application of ethical consideration to the harvesting and archiving of data. Whilst social media platforms such as 'Twitter' can be understood as a mediatised public digital 'agora', it is also subject to conditions of service and use that need to be interpreted and engaged with through the lens of an established frame of social science research. A key concern here is anonymity and data storage; the development and application of ethics in this context necessarily involves a consideration of principles and the development of protocols that manifest themselves not only as ethical procedures but also engineering applications. In this sense, the analysis of social media streams such as Twitter (via digital observatories and platforms that are, in turn, in a continual process of development) necessarily concerns itself with what Parker (2010) calls 'moral architecture', i.e. a practice where moral and ethical procedures become inscribed into the workflow of said digital platforms and observatories. The Association of Internet Researchers ethical guidelines (AoIR) highlight three key areas of tension: the question of human subjects online; data/text and personhood; and the public/private divide (AoIR 2012). Firstly, the notion of the 'human subject' is complicated when applied to online environments. For example, can we say 'avatars' are human subjects? Does digital representation and automation of some online 'behaviours' call into question the definition of human subjects in Internet based research? If so, then it may be more appropriate and relevant to talk of harms, vulnerabilities, personal identifiable information and so on. Secondly, the Internet complicates the conventional construction of 'personhood' and the 'self', questioning the presence of the human subject in online interactions. Again, can we say an avatar is a person with a self? Is digital information an extension of a person? In some cases this may be clear-cut: emails, instant message chat, newsgroup posts are easily attributable to the persons that produced them. However, when dealing with aggregate information in 'big social data' repositories, such as collective sentiment scores for sub-groups of twitter users, the connection between the object of research and the person who produced it is more indistinct. Attribute data on very large groups of anonymised twitter users could be said to constitute non-personalised information, more removed from the human subjects that produced the interactions as compared to, say, an online interview. In these cases, the AoIR (2012: 7) guidelines state 'it is possible to forget that there was ever a person somewhere in the process that could be directly or indirectly impacted by the research'. Anonymisation procedures for big social data are in their infancy and researchers are yet fully cognizant of the factors that may result in 'deductive disclosure' of identity and subsequent potential harms (Narayanan & Shmatikov 2008, 2009). Thirdly, it is accepted that people who use online 'public' spaces can perceive their interaction as private (Nissenbaum 2010). This can question the use of data aggregators that make accessible to the public, data on interactions that were intended for private consumption. The AoIR (2012: 7) guidelines state that social, academic and regulatory delineations of the public-private divide may not hold in online contexts and as such 'privacy is a concept that must include a consideration of expectations and consensus' within context. At this stage 'ethical engineering' is in its infancy as far as social scientific analysis of NSM is concerned. However, the use of proxies, data augmentation, archiving and harvesting need to be informed and develop within an emerging ethical context that is able to balance the digital public sphere with commercial interests, the privacy and protection of individual citizens and the requirements of critical and public social science.

Conclusion

7.1 In this paper, we demonstrated a range of techniques for collecting or estimating demographics from Twitter data including analysing gender, language and location. We showed that rich demographics can be derived from Twitter user profiles and tweets that are not explicitly stated in either of these sources. We also showed that Twitter users are found all over the world and tweet in at least 53 languages and our plots of the geographic distribution of Twitter users confirm that our search for a first-name gender database containing a wide variety of names from around the world was warranted.

7.2 Through the development of these methodologies and techniques we have demonstrated that it is possible to assess the representativeness of Twitter data to a limited degree. Within the social sciences this will provide a foundation for further research into issues of sampling and perhaps even inference. We have also provided reasonably robust solutions to estimating important demographic information either directly or by proxy from user profiles, tweets or metadata which will enable tweet data to be contextualised alongside 'traditional' social variables. Of particular salience are our methods for using geographical location as a link between virtual and terrestrial datasets, allowing the researcher to know the socio-economic characteristics of the area in which the person is tweeting or where the tweeter is from. The ability to link a tweet to an area with known demographic characteristics enables us to ask a plethora of new research questions. For example, we could identify tweets that use racist language, geolocate them and use Census data to investigate the ethnic composition of an area, level of migration, unemployment and deprivation in an attempt to understand the explanatory factors behind prejudice. Alternatively we could identify tweets referring to personal safety and security and cross-reference this with neighbourhood crime data to investigate the impact of criminal activity on individual lives.

7.3 The potential importance of social media to social science is too great to simply write off as an unviable source of useful and rich data and whilst these approaches are not without their problems and inaccuracies we hope that, by putting a flag in the ground at this early stage of methodological development, we will encourage others to critically engage and develop these proposals further.

Acknowledgements

This work is based on the project *Digital Social Research Tools, Tension Indicators and Safer Communities: a Demonstration of the Cardiff Online Social Media ObServatory (COSMOS)* which was funded by the Economic and Social Research Council (ESRC) under the Digital Social Research Demonstrator Programme (Grant Reference: ES/J009903/1). The parallel project *COSMOS: Supporting Empirical Social Scientific Research with a Virtual Research Environment* is funded by the Joint Information Systems Committee (JISC).

Notes

[1] However, recent analysis has revealed that the majority of users of these sites rarely contribute (post updates) instead using the services to keep up to date with social, professional and news related developments (see <http://blog.twitter.com/2011/09/one-hundred-million-voices.html>)

[2] A petabyte of information is equivalent to one thousand terabytes or one million gigabytes.

[3] 40N has two other categories called 'mostly male' and 'mostly female'. For simplicity, we consider 'mostly male' names to be male names and 'mostly female' names to be female names.

[4] A state is the level of political geography below country. Examples include US states such as California, UK constituent countries such as England, and Japanese prefectures such as Nara.

References

AOIR (2012) *Ethical Decision-Making and Internet Research: Version 2.0 – Recommendations from the Association of Internet Researchers Working Committee*. Available at: <http://aoir.org/reports/ethics2.pdf> [Accessed 22-02-13].

ARGAMON, S., Koppel, M., Fine, J. & Shimoni, A. R. (2006) 'Gender, genre, and writing style in formal written texts', *Text – Interdisciplinary Journal for the Study of Discourse*, 23(3) p. 321–346.

ASUR, S. & Huberman. B. (2010) *Predicting the Future with Social Media*. Social Computing Lab, HP Labs in Palo Alto.

BARRACUDA LABS Internet Security Blog (2012) *The Twitter Underground Economy: A Blooming Business*, <https://www.barracuda.com/blogs/labsblog?bid=2989> [accessed 11-03-13].

BRUNS, A., Burgess, J., Crawford, K. & Shaw, F. (2012) '#qldfloods and @QPSMedia: Crisis communication in the 2011 South East Queensland floods', *Research Report – Arc Centre of Excellence for Creative Industries and Innovation.* <http://www.mappingonlinepublics.net/dev/wp-content/uploads/2012/01/qldfloods-and-@QPSMedia.pdf> [Accessed 11-03-13].

BRUNS, A., Burgess, J., Highfield, T., Kirchhoff, L. & Nicolai, T. (2011) 'Mapping the Australian networked public sphere', *Social Science Computer Review*, 29(3) p. 277–287

BURGESS, J. & Bruns, A. (2012) 'Twitter archives and the challenges of "big social data" for media and communication research', *M/C Journal*, 15(5) p. 1–7.

BURNAP, P., Rana, O. & Avis, N. (2013) 'Making sense of self reported socially significant data using computational methods', in Housley, W. Edwards, A. Williams, M. & Williams, M. (Eds.), Special Issue, Computational Social Science: Research, Design and Methods, *International Journal of Social Research Methods*, 16(3) p. 215–230.

CHENG, Z., Caverlee, J. & Lee, K. (2010) 'You are where you tweet: a content-based approach to geo-locating twitter users', *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM '10)*.

CHENG, N., Chandramouli, R. & Subbalakshmi, K. P. (2011) 'Author gender identification from text', *Digital Investigation*, 8(1) p. 78–88

CHU, Z., Widjaja, I. & Wang, H. (2012) 'Detecting social spam campaigns on Twitter', Proceedings of the Tenth ACNS International Conference, [doi://dx.doi.org/10.1007/978-3-642-31284-7_27]

COLL, S., Glassey, O. & Balleys, C. (2011) 'Building social networks ethics beyond "privacy": a sociological perspective', *International Review of Information Ethics*, 16 p. 47–53.

DEBATIN, B., Lovejoy, J. P., Horn, A. & Hughes, B. N. (2009) 'Facebook and online privacy: attitudes, behaviours and unintended consequences', *Journal of Computer-Mediated Communication*, 15 p. 83–108.

DODDS, P. S. & Danforth, C. M. (2010) 'Measuring the happiness of large-scale written expression: songs, blogs, and presidents', *Journal of Happiness Studies*, 11(4) p. 441–456.

EDWARDS, A., Housley, W., Sloan, L., Williams, M. & Williams, M. (2013) 'Computational social science and methodological innovation: surrogacy, augmentation or reorientation?', in Housley, W. Edwards, A. Williams, M. & Williams, M. (Eds.) Special Issue, Computational Social Science: Research, Design and Methods, *International Journal of Social Research Methods*, 16(3) p. 245–260.

FININ, T., Murnane, W., Karandikar, A., Keller, N, Martineau, J. & Dredze, M. (2010) 'Annotating named entities in Twitter data with crowdsourcing', in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT '10)* (p. 80–88). Stroudsburg, PA: Association for Computational Linguistics.

GAYO-AVELLO, D. (2012) *I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper: A Balanced Survey on Election Prediction using Twitter Data*. Department of Computer Science, University of Oviedo (Spain).

GINSBERG, J., Mohebbi, M. H., Patel, R. S., Brammer, L. Smolinski, M. S. & Brilliant, L. (2009) 'Detecting influenza epidemics using search engine query data', *Nature*, 19(457) p. 1012–1014.

HALE, S., Gaffnet, D. & Graham, M. (2012) *Where in the world are you? Geolocation and language identification in Twitter*. Oxford Internet Institute Working Paper, <http://www.geospace.co.uk/files/icwsm_paper2.pdf> [Accessed 11-03-13]

HOUSLEY, W. Edwards, A. Williams, M. & Williams, M. (Eds.) (2013) Special Issue, Computational Social Science: Research, Design and Methods, *International Journal of Social Research Methods*, 16(3).

HUFFAKER, D. A. & Calvert, S. L. (2005) 'Gender, identity and language use in teenage blogs', *Journal of Computer-Mediated Communication*, 10(2) article 1. <http://jcmc.indiana.edu/vol10/issue2/huffaker.html>

JAFFE, J. M., Lee, Y. E., Huang, L. & Oshagan, H. (1999) 'Gender identification, interdependence and pseudonyms in CMC: language patterns in an electronic conference', *The Information Society*, 15 p. 221–234

KAPADIA, A., Henderson, T., Fielding, J. J. & Kotz, D. (2007) 'Virtual walls: protecting digital privacy in pervasive environments', *Pervasive Computing*, 4480 p. 162–179.

LANG, P. G. (2007) 'Publicly private and privately public: Social networking on YouTube', *Journal of Computer-Mediated Communication*, 13(1) article 18. <http://jcmc.indiana.edu/vol13/issue1/lange.html>

LEWIS, P., Newburn, T., Taylor, M., Mcgillivray, C., Greenhill, A., Frayman, H. & Proctor, R. (2011) 'Reading the riots: investigating England's summer of disorder', *Report – The Guardian and The London School of Economics*. <http://eprints.lse.ac.uk/46297/1/Reading%20the%20riots%28published%29.pdf> [Accessed 11-03-13].

MENDES, P., Passant, A. & Kapanipathi, P. (2010) 'Twarql: tapping into the wisdom of the crowd', Proceedings of the Sixth International Conference on Semantic Systems. <http://dl.acm.org/citation.cfm?id=1839762>

MICHAEL, J. (2007) *40000 Namen, Anredebestimmung anhand des Vornamens*, <http://www.heise.de/ct/ftp/07/17/182/> (in German) [accessed 15-10-12].

MISLOVE, A., Lehmann, S., Ahn, Y-Y., Onnela, J. P. & Rosenquist, J. N. (2011) 'Understanding the demographics of Twitter users', *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*.

NARAYANAN, A., & Shmatikov, V. (2009) 'De-anonymizing social networks', *IEEE Symposium on Security & Privacy*. Oakland, CA. Available: <http://www.cs.utexas.edu/~shmat/shmat_oak09.pdf>

NARAYANAN, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets (How to break anonymity of the Netflix prize dataset.) *IEEE Symposium on Security & Privacy*. Oakland, CA. Available: <http://arxiv.org/pdf/cs/0610105v2>

NISSENBAUM, H. (2010). *Privacy in context: Technology, Policy, and the Integrity of Social Life*. Stanford: Stanford University Press.

OFFICE FOR NATIONAL STATISTICS (2011) *2011 Census – Population and Household Estimates for England and Wales*.

PARKER, M. (2010) 'Ethical and moral dimensions of e-research', in Dutton, W. & Jeffreys, P. (Eds), *World Wide Research: Reshaping the Sciences and Humanities*. Cambridge, MA: MIT Press.

LANGUAGE DETECTION LIBRARY FOR JAVA, <http://code.google.com/p/language-detection/> [last accessed October 2012].

SAKAKI, T., Okazaki, M. & Matsuo, Y. (2010) 'Earthquake shakes Twitter users: real-time event detection by social sensors', presented at WWW 2010, April 26–30, Raleigh, NC, USA.

SAVAGE, M. & Burrows, R. (2007) 'The coming crisis in empirical sociology', *Sociology*, 41 p. 885–899.

SAVAGE, M. & Burrows, R. (2009) 'Some further reflections on the coming crisis of empirical sociology', *Sociology*, 43 p. 762–722.

SCARFI, M. (2012) *Social Media and the Big Data Explosion*. Forbes. Available at: <http://www.forbes.com/sites/onmarketing/2012/06/28/social-media-and-the-big-data-explosion/> [Accessed 15-11-12].

SMITH, D. (2012), *How Many People Use the Top Social Media?*. Digital Market Ramblings. Available at: <http://expandedramblings.com/index.php/resource-how-many-people-use-the-top-social-media/> [Accessed 15-07-12].

TAM, D. (2012) *Facebook Processes More than 500TB of Data Daily*. CNet. Available at: <http://news.cnet.com/8301-1023_3-57498531-93/facebook-processes-more-than-500-tb-of-data-daily/> [Accessed 15-11-12].

THOMSON, R. & Murachver, T. (2001) 'Predicting gender from electronic discourse', *British Journal of Social Psychology*, 40 p. 2193–2208.

THELWALL M., Buckley, K., Paltoglou, G., Cai, D. & Kappas, A. (2010) 'Sentiment strength detection in short informal text', *Journal of the American Society for Information Science and Technology*, 61(12) p. 2544–2558.

TUMASJAN, A., Sprenger, T. Sandner, P. & Welpe, I. (2010) 'Predicting elections with Twitter: what 140 characters reveal about political sentiment', *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*.

US SOCIAL SECURITY ADMINISTRATION *Top Names Over the Last 100 Years*. <http://www.ssa.gov/oact/babynames/decades/century.html> (US Social Security Administration Website) [Accessed 15-10-12].

YAHOO! PlaceFinder, <http://developer.yahoo.com/geo/placefinder/> [last accessed October 2012].

YUEN, M. C., King, I. & Leung.K, S. (2011) 'A survey of crowdsourcing systems', *In SocialCom '11: Proceedings of The Third IEEE International Conference on Social Computing*.