WARWICK
THE UNIVERSITY OF WARWICK

MENS TAT AGI MOLEM

UNIVERSITAS WARWICENSIS

Exploring Happiness Indicators In Cities and Industrial Sectors Using Twitter and Urban GIS Data

by

Neha Gupta

Thesis

Submitted to the University of Warwick

in partial fulfilment of the requirements

for admission to the degree of

Doctor of Philosophy

Supervised by: Prof. Stephen Jarvis and Dr. Weisi Guo

Department of Computer Science

September 2018

Contents

List of Tables

List of Figures

Acknowledgments

Thank you **Almighty** for granting me an opportunity to fulfil my aspiration and blessing me with everything needed to pursue this dream. Thank you to my spiritual gurus, **Mother** and **Sri Aurobindo** for their teachings and Grace. Emotionally, I dedicate this Thesis to my father, **Prof. (Dr.) Kaushal Kumar**, who taught me to aspire high and preserve a positivity for life in all circumstances. Affectionately, I am indebted by my brother, **Major. Shubham Agarwal**, for sharing those live words in a letter years ago -"A Brave Sister of an Army Officer...", which has kept me inspired and motivated. I hope this will make them both proud from wherever they are watching.

My special and heartily thanks to my supervisor, **Professor Stephen Jarvis** who always showed trust in me and has been my mentor in this academic journey. It is his ever encouraging support that has brought this work towards a completion. He has been truly a 'Guru', showing me light on the darkest nights by offering both professional and emotional support this arduous journey requires. I express my profound gratitude towards my supervisor **Dr. Weisi Guo** for his critical evaluation, timely supervision, structured guidance and subject matter expertise that guided and enhanced my skills as a researcher. He has enabled me to be a better version of myself both professionally and personally.

My sincere thanks to EPSRC for granting CDT scholarship and CAGE - University Of Warwick for their generous funding, which helped me to sustain financially. I would like to acknowledge caring attitude of the CDT administrator **Ms. Yvonne Colmer** whose prompt assistance took care of all the administrative matters. I appreciate **Ms. Katie Martin** too for her help in admin related things.

The generosity of the academic community at large has never failed to amaze me. I would like to praise **Dr. William Rupp** and **Ms. Sara Hattersley** from Learning and Development Centre, University of Warwick for offering their caring advice and re-assuring thoughts for me. WISC Centre for Doctoral Training has been a wonderful community of research students and I deeply acknowledge the support provided by my peer research students. Particularly, **Sarunkorn, Simone, Adam, Henry, David, Nataliya** and **Tejas**, sailing with them all together this journey has become a precious memory.

I'm grateful to my uncle **Dr. Nirankar S. Agarwal** for his fatherly support and sharing his academic experience with Universities all these years. He kept me motivated with his wisdom and his guidance always promoted my confidence. As a privileged daughter, I'm indebted to my mother **Mrs. Asha Agarwal**, who is ever available to help me to strive through this challenge. Nothing can match her care and selfless love and she is genuinely a 'hope', a promise for life, as her name means. A huge credit in completion of Thesis goes to my strongest pillar of strength, my husband, **Ajay Gupta** who has invariably comforted me. 'Thank You' alone cannot convey my gesture and love for his incredible support offered day after day unconditionally. I express my immense love and gratitude for my little champ, my son, **Yash**, who despite being so little cheered me to complete my PhD Thesis. I owe so much to him in return for showing this faith, trust and understanding. I thank **all my friends and extended family** in UK, US and India for listening me and being on my side whenever I needed.

Lastly, thanks to University of Warwick and all the people who maintain this amazing green campus, arrange thought provoking workshops and provide an opportunity to bring out best in an individual's academic journey. It was undoubtedly a cherished experience.

Declarations

I, Neha Gupta, declare that this Thesis and the scholarly work presented here is a result of my own original research conducted during the doctoral studies at Warwick Institute of Science Of Cities, Doctoral Training Centre in University Of Warwick. Additionally, I also confirm that where a small fraction of work was executed in collaboration with other peer researcher in the CDT, I have fully acknowledged what their contribution was and how it has functioned in the analysis performed by me.

Parts of this thesis have been previously published by the author in the following research articles:

1. Weisi Guo, Neha Gupta, Ganna Pogrebna, and Stephen Jarvis. Understanding happiness in cities using Twitter: Jobs, children, and transport. In *Smart Cities Conference (ISC2), 2016 IEEE International*, pages 1–7. IEEE, 2016 [1].

2. Neha Gupta, Henry Crosby, David Purser, Stephen Javis, and Weisi Guo. Twitter usage across industry: A spatiotemporal analysis. In *2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 64–71. IEEE, 2018 [2].

Following manuscript is currently under peer review:
3. Neha Gupta, Weisi Guo, and Stephen Jarvis. Analysing Twitter and UK Land Registry data to explore working hours and sentiment across industrial sectors. *Submitted to PLOS One*, 2019 [3].

Research that was performed in collaboration with other CDT researchers during the development of this thesis, but does not form part of the thesis:

4. Nataliya Tkachenko, Sarunkorn Chotvijit, Neha Gupta, Emma Bradley, Charlotte Gilks, Weisi Guo, Henry Crosby, Eliot Shore, Malkiat Thiarai, Rob Procter, et al. Google trends can improve surveillance of type 2 diabetes. *Scientific reports*, 7(1): 4993, 2017 [4].

Abstract

The changing demographics and landscape of cities emphasises to better understand the factors which influence citizen happiness. Inferring happiness (sentiment analysis) indicators from Twitter text and Urban GIS data offers a scalable solution. The current research is an exploratory study conducted to apply Natural Language Processing (NLP) and GIS techniques to geo-tagged Tweets in the Greater London area in order to investigate the underlying socioeconomic and urban geography features that potentially could influence happiness. Specifically, the present research devise a methodology to explore the aggregated sentiment of people engaged in various industrial sectors by joining diverse datasets (Twitter, INSPIRE polygons, Ordanance Survey AddressBase and UK Land Registry) which so far has existed in silos in order to monitor the working patterns and sentiment trends in industrial areas in urban space. Furthermore, the proposed methodology seek insights about the Brexit related Twitter sentiment trends in targetted industrial sectors. The results of this study could enable urban planners to move beyond planning services using traditional cost benefit analyses by incorporating openly available data sources. The novel data-driven approach developed in this work has an application in analysing the mood prevalent in various economic sectors and provides an evidence to incorporate social media analytics in organisational studies, thereby offering a mechanism to monitor working patterns in near real-time using tweet intensities. The procedure outlined can be used to extend more traditional survey and sample based methods in behavioural studies and also could be an enabler for policy makers to perceive the sentiment of a targeted sector of society in light of an existing social phenomenon.

Sponsorships and Grants

The work conducted in this Thesis has been majorly sponsored by EPSRC Centre for Doctoral Training in Urban Science and Progress, Grant Number - EP/L016400/1. Also, the first experiment in this work was additionally supported by Warwick Centre for Competitive Advantage in the Global Economy (CAGE), University Of Warwick, Coventry UK.

Acronyms

API Application Programming Interface.

BREXIT BRitain EXIT.

CRM Customer Relationship Management.

ESENER European Survey of Enterprises on New and Emerging Risks.

GIS Geographical Information Systems.

GML Geographical Markup Language.

GNH Gross National Happiness.

HTTP Hypertext Transfer Protocol.

INSPIRE Infrastructure for Spatial Information in Europe.

JSON Java Script Object Notification.

ML Machine Learning.

NLP Natural Language Processing.

OECD Organization forEconomic Co-operation and Development.

ONS Office Of National Statistics.

QGIS Quantum Geographical Information Systems.

SCRM Social Customer Relationship Management.

SIC Standard Industrial Classification.

STT Spatio Temporal Textual.

VGI Volunteered Geographic Information.

Chapter 1

Introduction

1.1 Motivation - Social Media Data to Study Sentiment

The internet has become an integral part of our life today. It provides an infrastructure to access unprecedented information about any and practically all aspects and topics about the world we live in. Nearly 47 percent, 3.5 billion, of the global population was using internet according to the United Nations' 2016 report, *The State of Broadcast* [5]. A wide range of our daily activities including banking, shopping, education, gaming, entertainment, exploring places of interest, accessing news and myriad facts and data for everything under the sun, are now possible due to the internet. As these countless patterns of life evolve, the advancement of internet technology is bringing a phenomenal change in the modes of social exchange of ideas as well. People are increasingly cooperating, intermingling and networking with not only their friends, colleagues and peers but also with complete strangers by resorting to novel communication channels commonly referred to as social media platforms, generating huge volumes of textual, spatial and temporal data expressing their opinions and sentiments regarding all happenings and events.

For decades, surveys or opinion polls have been the primary resource to ascertain public perception about products, services, policies, events and other research themes. But availability and ease of access of passive data from the social media today provides an unprecedented opportunity to assimilate public perceptions in real time. Monitoring social media provides the latest and ongoing information and perception of public reaction to diverse events which can assist in planning and decision making.

A noteworthy social media platform of comparatively recent origin is *Twitter*, a micro-blogging site launched in 2006 by Jack Dorsey and his associates from

San Francisco. However even in this short span of time, Twitter has become a truly global medium for public expression on current topics or events of general interest, e.g. politics, sports, disasters, celebrities, environment, government policies, business products or corporate social responsibility issues, current topics of interest and consequence, and other sundry fields of study [6]. All tweets have time fields associated with them and they can also be geo-tagged to pinpoint location of their inception. Consequently, Twitter data offers intriguing opportunities for uncovering spatio-temporal-textual (STT) trends about people's sentiments in reaction to events of consequence at minimal expense. It is no wonder, therefore, that researchers from diverse disciplines are using text mining (NLP), spatial and temporal analysis techniques to computationally process Twitter data in ever increasing situations.

As examples of this trend, scores of studies in diverse disciplines have explored Tweets, viz., for crisis response collaboration during flooding [7, 8]; monitoring earthquake events [9, 10]; surveillance of disease spread [11]; monitoring mobility patterns [12]; assisting data driven urban planning [13]; tracking community happiness [1, 14] and predicting election outcomes by assessing political orientation of the sample [15, 16], etc. A growing body of researchers applying similar methodology has also been active in uncovering emotions from Twitter data [17] and generating mood heat maps pertaining to city locations [18] as well as comparing and contrasting parameters of well-being and happiness between cities [19] [20]. Additionally, Twitter data is being examined in depth to gauge age, occupation, social class, race and locale of its users [21–23]. in order to devise the underlying demographics of Twitter users and proximate the behavioural trends of society allowing social scientists to make group comparisons.

Whilst social media and GIS data has been successfully used in many of the above mentioned research endeavours, there has been little discussion about how new forms of data such as large-scale geo-tagged Twitter data can be capitalised to study the trends and social phenomenon existing in industrial areas in urban spaces. Therefore, this Thesis is an exploratory study conducted to investigate if Twitter data can be used to study the happiness cues (sentiment analysis) in cities, in particular industrial area in urban spaces. Such Twitter based analysis sentiment analysis could offer a scalable and low cost mechanism in addition to the survey methods that could be employed as low cost interim approach in order to monitor the behaviour and happiness of people engaged in various industrial sectors.

1.2 Summary of Research Questions

The main research questions this thesis examines are discussed.

The demographics and landscape of cities is changing rapidly, and there is an emphasis to better understand the factors which influence citizen happiness. Inferring happiness (sentiment) from social media data is a scalable solution. Therefore in Chapter 3, I applied natural language processing to 0.4 million geo-tagged Tweets in the Greater London area to understand the underlying socioeconomic and urban geography parameters (as available from London Data Store Ward level data) that influence happiness. Particularly I explore **RQ1. Which demographic variables correlate strongly with the positive sentiments of aggregated tweet text and what is the statistical measure of influence of those variables on the tweet sentiment?**

High resolution geo-tagged social media data presents an opportunity to better understand behavioural patterns and sentiment of the populace. Whilst significant work has been conducted in various targeted social contexts, very little is understood about differentiated behaviour in diverse segments of society, for example, monitoring the sentiment of people linked to various industries. Devising a surveillance tool to perceive the workforce sentiment in industries can be valuable for policy makers and service providers to gauge an impact of new policies and can be used as a low cost alternative to survey methods in organisational studies. In chapter 4, I go on exploring **RQ2. Can social media data available from Twitter particularly, be used to understand happiness indicators in various industrial sectors?** Whilst RQ1 aims at generating a high level view of sentiment as a proxy of happiness in London Wards, in RQ2, more fine grained analysis of sentiment is targeted towards industrial sectors by utilizing spatial data available from UK Land Registry and Ordanance Survey AdressBase datasets. RQ2 is then split up into two sub-questions: **RQ2.1. how social media usage and general sentiment vary across the geographic and industry sector landscape,** which shall help us identify who are the main contributors to Twitter conversation in industry and **RQ2.2. What is the impact of working hours (estimated from Tweet intensities) on aggregated Twitter sentiment within each industrial sectors?**

The methodology proposed to answer the RQ2 enables us to link the real time geo-tagged Tweets with diverse industrial sectors. It will be really interesting to explore if this methodology can also be incorporated to mine event related sentiment within industrial landscape to perceive the mood of the population following an event that could impact the working conditions in a particular industry sector. Therefore, expanding the usage of Twitter data for event related studies my final research within the scope of this thesis explore: **RQ3.Can Twitter data be leveraged to ascertain the aggregate industrial sentiment about political events such as Brexit and what are the limitations, if any, Twitter analytics pose for any such analysis ?** Such exploration can aid policy makers to gainfully investigate the return of Twitter analytics and observe the sentiment of people in industries during an event like EU referendum.

1.3 Thesis Contribution

This thesis attempts to explore the use of social media and other openly available urban datasets to monitor and understand citizen sentiment (as a proxy of happiness) in urban spaces, in particular, at industrial locations. In effect, the positive sentiment observed in the aggregated tweet text within each industrial sectors during work hours is in-line with the conventional economics 'utility theory', defined by Rothbard and Edwards, which suggests that people who allocate more time to work derive more satisfaction from it [24]. This research incorporates new forms of data from social media sources to measure the sentiment of working population and unlock an opportunity to explore the possibility of using big data to answer research question which so far has been tested using qualitative or survey based methods only.

As a methodological contribution, this research forge a capability to integrate Twitter data with urban datasets (property data from Land registry and spatial INSPIRE Polygons) in order to passively mine and monitor people sentiment in industrial spaces. This method can be applied to monitor the people behaviours in urban spaces using state of the art social media mining techniques alongside survey based methods. In general, the research findings of this Thesis provide new evidence to strengthen the revelations of previous research and also provide a methodology and application as a possibility to incorporate big data in order to understand happiness in cities, specifically industrial sectors using social media data.

The results of the first experiment which examines the relationship between tweet sentiment of 0.4 million geo-tagged Tweets in the Greater London with London ward socio-economic characteristics verifies the established thinking: that job opportunities correlate strongly with positive sentiments. Additionally, the experiment generate other two insights that happiness has a negative relationship with the number of children and happiness has a parabolic relationship with access to public transportation. The latter implies that those that rely on, but do not have strong access to public transport are the least happy. Separately, the number of jobs, number of children, and accessibility to public transport explain 33-47% of the variations in tweet sentiment [1].

As a second valuable contribution to knowledge, this research reveals the usage of social media, particularly Twitter, across various industrial sectors, using the Greater London region as a case study. Unlike existing studies that incorporate surveys to study industrial trends, a novel geo-computational approach is devised to fuse location specific social media (Twitter) data with business locations (leveraging diverse property databases). The baseline results for Greater London identifies *Construction*, *Real Estate*, *Transport* and *Financial Services* industries consistently have stronger Twitter footprints. By applying NLP and geo-spatial techniques to Tweets the prevailing sentiment within each industry sector is uncovered. A further exploration of tweeting intensities is studied to explore the working hours and its effect on aggregated sentiment (as a proxy of happiness) within various industries. Computational Natural Language Processing methods were then employed to derive tweet sentiment and establish the operational working hours from the time series of tweet intensity. The methodology is tested on data from London UK and whilst the results reveal no general trends across industrial sectors, the scalable approach does show that for the *financial and insurance* sector, the duration of working hours as perceived using tweets intensity does display a co-relation with the tweet sentiment [2].

As a third contribution, a preliminary analysis about the Brexit related communication pattern and sentiment trends on Twitter space for industrial landscape is studied. The study based on geo-tagged London tweets spanning over the Brexit period (7 days before and 7 days after the date of EU referendum) to corroborate or highlight limitations of the proposed methodology for understanding political event related sentiment from industries. It was discovered that not all industrial sectors participate equally in event related conversations and there exist a minimum threshold of tweets necessary to conduct any industrial sector specific event related sentiment analysis. However, the sentiment trends from the following three industries, viz. *Construction*, *Financial and insurance Sector*, and *Accommodation and Food Services* can be gainfully investigated using Twitter data. For the remaining industrial sectors researchers would need to depend on survey methods given low social media participation level.

1.4 Conclusion and Thesis Outline

In this current **introduction chapter**, a brief overview and background of social media analytics is given highlighting the gaps in knowledge which this Thesis endeavour to address.

In **Chapter 2**, I present an overview of the literature. This shall cover the application domains where social media analytics has been applied and further dwell into the studies that has particularly inspired my research motivations. I also provide a description of state-of-the-art methodologies avaliable in NLP (sentiment analysis) and GIS analysis which has been employed in my research. **Chapter 3** presents my work conducted to explore sentiment in relation to socioeconomic and geographical feature of London Ward (RQ1); **Chapter 4** describe the spatio-temporal analysis that lead to the discovery of Twitter sentiment in industrial sectors and uncovers its relationship with working hours (RQ2); **Chapter 5** further builds up on the methodology proposed in Chapter 4 to understand communication pattern and sentiment in industrial sector during the time of a radical economy impacting event Brexit (RQ3). **Chapter 6** discusses the work on Twitter Analytics in cities and industrial sectors and evaluates the methodological contribution this research makes to the field of social media analytics with a proposal for future research directions. **Chapter 7** concludes this thesis, with the key contributions summarised.

Chapter 2

Literature Review

This chapter provides an introduction to the current status of social media analytics in general. It goes on then to furnish a brief background of the academic research that motivated my search of Twitter and urban GIS data for identifying gaps, if any, in the exploration of sentiment analysis in industry sectors particularly in city demographics. The chapter culminates with a technical overview of state-of-the-art methods employed in the analysis of tweets and urban GIS data.

2.1 Overview of Social Media Analytics

An ever escalating volume of data generated by social networking sites such as Facebook, Instagram, Flickr, YouTube and Twitter, etc. is facilitating the development of innovative research techniques for potentially isolating pervading social trends and assessing their impact on our society. One such procedure is the Social Media Analytics, a three step process of "capturing", "understanding" and "presenting" available information from social media websites [25]. In the first stage, data is acquired from a variety of social media sources by means of automated methods for example, third party tools, web crawling or data scrapping APIs. Secured data, after tidying up noise, is then subjected to diverse modelling and visualisation techniques (sentiment , temporal, statistical and spatial analysis) to discover underlying causes of the trends [25] [26]. Finally, results from these analytical steps are evaluated, summarised, and visually presented to the end-users for assistance in decision making [25]. The foregoing procedure of data-mining can be sub-divided in to three categories, social network structure mining, social media content mining and social media usage mining [27].

Since a large proportion of social media content is image, audio, text, location and time related data, researchers have developed an automated human age estimator analysing the image content which was crawled from Flickr [28] [29] and have fused audio, visual and textual features from YouTube and Facebook to perform multimodal sentiment analysis [30]. The growing number of users patronizing social media has encouraged social scientists to study the impact of social media itself on human psychology. One study sought to find the relationship between social media use and GPA score of students, and discovered no negative linear correlation [31]. In contrast to this, however, another study reported that significant social media usage negatively affects the state of happiness due to higher levels of techno-stress resulting in a lower task-performance [32]. Yet another study used social media data to assess the visibility of Americas most admired fortune 417 companies in digital public domain [33]. It is therefore evident that exploring the impact of social media on society is becoming a mainstream field of research in social sciences.

2.1.1 Application In Businesses

A number of industrial sectors have started using social media data for effective market analysis and making smart decisions by incorporating vital features in their products and services based on public opinions and appraisals garnered through social media conversations [34]. Some businesses are taking them in to consideration for effective pre-launch movie advertising [35] and blogging function to promote sales of their merchandise. Diverse firms hunt for feedbacks from consumers on product quality [36], track down strengths and weaknesses of their competitors [37]. The CRM function which once held the customer and organisation communication records in their invisible propriety databases archives is now being upgraded to monitor, store and address the communication taking place openly on social media platform - often termed as *social CRM* [38]. Organisations like International Business Machine Corporation (IBM) of Armonk, New York, USA, regularly perform content analysis of social media feeds to scrutinize queries and complaints to suggests actionable response based on sentiment analysis and topic analysis of such posts [39].

Business Intelligence is another significant area of social media analytic application for the business decision-making process from operational level to the strategic level. The records for knowledge discovery can be sourced from internal records of a company or can be obtained from external channels such as text-mining and web-mining from social media platforms to pinpoint and extract business relevant information [40]. Hence topic based sentiment analysis utilizing natural language processing (NLP) has been extensively exploited in business intelligence [41]. According to the report published in Techcrunch [42], the most successful businesses monitor structured and unstructured data from prominent social platforms and generate social media index of their own relevance. This process enables them scrutinize the pulse of ongoing consumer sentiment to a degree that traditional surveys were incapable of. One of the most successful companies to leverage social media intelligence is Dell, which has linked more than 21 million social connections on Facebook and Twitter alone. Through these sites, Dell has collected over 24,000 suggestions from crowd-sourcing social feedback, of which 550 relevant ideas since 2007 were implemented in their product line [43].

Though industry has rather extensively used social platforms to generate business value data for functions such as customer retention or customer relationship management, to our knowledge not much effort has gone in to monitoring behavioural trends in its own workforce. The only direction that *Human Resources* seems to have taken is to utilise Linked-In for recruiting quality professionals [44]. Generally speaking, for assessing the *well-being of the workforce*, it has been a general trend to resort to corporate surveys and online polls. Therefore, our study aims to explore the extent to which social media data, specifically Twitter, can be used to fathom sentiments and moods in industrial sectors.

2.1.2 Application In Government

Governments have also been active in utilising social media platforms to communicate digitally with general public. It may also be pertinent to explore the current scope of their operations in social media analytic and to determine the underutilised areas, if any.

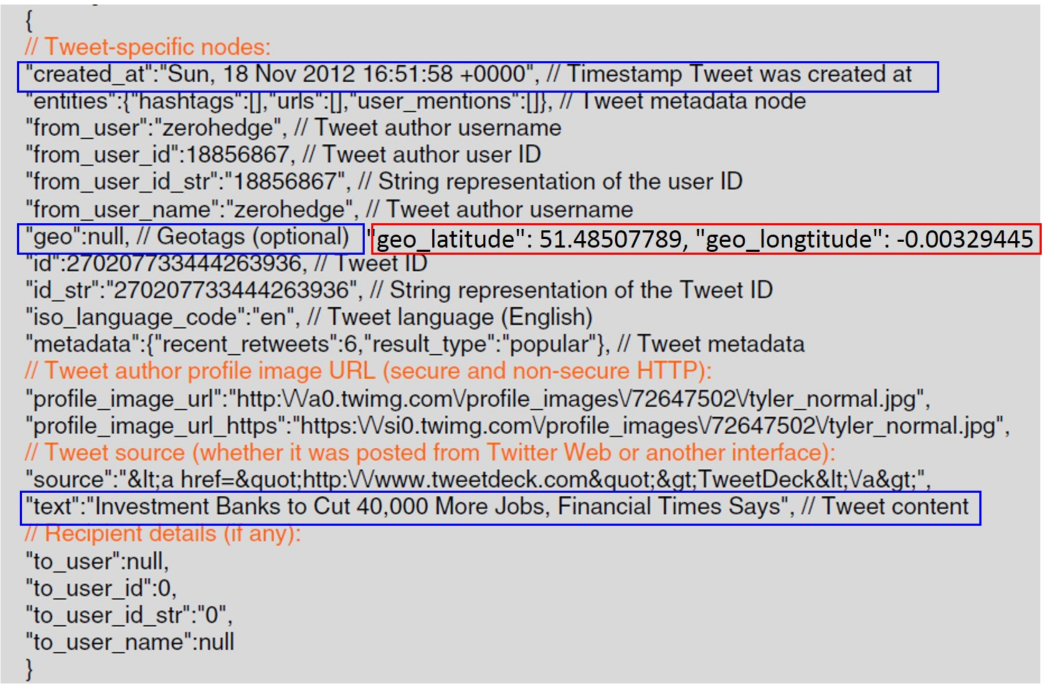
Earlier studies have highlighted a remarkable growth in usage of social networking tools by government departments; for example, in USA alone Facebook usage escalated to 87 percent in 2011 from a meagre 13 percent in 2009, and that of Twitter from 25 percent to 87 percent [45].In 31 most advanced Informational World Cities, Twitter was found to be predominantly used social media platform by government in comparison to Facebook and YouTube [46]. Twitter has been utilised in discussions for the purposes of *crisis management* and *emergency situational awareness* [47] [48] [49] in multiple cities and countries globally due to its potential for the rapid dissemination of information. Many studies have emphasised on the feasibility of quick response from the governments as well as people in crisis situations, and other civic problems. One study has reported the utilisation of computational Natural Language Processing techniques to study the implicit content available in tweets during Hurricane Sandy in US to identify the secondary resource needs of people and proposed a design of a responder system that captures early signs of crisis from social media channels [48]. In a similar manner, researchers provide overview of an Automated Web Text Mining (ESA-AWTM) system of Australian government Crisis Coordination Centre (CCC) that unveils the power of twitter messages in emergency situational awareness such as an earthquake [49].

Researchers evaluating the use of social media in government sector point to the fact that the government departments mostly disseminate or "push" information about their own policies and work often bypassing traditional mass media communication channels [50]. The extraction or "pull" of information by governments generally occurs during crisis situations as mentioned above. An Organization for Economic Co-operation and Development (OECD) report in the year 2014 states: "The purpose and returns of social media use by institutions are not as clear as they are for political personalities. This leads to uncertainty among government institutions about how to best leverage social media and about the implications for strategic objectives and day-to-day operations." This clearly means that social media ought to be used for "active listening" or "pulling" information which should help governments to a priori de-escalate problems and satisfy citizen necessities before they become a subject matter of discontent [51] [52].

Traditionally, governments have employed survey methods to perceive outcomes of their policies, but now with the availability of alternative sources of information a promising opportunity exists to "pull" views and sentiments of people in real time. However, experts have pointed out that social media users may not be representative of the entire population due to disparities of resources available to diverse segments of the masses. In light of this constraint, chapter 5 of this thesis explores the possibility of using social media (Twitter) analytics to monitor the sentiment of people engaged in industry during a radical event viz. EU referendum, that could drastically impact the UK economy. This methodology may provide a mechanism to "pull" opinion of the people in context of future research.

2.1.3 Twitter and GIS data

Twitter has achieved significant penetration (25 % of adult population in the UK), and usage (over 500 million messages per day worldwide) making it a powerful platform to share and measure messages about people behaviours and attitudes. A *tweet*, which was initially 140 characters and now 270 characters long text message, reveals the 'sentiment' of the tweeting user in real time, and therefore forms the essence of Twitter data. Additionally, meta data associated with the tweet textual content: *entities* and *places* are of great interest to researchers since it unveil the spatial enriched communication patterns that exist within our society. Tweet entities are essentially user mentions (starting with character '@'), hashtags (a keyword associated to social phenomenon starting with a character '#'), URL's and other media, whilst places can be geo-tagged location of a tweeting user expressed in 'latitude' and 'longitude' field of a tweet [53]. Finally, tweet *time-stamp* reveals the time of the user timezone when a particular tweet message was shared on the social media platform. All of the above mentioned tweet objects are available in a JSON format i.e. key-value pairs, which can be seen in a see Figure. 2.1 as described by earlier researchers [54]. Highlighted in the figure is the 'geo' field specifies the *location* of the user (if not null) by the values of 'geo_latitude' or 'geo_longitude' keys, 'created_at' reveals the *time-stamp* and 'text' key shows the *tweet* message of the user. These prime data objects enable an integrated spatio-temporal-textual (STT) [55] view for knowledge discovery in various social contexts.



```
{
// Tweet-specific nodes:
"created_at":"Sun, 18 Nov 2012 16:51:58 +0000", // Timestamp Tweet was created at
"entities":{"hashtags":[],"urls":[],"user_mentions":[]}, // Tweet metadata node
"from_user":"zerohedge", // Tweet author username
"from_user_id":18856867, // Tweet author user ID
"from_user_id_str":"18856867", // String representation of the user ID
"from_user_name":"zerohedge", // Tweet author username
"geo":null, // Geotags (optional) "geo_latitude": 51.48507789, "geo_longtitude": -0.00329445
"id":270207733444263936, // Tweet ID
"id_str":"270207733444263936", // String representation of the Tweet ID
"iso_language_code":"en", // Tweet language (English)
"metadata":{"recent_retweets":6,"result_type":"popular"}, // Tweet metadata
// Tweet author profile image URL (secure and non-secure HTTP):
"profile_image_url":"http:\/\/a0.twimg.com\/profile_images\/72647502\/tyler_normal.jpg",
"profile_image_url_https":"https:\/\/si0.twimg.com\/profile_images\/72647502\/tyler_normal.jpg",
// Tweet source (whether it was posted from Twitter Web or another interface):
"source":"&lt;a href=&quot;http:\/\/www.tweetdeck.com&quot;&gt;TweetDeck&lt;\/a&gt;",
"text":"Investment Banks to Cut 40,000 More Jobs, Financial Times Says", // Tweet content
// Recipient details (if any):
"to_user":null,
"to_user_id":0,
"to_user_id_str":"0",
"to_user_name":null
}
```

Figure 2.1: Tweet JSON [54], STT features of Tweet

Propriety tools in the business domain glean insights from this rich source of Twitter data and have developed bespoke software and web products such as - Sprout Social, Topsy, Trendsmap, Tweetreach , Statweestics, Brandwatch, Coosto, Pulsar Trak, Crimson Hexagon, Ripjar, Radian6, Traackr, Tweetdeck. Most of these tools offer point-in-time social media analysis and search services to help businesses set up alerts and monitor ongoing consumer sentiment. For example, the toolkits like Hootsuite, gives business organisations a visibility on their social engagement about their twitter account. These tools have built-in analytics in place to measure the number of tweets, replies to the tweets, mentions, number of likes, comments etc. It present graphs about a rise and fall of particular terms discussed on social media. IBM Social Media Analytic platform analyses billions of social media comments and provide personalised results through interactive dashboards and customised reports [56]. Oracle Social engagement and Monitoring Cloud Services helps organisations to create a unified, 360 degree view of their customers and protests by blending the social and private data of the enterprise [57]. Their service uses advanced statistical language modelling to address the inaccuracy and bluntness of keyword search. The resulting information is folded into Oracle existing data management tools to optimise business processes and improve business intelligence. SAS Sentiment Analysis Manager, which is part of SAS Text Analytic program, offers features to scrap content from social media outlets and creates reports which describe the consumers opinion in real time [54]. Although, this vast spectrum of commercial systems provides the retrieval of Twitter data with in-built analytical capabilities, they are primarily designed for business needs, are black-box and incurs a huge cost in their implementation. Such products being customised by nature these do not provide access to underlying raw Twitter data required for more fine grained analysis of research questions.

Scientific community therefore primarily use API scrapping techniques to retrieve raw Twitter data or purchase historic Tweets from Twitter and its commercial partners for their specific research needs. Tweets from Twitter's users account are available to download in JSON format through Twitters *Search API* for batch requests of past historical data and *Streaming API* for near real-time data [54]. Although, one may write a scrapping script (written in Python, Java, R etc) targeting towards a particular 'keyword' search in a given geographical 'bounding box' (geo-coordinates) using Search API and Streaming API, these techniques allows retrieval of only free 1% sample of the entire Twitter dataset that matches specified criteria. Such sources although has been used extensively in Twitter based research studies, is believed to incur bias and may be strongly dependant on the coverage and the type of analysis aimed for [58]. Hence, such analysis is based on only a fraction of the whole Twitter dataset. To address this uncertainty, another study incorporates the use of open source Twitter data called 'Sample API' to overcome the issues of bias in the Streaming API [59]. However, resorting to an additional data source which although fill gaps of existing data, can potentially increase complexity of the research design and is often time consuming.

In order to overcome the aforementioned data limitations of API's, additional approaches exist to acquire a full data sample from Twitter such as purchasing it from GNIP as a 'Firehose', a feed provided by Twitter's commercial partner that allows access to 100% of all public tweets [58] . The full sample of data for any time window can also be directly purchased from Twitter using their *Historical PowerTrack* that searches the entire Tweet archive, and examines each Tweet posted during the period of interest or any other selection criteria requested and generates a dataset of tweets matching the desired query [60]. This data is then available for download (in JSON file format) by the link shared by Twitter. The research outlined in this Thesis takes advantage of a full sample of Twitter activity (purchased datasets) and the details of this data is mentioned in respective contribution chapters (Chapter 3, 4 and 5).

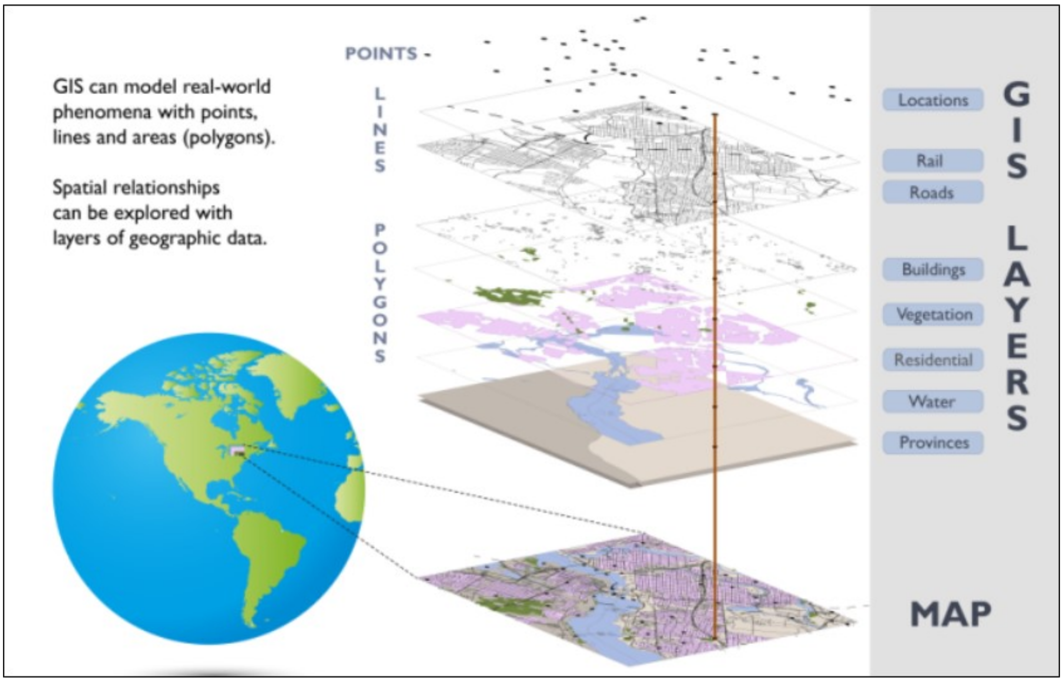

GIS can model real-world phenomena with points, lines and areas (polygons).
Spatial relationships can be explored with layers of geographic data.
POINTS
LINES
POLYGONS
Locations
Rail
Roads
Buildings
Vegetation
Residential
Water
Provinces
GIS LAYERS
MAP

Figure 2.2: GIS modelling of real world [61]

Tweet meta-data, that provides the location of the Twitter user creates an unprecedented opportunity to link this geo-tagged entity of social media (e.g. Twitter's *latitude* and *longitude* feature) to Earth observation data garnered from Geographic Information Systems (GIS), thereby offering a possibility to uncover how people can organise themselves in various societal contexts by readily knowing where everybody and everything are located [62]. The GIS system defines a place or a location on earth using geographic objects that have different geometries. Two such geometries of interests are: *Point* (zero dimension, i.e. a spatial occurrence of a single event for example disease outbreak, *a Tweet* etc.) and *Area* (polygons , i.e a two dimensional area with boundaries for example agriculture fields, *buildings* etc.) [63]. The geographical attributes (positioning and location) of any area, for example, buildings, can be coded in digital features in a *polygon* layer of a map (Figure.2.2) and is available as GML (Geographical Markup Language) shape files. These shape files can be read by GIS systems for any spatial analysis. London Data Store and Land-Registry UK facilities the open source access of such GML shapefile files (polygons or areas) for the Greator London Ward profiles and businesses locations respectively, which have been extensively used to explore the answers of the research questions mentioned in Section 1.2.

2.2 Twitter Sentiment and Urban Spatial Analysis

“Language lives braided with emotions. Language is never isolated, never separate form the world of moods and emotions”[64]. Twitter’s textual content sentiment characteristic is increasingly used as an indicator to mark people’s state of happiness and opinion at a given time, location and in various social contexts. *Sentiment Analysis*, the study to determine the subjectivity and polarity (positive, negative and neutral) of text data [65], has therefore attracted significant attention from the research community. Sentiment Analysis of a text can be performed using off-the-shelf tools or can be done by incorporating computational natural language processing methods - lexicon based text classification or supervised machine learning.

Some off-the-shelf software and web systems adapted in Twitter Analytics are mentioned in the Table 2.1. These tools offer impressive functionality to retrieve and analyse Tweets, powered by visual aggregated views of Tweet sentiment for a given timeline but are often limited to finding opinion about eminent political personalities, famous ongoing global events, products reviews etc. The trends emerging from these tools are rather complex to understand in relation to the social context and location given that these tools do not reveal the detail information about the Twitter’s meta-data and other data sources incorporated in their design. Additionally, their analytic metrics are often customised and they are designed as blackbox for the end users. Perhaps, earlier researchers reasonably have bench-marked the quality of Twitter sentiment analysis tools [66] and reported that their quality of performance and accuracy differ considerably, hence having implications on specific Twitter based analytics. Therefore, this research relies on scripts and tools coded in R and Python using state-of-the-art lexicon and machine learning based NLP techniques for Tweet sentiment classification, a brief background of which is described in next

Name	**Platform**	**Open Source / Commercial**	**Capabilities**
GATE (General Architecture for Text Engineering)	Desktop - Unix, Windows XP	Open Source	Language processing and with built-in capacity for comparing annotation data on documents and generating quantitative metrics.
Senti Strength	Desktop - Unix, Windows XP	Free for academic research	Estimates the strength of positive and negative sentiment in short texts on a scale of -5 to +5.
Netlytic	Cloud-Web Based	Free for academic research	Research Community supported text and social networks analyzer for social media to study public discourse on social media sites.
Pulsar	Web Based	Commercial	Scrape and Analyse Twitter text
Twitonomy	Web Based	Open Source	Visual analytics on anyone's tweets, retweets, replies, mentions, hashtags
Trendsmap	Web Based	Commercial	Shows the latest Twitter trending hashtags and topics from anywhere in the world using text and GIS anlaysis
Mozdeh	Desktop - Unix, Windows XP	Free for academic research	Gathers and provide sentiment analysis for upto two weeks old Twitter data.

Table 2.1: Twitter Analysis Off-the-shelf Tools

two sub-sections.

2.2.1 Sentiment Classification - Lexicon Based

Lexicon based sentiment classification of Tweets text relies on sentiment *lexicon*, which is a collection or corpus of words annotated with sentiment orientation viz. positive, negative and neutral. Creation of lexicon word lists can be done *manually* by annotating the words with sentiment orientating which is often time consuming, and be achieved using automated *dictionary- based* methods [67].

Dictionary-based approaches start with the collection of small set of seed words and an automated script looks for its synonyms iteratively in online sources (e.g. Thesuarus) or well know corpora WordNet [68] and adding words to the collection list, stopping only when no other words are found. One such immensely used lexicon in sentiment analysis task is Hu and Liu opinion lexicon [69] that classifies English words into positives (1967 positive words) or negatives (4783 negative words) sentiment wise. The advantage of using lexicon based methods for sentiment analysis is that it does not require training data and often claimed to be successful for domain independent sentiment classification. Researchers have therefore used lexicon-based algorithms to create tools to understand the sentiment orientation of Tweet texts. One of such tool is *Sentistrenght*, that interprets the peculiarities of the language frequently used on social media, and provides a broad range of sentiment score: -5 to +5 , i.e., from extremely negative to extremely positive classification of short texts, respectively and has been extensively used in many social media sentiment analysis tasks[70] [71] [72] [73].

In a similar vein, researchers draw on lexicon resources whilst coding a sentiment scoring algorithm during the design of their text mining process. Figure.2.3 from an earlier study [74] summarises the sentiment classification process that utilises opinion lexicons. Before applying any of the sentiment extraction methods, a common practice in text mining is to perform data *pre-processing* as it reduces the computational complexity during the word comparison stage (with lexicons) reached in a scoring algorithm. Typical text pre-processing procedure includes *tokenisation*, *stemming* and *stop-words removal.* Tokenisation process creates *unigram* words in a bag-of-words model that can be analysed further in the text mining process. Stemming is a technique of replacing words with their root or stem words , for example, 'organizes', and 'organizing' both originate from the stem word 'organize' will be replaced in text by 'organize' [75]. Similarly, the stop-words frequently used in English language that do not carry any meaning to themselves such as 'a', 'of', 'the', 'and' etc. are removed during the text pre-processing phase. The remaining words in the text sentence are then compared with words in various state-of-the-art opinion lexicons compiled by previous NLP researchers to obtain the sentiment label.

Lexicon based simple average sentiment scoring of Tweets has been used in earlier research [76], however it has not been applied to urban contexts to understand the underlying sources of happiness which is the aim of the present research.

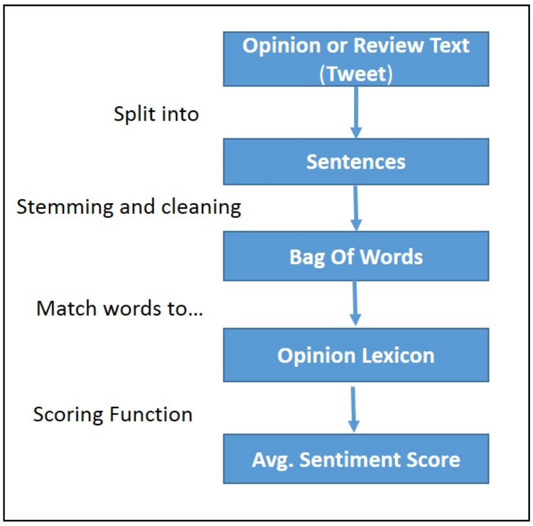


Figure 2.3: Lexicon-based sentiment analysis (Figure adapted form [74])

The purpose of this study is not to develop a highly accurate sentiment labelling algorithm using NLP techniques, but was to explore the impact of demographic variable on aggregated sentiment in various locations of the cities such as London wards. Therefore, in this thesis, existing state-of the-art simple average sentiment scoring lexicon-based sentiment analysis is employed in Chapter 3 to examine the sentiment of Tweets by utilising the Bing-Lui [69] sentiment lexicon.

2.2.2 Sentiment Classification - Machine Learning Methods

The classification of Tweet text by utilising machine learning method can be divided into two types - *supervised* and *unsupervised* learning methods. Figure. 2.4 provides a concise view of machine learning techniques and various state-of-the-art algorithm available in data mining. The supervised machine learning methods incorporate sentiment labelled tweets text called the *training data* to determine the sentiment classification of unseen sample of tweets . Conversely, for the unsupervised methods the input data examples of text are not class labelled. Typically, unsupervised methods use *clustering* of data to identify the classes within the data [77]. The field of data mining provides many clustering algorithms (such as K-Means, Hidden

Markov Model etc. as shown in Figure 2.4) that can be used to classify data in the context of textual data too. However, before the application of these algorithms to classify the text data, the text needs to be transformed and has to be represented in a binary vector, i.e. looking for a presence or absence of word in a sentence. Such text representation poses a challenge of very large dimensionality, but the underlying data is sparse. This complexity raises issues when such algorithms have to deal with short text messages of the length of 140-270 characters as tweets [78]. Since this study is based on short text tweet data, I ruled out an option to use the unsupervised methods and resorted to a tool that incorporates supervised machine learning sentiment classification for sentiment labelling used in Chapter 4 and Chapter 5.

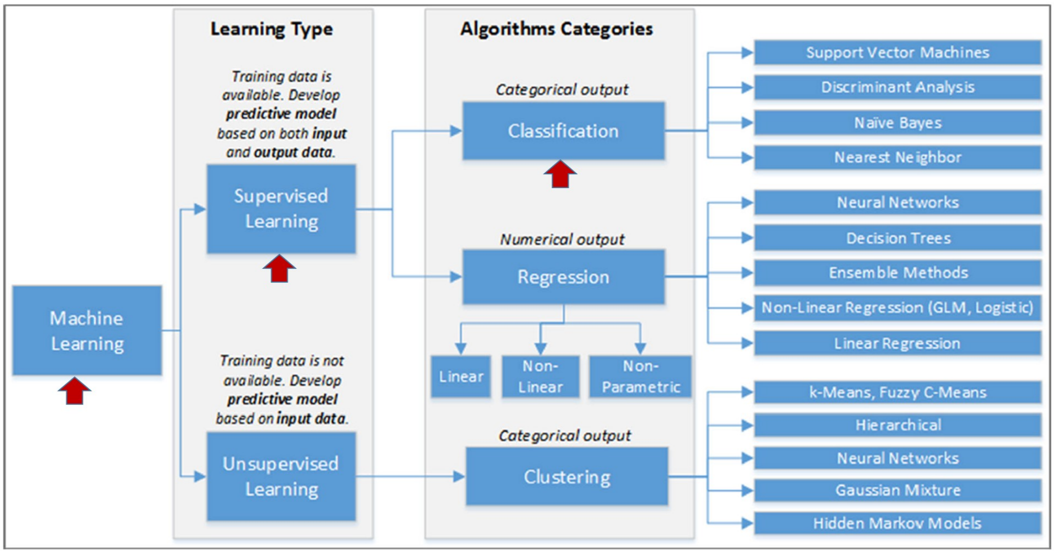


Figure 2.4: Machine Learning Techniques [26]

Naive Bayes, Support Vector Machine (SVM), Logistic Regression and Neural Network are the commonly used supervised ML sentiment analysis methods for texts analysis in NLP domain. In Naive Bayes, every feature which is extracted gets to determine what label should be assigned to an input value by checking the frequency of occurrence and prior probability for each label in training set. This classifier is based upon Bayes Theorem of conditional probability. Support Vector Machine, classifies a text data based on the training samples of positive and negative data given to it. It classifies text by creating a hyper plane of separation with maximum margin to separate the negative and positive data points available to it from the sample training datasets [79]. Logistics regression is a linear regression model using which an outcome label be predicted by using explanatory or independent variables.

Classifictaion Method	**Outlook For Sentiment Analysis**
Nave Bayes	Is advantageous to use as it requires small training data set and has been extensively used in document categorisation, e-mail spam detection and identify sentiment polarity.
Support Vector Machine (SVM)	It is also suited for smaller datasets as the training time using this method can be huge for large datasets.
Logistic Regression	In contrast to the Nave Bayes method, this methods has opposite assumptions about the features sets. In a scenario where a conditional independence assumption of Nave Bayes fails, this method might potentially yield a better result for sentiment analysis task.
Neural Network	Not ideal for short text messages classification task such as tweet and is intrinsically multi-layered process is a black box model.

Table 2.2: Classification Methods for Sentiment Analysis of Tweets

Ultimately, the performance of a classifier depends on how well the text feature engineering is formed for any sentiment labelling task. Table 2.2 highlights the outlook of the above four methods for Twitter sentiment analysis.

This paragraph explains the methodology adopted by the tool for classifying tweet sentiment. Sentiment classifier using supervised machine learning algorithms is built using a training data sample which consist of pre-annotated sentiment label for each text (tweets). Two labelled datasets for training and tests, provided by SemEval2015 (Semantic Evaluation) were used in this analysis, which is an ongoing series of NLP competitions [80]. These two data-sets are labelled tweets with different sentiment polarities - positive, negative and neutral. These datsets are then used to classify the unlabelled tweets of interest in this study for Chapter 4 and Chapter 5. In the next paragraph, I briefly explain the underlying functioning of the NLP tool used for this research that uses machine learning models for sentiment annotation of tweets.

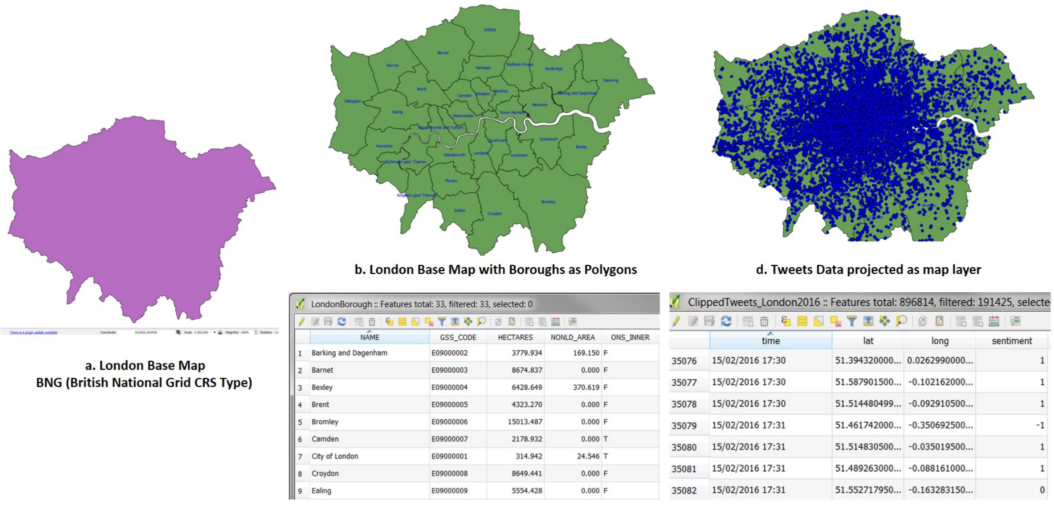
The pre-processing of the tweets was established using *Twokenize* package, which has become better in supporting Emojis and other Twitter specific language eg. #, replies [81]. Further in the process, the text features are derived from a training set following the example of state of the art machine learning approaches demonstrated in earlier research [75]. This feature matrix derived is used to allocate a value for each of the feature found in each tweet, including test dataset and unknown dataset. The features used are *unigrams*, *bigrams*, *Part of Speech(POS) tags*, *word vectors* and *sentiment lexica*. The first three are routine features in sentiment analysis tasks but are extended with a custom negator. *Unigrams* take each word in the training set and indicates its presence of absence in the Tweet. *Bigrams* take pairs of words and indicates its presence or absence in a Tweet. *POS Tags* indicate if the word in a this context is a noun, verb or vdjective to differentiate words that can appear with multiple meaning, which is determined by looking at the neighbouring words to establish that difference. For example, the word 'train' can be a noun or an adjective. For each seen context the absence of presence of each word is added as e.g. Train-noun = 1 and Train-Verb = 0. *Word Vectors* define a numerical vector to each word in their deciding its semantic meaning. The word vectors use average, max, min and count on each dimension of the 100 dimension GloVe dataset trained on two billion tweets [82]. 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 [83]. The four lexica that were consider are; Bing Liu opinion lexicon [84] mentioned before, the MPQA subjectivity lexicon [85], AFINN [86] and SentiWordNet [87], taking various counts and averages. Finally, the feature sets extracted are used to train the sentiment algorithms in NLP - Naive Bayes, SVM and Logistics Regression.

The k-Fold cross validation technique has been suggested to use for the parameter optimisation by earlier researchers [88] before testing the final models on the test set. The optimised model is then used to evaluate the unlabelled geo-tagged tweets available for this research and has been used as a proxy to detect happiness patterns in relation to various demographic variables and at business locations.

2.2.3 Analysing Spatial Patterns of Tweets

Analysis of tweets to detect and explore urban trends can be performed using many different approaches, one of which is to visualise spatial pattern of tweets on a geographical map. Such spatial analysis, relies on GIS systems, which, as defined by United States Geological Survey 2013, *is a computer system capable of assembling, storing, manipulating and displaying geographically referenced information, i.e. data identified according to its location* [89]. GIS system offers a synthesis of computer based cryptography and spatial statistical analysis [90] through which researchers can visualise data in a spatial environment to understand how a location of a place influence certain characteristics or attribute of the data, for example, positive *sentiment characteristic* of text at a particular *location* like park. The patterns exhibited by such characteristics, their relationship with other attributes can be studied in relation to the space using GIS mapping tools to aid decision making process. Many web and desktop based geo-analysis tools have emerged in recent years which visualise the real time data from social media sources on digital maps. CARTO (previously CartoDB) is one such web based cloud computing platform which provides GIS mapping services for location data emerging from tools like Twitter [91]. CartoDB has extended its capability to collect tweet and also provide geo-referencing services (Twitter Maps). However, it works on propriety black box algorithms and one cannot uncover the underlying process it is built on [92]. Similarly, other tools such as CASA's MapTube provides free resources for viewing, sharing, mixing and mashing maps and offers features like joining of data with users own data to build map visualisation form scratch [90] [93].

In contrast to the above mentioned web based spatial analysis tools, desktop GIS systems such as ArcGIS and open source QGIS tend to be more powerful as these do not suffer from performance issues which can pose a challenge when it comes to analysing huge Twitter data-set like ours (.5 million of Tweets), processing of which can take substantial amount of processing time if uploaded on web. Also, Twitter data agreement doesn't allow us to upload the purchased datasets on a cloud based web software. Therefore, this research resorted using open source desktop version of QGIS for spatial analysis.



a. London Base Map
BNG (British National Grid CRS Type)

b. London Base Map with Boroughs as Polygons

d. Tweets Data projected as map layer

LondonBorough :: Features total: 33, filtered: 33, selected: 0

	NAME	GSS_CODE	HECTARES	NONLD_AREA	ONS_INNER
1	Barking and Dagenham	E09000002	3779.934	169.150	F
2	Barnet	E09000003	8674.837	0.000	F
3	Bexley	E09000004	6428.649	370.619	F
4	Brent	E09000005	4323.270	0.000	F
5	Bromley	E09000006	15013.487	0.000	F
6	Camden	E09000007	2178.932	0.000	T
7	City of London	E09000001	314.942	24.546	T
8	Croydon	E09000008	8649.441	0.000	F
9	Ealing	E09000009	5554.428	0.000	F

c. Spatial Data Borough Attributes Table

ClippedTweets_London2016 :: Features total: 896814, filtered: 191425, selecte

	time	lat	long	sentiment
35076	15/02/2016 17:30	51.394320000...	0.0262990000...	1
35077	15/02/2016 17:30	51.587901500...	-0.102162000...	1
35078	15/02/2016 17:30	51.514480499...	-0.092910500...	1
35079	15/02/2016 17:31	51.461742000...	-0.350692500...	-1
35080	15/02/2016 17:31	51.514830500...	-0.035019500...	1
35081	15/02/2016 17:31	51.489263000...	-0.088161000...	1
35082	15/02/2016 17:31	51.552717950...	-0.163283150...	0

e. Imported Tweet sentiment csv file, projected as map layer

Figure 2.5: Spatial Analysis using QGIS

Spatial analysis using QGIS is done by initially downloading the base maps information available from variety of data sources into the desktop QGIS platform. For example, Figure 2.5 (a) shows a base map of London created using British National Grid (BNG) coordinate reference system (CRS) by downloading the statistical London Boundaries data available made available by Greator London Authority on London

Data Store which contains a range of key GIS boundary files for Map Info covering Greater London area [94]. This spatial map can then be overlayed with the Borough level information Figure 2.5(b) and Figure 2.5(c) for further granular analysis. A csv file which contains tweet sentiment information, Figure 2.5 (e) along with geo-coordinates can then be imported in QGIS system as shape file by providing XY coordinates (latitude and longitude) which help a user to see tweet information as point map layer Figure 2.5(d) over the baseline London Map. QGIS software provides number of features to perform spatial analysis of data such as analysing number of *points in a polygon*, generation of *heatmaps*, *clipping* etc. which are extensively used in this thesis and has been respectively detailed in each contribution Chapter 4 and Chapter 5.

2.3 Exploring urban sentiment

The examination of user generated content on Twitter pioneer diverse research that unveil many interesting urban trends and insights about our world - from the likes of helping urban planners plan their city better [95], aid the knowledge discovery processes to plan the strategies for civic participation [96], monitor public mood during specific events [97], predict stock markets [98] and many more. However, given the breadth of topics studied using this data pose a challenge to provide an account of all of them. Therefore in this section, I bound my literature review drawing on studies that predominantly motivated my first research question (RQ1) i.e delineation of studies that accessed the textual feature of the tweets to understand the *sentiment of people as a proxy of happiness* in urban locations and the ones that has explored demographic attributes of a place to establish co-relation with happiness indicators from social media text.

2.3.1 Tweet Sentiment - A Proxy of Happiness

Surveying citizen happiness is an important area of research [99][100][101]. Qualitatively, the pursuit of happiness is a cornerstone philosophies in the governance theory propelled by many ancient cultures. In modern history, quantitative measures such as the Gross National Happiness (GNH) gained traction after 2005. Given the subjective nature of happiness, it is typically measured through self-reported surveys that are validated and normalised against more objective metrics that are widely accepted as ones that support positive sentiment (i.e., income and lifespan). Existing research projects have pursued both qualitative and empirical experiments to understand the sentiment of urban spaces [99]. Indeed, census data is extensively used by governments to create well-being scores (an example can be found for London[1]). But this data from survey based methods are limited in their resolution (spatial-temporal). Alternative sentiment data collection methods employ wearable monitoring systems such as electro-dermal-activity sensors [102]. These systems will yield precise longitudinal data with high spatial-temporal accuracy. However, their expensive nature means that scaling to the general public and establishing pervasive and non-intrusive sensing remains a challenge.

Happiness or Subjective Well Being is accessed by researchers using 'evaluative' or 'hedonic' approach [103]. For 'evaluative' approach a cognitive judgement is made based on the response of individuals or population sample by measuring on a scale how someone feels or thinks about something (by recalling their feelings). For example, how happy you are with your current life?. Whereas, to monitor the happiness of population in 'hedonic' way, researchers measure people experiences and emotions as they are experienced moment by moment. Non- invasive methods such as population-level 'hedonometer' [104] therefore has been used to remotely sense and quantify emotional levels in order to measure happiness. The precision of self-reported methods in qualitative research has been contrasted with sentiment expressed in Facebook status updates and has exhibited a strong relationship [105]. As language is perceived to be never isolated from the world of moods and is braided in the emotional state of a person speaking [64], in socio-technical data mining domain, large-scale texts expressed on social media platforms has been used by researchers to evaluate emotional content of individual words in order to generate an overall score for the text [106] to measure the happiness level of person writing or tweeting the text.

The lexicon based (occurrence of positive and negative words in a sentence) sentiment analysis has been claimed to be high for domain specific sentiment analysis. Earlier researchers have confirmed that higher the normalised sentiment score is for a group of tweets (based on word count), the higher is the socio-economic well-being [14] for a population group. Interestingly, the other strand of NLP scientists discovered that changes in the demographics across urban areas can reflect the changes in happiness metrics as measured by words or language use on social media [107]. Generally speaking, a choice of dictionary used can impact the overall accuracy of the sentiment score. Therefore, NLP researchers endorse the use of several lexicon classifiers in order to predict the sentiment score and to achieve a high accuracy the final predicted sentiment score should be a sum of all the sentiment score calculated before [108].

[1] http://data.london.gov.uk/dataset/london-ward-well-being-scores

Nevertheless, the proliferation of online social interactions such as Twitter, has in recent years provided an opportunity to study sentiment of urban dwellers (residents, workers, tourists...etc.). Detecting sentiment using data from Twitter as a proxy incurs bias, does offer attractive benefits in scalability and there is growing research to validate and benchmark the sentiment labels of citizens in urban spaces. A growing body of research has conducted studies to detect emotions from Twitter data [109] and creating mood heat maps of city locations [110] as well as comparing emotions between cities [100][101].

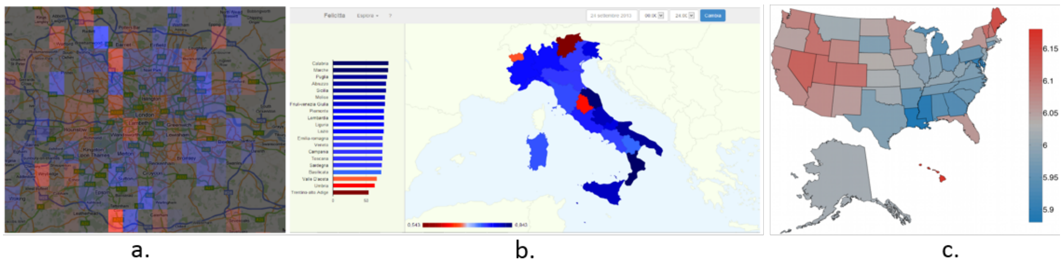


Figure 2.6: Tweet Sentiment to study happiness a) Communities in UK , b) Italian Cities and c) US States happiness scores [14] [111] [100]

Well-being or happiness of communities in Greater London area as been studied using Twitter sentiment and has been reported to correlate strongly with community socio-economic wellbeing as evident from London census data (Figure. 2.6a. highlights the happy and not happy communities in London red and blue) [14]. Similarly, happiness in Italian cities is investigated using geo-tagged tweets by summarising the sentiment scores in various Italian cities and is visualised through a platform called *Felicitta* (Figure. 2.6 b) [111]. Seminal studies by Lewis Mictchell et. al., explore the words usage of tweet content to arrive at happiness scores revealing happiest and unhappiest state in the US. (Figure. 2.6 c) [100], is an example of *choropleth* thematic map in which areas are shaded or patterned in proportion to the measurement of the statistical variable being displayed on the map such as average sentiment score. However, there is limited literature we could find where researchers have attempted to explore the happiness trends in London wards using Twitter data. As such, Chapter 3 of this Thesis goes on to overcome the challenge of understanding happiness indicators in London Wards using Twitter data.

2.3.2 Linking Twitter Happiness to Urban Demographic Features

Understanding of citizen sentiment in context to their urban environment offers urban planners to plan services using economic indicators beyond traditional costbenefit analysis and accurately perceive the consequential effects of those services on citizen sentiment [112]. In general, statistical averaging across large populations and across contexts yield reasonable understanding of an ongoing social phenomenon. For example, H. Andrew Schwartz et. al. [113] studied the correlation of word use in Tweet text with US county level subjective well being and found that the patterns discovered from this analysis supports the traditional happiness literature, such a positive effects on people sentiment when they are engaged in social activities, exercise, engagement at school and work etc. Additionally , the word topics were reported to have a strong correlation with demographic and socio-economic status variables [113]. Happiness indicators from Twitter language was found have strong *Spearman correlation* with wealth and negative with poverty when studied in conjunction with the US 2011 Census data [107] suggesting how Twitter data can be studied in relation to various socio-economics metrics at a population-scale. Not surprisingly Twitter and other big data sources dawns a new understanding about the socio-economic parameters prevailing in our environment [114].

The association between demographics variables of a place and area-level Twitter characteristics is studied using *linear regression models* uncovering how neighbourhoods with social and economic disadvantage, and with the availability of more fast food restaurants may exhibit lower happiness and fewer healthy behaviours [115]. Along the lines of the aforementioned research where researchers have studied the relationship between the sentiment and underlying demographic features of a place, the first experiment for this research entailed in Chapter 3 attempts to map and examine the relationship between large-scale Twitter sentiment data and urban geography features in London Wards. Different from previous work, we also determine the statistical significance of this relationship of socio-economic variables in London wards on Sentiment of Twitter using *coefficient of determination (R^2)*. This facilitates to an understanding about how some of the selected urban features of a place can potentially influence happiness in London wards.

2.4 Twitter Sentiment for Industrial Landscape

As the literature in Section 2.2.1 and 2.2.2 points to, the extraction and selection of textual and spatial features from Twitter enable us to glean a real time sentiment trend of users at a given location. Of which, understanding sentiment as a proxy of happiness has attracted a great attention by social media researchers. Though efforts to mine the sentiment trends are valuable contributions to the field of sentiment and big data analysis, very little is known about the Twitter sentiment in relation to the working hours prevailing in various type of industrial sectors, which is the second big research question (RQ2) we try to find an answer to in Chapter 4. Prior to this, we explored the studies in the field of Twitter sentiment analysis to know earlier endeavours to segment out the Twitter data, i.e *The usage of Twitter across* various population groups and how previous researchers have *perceived the happiness indicators for the group of people working in various industrial sectors* to inform the current research work. The below two sub-sections provides a brief these studies.

2.4.1 Dynamics of Twitter Usage - Who Contributes to Twitter

High spatial and temporal Twitter data captures the geographic information and posting time of the user in the *latitude-longitude* and *time* fields of a Tweet respectively, offering researchers a capability to understand the activity patterns and changing social dynamics. Earlier studies adopted a methodology to aggregate the tweet count for a certain short time of the day, by leveraging the time information available in tweets, revealing an indication about an urban land use profile to expose the dynamics of residences and businesses in an area [116] thereby segregating the Tweets usage at these two places. Juan Carlos et. al. aggregated tweets depending on the user ID for a specific spatial and temporal unit yielding activity profiles (Figure.2.7 a.) for each land use [117]. Together, such studies unearth the areas where twitter activity is more dominant in a city, for example offices, education and transport that can help in urban planning and forecasting business demands. In contrast, some researchers use a combination of geo-tagged Twitter and Wikipedia data and incorporate spatial clustering methods to perceive collective sense of a place [118]. A seminal study in this area segmented the geo-referenced tweets based on the topic of discussion using NLP methods to demonstrate the variability of Twitter content in accordance to the characteristics of a place and its users (Figure.2.7 b.) [119].

The aforementioned research endeavours concentrate on using Twitter data exclusively to discern the Twitter usage patterns in context to a place. However, the possibility of combining Twitter data with other sources of information to establish its key contributors is still at nascent stages. Only in some of the recent studies, researchers have recently linked geo-tagged activity of Twitter to other social networking site (LinkedIn) data in order to isolate user's home location and work location by streaming FriendFeed data [120]. One can also profitably combine diverse data types based on location information by fusing the geographical information about a place (Google Maps, Open Street Maps, Location specific spatial polygons etc.) to geo-tagged Twitter data to enhance our understanding about a social

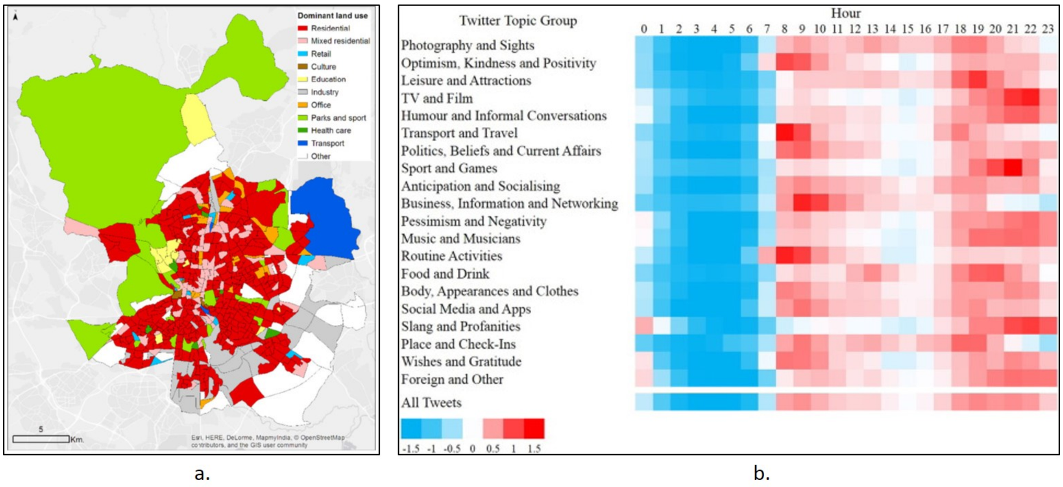


Figure 2.7: Twitter Usage and variability of its content a) Madrid land uses according to Tweets , b) Temporal frequency of Tweet topics [117][119]

phenomenon. For instance, business locations has been profiled by matching the geo-tagged tweets from Twitter against the venues from Foursquare enabling the study of mood heatmaps within various business regions [121]. The forecasting and surveillance of Influenza disease has been researched using the combination of data from multiple sources: Google searches, Twitter microblogs, nearly real-time hospital visit records, and data from a participatory surveillance system [122] and micro-blogging text messages has been examined in conjunction with sensor and hydrological authoritative data to establish the geographical relationship between flood phenomena and tweeting pattern to enhance the monitoring process during disaster events [123].

Understanding overall trends and usage patterns of micro-blogging has been deemed important to create recommendations and identify the social media targeted audiences [124]. It is important to be aware of the demographics of Twitters population in order to address the findings of the investigation with scrutiny, and as a basis for suggesting commonalities to the study's results. In practice, we need to augment the social media data with other data sources to conduct more high resolution studies to explore who uses such communication platforms. One approach to de-biasing 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 [125]. Likewise, prior work have attempted to infer demographic characteristics such as age, occupation and social class of Twitter users using profile description of Twitter users [21, 22] and gleaned insights about the race, place and gender of Twitter users by studying the intensity of the tweets [23].

Researchers have also quantified and analyzed the divergence between the demographics of the promoters of trends available from Twitter data and the sites overall population with a view to help government agencies understand different facets of public opinion [126]. The ethnic diversity of the population of three big cities, London, Paris and New York reveals that users of the Twitter are primarily male and are of Anglo-Saxon (English speaking) extraction [127]. In contrast, the spread of geo-located tweets in Greater London area is discovered to be uneven and appearing in accumulations posing a challenge to directly explain socio-demographic variables based on tweet counts in Londons wards [128].

Collectively, although these studies outline a critical role to understand the usage of Twitter and who contributes to Twitter trend, there has been no detail investigation about the *industrial level participation of Twitter* in the UK which is the subject of Chapter 4 of this Thesis. This exploration shall contribute to the ongoing research on understanding the usage of Twitter, but with a focus on industrial level participation (as identified using UK SIC code in this Thesis).

2.4.2 Perceiving Happiness Indicators in Industry

The literature on happiness research suggests that measurement of happiness varies across disciplines and one of its prominent factor being employment type and working conditions [129]. Of these work related determinants that contribute to overall life happiness are *working hours* and flexible working conditions [130]. On a global scale, the quantitative measurement of these parameters and their impact on happiness is conducted by organisations such as the Organisation for Economic Cooperation and Development(OECD), the United Nations Development Program, and the Gall Up program [131]. In Europe particularly, the European Agency for Safety and Health at Work, measures stress at work in a variety of industry sectors through the Enterprise survey (ESENER); the Office of National Statistics(ONS) in the UK measures subjective well-being of the the labour force on annual basis [132] [133]. However, the limited scale and temporal resolution of surveys perceiving the the happiness level of workforce in different industrial sectors on a more frequent basis remains a challenge. In UK, and Europe currently there exist no measure to monitor the sentiment of workforce engaged in various industrial sectors (identified by SIC Codes) on an ongoing and more frequent basic. Earlier studies has confirmed that language used in tweets from 1,300 different US counties was found to be predictive of the subjective well-being of people living in those counties as measured by representative surveys [134]. Also, research by Eszter Bokanyi et al.[135] illustrated that by decomposing aggregated activity time lines available from Twitter, it is possible to reveal US employment and unemployment statistics. They provide evidence that digital traces on Twitter can be explored to study employment trends.

The work conducted in Chapter 4 takes inspiration from these prior studies and propose a simplistic yet novel method that link the multitude of open data sources to identify tweets that belong to various industrial sectors in the UK. Earlier studies has estimated the location and user profile of tweets using computational methods resulting in demographic proxies [21]. However, the methodology devised resorts to fusion of open data sources, which existed in isolation so far, i.e.; the spatial property maps (INSPIRE Polygon) of business locations, business ownership data, which are the data sources available from the UK Land Registry, and uses SQL joins and GIS processing of geo-tagged tweets to create industry level summaries of tweets. I then study tweeting intensities at different time period to understand the prevailing working hours and what is the impact of working hours on aggregated tweets sentiment. By knowing a sentiment polarity of Tweets of people engaged in various industrial sectors, this experiment tries to overcome the challenge of monitoring people sentiment to assess the prevailing mood or happiness in industries in the UK as an alternative to the survey methods. Specifically, the Chapter 4 of Thesis would explore: *to what extent the sentiment analysis technique of Twitter can be incorporated to understand happiness factor in industries using London as case study.*

2.5 Twitter Sentiment during Political Events

As stated above in literature review, social scientists have expanded their exploration to use social media sources like Twitter to perceive an impact of a political events on society. Twitter data has been used to predict election results [136] [15], uncover political preferences of people [137], identify fluctuations in public sentiment prompted by political events such as United States presidential debate [136] and to track trade volume from Twitter subsequent to the broadcast of periodic business performance in terms of earnings and related parameters [138]. This section of literature provides a background of studies conducted using Twitter data , specifically during and after the Brexit vote. A brief literature review of studies that compare social media analytic to traditional survey is conducted to inform this research, since previous studies has argued the challenges and opportunities of both approaches in mining citizen opinion.

2.5.1 Twitter Sentiment during Brexit

European Union(EU) referendum in the United Kingdom, popularly called Brexit (Britain Exit) was held on 23 June 2016 to seek a mandate from British people regarding their wish to remain in the EU or withdraw from it its affiliation. Majority of Britishers opted for the withdrawal of EU membership. Social media Twitter space was flooded with public opinion tweets from across the globe regarding Brexit. These communications have been used to derive models of various study coordinates populism, nationalism, economism, globalism , and were found to express national sentiment and economic outlook of the people to a significant degree [139].

Differences of campaign methods of political parties and their impact has also been studied by using Twitter data [140]. Remain/Leave proportions of Brexit tweets have been categorized in relation to the variables of Age, Location and Education, giving an indication of how significant these parameters were in shaping the opinion pattern [141]. Brexit tweets have also been used to monitor variations in public emotions positive, negative, anger, anxiety, sadness prior to and subsequent to EU referendum [97]. Plotting geospatial distribution of Brexit tweets from their spatio-temporal-textual (STT) patterns, keywords and hashtags conveying positive, negative or neutral sentiments were discerned [142]. Another multidimensional study on EU referendum focused on the evolution of Brexit discussion in pre and post vote tweets, frequency of reference to major political players, variability of sentiment about the event within and without (especially Europe) U.K., and the significance of sentiments regarding the refugee crises [143].

Primary objective of above mentioned studies was to exploit either or combination of spatial, temporal or textual features of the Brexit tweets to uncover social phenomena at the political level or to fathom the spatial signatures they left on Britain and around the world. In the current study the endeavour is to explore a novel question: *can industrial sentiment about a political event like Brexit be obtained from social media sources such as Twitter* by analysing this organically available data focused around the period of the event, viz., one week before and one week after the EU referendum. A preliminary experiment to explore the possibility of such analysis is outlined in Chapter 5 of this thesis.

2.5.2 Comparing Surveys and Social Media Analytics

Surveys have been carried out in research sphere from decades by designing research specific question and soliciting the answers to those from a population sample. Traditional survey, was designated as 'designed data' by Groves [144] since the data collection process was principally planned keeping the definitive view of the research questions [90] and offers a mechanism to execute an experiment in a controlled manner [145]. However, passive collection of data from online social media sources such as Twitter offers an alternative tool to find answer to some research questions at a scale and speed that can complement long spaced survey methods. As Groves explains this 'organically' available data is capable of tracing transaction of all sorts in our society [144]. Sentiment analysis computational methods (as outlined in Section 2.2.1 and 2.2.2) utilises language analysis algorithm to create insights from public Tweets in various situation and contexts.

Nevertheless, there are outstanding and unsolved questions about the representativeness of Twitter data to determine the demographic profiles of the users of such platforms. Joe Murphy in the book 'Innovation in Digital Research Methods' argues that 'new digital technologies may not ultimately replace traditional approaches, it is important to continue evaluating the potential of new technologies and social media tool for their role in social research' [90]. Jasper et.al. advocates a combination strategy to use new data (such as social media Twitter data) and traditional survey data in synergy so that one may support research, analytics, and decision making [145]. Similarly, other researchers too recommends and illustrate by various business examples where the fusion of survey data with new form of data sources resulted in enhance analysis of a research question [146].

The last research question explored in this Thesis is to know the usage of Twitter in order to understand the public sentiment in Industrial landscape towards a political event like Brexit vote. Nonetheless, there remain an uncertainly about the participation level of each industry type in Twitter conversation. In such a scenario, if exists, a social scientist would still need to adopt to survey methods to understand people sentiment for certain industry types. Therefore, Chapter 5 of this Thesis evaluates the pros and cons of Twitter analytic and survey methods for mining Industrial sentiment and when it needs to resorts to combinatorial approach.

Chapter 3

Understanding Happiness in Cities using Twitter: Jobs, Children, and Transport

In this chapter, a Lexicon-based sentiment analysis techniques is applied to 0.4 million geo-tagged Tweets in the Greater London area in order to understand the underlying socioeconomic and urban geography parameters that influence happiness. Primarily, the experiment conducted in current chapter answers my research question *RQ1. Which socioeconomic variable correlate strongly with the positive sentiments of aggregated tweet text and what is the statistical measure of influence of those variables on the tweet sentiment?* The work disseminated in this chapter has been published in [1].

3.1 Introduction

The demographics and landscape of cities is changing rapidly, and there is an emphasis to better understand the factors which influence citizen happiness. As mentioned in the Section 2.3.1, inferring happiness (sentiment analysis) from social media data is a scalable solution to monitor people happiness in cities.

3.2 Methods

3.2.1 The Data

The data used in this chapter comes from two sources: (1) 0.4 million geo-tagged social media data purchased from Twitter, covering a 2 week period (see Fig. 3.1**a**),

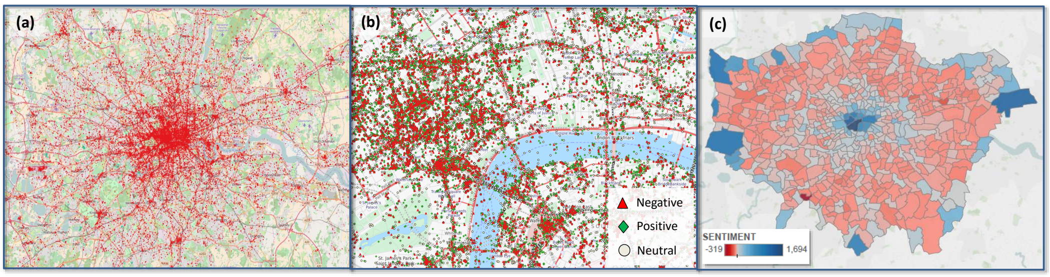


Figure 3.1: Mapping the Sentiment in London:(a) 0.4 million geo-tagged Tweets in Greater London over a 2-weeks period. (b) Tweets labelled as negative (red triangle), positive (green diamond), or neutral (pale circle) on a scale of 11. (c) Ward level sentiment where dark red indicates negative sentiment and dark blue indicates positive sentiment.

and (2) UK government ward-level socioeconomic and urban geographical data (open access) from the London Data Store[1].

0.4 million of Twitter data used for experimentation in this chapter range from 10th June 2012 to 24th June 2012. Fig. 3.2 shows the temporal distribution of tweets in the sample time window chosen for analysis. Previous research has claimed that Twitter tends to receive more tweets in summer than in winter, i.e. 60% of all tweets are received in summer and around 40% in winters [147]. The current dataset is based on summer time thereby giving the maximum benefit in terms of counts of tweets to analyse the co-relation of tweets sentiment with socioeconomic and urban geographic variables in the London ward area. However, the sentiment of the tweet could be subject to multitude of competing effects some of which are seasonal and others could be driven by external events. This study aims to demonstrate a case using a sample of Twitter data that cues from social media sources could be potentially exploited to understand the impact of London wards urban features on mood of people. Although it is legible to mention that such trends could change over time which is a limitation of using limited sample of social media data.

In terms of spatial resolution, the analysis in this chapter will focus on Greater London, which is made up of 628 wards, and are roughly analogous to a neighbourhood. Many services are delegated to the ward level, including policing; and a range of census statistics are available at the ward level. The ward level census data considers 64 key metrics, including demographics, education, housing, and business statistics.

This chapter's focus on using Twitter data (aggregated from all urban dwellers) as a proxy and comparing it to census data (mainly registered residential and business

[1] http://data.london.gov.uk/dataset/ward-profiles-and-atlas

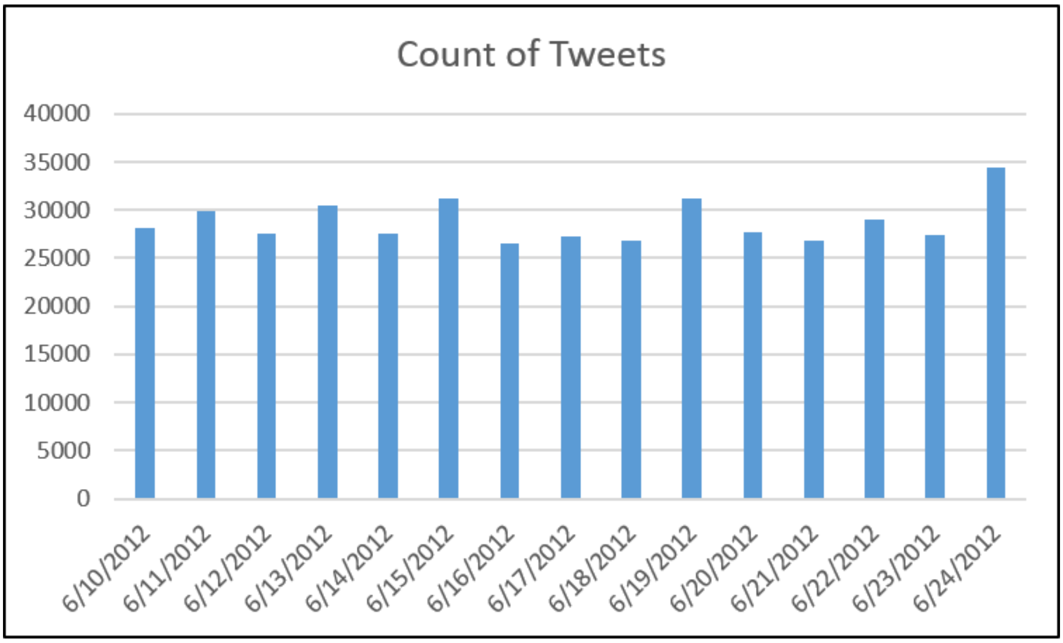


Figure 3.2: Distribution of 2012 Twitter data

data) means that we are concerned with how all people in London (including residents, workers, tourists) feel as a function of the urban geography and its socioeconomic parameters. It is extremely challenging to understand what distribution of the social media data belongs to which demographic, and in this chapter we treat all data as equally important (uniform weighting) and do not consider demographic categories within the sentiment data.

3.2.2 Sentiment Labelling using Lexicons

The primary aim of this research was to classify sentiment labels independent of a specific domain to infer polarity of tweets in a city and correlate this to existing citizen demographic data to explore life satisfaction patterns in London. Research to understand sentiment orientation can rely on two methods: machine learning and lexicon based. The polarity classification of a sentence using machine learning approach employs algorithm like Support Vector Machine, Maximum Entropy or Logistic Regression, Naive Bayes etc. These models are trained using multiple feature (like unigram, bi-gram, POS-Part of speech tags) and accounts for high accuracy for a specific domain but its performance is decreased significantly when these models are used to predict sentiment on different domain [148]. However advantage of using the second dictionary based approach is one can easily find large number of sentiment words with their semantic orientation. Since the lexicon based methods rely on linguistic knowledge they have been proved to be more robust if the sentiment classification task is independent of a domain [149]. Bing and Liu sentiment lexicon which is created using small set of sentiment word (seeds) with positive and negative orientation and then construct this set algorithmically through a bootstrapping process by searching WorldNet (http://wordnet.princeton.edu/) for their synonym and antonyms. Since lexicon method tend to be more robust across domains and text it was considered viable to choose this approach over the complex state of the art machine learning methods.

Given the aim of this study is not explore about peoples perception relevant to a particular topic or domain but is to get a general mood of the urban population based on the words they use within tweet sentences it was viable to use this robust method. The Bing Liu opinion lexicon dictionary of words is obtained through a bootstrapping process using WordNet, which offers semantic relations among words. It is built utilizing the adjectives synonym set and antonym set available in WordNet to predict the semantic orientations of text used in tweet sentence. This method enables us to find the average semantic orientation of tweets in simplistic ways of independent of the context of the text under analysis. This technique was successfully implemented in previous research to analyse sentiment [150], but has not been applied to urban contexts to understand the underlying sources of happiness.

To assign each tweet with a sentiment score we first apply *Tokenization* filtering to remove language noise and transform all text to a common lower case format with no punctuations. We then extract single words or features (unigrams) independently to determine the orientation of the tweet. Researchers in opinion mining have focused on trying to find suitable lexicon for classifying tweets sentiments by annotating tweets for negative or positive polarity (henceforth *happiness*) by recognising words as positive and negative sentiment. We apply the opinion lexicon [69] (full list is approximately 6800 words to each tweet. Our algorithm calculates the score of each tweet by simply subtracting the number of occurrences of negative words from the number of positive occurrences for each tweet. An example of the sentiment labelled Tweets is shown in Fig. 3.1**b**, and clustered to ward level in Fig. 3.1**c**. An interesting trend can be observed: that the happy wards (blue) are either in the centre or on the outer edges of Greater London, and the unhappy wards (red) are in the middle.

3.2.3 Metrics for Comparison

In order to conduct cross-dataset comparisons, the coefficient of determination, denoted R^2 is a number that indicates how well the statistical regression model fits the data or in other words: the percentage of variance in the data that can be explained by the proposed model. For a data vector $y = [y_1, y_2, ...y_K]$ (with mean $\overline{y}$) and a predicted data vector using the regression model $\hat{y}$, the residue vector is defined as $e = y - \hat{y}$. The coefficient of determination R^2 is defined as:

$$R^2 \equiv 1 - \frac{\sum_k e_k^2}{\sum_k (y_k - \overline{y})^2}, \tag{3.1}$$

where the numerator is the residual sum of squares and the denominator is the total sum of squares. In this chapter, we use the adjusted $R^2 = 1 - (1 - R^2)\frac{K-1}{K-P-1}$ to take discount against extra variables P in the model.

The adjusted R^2 is a modified version of R^2that has been adjusted for the number of predictors in the model. The adjusted R^2 value increases only if the new variable or term improves the model more than it would be expected by chance.

3.3 Results

3.3.1 Baseline Sentiment Data

We first present baseline sentiment data results, to gain a better understanding of the sentiment data of individual people and their tweets. It is of interest to understand the relationship between the number of tweets and aggregate sentiment of tweets. The results in Fig. 3.3 shows that people who tweet more also express stronger aggregate sentiments (absolute value: either positive or negative), but on average express a lower sentiment per tweet.

The chapter will now focus on 3 key areas that were identified through a correlation panel analysis (see Fig. 3.7): (1) Employment Opportunities, (2) Children and Fertility Rate, and (3) Accessibility to Public Transport. In particular, these are areas which affect urban lives on a daily/monthly basis and as such have a direct impact on the sentiment (see Table **??**). It is worth mentioning that for the results to be presented below, given the census data lists over 60 urban geography features that can potentially affect happiness, obtaining a coefficient of determination for a single feature that accounts for 33 to 47% of the variations in sentiment is a *significant result.*

3.3.2 Employment Opportunities

The two main attributes in employment opportunity measured by the census data are: (i) Number of jobs in a ward (data from businesses) and (ii) Number of jobs

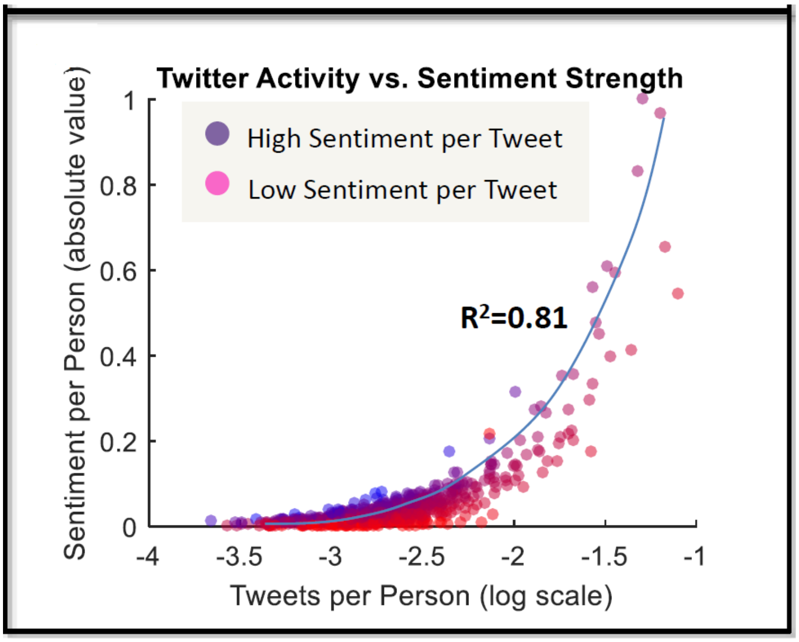


Figure 3.3: Sentiment Data Analysis: People who tweet more also express stronger aggregate sentiments, but on average express a lower sentiment per tweet.

normalised against the number of people in the working age (16-64) in a ward. Both sets of employment data are highly positively correlated with each other, as well with other crime and ambulance incident data (see Fig. 3.7). This reinforces the notion that increased opportunities often lead to an increase in the challenges [151]. In terms of how employment relates to online sentiment, Fig. 3.4**a** shows the number of jobs available in a ward is positively correlated with the sentiment in the ward (adjusted $R^2 = 0.45$). Similarly, Fig. 3.4**b** shows that the number of jobs normalised against working population is positively correlated with the sentiment in the ward (adjusted $R^2 = 0.47$). The adjusted $R^2 = 0.45 - 0.47$ indicates that the regressions (which both use quadratic functions, $P = 2$) explains for almost 50% of the variance in sentiment variations, and the remaining variations are due to other factors. In other words, this shows that the availability of jobs determines a significant 50% of the expressed sentiment. Yet, the sentiment is correlated with the number of jobs available and not with the number of employed people (see Fig. 3.7).

3.3.3 Number of Children

The main attributes in measuring the distribution of children in census data is the number and percentage of children (aged 0-15) in a ward. This percentage is negatively correlated with sentiment, as well with other data such as the general

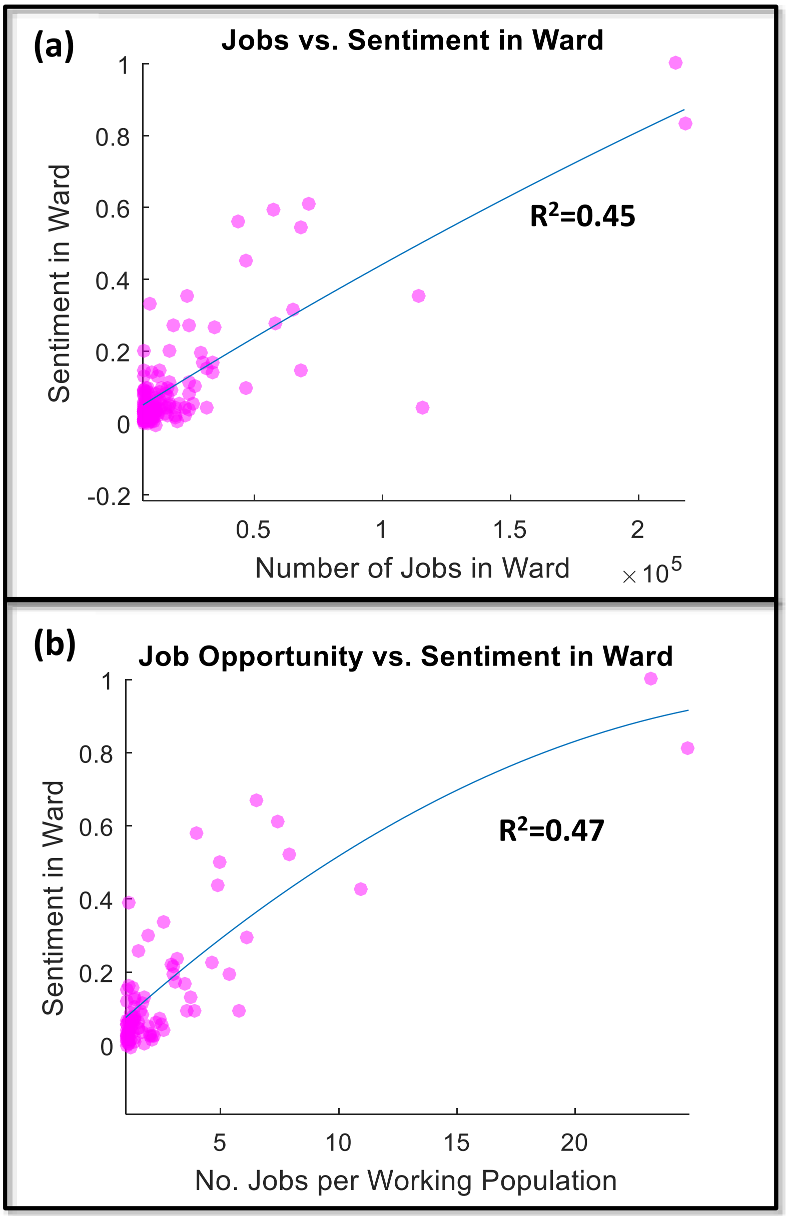


Figure 3.4: Relating Average Sentiment per Person to Jobs Opportunities in London: (a) The number of jobs available in a ward is positively correlated with the sentiment in the ward (adjusted $R^2 = 0.45$). (b) The number of jobs opportunities (jobs normalised against working population) in a ward is positively correlated with the sentiment in the ward (adjusted $R^2 = 0.47$).

fertility rate (see Fig. 3.7). Fig. 3.5**a** shows the percentage of population that are children in a ward is negatively correlated with the sentiment in the ward (adjusted $R^2 = 0.33$). This shows that the percentage of children determines a significant 33%

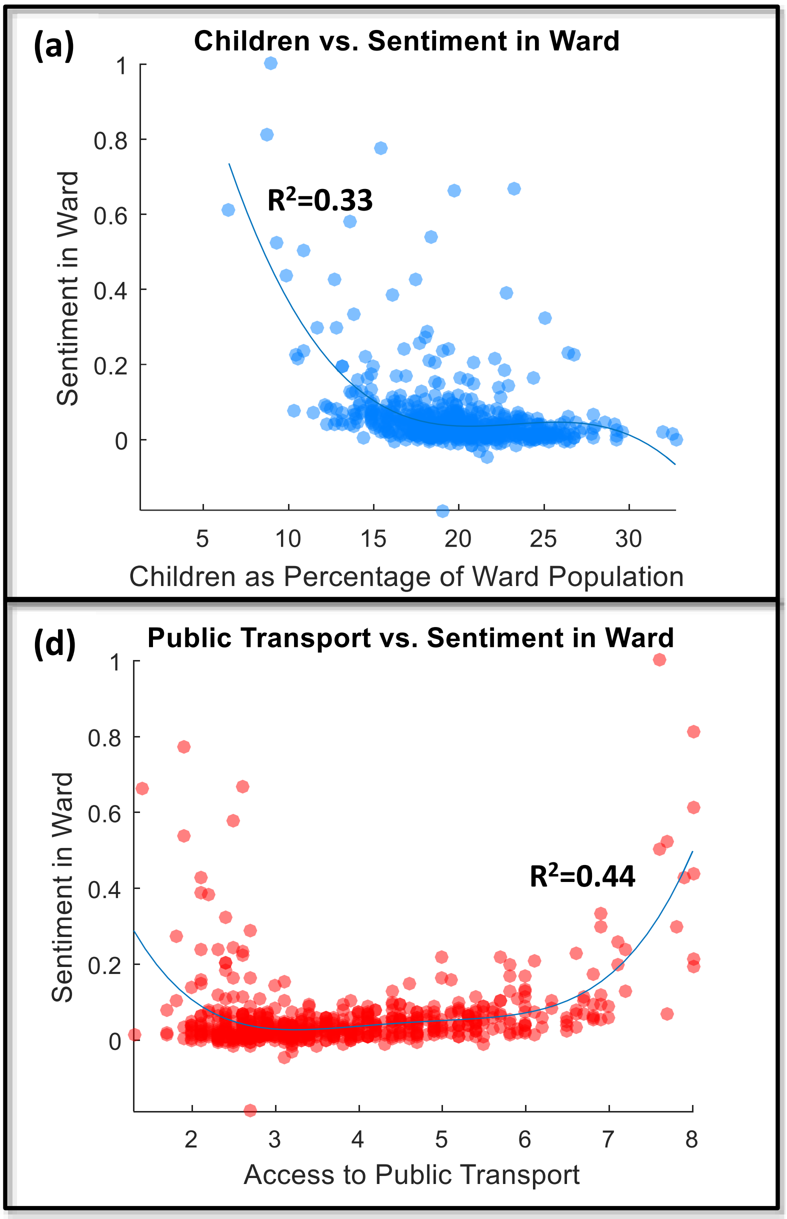


Figure 3.5: Relating Avg. Sentiment per Person to Number of Children and Access to Public Transport in London: (a) The percentage of population that are children in a ward is negatively correlated with the sentiment in the ward (adjusted $R^2 = 0.33$). (b) The accessibility to public transport in a ward has a parabolic relationship with the sentiment in the ward (adjusted $R^2 = 0.44$), such that those with good access to public transport are happy and those who are in areas with poor public transport are also happy (rely on personal transport), whilst those that are in between are generally less happy.

of the expressed sentiment. More specifically, it shows that there is a steep decline in sentiment from 5% to 15%, and the relationship saturates thereafter. It is worth noting that the percentage of children does not correlate with other socioeconomic factors such as the deprivation level in the ward, but is negatively correlated with the employment level in the ward. Without inferring causality, the data supports our previous finding that increased job availability leads to higher sentiment and a decrease in the percentage of children. We suspect that the wider applicability of this result will depend on the family cultural context.

3.3.4 Accessibility to Public Transport

The main attributes in measuring public transport availability in census data is the Public Transport Accessibility Levels (PTAL). It is a detailed and accurate measure of the accessibility of a point to the public transport network, taking into account walk access time and service availability. The method is essentially a way of measuring the density of the public transport network at any location within Greater London. The measure reflects 4 main attributes: (1) walking time to transport access point, (2) reliability of services, (3) number of services, and (4) the average waiting time. It does not consider the speed or utility of the service, crowding effects, and ease or efficiency of interchange. The PTAL methodology was developed for London where a dense integrated public transport network means that nearly all destinations can be reached within a reasonable amount of time. Research using the ATOS (Access to Opportunities and Services) methodology shows that there is a strong correlation between PTALs and the time taken to reach key services i.e., high PTAL areas generally have good access to services and low PTAL areas have poor access to services. Each area is graded between 0 and 6b, where a score of 0 is very poor access to public transport, and 6b is excellent access to public transport.

Fig. 3.5(b) The accessibility to public transport in a ward has a U-shaped (parabolic) relationship with the sentiment in the ward (adjusted $R^2 = 0.44$, $P = 4$), such that those with good access to public transport are happy and those who are in areas with poor public transport are also happy (possibly because they rely on personal transportation means), whilst those that are in-between are generally unhappy. Certainly the results in Fig. 3.6 seem to strongly support this hypothesis. The PTAL values explain for 71% of the variance in the number of private vehicles per household, showing that those with poor public transport access own up to four times more private vehicles per household. Therefore, the availability of public transport explains 44% of the variance in sentiment scores. The wider applicability of this result beyond London is difficult to determine. Yet, we speculate that economies

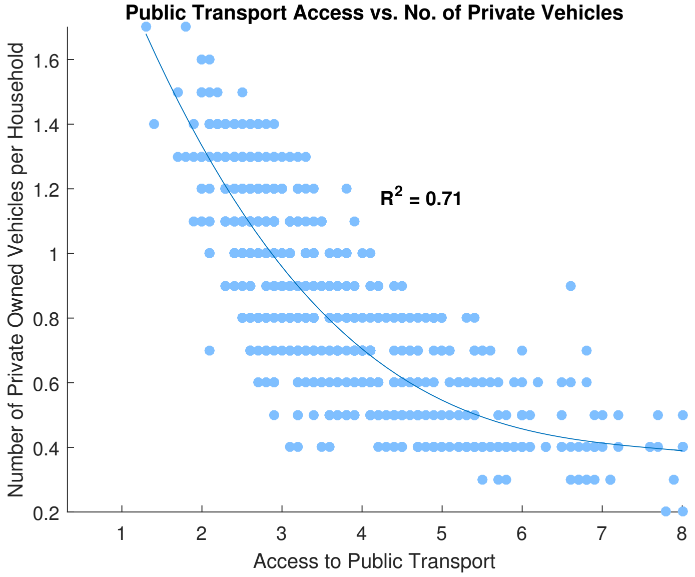


Figure 3.6: Public Transport Access vs. Number of Private Vehicles: Those with poor public transport access levels (PTALs) own up to 4x more private vehicles per household, and the PTALs explains 71% of the variance in car ownership numbers.

with a high number of privately owned vehicles will exhibit similar patterns, i.e., people are happy when they are either close to public transport or far removed, and struggle when they are in-between the choices.

3.3.5 Linear Regression of Sentiment vs. Ward Level Parameters

A linear regression of sentiment vs. ward level socioeconomic and infrastructure metrics is shown in Fig. 3.7. The linear regression does not uncover more complex parabolic relationships such as those found between sentiment and accessibility to public transportation. Nonetheless it serves as an overview of the first order relationship between all 67 parameters.

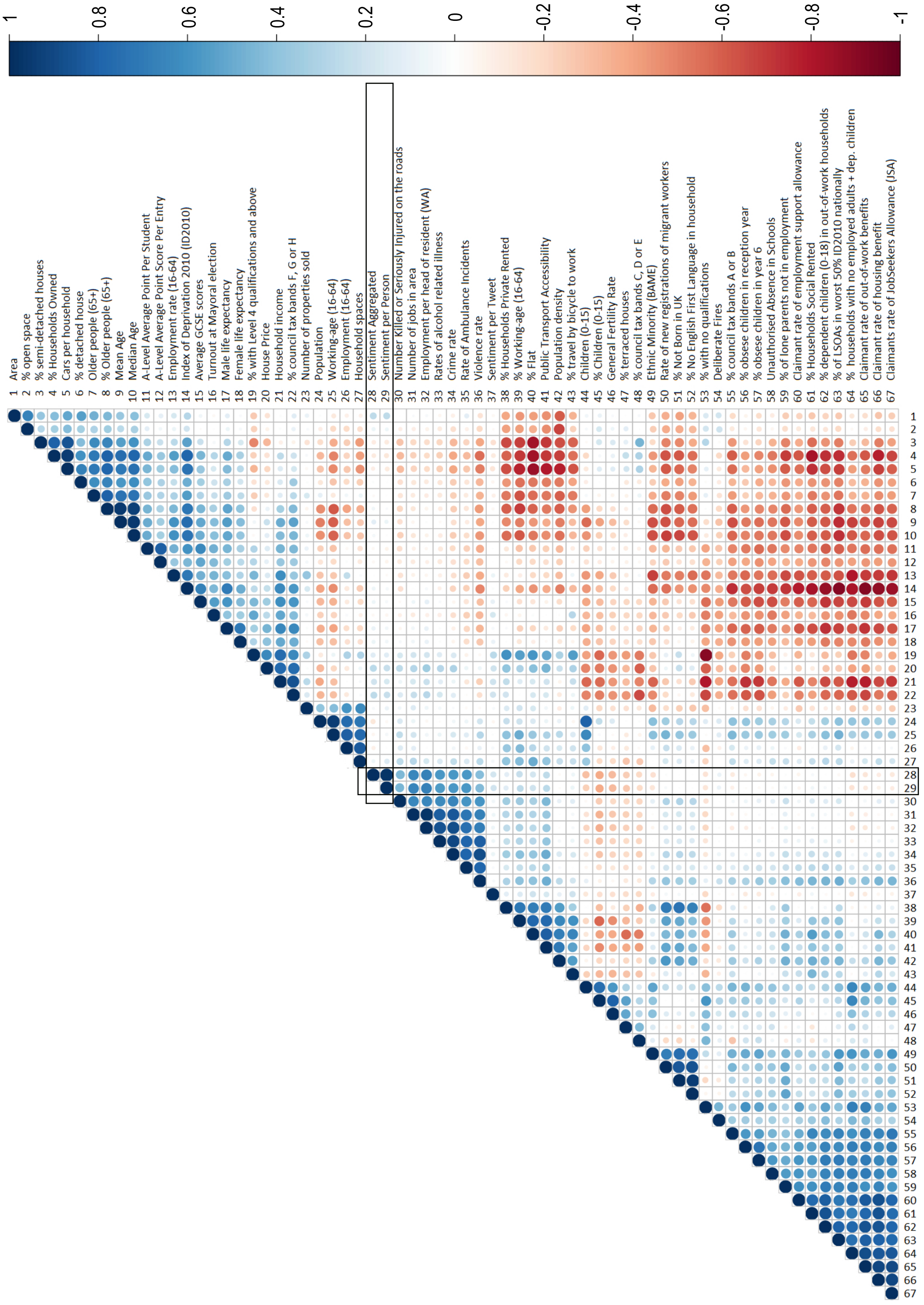


Figure 3.7: Linear Regression Matrix of Sentiment vs. Ward Level Socioeconomic and Infrastructure Metrics. Sentiment correlations are boxed.

3.4 Discussion

The demographics and landscape of cities are changing rapidly, and there is an emphasis to better understand the factors which influence citizen happiness in order to design smart urban systems. In this chapter, we apply natural language processing to 0.4 million geo-tagged tweets in the Greater London area to understand the underlying socioeconomic and urban geography parameters that influence happiness. Our results not only verify established thinking: that job opportunities explain 45-47% of the sentiment variations, but also reveal two additional insights: (1) happiness is negatively correlated with the number of children (accounts for 33% of sentiment variations) and (2) happiness derives from good access to public transport or high car ownership (44% of variations). The number of jobs and children, as well as accessibility to public transport are every day facets of urban living and individually explain up to 47% of the variations in happiness. Our results show that they influence happiness more significantly than more ambient parameters such as degradation and education quality.

Future work would need to examine the co-linearity between the independent variables studied (number of jobs, children and PTAL score) as this was not investigated in the current analysis. By definition, collinearity affect the coefficients of regression and p-values, but has limited influence on the predictions. Collinearity doesn't affect how well the model fits and if the model predicts the R-squared value satisfactorily, one can use high co-linear features or variables to make good predictions, which is the case in the current experiment. However, if the experiment is focused on using the coefficients of predictors variables, collinearity could make some variables statistically insignificant though they should be treated as significant. Therefore, measures such as VIF (Variance Inflation Factors) could be used to detect the multicollinearity in the data and the variables under study.

Additionally, as seen in the regression matrix in Fig. 3.7, 'crime rate' shows a high correlation with the sentiment score but was not a chosen variable to study in the current analysis. The research contemplated to explore the ward features that can be either managed by the government (e.g. access to public transport - PTAL, jobs) or in control of the population themselves (number of children). As such these features can be influenced with policy intervention or counselling respectively. Features, such as 'crime rate' could be of volatile nature and could occasionally be difficult to manage by government authorities. Hence, in urban growth context the research systematically plan to study aforementioned three variables only.

3.5 Conclusion

The experiment results of this chapter not only verify established thinking: that job opportunities correlate strongly with positive sentiments; but also reveal two additional insights: (1) happiness has a negative relationship with the number of children and (2) happiness has a parabolic relationship with access to public transportation. The latter implies that those that rely on, but do not have strong access to public transport are the least happy. Separately, the number of jobs, number of children, and accessibility to public transport explain 33-47% of the variations in sentiment. The results of this study will enable urban planners to move beyond planning services using traditional cost benefit analyses, and enable them to consider the consequences on citizens happiness.

The general study of how sentiment is linked to urban features and socioeconomic parameters is useful for urban planners and urban system designers. The results will allow decision makers to move beyond planning services using traditional costbenefit analyses, and enable them to consider the consequences on citizens' happiness. Further research on understanding how these patterns change with different cities and cultures is of interest, as well as how more reliable methods of labelling sentiment to social media data can be applied.

Chapter 4

Exploring Working Hours and Sentiment in Industry Using Twitter and UK Property Data

In the earlier chapter a high level view of sentiment was generated as a proxy of happiness in London Wards. The current chapter drills this exploration further to specifically look for happiness indicators in *industrial sectors* using Twitter and commercial properties spatial data. Particularly, this experiment shall enable to answer my second research question *RQ2. Can social media data available from Twitter particularly, be used to understand happiness indicators in various industrial sectors?* In order to arrive to the answer of the stated question , it would be useful to split this question in two parts, first: *RQ2.1. how social media usage and general sentiment vary across the geographic and industry sector landscape,* which shall help to identify who are the main contributors to Twitter conversation in industry identifiable using SIC Categories code, and second: *RQ2.2. What might be the impact of working hours (estimated from Tweet intensities) on aggregated Twitter sentiment within each industrial sectors?*

The machine learning tool implemented in this chapter to arrive at the Tweet sentiment labels was developed in collaboration with a peer researcher. There were two choices to perform sentiment analysis of Tweet text, one to use off the shelf black box machine learning tools or alternatively to write a code in Python, as Python provides powerful libraries for text sentiment analysis and the underlying approach adopted is white box and transparent. Since most of the research in this Thesis is conducted using R scripting, writing a code in Python required Python scripting expertise. Therefore, a collaboration with peer CDT researcher enabled me to incorporate a tool written in Python to arrive at Tweet sentiment labels. The first part of the work disseminated in this chapter has been published in [2]. Additionally, at the time of writing this Thesis, the second part of the work conducted is under review with PLOS ONE Journal.

4.1 Introduction

Studies outlined in the literature review Section 2.4.1 highlights the importance of investigating group differences of Twitter users, very little research that I'm aware of has investigated which business groups or sectors contribute to Twitter conversations and which industry sector the Twitter users belong to. Such isolation of Tweets in industrial sector shall enable to study the behaviour or sentiment of people in industrial landscape. The sentiment of people working in various industries can be studied in relation to various social and economics contexts, of which *working hours* if one such aspect which is understood to play a crucial impact on the happiness of people.

According to literature on happiness, the measurement of happiness varies across disciplines, with one of its major determinants being employment and working conditions [129]. These studies indicated that one of the prominent work related factors which contribute to overall life happiness are *working hours* and flexible working conditions [130]. Such economic research on working hours has shown that the potential adverse effects of variable and longer working hours on happiness, is offset by better income and job satisfaction. As discussed in Section 2.4.2, on global scale such studies to date have been conducted using data available from traditional sources such as surveys. Whilst surveys and self reported methods provide precise longitudinal data, they lack scalability and real-time monitoring. However, given the limited scale and temporal resolution of surveys, there are questions that can not be addressed: What is the happiness level of employees working in different sectors at different hours in the day? Does happiness relate to work start time and duration? Social media and workplace spatial data provides an opportunity to analyse work-related temporal dynamics throughout the day and its relationship to sentiment expressed through, for example, tweets.

As a result of these limitations, this Chapter therefore is set out to fill this knowledge gap by proposing a new method for incorporating secondary sources of data, both real-time and static, to evaluate the working hours and sentiment of employees across several industrial sectors. Research by Eszter Bokanyi et al.[152] illustrates that by aggregated activity time lines available from Twitter enables to understand people behaviour to study employment trends. To the best of my knowledge there has been no detailed investigation to utilize geo-tagged Twitter data to study working hours in relation to sentiment in various industrial sectors. Consequently, the methodology detailed in this Chapter provides a description of the process of linking geo-tagged Twitter data to different industrial sectors by using UK Land Registry and Ordnance Survey data-sets, to first understand the usage of Twitter across industrial sectors and then assess the impact of working hours (as calculated from twitter time intensities) on sentiment of people within each industrial sectors.

4.1.1 Describing the Study Hypothesis

To conduct the experiment, Greater London was used as a test-bed, and an integration of sampled geotagged Twitter data with large-scale data on commercial property ownership was performed. The aspiration was to derive industrial-sector-level indicators of sentiment and, based on the duration of work and established literature, the experiment hypothesize H_0 that working 'times' do not have any effect on tweet 'sentiment'. By combining Twitter and other administrative data this experiment envisage to design an alternative mechanism to survey methods, through a new data-driven, scalable methodology that has application beyond the analysis of mood prevalence in various economic sectors, and can be used to extend more traditional survey and sample based methods.

4.2 Methods

Since the experiment in this Chapter relies on large pool of diverse datasets, the Section 4.2.1 describes all the commercial and open data sources utilised for experimentation and the sources these datasets are acquired from. In the rest of the subsections a detailed description of the methodological steps taken to conduct sentiment analysis of Tweets , linking Tweets to various industrial sectors and assessing the impact of working hours on Tweet sentiment is explained.

4.2.1 Datasets Description

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 [153]. Additionally, the data set classifies whether an address of a property is representative of "residential", "commercial" or "land". From this dataset commercial properties in London were extracted.

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 [154]. 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 [155, 156]. As an example, Fig.4.1, displays an INSPIRE Index Polygon layer overlaying Google Street Map show the property area covered by Hammersmith and Fulham Council property.

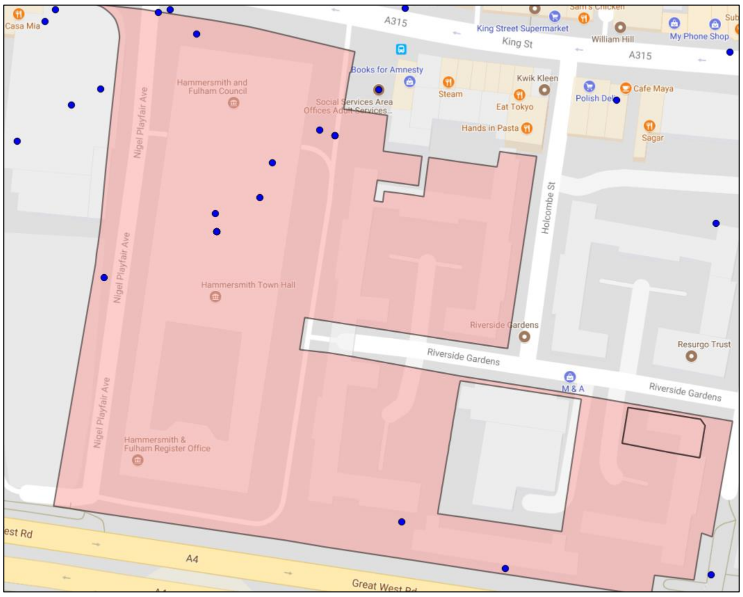


Figure 4.1: INSPIRE Polygon in pink - Blue dots as Tweeting activity

Commercial and corporate ownership data This data is collected by the Land Registry department in the UK as part of the land registration process [157]. 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-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 joined with the INSPIRE Polygon dataset using *title number*, all the registered property parcels (spatial polygon map) in UK 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.4.3 - 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 [158] 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. This research had access to a large commercially purchased data-set - 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 - 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 [159] which contains the information about the company (or business) number and type of economic activity the company is involved in. This data is used to categorise and aggregate tweets in different SIC Codes.

4.2.2 Creating Spatial Property Map for London

The HM Land Registry provides INSPIRE polygons for the whole of UK. Greater London shape file , an open source data which can be downloaded from London data store was used to filter the INSPIRE polygons for the London area only which can be seen in the Fig.4.2.

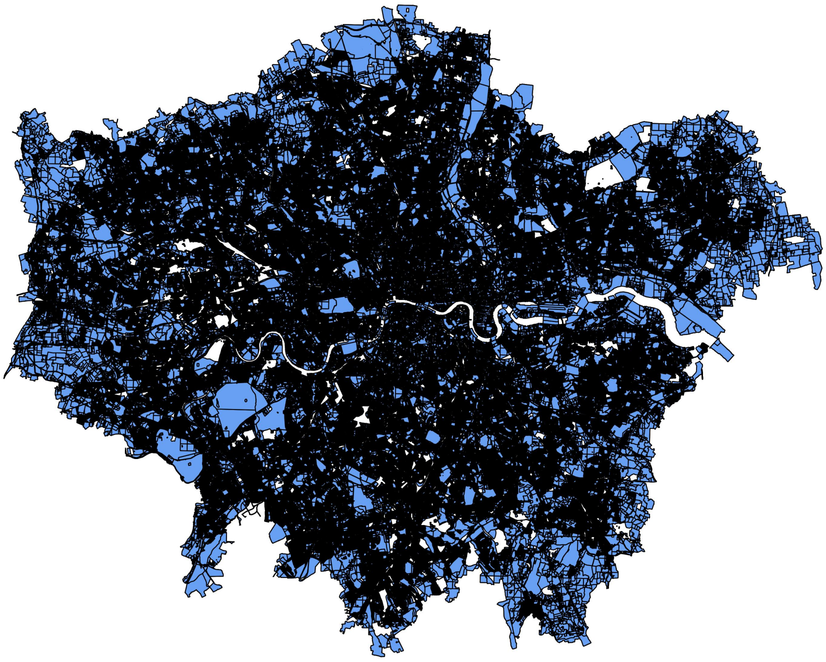


Figure 4.2: INSPIRE Polygon Map for London Properties

4.2.3 Extracting Commercial (only) Properties Spatial Layer

A key stage in this study was to create a commercial property polygon layer to isolate tweets which overlay **commercial property spaces** (referred to as *commercial polygons* in this Chapter) for Greater London. Fig.4.3 shows this stage consisting of two processes - *1) The Data Layer* joins diverse datasets, to create a link between the commercial property textual and spatial information, to tweets originating from the premises. *2) The Spatial Layer* joins commercial property information and tweets which overlay these premise spaces in London. The extraction of all the *commercial* INSPIRE polygons was done based on their address classification (*commercial*, *residential* or *land*) available in the Ordnance Survey dataset (1.1 in the Data Layer in Fig.4.3) to produce a spatial, *commercial-only* polygon layer (2.2 in Fig.4.3). The data layer of the INSPIRE Polygon (1.2 in Fig.4.3) was then joined with the Land Registry Commercial Ownership data (1.3 in Fig.4.3) using a unique field called *UPRN* (Unique Property Reference Number) common to both proprietary databases. The ownership details (which included the SIC Code information about the company) were taken from the Land Registry dataset, based on the *title number* associated with each commercial property. Finally, the SIC Code of the commercial properties was matched to the higher-level SIC Code type (Sections in [160]) in order to analyse the findings with respect to various industrial SIC code types (eg. Construction, Manufacturing, Real Estate etc.).

To explain it further, The 'Ordnance Survey' dataset had the following features of interests:

1. **Latitude-Longitude**, which are co-ordinates in *points* for each property record in the Ordanance Survey Data
2. **UPRN**, which a Unique Property Refrence Number for each property type
3. **Full Address Of the Property**
4. **Property Classification** (commercial, residential etc.)

'INSPIRE Polygon' is the polygon data-set for the UK contains the spatial maps for all the properties in the UK. However, since this research target is to only look at commercial places or industrial locations , limited number of commercial only property records need to be extracted from this dataset. The following feature in INSPIRE Polygon dataset shall help a spatial mapping between INSPIRE Polygons and Ordanance Survey data:

1. **INSPIRE ID**
2. **Property Title Number**

A spatial join between the 'Ordnance Survey' data and 'INSPIRE ID' data is conducted using QGIS 'point in polygon' operation which resulted in the *commercial only property polygons* layer labelled as 2.2 in Fig.4.3. The attribute layer of this spatial joins results in all the above enumerated features of the data in one table which can then be exported to SQL database for data layer join.

This dataset is then exported to SQL database where this table is joined to 'Commercial Property Ownership' data, available from UK Land Registry, using the *Property Title Number* field. 'Commercial Property Ownership' file contains following features of interest:

1. **Property Title Number**
2. **SIC Code**, which define the type of business type, i.e. the type of work umbrella this property is registered into for example, Financial or Manufacturing.

After the above join is accomplished, the commercial spatial property maps are now linked to the *SIC Code* entity, which describe the business or industry type of the commercial property. All the further analysis and aggregation of tweets are done using this SIC code feature.

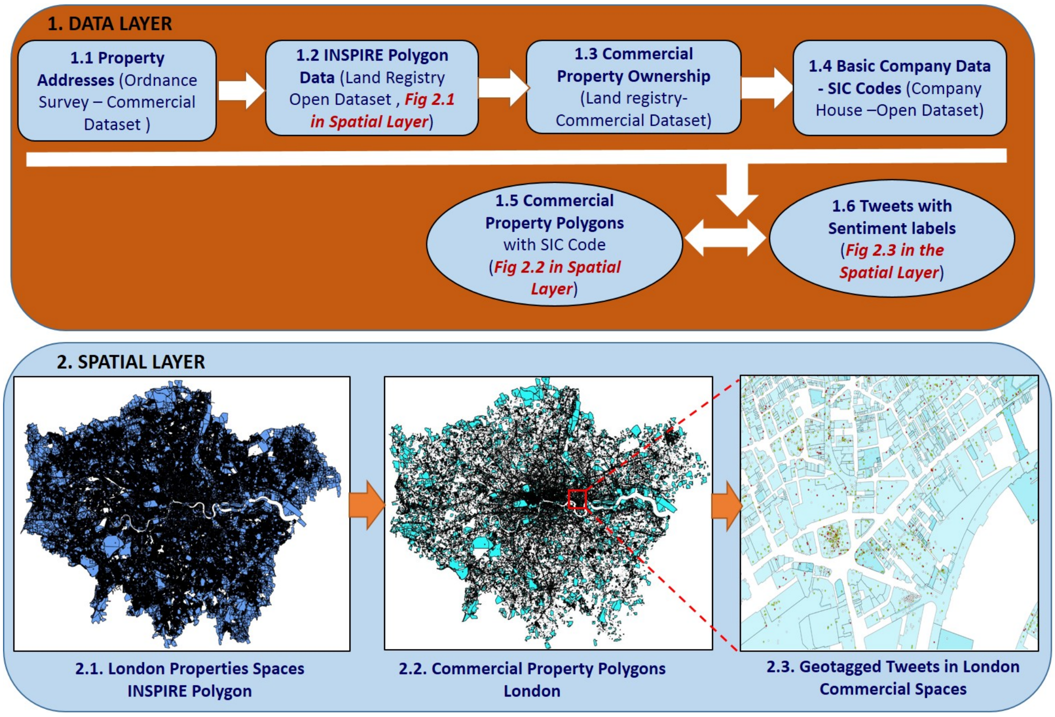

1. DATA LAYER
1.1 Property Addresses (Ordnance Survey – Commercial Dataset)
1.2 INSPIRE Polygon Data (Land Registry Open Dataset , Fig 2.1 in Spatial Layer)
1.3 Commercial Property Ownership (Land registry- Commercial Dataset)
1.4 Basic Company Data - SIC Codes (Company House –Open Dataset)
1.5 Commercial Property Polygons with SIC Code (Fig 2.2 in Spatial Layer)
1.6 Tweets with Sentiment labels (Fig 2.3 in the Spatial Layer)
2. SPATIAL LAYER
2.1. London Properties Spaces INSPIRE Polygon
2.2. Commercial Property Polygons London
2.3. Geotagged Tweets in London Commercial Spaces

Figure 4.3: Methodology (1) Link diverse sources of data-sets to create a commercial properties polygon layer for Greater London. (2) Conduct a spatial join between tweets and commercial property polygons

4.2.4 Sentiment Analysis using Machine Learning

Despite being character limited, the textual component of a tweet has been shown to be a rich source of data - for example as a proxy for well-being [14, 113, 161]. In this chapter sentiment analysis based on machine learning was employed. The two labelled datasets used for the training and test were sourced from SemEval2015, an ongoing series of NLP competitions [80]. These two data-sets constitute labelled tweets with different sentiment polarities - positive, negative and neutral. This model is then used to classify unlabelled tweets of interest in this study.

The tweets were pre-processed using the established *Twokenize* package [81], from which features are derived using state-of-the-art machine learning approaches, see [75]. 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, which itself has been trained on 2 billion tweets [82]. Sentiment lexica, features based on the objectivity or subjectivity of matched words present in the lexica, are chosen based on the research in [83]. The four lexica that were considered are; Bing Liu opinion lexicon [84], the MPQA subjectivity lexicon [85], AFINN [86] and SentiWordNet [87], 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 [162]. k-Fold cross validation [88] was used during parameter optimisation 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 final optimised model was used to predict the sentiment labels for the available London tweets. The sentiments labels (positive, neutral and negative) for two weeks of Twitter data were then aggregated based on the industry sector the tweets were linked to.

4.2.5 Spatially Filter Sentiment Labelled Tweets

The objective of this step is to create an association between tweets and the industry sector they belong to. After the tweets were classified with one of three sentiment labels (positive, neutral and negative), a *clipping* operation was performed using the QGIS software. This involved the spatial overlay of all tweets over the London Commercial polygons and then extract only those tweets that falls withing the commercial polygons boundaries (Fig.4.4). The assumption made is that a tweets location (latitude and longitude), which overlays the commercial property polygon, is linked to the industry classification (SIC) of that commercial property. Although, this approach shall encounter some ‘noise when filtering, and this limitation is talked about in more detail in the discussion section of this chapter. That said, industry sectors which primarily cater to visitors (retail and transport industries, for example), were excluded from the analysis. A *spatial join* is performed between the clipped layer of tweets and the commercial property polygons layer, to establish the SIC Code Type where the tweeting activity happened; note that this happens in both space and time for each tweet. The data layer is then exported and tweets are aggregated into SIC Code types for further analysis.

4.2.6 Extracting Temporal Variables from Tweet Timestamps

From the exploration of data, it was observed that the tweeting patterns within each industry sector (SIC Code type) varies over a 24 hour period. In order to understand

Data Cleaning Function	**2012 Tweets**	**2016 Tweets**
Total number of geotagged tweets available for study	432068	1017496
Tweets overlaying with commercial property polygons	137881	468733
Tweets with SIC(Standard Industrial Classification) information field not NULL	48607	51778

Table 4.1: Numbered of tweets filtered for analysis

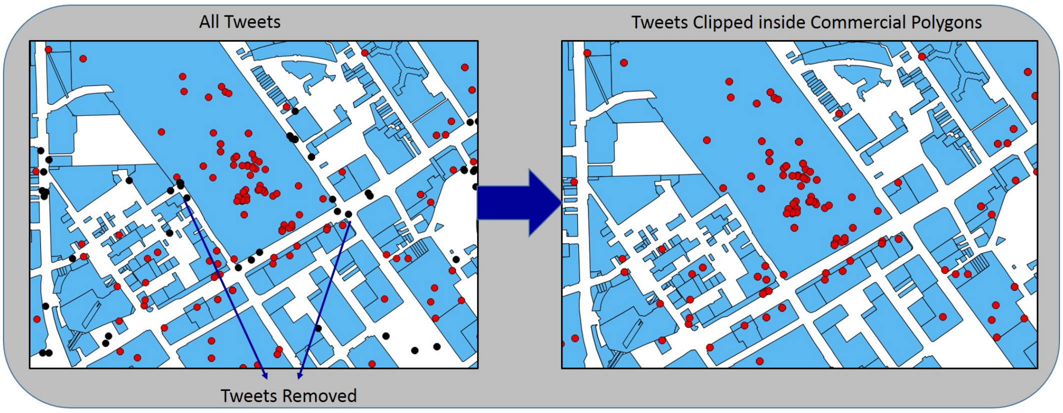


Figure 4.4: Clipping Tweets in Commercial Property Polygons

Parameters	Usage	Definition
$C_1, C_2 ... C_{24}$	Interim	C_i is the aggregated tweet count for each hour of the day
$S_1, S_2 ... S_{24}$	Interim	S_i is the aggregated tweet sentiment for each hour of the day
T_{sum}	Interim	Number of tweets between the hours specified
T_{max}	Interim	Maximum (peak) number of tweets during any hour of the day
T_{thresh}	Interim	Threshold value of tweets, set to 10% of T_{max} in the study
T_s	Regression	StartTime - Hour of day when $C_i >= T_{thresh}$
T_e	Interim	EndTime - Hour of day when $T_{thresh} < C_i$
D_n	Regression	Duration - Time duration in hours between T_s and T_e
SD_n	Regression	Sentiment Duration - Aggregated sentiment between T_s and T_e for duration (D_n)

Table 4.2: Parameters extracted from a tweets time stamp for each industry SIC code type

the relationship between operational hours at a location, and study its relation to the prevailing sentiment, we divided the aggregate tweets into hourly windows over a 24-hour period. This is a recognised approach to studying temporal trends available from social media data, whereby a sample of tweets at different time periods can be used to monitor changes in activity over time [163]. Three temporal parameters were extracted - *start-time*(T_s), *end-time*(T_e) and *duration*(D_n) from the time field of the aggregated tweets for each industry type. Definition of these parameters and others used in the study is shown in Table 4.2.

As shown in Fig 4.5, this experiment established the *threshold* value (T_{thresh}) of tweets as 10% of the maximum (peak) count T_{max} of the tweets in a 24-hour period. The point in time when the tweeting activity spikes above or drops below this threshold value (T_{thresh}) was recorded in the variable start-time (eg. in Fig 4.5 shown as T_{s1}, T_{s2} ...T_{sn}) and end-time (eg. in Fig 4.5 shown as T_{e1}, T_{e2} ...T_{en}) respectively. The number of hours between T_s and T_e is stored in a parameter called *duration* (eg. in Fig 4.5 shown as D_1, D_2 ... D_n). These temporal values were attained and stored in variables after grouping the tweets with their associated SIC code type i.e., a SIC-code-type-level summaries were created for the evident temporal patterns according to the Twitter time stamp. Since the aim this study is to measure the impact of duration (D_n) and start time (T_s) on the sentiment for duration (SD_n) prevailing within each SIC Code type, duration (D_n) and start time (T_s) were used as the *explanatory variable* and aggregated sentiment score for duration (SD_n) as a *response* variable to perform a multivariate linear regression.

A linear regression was then performed to examine the associations between sentiment score (SD_n) as derived from Twitter text and duration (SD_n) and start-time (T_{sn}) variable characteristics for all the industry categories identified by the SIC code. All statistics were calculated using the R package.

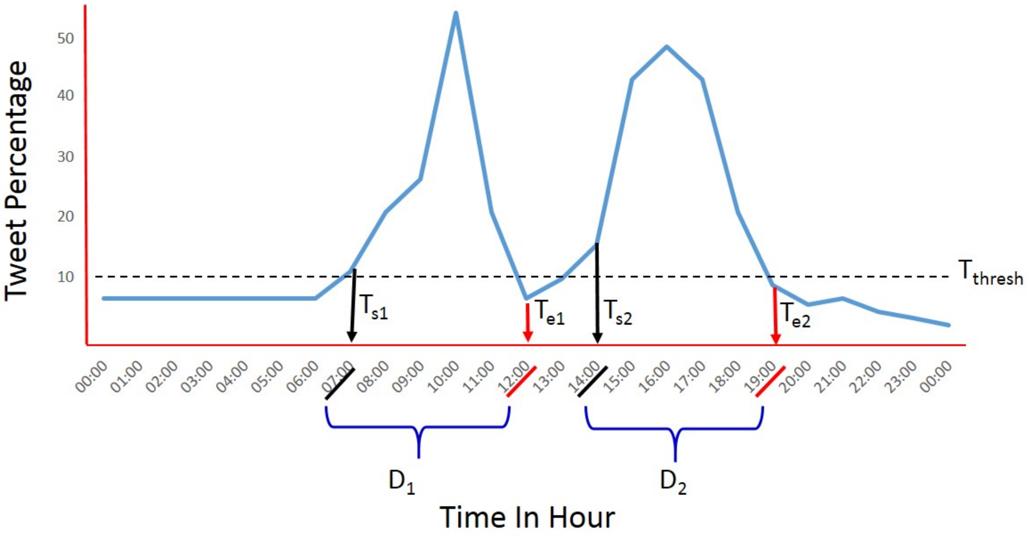


Figure 4.5: Extracting regression parameters from aggregated tweet timestamps for each SIC code type

To summarise, the method outline above uses a combination of INSPIRE map data, Properties SQL dartasets and Twitter data. In order to perform sentiment analysis and further regression analysis on the tweets temporal and sentiment feature this research incorporates spatial mapping techniques that clips the publically available or purchased sets of geo-tagged tweets to commercial property maps or the industrial areas in urban spaces. The commercial property used for this clipping purpose is generated by performing a spatial and sql joins been INSPIRE Polygons, Ordnance Survey and Property ownership datasets available from UK government.

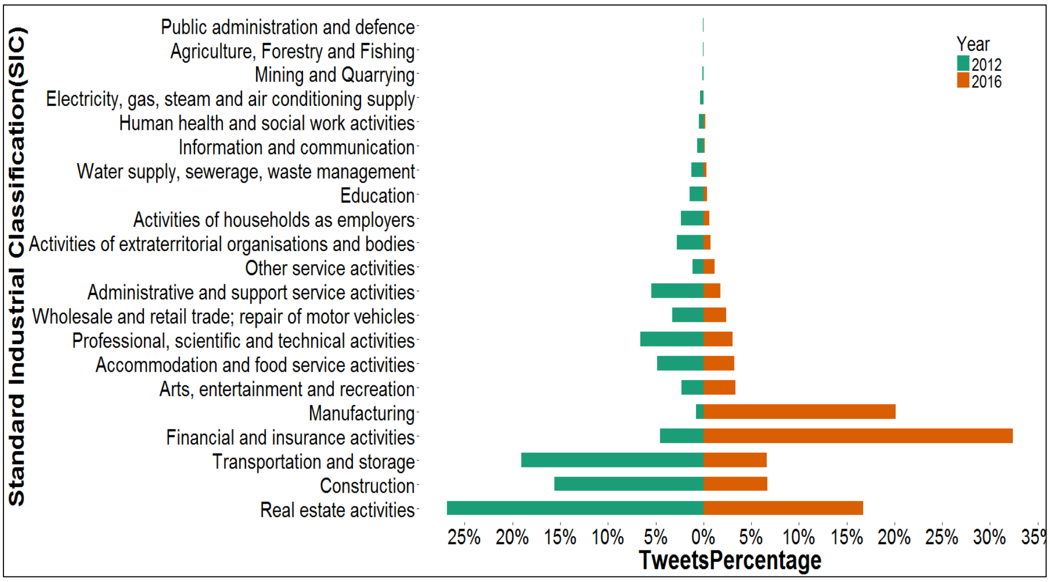


Figure 4.6: Two weeks Tweets Volume Comparison

4.3 Results

The goal in this research was to execute a data driven experiment to *integrate* geo-tagged tweets with diverse administrative datasets to analyse the impact of working hours within industries on the aggregated tweet sentiment. The final sample of tweets used (after executing the above filtering steps) 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 the findings are mentioned below.

4.3.1 Tweets Volume Comparison - 2012 and 2016 Data

The tweeting activity in each SIC Code for the two weeks of data was compared i.e 2012 and 2016 as shown in Fig 4.6. There is an observation of 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 [164–166]. 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. A possible reason 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

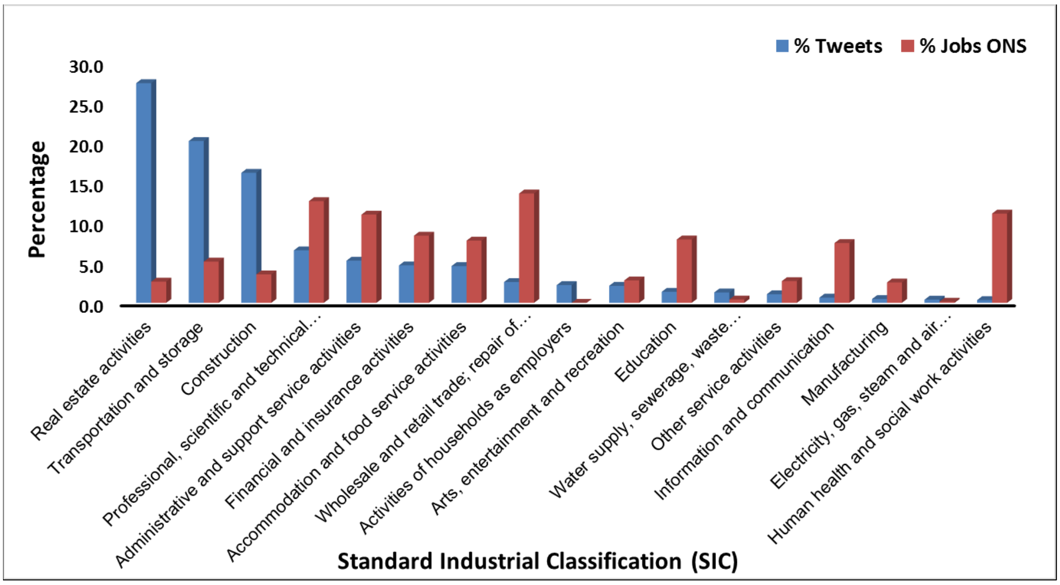


Figure 4.7: Tweets Percentage compared to percentage of Jobs in each sector - 2012

the Twitter data used for this analysis comes from the London city region, one 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 these four sectors were removed from any further analysis.

4.3.2 Tweets Volume Comparison to Proportion of People Working in the Business Sector

To study the penetration of social media in different sectors, the percentage of tweet volumes with the number of jobs available in each sector were compared as shown in Fig.4.7. To make this comparison an open source jobs sector data available from the Office of National Statistics [167] was downloaded for the London region. There was an interesting trend discovered 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. Further research should look into isolating the tweeting activity of visitors and commuters in these sectors and such implications are discussed in the discussion section of this chapter.

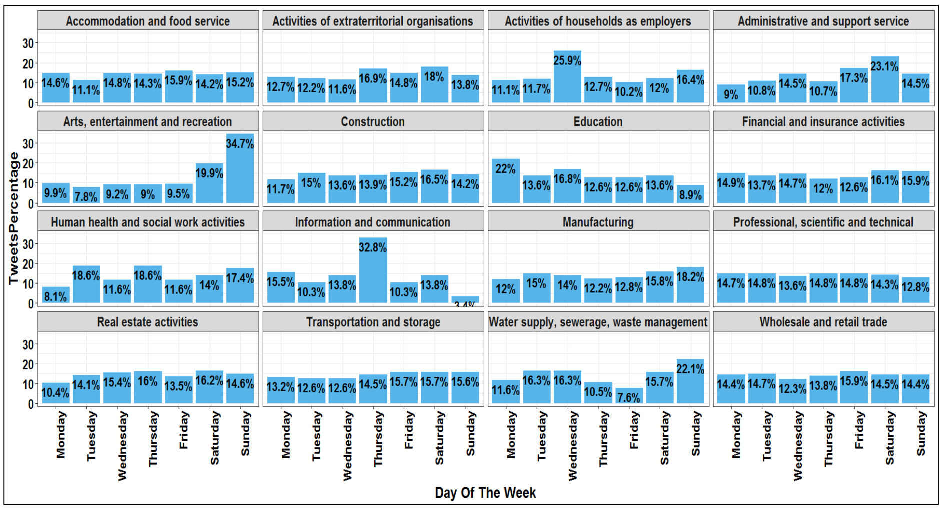


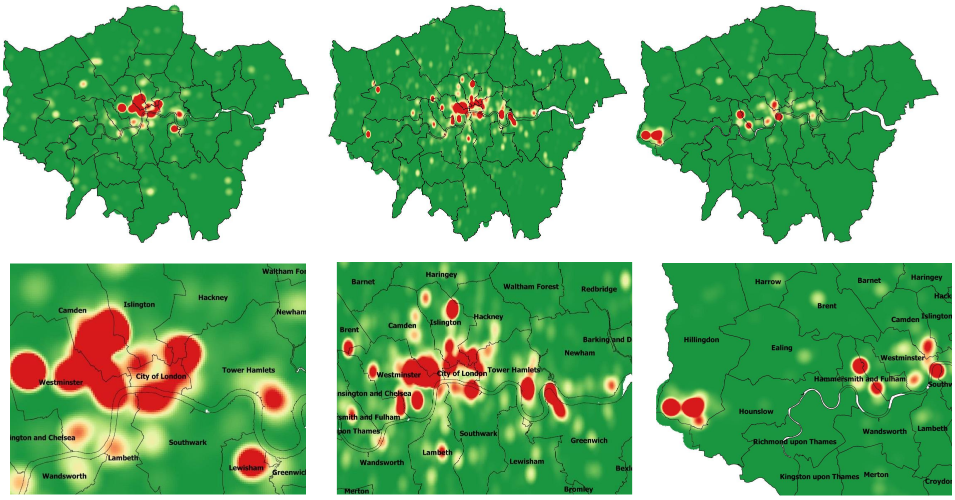
Figure 4.8: Percentage distribution of tweets during different days of the week - 2016

4.3.3 Tweeting Activity During the Week

The percentage tweet distribution pattern of a week Fig.4.8 in different sectors was examined 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 I appreciate that there exists an opportunity to further enhance the existing 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.

4.3.4 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. An aggregation of the geo-tagged tweets with the activity centres as commercial property polygons, associated with the SIC Codes was performed, 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 were analysed. The heat-maps are broken down into different sectors for both the year 2012 and 2016 data as shown in Fig.4.10. The brighter red colour



Construction-2012 Real Estate-2012 Transportation-2012

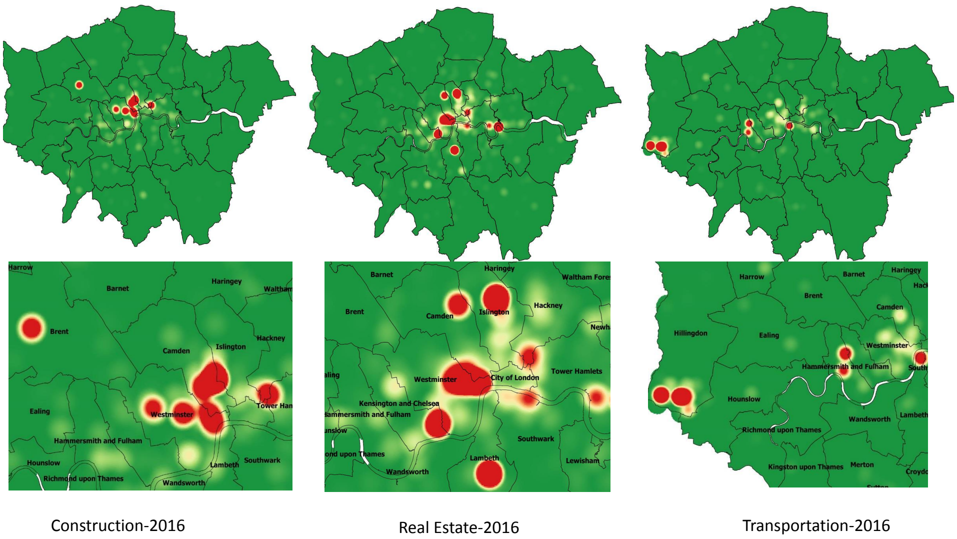
Figure 4.9: Heat-maps of Tweeting density in London's Industrial Sectors - 2012

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 [164–166]. 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.

4.3.5 Sentiment across SIC Categories

A general sentiment distribution across various industries in London using our two weeks of data for two distinct years was then examined. The tweets were first allocated in each SIC code. I then calculated the percentage of positive, neutral and negative tweets in each sector for both years of data. Fig. 4.11 and Fig.4.12 shows the sentiment distribution of individual industries. As can be seen in Fig.4.11 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.4.12, *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.



Construction-2016

Real Estate-2016

Transportation-2016

Figure 4.10: Heat-maps of Tweeting density in London's Industrial Sectors - 2016

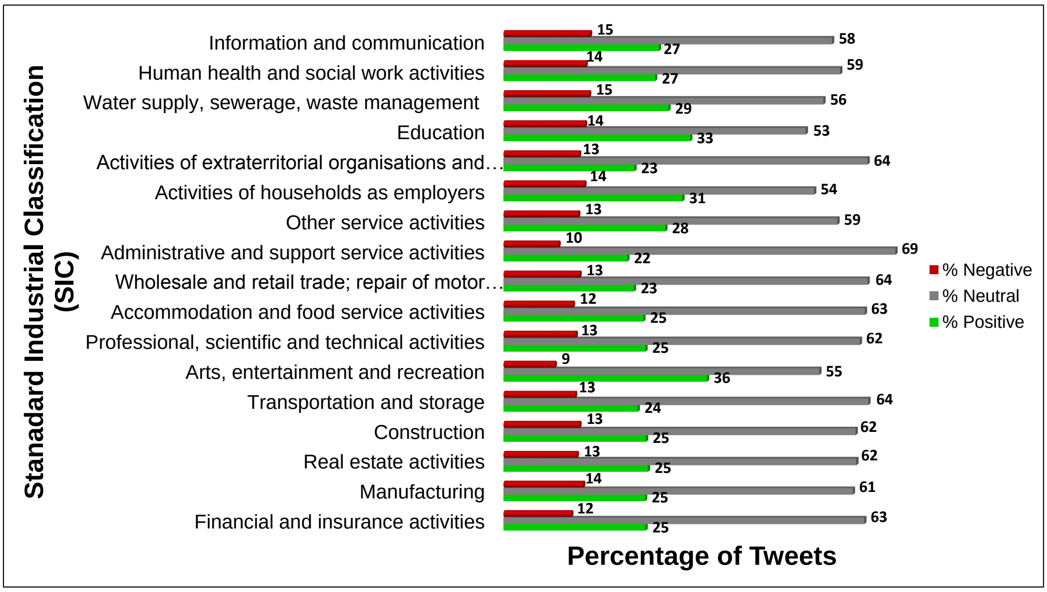


Figure 4.11: Sentiment in 2012

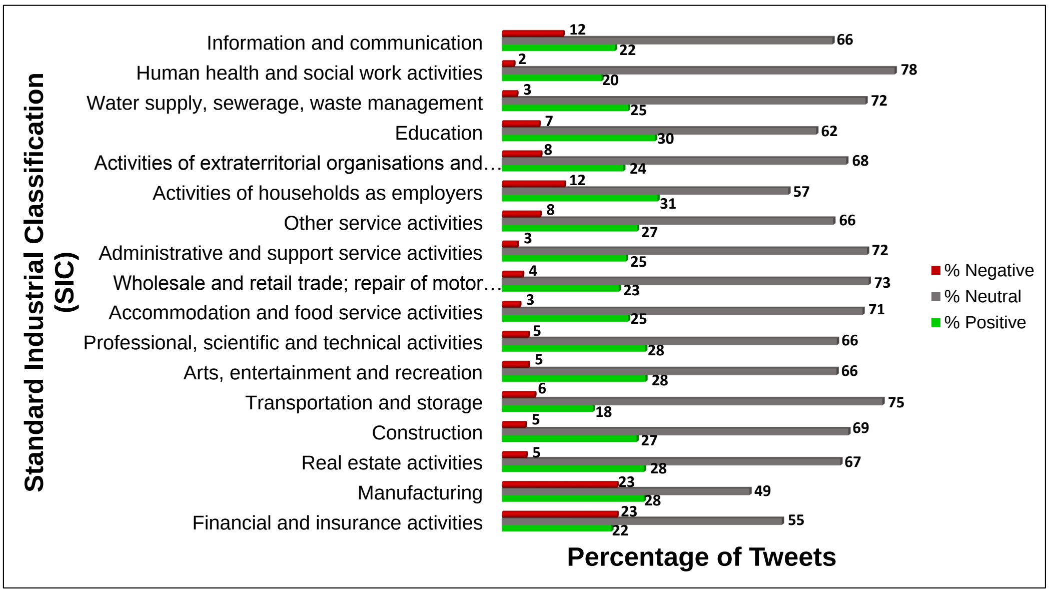


Figure 4.12: Sentiment in 2016

4.3.6 Tweet Hourly Trends within SIC Categories

In order to study the impact of working hours on sentiment, I made the decision to analyse 2016 data-set only, as it had more number of data points available in commercial areas for analysis in comparison to 2012 dataset Table 4.1. Fig 4.13 presents the normalised moving average of the tweet count within each industrial sector by hour of the day. For most of the industrial sectors studied a similar pattern was observed: the tweeting activity increases between 7:00 am and 9:00 am, when these commercial building locations start to be occupied by workers; there is also high tweeting activity in the second half of the day between 3:00 pm and 5:00 pm. However, in contrast to this general trend, the *human health and social work* sector peaks during the middle of the day between 11:00 am and 1:00 pm, and a further small peak occurs around 9:00 pm, highlighting the 24-hour work cycle of this industry (e.g., hospitals). The *water supply, sewerage and waste management* sector shows peak tweeting activity during the early hours of the day between 6:00 am and 9:00 am, again reflecting the working hours of this section of society.

4.3.7 Impact of Working Hours (using tweeting intensities) on Tweet Sentiment

After the tweets were clipped to the commercial areas in London, their aggregated temporal patterns were studied to establish the starting time and the duration of

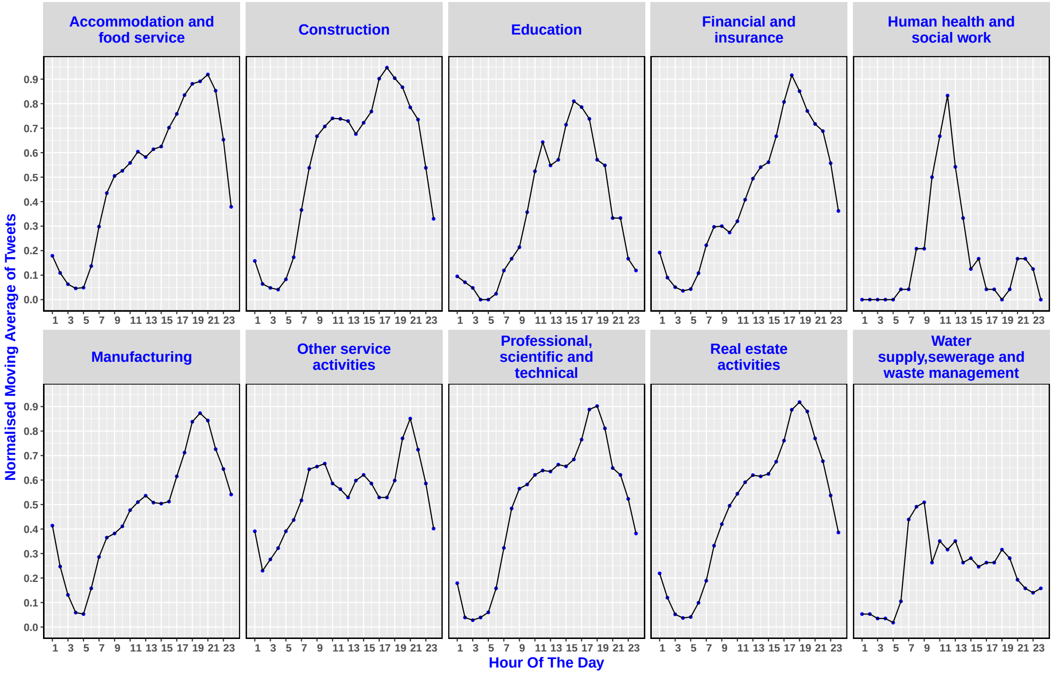


Figure 4.13: Hourly normalised moving average tweet count within industry sectors

individuals working in different industries using the threshold value of tweet activity times. To examine the impact of starting-time and duration on the sentiment score of the aggregated tweets, statistical measure of multivariate linear regression was employed.

After executing the linear regression model across different industrial sectors, the p-value for our predictor variables, *duration* and *start time* was calculated and was found to be greater ($p > .05$) for the majority of the industrial sectors studied as shown in Fig 4.14 (the exceptions to this are the *financial and Insurance* sector and the *professional, scientific and technical* sector). A high p-value for both predictor variables across the majority of other industrial sectors suggests that there is no linear dependence between the time variable and tweet sentiment. Therefore, based on the sample of datasets, the results lead to *accept the null hypothesis* H_0 (see Section 4.1.1.), i.e, working time has no impact on the sentiment of people working in the majority of industrial sectors in the Greater London area. The only exception to this result is where duration has some impact on the sentiment of people working in the *financial and insurance* sector (p-Value $< .05$) and the influence of start-time on the sentiment of people working in the *professional, scientific and technical* sector (p-Value $< .05$).

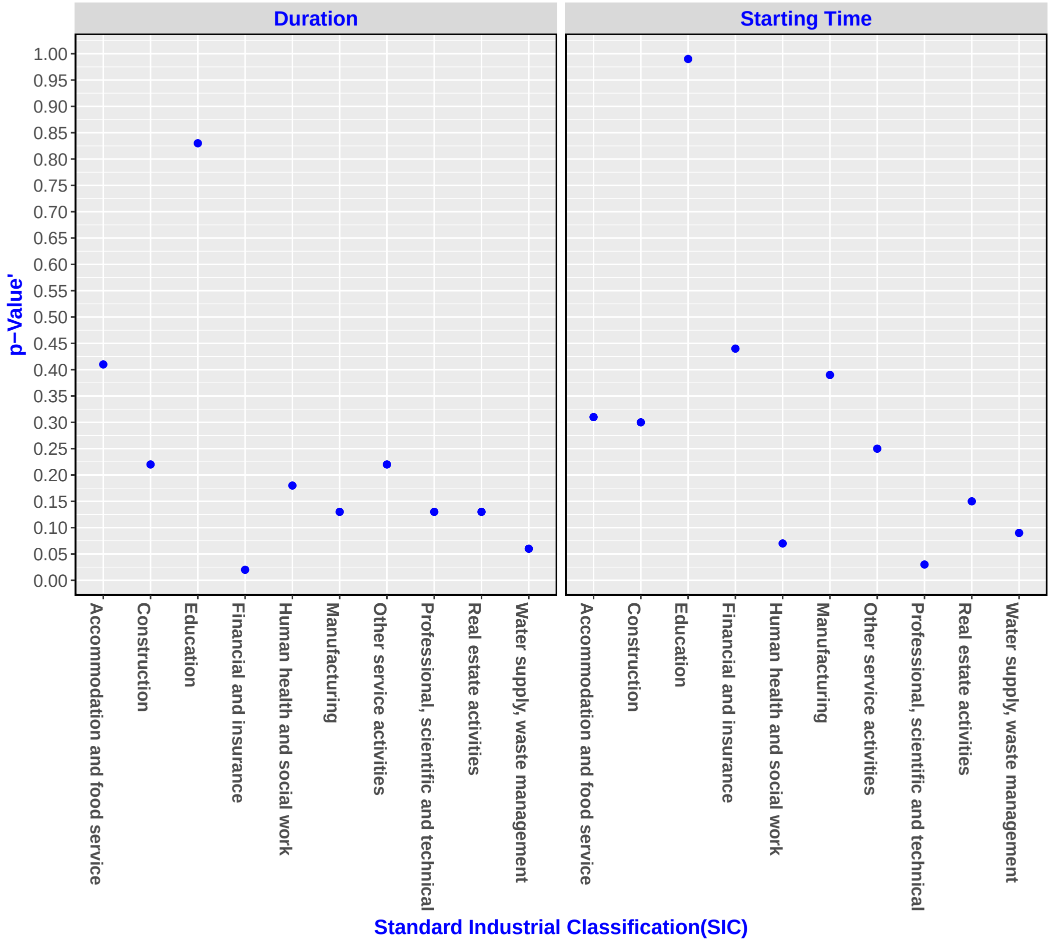


Figure 4.14: Multivariate regression p-values for the variables 'duration' and 'start time'

4.4 Discussion

To this end, previous studies have used survey data to research happiness in the work place [130, 168, 169]. In addition, the availability of digital traces at various urban locations has underpinned behavioural studies [118] and also the prediction of employment statistics from Twitter daily rhythms [170]. In this work an aggregation of large collections of tweet data at business locations to arrive at tweet sentiment and extracted temporal features for the understanding of group work patterns. We test the hypothesis that longer work hours and variable start-time has no impact on the happiness of people, using using tweet sentiment as a proxy happiness indicator.

The results of this experiment are consistent with the results of previous studies, which were conducted using survey sources and have reported that longer working hours do correlate with improved job satisfaction [171–173]. Conventional economics literature also suggests that the potential adverse effects of longer working hours are offset by the prospects of better income and job satisfaction [174]. Furthermore, research on 'utility theory, as defined by Rothbard and Edwards, suggests that people who allocate more time to work derive more satisfaction from it [24], which is echoed in the positive sentiment evident in our accumulated tweet text. Happiness has been identified as independent to the number of working hours [175] but associated to other positive factors such as social contacts at work [176].

The frequent count of positive sentiment evident at work places i.e. in different industrial sectors in London urban areas indicate a positive mood of population whilst working. Considering the notion that people's mood is reflected in their language (as discussed in literature review Chapter), policy makers can leverage this methodology as an interim measure between long spaced surveys to track the happiness cues within industrial segment from Twitter data. However, the current research is limited to analysis based on aggregated tweets and their text sentiment analysis within each industry sector. This methodology though onsets the use of big data and social media analytics application to mine the mood and sentiment in industries, yet it poses certain limitations. A detail investigation is required to understand the spatio-temporal aspect of a precise work location at more granular level in the interest of assessing the difference between geographical, temporal and tweets means and the influence of such factors on tweets sentiment. For example, future research potentially could look at only the Financial Sectors tweets by isolating the tweets emerging from offices involved in financial activities but spread across the different geographical areas in London.

4.5 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 the best of my knowledge there are limited studies that highlights the users of social media in context to various industry sectors they are linked to. The study in this Chapter integrates existing administrative data on businesses with geo-tagged tweets to uncover social media trends at work. Barring limitations which are future research avenues, the main contributions are as follows: *Firstly*, the experiment devised 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*, the visualisation of 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. *Thirdly*, for each sector a general mood of people evident from the tweets sentiment is uncovered using computational natural language processing (NLP) methods. The above mentioned findings confirm that not all industries participate equally in conversations on Twitter. The baseline results for the Greater London area identifies *Construction*, *Real Estate*, *Transport* and *Financial Services* industries consistently have stronger Twitter footprints. Although, the application of 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.

Additionally, the potential adverse effects of variable and longer working hours on happiness has long been studied in organisational and labour economics. Data from social media platforms (such as Twitter and Foursquare) is increasingly being used to conduct analysis on dynamic urban populations at detailed temporal and spatial resolution. In contrast to previous studies that employ survey-based methods to understand the happiness of people working in different economic sectors, the approach in this chapter employ both real time and static data to understand the relationship between working time patterns and aggregated sentiment in commercial buildings. I tested the hypothesis by joining INSPIRE polygon data (from the UK Land Registry), which provides high spatial resolution of registered properties in UK, to geo-spatial tweets, thereby establishing tweets 'belonging to different industrial sectors as classified by Standard Industrial Classification (SIC) code. I employed computational Natural Language Processing methods to derive tweet sentiment and establish the operational working hours from the time series of tweet intensity. I tested this approach on data from London UK and also studies the impact of *start-time* and *duration* on aggregated *sentiment score*. Whilst the results reveal no general trends across industrial sectors, which is consistent with previous literature, the scalable approach shows that for the *financial and insurance* sector, the *duration* of working hours does impact on employee sentiment. I believe that this new data-driven, scalable methodology has application beyond the analysis of mood prevalence in various economic sectors, and can be used to extend more traditional survey and sample based 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. In general, this research findings provide new evidence, and methodology, to strengthen the findings of previous research.

Chapter 5

Exploring Twitter Strengths and Limitations to Detect Event Related Sentiment in Industry - A Brexit Case Study

This chapter presents some preliminary work conducted to explore the opportunities as well as limitations of using Twitter data in order to unearth public perceptions and sentiment of the populace about an event like Brexit in the industrial landscape. The methodology detailed in a prior chapter empowers us to link the real time geo-tagged Tweets with diverse industrial sectors, potentially offering an alternative to survey methods for monitoring the ongoing sentiment and mood in different sectors of the industry. It is however not clear whether such data can be leveraged to ascertain the aggregate industrial sentiment during about political events such as Brexit, or will social scientists still need to conduct surveys of the industrial workforce to amass a collective feedback for the policy makers.

5.1 Introduction

As alluded to in the literature review chapter, Big Data has been used in several studies to glean new understanding and insights about various events. For instance, Twitter data has been utilised to comprehend urban sentiment about parks, restaurants, emotions of well-being status during disastrous events, as well as ascertaining political orientation about various civic events.

Perception of public sentiment during or subsequent to a political episode affecting the society in a drastic manner has been studied with great interest by computational social scientists using spatio-temporal-textual (STT) techniques as it unveils the behavioural pattern of the populace for usage in policy making. Though passively available social media data provides a great opportunity to identify a social phenomenon, survey focused social scientists are arguably doubtful about the quality and representativeness of such dynamically available data. They point to the limitations of bias, stability, privacy, selective reporting of findings and volatility to which this data is could be susceptible to [177]. In their opinion, surveys are indispensable for social science research but that methodology will need to adapt and incorporate the gains from Big Data inputs for perception of societal mores.A political move such as Brexit is almost certain to impact economic conditions in Britain and since the industrial sector contributes quite substantially to the UK economy, policy makers are bound to be interested in fathoming its real time sentiment trends.

At present such feedback from industries is solicited through organised survey methods conducted by organisations such as OECD or ONS (Office Of National Statistics, UK). But given the wider reach tools ingrained in the social media as exemplified by Twitter, more and more people are expressing their opinions about specific happenings using a geo-tagged tweet message. This means that signs of any upheaval within sectors of industry in response to political realignments can be facilely tapped through these tweets.

This chapter attempts to validate the advantages of going this route of data mining for event induced sentiment in diverse industry sectors. For instance, in context of Brexit, this task can be achieved by (i) analysing the volume of tweeting data during the time period of EU referendum and identifying the volume of Brexit related conversations within each industrial sector (ii) categorising the communication pattern and sentiment in industries that specifically participated in Brexit conversation and (iii) lastly contrasting the strengths, weaknesses, opportunities and challenges vis-a-vis survey data to monitor sentiment of people linked to different industrial sectors.

5.2 Data and Methodology

In this section we sketch the framework of the data used to conduct our exploratory experiment, and describe the methodology adapted to mine and aggregate Brexit related tweets.

5.2.1 Sourcing Twitter data

As stated in literature Review chapter there are multiple ways to mine Twitter data. For the present study we purchased tweets through the Historical Power Track (HPT) facility offered by Twitter. This facility searches the entire Tweet archive, and examines each Tweet posted during the period of interest, and generates a set of tweets matching the desired query [60]. We set our requirement for tweets posted from June 16-30, 2016 for the Greater London area in the bounding box (-0.5103 51.2868, 0.3340 51.6923). A bounding box defines the geographical extend (east-west-north-south combination of latitude and longitude parameters) of a place for which the tweets are to be mined. Twitter point of contact then ran a batch job for the query. These tweets were downloaded in JSON format following the steps mentioned in GNIP documentation [178]. The search query retrieved approximately 450,000 Tweets for the requested span of 15 days 7 days prior and 7 days after the referendum date of 23 June, 2016 for the Greater London bounding box of interest. This twitter data was then pre-processed to extract Tweet ID,Date-Time, Latitude, Longitude, and Tweet Text fields and the cleaned data was then stored in *csv* files. Subsequently this data was filtered with the geo-tagged information (latitude and longitude fields) NOT NULL generating a final sample of 118000 tweets as the target volume for our study. The tweet text was then annotated with the sentiment labels using machine learning Logistic Regression methods alluded to in an earlier chapter.

5.2.2 Filtering Tweets in Industrial Sectors using Land Registry and INSPIRE Polygon Data

Before extracting and analysing Brexit related conversations from tweets,the sample of 118000 geo-tagged tweets was linked to the SIC code by incorporating the method described in an earlier chapter. There was a total of 72292 tweets which could be linked to commercial INSPIRE polygons but only 36031 tweets were of interest for further analysis since only these had the SIC Code information field as NOT NULL, a requirement mandatory to link the tweet to an industrial sector.

The aggregated count of tweets in each industrial sector is shown in Table 5.1. The industries labelled 16-21 were excluded from analysis since (i) data of these industries for statistically significant analysis is insufficient. (ii) Non-urban industries, viz. Mining and Agriculture, due to geographical constraints were excluded as well.

After creating SIC code identified summaries of tweets, the tweeting density and count of the tweets were analysed in each industrial sector. The temporal distribution of tweeting activity within each industries was plotted using the Kernel

Sr.No.	**Industry Type**	**Count**
1	Financial and insurance activities	12033
2	Real estate activities	7984
3	Arts, entertainment and recreation	4325
4	Construction	3307
5	Accommodation and food service activities	2044
6	Transportation and storage	1669
7	Professional, scientific and technical activities	1114
8	Administrative and support service activities	1096
9	Wholesale and retail trade; repair of motor vehicles and motorcycles	880
10	Activities of extraterritorial organisations and bodies	318
11	Water supply, sewerage, waste management and remediation activities	274
12	Activities of households as employers	273
13	Other service activities	237
14	Education	169
15	Manufacturing	106
16	Human health and social work activities	82
17	Information and communication	71
18	Agriculture, Forestry andFishing	17
19	Electricity, gas, steam and air conditioning supply	15
20	Public administration and defence; compulsory social security	9
21	Mining and Quarrying	3

Table 5.1: Tweets Count in Industries during Brexit time window

density graph in R software. A kernel density graph prevaricates the problem of data discreteness which in the current experiment is the variable number of tweets count for each industrial sector, and provides a smoother estimated function by giving higher weights [179] to the data points that are proximate to each other. Fig.5.1 shows the Kernel density distribution of the sample dataaggregated within each industrial sector.

5.2.3 Screening Tweets to Identify Brexit Keywords and Hashes

For the analysis of conversations about the Brexit event in industrial space a literature review of Brexit studies that used Twitter data [140], [142], [141], [97], [143] was

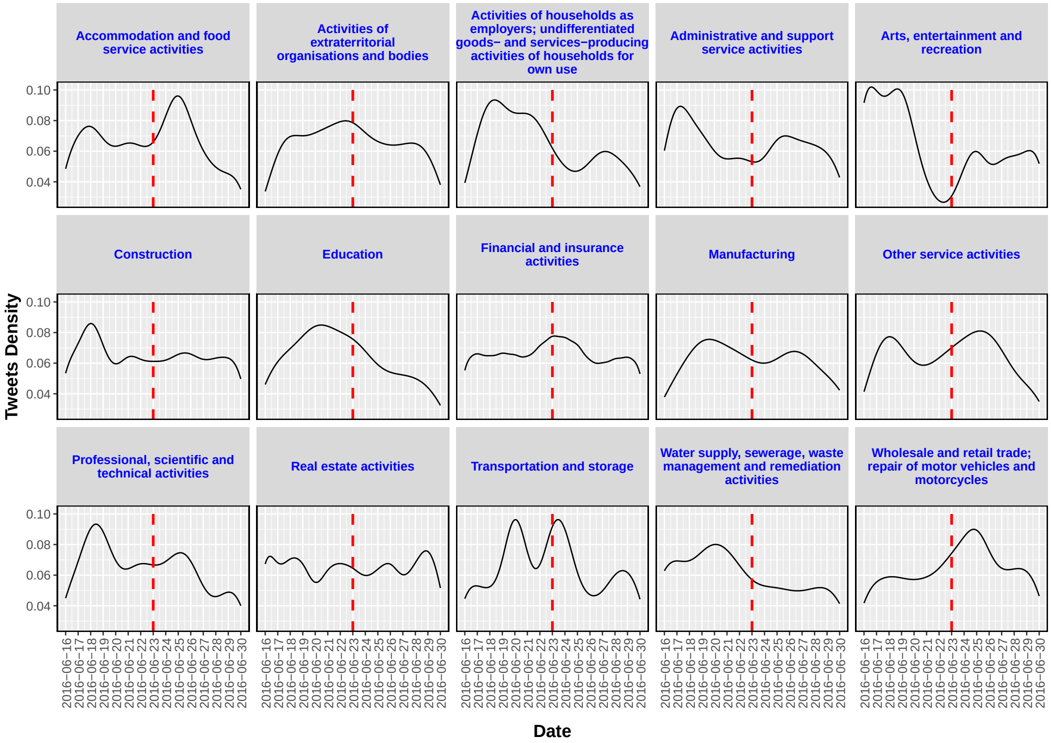
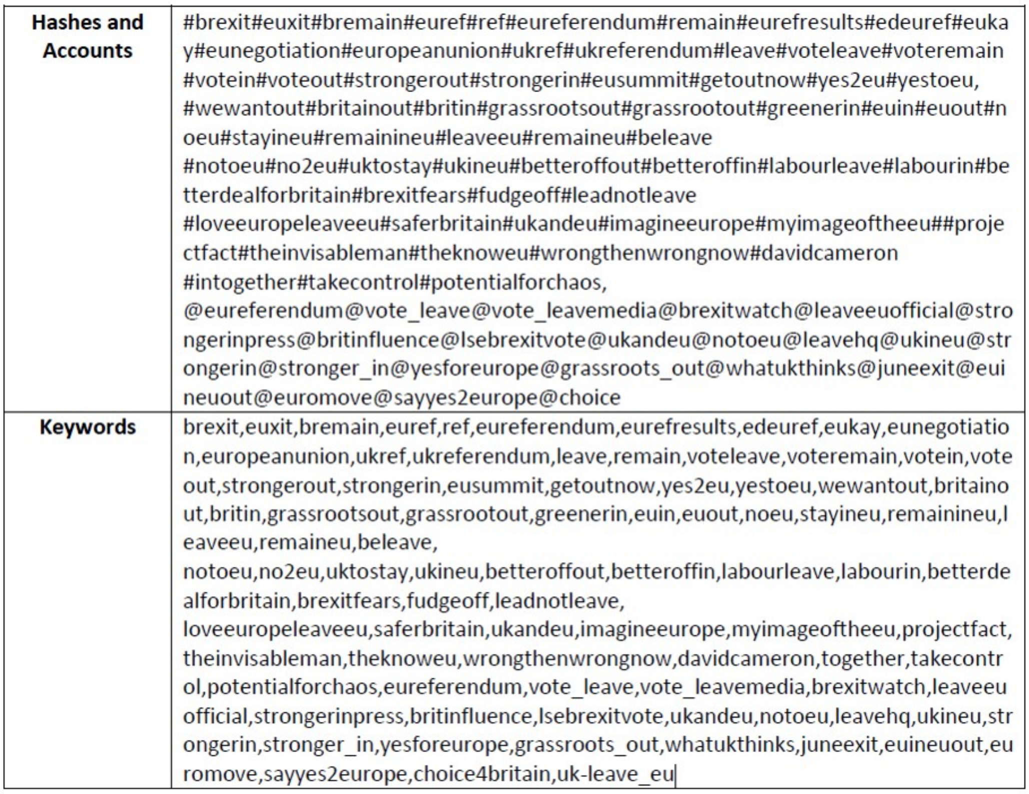


Figure 5.1: Tweeting density in different industries during Brexit

carried out and a list of words associated with Brexit (hashtags and keywords) was generated. The collected corpus of hashtags, user accounts and keywords used in this study are given in Figure. 5.2 .A NLP code was then written in R to automatically match the tweet text to hashtags and keywords in the list. Two flags (keyword and hashtag) added to the dataframe and the text matching algorithm updated these flags to be TRUE for the records where the tweet text contained the Brexit related keywords or hashtags. Surprisingly, only five industries, viz. *Financial and Insurance Activities*, *Real Estate*, *Arts and Entertainment*, *Construction* and *Accommodation and Food Services* had a significant number of Brexit related tweet counts (*threshold* arbitrarily set at a minimum of 50 tweets) were found suitable for further sentiment analysis. Tweets emerging from all other industry type polygons contained no mention of Brexit keywords and hashes. We were therefore constrained to focus further analyses on the industries with sufficient and relevant data samples. Furthermore, *Arts and Entertainment* industry had to be discarded from analysis since this sector is significantly skewed by a large number of visitors who are not a regular and typical part of this industry. All these provisos explain the rationale for our results and discussion being confined to the sentiment analysis in four industry sectors *Financial and Insurance Activities*, *Real Estate*, *Construction* and *Accommodation and Food Services* only.



Hashes and Accounts	#brexit#euxit#bremain#euref#ref#eureferendum#remain#eurefresults#edeuref#eukay#eunegotiation#europeanunion#ukref#ukreferendum#leave#voteleave#voteremain#votein#voteout#strongerout#strongerin#eusummit#getoutnow#yes2eu#yestoeu, #wewantout#britainout#britin#grassrootsout#grassrootout#greenerin#euin#euout#noeu#stayineu#remainineu#leaveeu#remaineu#beleave #notoeu#no2eu#uktostay#ukineu#betteroffout#betteroffin#labourleave#labourin#betterdealforbritain#brexitfears#fudgeoff#leadnotleave #loveeuropeleaveeu#saferbritain#ukandeu#imagineeurope#myimageoftheeu##projectfact#theinvisableman#theknoweu#wrongthenwrongnow#davidcameron #intogether#takecontrol#potentialforchaos, @eureferendum@vote_leave@vote_leavemedia@brexitwatch@leaveeuofficial@strongerinpress@britinfluence@lsebrexitvote@ukandeu@notoeu@leavehq@ukineu@strongerin@stronger_in@yesforeurope@grassroots_out@whatukthinks@juneexit@euineuout@euromove@sayyes2europe@choice
Keywords	brexit,euxit,bremain,euref,ref,eureferendum,eurefresults,edeuref,eukay,eunegotiation,europeanunion,ukref,ukreferendum,leave,remain,voteleave,voteremain,votein,voteout,strongerout,strongerin,eusummit,getoutnow,yes2eu,yestoeu,wewantout,britainout,britin,grassrootsout,grassrootout,greenerin,euin,euout,noeu,stayineu,remainineu,leaveeu,remaineu,beleave, notoeu,no2eu,uktostay,ukineu,betteroffout,betteroffin,labourleave,labourin,betterdealforbritain,brexitfears,fudgeoff,leadnotleave, loveeuropeleaveeu,saferbritain,ukandeu,imagineeurope,myimageoftheeu,projectfact, theinvisableman,theknoweu,wrongthenwrongnow,davidcameron,together,takecontrol,potentialforchaos,eureferendum,vote_leave,vote_leavemedia,brexitwatch,leaveeuofficial,strongerinpress,britinfluence,lsebrexitvote,ukandeu,notoeu,leavehq,ukineu,strongerin,stronger_in,yesforeurope,grassroots_out,whatukthinks,juneexit,euineuout,euromove,sayyes2europe,choice4britain,uk-leave_eu

Figure 5.2: Brexit Hashtags, Accounts and Keywords

5.3 Results and Discussion

This section relates preliminary findings from the analyses of Twitter platform carried out to evaluate the sentiment pattern in several industrial sectors. It also discusses limitations as well as the opportunities inherent in Twitter data as scrutinising tools for sentimental analysis in a social or political context.

5.3.1 Industrial Participation in Brexit Conversation

Initially, we analysed relative density of the 3 types of tweets tweets with Brexit hashtags, tweets with Brexit keywords, and tweets with no reference to Brexit for each of the 15 days for the 4 industries, *Financial and Insurance Activities*, *Real Estate*, *Construction* and *Accommodation and Food Services* in our truncated data set (see

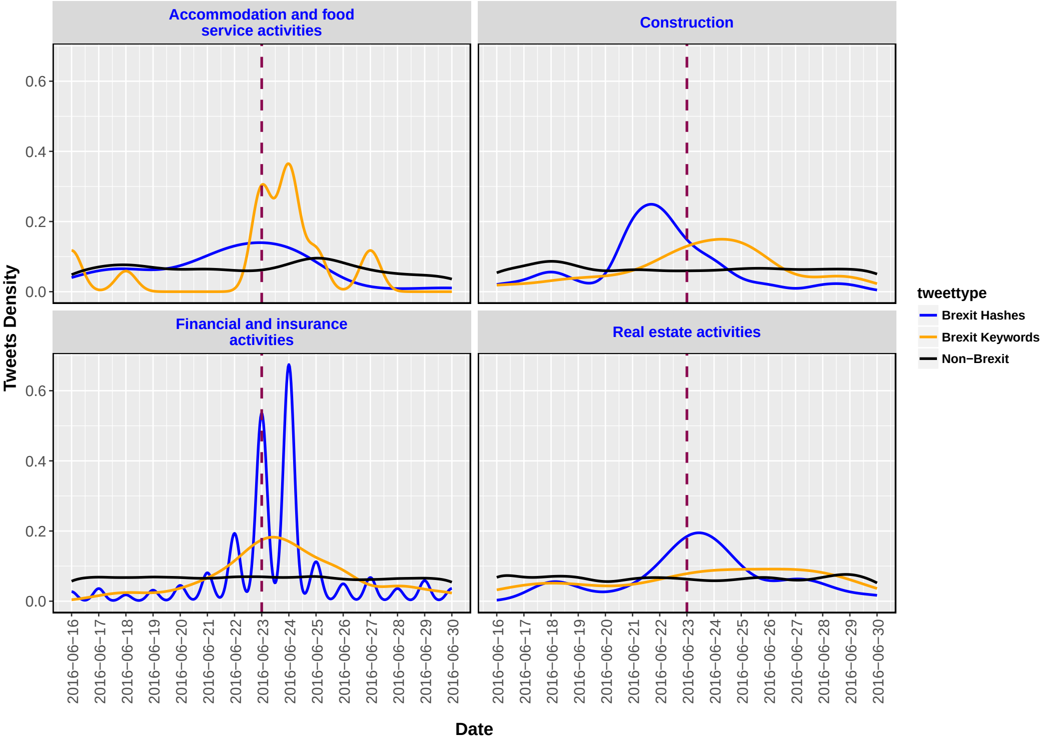


Figure 5.3: Tweet types during Brexit time window

Figure. 5.3) . As can readily be seen from the graphs, *Real Estate*, *Accommodation and Food Services* and *Construction* tweets had scant usage of hashtags but Brexit related keywords in the general language enabled their capture. On the other hand, *Financial and Insurance Activities* industry tweets make greater usage of hashtags demonstrating the fact that this sectors workforce is well versed in the protocols of using social media (specifically Twitter in the present case). Especially noteworthy from the graphs is the spike in the volume of tweets by *Accommodation and Food Services* and *Financial and Insurance Activities* industries subsequent to the broadcast of the result of referendum (Britain to exit EU) signifying that the outcome came as a distinct shock to this sector and they surmised that it may have an adverse economic impact on them. In contrast to this, the *Construction* sector had more conversations about the event before the date of referendum, and *Real Estate* sector discussed the event most on the day of referendum itself. Given the general tendency to think or converse more about things people consider important, this simple statistical analysis sheds light on the importance each sector assigns to the political event, viz. Brexit, in this case.

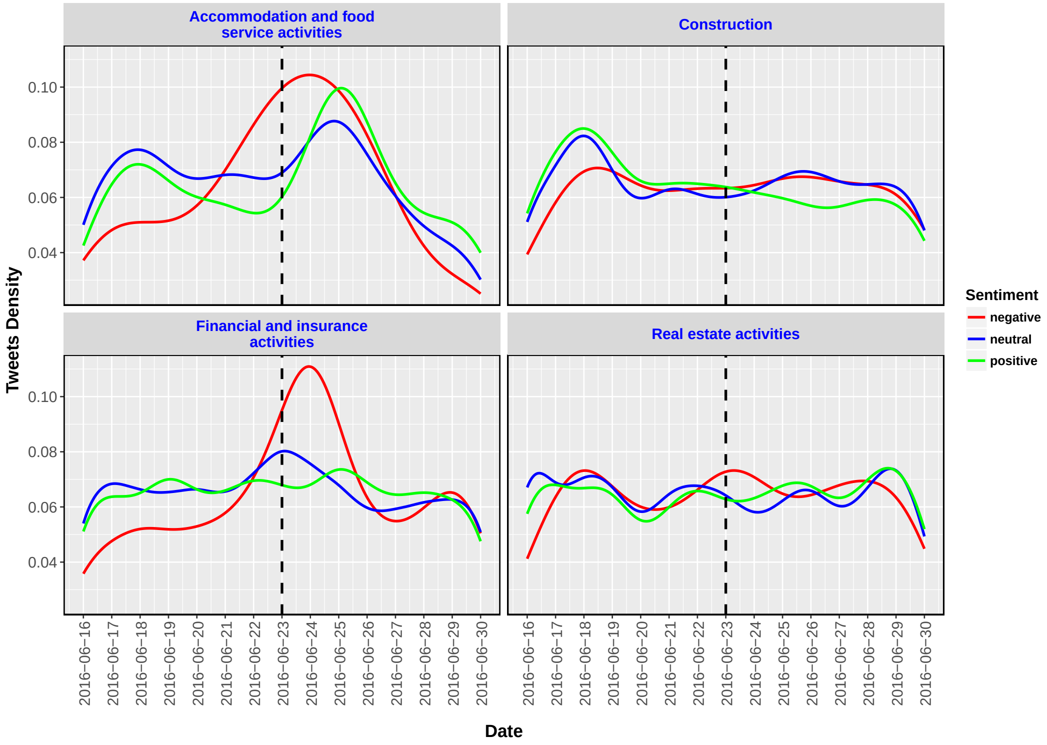


Figure 5.4: Tweet sentiment during Brexit time window

5.3.2 Brexit Related Tweet Sentiment in Selected Industry

Plots of positive, negative and neutral sentiments from the tweets are shown in Figure. 5.4) for the 4 industrial sectors, viz. *Financial and Insurance Activities* ,*Real Estate* , *Construction* and *Accommodation and Food Services*. As can be seen from the graphs, *Financial and Insurance Activities* sector shows a high volume of negative tweets after the referendum indicating a gloomy prospect following the exit of Britain from the EU. Similarly, *Accommodation and Food Services* industry also generates higher volume of negative sentiment on 24 June 2016, the day following the EU referendum but in this instance the negativity trend displays a declining pattern. *Real Estate* sector exhibits a fairly negative sentiment on the day of the referendum but on subsequent days blurs with the positive and neutral. Distribution of the three sentiments in *Accommodation and Food Services*, *Construction* and *Real Estate* industries tends to be relatively similar for both the week before and after the referendum.

Our results expose a temporal decay mechanism of a tweeting trend and also reveals the user preferences during and after EU refrendum, which is similar to the trends observed by previous studies that explored the Twitter activity trend during

London Homocide 2012 event [180]. Such analysis offers an indication as to how the social media mirrors the behaviour of people during the real-life events like Brexit and can immediately be tapped to seek real time sentiment of people working in some industry types.

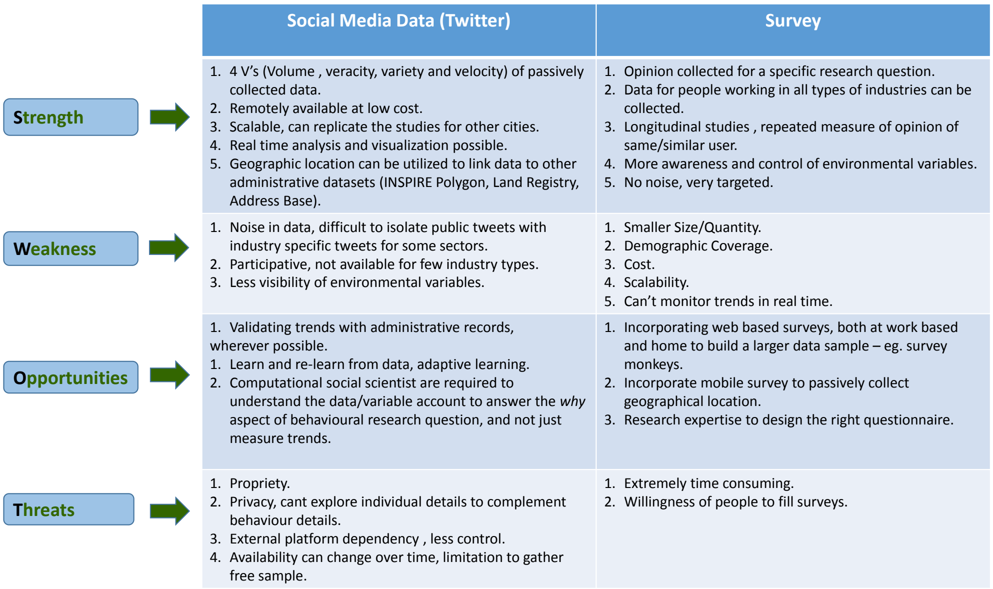
5.3.3 Discussing Limitations and Strengths - Twitter Data Linked to Industries

Digital data available from social media such as Twitter provide an illuminating view about the socio-political conversations people indulge in on the web. Hence, to probe the opinion of the people working in diverse industrial sectors, we experimentally evaluated the combination of Twitter data to other forms of organic data viz. Land Registry spatial property data. The results exhibit an imbalance of social media participation and we discovered that Twitter data is not representative for all industrial sectors. Since in some industries, for example, *Real Estate*, Twitter data can be significantly noisy. For, businesses identifiable by SIC category namely *Real Estate* are not necessarily home to 'Real Estate' type of work but may also have been rented out to other types of businesses. Hence we can expect a considerable error in the results. Further research is needed to devise mechanisms to address such shortcomings.

Therefore, in this section we discuss limitations and strengths our methodology offers whilst linking Tweets to industry type in order to sense people behaviour during a time of an event like Brexit. We compare our approach to traditionally available *Surveys* methods which has been codified by social scientist from decades to perform any such behavioural studies in industry segments.

In this section we also compare and contrast our methodology vis-a-vis survery methods in context of identifying people behaviour in industry due to a social or political event, e.g. Brexit. One of the strengths of Twitter data is its availability at minimal cost which offers an opportunity to perform a scalable visual analysis in real time. The geo-tagged data can be associated to other traditionally available information (for e.g. Land Registry and INSPIRE Polygon) computationally maximising an insight about a social phenomenon. However, in circumstances where Twitter data includes noise (for example Industries in public sphere such as Transportation and Wholesale) it is difficult to discern specific behavioural patterns about the workforce. Survey methods pose a clear-cut questionnaire and collect opinions and attitudes of people regarding specific research questions which big social media is unable to do [181]. In recent times, survey methods are leapfrogging to the digital methods of opinion collection (e.g. by the use of web applications like Survey

Monkeys and mobile surveys) which is the correct step to reach a greater number of individuals. An important constraint to using Twitter based data analysis is privacy issues since these proprietary media tools limit a scientist's ability to perform any fine-grained individual analysis. They can only report the trends available from data at an aggregated level and such collective trends can change over time. A number of researchers [145] [177] have affirmed that big data sources and surveys ought to be used in synergy for a better understanding of the behavioural pattern of the masses. Analogous to a previous study [182] we conclude this section by providing a SWOT analysis (Figure 5.5) in brief narrating our appraisal of the comparative use of Twitter and Survey data to study event specific behavioural patterns in industries.



	Social Media Data (Twitter)	Survey
Strength	1. 4 V's (Volume , veracity, variety and velocity) of passively collected data. 2. Remotely available at low cost. 3. Scalable, can replicate the studies for other cities. 4. Real time analysis and visualization possible. 5. Geographic location can be utilized to link data to other administrative datasets (INSPIRE Polygon, Land Registry, Address Base).	1. Opinion collected for a specific research question. 2. Data for people working in all types of industries can be collected. 3. Longitudinal studies , repeated measure of opinion of same/similar user. 4. More awareness and control of environmental variables. 5. No noise, very targeted.
Weakness	1. Noise in data, difficult to isolate public tweets with industry specific tweets for some sectors. 2. Participative, not available for few industry types. 3. Less visibility of environmental variables.	1. Smaller Size/Quantity. 2. Demographic Coverage. 3. Cost. 4. Scalability. 5. Can't monitor trends in real time.
Opportunities	1. Validating trends with administrative records, wherever possible. 1. Learn and re-learn from data, adaptive learning. 2. Computational social scientist are required to understand the data/variable account to answer the *why* aspect of behavioural research question, and not just measure trends.	1. Incorporating web based surveys, both at work based and home to build a larger data sample – eg. survey monkeys. 2. Incorporate mobile survey to passively collect geographical location. 3. Research expertise to design the right questionnaire.
Threats	1. Propriety. 2. Privacy, cant explore individual details to complement behaviour details. 3. External platform dependency , less control. 4. Availability can change over time, limitation to gather free sample.	1. Extremely time consuming. 2. Willingness of people to fill surveys.

Figure 5.5: SWOT analysis of Twitter and Survey Methods

5.4 Conclusion

The objective of the experiment detailed in this chapter was to determine and ascertain situations where Twitter data can be used advantageously to explore industry workforce sentiment in context of an event such as Brexit. We study the geo-tagged London tweets spanning over the Brexit period (7 days before and 7 days after the date of EU referendum) in diverse industrial sectors in order to corroborate or highlight limitations of this methodology. Using NLP methods we uncovered that not all industrial sectors participate equally in event related conversations. We also found that a minimum threshold of tweets is necessary to conduct industrial sector specific sentiment analysis. Furthermore, tweets originating from public-sphere industrial sectors (for e.g. Real Estate, Arts and Entertainment, Transportation) are difficult to process due to substantial admixture of opinions from visitors and industry workforce. But it is clear from our study that sentiment trends from the following three industries, viz. *Construction*, *Financial and insurance Sector*, and *Accommodation and Food Services* can be gainfully investigated using Twitter data.

Chapter 6

Discussion and Conclusion

This chapter concludes the Thesis. The text in the following section provides the reader with a chapter-by-chapter summary.

6.1 Thesis Summary

Chapter 2 provides a summarised theoretical background to social media analytics, specifically explaining the state-of-the art NLP methods existing for Twitter text analysis - lexicon and machine learning methods. Since an important aspect of this study is geo-spatial analysis using GIS techniques which enables us to filter geo-tagged twitter data at various locations, a brief account of geo-spatial analysis is described too. Finally, a significant reference to Twitter based text and geo-spatial studies is made which served as ground to inspire this study.

Chapter 3 presented the first and preliminary experiment conducted to explore the relationship between tweet sentiment achieved using lexicon based sentiment classification NLP methods to 0.4 million geo-tagged Tweets in the Greater London and London ward characteristics. The results verified the established thinking: that job opportunities correlate strongly with positive sentiments and additionally generated two other insights that happiness has a negative relationship with the number of children and happiness has a parabolic relationship with access to public transportation. The latter implies that those that rely on, but do not have strong access to public transport are the least happy. Separately, the number of jobs, number of children, and accessibility to public transport explain 33-47% of the variations in tweet sentiment.

Chapter 4, presented the experiment and results on how social media usage and general sentiment vary across the industry sector landscape for Greater London region. Unlike existing studies which use survey methods to study industrial trends, a novel geo-computational approach to link location specific Twitter data with business sectors (by leveraging the UK SIC Code) is devised and explained in detail. The baseline results for the Greater London area identifies *Construction*, *Real Estate*, *Transport* and *Financial Services* industries consistently have stronger Twitter footprints. The application of NLP methods and geo-spatial filtering of Tweets within industries using diverse open source datasets enable an understanding about prevailing sentiment within each industry sector and discuss how the evidence can contribute towards de-biasing Twitter data. Additionally, results of this chapter inform the usage of Twitter across various industrial sectors. A further exploration is done to verify if tweeting intensities can be studied to explore the working hours and its effect on aggregated sentiment (as a proxy of happiness) within various industries. Surveys and self reported methods has long existed in behavioural studies to make an investigation about such trends, however they lack scalability and real-time monitoring needs. As a result of these limitations this Chapter proposed a new method for incorporating secondary sources of data, both real-time and static, to evaluate the working hours and sentiment of employees across several industrial sectors. Based on the existing research, we hypothesise that duration of working hours and variable start time in different industrial sectors has no or very limited impact on the sentiment of the workforce. This hypothesis was tested by joining INSPIRE polygon data (from the UK Land Registry), which provides high spatial resolution of registered properties in UK, to geo-spatial tweets, thereby establishing tweets 'belonging to' different industrial sectors as classified by Standard Industrial Classification (SIC) code. Computational Natural Language Processing methods were then employed to derive tweet sentiment and establish the operational working hours from the time series of tweet intensity. This approach was tested on data from London UK and whilst the results reveal no general trends across industrial sectors, the scalable approach does shows that for the *financial and insurance* sector, the duration of working hours does impact on employee sentiment.

Chapter 5 presented a preliminary analysis about the Brexit related communication pattern and sentiment trends on Twitter for industrial landscape. The study was based on geo-tagged London tweets spanning over the Brexit period (7 days before and 7 days after the date of EU referendum) to corroborate or highlight limitations of the proposed methodology for understanding political event related sentiment from industries. It was discovered that not all industrial sectors participate equally in event related conversations and there exist a minimum threshold of tweets necessary (3000 in the current case) to conduct any industrial sector specific event related sentiment analysis. Although, this number can be variable given the dynamic nature of social media data. However, the sentiment trends from the following three industries, viz. *Construction* , *Financial and insurance Sector*, and *Accommodation and Food Services* can be gainfully investigated using Twitter data. For rest of the industrial sectors researchers still need to rely on survey methods given low social media participation level.

6.2 Discussion

In this section I present my results in light of the contributions this Thesis makes whilst exploring the feasibility of using a combination of Twitter and open source urban data to understand sentiment in cities with a particular focus on industrial sectors. Apart from highlighting the findings in a broader urban science context, some limitation of the current methodology are scrutinised too. Finally, I wrap up this chapter with some suggestions and ideas for the future work.

6.2.1 Research results and urban science

The results of this study suggests that Twitter data can be meaningfully linked and integrated to multitude of open data sources, be it urban socioeconomic data or spatial property data available from cities authorities to uncover and understand the leading indicators of happiness in cities. By correlating and measuring the influence of the socioeconomic parameters of London wards on aggregated Twitter sentiment, this research demonstrates how the arduous task of understanding people behaviour can be achieved in more real-time using new sources of data such as Twitter. The first research question this Thesis aimed for was to examine *which demographic variable strongly correlate with aggregated sentiment of tweets in London wards and what is the statistical measure of those variables on tweet sentiment.* The results of this question not only endorse the proven fact that the *availability of job opportunities*, as a variable explains substantial effect of sentiment variations but also provides an added insights that there exist a negative relationship of sentiment with the *number of children* socio-economic factor in London wards and a parabolic relationship of sentiment with *access to public transportation* factor prevailing in London wards. Since the Twitter data can be mined form hundred and thousands of citizens in a passive manner, such data source provide an advantage of low cost and real-time reporting of happiness indicators in the city landscape.

My exploration further focused to mine sentiment trends within various industrial sectors in an endeavour to first understand and compare the usage of social media in industries specifically asking the question *how social media usage and general sentiment vary across the geographic and industry sector landscape and what is the impact of working hours (as estimated from tweet temporal intensities) on aggregated Twitter sentiment within each industry segments.* The baseline results for the Greater London area identifies *Construction*, *Real Estate*, *Transport* and *Financial Services* industries consistently to have stronger Twitter footprints. Industries which are related to public places (*Transportation*, *Retail* and *Entertainment* Sector) presumably generate tweeting activity which is a results of combination of visitors and employees tweets. Whilst in such industries our results discover no general trends, however, the proposed methodology shows that for the *Financial and Insurance* sector, the duration of working hours does have an impact on employee sentiment.

Since the methodology devised in this research provides a possibility to isolate Twitter data within industrial domain, a further investigation was performed to seek the sentiment pattern in industries in relation to a political event like Brexit. Predominantly, the third and last research question explored in this Thesis is to understand *the usage of Twitter data to ascertain the aggregated industrial sentiment about political events and what are the limitations, if any, of Twitter analytics are in contrast to the traditional survey methods analysis.* In the results one can observe a limited use of Brexit related hashtags in the *Real Estate* and *Construction* sector , however *Financial and Insurance* sectors make greater use of tweet hashtags confirming how this particular industry is well versed with the social media protocols. The findings of this experiment reveal *Financial and Insurance* sector exhibit high volume of negative tweets after the event of Brexit pole which indicates a a unhappy prospect following the exit of Britain from the EU. Similarly, *Accommodation and Food Services* and *Real Estate* industry also generates negative sentiment on 24 June 2016, the day following the EU referendum. Additionally, the experiments affirm that Twitter data can be used to mine sentiment of people engaged in the above mentioned industrial sectors with greater ease since the workforce employed in these industries has higher participation level on this social media platform as is evident from the Twitter volume. However, other industrial sectors, labelled 9 to 21 in Table 5.1 of Chapter 5, are currently constraint to incorporate the proposed methodology as the volume of tweets is relatively low for any significant statistical analysis. This highlight the limitation of using Twitter data, i.e. the behavioural pattern and sentiment of people can be mined from Twitter for some industry types but is not achievable for all the industrial sectors currently. Therefore, to understand the prevailing sentiment for industries which has low social media participation, researchers has to pertain survey methods.

With an ever increasing digital data, scientists have an opportunity to process and analyse behavioural patterns of our society - individuals or collectives, by assimilating techniques available in varied disciplines. As a result, urban science researchers have started experimenting with text mining, spatial analysis and statistical evaluation techniques of computer science, geography and statistics respectively pioneering a new interdisciplinary field of research called "Computational Social Sciences"[183]. The citizen generated content on social media sites, also often termed as volunteered geographic information (VGI), is therefore progressively been used in variety of meaningful ways to monitor people behaviour in various social contexts - events [7, 8, 11], comparing people moods in cities, mobility scenarios [12], business gains [34–36], tracking community happiness [1, 14] and many more application discussed earlier in this thesis. Linking social media sentiment analysis to socio-economic features of a place allows urban planners to analyse the citizen sentiment in perspective of the geographical location and its features providing a novel understanding and knowledge about a social phenomenon to plan better urban system.

Equally, in light of the methodology proposed in this research, we are in a position to isolate twitter data in commercial places by leveraging the data which is organically available with the government authorities (UK Land Registry). Data from social media can be divided or fragmented and cannot represent the 360 degree view about a particular service in a urban context. When combining it with other datasets, features of a place - commercial or residential areas - not only enable the study of behaviour in a user current urban ecosystems but also provides a better representation about social media users. There are models designed by earlier researchers which uses language-based features to identify location indicative words in order to predict the location of Tweet user [184], or apply machine learning methods for user classification [185] whilst others use the tweet history and clustering of followers/followees to find the demographic estimation in order to profiling Twitter users [186]. Although these methods are important as they are able to predict the location of both geotagged and non-geotagged tweets, the methodology proposed in this Thesis systematically join databases to solve an important problem of mining industrial sentiment in a simplistic way. This fusion of diverse data sources, Twitter and spatial property data in the current case, unveil a mechanism to advance the application of state-of-the-art sentiment analysis methods in behavioural studies targeted towards industry sectors. Survey methods have long existed in social science research to explore the happiness of people at work places, however, they are high cost and time consuming processes. As widely being acclaimed by researchers in this domain [90, 145, 146], integrating both, survey and social media analytics, allows social scientists to analyse the changing sentiment of people in a novel way.

Exploring happiness indicators in cities, particularly industrial sectors using Twitter sentiment analysis and urban GIS data provides a new opportunities to explore the long-standing research queries such as *what influences happiness of people*. Studies based on survey methods has conveyed that although happiness correlates positively with hours worked but this relationship changes inversely in U-shape correlation when working hours are long [187]. Therefore, long working hours are believed to have a negative impact on happiness level and in some cases, for example a preliminary findings of a two-year experiment with a six-hour working day in Gothenburg, Sweden has been proven to improve employee well-being [188]. Conversely, the 'utility theory' acclaims more allocation of time at work drives satisfaction [24] and happiness has been identified as independent to the number of working hours [175]. Such existing studies has particularly relied on survey methods to enquire about happiness trends related to working patterns. The ongoing mood of the citizens engaged in various economic sectors and their working hours trends can be timely and passively obtained using aggregated tweeting intensities of people tweeting at work places. This research therefore is an endeavour to use a novel geo-computational approach incorporating social media data to ask how working hours (evident from tweeting intensities) influence the sentiment of people at work places. So far, surveys such as Enterprise survey (ESENER) measures stress at work in a variety of industry sectors and in UK ONS measures the subjective well-being of the the labour force on annual basis [132] [133]. The aforementioned surveys methods lacks scalablility and does not provide a timely solution to understand the happiness level of workforce in relation to day to day working hours operative at a work location or sentiment in reaction to an event. Importantly, in UK and Europe, to the best of my knowledge and inquiry there currently exist no measure to monitor the sentiment of workforce engaged in various industrial sectors on a frequent basis which this research endeavour to address using social media data.

In a scenario where a policy maker would like to introduce a policy change impacting working conditions in a specific industrial sector, authorities would be faced with a challenge to isolate the sentiment of working people engaged with that particular industry type. Therefore, isolating tweets in different industries not only aids in capturing the feedback of aforementioned policy changes but also presents an exciting opportunity for organisational studies to study and compare sentiment in relation to prevailing working hours proxies created using Twitter data. The work described in this study creates a sentiment analysis framework by integrating openly available property spatial datasets with socially generated data to monitor people sentiment , which can be taken as a proxy to attain a positive or native feedback during various events. *These powerful data sets, INSPIRE spatial property data and UK Land registry data, which has been existing in silos are meaningfully combined with social media data opening an opportunity to study the tweeting trends in industries and cities.* For businesses, such framework can be used in targeted advertising and viral marketing which is always looking for communities that is highly active in order to spread the relevant product information.

6.3 Study Limitations

Like any research, the work presented in this Thesis has some limitations. In this section I acknowledge such limitations, a few of which could be a subject of future work whilst others remain legitimate constraint of the study.

Firstly, the tweets filtered and spatially clipped inside the commercial property polygons, resulted in a very small data sample in comparison to the initial full data set, i.e. only 11% for the year 2012 and 6% for the year 2016 respectively (as illustrated in Table 4.1 in Chapter 4). A larger initial sample of geo-tagged tweets which is either purchased or collected for a longer duration (few months or a year) would clearly remedy this.

Secondly, tweets linked to industrial sectors - *Transport*, *Retail* and *Arts and entertainment* and some *Service sectors* like restaurants,hospitals etc are subject to noise as these industries operate in public sphere. A simple spatial filtering of tweets within these locations (which are identified as commercial properties in UK Land Registry and Address Base datasets) for these industries particularly shall not generate tweets linked to employees working in these sectors. Further investigation is required to device mechanism to isolate the tweeting activity of visitors and employees on these locations.

Moreover, although the methodology proposed in this study successfully created links between tweets and the type of industrial properties they originate from, the tweet data is subject to 'noise. Multi-story and mixed use (commercial and residential) properties will impact results. 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 there exist no height feature for the tweets, and the INSPIRE polygons do not have a height dimension, there is no way to establish which floor the tweet has originated from and as such for those properties , a tweet could have emerged form commercial or residential areas. This is an existing data restriction, which partially could be addressed using a tweet content analysis approach (e.g. Topic Modelling methods of NLP).

Another limitation is, 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, future work should consider ways to address such location specific planimetric accuracy issues.

Additionally, any Twitter-based analysis clearly reflects a sub-set of society who are engaged with such platforms. Although the aggregated patterns evident form tweets can be studied, it is challenging to comment on the representativeness of the results for individual working patterns of the population. Furthermore, working pattern and sentiment analysed using tweeting intensities and aggregated sentiment summaries, could not be bench-marked and evaluated against a known happiness measure available for different industry type SIC Codes as I couldn't identify any data source to compare my results with i.e. happiness measure of industries in the UK identifiable by its SIC section name.

I believe that the majority of limitations reflected are artefacts of the data. People in organisations and industrial sectors follow specific rules and their aggregated patterns develop over time as digital footprints. Social media tools like Twitter can uncover macro-level temporal activity, and therefore the methodology introduced in this study does have value and could be applicable to a diverse range of population studies.

6.4 Ideas for Future Research

The study described in Thesis is a proof of concept to use Twitter data for mining sentiment in relation to various socioeconomic patterns and also with respect to the working hours prevalent in each industrial sectors. The current work could be expanded for future studies in varied ways.

First, future work should endeavour to collect a larger volume of Twitter data which is targeted towards the working people specifically. This could be achieved by creating a work related hashtags on twitter such as '#working hours' or '#work'. Further research could be done in exploring how the language is used at work places and possibly mining those specific keywords from Twitter space for longer duration of time , in order to achieve a greater sample size for analysis. Furthermore, the current working style of our society is very fluid. There is an integration of work and life with people working from home and people work whilst they commute. According to one of the recent BBC article [189], the availability of wi-fi on trains has spread of mobile phones use which has extended the working day of people as they respond to e-mails on train. It will be quite interesting to expand our methodology and explore novel ways to integrate it to mobile phone data and email sentiment to capture such working patterns. However, such endeavour would require further collaboration with data hosting companies whilst at the same time ensuring an adoption to privacy protocols.

Secondly, the replication of the current experimentation with Twitter datasets for other big cities in the UK shall provide additional insight about the behaviour of the people in different cities. For example, one can expect that the availability of jobs is widely applicable across cultures, whereas the number of children will depend on the culture and the availability to public transport will depend on the ownership level of personal vehicles as well as the culture of transport. The INSPIRE property spatial data and UK Land Registry is currently available for the whole of the UK. The current methodology could be replicated for other urban cities such as Manchester and Birmingham to assess the reported trends and study its variation to London. Other countries in EU where the INSPIRE property and land use data is available could be studied too. However, one has to take into account the geographic and cultural features of those cities, as the working pattern, social and geographical dynamics of place shall vary to what is evident in London.

Future work can also focus on creating proprietary sentiment labels for each city by combining meta-data for boosting sentiment analysis accuracy [190] which can enable large-scale cross-country or city comparisons to be made [191].

Lastly, scientists have so far examined happiness in cities using Gallup-Healthways Surveys to discover human capital plays a central role in happiness of cities in addition to the income parameter [192]. The results of this Thesis confirms a positive correlation between income and Tweets positive sentiment. Perhaps there are prospects to mine group sentiment at various locations for example entertainment industry , and employ Topic Modelling NLP techniques to enquire topics that relate to postive sentiment.

6.5 General Conclusion

In general, Twitter offers urban scientist an opportunity to efficiently collect or obtain enormous amount of data about the people activity at various locations unobtrusively thereby offering a possibility to monitor a true reaction and behaviour of the people during a social phenomenon. Such data can be meaningfully combined with socio-economic, commercial or spatial open data sources to understand citizens happiness in context to their urban ecosystem. Not only such insights will prove a valuable surveillance tool for policy makers and service providers to monitor ongoing sentiment in different domains of society but can also be incorporated as a low cost alternative to survey methods in organisational studies to study the sentiment of people. In essence, this new data-driven, scalable methodology has application beyond the analysis of mood prevalence in various segments of society (London Wards or particular industrial sector), and can be used to extend more traditional survey and sample based methods.

Bibliography

[1] Weisi Guo, Neha Gupta, Ganna Pogrebna, and Stephen Jarvis. Understanding happiness in cities using Twitter: Jobs, children, and transport. In *Smart Cities Conference (ISC2), 2016 IEEE International*, pages 1–7. IEEE, 2016.

[2] Neha Gupta, Henry Crosby, David Purser, Stephen Javis, and Weisi Guo. Twitter usage across industry: A spatiotemporal analysis. In *2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 64–71. IEEE, 2018.

[3] Neha Gupta, Weisi Guo, and Stephen Jarvis. Analysing Twitter and UK Land Registry data to explore working hours and sentiment across industrial sectors. *Submitted to PLOS One*, 2019.

[4] Nataliya Tkachenko, Sarunkorn Chotvijit, Neha Gupta, Emma Bradley, Charlotte Gilks, Weisi Guo, Henry Crosby, Eliot Shore, Malkiat Thiarai, Rob Procter, et al. Google trends can improve surveillance of type 2 diabetes. *Scientific reports*, 7(1):4993, 2017.

[5] Twitter online stats. URL `https://news.un.org/en/story/2016/09/539112-nearly-47-cent-global-population-now-online-un-report`.

[6] Using twitter, . URL `https://help.twitter.com/en/using-twitter`.

[7] Antonia Saravanou, George Valkanas, Dimitrios Gunopulos, and Gennady 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*, pages 1233–1238. ACM, 2015.

[8] Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann. *Twitter and society*, volume 89. P. Lang, 2014.

[9] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[10] Bella Robinson, Robert Power, and Mark Cameron. A sensitive Twitter earthquake detector. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 999–1002. ACM, 2013.

[11] Chris Allen, Ming-Hsiang Tsou, Anoshe Aslam, Anna Nagel, and Jean-Mark Gawron. Applying GIS and Machine Learning methods to Twitter data for multiscale surveillance of influenza. *PloS one*, 11(7):e0157734, 2016.

[12] Bartosz Hawelka, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti. Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3): 260–271, 2014.

[13] Vanessa Frias-Martinez and Enrique Frias-Martinez. Spectral clustering for sensing urban land use using twitter activity. *Engineering Applications of Artificial Intelligence*, 35:237–245, 2014.

[14] Daniele Quercia, Jonathan Ellis, Licia Capra, and Jon Crowcroft. Tracking gross community happiness from tweets. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 965–968. ACM, 2012.

[15] Adam Tsakalidis, Symeon Papadopoulos, Alexandra I Cristea, and Yiannis Kompatsiaris. Predicting elections for multiple countries using Twitter and polls. *IEEE Intelligent Systems*, 30(2):10–17, 2015.

[16] Diana Maynard and Adam Funk. Automatic detection of political opinions in tweets. In *Extended Semantic Web Conference*, pages 88–99. Springer, 2011.

[17] R.Mitchell and F. Popham. Greenspace, urbanity and health: relationships in england. *Journal of Epidemiology and Community Health*, 61(8):Journal of Epidemiology and Community Health, 2007.

[18] Lansdall-Welfare Thomas, Lampos Vasileios, and Cristianini Nello. Nowcasting the mood of the nation. *Significance*, 9(4):26–28.

[19] Lewis Mitchell, Morgan R. Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M. Danforth. The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics

of place. *PLOS ONE*, 8:1–15, 05 2013. URL https://doi.org/10.1371/journal.pone.0064417.

[20] Morgan R. Frank, Lewis Mitchell, Peter Sheridan Dodds, and Christopher M. Danforth. Happiness and the patterns of life: A study of geolocated tweets. *Scientific Reports*, 3, 09 2013.

[21] Luke Sloan, Jeffrey Morgan, Pete Burnap, and Matthew Williams. Who tweets? deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *PloS one*, 10(3):e0115545, 2015.

[22] Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. An analysis of the user occupational class through Twitter content. The Association for Computational Linguistics, 2015.

[23] Dhiraj Murthy, Alexander Gross, and Alexander Pensavalle. Urban social media demographics: An exploration of Twitter use in major American cities. *Journal of Computer-Mediated Communication*, 21(1):33–49, 2016.

[24] Nancy P. Rothbard and Jeffrey R. Edwards. Investment in work and family roles: A test of identity and utilitarian motives. *Personnel Psychology*, 56 (3):699–729, 2003. ISSN 1744-6570. doi: 10.1111/j.1744-6570.2003.tb00755.x. URL http://dx.doi.org/10.1111/j.1744-6570.2003.tb00755.x.

[25] Weiguo Fan and Michael D. Gordon. The power of social media analytics. *Commun. ACM*, 57(6):74–81, June 2014. ISSN 0001-0782. doi: 10.1145/2602574. URL http://doi.acm.org/10.1145/2602574.

[26] Bogdan Batrinca and Philip C. Treleaven. Social media analytics: a survey of techniques, tools and platforms. *AI & SOCIETY*, 30(1):89–116, Feb 2015. ISSN 1435-5655. doi: 10.1007/s00146-014-0549-4. URL https://doi.org/10.1007/s00146-014-0549-4.

[27] Raymond Kosala and Hendrik Blockeel. Web mining research: A survey. *SIGKDD Explor. Newsl.*, 2(1):1–15, June 2000. ISSN 1931-0145. doi: 10.1145/360402.360406. URL http://doi.acm.org/10.1145/360402.360406.

[28] Bingbing Ni, Zheng Song, and Shuicheng Yan. Web image mining towards universal age estimator. In *Proceedings of the 17th ACM International Conference on Multimedia*, MM '09, pages 85–94, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-608-3. doi: 10.1145/1631272.1631287. URL http://doi.acm.org/10.1145/1631272.1631287.

[29] Yun Fu and T. S. Huang. Human age estimation with regression on discriminative aging manifold. *Trans. Multi.*, 10(4):578–584, June 2008. ISSN 1520-9210. doi: 10.1109/TMM.2008.921847. URL http://dx.doi.org/10.1109/TMM.2008.921847.

[30] Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174:50 – 59, 2016. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2015.01.095. URL http://www.sciencedirect.com/science/article/pii/S0925231215011297.

[31] Esam Alwagait, Basit Shahzad, and Sophia Alim. Impact of social media usage on students academic performance in saudi arabia. *Computers in Human Behavior*, 51:1092 – 1097, 2015. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2014.09.028. URL http://www.sciencedirect.com/science/article/pii/S0747563214004804. Computing for Human Learning, Behaviour and Collaboration in the Social and Mobile Networks Era.

[32] Stoney Brooks. Does personal social media usage affect efficiency and well-being? *Computers in Human Behavior*, 46:26 – 37, 2015. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2014.12.053. URL http://www.sciencedirect.com/science/article/pii/S0747563215000096.

[33] Marcia W. Distaso, Tina McCorkindale, and Alexa Agugliaro. America's most admired companies social media industry divide. *Journal of Promotion Management*, 21(2):163–189, 2015. doi: 10.1080/10496491.2014.996799.

[34] Sinan Aral and Dylan Walker. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science*, 57(9):1623–1639, 2011. doi: 10.1287/mnsc.1110.1421. URL https://doi.org/10.1287/mnsc.1110.1421.

[35] Hiroshi Onishi and Puneet Manchanda. Marketing activity, blogging and sales. *International Journal of Research in Marketing*, 29(3):221 – 234, 2012. ISSN 0167-8116. doi: https://doi.org/10.1016/j.ijresmar.2011.11.003. URL http://www.sciencedirect.com/science/article/pii/S0167811612000419.

[36] W. Glynn Mangold and David J. Faulds. Social media: The new hybrid element of the promotion mix. *Business Horizons*, 52(4):357 – 365, 2009. ISSN 0007-6813. doi: https://doi.org/10.1016/j.bushor.2009.03.002. URL http://www.sciencedirect.com/science/article/pii/S0007681309000329.

[37] Wu He, Shenghua Zha, and Ling Li. Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, 33(3):464 – 472, 2013. ISSN 0268-4012. doi: https://doi.org/10.1016/j.ijinfomgt.2013.01.001. URL http://www.sciencedirect.com/science/article/pii/S0268401213000030.

[38] Pvels Gonarovs. Data analytics in crm processes: A literature review. *Information Technology and Management Science*, 20(1), 2017. URL https://content.sciendo.com/view/journals/itms/20/1/article-p103.xml.

[39] Jitendra Ajmera, Hyung-iL Ahn, Meena Nagarajan, Ashish Verma, Danish Contractor, Stephen Dill, and Matthew Denesuk. A crm system for social media: Challenges and experiences. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 49–58, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2035-1. doi: 10.1145/2488388.2488394. URL http://doi.acm.org/10.1145/2488388.2488394.

[40] Rajiv Sabherwal and Irma Becerra-Fernandez. *Business Intelligence: Practices, Technologies, and Management.* John Wiley and Sons Inc., 2011.

[41] E. Qualman. *Socialnomics: How Social Media Transforms the Way We Live and Do Business.* Business professional collection. Wiley, 2010. ISBN 9780470901229. URL https://books.google.co.uk/books?id=yAqD19i2U0UC.

[42] achis group debuts social business index think of it as klout for companies. URL https://techcrunch.com/2011/09/13/dachis-group-debuts-social-business-index-think-of-it-as-klout-for-companies/.

[43] Dell ideastorm. URL https://www.ideaconnection.com/crowdsourcing/dell-00069.html.

[44] Richard Doherty. Getting social with recruitment. *Strategic HR Review*, 9 (6):11–15, 2010. doi: 10.1108/14754391011078063. URL https://doi.org/10.1108/14754391011078063.

[45] Karen Mossberger, Yonghong Wu, and Jared Crawford. Connecting citizens and local governments? social media and interactivity in major u.s. cities. *Government Information Quarterly*, 30(4):351 – 358, 2013. ISSN 0740-624X. doi: https://doi.org/10.1016/j.giq.2013.05.016. URL http://www.sciencedirect.com/science/article/pii/S0740624X13000774.

[46] A. Mainka, S. Hartmann, W. G. Stock, and I. Peters. Government and social media: A case study of 31 informational world cities. In *2014 47th Hawaii International Conference on System Sciences*, pages 1715–1724, Jan 2014. doi: 10.1109/HICSS.2014.219.

[47] Sebastian Denef, Petra S. Bayerl, and Nico A. Kaptein. Social media and the police: Tweeting practices of british police forces during the august 2011 riots. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 3471–3480. ACM, 2013. ISBN 978-1-4503-1899-0. doi: 10.1145/2470654.2466477. URL http://doi.acm.org/10.1145/2470654.2466477.

[48] Shreyansh P. Bhatt, Hemant Purohit, Andrew Hampton, Valerie Shalin, Amit Sheth, and John Flach. Assisting coordination during crisis: A domain ontology based approach to infer resource needs from tweets. In *Proceedings of the 2014 ACM Conference on Web Science*, WebSci '14, pages 297–298. ACM, 2014. ISBN 978-1-4503-2622-3. doi: 10.1145/2615569.2615652. URL http://doi.acm.org/10.1145/2615569.2615652.

[49] Mark A. Cameron, Robert Power, Bella Robinson, and Jie Yin. Emergency situation awareness from Twitter for crisis management. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, pages 695–698. ACM, 2012. ISBN 978-1-4503-1230-1. doi: 10.1145/2187980.2188183. URL http://doi.acm.org/10.1145/2187980.2188183.

[50] Ruth Garland, Damian Tambini, and Nick Couldry. Has government been mediatized? a uk perspective. *Media, Culture & Society*, 40(4):496–513, 2018. doi: 10.1177/0163443717713261. URL https://doi.org/10.1177/0163443717713261.

[51] Arthur Mickoleit. Social media use by governments. (26), 2014. doi: https://doi.org/https://doi.org/10.1787/5jxrcmghmk0s-en. URL https://www.oecd-ilibrary.org/content/paper/5jxrcmghmk0s-en.

[52] John C. Bertot, Paul T. Jaeger, and Justin M. Grimes. Using icts to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies. *Government Information Quarterly*, 27(3):264 – 271, 2010. ISSN 0740-624X. doi: https://doi.org/10.1016/j.giq.2010.03.001. URL http://www.sciencedirect.com/science/article/pii/S0740624X10000201.

[53] Matthew A Russell. *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More.* " O'Reilly Media, Inc.", 2013.

[54] Bogdan Batrinca and Philip C Treleaven. Social media analytics: a survey of techniques, tools and platforms. *Ai & Society*, 30(1):89–116, 2015.

[55] Bingsheng Wang, Haili Dong, Arnold P. Boedihardjo, Chang-Tien Lu, Harland Yu, Ing-Ray Chen, and Jing Dai. An integrated framework for spatio-temporal-textual search and mining. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '12, pages 570–573, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1691-0. doi: 10.1145/2424321.2424418. URL `http://doi.acm.org/10.1145/2424321.2424418`.

[56] IBM Software Business Analytics. Social media analytics making customer insights actionable, Feb 2013. URL `http://www.mcrinc.com/Documents/Newsletters/201401_SocialMediaAsMarketingTool.pdf`.

[57] An Oracle White Paper. Social media and business intelligence: Creating the integrated customer hub, September 2012. URL `http://www.oracle.com/us/products/social-media-and-bi-1845281.pdf`.

[58] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *ICWSM*, 2013.

[59] Fred Morstatter, Jürgen Pfeffer, and Huan Liu. When is it biased?: Assessing the representativeness of twitter's streaming api. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 555–556, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2745-9. doi: 10.1145/2567948.2576952. URL `http://doi.acm.org/10.1145/2567948.2576952`.

[60] Tweets delivery with Historical Power Track, . URL `http://support.gnip.com/articles/choosing-historical-api.html`.

[61] Ottawa. How gis models the real world. `https://hillnotes.ca/2016/11/16/gis-a-tool-for-visualizing-data-with-a-geographic-element-location-matters/`.

[62] J Li, JA Benediktsson, B Zhang, T Yang, A Plaza, G Cheng, J Han, X Lu, S Lefèvre, D Tuia, et al. Spatial technology and social media. *Proceedings of the IEEE*, 105(10), 2017.

[63] Paul A Longley, Michael F Goodchild, David J Maguire, and David W Rhind. *Geographic information systems and science.* John Wiley & Sons, 2005.

[64] Chalmers Brothers. Language and the pursuit of happiness, 2005.

[65] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January 2008. ISSN 1554-0669. doi: 10.1561/1500000011. URL `http://dx.doi.org/10.1561/1500000011`.

[66] Ahmed Abbasi, Ammar Hassan, and Milan Dhar. Benchmarking twitter sentiment analysis tools. In *LREC*, volume 14, pages 26–31, 2014.

[67] Tong Richard. M.: An operational system for detecting and tracking opinions in on-line discussions. In *Working Notes of the ACM SIGIR Workshop on Operational Text Classification, New York*, pages 1–6, 2001.

[68] George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244, 1990.

[69] M. Hu, B. Liu, and S. M. Street. Mining and Summarizing Customer Reviews. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2004.

[70] Mike Thelwall. The heart and soul of the web? sentiment strength detection in the social web with sentistrength. In *Cyberemotions*, pages 119–134. Springer, 2017.

[71] Pavlos Fafalios, Vasileios Iosifidis, Eirini Ntoutsi, and Stefan Dietze. Tweetskb: A public and large-scale rdf corpus of annotated tweets. In *European Semantic Web Conference*, pages 177–190. Springer, 2018.

[72] Atte Oksanen, David Garcia, and Pekka Räsänen. Proanorexia communities on social media. *Pediatrics*, pages peds–2015, 2015.

[73] Ilya Zheludev, Robert Smith, and Tomaso Aste. When can social media lead financial markets? *Scientific reports*, 4:4213, 2014.

[74] Eman MG Younis. Sentiment analysis and text mining for social media microblogs using open source tools: An empirical study. *International Journal of Computer Applications*, 112(5), 2015.

[75] Olga Kolchyna, Tharsis TP Souza, Philip Treleaven, and Tomaso Aste. Twitter sentiment analysis: Lexicon method, machine learning method and their combination. *arXiv preprint arXiv:1507.00955*, 2015.

[76] Jinan Fiaidhi, Osama Mohammed, Sabah Mohammed, Simon Fong, and Tai hoon Kim. Opinion mining over twitterspace: Classifying tweets programmatically using the r approach. In *Digital Information Management (ICDIM), 2012 Seventh International Conference on*, pages 313–319. IEEE, 2012.

[77] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[78] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*, 2017.

[79] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[80] Sara Rosenthal et. al. Semeval 2015 task 10, 2015. URL `http://alt.qcri.org/semeval2015/task10/index.php?id=data-and-tools`.

[81] Myle Ott. Python port of the twokenize class from ark-tweet-nlp, 2013. URL `https://github.com/myleott/ark-twokenize-py`.

[82] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL `http://www.aclweb.org/anthology/D14-1162`.

[83] Hussam Hamdan, Patrice Bellot, and Frederic Bechet. lsislif: Feature extraction and label weighting for sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 568–573, 2015.

[84] Bing Liu. Opinion lexicon, 2004. URL `https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html`.

[85] Theresa Wilson, Janyce Wiebe, and Paul 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*, pages 347–354. Association for Computational Linguistics, 2005.

[86] F. Nielsen. Afinn, 2011. URL `http://www2.imm.dtu.dk/pubdb/p.php?6010`.

[87] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.

[88] Payam Refaeilzadeh, Lei Tang, and Huan Liu. *Cross-Validation*, pages 532–538. Springer US, Boston, MA, 2009. ISBN 978-0-387-39940-9. doi: 10.1007/978-0-387-39940-9_565. URL `http://dx.doi.org/10.1007/978-0-387-39940-9_565`.

[89] Peter A Burrough, Rachael McDonnell, Rachael A McDonnell, and Christopher D Lloyd. *Principles of geographical information systems*. Oxford University Press, 2015.

[90] Peter Halfpenny and Rob Procter. *Innovations in digital research methods*. Sage, 2015.

[91] CartoDB Interactive Mapping, 2014. URL `developers.cartodb.com`.

[92] Bahareh Rahmanzadeh Heravi and Ihab Salawdeh. Tweet location detection. In *Computation+ Journalism Symposium*, 2015.

[93] UCL CASA. Welcome to maptube. URL `http://www.maptube.org/home.aspx`.

[94] Greater London Authority (GLA). London datastore. URL `https://data.london.gov.uk/dataset/statistical-gis-boundary-files-london`.

[95] Dan Tasse and Jason I. Hong. *Using User-Generated Content to Understand Cities*, pages 49–64. Springer International Publishing, Cham, 2017. ISBN 978-3-319-40902-3. doi: 10.1007/978-3-319-40902-3_3. URL `https://doi.org/10.1007/978-3-319-40902-3_3`.

[96] Piyushimita (Vonu) Thakuriah, Nebiyou Y. Tilahun, and Moira Zellner. *Big Data and Urban Informatics: Innovations and Challenges to Urban Planning and Knowledge Discovery*, pages 11–45. Springer International Publishing, Cham, 2017. ISBN 978-3-319-40902-3. doi: 10.1007/978-3-319-40902-3_2. URL `https://doi.org/10.1007/978-3-319-40902-3_2`.

[97] Thomas Lansdall-Welfare, Fabon Dzogang, and Nello Cristianini. Change-point analysis of the public mood in uk twitter during the brexit referendum. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2016.

[98] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1 – 8, 2011. ISSN 1877-7503. doi: https://doi.org/10.1016/j.jocs.2010.12.007. URL `http://www.sciencedirect.com/science/article/pii/S187775031100007X`.

[99] H. Engelbrecht. Natural capital, subjective well-being, and the new welfare economics of sustainability: Some evidence from cross-country regressions. *Ecological Economics*, 2009.

[100] L. Mitchell, M. Frank, K. Harris, P. Dodds, and C. Danforth. The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place. *PLOS ONE*, 8, 2013.

[101] M. Frank, L. Mitchell, P. Dodds, and C. Danforth. Happiness and the Patterns of Life: A Study of Geolocated tweets. *Scientific Reports*, 3, 2013.

[102] E. Kanjo and A. Chamberlain. Emotions in context: examining pervasive affective sensing systems, applications, and analyses. *Personal and Ubiquitous Computing*, 2015.

[103] Daniel Kahneman and Jason Riis. Living, and thinking about it: Two perspectives on life. *The science of well-being*, 1, 2005.

[104] Francis Ysidro Edgeworth. *Mathematical psychics: An essay on the application of mathematics to the moral sciences*, volume 10. Kegan Paul, 1881.

[105] Adam DI Kramer. An unobtrusive behavioral model of gross national happiness. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 287–290. ACM, 2010.

[106] Peter Sheridan Dodds and Christopher M. Danforth. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, 11(4):441–456, Aug 2010. ISSN 1573-7780. doi: 10.1007/s10902-009-9150-9. URL `https://doi.org/10.1007/s10902-009-9150-9`.

[107] Lewis Mitchell, Morgan R Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M Danforth. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one*, 8(5):e64417, 2013.

[108] Bruno Ohana, Brendan Tierney, and Sarah-Jane Delany. Domain independent sentiment classification with many lexicons. In *2011 Workshops of International*

Conference on Advanced Information Networking and Applications, pages 632–637. IEEE, 2011.

[109] R. Mitchell and F. Popham. Greenspace, urbanity and health: Relationships in England. *Journal of Epidemiology and Community Health*, 2007.

[110] T. Lansdall-Welfare, V. Lampos, and N. Cristianini. Nowcasting the mood of the nation. *Significance*, 9, 2012.

[111] Leonardo Allisio, Valeria Mussa, Cristina Bosco, Viviana Patti, and Giancarlo Francesco Ruffo. Felicittà: Visualizing and estimating happiness in italian cities from geotagged tweets. In *CEUR WORKSHOP PROCEEDINGS*, volume 1096, pages 95–106. CEUR Workshop Porceedings, 2013.

[112] A. Duarte, C. Garcia, G. Giannarakis, S. Limao, A. Polydoropoulou, and N. Litinas. New approaches in transportation planning: happiness and transport economics. *Economic Research and Electronic Networking*, 10, 2010.

[113] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Richard E Lucas, Megha Agrawal, Gregory J Park, Shrinidhi K Lakshmikanth, Sneha Jha, Martin EP Seligman, et al. Characterizing geographic variation in well-being using tweets. In *ICWSM*, pages 583–591, 2013.

[114] Yu Liu, Xi Liu, Song Gao, Li Gong, Chaogui Kang, Ye Zhi, Guanghua Chi, and Li Shi. Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers*, 105(3): 512–530, 2015.

[115] Quynh C Nguyen, Dapeng Li, Hsien-Wen Meng, Suraj Kath, Elaine Nsoesie, Feifei Li, and Ming Wen. Building a national neighborhood dataset from geotagged twitter data for indicators of happiness, diet, and physical activity. *JMIR public health and surveillance*, 2(2), 2016.

[116] John Krumm, Andrew L. Kun, and Petra Varsányi. Tweetcount: Urban insights by counting tweets. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, UbiComp '17, pages 403–411, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5190-4. doi: 10.1145/3123024.3124412. URL http://doi.acm.org/10.1145/3123024.3124412.

[117] Juan Carlos Garca-Palomares, Mara Henar Salas-Olmedo, Borja Moya-Gmez, Ana Condeo-Melhorado, and Javier Gutirrez. City dynamics through twitter: Relationships between land use and spatiotemporal demographics. *Cities*, 72:310 – 319, 2018. ISSN 0264-2751. doi: https://doi.org/10.1016/j.cities.2017.09.007. URL http://www.sciencedirect.com/science/article/pii/S0264275117306303.

[118] Andrew Jenkins, Arie Croitoru, Andrew T. Crooks, and Anthony Stefanidis. Crowdsourcing a collective sense of place. *PLOS ONE*, 11(4):1–20, 04 2016. doi: 10.1371/journal.pone.0152932. URL https://doi.org/10.1371/journal.pone.0152932.

[119] Guy Lansley and Paul A. Longley. The geography of twitter topics in london. *Computers, Environment and Urban Systems*, 58:85 – 96, 2016. ISSN 0198-9715. doi: https://doi.org/10.1016/j.compenvurbsys.2016.04.002. URL http://www.sciencedirect.com/science/article/pii/S0198971516300394.

[120] Hariton Efstathiades, Demetris Antoniades, George Pallis, and Marios D. Dikaiakos. Users key locations in online social networks: identification and applications. *Social Network Analysis and Mining*, 6(1):66, Aug 2016. doi: 10.1007/s13278-016-0376-3. URL https://doi.org/10.1007/s13278-016-0376-3.

[121] Francine Chen, Dhiraj Joshi, Yasuhide Miura, and Tomoko Ohkuma. Social media-based profiling of business locations. In *Proceedings of the 3rd ACM Multimedia Workshop on Geotagging and Its Applications in Multimedia*, GeoMM '14, pages 1–6, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-3127-2. doi: 10.1145/2661118.2661119. URL http://doi.acm.org/10.1145/2661118.2661119.

[122] Mauricio Santillana, Andr T. Nguyen, Mark Dredze, Michael J. Paul, Elaine O. Nsoesie, and John S. Brownstein. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLOS Computational Biology*, 11(10):1–15, 10 2015. doi: 10.1371/journal.pcbi.1004513. URL https://doi.org/10.1371/journal.pcbi.1004513.

[123] Joo Porto de Albuquerque, Benjamin Herfort, Alexander Brenning, and Alexander Zipf. A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International Journal of Geographical Information Science*, 29(4):667–689, 2015. doi: 10.1080/13658816.2014.996567. URL https://doi.org/10.1080/13658816.2014.996567.

[124] Efe Sevin. Places going viral: Twitter usage patterns in destination marketing and place branding. *Journal of Place Management and Development*, 6(3): 227–239, 2013. doi: 10.1108/JPMD-10-2012-0037. URL `https://doi.org/10.1108/JPMD-10-2012-0037`.

[125] Grant Blank. The digital divide among Twitter users and its implications for social research. *Social Science Computer Review*, page 0894439316671698, 2016.

[126] Abhijnan Chakraborty, Johnnatan Messias, Fabrício Benevenuto, Saptarshi Ghosh, Niloy Ganguly, and Krishna P. Gummadi. Who makes trends? understanding demographic biases in crowdsourced recommendations. *CoRR*, abs/1704.00139, 2017. URL `http://arxiv.org/abs/1704.00139`.

[127] Muhammad Adnan and Paul Longley. Analysis of twitter usage in london, paris, and new york city. In *16th AGILE international conference on geographic information science, Leuven*, pages 1–7, 2013.

[128] Barbara Hofer, Thomas J. Lampoltshammer, and Mariana Belgiu. *Demography of Twitter Users in the City of London: An Exploratory Spatial Data Analysis Approach*, pages 199–211. Springer International Publishing, Cham, 2015. ISBN 978-3-319-07926-4. doi: 10.1007/978-3-319-07926-4_16. URL `https://doi.org/10.1007/978-3-319-07926-4_16`.

[129] Bruno S Frey and Alois Stutzer. *Happiness and economics: How the economy and institutions affect human well-being.* Princeton University Press, 2010.

[130] Neil Thin. *Social happiness: Theory into policy and practice.* Policy Press, 2012.

[131] Luigino Bruni and Pier Luigi Porta. *Handbook of Research Methods and Applications in Happiness and Quality of Life.* Edward Elgar Publishing, 2016.

[132] Ons:labour market survey. `"http://bit.ly/2GdF9mm"`, 2018.

[133] Esener:the interactive survey dashboard. `"https://osha.europa.eu/en/surveys-and-statistics-osh/esener/2014"`, 2014.

[134] Hansen Schwartz, Johannes Eichstaedt, Margaret Kern, Lukasz Dziurzynski, Richard Lucas, Megha Agrawal, Gregory Park, Shrinidhi Lakshmikanth, Sneha Jha, Martin Seligman, and Lyle Ungar. Characterizing geographic variation in well-being using tweets, 2013. URL `https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6138`.

[135] Eszter Bokányi, Zoltán Lábszki, and Gábor Vattay. Prediction of employment and unemployment rates from twitter daily rhythms in the us. *EPJ Data Science*, 6(1):14, 2017.

[136] Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. A system for real-time Twitter sentiment analysis of 2012 u.s. presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, ACL '12, pages 115–120, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=2390470.2390490`.

[137] Andrea Ceron, Luigi Curini, Stefano M Iacus, and Giuseppe Porro. Every tweet counts? how sentiment analysis of social media can improve our knowledge of citizens political preferences with an application to italy and france. *New Media & Society*, 16(2):340–358, 2014. doi: 10.1177/1461444813480466. URL `https://doi.org/10.1177/1461444813480466`.

[138] Peter Gabrovek, Darko Aleksovski, Igor Mozeti, and Miha Grar. Twitter sentiment around the earnings announcement events. *PLOS ONE*, 12(2):1–21, 02 2017. doi: 10.1371/journal.pone.0173151. URL `https://doi.org/10.1371/journal.pone.0173151`.

[139] Marco Bastos and Dan Mercea. Parametrizing Brexit: mapping Twitter political space to parliamentary constituencies. *Information, Communication & Society*, 21(7):921–939, 2018. doi: 10.1080/1369118X.2018.1433224. URL `https://doi.org/10.1080/1369118X.2018.1433224`.

[140] Clare Llewellyn and Laura Cram. Brexit? Analyzing opinion on the uk-eu referendum within twitter. 3 2016.

[141] Frederico Caroli Siegfried Handschuh Ross McDermott Manel Zarrouk Manuela Hrlimann Brian Davis Tobias Daudert Malek Ben Khaled David Byrne Sergio Fernndez Angelo Cavallini Laurentiu Vasiliu, Andr Freitas. In or Out? Realtime monitoring of BREXIT sentiment on Twitter. SEMANTiCS 2016 CEUR Workshop Proceedings, 2016. URL `http://hdl.handle.net/10379/6239`.

[142] Amit Agarwal, Ritu Singh, and Durga Toshniwal. Geospatial sentiment analysis using twitter data for uk-eu referendum. *Journal of Information and Optimization Sciences*, 39(1):303–317, 2018. doi: 10.1080/02522667.2017.1374735. URL `https://doi.org/10.1080/02522667.2017.1374735`.

[143] Mauro Coletto, Andrea Esuli, Claudio Lucchese, Cristina Ioana Muntean, Franco Maria Nardini, Raffaele Perego, and Chiara Renso. Perception

of social phenomena through the multidimensional analysis of online social networks. *Online Social Networks and Media*, 1:14 – 32, 2017. ISSN 2468-6964. doi: https://doi.org/10.1016/j.osnem.2017.03.001. URL http://www.sciencedirect.com/science/article/pii/S246869641630009X.

[144] Robert M Groves, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer, and Roger Tourangeau. *Survey methodology*, volume 561. John Wiley & Sons, 2011.

[145] Lilli Japec, Frauke Kreuter, Marcus Berg, Paul Biemer, Paul Decker, Cliff Lampe, Julia Lane, Cathy ONeil, and Abe Usher. Big data in survey research:aapor task force report. *Public Opinion Quarterly*, 79(4):839–880, 2015. doi: 10.1093/poq/nfv039. URL http://dx.doi.org/10.1093/poq/nfv039.

[146] Theresa A Pardo. Making data more available and usable: A getting started guide for public officials. *Presentation at the Privacy, Big Data and the Public Good Book Launch*, 16, 2014.

[147] Helen Roberts, Jon Sadler, and Lee Chapman. Using Twitter to investigate seasonal variation in physical activity in urban green space. *Geo: Geography and Environment*, 4(2):e00041. doi: 10.1002/geo2.41. URL https://rgs-ibg.onlinelibrary.wiley.com/doi/abs/10.1002/geo2.41.

[148] A. Aue and M. Gammon. Customizing Sentiment Classifiers to New Domains: a Case Study. In *Microsoft Research*, 2005.

[149] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. *Journal Computational Linguistics*, 37, 2011.

[150] J. Fiaidhi, O. Mohammed, S. Mohammed, S. Fong, and T. H. Kim. Opinion mining over twitterspace: Classifying tweets programmatically using the R approach. In *ACM Int. Conf. Digit. Inf. Manag. (ICDIM 2012)*, 2012.

[151] Luís MA Bettencourt. The origins of scaling in cities. *science*, 340(6139): 1438–1441, 2013.

[152] Xiaodong Cao, Piers MacNaughton, Zhengyi Deng, Jie Yin, Xi Zhang, and Joseph G. Allen. Using Twitter to better understand the spatiotemporal patterns of public sentiment: A case study in Massachusetts, USA. *International Journal of Environmental Research and Public Health*, 15(2), 2018. ISSN 1660-4601. URL http://www.mdpi.com/1660-4601/15/2/250.

[153] Ordanance Survey, 2017. URL https://www.ordnancesurvey.co.uk/.

[154] Europe INSPIRE Directive. Infrastructure for spatial information in europe, 2017. URL https://inspire.ec.europa.eu/.

[155] UK INSPIRE. The uk inspire regulations, 2009. URL https://data.gov.uk/inspire.

[156] HM Land Registry. Inspire index polygons spatial data, 2014. URL https://www.gov.uk/guidance/inspire-index-polygons-spatial-data.

[157] HM Land Registry commercial services. Commercial and corporate ownership data, 2014. URL https://www.gov.uk/guidance/commercial-and-corporate-ownership-data.

[158] Twitter. Twitter developer documentation, 2017. URL https://dev.twitter.com/overview/api/tweets.

[159] Companies House. Free company data product, 2017. URL http://download.companieshouse.gov.uk/en_output.html.

[160] ONS-Office Of National Statistics. UK Standard Industrial Classification of economic activities 2007. URL https://www.ons.gov.uk/file?uri=/methodology/classificationsandstandards/ukstandardindustrialclassificationofeconomicactivities/uksic2007/uksic2007web.pdf.

[161] Anthony Brew, Derek Greene, Daniel Archambault, et al. Deriving insights from national happiness indices. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 53–60. IEEE, 2011.

[162] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

[163] K. Weller, Bruns, Burgess A., M. J., Mahrt, and C Puschmann. *Twitter and Society*, chapter Sentiment Analysis and Time Series with Twitter. Peter Lang, New York, NY., 2014.

[164] BBC. 2012 as job creator, 2009. URL http://news.bbc.co.uk/1/hi/uk/7831847.stm.

[165] Telegraph. London 2012 olympics-the olympic stadium, 2011. URL http://www.telegraph.co.uk/finance/london-olympics-business/8641977/London-2012-Olympics-The-Olympic-Stadium-made-in-Britain.html.

[166] Telegraph. Housing boom, 2012. URL http://bit.ly/2BoAZWj.

[167] ONS - Office of National Statistics. Employee jobs by industry sector - london, 2017. URL http://www.nomisweb.co.uk/.

[168] Sabine AE Geurts, Debby GJ Beckers, Toon W Taris, Michiel AJ Kompier, and Peter GW Smulders. Worktime demands and work-family interference: Does worktime control buffer the adverse effects of high demands? *Journal of Business Ethics*, 84(2):229–241, 2009.

[169] Marianna Virtanen, Stephen A Stansfeld, Rebecca Fuhrer, Jane E Ferrie, and Mika Kivimäki. Overtime work as a predictor of major depressive episode: a 5-year follow-up of the whitehall ii study. *PloS one*, 7(1):e30719, 2012.

[170] Zoltn Lbszki Eszter Boknyi and Gbor Vattay. Prediction of employment and unemployment rates from Twitter daily rhythms in the us. *EPJ Data Science*, 6(1):14, 2017. doi: 10.1140/epjds/s13688-017-0112-x. URL https://doi.org/10.1140/epjds/s13688-017-0112-x.

[171] Francis Green. Work intensification, discretion, and the decline in well-being at work. *Eastern Economic Journal*, 30(4):615–625, 2004. ISSN 00945056, 19394632. URL http://www.jstor.org/stable/40326152.

[172] Stephan Meier and Alois Stutzer. Forthcoming in economica is volunteering rewarding in itself?, 2006.

[173] Ingrid Schoon, Leeni Hansson, and Katariina Salmela-Aro. Combining work and family life. *European Psychologist*, 10(4):309–319, 2005. doi: 10.1027/1016-9040.10.4.309. URL https://doi.org/10.1027/1016-9040.10.4.309.

[174] Lonnie Golden and Barbara Wiens-Tuers. To your happiness? extra hours of labor supply and worker well-being. *The Journal of Socio-Economics*, 35(2):382 – 397, 2006. ISSN 1053-5357. doi: https://doi.org/10.1016/j.socec.2005.11.039. URL http://www.sciencedirect.com/science/article/pii/S1053535705001691. The Socio-Economics of Happiness.

[175] Andreas Knabe and Steffen Rtzel. Income, happiness, and the disutility of labour. *Economics Letters*, 107(1):77 – 79, 2010. ISSN 0165-1765. doi: https://

doi.org/10.1016/j.econlet.2009.12.032. URL http://www.sciencedirect.com/science/article/pii/S0165176509004364.

[176] Alison L. Booth and Jan C. Van Ours. Job satisfaction and family happiness: The part time work puzzle. *The Economic Journal*, 118(526):F77–F99. doi: 10.1111/j.1468-0297.2007.02117.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0297.2007.02117.x.

[177] Mick Couper. Is the sky falling? new technology, changing media, and the future of surveys. *Survey Research Methods*, 7(3):145–156, 2013. ISSN 1864-3361. URL https://ojs.ub.uni-konstanz.de/srm/article/view/5751.

[178] Downloading JSON from Twitter. URL http://support.gnip.com/articles/downloading-hpt-files.html.

[179] Kernel smoother. URL https://en.wikipedia.org/wiki/Kernel_smoother.

[180] Ourania Kounadi, Thomas J. Lampoltshammer, Elizabeth Groff, Izabela Sitko, and Michael Leitner. Exploring twitter to analyze the publics reaction patterns to recently reported homicides in london. *PLOS ONE*, 10(3):1–17, 03 2015. doi: 10.1371/journal.pone.0121848. URL https://doi.org/10.1371/journal.pone.0121848.

[181] Mario Callegaro and Yongwei Yang. *The Role of Surveys in the Era of "Big Data"*, pages 175–192. Springer International Publishing, Cham, 2018. ISBN 978-3-319-54395-6. doi: 10.1007/978-3-319-54395-6_23. URL https://doi.org/10.1007/978-3-319-54395-6_23.

[182] Brendan Collins. Big data and health economics: Strengths, weaknesses, opportunities and threats. *PharmacoEconomics*, 34(2):101–106, Feb 2016. ISSN 1179-2027. doi: 10.1007/s40273-015-0306-7. URL https://doi.org/10.1007/s40273-015-0306-7.

[183] Ines Amaral. *Computational Social Sciences*, pages 1–3. Springer International Publishing, Cham, 2017. ISBN 978-3-319-32001-4. doi: 10.1007/978-3-319-32001-4_41-1. URL https://doi.org/10.1007/978-3-319-32001-4_41-1.

[184] Bo Han, Paul Cook, and Timothy Baldwin. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49:451–500, 2014.

[185] Marco Pennacchiotti and Ana-Maria Popescu. A machine learning approach to twitter user classification. *Icwsm*, 11(1):281–288, 2011.

[186] Kazushi Ikeda, Gen Hattori, Chihiro Ono, Hideki Asoh, and Teruo Higashino. Twitter user profiling based on text and community mining for market analysis. *Knowledge-Based Systems*, 51:35–47, 2013.

[187] Steffen Rtzel. Revisiting the neoclassical theory of labour supply Disutility of labour, working hours, and happiness. FEMM Working Papers 09005, Otto-von-Guericke University Magdeburg, Faculty of Economics and Management, February 2009. URL `https://ideas.repec.org/p/mag/wpaper/09005.html`.

[188] Sweden sees benefits of six-hour working day. `https://www.theguardian.com/world/2017/jan/04/sweden-sees-benefits-six-hour-working-day-trial-care-workers`, 2017.

[189] Emails while commuting 'should count as work'. 2018. URL `https://www.bbc.co.uk/news/education-45333270`.

[190] Felipe Bravo-Marquez, Marcelo Mendoza, and Barbara Poblete. Combining strengths, emotions and polarities for boosting twitter sentiment analysis. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 2. ACM, 2013.

[191] Hans-Jürgen Engelbrecht. Natural capital, subjective well-being, and the new welfare economics of sustainability: Some evidence from cross-country regressions. *Ecological Economics*, 69(2):380–388, 2009.

[192] Richard Florida, Charlotta Mellander, and Peter J Rentfrow. The happiness of cities. *Regional Studies*, 47(4):613–627, 2013.