Computers, Environment and Urban Systems 58 (2016) 85–96

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

The geography of Twitter topics in London

Guy Lansley *, Paul A. Longley

Department of Geography, Pearson Building, University College London, Gower Street, London WC1E 6BT, United Kingdom

ARTICLE INFO

Article history:
Received 19 October 2015
Received in revised form 18 March 2016
Accepted 4 April 2016
Available online 17 April 2016

Keywords:
Social media
Twitter
Topic modelling
Latent Dirichlet Allocation
Geotemporal

ABSTRACT

Social media data are increasingly perceived as alternative sources to public attitude surveys because of the volume of available data that are time-stamped and (sometimes) precisely located. Such data can be mined to provide planners, marketers and researchers with useful information about activities and opinions across time and space. However, in their raw form, textual data are still difficult to analyse coherently and Twitter streams pose particular interpretive challenges because they are restricted to just 140 characters. This paper explores the use of an unsupervised learning algorithm to classify geo-tagged Tweets from Inner London recorded during typical weekdays throughout 2013 into a small number of groups, following extensive text cleaning techniques. Our classification identifies 20 distinctive and interpretive topic groupings, which represent key types of Tweets, from describing activities or informal conversations between users, to the use of check-in applets. Our motivation is to use the classification to demonstrate how the nature of the content posted on Twitter varies according to the characteristics of places and users. Topics and attitudes expressed through Tweets are found to vary substantially across Inner London, and by time of day. Some observed variations in behaviour on Twitter can be attributed to the inferred demographic and socio-economic characteristics of users, but place and local activities can also exert a considerable influence. Overall, the classification was found to provide a valuable framework for investigating the content and coverage of Twitter usage across Inner London.

1. Introduction

With a global reach of 500 million Tweets transmitted each day by over 300 million users globally (Twitter, 2015), Twitter could potentially provide a valuable source of social data. Unfortunately the structure of Twitter content is inherently hard to interpret and analyse, not least because the messages are exclusively text strings restricted to just 144 characters. There have been many recent developments in text mining techniques applied to Tweets in order to quantifiably interpret their content. Topic modelling poses as an interesting opportunity to develop a generalised understanding of the dynamics of Twitter usage, and is used here to explore variation in Twitter usage across Inner London. This work provides in turn, a platform for developing a more thorough understanding of the relationship between human activities, user characteristics and behaviour on social media.

Roughly 80% of active Twitter users access the service via a mobile telephone (Twitter, 2015), and about 1% of users opt to share their locations based on the coordinates of their devices. It is probable that the nature of posts on Twitter varies systematically according to location, and also the time of day — because of the nature of popular activities, and the loci of activities of individuals that have different social characteristics. Inductive generalisation about the geography of topics on social media can thus contribute to understanding the social dynamics of urban areas. Moreover, the ability to quantify observed social trends across time and space is of great value to retailers and marketers, including out-of-home advertising companies who rent space on digital billboards that can be updated in real-time.

In this paper we investigate the differences in observed Tweeting behaviour as users move around the city. Our research seeks to link typical behaviours to observable characteristics, in the broad analytic tradition of geodemographics, which extends from factorial ecology of the 1960s (see Harris, Sleight, & Webber, 2006), to novel nomenclatures based upon data mining (Spielman & Thill, 2008). We hypothesise that the content of Tweets bears an identifiable correspondence with personal characteristics, location, and activity. There is also likely to be a temporal rhythm to such activities. Such variations are unlikely to be picked up in conventional geodemographic classifications, which associate individuals only with night-time residence (see Singleton & Longley, 2015). As such, our motivation is to identify the trends on Twitter during typical weekdays in Inner London using a large sample of geo-tagged Tweets. Our primary aim is to investigate how the key trends in behaviour on Twitter vary across space and time, and also according to different user characteristics. Our unsupervised topic modelling approach produces a readily interpretable classification of Tweets based on the use of words, and we discuss how key Twitter topics vary across the dataset.

* Corresponding author.
E-mail address: g.lansley@ucl.ac.uk (G. Lansley).

http://dx.doi.org/10.1016/j.compenvurbsys.2016.04.002

2. Background

Largely because of their ready availability, analysis of Twitter data has received much attention from the academic community. Most research has focused on the content of Tweet messages and the characteristics of Twitter users (Williams, Terras, & Warwick, 2013), and Twitter has become a popular data source for opinion mining and trend tracking. Furthermore, the sub-sample of Tweets that are geolocated facilitate profiling of usage across space as well as time — although we are unaware of any attempt to use topic modelling to do this over any geographically extensive area. Despite the sample bias, geo-tagged Tweets have been found to be a useful tool for urban research (Longley, Adnan, & Lansley, 2015).

Research into the content of geotagged Tweets has ranged from identifying new trends across time and space (Kwak, Lee, Park, & Moon, 2010) to tracking specific topics (Signorini, Segre, & Polgreen, 2011). Due to the volume of data, many researchers have employed various text mining techniques to achieve quantitative insights from Tweets, including sentiment analysis and topic modelling (e.g. Chamlertwat, Bhattarakosol, & Rungkasiri, 2012; Hong & Davison, 2010). Research has used such techniques to accommodate Twitter data into opinion polling (O'Connor, Balasubramanyan, Routledge, & Smith, 2010; Tumasjan, Sprenger, Sandner, & Welpe, 2010), although validity of the findings have been criticised (Gayo-Avello, Metaxas, & Mustafaraj, 2011). Still other research has investigated the association between sentiment on Twitter and short-term stock market trends (Bollen, Mao, & Zeng, 2011). Whilst Twitter data have previously been used to identify unusual events in space and time (Chae et al., 2012), and as a tool to model sentiment across space (Quercia, Ellis, Capra, & Crowcroft, 2012), there remains a dearth of research at intra-urban scales of analysis –one notable example is Andrienko et al.'s (2013) study of Seattle, but this falls short of linking spatial variation in content to user characteristics.

One of the most fascinating and useful directions of social media data research has focused on social sensing (Liu et al., 2015). Through social media platforms and web services individuals can record data on their surroundings, activities and/or opinions, effectively acting as sensors themselves (Goodchild, 2007). Whilst much research has considered social media data as a means of identifying and understanding unusual and unique phenomena, there is perhaps greater value in using them to monitor places and their typical activity patterns. Such data could, therefore, provide unique information about the social dynamics of places that is not obtainable at a grand scale from traditional approaches (Liu et al., 2015). Consequently, some research has attempted to use georeferenced social media data to estimate daily footfall profiles of locations and to predict land uses (McKenzie, Janowicz, Gao, & Gong, 2015; Jiang, Alves, Rodrigues, Ferreira, & Pereira, 2015). However, there have been limited attempts to explore typical activities from textual components of the data.

In this paper we extend such research by using georeferenced Tweet content to reveal the typical weekday patterns of social media activity across a city, differentiated by modelled user characteristics. We illustrate how segmenting Tweets by content type and linkage of results with other datasets (specifically land-use and Census statistics) can contribute to our understanding of urban dynamics.

3. Data

The Twitter data for this study were sourced using Twitter's filtered streaming API between 1st January 2013 and 31st December 2013. The stream collected georeferenced Tweets only. Whilst the streaming API only obtains 1–2% of the complete feed, previous research has found it can still extract over 90% of all geo-tagged posts (Morstatter, Pfeffer, Liu, & Carley, 2013). Our case study is Inner London (Fig. 1) as the rate and density of Tweeting activity are great enough to identify typical trends at intra-urban scales of analysis.

The 2011 Census of Population recorded just over 3.2 million residents within this study area, and a workplace population of almost 2.7 million. The density of the work day population (comprising those who work there plus other non-working residents) is 213.3 persons per hectare. This is much higher than the over-all average of 114.0 for Greater London. Partly as a consequence, the density of Tweets sent from Inner London over the study period was 2.27 times higher than the average for Greater London.

3.1. Data cleaning

Given the focus of this paper upon typical weekday topics in London, Tweets transmitted during the weekends were omitted from the study. Those sent on Mondays and Fridays were also removed as parts of these days are influenced by weekend activities, rendering them unrepresentative of a full 24 hour weekday cycle.

The Twitter data required a thorough clean to ensure that text mining would identify valid and representative patterns of user opinions. First, the following words were removed from the data:

- Words with fewer than three characters as they would not be informative in a topic model
- Words with more than 16 characters to reduce the occurrence of words with low frequencies, many of which would be bespoke tags
- Stopwords (selected from the R English Stopword library) as they are largely uninformative in a topic model and their removal would improve the efficiency of the text mining
- URLs
- Words containing non-Latin characters

The following Tweets were also removed:

- Tweets with fewer than 3 words
- Tweets from users with over 3000 Tweets within the sample, who would tend to dominate the analysis
- Tweets from users who have posted identical messages more than three times in the data as these are likely to be fake accounts
- Tweets with uncertain coordinates, arising because of rounding of recorded values by certain user devices

All words were converted to lower case, and numbers and punctuation were removed. The cumulative effects of these changes upon the number of Tweets used in the subsequent analysis are shown in Table 1.

In total our remaining Tweet database consisted of 153,397 unique users. The most active remaining user made precisely 3000 Tweets, and almost 58,000 users only had one Tweet in the sample. The dataset was extremely positively skewed, with a median number of Tweets per user of just three.

It is desirable to relate the remaining data to some identifiable population in terms of residence, workplace and visitor status. This is in practice extremely difficult, not least because there are no recorded aggregate demographic characteristics of users. Previous research has established that Twitter users who share coordinates with Tweet messages over represent younger age cohorts, particularly those between the ages of 15 and 30 (Longley et al., 2015). This work also found over-representation of White British users. These findings are supported by market research (Ipsos, 2013), although the relevant surveys do not investigate users with geolocation enabled.

More generally, questions of uncertainty can be raised about the representativeness of Twitter as a source for opinion mining. Tweets are assumed to represent the users' opinions, but many users may be influenced by audience and tailor their comments accordingly (Marwick & Boyd, 2010). Additionally, not all Twitter accounts represent individuals and their personal opinions, but may instead represent businesses

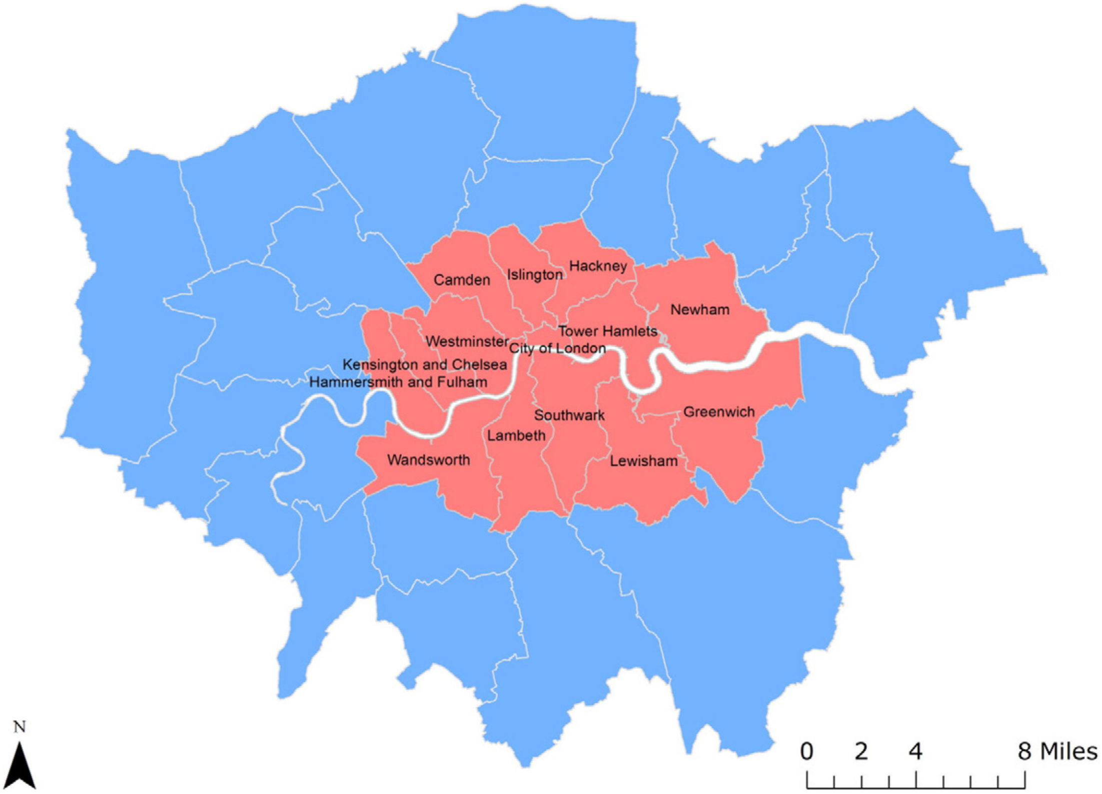


Fig. 1. The Inner London Boroughs used to define the study area (red) within Greater London (blue) (The extent of the scale bar, 8 miles, is 12.875 km.).

or journalists. It is therefore important to consider that the results from this paper reflect Twitter users, not the population at large.

4. The spatial distribution of Tweets in London

Georeferenced Tweets are of potential value in understanding the characteristics and movements of populations because they provide highly granular observations that are not restricted to residences or to workplaces. They may be particularly useful for recording specific travel routes or other places visited for services or leisure purposes. The density of all Tweets from our dataset is shown in Fig. 2.

From the map, it is apparent that the geotemporal distribution of Tweets in Inner London may reflect much more than workplace or residential locations. The highest concentration of Tweets is in Central London. It is also notable that Western Central London has a higher density of Tweets than the City of London. This is probably because the City of London offers fewer activities additional to those catering for the working population. By contrast, the western part of Central London has more tourist and leisure attractions, as well as hosting a sizeable working population. To test this assumption a standardised difference map was produced to compare the distribution of Tweets arising from the working day (10:00–16:00) and the night-time (19:00–7:00). Tweets from the rush hours have not been considered as they are not reflective of either time sample. The frequencies of Tweets from each sample were calculated across a 200 m grid and were standardised as *Z*-scores to take into consideration the uneven sample sizes. Fig. 3 presents the standardised differences between the two samples, the day-time standardised frequencies have been subtracted from the night-time equivalent.

Table 1
Residual numbers of Tweets remaining at each stage of the data cleaning.

Function	Count
Harvested Tweets from the Greater London Area in 2013	8,005,874
Tuesdays, Wednesdays and Thursdays only	3,538,308
Tweets from Inner London only	1,679,571
Coordinates cleaned	1,557,282
Fake users and users with 3000+ Tweets removed	1,545,899
Tweets with fewer than 3 words removed after cleaning	1,301,004

The places with the greatest over-representations of Tweets in the evening are not just residential areas, but also those associated with nightlife activities. This is demonstrated by the dominance of night-time Tweets in Central London's Soho district (in Westminster) and the area surrounding the O2 Arena in Greenwich. Daytime overrepresentation of Tweets largely occurs at employment hubs, such as the rest of Central London and Canary Wharf. This is consistent with previous research (Longley et al., 2015), which we develop and extend here through examination of the association between land-use and user content.

5. Topic modelling

Twitter data are challenging to model as the content is restricted to just 140 characters and is likely to include non-standard uses of language (Ramage, Dumais, & Liebling, 2010). Some have avoided the use of generative models by using lists of words to look-up and assign topics (Michelson & Macskassy, 2010). However, there is little prior knowledge of how Tweets might be segmented in London and therefore an unsupervised modelling technique was pursued.

Since its introduction in 2003, Latent Dirichlet Allocation (LDA) has become a widely used tool for probabilistic topic modelling. LDA is an unsupervised model which can be used to identify probable topics from collections of text. The approach is described in Blei, Ng, and Jordan (2003), and essentially formulates semantic groups from text documents. LDA develops probabilistic topics (or groups of words) from large collections of discrete data, and is therefore appropriate for textual analysis (Blei et al., 2003). Each value, or word, is given a probability of falling within each of the generated topics. From viewing the

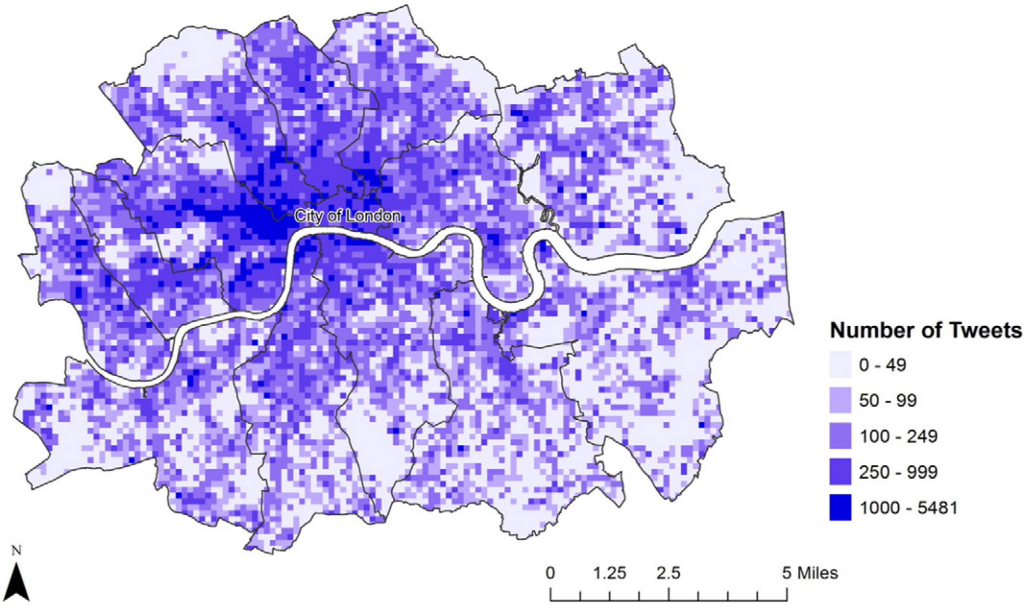


Fig. 2. The density of Tweets in Inner London in 2013. The data are displayed as the number of Tweets across a 200 m grid.

words with the highest probabilities for each group, the topic or underlying theme of the produced groups, can be deduced.

LDA has been most commonly applied on longer extracts of corpora such as online news articles and its validity on shorter document types such as Tweets has been questioned (Andrienko et al., 2013). Various extensions have been made to the basic methodology, such as a labelled LDA approach (Ramage et al., 2010). A related development is the author-topic model which groups all the Tweets belonging to individual users into single documents (Steyvers, Smyth, Rosen-Zvi, & Griffiths, 2004). In the spirit of Chae et al. (2012) we do not aggregate Tweets as individual users may Tweet about multiple topics over the course of the year.

The number of topics (k) to be generated must be specified by the user and, after experimentation, we specified 20 groups, in order that each of the groups would comprise a sizeable number of Tweets (65,000 on average). This was found to render each class representative of a broad yet distinctive theme. This approach was favoured over a hierarchical LDA which automatically determines the number groups and subgroups because this limits each word to representing only a single path (Blei, Griffiths, Jordan, & Tenenbaum, 2004). Tweets are very short, and in this latter approach a mislabelling of a single word within a Tweet could result in a misallocation of the overall post.

LDA was used to record the numbers of times words in each Tweet were assigned to each of the groups, and this was used to assign each Tweet to its most probable topic. The LDA was run 20 times to create 20 different classifications, and the optimal classification was selected in terms of both the balance between groups in sizes and subjective distinctiveness of each of the groups. Fortunately, the technique proved to be robust and there were only extremely subtle differences between each of the iterations.

With the 20 group classification finalised, the process was repeated to create subgroups, by iterating the Tweets for each of the groups

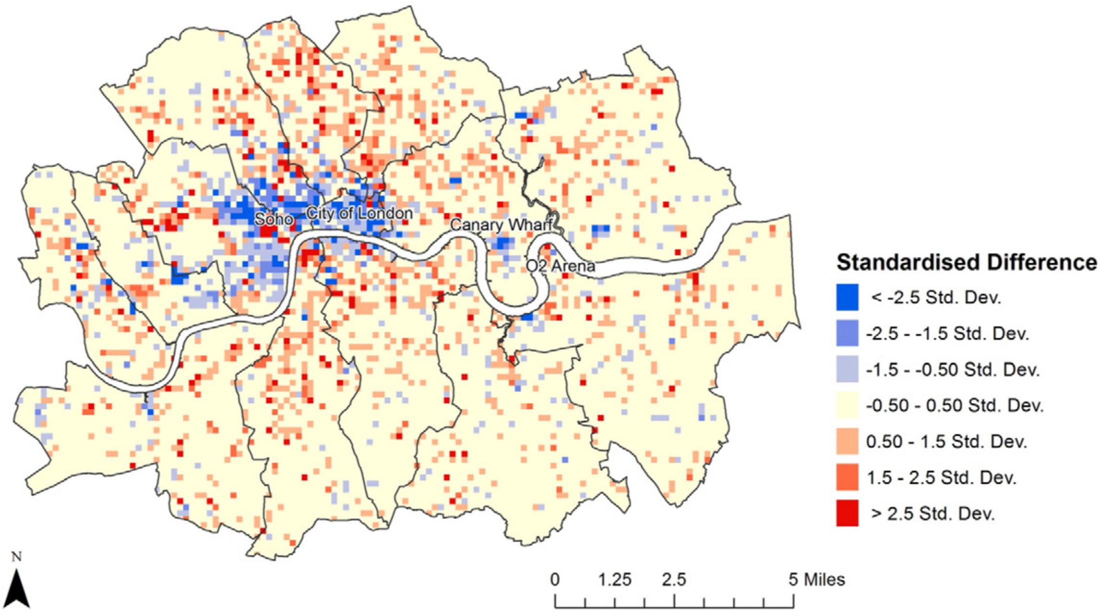


Fig. 3. The standardised difference between the densities of Tweets during the night-time (19:00–7:00) and the day-time (10:00–16:00).

through the LDA to produce 5 subgroups in each. In smaller groups, comprising less than 50,000 Tweets, a four subgroup solution was tested if the five subgroups were difficult to distinguish. This was only found to be appropriate for one group. A six subgroup solution was adopted for the largest group as each subgroup was deemed to be sufficiently distinctive.

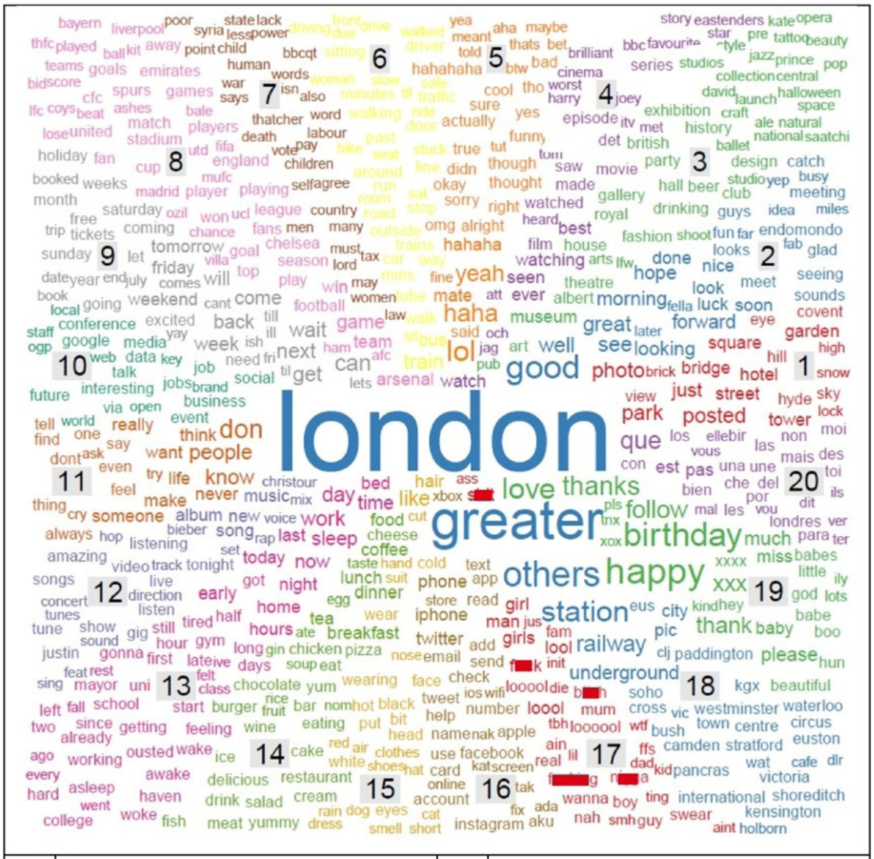
Each group was subjectively labelled to ease understanding and interpretation in subsequent geographical analysis. The LDA algorithm ranked unique words by their probabilities for each of the groups and the subgroups. These data and also the most frequently occurring words per group were used to determine the contents of each group. The labels were also validated by reviewing a small number of randomly selected Tweet messages, and checking if their assigned label was appropriate.

6. Results

The text mining and subsequent LDA method identified distinctive groups of Tweets based on the use of words. The Twitter groups were visualised using a comparison word cloud of the most popular words in the whole dataset (Fig. 4). The comparison cloud partitioned the most common words by the groups in which they are most abundant.

Each of the 20 groups comprises unique assemblages of words. Some of the groups can be associated with the discussion of popular interests, as exemplified by the groups labelled ‘TV and Film’ or ‘Sport and Games’. Other groups are more specific to comments about day-to-day activities, a good example of this being the ‘Routine Activities’ group. The model also created two groups that were more distinguishable by sentiment, one more positive and another more pessimistic. The classification has identified two groups which emerged because of uses of the Twitter service that are not primarily focused upon textual communication; first, uploading and sharing photos and second, using check-in services. Finally, the model also identified two groups that are distinguishable by their use of dialect and foreign languages: one group comprises youth slang and text abbreviations and the other is largely a non-English language group.

The classification is quite evenly balanced, with only two groups accounting for more than 6% of the data (Table 2). The largest group is the Foreign and Other group, which represents 8.9% of Tweets. Within this group only one subgroup is largely composed of English language words, and the remaining ‘foreign language’ subgroups consist of 7.25% of all Tweets. A previous study estimated that roughly 7.5% of georeferenced Tweets from London were written in a foreign language, based on a large sample of 3.3 million message from 2012 (Manley &



N	Name	N	Name
1	Photography and Sights	11	Pessimism and Negativity
2	Optimism, Kindness and Positivity	12	Music and Musicians
3	Leisure and Attractions	13	Routine Activities
4	TV and Film	14	Food and Drink
5	Humour and Informal Conversations	15	Body, Appearances and Clothes
6	Transport and Travel	16	Social Media and Apps
7	Politics, Beliefs and Current Affairs	17	Slang and Profanities
8	Sport and Games	18	Place and Check-Ins
9	Anticipation and Socialising	19	Wishes and Gratitude
10	Business, Information and Networking	20	Foreign and Other

Fig. 4. A comparison word cloud of the 20 Twitter topic groups. Offensive language has been removed from the word cloud.

Table 2
The size of each of the 20 Twitter topic groups.

Twitter topic group	Number of Tweets	Percent (%)
Photography and Tourism	65,799	5.06
Optimism, Kindness and Positivity	67,417	5.18
Leisure and Attractions	57,066	4.39
TV And Film	58,900	4.53
Humour and Informal Conversations	71,538	5.50
Transport and Travel	71,785	5.52
Politics, Beliefs and Current Affairs	67,330	5.18
Sport and Games	62,331	4.79
Anticipation and Socialising	66,084	5.08
Business, Information and Networking	55,364	4.26
Pessimism and Negativity	85,107	6.54
Music and Musicians	46,931	3.61
Routine Activities	73,239	5.63
Food and Drink	54,192	4.17
Body, Appearances and Clothes	64,151	4.93
Social Media and Apps	47,962	3.69
Slang and Profanities	65,785	5.06
Place and Check-Ins	56,999	4.38
Wishes and Gratitude	46,162	3.55
Foreign and Other	116,862	8.98
All Tweets	1,301,004	100

Cheshire, 2013). This figure is very close to our own findings despite being calculated by using an entirely different methodological approach. This also reveals that of 1.3 million Tweets, only 24,000 could not be allocated to a homogenous topic.

6.1. The temporal dimension

The temporal distributions of Tweets from each of the 20 groups, along with the over-all temporal distribution for comparison, are shown as a heat map in Fig. 5. It is clear that each of the groups manifests a distinctive temporal pattern.

The temporal patterns of each of the Twitter groups correspond with existing knowledge on their associated activities, reinforcing the validity of the classification and use of Twitter data. For instance, the Food and Drink group shows a peak during midday and again in the evening, presumably because of the occurrence of meal times. The Transport and Travel group highlights the morning rush hour, and the evening peak to a lesser extent and corresponds with Transport for London's passenger frequency data (Ceapa, Smith, & Capra, 2012). The Sport and Games group is most prominent during the evenings (19:00–22:00) when professional football games are televised live. The TV and Film group is also active during the evenings, which is reasonable as the data only consider working days.

6.2. Spatial distributions

We would also expect the composition of Twitter topics to vary across space. This is largely because of the influence of local activities and also a consequence of the uneven geodemographic distributions of people across London at all times of the day. The section first compares the distribution of Twitter topics to a general land-use classification, then explores patterns across selected places characterised by distinctive activities.

6.2.1. Land-use

Using the Generalised Land-use Database (GLUD), Tweets were filtered by land-use (DCLG, 2006). The method is limited by the accuracy of coordinates of Tweets (with an error of up to 10 m), and the fact that many land-use parcels in the GLUD can be very thin. The GLUD is a recoding of Ordnance Survey MasterMap into polygons of 9 key land-uses (including a catch-all ‘other’ category), and is precise to one metre resolution (DCLG, 2006). From the GLUD we consider three main aggregated land-use categories: residential (domestic buildings and gardens); non-domestic buildings; and public open space (or non-domestic green space). In total, 289,240 Tweets from our final dataset were recorded from residential land-uses, 241,095 from non-domestic buildings and 115,522 Tweets from areas of public green space. The three land-use categories are shown in Fig. 6.

To compare the cases, location quotients were produced by dividing the proportions of each topic out of all Tweets within each of the land-use categories by the overall proportion of the same topic across the

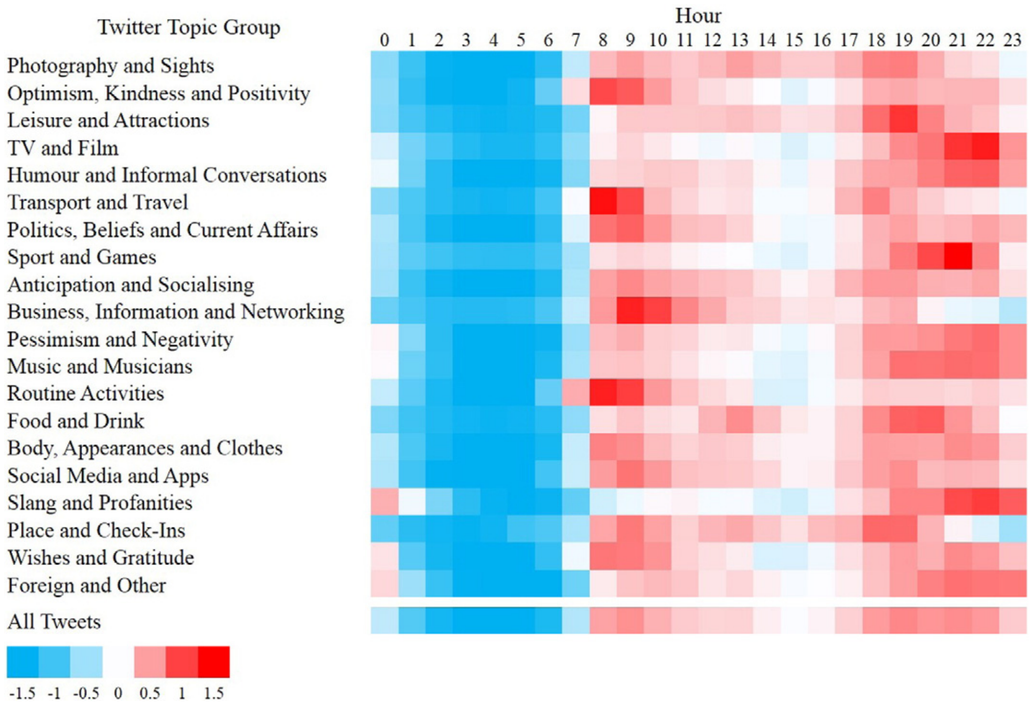


Fig. 5. A heat map of the temporal frequency of Tweet topics across the whole weekday sample by hour of the day. The data from each of the Twitter groups has been standardised as Z-scores to account for variations in their sizes. Larger numbers (red) therefore indicate overrepresentation.

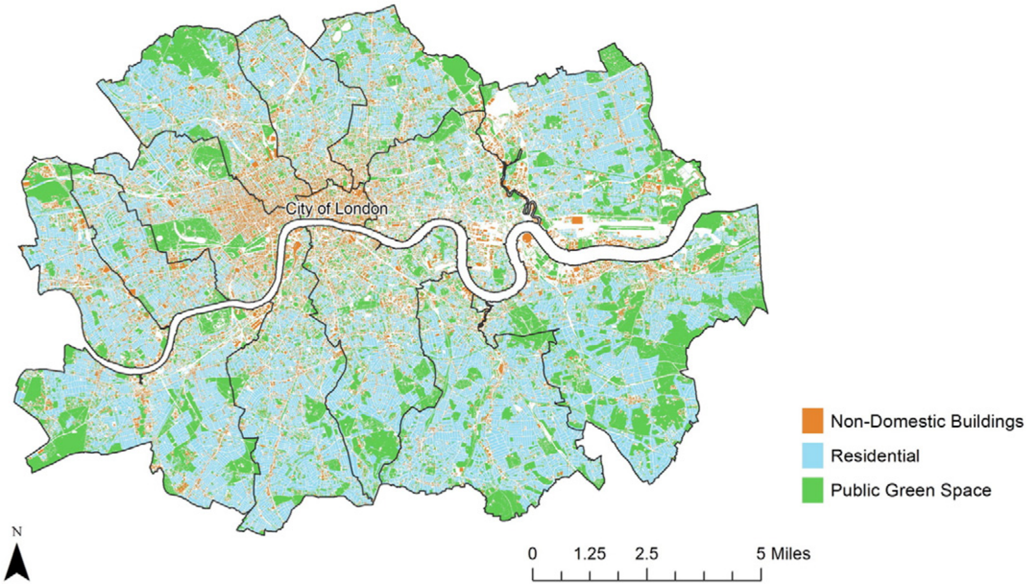


Fig. 6. The spatial distribution of three key land-use categories in Inner London.

entire sample to produce a ratio (Table 3). Therefore 1 is equal to the expected proportion, 2 is double and 0.5 is half.

The variations between topic compositions at residential and the other land-uses identifies that users typically Tweet slightly differently when at home. The groups most predominant in residential locations are those associated with household leisure activities (as exemplified by the TV and Film group), and to a lesser extent, those that are more consistent with general conversation. The Business, Information and Networking group is overrepresented in non-domestic buildings, presumably because of the abundance of Tweets from places of work and event locations. The Places and Check-Ins and Leisure and Attractions groups are also overrepresented. Green space has a greater proportion of photography based Tweets which is unsurprising. Interestingly, the Slang and Profanities group is also overrepresented here, possibly because of a higher presence of younger age groups in these places.

6.2.2. Key places

Segmenting Tweets by general land-use categories could fail to capture the potentially substantial variations in patterns of Tweet content between different types of non-domestic places. For instance, Tweets made at a sporting stadium are likely to differ substantially in content from those sent from shopping centres. To demonstrate this feature, Tweets were selected from 100 m buffers around the polygons of six key locations of different activities. The key places selected comprised a football stadium (the Emirates Stadium), a large entertainments arena (the O2 Arena), a train station (Waterloo Station), a shopping centre (Westfield Stratford), an area with a vibrate night life (Soho), and a place of work (Canary Wharf). The data from these locations are presented in Table 4 as location quotients.

Table 3
Location quotients for each of the Twitter groups by three key land-uses.

Twitter topic group	Residential	Non-domestic buildings	Public green space
Photography and Tourism	0.59	1.15	1.25
Optimism, Kindness and Positivity	0.98	0.96	0.96
Leisure and Attractions	0.73	1.44	0.80
TV and film	1.10	0.91	0.97
Humour and Informal Conversations	1.22	0.76	1.05
Transport and Travelling	0.87	1.01	1.00
Politics, Beliefs and Current Affairs	1.04	0.95	1.02
Sport and Games	1.11	0.86	1.11
Anticipation and Socialising	1.04	0.94	0.98
business, Information and Networking	0.73	1.48	0.82
Pessimism and Negativity	1.18	0.81	1.12
Music and Musicians	1.05	1.08	0.91
Routine Activities	1.10	0.87	1.07
Food and Drink	0.79	1.08	0.77
Body, Appearances and Clothes	1.07	0.87	1.05
Social Media and Apps	1.06	0.95	0.98
Slang and Profanities	1.40	0.59	1.21
Place and Check-Ins	0.46	1.94	0.73
Wishes and Gratitude	1.18	0.85	1.04
Foreign and Other	1.08	0.91	0.98

Tweets from each of the key places vary considerably in topic composition. Relative to the whole sample, Tweets from the Emirates Stadium area are 11.7 times more likely to be from the Sports and Games group, and those from and around the O2 Area are almost 6 times more likely to be from the Music and Musicians group. One finding which may appear unexpected is the fact that the Leisure and Attractions group is overrepresented at Stratford shopping centre. However, discussions of shopping are included within this group as identified at the subgroup level. The findings from the key places are largely self-explanatory. This suggests that behaviour on Twitter is influenced by local activity, and that the methodology for this study is an appropriate means of segmenting Tweets.

6.3. Subgroups

We next present the inferred labels of subgroups from the Twitter topic classification (Table 5).

All 100 subgroups are distinctive of each other. For example, within the Leisure and Attractions group, each subgroup represents comments about distinctive types of activities. Consequently, the spatial distributions of each subgroup also differ. Fig. 7 illustrates the spatial distribution of four of these subgroups in Central London, visualised using kernel density estimation.

All four of the examples above identified spatial clusters in areas of known associated activity. For instance, the Museums and Galleries subtopic attains highest densities in South Kensington where three large museums are located. The Fashion and Shopping subgroup is densest at popular upmarket fashion shopping locations such as Regent's Street and also locations where there are fashion exhibitions, such as Somerset House. The results also highlight the benefits of the high accuracy and precision of geo-tagged Tweets.

Table 4
Location quotients for each of the Twitter topic groups by six key places in London.

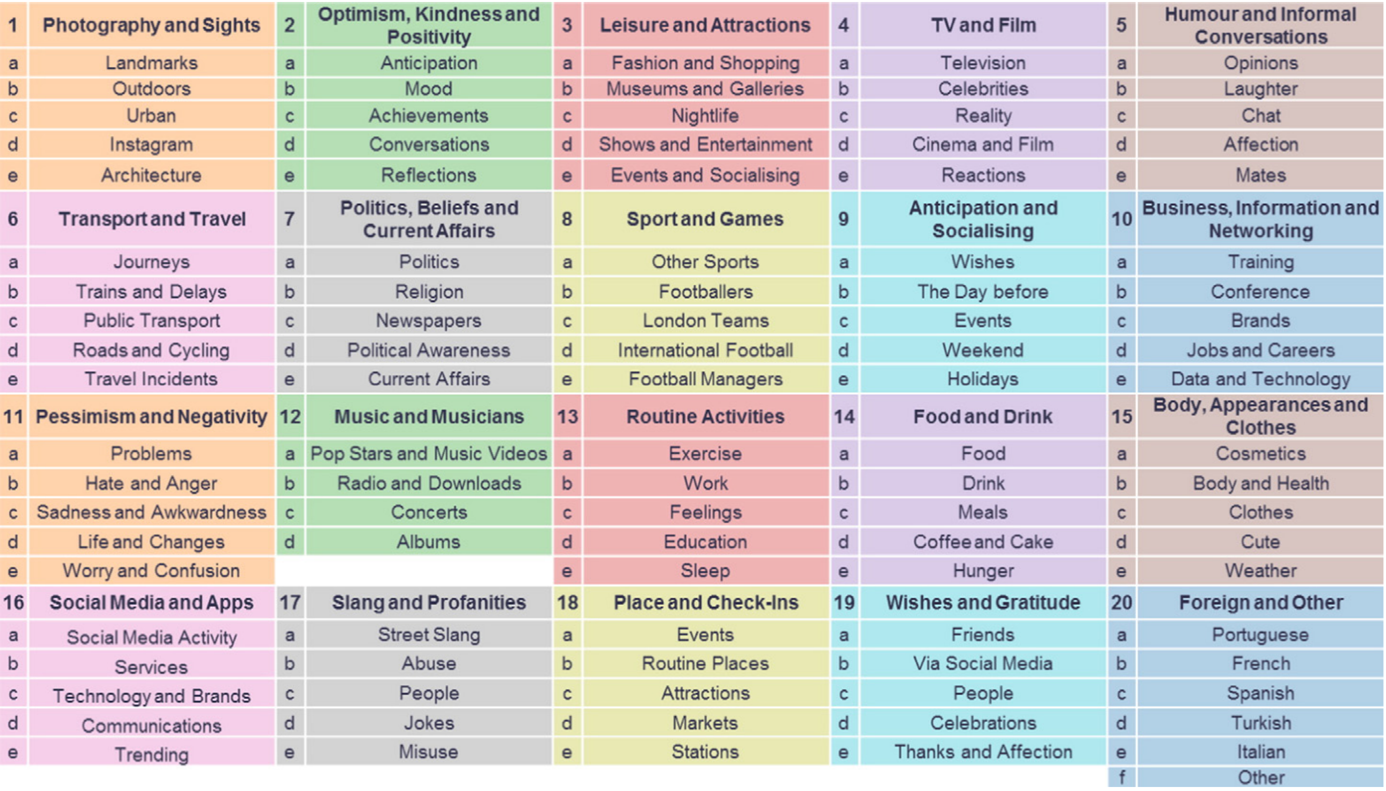
Twitter topic group	The Emirates Stadium	The O2 Arena	Waterloo Station	WestField Stratford	Soho	Canary Wharf
Photography and Tourism	0.82	0.83	0.73	0.73	1.46	2.49
Optimism, kindness and positivity	0.50	0.75	1.06	0.93	1.02	1.08
Leisure and attractions	0.47	1.00	0.73	1.91	3.35	0.55
TV and Film	0.49	1.39	0.82	0.84	1.08	0.66
Humour and Informal Conversations	0.39	0.69	0.62	1.02	0.56	0.80
Transport and Travelling	0.49	0.89	2.19	1.18	0.69	1.09
Politics, Beliefs and Current Affairs	0.33	0.43	0.87	0.46	0.68	0.89
Sport and Games	11.73	0.71	0.73	0.81	0.44	1.13
Anticipation and Socialising	0.28	1.02	0.98	1.14	0.81	1.22
Business, Information and Networking	0.66	1.15	0.87	0.59	0.87	2.86
Pessimism and Negativity	0.29	0.57	0.69	0.81	0.63	0.75
Music and Musicians	0.28	5.79	0.72	0.67	1.00	0.97
Routine Activities	0.45	0.53	0.96	0.82	0.60	1.03
Food and Drink	0.21	0.77	0.91	1.87	3.52	1.20
Body, Appearances and Clothes	0.27	0.65	0.95	1.77	0.79	0.95
Social Media and Apps	0.37	0.71	0.98	1.00	1.19	1.25
Slang and Profanities	0.25	0.40	0.40	0.69	0.28	0.34
Place and Check-Ins	1.13	2.03	4.41	3.15	1.85	1.00
Wishes and Gratitude	0.41	0.82	0.68	1.03	0.79	0.76
Foreign and Other	0.44	0.52	0.27	0.55	0.56	0.37

Focusing on South Kensington, the Twitter classification highlights a clear differentiation of Twitter topics between Tweets from the Royal Albert Hall, and those from the cluster of museums just one block south. The subgroups thus provide a more nuanced view of Twitter use than the group level alone.

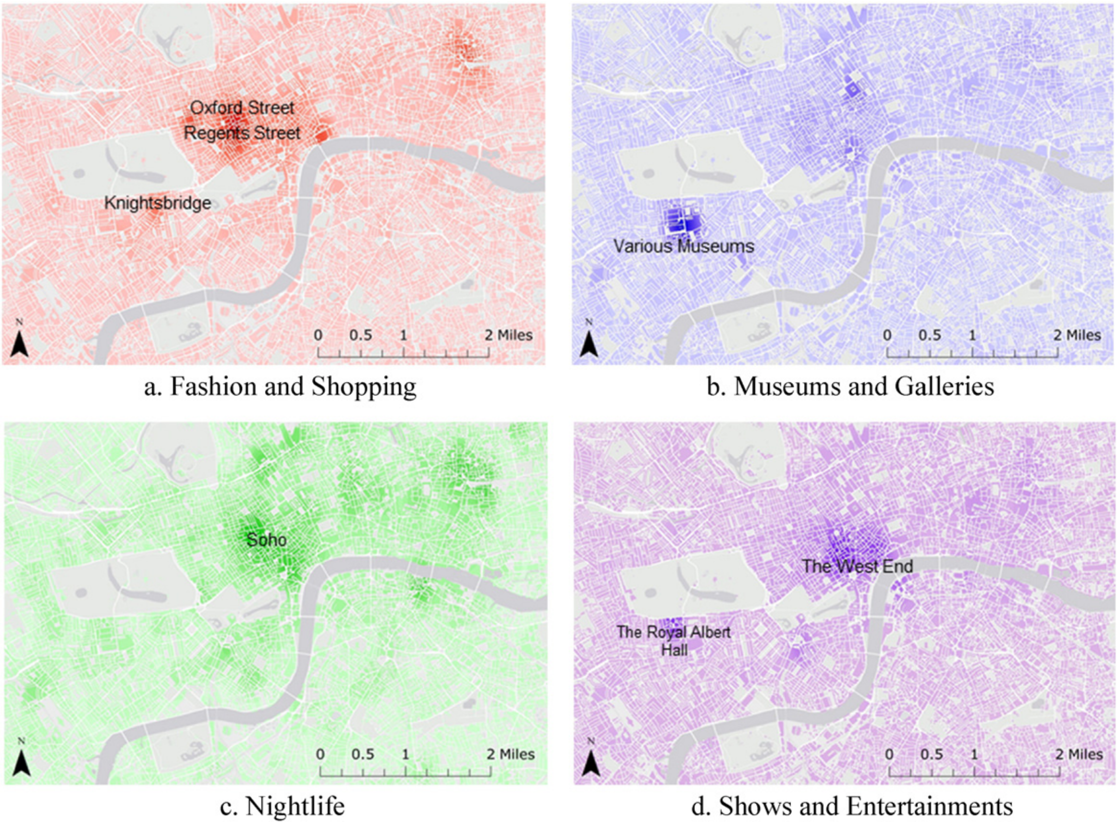
6.4. Twitter users and topics

Having established the pattern of associations between types in the Twitter Classification and geotemporal activity patterns across London, we next seek to link the typology to user characteristics. It may be taken as highly probable that topics are not randomly distributed between users. To test this hypothesis, we extracted all of the users with over 100 Tweets in our final dataset. 100 Tweets was considered enough to capture variance as most users would most likely Tweet about a number of different topics over the data capture period. We have subsequently termed this sample, the regular users. As the sample only contains 1.3 million Tweets and Tweets were very positively skewed between unique users, only 1750 accounts met this criterion. An index of dissimilarity was produced for each of the topics to observe their disassociations (Fig. 8). The index, *D*, was proposed by Duncan and Duncan (1955) to quantify racial

Table 5
Labels for all 100 of the Twitter topic subgroups.



1	**Photography and Sights**	2	**Optimism, Kindness and Positivity**	3	**Leisure and Attractions**	4	**TV and Film**	5	**Humour and Informal Conversations**
a	Landmarks	a	Anticipation	a	Fashion and Shopping	a	Television	a	Opinions
b	Outdoors	b	Mood	b	Museums and Galleries	b	Celebrities	b	Laughter
c	Urban	c	Achievements	c	Nightlife	c	Reality	c	Chat
d	Instagram	d	Conversations	d	Shows and Entertainment	d	Cinema and Film	d	Affection
e	Architecture	e	Reflections	e	Events and Socialising	e	Reactions	e	Mates
6	**Transport and Travel**	7	**Politics, Beliefs and Current Affairs**	8	**Sport and Games**	9	**Anticipation and Socialising**	10	**Business, Information and Networking**
a	Journeys	a	Politics	a	Other Sports	a	Wishes	a	Training
b	Trains and Delays	b	Religion	b	Footballers	b	The Day before	b	Conference
c	Public Transport	c	Newspapers	c	London Teams	c	Events	c	Brands
d	Roads and Cycling	d	Political Awareness	d	International Football	d	Weekend	d	Jobs and Careers
e	Travel Incidents	e	Current Affairs	e	Football Managers	e	Holidays	e	Data and Technology
11	**Pessimism and Negativity**	12	**Music and Musicians**	13	**Routine Activities**	14	**Food and Drink**	15	**Body, Appearances and Clothes**
a	Problems	a	Pop Stars and Music Videos	a	Exercise	a	Food	a	Cosmetics
b	Hate and Anger	b	Radio and Downloads	b	Work	b	Drink	b	Body and Health
c	Sadness and Awkwardness	c	Concerts	c	Feelings	c	Meals	c	Clothes
d	Life and Changes	d	Albums	d	Education	d	Coffee and Cake	d	Cute
e	Worry and Confusion			e	Sleep	e	Hunger	e	Weather
16	**Social Media and Apps**	17	**Slang and Profanities**	18	**Place and Check-Ins**	19	**Wishes and Gratitude**	20	**Foreign and Other**
a	Social Media Activity	a	Street Slang	a	Events	a	Friends	a	Portuguese
b	Services	b	Abuse	b	Routine Places	b	Via Social Media	b	French
c	Technology and Brands	c	People	c	Attractions	c	People	c	Spanish
d	Communications	d	Jokes	d	Markets	d	Celebrations	d	Turkish
e	Trending	e	Misuse	e	Stations	e	Thanks and Affection	e	Italian
								f	Other



a. Fashion and Shopping

b. Museums and Galleries

c. Nightlife

d. Shows and Entertainments

Fig. 7. A kernel density smoothing of georeferenced Tweets from four of the subcategories of Group 3. Darker tones correspond with higher densities. (Basemap courtesy OpenStreetMap.)

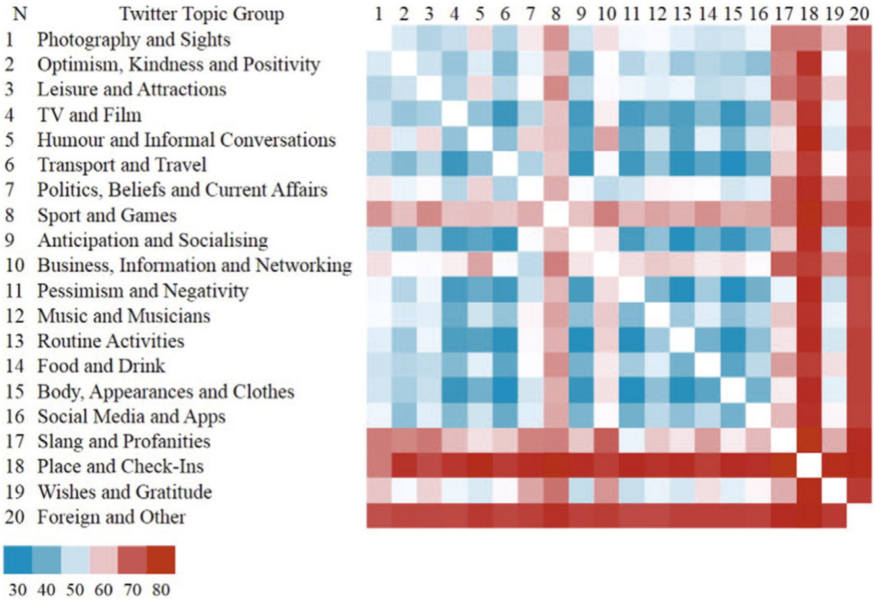
segregation as a measure of evenness between pairs of population groups across aggregate units. In the two group case:

$$D = \frac{1}{2}\sum_{i=1}^{n}\left|\frac{a_i}{A_T}-\frac{b_i}{B_T}\right|$$

where $n =$ Number of users;
$a_i =$ Number of user's Tweets in group A;
$A_T =$ Total number of Tweets in Group A from regular users;
$b_i =$ Number of user's Tweets in group B; and
$B_T =$ Total number of Tweets in Group B from regular users

The index thus measures the distance between pairs of topics based on the composition of topics from all of the users in the sample, relative to the overall composition of topics. Higher values correspond with greater dissimilarity.

The dissimilarity matrix revealed that all of the groups are uniquely distributed across the regular users. The Place and Check-Ins group is the most isolated group overall: that is to say, users who tweeted in



N	Twitter Topic Group
1	Photography and Sights
2	Optimism, Kindness and Positivity
3	Leisure and Attractions
4	TV and Film
5	Humour and Informal Conversations
6	Transport and Travel
7	Politics, Beliefs and Current Affairs
8	Sport and Games
9	Anticipation and Socialising
10	Business, Information and Networking
11	Pessimism and Negativity
12	Music and Musicians
13	Routine Activities
14	Food and Drink
15	Body, Appearances and Clothes
16	Social Media and Apps
17	Slang and Profanities
18	Place and Check-Ins
19	Wishes and Gratitude
20	Foreign and Other

Fig. 8. Dissimilarity matrix of all Twitter topics from users who tweeted at least 100 times in the dataset. The data have been visualised as a heat map.

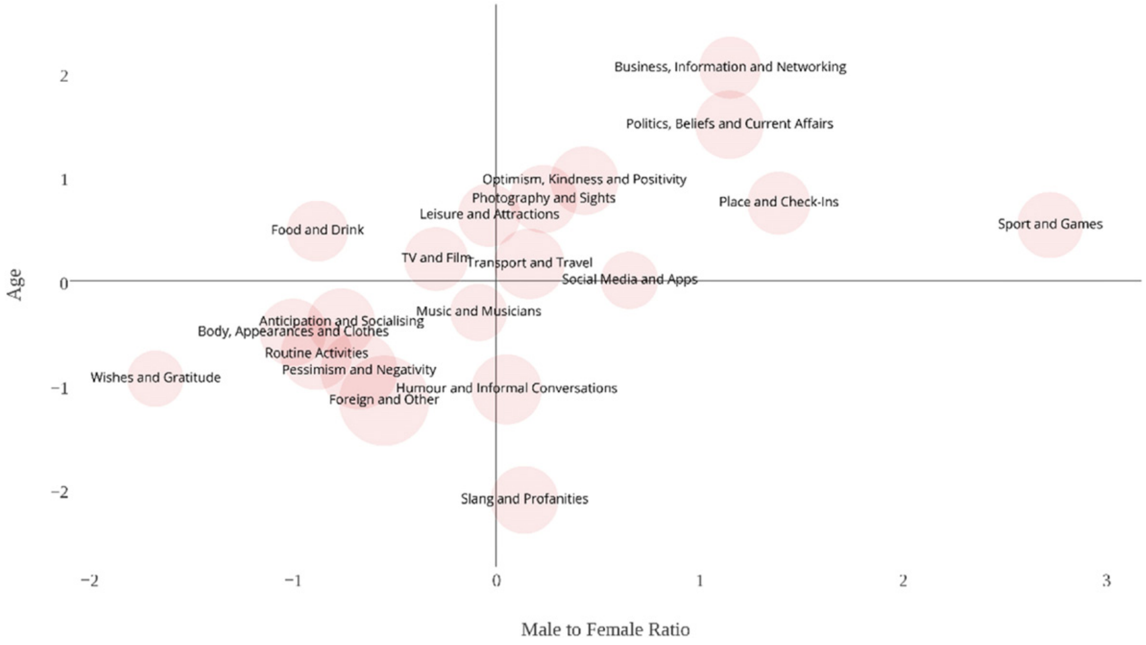


Fig. 9. A bubble plot of the average gender and age distributions for each Twitter group. The size of the bubbles corresponds with the number of Tweets assigned to that group. The data are represented as standard deviations.

this category were less likely to Tweet in any of the other categories as well. Considering that this dataset is comprised entirely of geotagged Tweets, it is probable that some users mainly use the geotag function when submitting check-ins via applets. The second most isolated group is Foreign and Other. Given that a very high proportion of Tweets from this category are written in a foreign language, this result is unsurprising. It is also interesting to observe the dissimilarities experienced between the Twitter groups in terms of how they are distributed between users. The Business, Information and Networking group is quite isolated, although its users are also likely to tweet about Politics, Beliefs and Current Affairs. It is possible that Tweeters who are largely associated with these two groups use the social network more formally as a means of promoting information. In contrast, the groups associated with day to day activities and opinions share lower dissimilarities scores between each other.

Our analysis confirms that topics are not randomly distributed between users. It is likely that the unobserved characteristics of Twitter users are likely to influence what they Tweet about. We therefore extend the research to identify linkages between variabilities in Twitter groups and the demographic and socio-economic characteristics of users. As Twitter does not require users to record such details about themselves, characteristics have been inferred using two approaches described below.

Table 6
Location quotients for each of the Twitter topic groups by NS-SEC groups assigned to their users.

Twitter topic group	Higher managerial, & professional occupations	Lower managerial & professional occupations	Intermediate occupations	Lower supervisory & technical occupations	Semi routine occupations	Routine occupations	Never worked & long term unemployed
Photography and Tourism	1.16	1.05	0.89	0.90	0.86	0.87	0.92
Optimism, Kindness and Positivity	1.04	1.02	1.00	0.98	0.95	0.95	0.95
Leisure and Attractions	1.13	1.06	0.93	0.89	0.88	0.89	0.91
TV and Film	1.02	1.01	1.02	0.99	0.98	0.98	0.97
Humour and Informal Conversations	0.91	0.95	1.05	1.07	1.10	1.07	1.05
Transport and Travelling	1.02	1.01	1.01	0.99	0.99	0.98	0.98
Politics, Beliefs and Current Affairs	1.06	1.03	0.99	0.94	0.95	0.97	0.97
Sport and Games	1.00	0.99	1.04	1.00	1.01	1.01	1.01
Anticipation and Socialising	1.00	1.00	1.01	1.01	1.01	1.00	0.98
Business, Info' and Networking	1.15	1.06	0.95	0.91	0.87	0.87	0.89
Pessimism and Negativity	0.95	0.97	1.03	1.04	1.05	1.05	1.03
Music and Musicians	0.98	1.00	1.00	1.01	1.02	1.02	1.00
Routine Activities	0.96	0.98	1.03	1.03	1.04	1.02	1.01
Food and Drink	1.10	1.05	0.96	0.92	0.91	0.91	0.94
Body, Appearances and Clothes	0.98	0.99	1.02	1.01	1.03	1.02	1.02
Social Media and Apps	1.03	1.01	0.99	0.99	0.99	0.98	0.99
Slang and Profanities	0.83	0.92	1.05	1.09	1.16	1.16	1.14
Place and Check-Ins	1.18	1.12	0.83	0.81	0.79	0.82	0.89
Wishes and Gratitude	0.94	0.97	1.04	1.06	1.06	1.04	1.02
Foreign and Other	0.97	1.00	0.97	1.05	0.99	1.04	1.05

6.4.1. Age and gender

Gender and probable age distributions can be inferred from the registered forenames of users because of trends in baby naming and migration. Using a name database outlined in Lansley and Longley (2016), the probable gender and age distributions of given names can be assigned to Twitter users. The database consists of demographic structure estimates for over 30,000 given names and was produced from a combination of birth certificate records and market data representing over 17 million individuals from the UK. As Twitter does not record individuals' forenames as a unique variable, an algorithm that extracts the probable forenames and surnames from their registered names was implemented (as described in Longley et al., 2015).

Using this approach, demographics were extracted from as many Twitter usernames as possible. As a large proportion of users have opted to use non-conventional names such as nicknames or the names of companies, about 36% of users could not be matched to the names database. Consequently, only the ages and genders of 98,409 users could be modelled.

The 20 groups have been plotted by the proportion of male users and their average estimated ages (Fig. 9). The findings reveal that the typical demographic characteristics of Twitter users vary between each of the topics. For instance, the Slang and Profanities group has the youngest average age across the whole sample, possibly due to the content's association with youth culture. In contrast, the Business, Information and Networking group has the oldest population on average. This result seems logical as this is the only group which is more restricted to the working age population, and possibly overrepresented by those in established careers. The Sport and Games group has the highest proportion of male Tweeters, with over 80% of Tweets originating with male users. The majority of the contents of this group is related to football, a sport which is disproportionately popular amongst males. The most female dominated group is Wishes and Gratitude, followed by Routine Activities and Body, Appearances and Clothes.

6.4.2. Neighbourhood socio-economic characteristics

Additional inferences can be drawn based upon the characteristics of the neighbourhoods in which the Twitter users are likely to reside. We ascribe socio-economics using the National Statistics Socio-Economic Classification (NS-SEC) from the 2011 Census. This approach has a number of limitations as by no means all of the residents recorded in the 2011 Census are Twitter users. However, our aim was to identify broader neighbourhood characteristics which in turn, may correspond to a large extent with Twitter users.

The 289,240 Tweets submitted from residential land-uses (as identified from the GLUD data) were assigned to census Output Areas via a spatial interpolation. Output Areas (OAs) are the smallest geographic unit for which 2011 Census data are available, and represent a mean of 309 individuals. The users who had tweeted from multiple residential locations were each then assigned to the OA from which they had tweeted most frequently. Following this, each user was assigned statistics on the relative proportions of each NS-SEC group from their inferred residential OA.

These characteristics were then appended to the rest of the Tweets also sent from these users so the subsequent analysis was not restricted to only those sent from residential locations. In total, just over 740,000 Tweets could be assigned NS-SEC data. Finally, these Tweets were used to cross-tabulate the Twitter topics by the average proportion of each NS-SEC group. The results have been presented as location quotients (Table 6). The NS-SEC group small employers and own account workers has been excluded from the analysis as it does not reflect a homogenous social group (Rose & Pevalin, 2001).

The analysis identified an association between neighbourhood characteristics of users and Twitter topics. Although the associations are not as great as those identified by name inferred ages.

An interesting finding is that users from neighbourhoods with higher proportions of population in higher socio-economic echelons are more likely to Tweet optimistically, whereas users from lower social status neighbourhoods are more likely to be pessimistic as demonstrated by variations between the Optimism, Kindness and Positivity, and Pessimism and Negativity groups. The results also demonstrate a distinctive divide in the uses of language between classes. Those from lower socio-economic status neighbourhoods were more likely to send Tweets categorised as Slang and Profanities, and also Humour and Informal Conversations. Both of these groups are notable for informal uses of language. By contrast, users from higher socio-economic neighbourhoods were more likely to discuss Business, Information and Networking and Leisure and Attractions. However, it is also reasonable to assume that persons of different classes may still discuss the same topics, but approach them differently.

7. Conclusions

The research has demonstrated that although unregulated and non-conventional for quantitative analysis, Twitter data can be harvested into a simple classification which can be useful to planners, marketers and researchers. The findings revealed distinctive traits of Tweets across space and time, and also between Tweeters themselves. It identified the influence of land-use and activity on the content of Tweets which, whilst not surprising in many instances, documents influences that have not been explored on the extensive scale enabled by the methodological approaches presented. Furthermore, the analysis demonstrated that social media data can reveal insights into urban dynamics which are not available from traditional datasets.

Based on the sample of geo-tagged Tweets from Inner London, the analysis demonstrated that users do not Tweet evenly across space. Twitter is a means of spreading information, but the type of information and the nature of how it is communicated vary between users. A wide range of individual characteristics are likely to be associated with variations in what is communicated through Twitter, including demographics and socio-economics. As there are no data on the users beyond what is provided from Twitter, it is not possible to measure any of these traits conclusively. However, using novel modelled age and gender estimates from forenames of Twitter users, the research identified key variances in the uses of Twitter based on inferred demographic characteristics. It should be noted that the use of names is an uncertain determinant of identity, however (Longley et al., 2015). It was also possible to infer neighbourhood characteristics of users by estimating probable residential locations from land-use data.

This paper has developed an operational classification for georeferenced Tweets from London. The presented Twitter segmentation demonstrates that with a large enough sample of data and robust text cleaning techniques, LDA can suitably segment Tweets. Whilst the Tweets from this study have been restricted to geo-tagged Tweets from Inner London only, the methodological approaches to topic modelling can be applied to any sample of Tweets. In so doing, selection of the optimal number of classes is a subjective process and demographic data may need to be sourced from alternative locations. Moreover, the topic groups from our model can also be applied to other data by utilising the probabilities for each of the words created by the LDA as a look-up file — although it would of course only be possible to assign words which appeared in our modelled data. Furthermore, future research may build upon the associations identified in this paper to develop predictive tools to estimate either land use and activity, or the likely content of social media posts across space and time, and between users.

Acknowledgements

This work is funded by the UK ESRC Consumer Data Research Centre (CDRC) grant reference ES/L011840/1 and 'The analysis of names from the 2011 Census of Population' (ES/L013800/1).

References

Andrienko, G., Andrienko, N., Bosch, H., Ertl, T., Fuchs, G., Jankowski, P., & Thom, D. (2013). Thematic patterns in georeferenced tweets through space-time visual analytics. *Computing in Science and Engineering*, *15*(3), 72–82.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Blei, D. M., Griffiths, T., Jordan, M. I., & Tenenbaum, J. B. (2004). Hierarchical topic models and the nested Chinese restaurant process. In S. Thrun, L. K. Saul, & B. Schölkopf (Eds.), *Advances in neural information processing systems 16*. Cambridge, MA: MIT Press.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8.

Ceapa, I., Smith, C., & Capra, L. (2012). *Avoiding the crowds: Understanding tube station congestion patterns from trip data. In proceedings of the ACM SIGKDD International Workshop on Urban Computing, (Beijing, China)*.

Chae, J., Thom, D., Bosch, H., Jang, Y., Maciejewski, R., Ebert, D. S., & Ertl, T. (2012). *Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In proceedings of the 2012 EIEE Conference on Visual Analytics Scieme and Technology*. USA: Seattle.

Chamlertwat, W., Bhattarakosol, P., & Rungkasiri, T. (2012). Discovering consumer insight from Twitter via sentiment analysis. *Journal of Universal Computer Science*, *18*(8), 973–992.

DCLG (2006). Generalised land-use database statistics for England 2005 Department for Communities and Local Government, London. Online: http://webarchive.nationalarchives.gov.uk/20120919132719/http://communities.gov.uk/documents/planningandbuilding/pdf/pdf/154941.pdf (Access 05/03/2014)

Duncan, O. D., & Duncan, B. (1955). A methodological analysis of segregation indexes. *American Sociological Review*, *20*(2), 210–217.

Gayo-Avello, D., Metaxas, P. T., & Mustafaraj, E. (2011). *Limits of electoral predictions using social media data. Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. Spain: Barcelona.

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, *69*(4), 211–221.

Harris, R., Sleight, P., & Webber, R. (2006). *Geodemographics, GIS and neighbourhood targeting*. Chichester: Wiley.

Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in Twitter. *In proceedings of the SIGKDD Workshop on SMA* (Online: http://snap.stanford.edu/soma2010/papers/soma2010_12.pdf (Access 06/04/2015)).

Ipsos, M. O. R. I. (2013). Ipsos MediaCT Tech tracker Q4 2013. Online: https://www.ipsos-mori.com/Assets/Docs/Publications/IpsosMediaCT_Techtracker_Q4_2013.pdf (Access 03/09/2016)

Jiang, S., Alves, A., Rodrigues, F., Ferreira, J., & Pereira, F. C. (2015). Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Computers, Environment and Urban Systems*, *53*, 36–46.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? *Proceedings of the 19th International Conference on World Wide Web (WWW), 2010*.

Lansley, G., & Longley, P. (2016). Deriving age and gender from forenames for consumer analytics. *Journal of Retailing and Consumer Services*, *30*, 271–278.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., ... Shi, L. (2015). Social sensing: A new approach to understanding our socio-economic environments. *Annals of the Association of American Geographers*, *105*(3), 512–530.

Longley, P. A., Adnan, M., & Lansley, G. (2015). The geotemporal demographics of. *Twitter usage Environment and Planning A*, *47*(2), 465–484.

Manley, E., & Cheshire, J. (2013). The utility of geolocated Twitter data for mapping language diversity. *The Royal Geographical Society Annual International Conference*. UK: London.

Marwick, A. E. Boyd, D. (2010) I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. New media & society, 13 (1), 114-133.

McKenzie, G., Janowicz, K., Gao, S., & Gong, L. (2015). How where is when? On the regional variability and resolution of geosocial temporal signatures for points of interest. *Computers, Environment and Urban Systems*, *54*, 336–346.

Michelson, M., & Macskassy, S. a. (2010). *Discovering users' topics of interest on twitter: a first look. AND '10: Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data, 73–80*.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. (2013). Is the sample good enough? *Comparing data from Twitter's streaming API with Twitter's firehose* (ICWSM '13).

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). *From tweets to polls: Linking text sentiment to public opinion time series. In Fourth International AAAI Conference on Weblogs and Social Media*. USA: Washington D.C.

Quercia, D., Ellis, J., Capra, L., & Crowcroft, J. (2012). *Tracking "gross community happiness" from tweets. In Proceedings of the 15th ACM Conference on Computer Supported Cooperative Work*. USA: Seattle.

Ramage, D., Dumais, S., & Liebling, D. (2010). *Characterizing microblogs with topic models. In Fourth International AAAI Conference on Weblogs and Social Media*. USA: Washington D.C.

Rose, D., & Pevalin, D. J. (2001). *The National Statistics Socio-economic Classification: Unifying official and sociological approaches to the conceptualisation and measurement of social class, paper 2001–04*. Working papers of the Institute for Social and Economic Research. Colchester: University of Essex.

Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The use of twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *PloS One*, *6*(5), e19467.

Singleton, A. D., & Longley, P. (2015). The internal structure of Greater London: A comparison of national and regional geodemographic models. *Geo: Geography and Environment*, *2*(1), 69–87.

Spielman, S. E., & Thill, J. C. (2008). Social area analysis, data mining, and GIS. *Computers, Environment and Urban Systems*, *32*(2), 110–122.

Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004). Probabilistic author-topic models for information discovery. *The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. *In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*.

Twitter (2015). Twitter usage/company facts (updated March 31, 2015). Online. https://about.twitter.com/company (Accessed 02/06/2015)

Williams, S. A., Terras, M., & Warwick, C. (2013). What people study when they study Twitter? Classifying Twitter related academic papers. *Journal of Documentation*, *69*(3), 384–410.