# Twitter Usage Across Industry: A Spatiotemporal Analysis

Neha Gupta*, Henry Crosby*, David Purser*, Stephen Jarvis* and Weisi Guo*
*Warwick Institute for the Science of Cities
University of Warwick
Coventry, UK
{Neha.Gupta, H.J.Crosby, D.J.Purser, S.A.Jarvis, Weisi.Guo}@warwick.ac.uk

***Abstract*—High resolution social media data presents an opportunity to better understand people's behavioural patterns and sentiment. Whilst significant work has been conducted in various targeted social contexts, very little is understood about differentiated behaviour in different industrial sectors. In this paper, we present results on how social media usage and general sentiment vary across the geographic and industry sector landscape. Unlike existing studies, we use a novel geo-computational approach to link location specific Twitter data with business sectors by leveraging the UK Standard Industrial Classification Code (SIC Code). Our baseline results for the Greater London area identifies *Construction*, *Real Estate*, *Transport* and *Financial Services* industries consistently have stronger Twitter footprints. We go on to apply natural language processing (NLP) techniques to understand the prevailing sentiment within each business sector and discuss how the evidence can contribute towards de-biasing Twitter data. We believe this research will prove a valuable surveillance tool for policy makers and service providers to monitor ongoing sentiment in different industry sectors, perceive the impact of new policies and can be used as a low cost alternative to survey methods in organisational studies.**

***Index Terms*—Social Media Analysis in industry, Big Data Visualisation, Knowledge Integration, GIS**

## I. Introduction

The connected society on social media platforms such as Twitter generate large volumes of data everyday providing an unprecedented opportunity to perform social scientific analysis both at an individual and aggregated community level. Several recent studies have exploited the geo-spatial property of Twitter data for a variety of event detection and service support roles. For example, Twitter has been used for: measuring the perceived psycho-demographic of people [1]; crisis response collaboration required during urban flooding [2], [3]; monitoring earthquake events [4], [5]; surveillance of disease spread [6]; monitoring mobility patterns [7]; assisting data-driven urban planning [8]; tracking community happiness [9], [10] and also predicting election outcomes by understanding the political orientation of population [11], [12].

Whilst Twitter data has been successfully used in the many of the research scenarios outlined above, we know very little about the underlying demographics of the users. To understand the key contributors of the conversations taking place on social media platforms like Twitter is a research challenge. In general, statistical averaging across large populations and across contexts yield reasonable understanding. In practice, we need to augment the social media data with other data sources to conduct more high resolution studies. One approach to debiasing is for Twitter data to be compared to either established knowledge or situational context. Existing research has studied survey data from the well known Oxford Internet Survey (OxIS) and America's Pew Internet survey to understand the population representativeness of Twitter Data and has reported Twitter users as disproportionate members of elites in both countries [13]. Likewise, other studies have attempted to infer demographic characteristics such as age, occupation and social class of Twitter users using profile description of Twitter users [14], [15] and gleaned insights about the race, place and gender of Twitter users by studying the intensity of the tweets [16].

### A. Twitter Usage in Industrial Sectors

Although these studies outline the importance of investigating group differences of Twitter users, very little research that we are aware of has investigated which business groups or sectors contribute to Twitter conversations and which economic sector the Twitter users belong to. Consider a scenario where a policy maker would like to introduce policy change impacting working conditions in a targeted sector. One would be faced with a challenge to isolate the public opinion with respect to the sectors they are originating from. This study therefore is set out to fill this knowledge gap by providing: 1) a methodology to link geo-tagged Twitter data to different sectors by using very diverse sets of datasets owned by government bodies (the UK Land Registry) and Ordnance Survey, and 2) provide empirical evidence to help debias future Twitter data analysis. Such isolation of tweets in different sectors not only aids in capturing the feedback of aforementioned policy changes but also presents an exiting opportunity for organisational studies to study and compare sentiment and mood in different sectors.

To conduct this data driven experimental study we use the geo-tagged twitter data from the London region ranging over a two week period for the years 2012 and 2016. Given the availability of more social media data both longitudinal and spread across other geographies, this methodology can easily be translated to perform similar studies for other cities in the UK. These tweets are profiled into *UK standard industrial classification Code 2007 (SIC Code)* sectors, which are used in

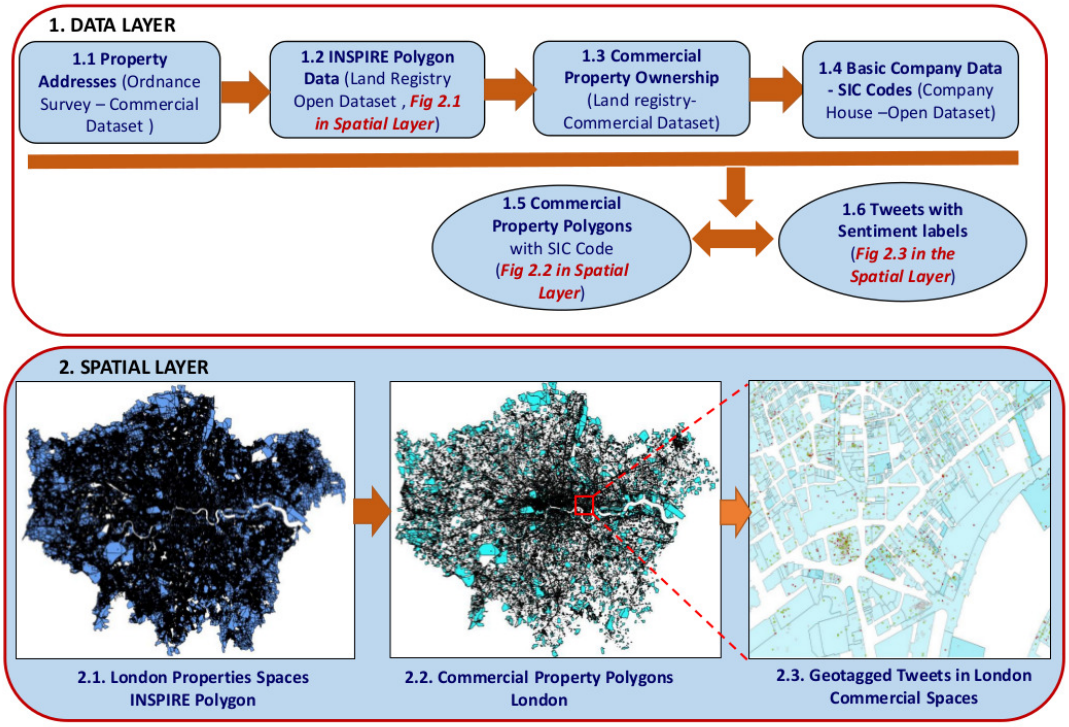


Fig. 1. **Methodology:** 1) The data layer shows joins between diverse data-sets to create commercial properties polygons layer for London and clip tweets inside those polygons. 2) The spatial layer join between tweets and commercial property polygons.

classifying business establishments and other statistical units by the type of economic activity in which they are engaged in the UK [17]. The most relevant recent example is [18], which is primarily focused towards understanding opinion and sentiment towards popular industry brands.

The next two sections provide a description of the data sources used in this research and outline the framework designed to extract tweets linked to the business sectors.

## II. Data Sources

Our research relies on a large pool of diverse datasets, some collected through open sources and others available commercially in proprietary databases. The data description and the sources by which they were acquired from are given below.

*Property Addresses from Ordnance Survey* - Property Addresses from Ordnance Survey data provides a unique property reference number (UPRN) for every property address in the UK as well as longitude and latitude for each address [19]. Additionally, the data set classifies whether an address of a property is representative of "residential", "commercial" or "land". From this we extracted the commercial properties in London.

*INSPIRE Index Polygons* - the INSPIRE (Infrastructure for Spatial Information in Europe) directive came into force on 15 May 2007 with an aim to create a European Union spatial data infrastructure for the purposes of EU environmental policies and policies or activities which may have an impact on the environment [20]. To comply with the INSPIRE EU directive, the UK has developed an open source data-set called the INSPIRE Index Polygon which contains the locations of freehold registered property in England and Wales, a sub-set of UK government Index Polygons for all freehold land and property. These polygons are the shapes files that show the position and indicative extent of a registered property. This data comprises a set of polygons which represent land parcel use. Each INSPIRE Index Polygon has a unique identification number called the Land Registry-INSPIRE ID that relates to a registered title of the property number [21], [22]. As an example, Fig.2, displays an INSPIRE Index Polygon layer overlaying Google Street Map show the property area covered by Hammersmith and Fulham Council property. This GML data-set was converted to shape files for analysis using the open source software QGIS (Quantum Geographical Information System) for the whole of the London region (Fig.1 - the map labelled 2.1 shows all INSPIRE Polygons for the London area).

*Commercial and corporate ownership data* - This data is collected by the Land Registry department in the UK as part of the land registration process [23]. This data contains information about 13.8 million title records of freehold and leasehold properties in England and Wales, and contains information about the location of companies registered by companies house, a UK government department. Features of this data-

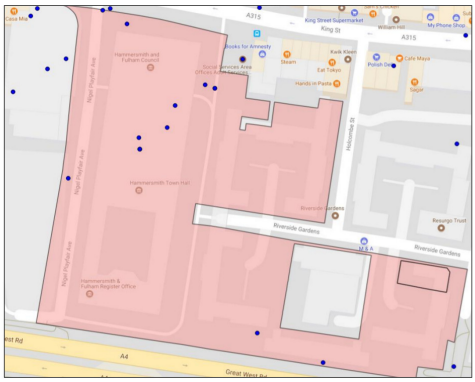

Fig. 2. INSPIRE Polygon in pink - Tweeting activity as blue dots in public sector industry (Hammersmith and Fulham Council Building).

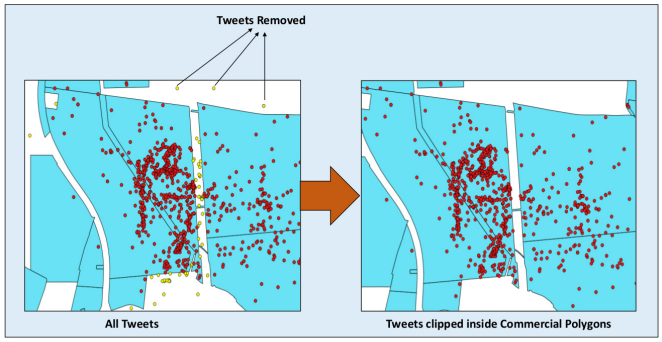


Fig. 3. Clipping Tweets in Commercial Property Polygons

set include the property address with the administrative area, name and address of the legal owner, title numbers, tenure (freehold/leasehold) and SIC code which defines the type of businesses hosted at the property address. When combined with the INSPIRE Polygon data, all the registered property parcels (spatial polygon map) in London can be uncovered. Commercial and corporate ownership data will be used for filtering the INSPIRE polygons and extracting only those polygons from the full data-set of the INSPIRE polygons which constitute *commercial areas* in London, i.e a limited number of polygons which are labelled as organisations or registered companies (Fig.1 - the map labelled 2.2. Commercial Property Polygons London).

*Twitter data* - People post their thoughts, observations about an event or everyday encounters as a 140 character text message on Twitter. The key data fields contained in the full tweet data [24] include information such as: username, Tweet text, time stamp, geo-location (latitude and longitude of the place from where the tweet was posted). The text data field can be further analysed using natural language processing (NLP) techniques to identify the sentiment polarity (positive, negative and neutral) of the text message. The time and geo-location of the Tweet can be processed to study the spatio-temporal characteristics of the tweets. For this research we had access to a data-set of nearly half a million geo-tagged Tweets covering a two week period in 2012 and 1.2 million of geo-tagged Tweets for two weeks in 2016 obtained from Twitter.

*SIC Code Information* - A SIC Code is a **S**tandard **I**ndustrial **C**lassification code used by UK Companies House (which is a government agency falling under the Department for Business, Energy and Industrial Strategy) to classify the type of economic activity in which a company or business are is engaged. This open source data is freely available to download [25] which contains the information about the company (or business) number and type of economic activity the company is involved in. We use this data to categorise and aggregate tweets in different SIC Codes.

## III. Methodology

In order to analyse tweets belonging to different sectors our experimental design involved three sequential steps. 1) Create the commercial polygon layer for London, to identify commercial or industry specific places in the London region; 2) Spatial filter geo-tagged tweets inside the commercial polygons, to isolate tweets that could be potentially related to the business premises they are geographically co-located with; 3) Sentiment Analysis of tweets using NLP (Natural Language Processing) computational techniques, to characterise and compare the distribution of tweet text sentiment (negative, neutral and positive) inside each sector.

### A. Commercial Polygon Layers

One of the key steps for this framework is to create a commercial property polygons spatial layer to isolate tweets which overlay with the **commercial property spaces** (referred to *commercial polygons* in this paper) for the London region. Fig.1 displays the methodology broken into two processes - *1) The Data Layer* join between diverse datasets to create a link between the commercial property textual and spatial information to tweets originating from the premises. *2) The Spatial Layer* join between commercial property information and tweets which overlay these premise spaces in London. We extracted all *commercial* INSPIRE polygons based on their address classification (commercial, residential or land) presented in the Ordnance Survey dataset to produce a spatial commercial polygon layer (map 2.2 in Fig.1). We then took the ownership details from the Land Registry dataset based on title number associated with each of the commercial INSPIRE polygons extracted. We use the company details in ownership data from the Land Registry to obtain the SIC Code and description reported by Company House.

### B. Sentiment Analysis

The textual component of Tweets often reflects the end user's sentiment and can be used as a proxy for well-being [9], [26], [27]. A common sentiment analysis approach is to use machine learning. We use two labelled datasets for training and test, provided by SemEval2015 (Semantic Evaluation), which

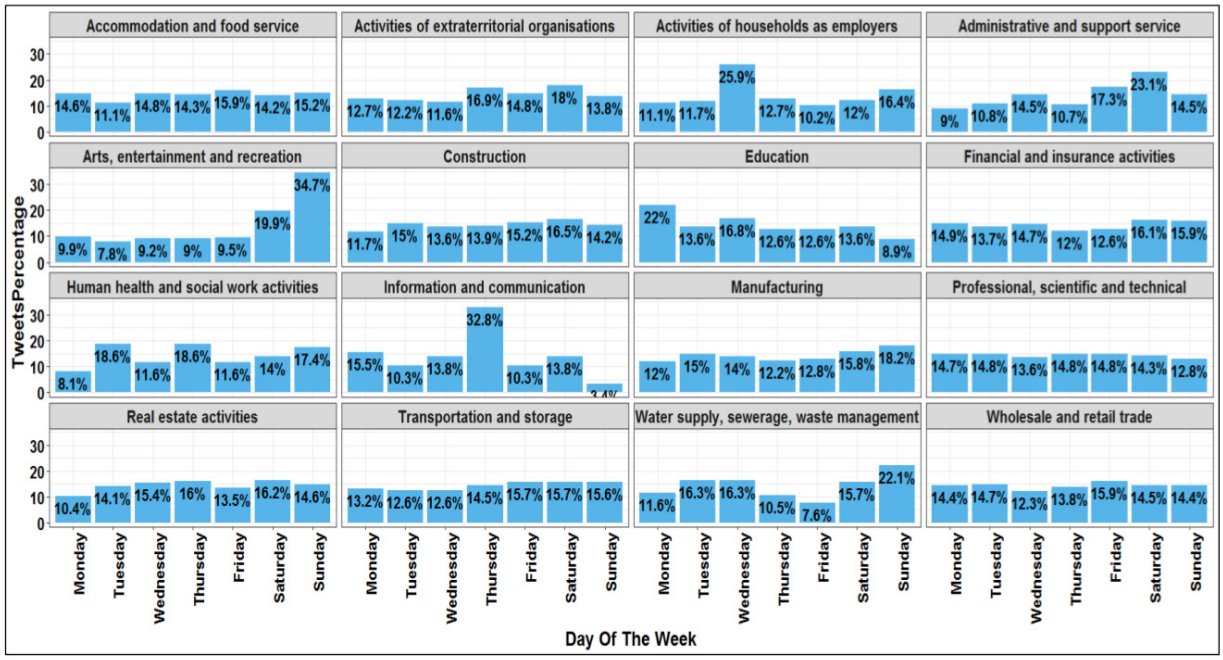


Fig. 4. **Percentage distribution of tweets during different days of the week - 2016**

is an ongoing series of NLP competitions [28]. These two data-sets are labelled tweets with different sentiment polarities - positive, negative and neutral. This model is then be used to classify our unlabelled tweets of interest in this study.

We pre-process the tweets using the established *Twokenize* package [29], from which features are derived following the example of state of the art machine learning approaches demonstrated in earlier research [30]. The features used are *unigrams*, *bigrams*, *part of speech tags*, *word vectors* and *sentiment lexica*. The first three are routine, but are extended with a custom negator. The word vectors use average, max, min and count on each dimension of the 100 dimension GloVe dataset trained on two billion tweets [31]. Sentiment lexica, which are features based on the objectivity or subjectivity of matched words present in the lexica are recommended by researchers participating in SemEval 2015 [32]. The four lexica we consider are; Bing Liu opinion lexicon [33], the MPQA subjectivity lexicon [34], AFINN [35] and SentiWordNet [36], taking various counts and averages. In total 6,033 features are derived and used for training on 7,970 tweets using some of the most adopted algorithms in NLP - Naive Bayes, SVM and Logistics Regression, using the Scikit learn package in Python [37].

The $k$-Fold cross validation technique [38] was used during the parameter optimisation search before testing the final models on the remaining 1,374 tweets, achieving a 0.66 macro average F1 score (excluding neutral) using a logistic regression classifier. The choice of metric was used to match the competition, eventually outperforming the 2015 contest winners, although tested on a slightly different dataset. The final optimised model was developed using Python scripts and used to predict the aforementioned sentiment labels for London tweets. The sentiments labels (positive, neutral and negative) for two weeks of tweet data (2012 and 2016) are then aggregated based on the industry sector these were linked to and discussed in the results section.

### C. Spatial Filtering

The objective of this step is to create an association between the tweets and the business sector they belong to. We used QGIS software for all the spatial analysis. After the tweets had been classified within the three sentiment labels (positive, neutral and negative) in the above step, we clipped the tweets which are present inside the commercial polygons boundaries (Fig.3). The basic assumption is that a tweets location (latitude and longitude) which overlays with the commercial property polygons are related or linked to the industry classification (SIC) of the commercial property. (We appreciate that we will encounter some noise using this filtering approach as industries like *retail* and *transport* will have data generated by visitors, the limitation section discusses this issue in detail). We then performed a spatial join between the clipped subset of tweets and commercial property polygons. This helped us established the SIC Code where the tweeting activity happened in space and time for each tweet. The data layer was then exported and we aggregated the tweets into all the available SIC Code sectors to perform our analysis.

## IV. Discussion and Results

Our goal in this research was to execute a data driven experiment to *integrate* geo-tagged tweets with diverse admin-

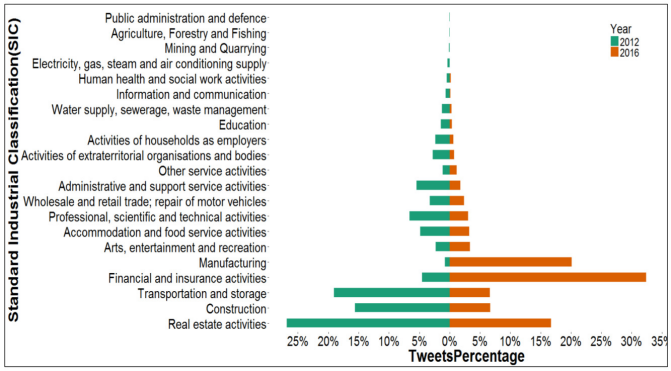


Fig. 5. Two weeks Tweets Volume Comparison

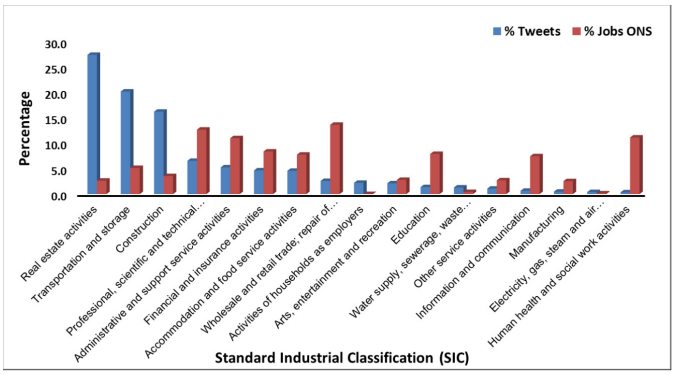


Fig. 6. Tweets Percentage compared to percentage of Jobs in each sector - 2012

istrative datasets. Our sample of tweets used for this research, associated with SIC Codes, consisted of 48,607 tweets for two a weeks period in 2012 and 51,778 for a two weeks period in 2016. The highlights of our findings are mentioned below:

### *A. Tweets Volume Comparison - 2012 and 2016 Data*

We first compared the tweeting activity in each SIC Code for the two weeks of data we have for this research i.e 2012 and 2016 as shown in Fig 5. We see a higher tweet volume for *Construction* and *Real estate* sectors in 2012 than in year 2016. This could possibly be accounted for more people working and tweeting in these two sectors since London hosted the Olympics in the year 2012, which, as per media sources created more jobs in these sectors [39]–[41]. Also, there are very limited tweeting activity in the *Information and Communication* sector and *Human Health and Social work activities* for both year 2012 and 2016. We suspect this could be down to the non-marketing nature of these industry sectors. A further comparison of proportion of jobs (an indicative of number of people employed) available in these sectors for the London region is discussed in the next section, which supports our findings. Since the Twitter data we use for this analysis comes from the London city region, we can see negligible tweet volume in both the years of data for *Mining and Quarrying*, *Agriculture, Forestry and Fishing*, *Electricity, Gas, Steam* and *Air Conditioning Supply* and *Public Administration and Defence* sectors and hence we removed these four sectors from our analysis.

### *B. Tweets Volume Comparison to Proportion of People Working in the Business Sector*

To study the penetration of social media in different sectors we compared the percentage of tweet volumes with the number of jobs available in each sector as shown in Fig.6. To make this comparison we downloaded open source jobs sector data available from the Office of National Statistics [42] for the London region. We detect an interesting trend that a relatively smaller number of employees in *Real Estate* and *Construction* sectors contribute to the greater volume of tweets. Understandably, the high tweeting volume from *Transportation* and *Retail* sector is still biased due to the public access of these places. Our further research plans to address and isolate the tweeting activity of visitors and commuters in these sectors and is discussed in the limitations section.

### *C. Tweeting Activity During the Week*

We finally examine the percentage tweet distribution pattern of a week Fig.4 in different sectors to understand the population dynamics in different business sectors and how it varies over the week. Some of the key highlights of this analysis shows a large percentage (nearly 50%) of tweeting activity in the *Art and Entertainment* sector happen at the weekend which reasonably can be accounted by the number of visitors in these sectors over the weekend. Though we appreciate we have an opportunity to enhance our analysis to filter those who participate in or those who are employed in this industry sector, nevertheless, such findings shed light on interesting dynamics of the people in the city.The *Information and Communication* sector tends to tweet more on Thursday and *Water Supply and Sewerage* activity appears at greater volume on Sunday then any other day.

### *D. Heatmaps of Tweeting Activity Across Sectors*

An activity centre is a place or location where an individual visits for a special purpose, such as work or home. We aggregated the geo-tagged tweets with the activity centres commercial property polygons, which are also associated with the SIC Codes, to create kernel density heat maps using a Gaussian kernel with a 500m bandwidth in the QGIS software. The spatial patterns of two weeks of tweeting activity are analysed. The heat-maps are broken down into different sectors for both the year 2012 and 2016 data as shown in Fig.8. The brighter red colour indicates the higher density of tweets in these SIC Code sectors. Some of the important observations illustrate there are more *real estate*, *construction* and *manufacturing* activities in London in 2012 than in 2016 [39]–[41]. Also, the western side of the London map (near Heathrow Airport) in both Transport 2012 and Transport 2016 shows quite similar tweet density and has not changed much between these two sample years.

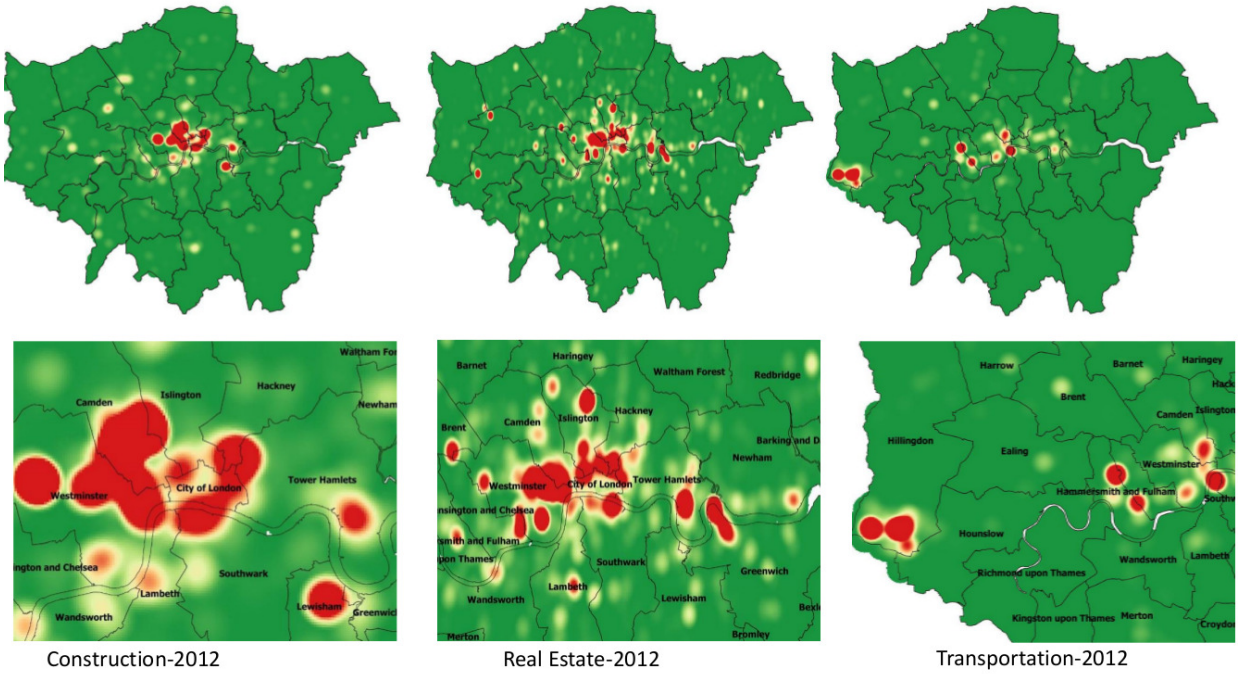


Fig. 7. **Heat-maps of Tweeting density in London's Different Business Sectors - 2012**

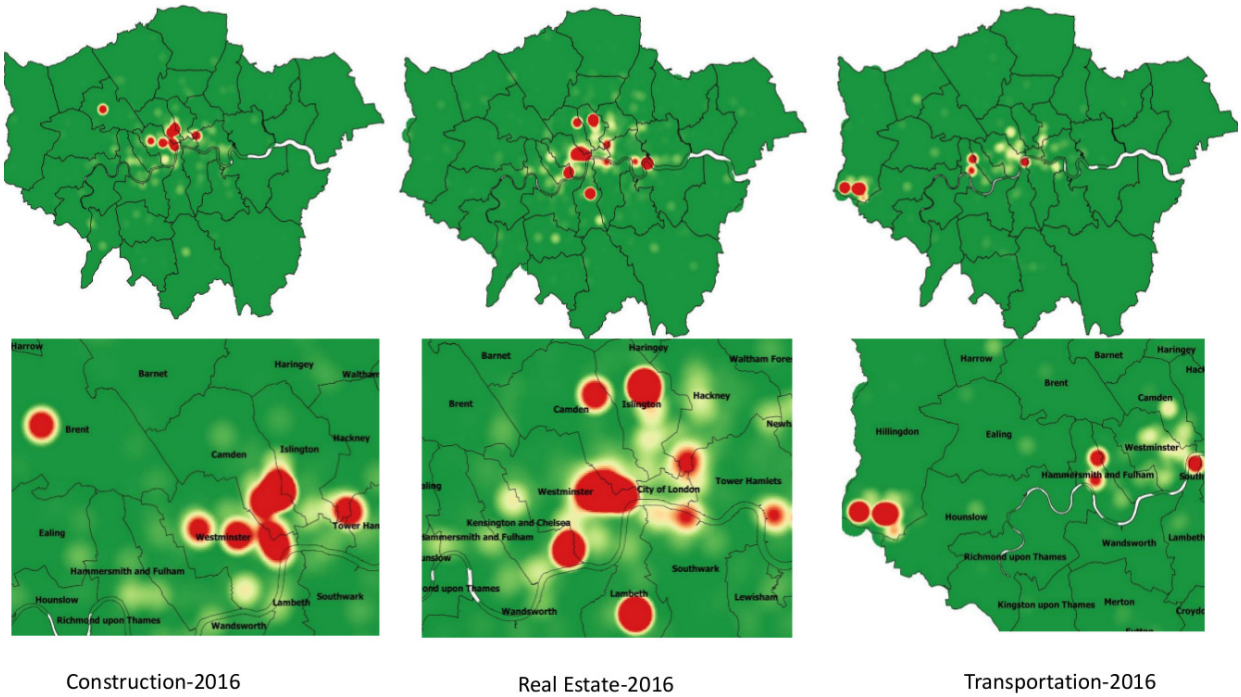


Fig. 8. **Heat-maps of Tweeting density in London's Different Business Sectors - 2016**

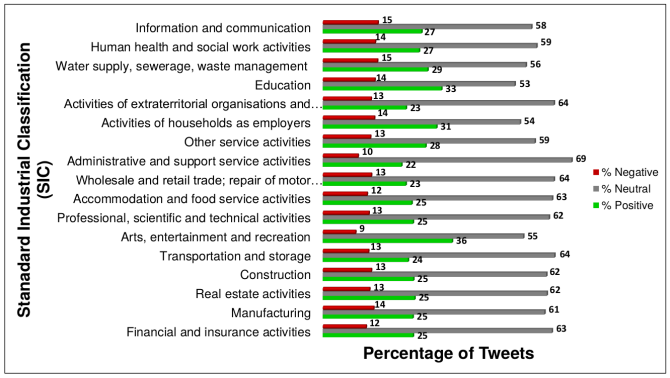


Fig. 9. **Sentiment in 2012**

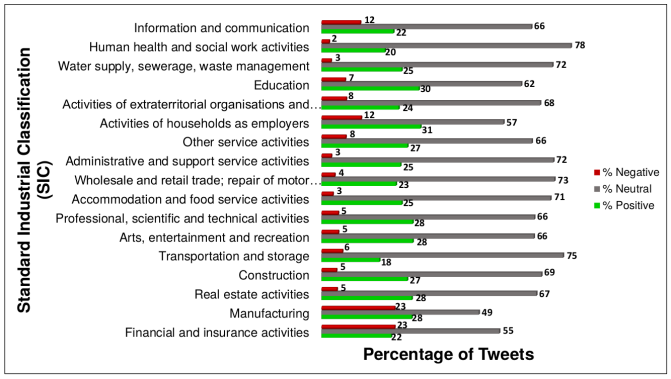


Fig. 10. **Sentiment in 2016**

### *E. Sentiment across Sectors*

Lastly, we examine the sentiment distribution across various industries in London using our two weeks of data for two distinct years. The tweets were first allocated in each SIC code. We then calculated the percentage of positive, neutral and negative tweets in each sector for both years of data. Fig. 9 and Fig.10 shows the sentiment distribution of individual industries. We note in Fig.9 that the most positive tweets are emerging from *Arts, Entertainment and Recreation* sector (36%) which also has the least negative sentiment (9%) for year 2012. Additionally the industry with the highest negative tweet for year 2012 is the *Information and Communication* sector (15%). This trend has changed in 2016 as shown in Fig.10, *Activities of Household As Employers* displays the most positive sentiment (31%) amongst all sectors and *Manufacturing* and *Financial and Insurance* industries show significant negative sentiment (23%) in general as compared to other industries.

## V. Limitations and Future Work

Whilst we have successfully covered much ground to understand the Twitter usage in industrial sector for London area, additional study is required to fill the gaps in the current analysis and answer further questions.

- It will be interesting to validate and compare the aforementioned trends using data for other two big UK cities - Birmingham and Manchester, which we plan to address in future work.
- We filtered the tweets which overlays with the commercial building polygons of the businesses and assumed that those are tweets of people working in these industrial sectors. However, there will be some noise in this approach which we plan to address using content analysis approach like *topic modelling*.
- The accuracy of the coordinates supplied by Twitter are likely reported by smart phones which may not always be entirely accurate. As such there is potential for some noise generated by people on the street appearing to come from the site in question and people in the site being missed by reporting a location away from the site. Therefore, our future work will consider ways to address such location specific planimetric accuracy issues.
- An INSPIRE polygon can have multiple uses. For example, this could include a shop on the ground floor and flats on the 2nd or 3rd floor. Since we do not have height values for the tweets, and the inspire polygons do not have a height dimension, we are not able to establish which floor the tweet has come from and as such we do not know whether the tweet is commercial or residential. This is an existing data restriction we have, however this can be partially addressed using a tweet content analysis approach which we mentioned above.
- We had two weeks data for analysis, however the trends might change over time.

## VI. Conclusions

A good deal of previous research in industry and academia has attempted to understand who are the contributors to social media data. However, to our knowledge there are limited studies that highlights the users of social media in context to various industry sectors they are linked to. This study integrates existing administrative data on businesses with geo-tagged tweets to uncover social media trends at work. Barring limitations which are future research avenues, our **main contributions** are as follows: **Firstly**, we provide a novel methodology to integrate three diverse data-sets: Twitter data, Land Registry data for London and geo-spatial polygon data of commercial business sectors which shall helps social scientists to study the behavioural trends available from social media data in the context of the industry sector the tweet users belong to. **Secondly**, we visualise the higher activity zones of Twitter data using geo-referenced tweets for London which can helps urban data scientists to understand the spatial distribution of various economic activities in a city to aid better urban planning and service delivery. **Finally**, for each sector we shed light on the mood of people evident from the tweets text using computational natural language processing (NLP) methods. Such an approach can be used in organisational studies which aim to monitor the general sentiment of the people working in different sectors as an longitudinal and low cost mechanism to supplement survey methods.

## ACKNOWLEDEGMENTS

We thank Paul Davis from Assured Property Group for providing the access to Land Registry Commercial Ownership data to conduct this research. This study is funded by the EPSRC (Engineering and Physical Sciences Research Council) Centre for Doctoral Training in Urban Science under the research grant EP/L016400/1.

## REFERENCES

[1] S. Volkova, Y. Bachrach, and B. V. Durme, "Mining user interests to predict perceived psycho-demographic traits on twitter," in *2016 IEEE Second International Conference on Big Data Computing Service and Applications*. IEEE, 2016, pp. 1–8.

[2] A. Saravanou, G. Valkanas, D. Gunopulos, and G. Andrienko, "Twitter floods when it rains: A case study of the uk floods in early 2014," in *Proceedings of the 24th International Conference on World Wide Web*. ACM, 2015, pp. 1233–1238.

[3] K. Weller, A. Bruns, J. Burgess, M. Mahrt, and C. Puschmann, *Twitter and society*. P. Lang, 2014, vol. 89.

[4] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 851–860.

[5] B. Robinson, R. Power, and M. Cameron, "A sensitive twitter earthquake detector," in *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 2013, pp. 999–1002.

[6] C. Allen, M.-H. Tsou, A. Aslam, A. Nagel, and J.-M. Gawron, "Applying gis and machine learning methods to twitter data for multiscale surveillance of influenza," *PloS one*, vol. 11, no. 7, p. e0157734, 2016.

[7] B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, and C. Ratti, "Geo-located twitter as proxy for global mobility patterns," *Cartography and Geographic Information Science*, vol. 41, no. 3, pp. 260–271, 2014.

[8] V. Frias-Martinez and E. Frias-Martinez, "Spectral clustering for sensing urban land use using twitter activity," *Engineering Applications of Artificial Intelligence*, vol. 35, pp. 237–245, 2014.

[9] D. Quercia, J. Ellis, L. Capra, and J. Crowcroft, "Tracking gross community happiness from tweets," in *Proceedings of the ACM 2012 conference on computer supported cooperative work*. ACM, 2012, pp. 965–968.

[10] W. Guo, N. Gupta, G. Pogrebna, and S. Jarvis, "Understanding happiness in cities using twitter: Jobs, children, and transport," in *Smart Cities Conference (ISC2), 2016 IEEE International*. IEEE, 2016, pp. 1–7.

[11] A. Tsakalidis, S. Papadopoulos, A. I. Cristea, and Y. Kompatsiaris, "Predicting elections for multiple countries using twitter and polls," *IEEE Intelligent Systems*, vol. 30, no. 2, pp. 10–17, 2015.

[12] D. Maynard and A. Funk, "Automatic detection of political opinions in tweets," in *Extended Semantic Web Conference*. Springer, 2011, pp. 88–99.

[13] G. Blank, "The digital divide among twitter users and its implications for social research," *Social Science Computer Review*, p. 0894439316671698, 2016.

[14] L. Sloan, J. Morgan, P. Burnap, and M. Williams, "Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data," *PloS one*, vol. 10, no. 3, p. e0115545, 2015.

[15] D. Preoţiuc-Pietro, V. Lampos, and N. Aletras, "An analysis of the user occupational class through twitter content." The Association for Computational Linguistics, 2015.

[16] D. Murthy, A. Gross, and A. Pensavalle, "Urban social media demographics: An exploration of twitter use in major american cities," *Journal of Computer-Mediated Communication*, vol. 21, no. 1, pp. 33–49, 2016.

[17] Office For National statistics. (2007). [Online]. Available: https://www.ons.gov.uk/methodology/classificationsandstandards/ukstandardindustrialclassificationofeconomicactivities/uksic2007

[18] P. B. et. al. (2017) Analyzing users' sentiment towards popular consumer industries and brands on twitter. [Online]. Available: https://arxiv.org/abs/1709.07434

[19] Ordanance Survey. (2017). [Online]. Available: https://www.ordnancesurvey.co.uk/

[20] Europe INSPIRE Directive. (2017) Infrastructure for spatial information in europe. [Online]. Available: https://inspire.ec.europa.eu/

[21] UK INSPIRE. (2009) The uk inspire regulations. [Online]. Available: https://data.gov.uk/inspire

[22] HM Land Registry. (2014) Inspire index polygons spatial data. [Online]. Available: https://www.gov.uk/guidance/inspire-index-polygons-spatial-data

[23] HM Land Registry commercial services. (2014) Commercial and corporate ownership data. [Online]. Available: https://www.gov.uk/guidance/commercial-and-corporate-ownership-data

[24] Twitter. (2017) Twitter developer documentation. [Online]. Available: https://dev.twitter.com/overview/api/tweets

[25] Companies House. (2017) Free company data product. [Online]. Available: http://download.companieshouse.gov.uk/en_output.html

[26] A. Brew, D. Greene, D. Archambault *et al.*, "Deriving insights from national happiness indices," in *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 2011, pp. 53–60.

[27] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, R. E. Lucas, M. Agrawal, G. J. Park, S. K. Lakshmikanth, S. Jha, M. E. Seligman *et al.*, "Characterizing geographic variation in well-being using tweets." in *ICWSM*, 2013.

[28] S. R. et. al. (2015) Semeval 2015 task 10. [Online]. Available: http://alt.qcri.org/semeval2015/task10/index.php?id=data-and-tools

[29] M. Ott. (2013) Python port of the twokenize class from ark-tweet-nlp. [Online]. Available: https://github.com/myleott/ark-twokenize-py

[30] O. Kolchyna, T. T. Souza, P. Treleaven, and T. Aste, "Twitter sentiment analysis: Lexicon method, machine learning method and their combination," *arXiv preprint arXiv:1507.00955*, 2015.

[31] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[32] H. Hamdan, P. Bellot, and F. Bechet, "lsislif: Feature extraction and label weighting for sentiment analysis in twitter," in *Proceedings of the 9th International Workshop on Semantic Evaluation*, 2015, pp. 568–573.

[33] Bing-Liu. (2004) Opinion lexicon. [Online]. Available: https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon

[34] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, 2005, pp. 347–354.

[35] F. Nielsen. (2011) Afinn. Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby. [Online]. Available: http://www2.imm.dtu.dk/pubdb/p.php?6010

[36] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining." in *LREC*, vol. 10, 2010, pp. 2200–2204.

[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[38] P. Refaeilzadeh, L. Tang, and H. Liu, *Cross-Validation*. Boston, MA: Springer US, 2009, pp. 532–538. [Online]. Available: http://dx.doi.org/10.1007/978-0-387-39940-9_565

[39] BBC. (2009) 2012 as job creator. [Online]. Available: http://news.bbc.co.uk/1/hi/uk/7831847.stm

[40] Telegraph. (2011) London 2012 olympics-the olympic stadium. [Online]. Available: http://www.telegraph.co.uk/finance/london-olympics-business/8641977/London-2012-Olympics-The-Olympic-Stadium-made-in-Britain.html

[41] Telegraph. (2012) Housing boom. [Online]. Available: http://bit.ly/2BoAZWj

[42] ONS - Office of National Statistics. (2017) Employee jobs by industry sector - london. [Online]. Available: http://www.nomisweb.co.uk/