

UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

Electronics and Computer Science

**Data Mining Locative Thematic Narratives: Analysing Twitter and
POI Data to Extract Precise Spatial Themes**

by

Nicholas C. Bennett

Thesis for the degree of Doctor of Philosophy

June 2020

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

Electronics and Computer Science

Doctor of Philosophy

DATA MINING LOCATIVE THEMATIC NARRATIVES: ANALYSING TWITTER
AND POI DATA TO EXTRACT PRECISE SPATIAL THEMES

by **Nicholas C. Bennett**

This thesis produces a novel framework for constructing thematic narratives of space. Existing research into extracting information from Twitter data is fraught with inconsistencies and a general lack of attention to metadata. This thesis conducts five experiments to delve deeply into how to extract precise location-based themes from Twitter data, with a focus on under-researched metadata fields and exploration of external APIs to enrich the location data of previously assumed ‘precisely’ geolocated tweets. The analysis is based on Twitter data from June 2016 to August 2018. A key argument is made for the analysis of third-party sources and the stratification of tweet granularity to better understand the scale of location-based narratives. These experiments form a framework that is tested in a case study, then validated and evaluated with qualitative interviews with HM Treasury. This framework is the main contribution of the thesis.

Contents

Declaration of Authorship	xiii
Acknowledgements	xv
1 Introduction	3
1.1 Narratives & Social Media	4
1.1.1 Applications of Twitter Data	5
1.2 Volunteered Geographical Information (VGI)	6
1.3 The Twitter API	7
1.4 Research Goals	8
1.4.1 Research Questions	8
1.5 Thesis Methodology	9
1.6 Thesis Structure	11
2 Literature Review	13
2.1 Narrative Analysis	13
2.1.1 Formalism	13
2.1.2 Structuralism	14
2.1.3 Thematic Narratives	14
2.1.4 Space and Place	14
2.1.5 Placelessness	16
2.1.6 Location-based Storytelling	17
2.1.7 Location as Identity	17
2.1.8 Theoretical Perspectives	18
2.2 Topic Modelling	19
2.2.1 Introduction	19
2.2.2 Latent Semantic Analysis (LSA)	20
2.2.3 Probabilistic LSA (pLSA)	21
2.2.4 Latent Dirichlet Allocation (LDA)	21
2.2.5 K-Means	22
2.2.6 Applications of LDA	22
2.2.7 LDA on Twitter Data	23
2.2.8 Keyword Matching	26
2.3 Spatial Analyses	27
2.3.1 Introduction	27
2.3.2 Kernel Density Estimation	29
2.3.3 Getis Ord and Moran's I	29

2.3.4	Location and Social Media Data	30
2.3.4.1	Application of KDE	31
2.3.4.2	Resolving Twitter Locations	33
2.4	Conclusions from the Literature	34
3	Developing a Framework for Location-Based Narrative Extraction	35
3.1	Experiment 1: Prototype Event Classification Model	36
3.1.1	Methodology	36
3.1.1.1	Pre-processing	37
3.1.2	Results	38
3.1.2.1	BBC News Articles	39
3.1.2.2	Parallel Coordinate Representations	39
3.1.3	Discussion	43
3.1.4	Concluding Remarks	43
3.2	Experiment 2: Thematic Kernel Density Estimation	44
3.2.1	Methodology	46
3.2.1.1	Thematic Tagging	46
3.2.1.2	Kernel Density Estimation	47
3.2.2	Results	48
3.2.2.1	Comparison with Second Dataset	50
3.2.3	Discussion	52
3.2.4	Concluding Remarks	54
3.3	Experiment 3: Stratifying Location Resolutions	55
3.3.1	Methodology	55
3.3.1.1	Pre-processing	56
3.3.1.2	Source Analysis	56
3.3.1.3	Stratifying Coordinate Spaces	58
3.3.2	Results	58
3.3.2.1	Applying Census Data	60
3.3.3	Discussion	62
3.3.4	Concluding Remarks	62
3.4	Experiment 4: Resolving Third-Party POIs	63
3.4.1	Methodology	63
3.4.1.1	Analysing Third-Party URLs	64
3.4.1.2	Technical Approach	64
3.4.2	Discussion	66
3.4.3	Concluding Remarks	67
3.5	Experiment 5: API Calling	67
3.5.1	Methodology	68
3.5.2	Calling the APIs	68
3.5.3	Formulating Accuracies	69
3.5.4	Results	70
3.5.5	Foursquare Checkin Resolve	71
3.5.6	API Evaluation	71
3.5.7	Discussion	73
3.5.8	Concluding Remarks	74
3.6	The Proposed Framework	74

4	Case Study: HM Treasury	77
4.1	Introduction	77
4.1.1	Existing Methods	78
4.1.1.1	Gross Domestic Project	78
4.1.1.2	Gross Value Added	79
4.1.1.3	Other Datasets	79
4.1.2	Real-Time Economic Indicators	80
4.1.2.1	Price Indices	80
4.2	Previous Work	81
4.2.1	Economic Representivity of Twitter	81
4.3	Methodology	82
4.3.1	Data Collection	83
4.3.2	Applying the Framework	84
4.3.2.1	Tweet Cleaning	85
4.3.2.2	Cluster Analysis	85
4.3.2.3	Keyword Matching	86
4.4	Results	87
4.4.1	Source Analysis	88
4.4.2	Cluster Analysis	89
4.4.3	Keyword Matches	90
4.4.3.1	Keyword Set Maps	92
4.4.4	LDA	93
4.4.4.1	LDA Results	93
4.4.4.2	LDA Resolutions	96
4.5	Summary	99
4.5.1	Key Outcomes	100
5	Evaluative Interviews	103
5.1	Interview Design	103
5.1.1	Coding Theory	104
5.2	Interview Method	105
5.2.1	Questions and Codes	106
5.2.2	Transcribing and Coding	106
5.2.2.1	Interview Overview	107
5.3	Evaluating the Framework	108
5.3.1	Importance of Location Analysis	109
5.3.2	Importance of Timeliness	111
5.3.3	Spatial and Temporal Scalability	112
5.3.4	Usefulness of the Framework	113
5.4	Suggested Improvements and Future Work	115
5.5	Summary	117
5.5.1	Outcomes	118
5.5.2	Advice and Improvements	119
6	Conclusions	121
6.1	Overall Thesis Summary	121
6.1.1	Addressing the Research Questions	122

6.2	Contribution	125
6.2.1	Twitter Cleaning Methodology	126
6.2.1.1	Source Analysis	126
6.2.1.2	Cluster Analysis	126
6.2.1.3	Topic Analysis	127
6.2.2	Limitations	127
6.2.2.1	Spatial Analyses	128
6.2.2.2	Topic Modelling	128
6.3	Future Work	129
6.3.1	Future Applications	131
6.4	Final Thoughts	132
A	Appendix A - Tweet Ethics Application	135
B	Appendix B - Treasury Interview Ethics Application	147
C	Appendix C - LDA Analysis on Treasury Tweets	165
C.1	Concluding Remarks	169
	References	171

List of Figures

1.1	An example of a tweet. It shows the screen name, username, text entry, timestamp, geographical location and first- or third-party origin of the tweet.	7
1.2	A simplified flowchart showing how the Twitter API is accessed.	8
1.3	Flowchart depicting the approach of the thesis to create a flowchart for identifying and analysing locative thematic narratives.	9
3.1	Graph representing total tweets over time for the 9-20 June 2016 Southampton tweet dataset, with retweets included. Figure generated using the Python module Matplotlib.	39
3.2	Application of the remain (Brexit) debate event to the model. Population here refers to total audience size — viewership was 3M.	41
3.3	Model with representation of the Orlando vigil events, a vigil in memory of those shot during a night club attack in Orlando, Florida, USA.	42
3.4	Model with representation of the shooting of MP Jo Cox.	42
3.5	Map of Southampton, UK, with workplace zone classification data obtained from ONS. Used as reference to specific locations within the city.	45
3.6	Cleaned KDE heatmap of geotagged tweets, centred on Southampton, UK.	48
3.7	Cleaned KDE heatmap of geotagged Southampton tweets, tagged as (a) commerce and (b) entertainment.	49
3.8	Maps of (a) Eastleigh COWZ classification, (b) KDE of cleaned tweets, (c) KDE of commerce tweets, (d) KDE of entertainment tweets.	51
3.9	Maps of Winchester with (a) the COWZ classification, (b) a zoomed in area to better show the geography, (c) KDE of commerce tweets, (d) KDE of entertainment tweets.	52
3.10	Map of Southampton showing the tweet clusters comprising of fewer than 4 tweets, a total of 13,465 tweets.	59
3.11	Map of Southampton showing the clusters comprising more than 4 tweets, a total of 96,518 tweets.	59
3.12	Map showing the clusters classified as neighbourhoods. The shading relates to the population count within the Output Area.	61
3.13	Map showing the clusters of POIs, emphasising how many there are in the dataset that would otherwise have been overlooked. The shading relates to the population count within the Output Area.	61
3.14	Visual workflow of methodology chapter.	75
4.1	Flowchart depicting the application of the framework to this specific case study to extract locative economy-themed narratives.	83

4.2	Map of the UK showing the study area in red. The counties in (b) from bottom-left clockwise: Dorset, Wiltshire, Hampshire, Isle of Wight.	84
4.3	Comparison map of the cleaned Twitter dataset collected from the 1st November 2016 until 4th August 2018 (a) and the same dataset with ‘ClusterOfOne’ tweets removed (b).	86
4.4	Map of all the raw tweets compared with the cleaned tweets within the southern UK counties related to the 26-mile radius Twitter search collected from the 1st November 2016 until the 4th August 2018.	89
4.5	Map of tweets within Southampton collected from 1st November 2016 to 4th August 2018, with KDE applied. (a) shows all the tweets, (b) shows the tweets that do not belong to any cluster, (c) shows tweets at POI resolution, (d) shows tweets at non-POI resolution.	91
4.6	Map showing all the cleaned geolocated tweets within Southampton from 1st November 2016 to 4th August 2018 tagged with the Harvard (a), Thesaurus (b) and Treasury (c) keyword tags.	92
4.7	The output of the LDA analysis on the Thesaurus keyword set, visualised using the Python module LDAVis and with the 9th topic highlighted. . . .	94
4.8	The output of the LDA analysis on the Harvard keyword set, visualised using the Python module LDAVis and with the first topic highlighted. . . .	95
4.9	The output of the LDA analysis on the Treasury keyword set, visualised using the Python module LDAVis and with the first topic highlighted. . . .	96
4.10	The output of the LDA analysis on the Thesaurus keyword set classified as originating from a non-POI, visualised using the Python module LDAVis and with the first topic highlighted. Image represents 294 tweets.	97
4.11	The output of the LDA analysis on the Harvard keyword set classified as originating from a POI, visualised using the Python module LDAVis and with the fifth topic highlighted. Image represents 5,618 tweets.	98
4.12	The output of the LDA analysis on the Treasury keyword set classified as originating from a POI, visualised using the Python module LDAVis and with the third topic highlighted. Image represents 77 tweets.	98
C.1	The output of the LDA analysis on the Thesaurus keyword set classified as originating from a POI, visualised using the Python module LDAVis and with the forth topic highlighted. Image represents 725 tweets.	165
C.2	The output of the LDA analysis on the Thesaurus keyword set classified as precise, visualised using the Python module LDAVis and with the forth topic highlighted. Image represents 1,019 tweets.	166
C.3	The output of the LDA analysis on the Harvard keyword set classified as originating from a non-POI, visualised using the Python module LDAVis and with the forth topic highlighted. Image represents 1,782 tweets.	167
C.4	The output of the LDA analysis on the Harvard keyword set classified as precise, visualised using the Python module LDAVis and with the forth topic highlighted. Image represents 7,400 tweets.	167
C.5	The output of the LDA analysis on the Treasury keyword set classified as originating from a non-POI, visualised using the Python module LDAVis and with the forth topic highlighted. Image represents 15 tweets.	168
C.6	The output of the LDA analysis on the Treasury keyword set classified as precise, visualised using the Python module LDAVis and with the forth topic highlighted. Image represents 92 tweets.	169

List of Tables

3.1	Table showing tweet numbers before and after cleaning. Mean relates to the tweets per user and SD stands for standard deviation.	37
3.2	Table showing the most prevalent LDA topics within the 9-20 June tweet dataset, with retweets removed. The associated events obtained from manually searching the BBC News article dataset.	38
3.3	Table showing a sample of main BBC News articles from 9-20 June 2016 and how many unique tweets collected from the Southampton bounding circle during the same period matched the keywords. Only tweets explicitly matching the keywords were counted. This avoided over-representation of popular topics that last several days.	40
3.4	Table showing tweet numbers before and after cleaning for the main dataset.	46
3.5	Table showing the top 15 synonyms returned from the online thesaurus for the two themes. In total there were 135 terms for Commerce and 433 for Entertainment.	48
3.6	Table showing tweet numbers before and after retweets and noisy users were removed for ‘Eastleigh’ and ‘Winchester’ keyword searches. Note: these are global values. There are mentions of ‘Eastleigh’ in Kenya and ‘Winchester’ in the USA.	50
3.7	Breakdown of top 10 Sources	57
3.8	Breakdown of top 10 Geo Sources	57
3.9	Example of geo sources with 5 or fewer users (text from human users anonymised).	58
3.10	Tweet sources before and after resolving POI URLs. Some were private, had no venue data or venue coordinates and are grouped as “Error”. +Text means if the URL-scraped text was longer than the original tweet text and +Geo means if the scraped POI coordinates differed from the original ones.	65
3.11	Text-based POI matches between each data source.	65
3.12	Table showing the results of matching the 7,039 tweet clusters to the three APIs, along with how many responses were within 25 metres of the original tweet and of those responses how many were POIs. Tweets above 25m are automatically classified as neighbourhoods (N).	70
3.13	Table showing the total number of Foursquare checkins within the dataset, along with how many non-Foursquare tweets shared the same tweet co-ordinates and how many Foursquare venue coordinates matched existing tweets.	70
3.14	Table showing the top 5 checkin types from Foursquare’s Checkin Resolve API.	70
3.15	The results of the API classification evaluation, including adjustment. . .	72

3.16	A detailed explanation of the framework.	76
4.1	List of keyword sets, their unique terms and subsequent matches in the raw, cleaned and Southampton area datasets.	87
4.2	Characteristics of the final dataset, each row a subset of the previous one. Here ‘Boundary Box’ refers to any user who has only allowed for their tweets to be at neighbourhood or city level, therefore do not produce any coordinate data aside from the overarching polygon. ‘Native check-ins’ refers to tweets from native Twitter apps that do not have precise coordinates but do have matching bounding box coordinates.	87
4.3	Top 10 sources that contain geolocated tweets ranked by tweet to user ratio. After removing sources with fewer than 5 users and 10 tweets, 26 total sources remained.	88
4.4	Top five sources ranked by tweet count. Location labels obtained by manually searching Google Maps and OpenStreetMaps.	89
5.1	Table showing the codes derived from the questions and interviews.	107

Declaration of Authorship

I, **Nicholas C. Bennett** , declare that the thesis entitled *Data Mining Locative Thematic Narratives: Analysing Twitter and POI Data to Extract Precise Spatial Themes* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published as: ([Bennett et al., 2016](#)), ([Bennett et al., 2017b](#)), ([Bennett et al., 2017a](#)) and ([Bennett et al., 2018](#)).

Signed:.....

Date:.....

Acknowledgements

Thanks to Ordnance Survey and EPSRC for sponsoring this PhD process. The PhD began in September 2015 after a year-long MSc conversion course. The MSc gave us the skills and background knowledge of how Web Science evolved and is applied, which was then taken forward for a further 4.5 years of PhD study, including 3 funded years, a nominal year and 6 months of corrections.

Thanks to my supervisors Dr. David Millard, Professor David Martin and Dr. Jeremy Morley for their expertise and patience, especially when guiding me through the switch from event detection to thematic narratives. Thanks to my examiners who have given me constructive feedback throughout the process. Thanks to my fellow DTC/CDT colleagues for their companionship and advice. Thanks to my bay buddies Sarah and Tom. Thanks to my fiancée Sasha who had to listen to all of this for over four years. Thanks to all of my friends for their help and advice, especially Briony, Nikko, Will, Rob, Gemma, Louise, Allison, Ed, Sam and the wider Web and Internet Science (WAIS) research group.

Aside from the 4 previously published works mentioned in this thesis, there are no current plans for further publications.

I would like to dedicate this PhD to Dr. Pouria Amirian, a former supervisor from Ordnance Survey who passed away in 2017.

Todo list

Chapter 1

Introduction

Since the creation and widespread adoption of the World Wide Web in 1989 ([Berners-Lee and Fischetti, 1999](#)) to the early 2000s, users have been able to digest digital data at scale. From the early 2000s and the rise of social media, users have been able to not only digest but also create highly personalised data with which they can share their everyday lives, interests, activities and connect to others with similar (or totally different) lifestyles. This user-centric social media phenomenon has connected millions of people across the world, creating vast data streams of geographically influenced social activity. Contemporaneously, GPS- and Web-enabled smartphones became prevalent with a significant proportion of the UK population. As of 2018, 89% of all UK adults use the internet at least weekly, with 86% of this sample using it at least daily and 78% accessing the Web via smartphones ([Office for National Statistics, 2018a](#)). According to a further ONS report, 66% of all UK adults use social media with 96% of 16-24, 88% of 25-34 and 83% of 35-44 year olds using it before dropping sharply to 68% of 45-54 year olds until just 27% of 65 and over use it ([Office for National Statistics, 2018c](#)). This therefore means that from a UK population of around 66 million, over 53 million adults use the internet with over 43 million social media users ([Office for National Statistics, 2018b](#)).

As social media allow its users to post about their interests, location and activities, there is potential to capture the online social information of 43 million people in the UK. Currently, there is a plethora of different social media platforms within which people generate content, therefore to obtain all this information at the same time would require both a sophisticated Web scraper and the participating platforms and users allowing their information to be collected. Despite the benefits this would bring to social research and understanding, many of the platforms' Application Program Interfaces (APIs) are closed by default, such as Facebook and Instagram, meaning information either cannot be accessed at all or only the creator's profile can be accessed. However, other platforms are open by default, such as Twitter, allowing researchers access to an unprecedented amount of real-time social data. Twitter is one of the largest social media platforms

with an estimated 335 million monthly global users. While the number of UK users is not released by Twitter itself, some estimate there are 13 million registered UK accounts (EMarketer, 2015; Omnicoreagency, 2018), representing 30.2% of all UK adults who use social media.

The ability to understand a population is of interest to a wide variety of commercial agencies and government organisations to improve marketing campaigns, predict consumer behaviour or to plan for policy changes in future years. The significant amount of data produced is also of interest to mapping agencies such as Ordnance Survey, who sponsored this PhD. Large quantities of geolocated information help to populate social-centric maps and give semantic meaning to spaces rather than the standard geographic representation. These data are also of interest to government agencies who wish to understand economic impact at scale, as social media users will post their reflections on their habits and surroundings, both of which are influenced by the current economic climate.

This thesis explores the value of Twitter for discovering and analysing narratives of social space. These narratives, generated by individual users, have the potential to provide key insights into real-time location-based activities at scale. This approach is attractive to policy makers in government organisations due to its cheap implementation, low up-keep cost and passive collection.

1.1 Narratives & Social Media

Communication is a key part of human society. Narratives are defined as a sequence of events or story that we convey either to ourselves or others to help make sense of our social and physical interactions with our environments (Farrow et al., 2015). Traditionally, narrative study (narratology) is in fictional literature, where this approach is different; the distance between the author and audience creates an impersonal feel, such that the reader knows the narrative is constructed regardless of its believability (Fludernik, 2009), exemplified by the reader being able to pick up a book after a break and continuing the story. This position changes dramatically with the adoption of social media. The narrative is created in real-time by authors contributing their own micro-narratives about their experiences and environments. This narrative is ever-changing such that a reader cannot directly apply narratology with the same passiveness as with literature. Social media are both immediate and networked (e.g. followers, groups, propagated messages), thus to sufficiently understand the emerging narratives, especially in the era of Big Data, one must apply computational approaches.

The ability for social media to broadcast the author's stories in real-time, for them to be networked with millions of others and for interested parties to access and analyse this data in real-time makes narratology and social media key assets for understanding

human activity and behaviour at scale. Additionally, the “high volume” (Xu et al., 2016), immediacy, often open access nature and broad geographical spread of social media make them key resources for research and information, especially due to the relatively low cost of data acquisition (Gu et al., 2016) in terms of computing power, skill and time required to obtain the data.

As social media have become popular a vast quantity of textual representations of activities, events and experiences are being produced by users. These narratives are increasingly used by academic and industry researchers to understand the context of and behaviours in social interactions (Tamburrini et al., 2015). As social media are often used to report on individual experiences, they are a rich source of narrative information at personal and societal levels, both in real-time and historically. Social media platforms are increasingly seen as bridging the divide between databases and personal stories (Farrow et al., 2015). As with Web 2.0 constructs these narratives are dynamic and change over time, therefore constant study is required to understand and track the changing nature of narratives.

Changing spatial narratives have been investigated in regards to political events (Himelboim et al., 2016; Tremayne, 2014), and tracking environmental disasters (Sakaki et al., 2010; Zielinski et al., 2013), but relating to general societal change these studies are lacking. The application of archaeological or historical perspectives is useful in this case; Twitter and other social media are creating digital artefacts that intricately map societies around the world, contributing to a vast dataset of current human activity. The ability for Twitter to reveal societal practices at scale makes it an appropriate data source for analysing narrative information within various locations, herein referred to as locative thematic narratives.

1.1.1 Applications of Twitter Data

Despite the popularity of image-based social media platforms¹, posts still predominantly contain text in some form (photo descriptions, links and general text). Current opinions, trends and market habits are therefore voluntarily contributed at a large scale without the need for complex image processing software. With the development of smartphones powerful enough to run Global Positioning Systems (GPS) technology, users are able to post these key insights alongside geographical location references. There exist other techniques for establishing location, such as cell tower triangulation or WiFi router placements, though the range of GPS satellites allows smartphones to post location data almost anywhere in the world, provided the user has line of sight to at least 4 satellites (Wing et al., 2005; Bauer, 2013). The digital micro-narrative is enriched with these geotags that contributes both to a societal narrative as well as pinpoints these emerging

¹Instagram hit 1 billion monthly users in 2018: <https://techcrunch.com/2018/06/20/instagram-1-billion-users/> [Accessed 5 April 2019].

stories to specific areas. This Volunteered Geographic Information (VGI) (Goodchild, 2007) has emerged over time, with various technological limitations that are important to understand by researchers with interests in locative thematic narratives.

1.2 Volunteered Geographical Information (VGI)

Since the 2000s, smartphones have been powerful enough to communicate with GPS satellites to relay their geographic location at any given time. These metadata can be shared with social media platforms to allow the user to create a post that is linked to their location, a process called Volunteered Geographic Information (VGI). The term VGI was coined by Goodchild (2007) to describe a subset of Web-based geographic content created and uploaded by Web users. This differed from previous location data originating from governments or private companies (Quesnot and Roche, 2015) and gave rise to services like OpenStreetMap (OSM) and Foursquare². Location data were no longer restricted behind paywalls and could be consumed by the average Web user. VGI has been used by academics to understand location-based activities and have allowed for the development of early-warning systems for natural disasters like earthquakes (Sakaki et al., 2010) and floods (Smith et al., 2017). VGI also affords insights into social activities and is used by mobility pattern analysts to understand deprivation (Quercia and Saez, 2014), general temporal-based user activities (Hasan et al., 2013) and the activities of ‘mapping parties’ where users collaboratively contribute to open sources such as OSM (Mooney and Corcoran, 2014; Fritz et al., 2017). VGI has been adapted and integrated within social media platforms, allowing users to digitally check-in to certain establishments or upload text and photo with their corresponding geographic location.

On Facebook, profiles are private by default, therefore obtaining a user’s location data is only permitted if the user has allowed it. On Swarm, location data from a check-in (a user notifying the app that they are in a particular area) are semi-public by default, the researcher will need to create an account and install the app to see the data. These check-ins are merely snapshots and the whole premise of Swarm (and previously Foursquare) is to gamify location data and therefore this biases any potential conclusions about realistic mobility patterns. For Twitter, all tweets are public by default and can be seen without the need for an account, despite the current homepage³ not indicating this possibility. Location data can be applied to a tweet at the user’s discretion and can be at five distinct resolutions: precise (device location), point of interest (nearby monument or shop), neighbourhood (nearby neighbourhood cluster), city or country (both defined by an arbitrary centroid). While Twitter allows for the user to apply a location to each tweet at the time of composition, the user can also enable automatic geolocation on

²<http://www.foursquare.com> (Check-in feature spun out to Swarm in 2014.)

³<http://www.twitter.com> [Accessed 29 June 2018]

every tweet to one of the five degrees. The text and geolocation metadata are implicitly embedded in the tweet, thus providing rapid access for researchers (Mooney et al., 2016).

Twitter therefore creates both textual representations and geographically mappable individual and societal behaviours. Its open nature and popularity make it ideal for research. Naturally, any information gathered from Twitter comes with selection bias; the user creates their own content and thus shapes the impact this will have on their audience. This is true of any user-generated narrative. As social media in general, specifically smartphone-centric ones, are used by young to middle-aged adults more than the elderly this will also impact the sample bias as issues felt by the older generations will be less frequently represented. However, the sheer volume of information generated by Twitter users creates a rich dataset full of discoverable locative thematic narratives. While VGI is not discussed in great depth later in the thesis, this section has outlined the background of social media data, the properties of Twitter and its characteristics as a source of VGI which make it an appropriate source for this research.

1.3 The Twitter API



Figure 1.1: An example of a tweet. It shows the screen name, username, text entry, timestamp, geographical location and first- or third-party origin of the tweet.

An integral part of this research relies on understanding how the Twitter API works to create the tweet shown in figure 1.1. As emphasised by Halford et al. (2018), Twitter as a company has protected itself from hacks, DDoS (Direct Denial of Service) attacks and server overloads by carefully creating a system of APIs and data centres to handle content creation and retrieval. This networked system has corresponding impacts on the nature and quality of information requests. There has been research into how representative Twitter is of its own users (Mislove et al., 2011; Morstatter et al., 2014) as well as the extent to which researchers can access the API and retrieve data, a value assumed to be a maximum of 1% all traffic if using the Stream or Search APIs (Morstatter et al., 2013). Furthermore, the micro-blogs (referred to as tweets) that contain geolocation information is assumed to be around 1% of all tweets, though when specifically querying the Twitter API for a certain geographic area, the API has been shown to consistently

return up to 100% of all geolocated tweets (Morstatter et al., 2013). Until early 2018, all tweets were restricted to a maximum of 140 characters; during 2018, a slow roll-out of an extension to 280 characters⁴ was carried out, thus the latter stages of the tweet analyses include this changed limit. Despite the increased character limit, interest in attributing geolocation information was unchanged as geolocated tweets still comprised around 1% of tweets (Casadei and Lee, 2020). An example flowchart of how to access the Twitter API is shown in figure 1.2.

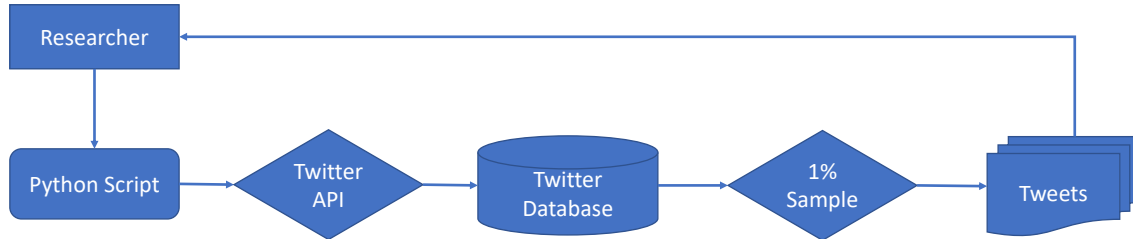


Figure 1.2: A simplified flowchart showing how the Twitter API is accessed.

1.4 Research Goals

Existing methods, covered in the next chapter, focus significant effort on summarising topics of conversation rather than expanding them into a locative thematic narrative that aligns more closely with natural themes. While topic modelling is a useful natural language processing step, and is often necessary on large datasets, most work stops there and does not take it further. This thesis aims to understand the strengths of topic modelling, spatial clustering and narratology techniques to create a unified framework for discovering locative thematic narratives. It will develop this framework by conducting five experiments that address topic modelling, spatial clustering and narratology. The thesis will then apply the framework to a case study with HM Treasury, the UK government’s economics department, to discover economy-themed locative narratives. The framework and case study achieve these goals by addressing three research questions.

1.4.1 Research Questions

RQ1: To what extent does a thematic approach afford a richer understanding of location-based activity than topic modelling alone?

Topic modelling summarises textual datasets to extract salient themes. However, this removes contextual data and creates results that often require subjective interpretation.

⁴https://blog.twitter.com/en_us/topics/product/2017/tweetingmadeeasier.html [Accessed 10 October 2018]

Work is needed to properly prepare the data for thematic extraction, for which this thesis contributes a validated approach.

RQ2: Can locative thematic narrative modelling be automated?

Text and spatial modelling can be computationally intensive and a challenge to automate without subjectivity. A framework can give structure to this combined approach, which is created through results from the five experiments outlined in the thesis.

RQ3: Can this approach be applied to a real-world situation with actionable results?

The previous two research questions contribute to a framework for extracting locative thematic narratives, shown below in figure 1.3. The third research question covers a real-world application of the framework which is then validated and evaluated by two UK government departments.

1.5 Thesis Methodology

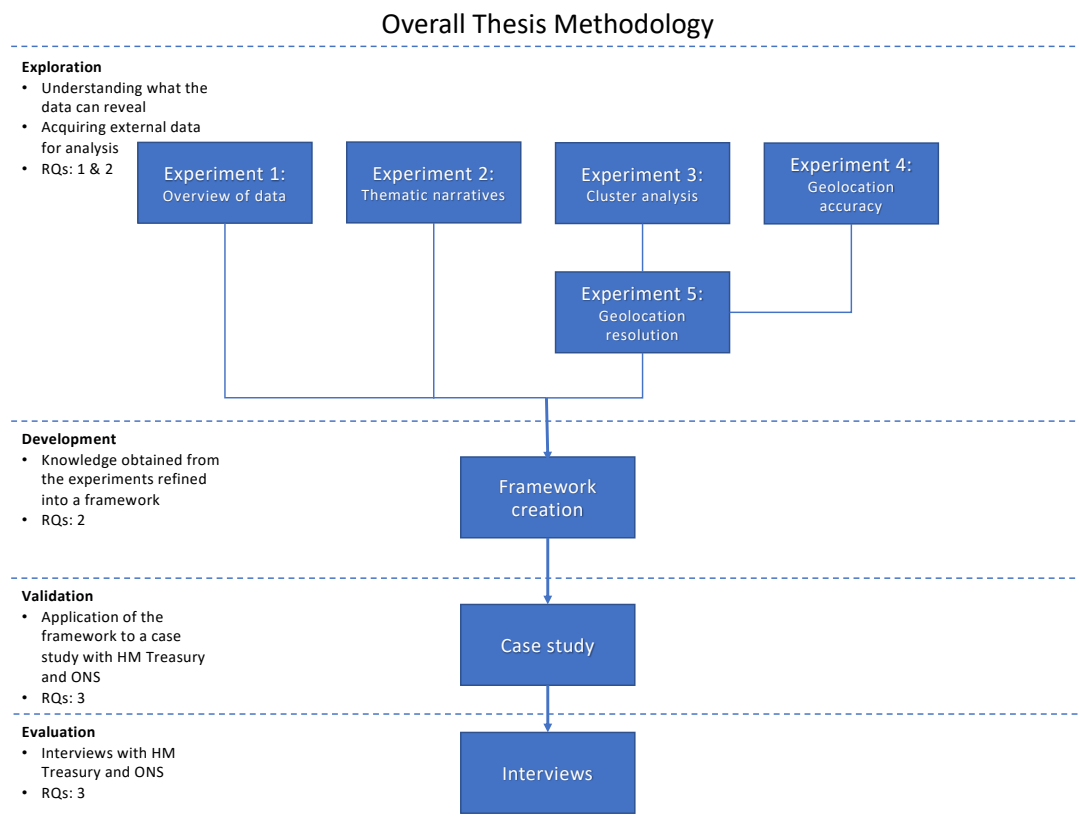


Figure 1.3: Flowchart depicting the approach of the thesis to create a flowchart for identifying and analysing locative thematic narratives.

Figure 1.3 gives an overview of the experiments and how they contribute towards a framework for locative thematic narrative analysis. As literature describing the obtainable properties from social media differ, it was appropriate to analyse these gaps in the research and build a holistic model of what the data can reveal. This model serves as an overview of how the thesis is structured. The first experiment revealed the challenges in extracting meaningful event data and guided the research towards a more inclusive narrative approach, further developed in the second experiment which extracted locative thematic narratives. General inconsistencies in the observed location data warranted further exploration into how geolocated tweets were created, as well as if they can be matched with address data to extract their original geographic resolution. Experiments 3, 4 and 5 showed that geolocated tweet data cannot be assumed accurate, and careful consideration is required when deriving meaningful locative thematic narratives.

New knowledge obtained through the exploratory experiments was condensed into a widely applicable framework to best analyse geolocated text data. The framework was validated through a case study with HM Treasury and the Office for National Statistics (ONS), two UK government departments with an interest in exploring novel approaches. The results and overall thesis methodology was evaluated through qualitative interviews with employees from both departments.

Part of Experiment 1 was published in: Bennett, N. C., Millard, D. and Martin, D. (2016) Narrative Extraction through the Detection and Characterisation of National and Local Events, in Gartner, G. and Huang, H. (eds) *Proceedings of the 13th International Conference on Location Based Services*. Vienna, Austria: Vienna University of Technology, pp. 196-200.

Part of Experiment 2 was published in: Bennett, N. C. et al. (2017) Towards a Unified Narrative-Centric Spatial Clustering Model of Social Media Volunteered Geographic Information, in *Proceedings of the 25th GIS Research UK (GISRUK) 2017*. Manchester, UK, pp. 1-7.

Part of Experiment 3 was published in: Bennett, N. C. et al. (2017) Spatial Narrative Construction using Thematic KDE, in *Proceedings of the 2017 International Conference on GeoComputation*. Leeds, UK: Centre for Computational Geography, University of Leeds, pp. 1-8.

Part of Experiment 4 was published in: Bennett, N. C., Millard, D. E. and Martin, D. (2018) Assessing Twitter Geolocation Resolution, in Akkermans, H., Fontaine, C., and Vermeulen, I. (eds) *Proceedings of the 10th ACM Conference on Web Science*. Amsterdam, Netherlands: ACM, pp. 239-243.

1.6 Thesis Structure

With the rationale behind the thesis explored, the rest of the thesis will be structured in the following way: Chapter 2 will explore existing literature into narrative, textual and spatial analyses to situate the thesis within the wider research fields; Chapter 3 covers five experiments into how locative thematic narratives can be discovered and presented in a framework for wider use; Chapter 4 applied the framework to a case study in partnership with HM Treasury and Office for National Statistics (ONS) to discover and map real-time economic indicators (locative narratives with an economic theme); Chapter 5 evaluates the framework by interviewing relevant HM Treasury and ONS employees; finally, Chapter 6 draws the thesis to a close with a summary of the completed work, discussion of its limitations and contribution, future work to improve the framework and how it can be applied to future experiments.

Chapter 2

Literature Review

In the previous chapter, Twitter was situated within the landscape of social media and its useful features were highlighted to understand why it a key source for location-based narrative research. This chapter outlines the existing research into how Twitter has been analysed to extract narrative and location information. It then systematically analyses the existing quantitative and narrative techniques used in these works, discussing the strengths and flaws for each method. The chapter concludes by suggesting how these analytical approaches can be united in their analysis of Twitter data to extract ambient location-based narratives.

2.1 Narrative Analysis

The study of narratology helps academics and researchers to understand the philosophical perspectives inherent in constructing narratives. This section outlines key frameworks for narrative analyses, discussing the merits of each one before deriving an approach that focuses on location-based thematic narratives.

2.1.1 Formalism

Narratives are frequently associated with literature due to the work having a story or narrator ([Fludernik, 2009](#)). Work into understanding narratives in literature is well established, with formalism being a key analytical framework. Formalism dictates narratives must conform to prescribed structures without outside influences like politics or history ([Pavel, 1988](#)). Formalism analyses inherent features within the text, such as grammar and syntax, highlighting subtleties in how narratives are composed. This is advantageous for authors who do not wish to politically align their work as their narratives are unbiased. However, formalist approaches have been criticised for not accepting

the inherent influences of personal history and opinion on narrative, as well as their adherence to formal procedures failing to produce interesting work (Pavel, 1988).

2.1.2 Structuralism

Structuralism takes the opposite approach in that works must regard and be placed within their larger contexts (Landau, 1984). Structuralism concerns itself with the inherent patterns within text and how they relate to each other, though its approaches have been criticised for being overly formulaic (Pavel, 1988). Where structuralism greatly differs from formalism is its inclusion of human experiences as narrative components (Barthes, 2002) and their subsequent portrayal. Structuralists differentiate between what was discussed within a narrative and how it was told, eventually using 'Story' and 'Discourse' respectively as appropriate descriptors (Hargood, 2011); the story comprises the events within the narrative and the discourse is how these events are portrayed.

2.1.3 Thematic Narratives

Another alternative is thematics, which concerns itself with the underlying meaning of the narrative. Russian formalist Tomashevsky (1965) outlined a framework for analysing narratives within literature; narratives, he claimed, are comprised of features, motifs and themes, each comprising the next. Features are the smallest unit of meaning within this framework, such as individual words; motifs are comprised of these features, such as 'bustling market'; lastly, themes are comprised of motifs, such as the motif of 'bustling market' alluding to a theme of 'commerce'.

A thematic analysis of narrative is a highly conceptual approach to extracting meaning. Features can have several meanings, motifs can be themes within themselves and themes can be subjective. However, thematic analysis of narrative is especially useful when applied to large bodies of text as it can track popular themes, when they change and for how long they are prevalent. This dynamic approach to understanding narrative is also useful when analysing Twitter messages. As previously mentioned, narratives help us make sense of the world (Cambria and White, 2014) and Twitter allows users to reflect on their surroundings, therefore applying thematic analysis to tweets will extract both what people are saying about their environment and how they are saying it. However, analysing text alone is not sufficient for understanding these situated narratives; to fully appreciate narratives one must also consider location.

2.1.4 Space and Place

To fully appreciate narrative analysis of social media, one must first understand the context within which it resides. Space and Place theory is prominent amongst the

literature, specifically the differences between the two terms and how they are reflected in our daily lives. The distinction and understanding of space and place is important when considering that our social media contributions are digital reflections of physical activities; therefore, to analyse them as proxy to real-world happenings requires careful consideration of the theoretical perspectives.

It has long been argued that space allows for the structuring of place. In his work, [Agnew \(2011\)](#) describes space as a dimension for places to exist. This view is strengthened by [Cranshaw et al. \(2016\)](#), who emphasises the cultural void of a placeless space. However, [Relph \(1976\)](#) originally argued for space being a spectrum with direct experience and abstract thought at either end, for example perceptual versus cognitive space.

[Agnew \(2011, p.6\)](#) defines *place* in its simplest form as referring to “either a location somewhere or to the occupation of that location”. This is a deterministic approach, insofar as an object must have a physical presence to be a *place*. Alternatively, [Gao et al. \(2017\)](#) argue that *place* only exists as human-defined social space, created by their perceptions and experiences of the world. Whilst these definitions are not mutually exclusive, there is evidently an argument as to the extent to which human involvement creates the existence of place, an argument lacking in both definitions. In essence, as admitted by [Agnew \(2011\)](#), space and place are often used interchangeably, deliberately or not. The argument for their specific definitions has spanned decades with no concrete conclusion.

In later works, with the advent of social media and smartphones, space and place have been gamified through check-ins and ratings, as seen in TripAdvisor (an online reviewing platform for venues including hotels ([Ghermandi and Sinclair, 2019](#))) and Foursquare ([Saker and Evans, 2016](#); [García-Palomares et al., 2018](#)). These platforms allow users to comment and rate on physical places, but do so descriptively rather than conceptually. This is a trend seen in the literature: many reference Agnew’s space and place work but do not advance or refute it. Conversely, tied in with narrative structures discussed later in the chapter, research is ongoing into exploring places in real-time, giving the user a feel for the history and changing nature of spatial narratives while the story unfurls through their smartphone, both through image, text and audio ([Millard and Hargood, 2015](#); [Bassano et al., 2019](#)). These works expand the definition of *place* to include abstract concepts such as historical recreations and sensory experiences.

[Liu and Fuhrmann \(2018\)](#) describe space and place as opposite to [Agnew \(2011\)](#), where *space* is a physical location and *place* is the abstract construct created by our associations with the space. The aim of their paper was to promote the union of augmented reality (AR) and location-based social networks, but more significantly contribute a description of the increasingly popular augmented reality paradigm as being a means through which users can experience other users’ perceptions of *place* (as described by Agnew). The fusion of AR and social media is an interesting avenue of research, both for space and

place theory and the exploration of location-based storytelling as discussed in a later section.

The work presented in this thesis shall build upon existing arguments and pose a theoretical perspective from which the work shall follow: space is an abstract three-dimensional entity within which objects coexist and have relations, but only the objects with which humans relate can be classed as *places* as it is humans that give these places meaning. These places, through their human interactions, form the social space and their related narratives. Space and place exist independently of each other, but the former comprises the latter and the latter defines the former, a view that builds upon the generalist work by Agnew (2011) by modifying it with the specific creation of *place* by humans (Gao et al., 2017).

2.1.5 Placelessness

A strong challenge to this conceptual model is given by post-modernists. Arguments from Arefi (1999) focus on the ubiquitous nature of modern expansion, standardising the landscape and social space with monotonous motorways, train stations and airports. The journey made obsolete, travelling vast distances had become trivial and thus place lost its meaning. This argument is also relevant to ubiquitous computing. Friedman (2005) argued that space is consuming place: the ‘localness’ of place is irrelevant in the ‘globalness’ of the digital age. This is true inasmuch as social networking sites allow for people to instantly contact each other across the world, but it is naive to say the Internet or the Web makes place as a physical property obsolete. ‘Placelessness’ has arisen as a byproduct of the human drive to modernise, creating non-unique road networks, hospitals, restaurant chains and other aspects of our daily lives that deny any sense of difference. *Placelessness* dominates *place* (Agnew, 2011). In his original work, Relph (1976) discusses ‘placelessness’ as a result of the fixation on efficiency for its own sake, thus polluting human relationships with places by removing their character, moving towards a feeling of ‘outsideness’; the person feels a separation between themselves and their location.

Whilst some would argue it’s all semantics, as frequently mentioned in Agnew (2011), it is important to distinguish between space and place to understand society at a philosophical and sociological level, particularly with the rise and popular adoption of the Web. Social media is a prime example of how the world is rapidly moving towards ‘placelessness’. Users from across the world can communicate with friends and the public in real-time without needing to move. Whilst the rise of the Web has led to the demise of physical place as a necessity for communication, digital place must be the logical successor. If one considers the different social media as places within which to communicate, then this defies the post-modernist theory of ‘placelessness’ by creating virtual layers

over physical space, thus digital communication embodies the same cultural significance and social impact as traditional physical places.

2.1.6 Location-based Storytelling

There is an increasing body of work studying the storytelling aspects of location-based narratives, building upon the philosophical foundations laid out by Space and Place theory. These stories can be delivered by a number of systems, but most recently has been incorporated into the smartphone industry. Work into fictional and non-fictional storytelling ([Millard and Hargood, 2015](#)) gives us insight into how digital storytelling is evolving. This study focuses on the power of geo-enabled Web-connected smartphones to unlock new narrative content as the user moves through their space, as well as how the location affects the narrative and vice versa. Whilst this work uses pre-generated narrative it highlights the relationship between location and narrative and the ability for smartphone devices to capture and emphasise the significance of this relationship.

While the work by [Bassano et al. \(2019\)](#) involved location-based storytelling, their approach was to use it as a marketing tool. They bridge the gap between narratives as a fundamental aspect of human nature, and governmental and commercial interests in increasing footfall to particular areas. The assumption is if a user has an extra motivation to visit a location, such as an interest in their story, then this will increase traffic to that location. Their work is theoretical in nature and tested through two case studies, both aimed at companies to improve their competitive advantage through leveraging their customs and traditions into interesting stories, though no evaluation is made. A better angle would be the one proposed by [Liu and Fuhrmann \(2018\)](#), where augmented reality is used to overlay social media posts by nearby users to enrich an area with their own stories.

Obtaining and managing location data is not without issues. As location is a subclass of personal data and can be used to locate people who may otherwise not wish to be located, location data need to be carefully analysed ([Zafeiropoulou et al., 2012](#)). Similarly, in their work [Narayanan and Felten \(2014\)](#) argue there is no effective way of anonymising location data, emphasising the ease with which they identify a person from their metadata by knowing just one of the person's locations. It is therefore advantageous to use a thematic narrative approach; an understanding of the social use of space is extracted whilst individual users are not specifically identified.

2.1.7 Location as Identity

A related area for understanding how and why people create online content is identity. The ability to have near-total control over how one is represented online affects what sort

of content they will upload and when. This constructed, curated and deliberate content creation affects not only how one's online presence is reflected upon by the creator, but also how the recipients, or 'consumers', of the data view the creator.

In this light, [Hecht et al. \(2011\)](#) analysed how Twitter users presented their location through their profiles. Previous work, they claimed, took this field as a trusted value and subsequent analyses did not investigate further into its validity. The work by [Hecht et al. \(2011\)](#) concluded that only 66% of Twitter users entered a location that returned a valid coordinate when queried against Yahoo's Geocoding API¹. While a more technical paper, [Hecht et al. \(2011\)](#) does create useful insights into the nature of profile locations. Their methodology, however, relied on old or defunct technology such as aforementioned Yahoo Geocoding API, as well as UberTwitter (a Twitter client for Blackberries) and the Twitter API before it had an integrated location metadata field. This shows that constant revision is required as social media APIs can and do change over time, as well as supporting software being discontinued or new software emerging.

Similarly, and more recently, [Saker \(2017\)](#) studied how users promote their daily activities using Foursquare, and focus on how physical location and the idea of location influences users' activities. For instance, their findings show that users deliberately visited areas they thought their friends would find impressive to give an impression of themselves as well-travelled or cultured. Another example is the promotion of physical activity as a competition with friends; one user only went to the gym to keep up with his other friend who had 'checked-in' to that location. At the time, Foursquare had gamified the 'check-in' feature to generate points and badges for frequent visits, rewarding the top user with a mayorship. The gamified approach to Foursquare led [Saker \(2017\)](#) to conclude that 'check-ins' do not always have to pertain to the location at which the user is posting, as any post will give them in-game points; however, they counter this by saying fake 'check-ins' would be "against the spirit of Foursquare" ([Saker, 2017](#), p.946). Their unhelpfully vague conclusion, while not revealing how many of their participants were gaming the system, does shed an important light on the nature of the data: when analysing the data one must be aware of the self-curated, deliberate creation of social media posts and the impact this could have on future analyses.

2.1.8 Theoretical Perspectives

It is clear from the literature on theoretical perspectives that thematic narratives and location are closely related. Narratives are constructed from user experiences of their environments, allowing them to understand and relate their social and spatial information. This helps them construct an identity of themselves and their relationships with their environment. These spatial relationships have been impacted by the advent of digital

¹This is now deprecated, replaced by Yahoo PlaceFinder.

social media, adding layers of gamification and wider networking that have opened these narratives to researchers.

Therefore, to appropriately analyse narratives displayed through social media a twofold approach is required: textual and spatial. The union of these contexts will enable a holistic yet critical analysis of narratives of space, thus requiring analytical techniques from both of these areas. The following sections will cover topic modelling and spatial analyses, discussing their strengths and weaknesses and application to the thesis.

2.2 Topic Modelling

This section describes the behaviour of Twitter data and the effect it has on applicable analytical methods. It then outlines the development of topic modelling before concluding on the best available method and how it has been used in the literature.

2.2.1 Introduction

Topic modelling is the process of extracting similar patterns of text from abstract structures hidden within documents. There are two main approaches to defining topics: supervised and unsupervised (Mahmood, 2009; Beigi et al., 2016; Zhao et al., 2017). Supervised topic classification, otherwise called supervised learning within the machine learning community, relies on an existing dataset of labelled topics from which to match against the text of interest. If there is a match, such as keywords or sentence structure, then the new text is labelled appropriately. This is a rigid approach, and the training data must be relevant to the topic otherwise inconsistencies can arise (Kaneko and Yanai, 2015; Ibrahim et al., 2017) such as partial or erroneous labelling. The alternative is to use unsupervised learning, which typically implements a probabilistic model to analyse sentence structure and word co-occurrences. As this does not use any training data, unseen data can be used without prior data collection and is advantageous when analysing text in real-time. When the algorithm detects similar sentence structures or keywords, it can cluster the related terms together to form a topic. This is based on probability, with sentences that share more similarities with each other having a higher probability of being related to the same topic. This does risk incorporating similarly structured sentences together just because their structure is the same, but their word co-occurrence score would be low and therefore the algorithm either would not cluster the sentences or the investigator can see that particular topic has a disproportionate score.

With probabilistic models such as Latent Dirichlet Allocation (Blei et al., 2003) and k-means, topics are assumed to be probability distributions of words comprising that topic, thus if frequently co-occurring words or similarly constructed sentences are found

within the text they are matched to their representative topic. While both methods require the user to input k number of topics for clustering, LDA allows for words to belong to several different topics whereas k-means prescribes words to only one topic. The advantage of allowing flexibility with topic allocation is it aligns more closely with natural language; for example ‘chair’ could relate to the physical object or the head of a panel, thus LDA would allow for different documents containing the word ‘chair’ to be allocated to the two (or more) different topics. K-means, while not designed for topic modelling but sometimes used in the literature ([Ibrahim et al., 2017](#)), clusters similar documents together but is less flexible as each point within its vector space is fixed, thus is arguably less representative of natural language as ‘chair’ can only belong to one topic. These concepts are explored in greater detail in Chapter 3.

Therefore, topic models are useful and their machine-learning capabilities often necessary for discerning and extracting the ambient topics from noisy and large-volume text-based datasets. Twitter allows its users to publish free-form text, creating a large dataset of noisy and unstructured data. For researchers wishing to extract topics and hidden nuances from these noisy tweets, topic modelling provides an appropriate angle for analysis.

Twitter allows its users to publish 140-280 character² posts about a subject or subjects of their choice. This flexibility in textual input creates a problem for analysis: no fixed technique will be able to deal with the nuances of natural language. Thus, the method that is applied to tweets should be flexible, able to handle unseen data yet still extract meaningful topics. Furthermore, the volume of tweets being produced even within a small area requires a computational approach, thus the algorithm must also be able to handle large datasets. As tweets are also quite short compared with conventional documents such as news reports or blogs, there is less innate structure due to authors needing to be brief. This results in a large quantity of poorly or completely unstructured text, necessitating a process that applies nuanced structure to extract meaning ([Casas and Delmelle, 2017](#)). To understand these processes, the rest of this section will outline the utility of topic modelling algorithms, decide upon their applicability to this project, justify this with examples from the literature and conclude with examples of other approaches.

2.2.2 Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) is often heralded as the first iteration of modern topic modelling ([Deerwester et al., 1990](#)). LSA is used to find latent topics within a corpus at the semantic level ([Yalcinkaya and Singh, 2015](#)), in other words it’s used to find similar structures of meaning within the text. It achieves this by converting the unstructured

²At the beginning of this thesis users were limited to 140 characters. In November 2017 this limit was increased to 280.

text into a matrix, with rows representing terms and columns representing documents, to discover similar patterns of words and then assigning them to a latent concept based on their similarity.

A key advantage of LSA is its speed; as the algorithm is only creating one matrix, the process can usually be completed faster than other methods even on large data. Furthermore, as the user need only submit a value for k the process is consistent and results are similar each time. Disadvantages with LSA include its inability to differentiate between different meanings of the same word (polysemy); as the matrix is a fixed structure, words with several meanings like ‘chair’ are either ignored or obfuscate the results as LSA cannot reorder the matrix to accommodate the different meanings. This is a significant downside when dealing with text as polysemy is an important concept in language. Regarding tweets, where there is a limitation on the number of words possible within a tweet, each word has a significant impact on the meaning of the tweet thus any topic modelling technique must be able to consider polysemy.

2.2.3 Probabilistic LSA (pLSA)

Probabilistic Latent Semantic Allocation (pLSA), proposed by [Hofmann \(1999\)](#), is one method of improving the conceptual clustering process. Whereas LSA assigns documents to conceptual topics based on co-occurrence similarities, pLSA assumes each document is a probability distribution of topics and fits similar document profiles to related clusters. This approach works well with a known corpus of documents and can be repeated to obtain similar results. However, it has been criticised for over-fitting documents to topics and incorrectly clustering new documents due to its rigidity ([Blei et al., 2003](#)).

2.2.4 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation, proposed by [Blei et al. \(2003\)](#), improves upon the pLSA model by incorporating a Dirichlet prior to assert each topic is comprised of only a sparse number of words. This prevents over-fitting by restricting the number of words per topic and hence documents are not so aggressively clustered. Furthermore, as it defines documents as a collection of topics and enables terms within documents to be assigned to more than one topic, it naturally addresses polysemy. As LDA analyses the word-to-topic and topic-to-document relationships, this enables the thematic analysis of textual datasets (corpora) ([Eickhoff and Wieneke, 2018](#)). This is advantageous when analysing free-form text such as newspapers, articles or social media posts as these can comprise several different topics or themes in their content. While social media posts can be short, the vast quantity of them and their unstructured nature makes LDA a powerful tool in extracting hidden (latent) themes within the corpora.

A disadvantage present in many topic modelling algorithms including LDA is the requirement by the user to define k . This relates to the number of expected topics within the corpus and is subjective by nature unless additional means are used to computationally derive k ; for example, [Toubia et al. \(2018\)](#) calculate a “goodness-of-fit” value for use in their LDA model, though only obtain 61% - 71% precision on unseen data. Without this method, trial-and-error is often required to obtain an ideal value of k , a problem discussed later in [Chapter 4](#).

2.2.5 K-Means

Lastly, k-means is another clustering algorithm that works in vector space. It aims to partition the document set into k clusters. It does this by randomly plotting the k clusters then calculating the mean distance of each document to its closest centroid, then moves the centroid closer to the cluster mean before recalculating distances between the documents and the new centroid position. It then recalculates which document belongs to which cluster based on this new mean until the centroids and documents reach equilibrium. The final clusters represent documents with the least variance to each other. Due to the simple nature of its algorithm it can run quickly over a large dataset, but due to its clustering process it can only assign a document to a maximum of one node ([Singh et al., 2011](#)), so for NLP it can only assign a document to one topic, rather than assigning terms within the document to several topics. The k-means process starts by randomly plotting clusters in vector space before calculating distances, so each run of the algorithm is different and thus can produce slightly different results each time. Therefore, due to its single-cluster document classification, k-means is inappropriate for extracting narratives from documents that could contain several topics and would need to belong to several different clusters.

2.2.6 Applications of LDA

LDA has been successfully used in a wide variety of research projects to extract topics from noisy text-based datasets. [Guo et al. \(2017\)](#) used LDA to analyse hotel reviews, thus not requiring any form of training data or location parameters prior to the study unlike [Ashktorab et al. \(2014\)](#) or [Fang et al. \(2015\)](#). They argue that low-dimensional data such as hotel reviews benefit from LDA due to its Dirichlet prior affording flexibility with sparse topics spread over time ([Blei et al., 2003](#)). In their methodology, [Guo et al. \(2017\)](#) used the natural language toolkit (NLTK) package from Python to stem and tokenise each of the words in the reviews to further reduce noise and homogenise the dataset. This extra methodological step improved the topic accuracy by reducing the term variance, a key process missing from earlier work such as [Ashktorab et al. \(2014\)](#) and [Fang et al. \(2015\)](#) as well as more recent work such as [Yu et al. \(2019\)](#). The authors’

results, however, did not take into account any spatio-temporal attributes which seems important when analysing tourist destinations.

LDA has been used in other fields such as psychology (Toubia et al., 2018) and mathematics (Yasunaga and Lafferty, 2019). The underlying features of LDA make it viable for any form of clustering problem, affording it widespread application. The use of LDA on Twitter data also successfully extracts topics from the noise, but as discussed below, there are several factors that must be considered when applying and presenting the results.

2.2.7 LDA on Twitter Data

Extracting topics from social media data allows for intricate semantic analysis of social activities, and for the construction of narratives over time. By far the most popular topic modelling algorithm in the literature for tweet data is Latent Dirichlet Allocation (LDA) due to its ability to understand low-dimensional data, such as short tweet texts. As previously discussed, LDA allows words the flexibility to belong to many different topics rather than restricting each word to one topic, better mirroring natural language.

LDA has been used as a key research method to extract topics about disaster management (Ashktorab et al., 2014), tourism (Guo et al., 2017) and obesity (Ghosh and Guha, 2013), as well as in more generalised data extraction research into location-based topics (Lansley and Longley, 2016), event detection (Chae et al., 2012; Zhang and Eick, 2019) and political leanings (Fang et al., 2015). For all of these papers, Twitter was used as either the sole or a major contributor of their data.

In their work on obtaining actionable information for first responders in disaster situations within North America, Ashktorab et al. (2014) highlighted the noisy nature of Twitter often obscuring vital information. To combat this, they modified the LDA algorithm to allow for seeded topics — predetermined topics with relevant keywords around which related tweets would cluster during the LDA process. Though this process was successful in discovering known topics, the precision and recall values of 42% and 65% respectively emphasised how poorly this method would perform on unseen data. They also do not discuss any forms of data cleaning, such as removing bots, spam or otherwise irrelevant data, thus their topic accuracy would have been impacted. Improving the topic accuracy by extensive cleaning is a major part of this thesis contribution. A similar approach by Pereira et al. (2017) to extract actionable information within Brazilian cities focused more on data cleaning to remove noise by tokenising, lemmatising and removing stopwords. Though their results showed an extensive list of 50 topics within the tweet dataset, their subjective classification included topics such as “mood” and “routine activities” which are too vague to be of use. For studies using LDA over such a

large geographic area, it would be more effective to first target a topic of interest, such as via keywords, then apply LDA to create more actionable results.

[Ghosh and Guha \(2013\)](#) extracted spatially-relevant LDA topics from a corpus of tweets relating to obesity. This allowed for both discovering what people felt about the topic but also which areas had the greatest level of conversation. Their results showed the extent to which obesity-related tweets were prevalent within US counties using ArcGIS. Therefore, their methodology allowed for both a ‘what’ and a ‘where’ in terms of their topic of interest, and although ArcGIS is a black-box in terms of underlying functionality, it is an industry-leading software suite and thus produces reliable results. However, the notion “garbage in, garbage out” ([Kim et al., 2016](#)) applies, thus careful cleaning of the data is required before using such software. [Ghosh and Guha \(2013\)](#) removed stopwords and used tokenisation and bi-grams to create a homogenous data source, though warned of the increase in computational power and time required to analyse bi-grams, as well as a risk of creating sparse topics. However, as they were analysing obesity in relation to other factors such as fast-food chains or schools, this was a logical step. For other investigations into more generalised topics, using bi-grams or greater would not be appropriate.

[Lansley and Longley \(2016\)](#) approached tweet cleaning with a more rigorous manner. As their area of interest was weekday activity in London, they first removed tweets originating at the weekend then Fridays and Mondays, arguing there were overlapping timescales and therefore would skew the representation. Their approach to removing irrelevant or spam tweets differed from those previously analysed, as they took a more pragmatic approach to what constituted a meaningful tweet; they removed tweets with fewer than three words and words with fewer than three characters as they argued they did not meaningfully contribute to a topic model, and users with more than 3,000 tweets to reduce bias; however, this value was arbitrary and not generalisable to all Twitter datasets. They also removed stopwords, a feature present in the earlier works. They discovered that 20 topics comprising about 65,000 tweets per topic generated representative and distinct results, a part of the results not mentioned in the previous works. Despite their more rigorous data cleaning methods, they do not focus on third-party tweet sources which are often a source of irrelevant information, nor do they attempt to resolve third-party geolocation information that may have influenced their maps of London (both methods discussed later in Chapter 3). Without these two focal points in their data cleaning method, their maps and topics could be significantly biased. A key contribution of this thesis is to focus on both source and location analyses as part of the cleaning process.

[Chae et al. \(2012\)](#) create a piece of software capable of extracting location-based topics. They deliberately did not automate the entire process to allow for user interaction and subjective processing ([Chae et al., 2012](#), p.2). While this method may be of use to niche researchers with a knowledge of how LDA works, including topic number and spatial

extents, their software is not applicable to the general public without first training the users. Spatial extents and topic count can significantly impact the output, thus each time a user uses their program the results will be subjectively biased. Furthermore, while they use Seasonal Trend Decomposition with Loess, a method that identifies topics that have longevity within a dataset, to filter out background chatter and highlight ephemeral events, they still do not consider automated or otherwise irrelevant users or sources thus their results can be skewed by particularly active but short-lived bots. Therefore, their work is not generalisable to other projects and is only of use to niche users with knowledge of topic modelling algorithms, making it undesirable for widespread adoption.

[Zhang and Eick \(2019\)](#) created an event detection system using geolocated Twitter data, LDA and kernel density analyses. Their framework collects tweets in temporal batches of a fixed length (configurable by the user), applies LDA to each batch to extract event labels, then analyses the hotspots found within the geolocated tweets to create event seeds. The next batch is analysed, and similar topics or locations are grouped together to form events while new topics or hotspots form new seeds. This framework then produced maps of the events, such as the Ferguson riots in the USA in November 2014, and how they changed over time. However, their spatial analyses are only at city or state level, and their LDA labelling has no human intervention so if a batch were improperly labelled then this would have a systemic impact on the results, creating alerts over meaningless topics. However, their work does include a reflexive section about the pitfalls of LDA, which is refreshing. They highlight the importance of cleaning tweets, otherwise LDA outputs will include hashtags or mentions; they also discuss dominating topics such as a sporting event skewing the LDA results, but are unable to resolve this issue. As the work created an topic-ambivalent event detection system, they were unable to refine the tweets to a particular event of interest. Had they had some method to refine the tweets initially, such as keyword matching, then undesirable topics would not have dominated the results.

[Fang et al. \(2015\)](#) attempted to mix keyword training sets and LDA topic modelling. They matched pro- and anti-referendum hashtags to tweets from Scottish residents during the 2014 independence referendum, then created a dataset of related tweets without the hashtags. This created a generalised training set that could be applied to other fields. Though this was a novel approach, they clustered individual users' tweets into one document, with an assumption that they will tweet about the same political approach. However, this is an approach criticised by [Lansley and Longley \(2016\)](#) as reductive as people are likely to talk about a variety of topics, therefore an LDA model built upon an aggregated tweet dataset risks creating vague and irrelevant topics. This is shown by the increasing topic count reducing the accuracy of their classifier once the topic count reached 30, indicating that as the topics get more diverse the vague training data inhibits more complex clustering.

[Juntao et al. \(2015\)](#) successfully modelled tweets and their topics around the vicinity of London tube stations using LDA and k-means. Their approach created maps of topics of interest by commuters and travellers which were clustered around their nearest station, highlight particular groups such as tourists or topics such as sports. While this is useful in understanding which topics are discussed at which tube stations, they do not specify how they obtained their geolocated tweets, nor any method of cleaning the data aside from dividing the tweets into weekday (Tuesday-Thursday) and weekend (Saturday-Sunday). Their results are therefore difficult to reproduce as it is unclear how they were generated.

2.2.8 Keyword Matching

Keyword matching is an area of research that has been previously discussed but within the context of LDA ([Ashktorab et al., 2014](#); [Fang et al., 2015](#)). Within research into text analysis, careful cleaning of the data is a key step that many of the works ignored. The notion of “garbage in, garbage out” ([Kim et al., 2016](#)) applies strongly to topic modelling algorithms as they rely on relationships generated between words in documents and between all documents in a collection. If the underlying data are compromised by noise then the results will be equally hindered.

Keyword matching provides an alternative to LDA. This is a simpler approach where the user has a set of existing keywords that are relevant to the topic of interest, and in matching these to the tweets subsequently extracts related information. Keyword matching is often the focus of search engine optimisation ([Gong and Liu, 2009](#); [Abilhoa and De Castro, 2014](#)) as they use a networked graph of related terms to return relevant information.

Where there is no existing set of keywords, care must be taken when generating one, as subjectively creating keyword sets risks introducing bias into the results ([Habernal et al., 2013](#); [Ghermandi and Sinclair, 2019](#)). Matching keywords to large datasets can also create misleading results as the datasets can contain numerous different topics. This is shown in [Fang et al. \(2015\)](#) who created one document for each user’s tweets and aggregated all users’ narratives, an accumulation task criticised by [Lansley and Longley \(2016\)](#) for diluting the topics inherent within individual tweets. Therefore, how keywords are applied to tweets must be carefully considered.

Where keyword matching has been successfully used is in disaster management, as shown by [Smith et al. \(2017\)](#). In their work, they use flood keywords generated through domain expertise such as ‘water’, ‘inundated’, ‘flowing’ to identify the theme of flood-related tweets within a spatial boundary box in Newcastle. This allowed them to study the spread of the flood and how widely it affected the population. While their approach was successful in mapping the affected areas and creating actionable results for first

responders, they admitted that only a few hundred tweets matched the keyword set, causing issues with their summaries. However, they were able to generate reliable results from this small sample and therefore keyword matching can be used as a viable approach for extracting specific thematic data.

Another application of keyword matching is with specific studies into particular events. [Jeske et al. \(2017\)](#) investigated the discussion on Twitter surrounding the “#Heart-bleed” event — a bug in OpenSSL which created vulnerabilities across client to server connections, allowing interested parties to intercept the data. This had widespread implications for all technological sectors and was thus a large topic of conversation on social media. Their approach applied a set of keywords relating to vocational sectors, such as ‘IT Professional’ or ‘barrister’ for IT professionals and legal professionals respectively ([Jeske et al., 2017](#), p.177), to the user’s profile information to identify users who would most likely be affected by the attack. Their tweets were then classified as belonging to that sector and investigated for further topics of interest, such as advice or reactions. Their results were promising in understanding how tweets formed around a particular issue and the resulting discussions, though the authors admitted that future work would be key in further validating their approach.

2.3 Spatial Analyses

To properly identify location-based narratives, both textual and spatial analyses are required. Twitter data is noisy by nature due to it allowing users to post at any time or place. For tweets with geolocation information, the accuracy of this metadata is not straight forward. Inherent GPS inaccuracies generated by poor satellite coverage or densely-packed buildings blocking the signal can cause the tweets to be several metres off their intended location, on top of existing inaccuracies from the device’s built-in GPS. This creates a spatially noisy representation, thus methods that directly address this behaviour are of high importance for Twitter analysis.

2.3.1 Introduction

As defined by the spatial literature, tweets represent marked point patterns ([Lloyd, 2010](#)). These are data points with extra attributes, such as usernames, text or timestamps. Point pattern analysis is a substantial area of study within the geographical research community, and several relevant methods to analyse point patterns have emerged. These methods analyse spatial relationships, such as distances from each point as well as any discernible cluster centroid. [Lloyd \(2010\)](#) also includes two further types of relationships: first-order and second-order effects; the former relates to the intensity of events per unit area, the latter analyses the relationship between paired points in the

entire study area ([Bailey and Gatrell, 1995](#)). These effects are often applied to disease outbreaks; a first-order effect would be poor sanitation leading to an increase in mosquitoes, with a second-order effect being a rise in cases of malaria in nearby places. An example of first-order effects for Twitter data would be an event at a location causing more people go there and tweet about it. A second-order effect would be the nature of the information dispersal from the location of the event causing users to talk about the event at other locations.

A further distinction that is required to understand the nature of Twitter data is parametric versus non-parametric. Parametric data are assumed to be drawn from a population that can be sufficiently described by a probability distribution with fixed parameters. The converse is that non-parametric data are assumed to be drawn from a population whose probability distribution cannot be described with fixed parameters. Twitter data can be defined as non-parametric, due to the large volume of tweets produced each day by different individuals at many different locations not conforming to any fixed distribution. Therefore, non-parametric statistical tests should be used to analyse this data, in particular regarding spatial distributions.

Methods to discover and analyse these data are grounded in mathematics but have since been adopted into existing systems such as ArcGIS or packages for programming languages. Kernel density estimation (KDE) and K-nearest neighbour (KNN) are two popular non-parametric spatial analysis algorithms that calculate the distances between each data point compared to a unit area to discover densities. Both approaches require user input to function, but they deal with densities differently. KNN calculates the distance between each point within a window, but the shape of the window changes depending on the distance between the point of interest and its nearest neighbouring point. Therefore, the result can be erratic and non-uniform due to the continuous recalculation of densities. KDE applies a fixed window over each point and calculates the distances within, creating a smoother result due to having a standard window size at its core. The risk with KDE is over-smoothing and under-smoothing of the densities due to the window not being an appropriate size for the related densities, therefore there is an element of subjectivity when selecting a this window and a corresponding impact on the results. Despite this, it deals with Twitter noise more effectively than other methods by smoothing out erratic tweeting behaviours from either GPS inaccuracies or many users tweeting from a similar location. In sync with the linguistically unstructured nature of tweets discussed in [Section 2.2.1](#), in order to extract meaning from the noise one must first discover structure from or attribute structure to the data, such as overlaying a spatial grid and calculating the relationship between the constituent points.

The following section further describes the functionality of KDE as well as subsequent spatial analyses. The section then covers how KDE has been used within the literature before describing how the nuances present in Twitter's API require novel analytical approaches before KDE can be applied.

2.3.2 Kernel Density Estimation

Kernel Density Estimation (KDE) is a non-parametric statistical method that generates a probability density for spatially related data. It calculates distances between each point within a window, generates a density value based on the closeness of these points, then smooths the output based on the bandwidth parameter. The bandwidth determines how closely the points match the distribution; a smaller bandwidth creates a more granular output, whereas a larger bandwidth smooths out the irregularities. The output of applying KDE to spatial data is similar to a heatmap. To understand the benefits of KDE, one must understand its predecessor: histograms. Histograms represent discrete collections of data with similar properties, and require an input to define the extent of the clustering of points with similar properties. For example, a histogram can represent a tweet dataset where each category, or bin, can be a range of how many tweets a user has made. The resulting histogram will therefore represent the distribution of tweets per user. When applied to spatial data, each bin could reflect the distances between points. KDE applies the same approach to spatial data by calculating the distances between each point within the window and categorising these distances in bins, then smooths the outcome of these bins by the bandwidth value.

As a raw tweet dataset does not conform to known distributions, KDE is ideal for analysing noisy Twitter data and extracting clusters of activity. The downside is the reliance on a user-defined bandwidth value as this can risk over- or under-smoothing results if the window is ill fitting. There has been work into adaptive KDE, where the bandwidth changes its value based on how many tweets are within the area ([Lloyd and Cheshire, 2017](#)), but this approach is too complex for the scope of the thesis.

2.3.3 Getis Ord and Moran's I

Two further spatial models that are relevant but out of scope for the thesis are Getis Ord and Moran's I. Getis Ord is a statistical method for calculating spatial autocorrelation. It calculates the cluster strength of one area and compares it with neighbouring areas, then compares these neighbourhoods to the overall dataset. The main advantage is with its incorporation of confidence values, as Getis Ord uses its own confidence values to calculate the spatial similarities of clustered areas as opposed to relying on other models such as KDE.

Similarly to Getis Ord, Moran's I is a statistical method for calculating spatial autocorrelation ([Anselin, 1995](#)). When applied to tweet clusters, this method allows for the measurement of cluster strengths within an area compared to its neighbouring areas. Where Local Moran's differs is its ability to remove each spatial component within a neighbourhood area and re-calculate the neighbourhood cluster strength. This allows for the discovery of outliers, for example abnormally high values within sparse areas or

low values within otherwise high clustered neighbourhoods. An example use-case of this method is in the retail industry to discover popular or under-performing shops within a dense retail area (Reigadinha et al., 2017). As KDE is predominantly used for visualisation purposes rather than statistical, additional time required for the implementation of these supplementary methods could not be justified within the context of the thesis. Furthermore, as identified in the case study in Chapter 4, parties such as HM Treasury are not currently interested in hyper-local results. However, the detailed analysis of already precise tweets would provide further interesting statistical insights into areas of activity highlighted by KDE, discussed further in the Future Work section of Chapter 6.

2.3.4 Location and Social Media Data

With the various spatial techniques being investigated, analysed and their usefulness understood, it is important to appreciate how these techniques have been used in social media analyses. Work in this area has focused on extracting home and work locations, discovering retail catchment areas and location-based topics from Twitter data.

Previous work on extracting home locations has raised several issues regarding cluster analysis and the quality of conclusion that can be made from the data (Cheng et al., 2010; Mahmud et al., 2012, 2014; Lin and Cromley, 2018). Work that focused on city-level classifications clustered the tweets based on a location hierarchy such as state, province, city (Mahmud et al., 2012, 2014) or a probabilistic model of the origin city from text analyses (Cheng et al., 2010). These works relied on raw text data or profile information from Twitter and Foursquare to determine a user's location, ignoring precise coordinate data and assuming profile locations are trustworthy, which has previously been refuted (Hecht et al., 2011). While these methods are advantageous when tweets are missing coordinate data, city-level classifications are vague in comparison with tweets that include coordinate data, and are therefore not suited to street-level analysis. Lin and Cromley (2018) used a combination of parcel data (administrative classifications like residential, industrial) and tweets with precise location enabled to classify a user's home location. They applied a series of tests, namely a weighted most-frequently visited (WMFV) algorithm and a Support Vector Machine. Their results show a 60-85% accuracy when classifying home locations using the two approaches to analyse tweet clusters generated at different times of day. While these values are relatively favourable, they decided to include clusters of 1 tweet into their analyses. As these tweets could be anything from check-ins, precise tweets or ones generated by bots (which the authors do not address), this hinders the accuracy of their results.

To achieve street-level analyses, Middleton et al. (2014) and Smith et al. (2017) also use precisely geolocated tweets. With a focus on modelling real-time disaster response in at-risk areas, Middleton et al. (2014) used named entity extraction to match placenames

to a pre-loaded gazetteer of streets and buildings. This approach generated a high precision of 100% when applied at city-level, but when the area was reduced to street-level the precision dropped to 33%. This was due to the precision relying on named entity matching, thus with a reduced baseline dictionary the precision fell. Furthermore, their methodology is only applicable for at-risk areas that generate a strong Twitter footprint. Therefore, relying on named entity extraction risks poor results when not in built-up areas. Also related to disaster management, [Smith et al. \(2017\)](#) analysed flooding in densely populated Newcastle, UK. They acquired tweets from the Streaming API using keywords and a boundary box, neither of which they specified in the paper. They used a keyword-matching approach to discover flood-related tweets then mapped them using their geolocation data and metadata extracted from Ordnance Survey vector maps, OSM and the Royal Mail postcode address database. This had the advantage of attributing the tweets to a known address, but was a costly endeavour both financially and technically as the authors admitted to using 4 high-powered graphics cards (NVIDIA Tesla M2075) to render the flooding simulations, an approach not always available to other researchers. However, the initial step of attributing tweets with a known location is a useful contribution and strengthens the argument for a keyword-based approach, as well as promoting it as a useful step in future investigations. The ability to enrich tweets with real-world addresses makes the dataset useful for a wider audience such as policy makers.

These approaches have shown the weaknesses in named entity extraction, namely computation power required and lack of domain ambivalent applications. Using existing spatial data analysis methods such as kernel density estimation (KDE) has shown an increase in applicability and a reduced computation cost. KDE relies on just geospatial data, such as coordinates, therefore is preferred over the types of approaches mentioned above that rely on nuanced machine understanding of placenames or bespoke gazetteers.

2.3.4.1 Application of KDE

Other work that has matched social media data to known locations has employed kernel density estimation (KDE), a popular spatial data analysis method explored previously in Section 2.3.2. Some of these have focused on event detection ([Chae et al., 2012](#); [Zhao et al., 2017](#); [Gao et al., 2018](#)), individual and urban mobility ([Hasan et al., 2013](#); [Lichman and Smyth, 2014](#); [Boy and Uitermark, 2017](#)), retail zones ([Lloyd and Cheshire, 2017](#)) and crime prediction ([Gerber, 2014](#)). All of these works have used KDE to extract activity spaces and visualise location-based narratives, thus providing a key foundation for further study.

In their work, [Chae et al. \(2012\)](#) developed an analytical program that could render a KDE map based on user input. Previously geolocated topics could have their densities analysed and extracted for further analysis, a similar approach to [Zhao et al. \(2017\)](#).

However, these authors used KDE as a visualisation tool rather than an analytical one. As KDE was only a small function of their overall methodology, the authors missed the analytical potential for more intricate spatial relationships, unlike [Gao et al. \(2018\)](#) who used KDE as a focal point in their analysis. As methods for displaying density values rely on enough data within each raster cell to create bins, these systems fail when data are sparse ([Gao et al., 2018](#)). Therefore, as KDE uses comparisons between windows to create a smoothed probability density distribution, it is more robust to use for social media data as these data are also often sparse. [Gao et al. \(2018\)](#) used KDE to track and represent the migration of influenza across the United States which, due to its size and sparsity of Twitter data, necessitated a KDE approach.

In the above examples, KDE was used over a very large geographic area; however, it has also been applied to urban areas with equal success. In their approach, [Lichman and Smyth \(2014\)](#) experimented with different window sizes when analysing Twitter and Gowalla data from southern California, arguing that KDE as a whole is better than previously-used Gaussian approaches as the latter oversimplified the intricate spatial patterns seen in human mobility data. They proposed an adaptive model, one that mitigated against KDE's tendency to over- or under-smooth areas with naturally differing spatial densities (e.g. an urban area next to a rural zone). Though they argued this produced more accurate maps, they did not produce any adaptive KDE maps in their paper. Furthermore, this approach is overly complex when dealing with singular geographic densities and thus is not applicable when studying purely urban areas. Adaptive-KDE was also used by [Lloyd and Cheshire \(2017\)](#) who studied the catchment area of retail centres using data from Twitter and the IRUK Retailing top 500 Annual Report, which outlined the number of shops and their locations. Their adaptive-KDE approach identified hotspots of activity around retail centres, tracked the user's tweets back to their predicted origin points, and concluded with maps showing retail centre catchment areas and the distances its users travelled to reach them. This work showed how Twitter data and KDE can have an impact in retail studies and urban planning by modelling location and mobility patterns at scale. [Hasan et al. \(2013\)](#) also studied urban movements using check-in data from Twitter data and analysed it using KDE. However, their approach differed from the previously mentioned ones as they incorporated a temporal aspect that modelled the 'pulse' of the POI. However, like [Lichman and Smyth \(2014\)](#) and [Gerber \(2014\)](#) their work did not mention any method to remove undesired tweets before applying their analyses, unlike [Lloyd and Cheshire \(2017\)](#) who set frequency thresholds for users within their dataset to mitigate against spam. These methods were simplistic, thus more work is needed to detect automated messages, bots, spam and other undesirable or irrelevant tweets, especially when dealing with location data whose precise value and spatial relationships can become obfuscated if automated posts are included.

[Boy and Uitermark \(2017\)](#) took an interesting approach to analysing social media data.

Their focus was on Instagram rather than Twitter, and aimed to map activities within the city of Amsterdam. Similarly to [Li et al. \(2015\)](#), they aimed to match the photos and coordinate data with POIs within the city. However, only [Boy and Uitermark \(2017\)](#) generated a visual map of their results, whereas [Li et al. \(2015\)](#) only produced stats and graphs. [Boy and Uitermark \(2017\)](#) produced a density map of Instagram tweets within Amsterdam city centre, and though they did not specify the methodology with which they derived the map, the outcome is similar to a KDE distribution, showing that KDE implementations have widespread application. The authors also note how roads and paths that are frequently used are identifiable from the density map, a trait shared by [Lloyd and Cheshire \(2017\)](#).

2.3.4.2 Resolving Twitter Locations

Though the above works included tweets linked from other sources (for example Instagram and Foursquare), none of them mention the spatial resolution of the data nor the creation of the posts from the original source. With Instagram and Foursquare, users "check-in" to a location and can share this information with Twitter. Twitter then displays the text of the message, often cut short to 140 or 280 characters, with a geolocation metadata tag that corresponds to the POI at which the user has checked in. So while the spatial resolution of the data is apparently at POI-level, when directly comparing the coordinates found in the tweet metadata to the original post there are often discrepancies: the coordinates are not always the same. Literature on this is non-existent, and the various API documentation do not mention the translation of third-party location information to Twitter's format. During this process, the location information is matched against Twitter's gazetteer and a location attribute added to the tweet, either because no location information is shared from the source to Twitter or Twitter does not recognise the point of interest. This can result in the coordinates being mismatched, thus creating an inaccurate representation of location data received from third-party sources. As Twitter uses Foursquare to resolve its POI information, the tweets from Foursquare are largely untouched. Check-ins or posts from Instagram use an older version of Foursquare's gazetteer and more recently use Facebook's database, thus these tweets are more likely to be affected. While the experiment to prove this is not mentioned until Chapter 3 Section 3.4, the lack of discussion of this discrepancy forms a key part of the research gap and subsequent proposed framework. Furthermore, as Instagram and Foursquare allow for significantly larger text fields than Twitter, any form of text analysis on tweets must be aware of this additional limitation, but is equally missing from previous work.

2.4 Conclusions from the Literature

From the technical literature it is clear that topic modelling and spatial clustering is useful for understanding Twitter data. Above all, LDA and KDE are two main methods that emerged as most useful. Keyword matching was used successfully to extract thematically-relevant information, therefore this will also be investigated. However, a key area that is lacking targeted research is data cleaning and preparation; little to no mention of removing bots, spam or otherwise irrelevant tweets was part of the literature, and no mention of analysing location-based automated accounts was made at all. It is therefore certain that works published by the authors above included irrelevant data that will have impacted their results, particularly relating to location modelling as seen in [Lin and Cromley \(2018\)](#).

In close partnership with this technical method is the underlying philosophical approach. Analysis of narratological approaches showed how a thematic slant was best suited to Twitter analysis, thus the aforementioned computational methodologies are used to support this perspective. Keyword matching and LDA will contribute towards Research Question 1 (*To what extent does a thematic approach afford a richer understanding of location-based activity than topic modelling?*). Data cleaning processes and KDE will contribute towards Research Question 2 (*Can location-based thematic modelling be automated?*). A practical application of the methodology created by this thesis based on the answers to RQs 1 and 2 will be used to investigate real-time economic indicators on Twitter to answer Research Question 3 (*Can this approach be applied to a real-world situation with actionable results?*).

From the literature reviewed in this chapter it is clear that work analysing location-based Twitter data is both complex and in need of improvement. The combination of narratology, topic modelling and spatial clustering were key in understanding the nuanced nature of the data produced by the social media platform; however, gaps emerged in the literature where insufficient research had been conducted on how tweets were clustered, their behaviours when generated from third-party sources and where topic modelling had removed important contextual and narrative information. These key issues impact the quality of conclusions that can be drawn from such data, thus careful consideration of metadata is vital when producing trustworthy results.

Chapter 3

Developing a Framework for Location-Based Narrative Extraction

This chapter outlines five preliminary experiments that addressed oversights in the literature identified in the previous chapter. The experiments dealt with different aspects of discovering, modelling and analysing location-based Twitter data, with an aim to generate a framework for gathering, extracting and mapping location-based narratives. The first experiment aimed to model the multivariate nature of social events, merging spatial and social aspects into a parallel coordinate model for appropriate visualisation. It then modelled several real-world events that occurred during the summer of 2016 and discussed the viability of detecting and modelling these events. The second built upon the topic modelling from the first experiment and focused on the location aspect, moving away from specific events and towards modelling ambient narratives. This was an important step in understanding ‘real’ narratives, as the events detected in the previous experiment were only snapshots of an overall location-based narrative. The third experiment focused specifically on the location metadata, due to Twitter’s API having a greater impact on the location accuracy than previously thought. A new method was created to deal with artificial location clustering caused by improperly expressed POI metadata or the user’s privacy settings. The fourth experiment arose from further investigations into how Twitter’s POI accuracy compared with established systems, such as Ordnance Survey’s POI database. It discovered interesting discrepancies between expected POIs and to where the tweet was geolocated. The fifth experiment is a successor to the third and fourth, in which geolocated tweet clusters were scrutinised for their accuracy. The cluster coordinates were compared against three open APIs that returned address information indicating to which resolution the tweet cluster belonged. Finally, the chapter concludes with a discussion of the five experiments and how they fulfilled

the aim of this chapter to the construct a holistic framework to extract location-based narratives. Parts of this chapter are published in [Bennett et al. \(2016, 2017a,b, 2018\)](#).

3.1 Experiment 1: Prototype Event Classification Model

Throughout the reviewed literature, there was no attempt made to consolidate the shared attributes of events, such as social impact, spatial scale or length of event as discussed in the event detection literature (such as [Sakaki et al. \(2010\)](#); [Chae et al. \(2012\)](#); [Zhao et al. \(2017\)](#)). It was therefore necessary to first understand which types of events can be identified within tweets, and whether it was necessary to combine them with external sources such as news articles to understand an event’s impact on the narratives.

To identify events present in tweets this experiment applied topic modelling to identify similar themes and extract social and locative aspects from them. Ground-truth datasets are a feature present in much of the event detection literature and offer a comparison set of known locations with which to compare and evaluate the results. For this experiment, ground-truth articles are sourced from BBC News¹ for two reasons; primarily they confirm whether the events detected in the Twitter dataset are ‘real’; secondly they act as a comparison to show the impact of national and local news on discussions.

3.1.1 Methodology

To discover the impact of events on a population, the city of Southampton, UK, was used as a data source of tweets. Ethical approval was obtained for collecting data; this process took one working week, during which the plan for the experiment was finalised (Appendix A for the ethics application). A bounding circle with a radius of 3km was placed over the city to collect tweets from the 9th to 20th June 2016. This size was used as it included some of the surrounding urban areas, creating a dataset large enough to analyse and one that captured a variety of social spaces. These dates were selected as the 9th was the first day after the ethic approval was obtained, with the 20th being a sufficient number of days after to obtain a large enough dataset to analyse, irrespective of current affairs. The public Stream API was used to collect the data, therefore the data represents a random sample of the tweets from the area, up to a maximum of 100% of all tweets but a potential 1% if the sample exceeds 1% of all tweets generated globally during the time period. However, as this experiment used a bounding circle it restricted the global area to a much smaller region, thus there is potential for 100% of the tweets to be collected ([Morstatter et al., 2013, 2014](#)). In total, 348,129 tweets from 73,586 users were collected. To compare the events discovered within the tweets to a ground-truth dataset, BBC News articles were collected from the same period. These

¹<http://www.bbc.co.uk/news> [Accessed 20 June 2016]

were obtained manually using the Wayback Machine² to accurately temporally situate them as, at the time, whenever a BBC News article was edited the date of publication was also updated and thus it could be several hours or days out of sync. In total, 223 BBC News articles were obtained from the front page³ and UK section⁴ and stored in a CSV file with the name of the article, a permalink to it and its related section. All tweet gathering, processing and result generation is done in Python 2.7 with the topic modelling via Gensim⁵.

3.1.1.1 Pre-processing

To obtain representative tweets, users who posted frequently were removed from the dataset to avoid bias. To discover these accounts, users were ranked by how many tweets they made during the period and those who made over 1,000 were removed, a similar approach taken by [Lansley and Longley \(2016\)](#) and [Lloyd and Cheshire \(2017\)](#) to reduce representation bias. In total, 2 users were removed as they had a markedly higher number of tweets than the rest of the sample (2,273 and 1,494 respectively compared with 913 of the next user). As retweets offer no unique information ([McMinn et al., 2013](#)), 111,459 retweets were removed. Lastly, tweets shorter than three words and those only containing URLs were removed as these provided little to no information. After this last stage, 15,975 users and 207,623 tweets remained, emphasising the importance of data cleaning. The cleaning process was automated and is summarised below in Table 3.1. These tweets were further broken down into the individual days spanning 9-20 June and LDA was applied to each set to extract the dominant topics. In total, 223 unique BBC News articles were published between 9-20 June. To refine the articles to those concerned with a location, a manual inspection of the articles identified only those that explicitly mentioned a venue, town or city, reducing the total number of articles to 83 and those that were UK-based to 45.

Total	Raw	Cleaned
Tweets	348,129	207,623
Retweets	111,459	0
Geotagged	2,953	2,904
Users	73,586	15,975
Mean	4.7	13.0
SD	22.0	40

Table 3.1: Table showing tweet numbers before and after cleaning. Mean relates to the tweets per user and SD stands for standard deviation.

²<http://archive.org/web> [Accessed 20 June 2016]

³<http://bbc.co.uk/news>

⁴<http://bbc.co.uk/news/uk>

⁵<http://radimrehurek.com/gensim/index.html>

3.1.2 Results

To compare the popular topics within the tweet dataset to the BBC News articles, and to discover threads more specific to local news, LDA was used to analyse the data. Topic modelling has been argued as effective in generating summaries of textual datasets, as previously discussed in Chapter 2 section 2.2; however, as this experiment is interested in a daily breakdown of news articles, and as LDA does not natively handle temporal attributes, the algorithm is run on each days' worth of tweets. These results are shown in Table 3.2. For each topic that LDA identified, the keywords were manually matched against the existing BBC News article dataset.

Date	LDA Results	Topic(s)
June 09	izzard farage #itveuref	Izzard and Farage debating EU on Question Time Boris Johnson debating EU on ITV
June 10	stewart university football	Rod Steward concert in Southampton University fined for not flying EU flag Start of Euro 2016
June 11	engrus, marseille, football	England vs Russia in the Euro 2016
June 12	adamlambert orlando	Adam Lambert performing with Queen at IOW2016 Orlando shooting
June 13	e32016 loveisland ropelet	E3 gaming convention ITV reality show Sales spam
June 14	fathersday orlando	Father's Day sales spam Orlando vigil held at Southampton Guildhall
June 15	rickastley50 orlando	Rick Astley releases new album Further conversation about Orlando vigil
June 16	england, game, watching	England vs Wales in the Euro 2016
June 17	southampton, house, dance	No clear topic
June 18	universal, 1955, hoy	No clear topic
June 19	father's day 12.99, soundcloud	Father's day Soundcloud's £12.99 streaming subscription fee
June 20	westquay, southampton, ghost	No clear topic

Table 3.2: Table showing the most prevalent LDA topics within the 9-20 June tweet dataset, with retweets removed. The associated events obtained from manually searching the BBC News article dataset.

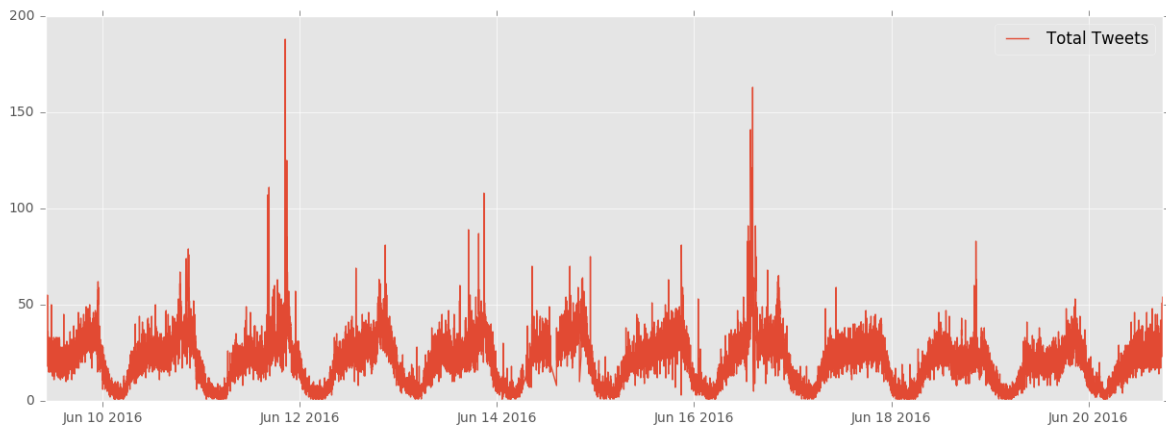


Figure 3.1: Graph representing total tweets over time for the 9-20 June 2016 Southampton tweet dataset, with retweets included. Figure generated using the Python module Matplotlib.

3.1.2.1 BBC News Articles

To compare the popularity of local versus national news articles within the Twitter dataset, keywords were manually derived from the articles and matched to the tweets using Python. For example, the title for the original article covering the shooting of Jo Cox, the local Labour MP for Batley and Spen on 16 June 2016, did not mention her name so the keywords ‘shoot’, ‘shot’, ‘MP’ and ‘Jo Cox’ were used to find matches as these terms were prominent throughout the article text. For articles that had more specific titles the latter was all that formed the keyword set. The results are shown in Table 3.3.

3.1.2.2 Parallel Coordinate Representations

The LDA results showed the prevalent topics within the Southampton dataset. While this is useful in understanding which events and discussions are popular, it does not convey many attributes that belong to each event. As highlighted in the literature review, previous work into presenting events do so inconsistently, therefore there was a need to create a holistic model that was capable of representing both ground-truth and Twitter data simultaneously. These models can be used to understand the relationship between real-world events and their representation on Twitter.

A parallel coordinate graph was chosen as appropriate representation of the multivariate nature of events. As events can share similar attributes, such as attendance, duration and graphical scale, these attributes can be visually modelled on a parallel coordinate graph. A parallel coordinate graph is similar to a standard X/Y graph but has several

Date	Article Title	Tweets	Keywords
June 09	Remain target Johnson in TV debate	47	boris, johnson, remain
June 10	Rod Stewart and Tim Peake head honours	94	stewart, peake, honours
	Service marks Queen's devotion at 90	37	queen, birthday, service
June 11	Trooping the Colour marks Queen at 90	36	queen, trooping, parade
June 12	Guests brave rain at Queen's picnic lunch	18	queen, picnic, lunch
June 13	Vigils for Orlando held across UK	78	orlando, vigil
June 14	Vigils for Orlando held across UK	68	orlando, vigil
June 15	Farage and Geldof in flotilla face-off	39	farage, geldof, flotilla
June 16	MP dies in 'horrific' shooting attack	258	jo cox, MP, shoot
	England earn thrilling win over Wales	660	england, wales, euro
June 17	Potash mine worker dies in 'gas blowout'	1	potash, gas, explosion
	Disruption continues as train derails	0	train, derail, crash
	Human leg found under platform	0	human, leg, platform
June 18	Man in court over killing of MP Jo Cox	8	court, killer, cox
	UK astronaut Tim Peake returns to Earth	47	tim, peake, return
June 19	Jo Cox service mourns 'Good Samaritan'	28	cox, service, mourn
	Murray wins record fifth Queen's title	7	murray, tennis, title
June 20	MPs to pay tribute to killed MP Jo Cox	109	MP, tribute, cox

Table 3.3: Table showing a sample of main BBC News articles from 9-20 June 2016 and how many unique tweets collected from the Southampton bounding circle during the same period matched the keywords. Only tweets explicitly matching the keywords were counted. This avoided over-representation of popular topics that last several days.

Y-axes, allowing for more than one aspect to be represented and facilitating the modelling of relationships between these aspects. Events modelled in such a way can have their profiles compared against each other, highlighting the degree to which events are represented on Twitter.

The parallel coordinate axes labels are derived from existing literature; [Sakaki et al. \(2010\)](#) discussed spatial scale and affected population; [Middleton et al. \(2014\)](#) and [Smith et al. \(2017\)](#) focus on population and local instances but took for granted social media presence; lastly, [Pohl et al. \(2012\)](#), [Dou et al. \(2012\)](#) and [Middleton et al. \(2014\)](#) emphasised the presence of sub-events that comprise overarching events. Creating a parallel coordinate model capable of representing these previously separate aspects is beneficial for a holistic understanding of an event profile.

Therefore, derived from the literature, Spatial Scale reflects the total geographic impact; Spatial Instance is the scale of the specific event; Duration is how long the event lasted; Repeated is whether it is a one-off event or a sub-event that is part of a regular schedule;

Population is the number of people directly or indirectly affected (depending on which aspect is emphasised within the graph); lastly, Social Media Presence reflects the number of unique messages pertaining to that event. All of these aspects are derived and applied manually to the graphs, which were created in Microsoft PowerPoint.

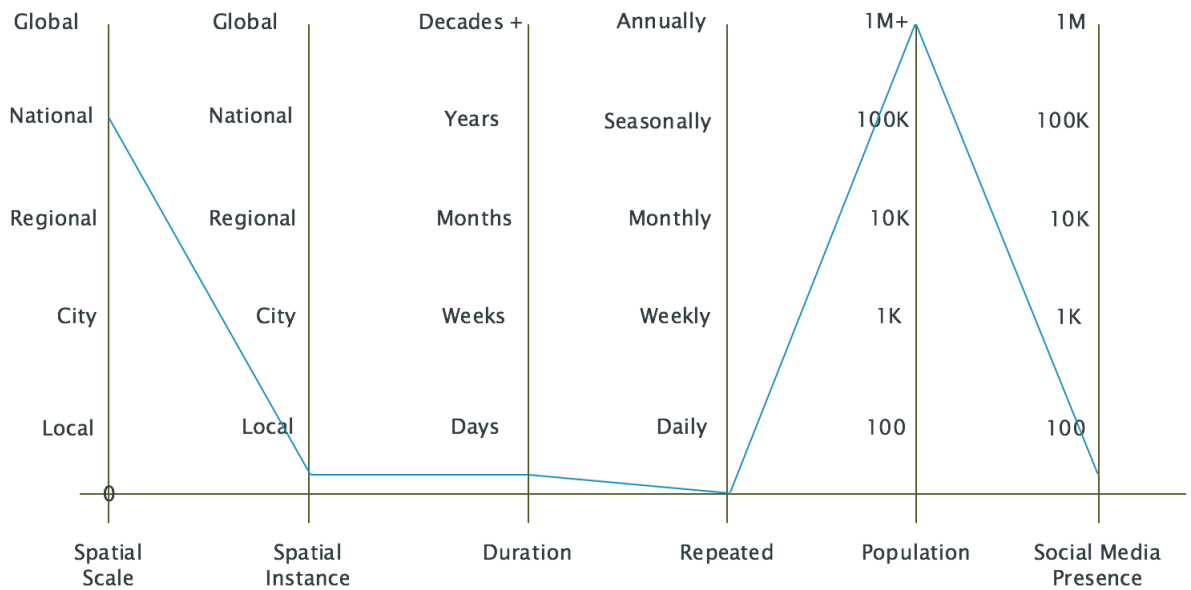


Figure 3.2: Application of the remain (Brexit) debate event to the model. Population here refers to total audience size — viewership was 3M.

Figure 3.2 covers the EU referendum debate aired on ITV on the 9th June 2016. As the debate is regarding Britain, the Spatial Scale was set to National, though arguably could be larger if the EU were included. The debate was filmed in a studio over the course of an hour, thus the Spatial Instance and Duration are appropriately low; additionally, the debate itself was a one-off thus is classed as a sub-event. The Population axis was harder to model; as the event was televised a count of 3M was taken for those impacted, derived from the viewership statistics⁶. This was taken as the number as those who watched a repeat or recorded it are not part of the sample. Lastly, only 12 tweets mentioned the event, thus it had a surprisingly low footprint.

Figure 3.3 represents the shooting at a nightclub in Orlando, Florida, USA, on the 12th June 2016. While the Spatial Scale could arguably be global, the event itself was in reference to vigils held across cities the UK in support of those affected, thus the Scale, Instance, Duration and Repeated axes reflect these attributes. The Population axis was an estimation of the attendance at each city event, as the Instance was at city level, though could have been raised to include all those who attended. The Social Media Presence was higher than the ITV debate due to more users sharing the article and expressing their support.

⁶<https://www.theguardian.com/media/2016/jun/10/boris-johnson-itv-eu-referendum-debate-gets-3m-viewers-nicola-sturgeon> [Accessed 24th June 2016]

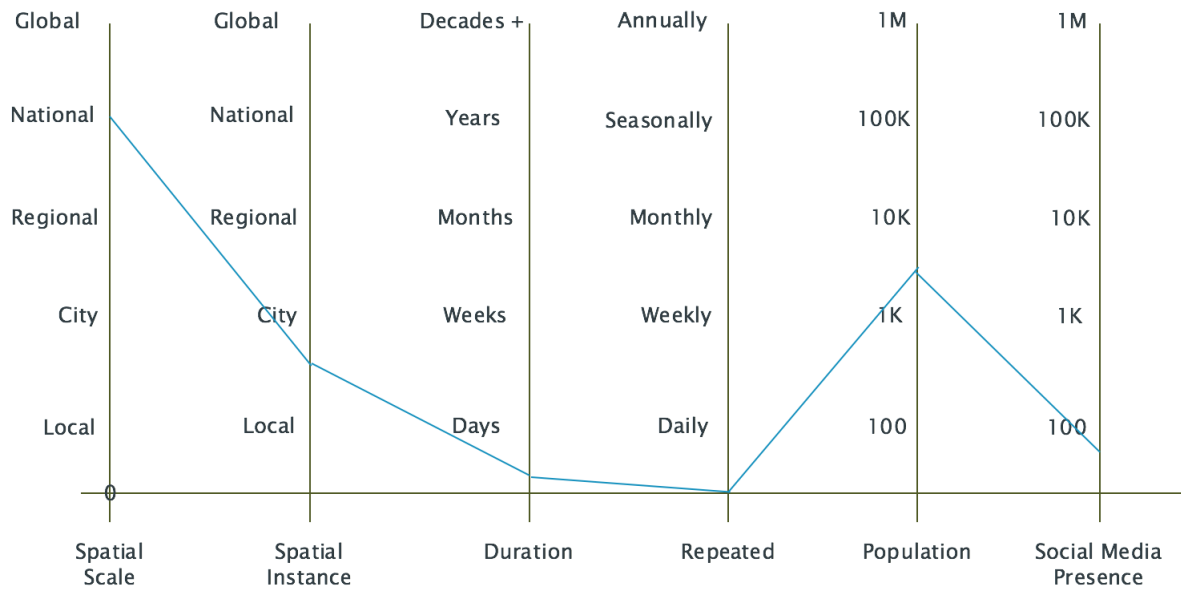


Figure 3.3: Model with representation of the Orlando vigil events, a vigil in memory of those shot during a night club attack in Orlando, Florida, USA.

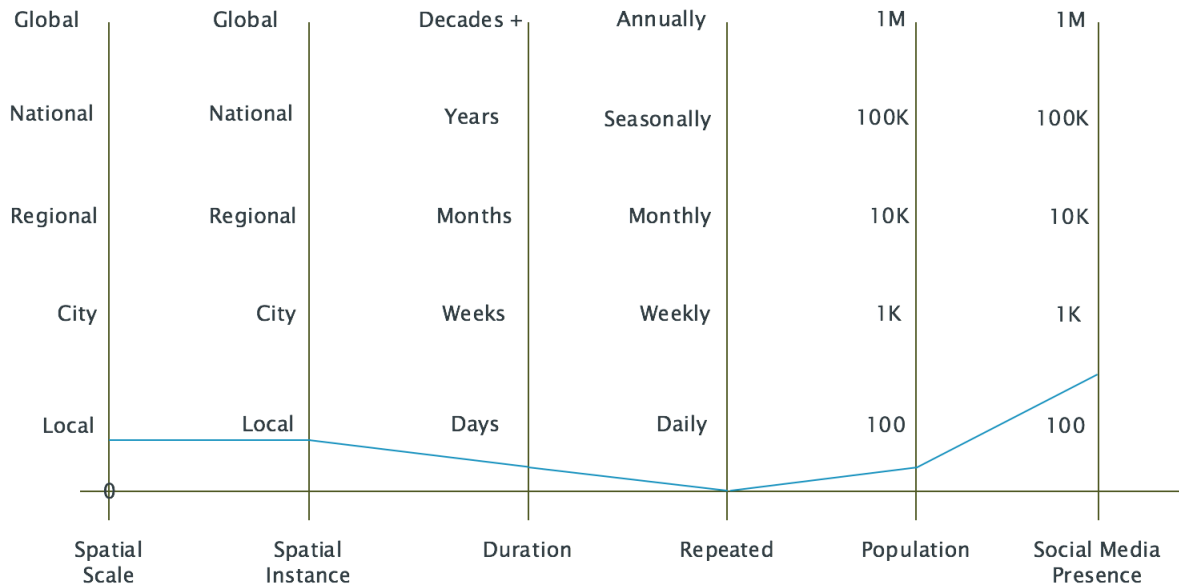


Figure 3.4: Model with representation of the shooting of MP Jo Cox.

Lastly, figure 3.4 shows the event profile for the shooting of Jo Cox on the 16th June 2016. The event was of national significance, as Jo Cox was a member of Parliament, the UK's legislative body. The shooting itself only happened within a small village called Birstall in the UK. The event was over quickly, and was a one-off. However, as the story impacted many people in the UK, there is a rise in Social Media Presence. These attributes are reflected on the graph.

3.1.3 Discussion

It is clear from tables 3.2 and 3.3 and graphs in Subsection 3.1.2.2 that the topics of conversation differ. When visualising a sample of these events, a challenge for the parallel coordinate model was modelling population values; these can be extrapolated from social media posts or television views, but establishing an accurate representation would require additional data, such as from the census or population survey. It did succeed in creating a novel way to graphically present the multivariate nature of events, as well as visualise the relationships between each of the components.

The ability for topic modelling to detect and extract national and local events from the tweets allowed for a direct comparison of topics deemed locally important with those published by a national news source. While many of the BBC News articles featured prominently in the Southampton tweet dataset, there were also local events of equal or greater popularity, such as the Isle of Wight 2016 music festival (IOW2016) which was prominent amongst users in Southampton but did not feature at all in the BBC News articles. The downside to this approach is the temporal aspect is too specific; the events discovered and topics detected are only snapshots of that time period, after which they fade out of local importance. These events are intrinsically interesting because of this, but in order to understand true local narratives a more thematic approach is needed that retains the original tweet context. Furthermore, estimating a population number proved challenging as ground-truth values could not be obtained for all of the events. Viewership figures exist for televised events but are not always published. Similarly, for those attending the Orlando vigils a count was likely taken by law enforcement, but is not publicly available.

3.1.4 Concluding Remarks

Data cleaning is a vital step in analysing Twitter data. As seen with in table 3.1, 57,611 users and 140,506 tweets were removed from the dataset as they had been classified as either a retweet or a content-poor tweet, with users removed if that content was all they provided. These high values were largely due to an undocumented bug in the API; when a user geolocated their account to Southampton, and when their tweets were retweeted (geolocated or not) the retweets appeared in the Southampton collection, regardless of where they originated. This was quickly fixed and was not an issue for any of the following experiments.

Thanks to the data cleaning methods, this experiment successfully identified prominent topics within the tweet dataset for Southampton from 9-20 June 2016. However, due to the tendency of topic modelling algorithms to remove contextual features from the text to create topics that require subjective interpretation (Pereira et al., 2017), this approach is unlikely to be preferred as the sole method of extracting specific topics.

As discussed in Chapter 2, a more specific keyword-matching approach was seen in the work by [Jeske et al. \(2017\)](#) to analyse the ‘#Heartbleed’ topic on Twitter. This reduced the subjectivity and easily identified tweets that related to this topic. The downside with this approach is that it would only extract tweets with that specific hashtag, thus tweets that may discuss the topic but use different terms would be missed. A term expansion method, proposed in the next section, aimed to improve this approach. The parallel coordinate model was useful in understanding which event features can be extracted from social media and online news publications, but did not produce the desired results when constructing narratives.

Therefore, to build upon the lessons learnt during this experiment and to better answer RQ1, the following experiment used keyword-matching including term-expansion to identify tweets relating to two specific themes. This approach departs from the notion of event detection and aligns more closely with identifying location-based thematic narratives from features ([Tomashevsky, 1965](#)).

3.2 Experiment 2: Thematic Kernel Density Estimation

After the previous experiment it was evident that an event-centric approach was not conducive to a holistic understanding of thematic spatial narratives as it was too restrictive and lost sight of trends and themes within the sample. It was then clear that a new method was needed, one that approached narratives from a more thematic and less ephemeral angle. Using the concept of features connoting motifs connoting themes from [Tomashevsky \(1965\)](#) and [Hargood \(2011\)](#), the experiment outlined in this section built upon the previous one by moving away from topic modelling as a primary focus and instead used keyword matching and term-expansion to identify relevant tweets, and expanded upon the location features present in the tweet metadata to construct location-based thematic narratives.

To achieve this the themes of ‘commerce’ and ‘entertainment’, two prominent themes identified by [Gattani et al. \(2013\)](#) and [Habernal et al. \(2013\)](#), are explored in a larger Twitter dataset. Instead of manually deriving keywords from news articles, an objective term-expansion model is proposed where the theme of interest is entered into an online thesaurus and related terms are collated into a keyword set. The thesaurus was used as a Web-based objective source of data and as an investigative process into how effective this type of data can be for discovering thematic narratives. Thematic tagging is proposed as an appropriate alternative to the event-centric approach of the previous experiment due to its closer relationship with thematic narratives. This approach is tested to see if the process creates a more thematic location-based narrative than the event-centric one from the previous section. The validity is tested by identifying which locations are popular for either theme and manually inspecting in which areas the nodes of activity

appear, such as within shopping centres, restaurant districts or office complexes. Part of this work is published in [Bennett et al. \(2017a,b\)](#).

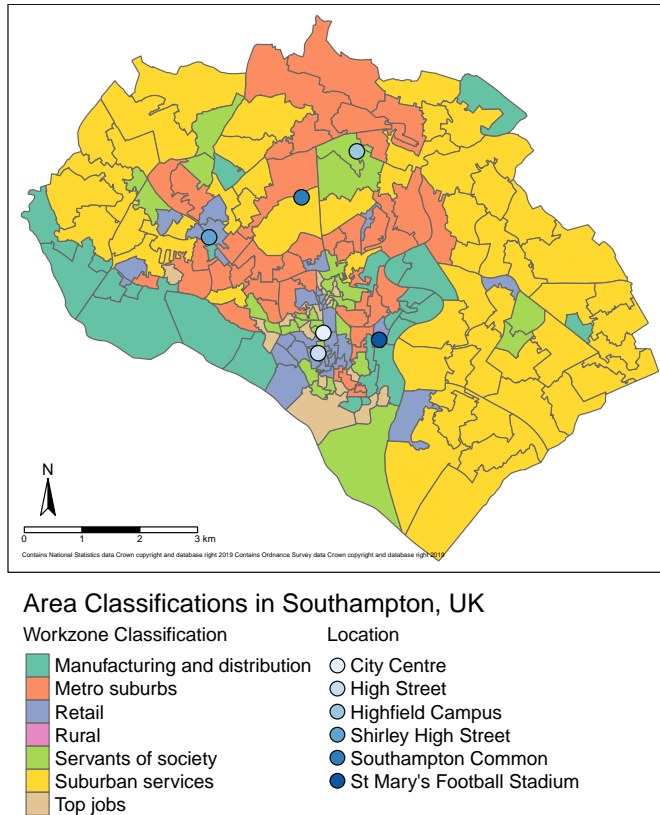


Figure 3.5: Map of Southampton, UK, with workplace zone classification data obtained from ONS. Used as reference to specific locations within the city.

Figure 3.5 shows the city of Southampton, UK, with the workplace zone (COWZ) data and specific locations highlighted, created using the programming language R and the software suite RStudio, for use as a reference point throughout the thesis. Additionally, all subsequent maps are generated using R. Workplace zones are derived from census and residential classification data published by the ONS and created by the University of Southampton ([Martin et al., 2013](#)) to visualise the areas of commerce and industry with greater precision than the previous residential classifications. As such, the method in which Southampton is divided is different, reflecting the workplace zones themselves. Of note are the retail zones including shops, restaurants and entertainment areas, which will feature prominently in the Twitter data analysis within this experiment. For future representations of Southampton in this thesis, the regular output area (OA) maps will be used as default as these areas are the smallest geographic output from the census and better reflect the overall physical geography. An ideal combination is applying COWZ to the OA maps, but it is currently not possible to directly map COWZ data onto OA divisions. Where appropriate, such as with Eastleigh and Winchester maps further on

in this experiment, the higher precision COWZ data was used to link tweets with their relevant area classifications

3.2.1 Methodology

To create a thematic dataset of tweets a larger collection was required to allow comparisons between cities. A bounding circle was placed over Southampton with a radius of 26 miles in order to achieve this, as this included additional cities and towns. In total, 3,375,107 tweets were collected between 1st November 2016 and 2nd March 2017. The tweets for this primary dataset were collected using Python and Twitter’s Search API. To compare against the bounding circle method of data collection and justify its expansion, two separate tweet collections were established using ‘Eastleigh’ and ‘Winchester’ keywords from the 19th January until the 28th March 2016. Eastleigh is a town approximately 8km north of Southampton city centre and Winchester is a separate city approximately 18km north of Southampton, both within Hampshire county in the south of the UK. The location of the whole study area will be shown in more detail in Figure 4.2. Collection of additional tweets aimed to understand whether searching for a town or city, which could match a tweet from anywhere in the world, could create a more localised and useful dataset than with the bounding circle. For this secondary dataset, 435,981 tweets were collected.

In keeping with the previous experiment and the suggestions of [Lansley and Longley \(2016\)](#) and [Lloyd and Cheshire \(2017\)](#), users with more than 3,000 tweets were capped at that level. Retweets were also removed from the primary dataset. In total, 993,109 tweets were removed, of which 3 were geolocated. This data is summarised in Table 3.4.

Tweets	Raw	Cleaned
Total	3,375,107	2,381,998 (70.6%)
Geotagged	29,181 (0.9%)	29,178 (1.2%)

Table 3.4: Table showing tweet numbers before and after cleaning for the main dataset.

3.2.1.1 Thematic Tagging

Manually obtaining keywords for natural language processing risks creating highly subjective results ([Habernal et al., 2013](#); [Ghermandi and Sinclair, 2019](#)); therefore, an objective approach was necessary to avoid biasing future results with subjective terms. To achieve this, a leading online thesaurus “Thesaurus”⁷ was chosen to obtain the keywords necessary for thematic tagging. The online database has an association value between

⁷<http://www.thesaurus.com> [Accessed 3rd March 2017]

terms; for example, if a term is strongly related to the original word it is given a 3, if it is weakly related a 1, such as ‘commerce’ having a strong relation to ‘business’ and a weak one to ‘merchandise’. This facilitated the population of a weighted thematic dataset. Furthermore, each of the strongly-related terms were also searched and their related terms saved to the dataset, along with their weighted relation score. Due to strongly-related terms appearing frequently within each search, once the final thematic dataset was collected terms that only appeared once were removed. This reduced the risk of noise created by weakly-related terms. Web scraping using the Python libraries “BeautifulSoup” and “Requests” was used to obtain the related terms, with each term in turn also being scraped to obtain a semantically expanded synonym set.

This process is carried out for the two terms ‘commerce’ and ‘entertainment’ based on the work by [Gattani et al. \(2013\)](#) and [Hasan et al. \(2013\)](#) who identified these themes as prominent when manually labelling tweet classifications. Using an online thesaurus as a source of keywords is proposed as appropriate as it is predominantly objective, barring the removal of weakly-related terms to increase relevance. The finalised thematic keyword datasets are used to search through the tweets, and tweets that match any of the keywords and have geolocation metadata are extracted. Geolocated tweets are preferred as they give a more direct representation of localised narratives than merely keyword matching alone, as well as geoparsing free-form text to extract location data being fraught with selection bias ([Middleton et al., 2014](#)); however, this could prove fruitful future research.

3.2.1.2 Kernel Density Estimation

To analyse the spatial relationships within the dataset, Kernel Density Estimation (KDE) is chosen as an appropriate statistical method (its merits previously argued in Chapter 2 Section 2.3). KDE creates a probability function to cluster tweets within certain areas and estimate these spatial densities. It is appropriate to use on Twitter data as tweets are non-parametric, meaning they do not conform to any theoretical distribution model. KDE is used to discover areas of density, which correlate to areas of Twitter activity. By using the GISTools library and the programming language R, a KDE map is created for each of the thematic tagging result sets, an example of which is shown in Figure 3.6. All KDE maps in this thesis use a grid of 200x200m and a kernel bandwidth of 0.002, which represents the level of smoothness rather than a distance. While calculating the ideal bandwidth was not a goal of the thesis, these values produce the most visually clear maps for analyses. For visualisation purposes, all density values lower than 0.001 are removed to tighten the plot around the points. If this step were not taken, the visualisation would create false positives of activity around areas without tweets. Similarly for the sake of communication, point density values are capped at 3,000; values greater than this only appeared in already high density areas and including them dramatically diluted the lower values to an extent where only the high values

could be identified on the map. The higher values likely identified areas of local density within already dense areas, thus future work into using Getis Ord and Local Moran's would better analyse these phenomena.

3.2.2 Results

Commerce	Entertainment
trade	fun
traffic	amusement
business	diversion
commerce	pleasure
barter	festivity
exchange	entertainment
truck	recreation
swap	celebration
market	enjoyment
dealing	joy
industry	frolic
retailing	gaiety
transaction	hilarity
transfer	refreshment
distribute	satisfaction

Table 3.5: Table showing the top 15 synonyms returned from the online thesaurus for the two themes. In total there were 135 terms for Commerce and 433 for Entertainment.

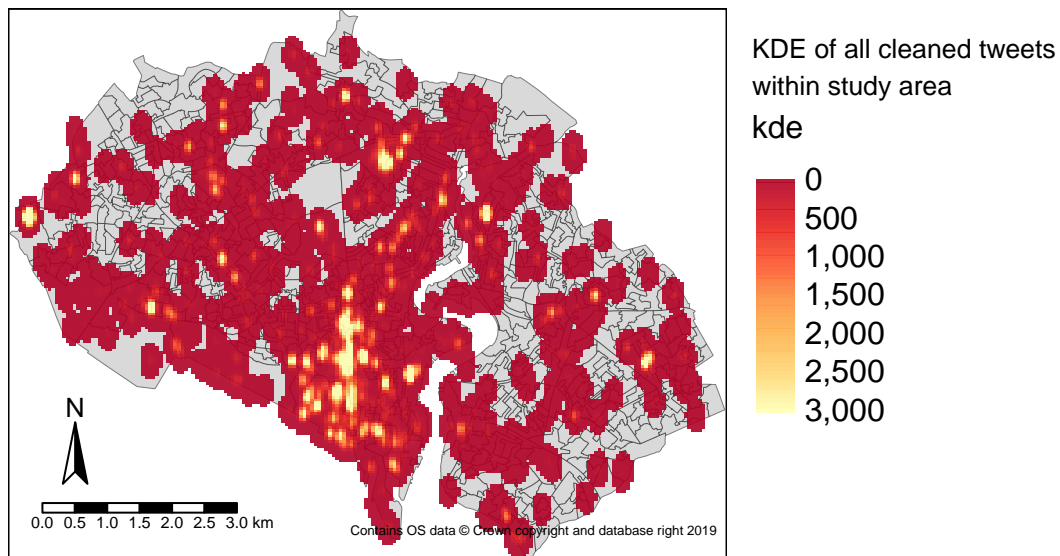


Figure 3.6: Cleaned KDE heatmap of geotagged tweets, centred on Southampton, UK.

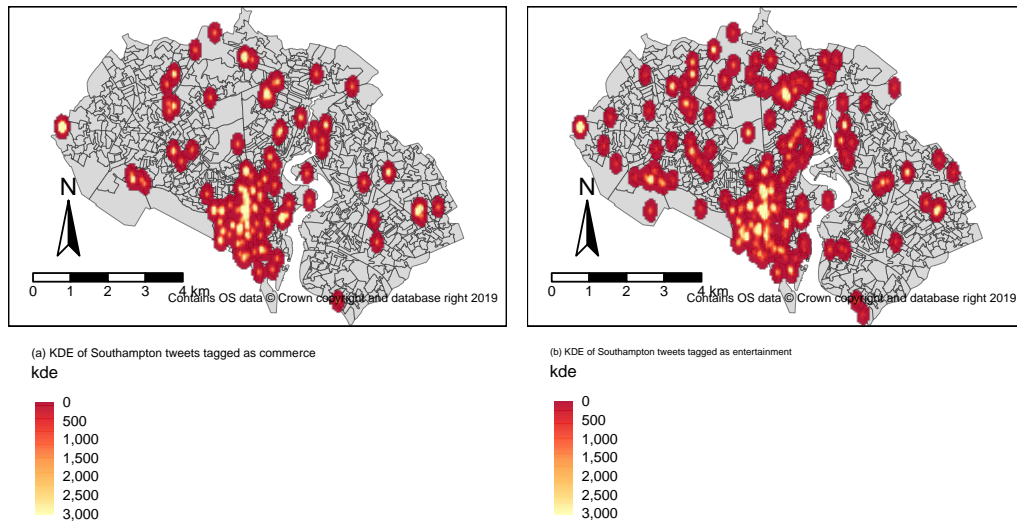


Figure 3.7: Cleaned KDE heatmap of geotagged Southampton tweets, tagged as (a) commerce and (b) entertainment.

The commerce tweets shown in Figure 3.7a follow a similar distribution to the overall dataset shown in Figure 3.6, with large clusters seen over the city centre. However, when comparing the ‘entertainment’ distribution seen in Figure 3.7b, the clusters are tighter around individual entertainment buildings, exemplified by fewer clusters and noticeable gaps along the centre. This is likely reflecting the respective patterns of population densities during commercial and leisure activities.

3.2.2.1 Comparison with Second Dataset

Tweets	Eastleigh	Winchester
Raw	40,770	395,211
Cleaned	37,844 (92.8%)	240,758 (60.9%)
Cleaned Geotagged	267 (0.7%)	10,059 (4.2%)

Table 3.6: Table showing tweet numbers before and after retweets and noisy users were removed for ‘Eastleigh’ and ‘Winchester’ keyword searches. Note: these are global values. There are mentions of ‘Eastleigh’ in Kenya and ‘Winchester’ in the USA.

Figure 3.8 shows the Twitter activity in Eastleigh, UK, collected using the keyword “eastleigh”. The tweets are concentrated around the large retail areas to the north, with some originating from within the manufacturing districts; this therefore explains why the KDE of the tweets shown in 3.8c are within the zone. The entertainment tweets in 3.8d are slightly more spread out, covering the same manufacturing zone as well as one of the retail areas.

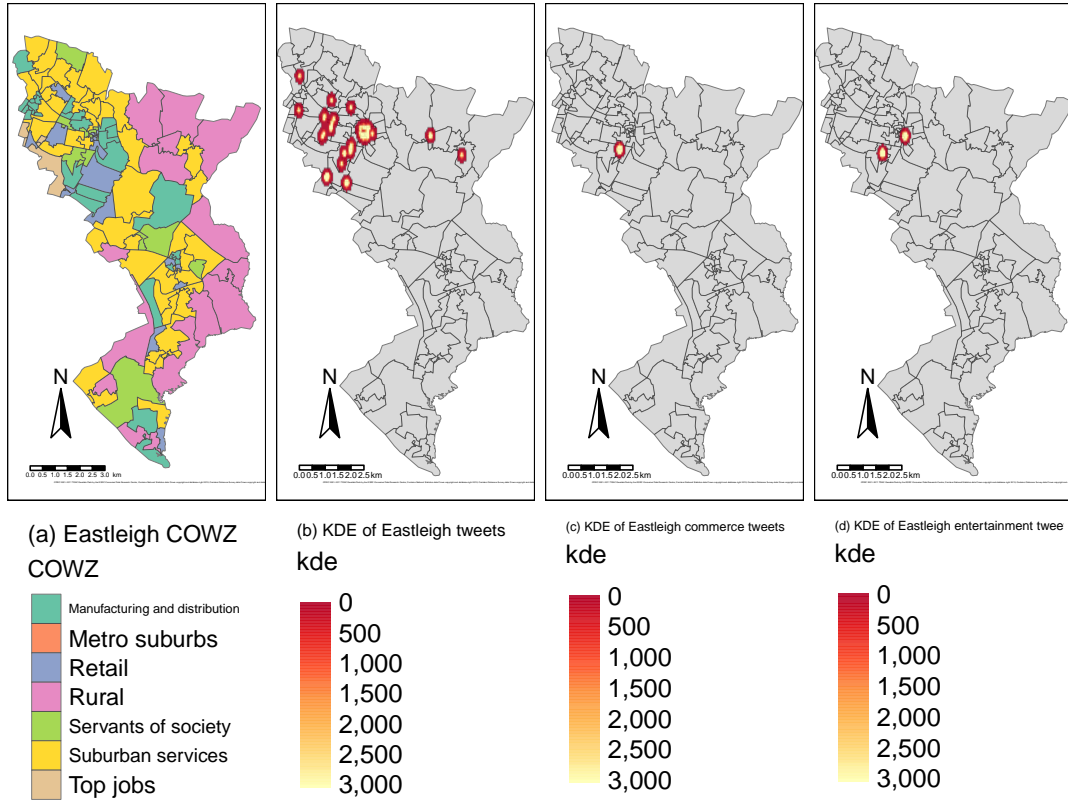


Figure 3.8: Maps of (a) Eastleigh COWZ classification, (b) KDE of cleaned tweets, (c) KDE of commerce tweets, (d) KDE of entertainment tweets.

Figure 3.9 shows the tweets collected using the “winchester” keyword. For a better visual representation, 3.9a shows the full extent of the Winchester city border with the subsequent maps zoomed in for clarity. There is a much stronger signal than in the tweets from Eastleigh shown in figure 3.8 due to the dataset being larger. As can be seen from map 3.9c, the commerce tweets are almost exclusively over the retail zones. Map 3.9d shows the entertainment tweets are spread out across the city centre, with most of them coming from the retail zones but also from many of the other zones. This is likely due to Winchester having its campus in the centre, suggesting that some of the clusters generated within the ‘Servants of society’ zones, which includes education (Martin et al., 2018), will be from the decentralised campus.

It is interesting to see the distinct difference in the proportion of geolocated tweets between Eastleigh and Winchester. This is likely due to ‘Winchester’ referring to a type of gun, with 502 tweets containing the word ‘rifle’ alongside ‘Winchester’ (these tweets

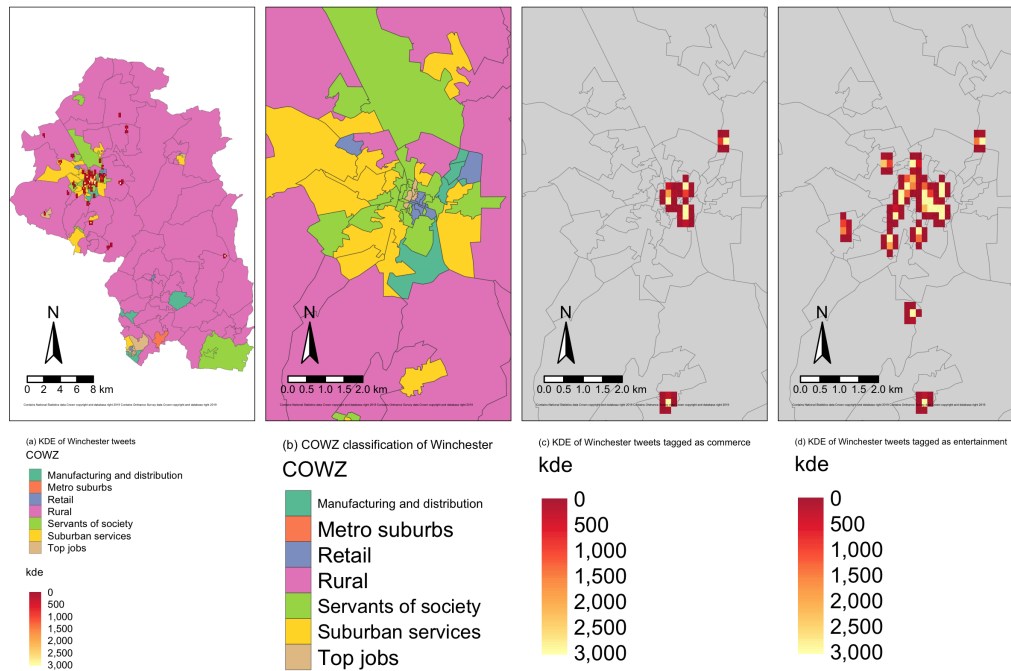


Figure 3.9: Maps of Winchester with (a) the COWZ classification, (b) a zoomed in area to better show the geography, (c) KDE of commerce tweets, (d) KDE of entertainment tweets.

were not mapped), as well as Winchester being a city with a larger geographic spread and therefore a likely larger social media footprint. The spatial distributions of the tweets are also different from those in Southampton, seen in figures 3.8 and 3.9, due to the different structural organisation of their centres.

3.2.3 Discussion

This experiment aimed to discover location-based narratives using term expansion as an improved approach to topic detection. The experiment was successful in modelling the different themes present within the two tweet datasets and clearly visualised their spatial relationships, suggesting a strong link between thematic content and location. It is to be expected that the retail areas would have the strongest signal due to the high daytime population density, which is reflected in the ‘commerce’ and ‘entertainment’ tweets and in the KDE maps. This is a promising suggestion that Twitter may be representative of the general population — at least within town and city centres (Mislove et al., 2011). More work would be needed to ascertain whether other themes, ones less

related to commercial districts, are as easily identifiable and have as strong a signal. This experiment was successful in using an objectively-obtained keyword dataset to create maps of text-based content, therefore is transferable to other geo-enabled media. The visualisations follow logical patterns within the city centres, for example in Figures 3.6 and 3.7 which focus on Southampton, there is a clear north to south pattern that correlates with the main high street. A similar pattern is seen for Winchester in figure 3.9. The same cannot be said for Eastleigh, shown in figure 3.8, due to the relatively low volume of geolocated tweets originating from the area. It can therefore be deduced that using datasets formed by keyword parameters alone rather than bounding circles are unreliable for constructing location-based narratives over short time frames in smaller urban centres.

An issue prevalent in NLP is the ambiguity of words. This was demonstrated by the term ‘work’, which was part of both the ‘Commerce’ and ‘Entertainment’ datasets, albeit with commerce having the stronger weight. This can cause false positives and over-fitting of keyword matches and therefore needs review for future experiments, such as using part-of-speech tagging to ascertain the manner in which ‘work’ is used and subsequently removing tweets that use it in undesirable contexts. While weighting of the term was used in collection of the thematic datasets and the tagging process, when mapping the tweets no weight was assigned to the tweets. Further work would need to incorporate this weighting to enable a more representative location-based narrative.

Using an online thesaurus as training data for thematic tagging has the benefit of being objective but the disadvantage of being too specific. When tagging the tweets, often those that related to commerce such as “I’m out shopping today” were not properly classified as commerce as the term ‘shopping’ is not strictly a synonym of commerce despite being related. In applying the narrative analysis from Tomashevsky (1965) and Hargood (2011), it became apparent that creating abstract motifs from features was very difficult to automate. ‘Shopping’, in this sense, would be a motif connoting a theme of commerce and concurrently a feature within tweets. To construct the abstract motifs a complex NLP algorithm would be required to understand the nuances present in the relationships between features. This is an expensive method to create a step that could conceivably be skipped in the narrative journey; themes can be made from motifs but also directly from features, thus this approach is favourable when considering the computational complexity of generating motifs. To further promote the removal of motifs in thematic generation, this experiment has shown how features generated from term expansion identified location-based commerce and entertainment themes without the need for abstract motifs.

Mapping themed tweets using KDE proved advantageous in reducing the noisy data and producing representative density surfaces of both high and low activity, showing that a thematic approach generated more realistic locative narratives than topic modelling alone, answering RQ1. A limitation of KDE is its dependency on a subjectively chosen

fixed kernel value for its smoothing algorithm. For instance, a kernel window of 20m^2 would create many separate modes (most popular locations) that would only represent a few tweets, whereas a kernel window of 200m^2 created the KDE images shown in this experiment. 200m was more representative of the modes and generated comprehensible visualisations. While a kernel window of 200m^2 was appropriate for this data, it is still set manually and thus risks over-smoothing busy areas and under-smoothing sparsely populated areas. Adaptive KDE discussed in Chapter 2 Section 2.3 could potentially solve this issue but was too computationally complex to fit within the scope of this thesis.

3.2.4 Concluding Remarks

While the KDE maps show distinct activities within their applied locations, the coordinate data obtained from the tweet metadata are not uniform across the dataset. After this experiment concluded it was found that tweets linked to points of interest (POIs) often do not include the correct coordinate metadata, thus these clustered tweets were included in the main set as if they were genuinely user-generated when in fact they were computationally derived. This is exemplified by the strong central clusters seen in figures 3.6 and 3.7. While it might appear that these clusters represent hubs of activity, the tweets are artificially placed there due to users ‘checking in’ to the city, with the tweet coordinates derived from Twitter’s gazetteer of city centroids⁸.

This is a similar issue for tweets from third-party sources, such as Instagram, where a user will share several photos but each tweet will have the exact same pair of coordinates, artificially inflating that particular location. Furthermore, user-specified privacy settings can also blur the true locations of tweets by artificially clustering around nearby neighbourhoods or at city level without corresponding metadata. These features artificially inflate the supposedly natural clusters and thus the KDE representation is skewed. The following experiment outlines this discovery, develops a new method to identify unwanted third-party sources and tweets that are missing key metadata, and proceeds to separate these misleading tweets from the main dataset. The remaining tweets are therefore more representative of the true location of the user and thus reflect a more accurate location-based narrative. That is not to say the clustered tweets do not form part of their own location-based narrative but, as the next section will show, the geographic resolution of these clusters cannot easily be deduced.

⁸This is an assumption, there is no official documentation to explicitly explain centroid locations, but it is a repeated pattern seen in later experiments when analysing clusters in more detail.

3.3 Experiment 3: Stratifying Location Resolutions

The previous two experiments showed it was possible to discover location-based narratives from Twitter data. The key lessons learnt from them were that an event-centric approach only creates snapshots of narratives and that spatial analyses are highly reliant on the quality of the geolocation metadata within the tweet. Since the completion of the second experiment, it became clear that there were some anomalies with the geolocation information. These included the large clusters in the centre of the city being created in part by Twitter normalising tweets to the city's central point, as well as tweets being artificially clustered to nearby neighbourhoods but without corresponding metadata declaring the altered type. Additionally, tweets from some third-party sources flood the dataset with hundreds of tweets with identical coordinates, such as check-ins or advertisements. This created misleading impressions of certain areas having high activity. It was therefore necessary to improve the methods to recognise this quirk of the Twitter API and resolve the different spatial granularities.

This experiment achieved this by critically analysing the quality of the tweet coordinates. Tweets were clustered by the coordinate metadata field and queried against an open API to extract address information. Results showed that some clusters had been created by users checking in to a particular venue, though there was no indication of this within the metadata. Analysing the data in this way aided in identifying the spatial granularities inherent in tweets, a technical approach that was not sufficiently explored in the literature. This led to more reliable conclusions as the spatial data were no longer taken at face value. Furthermore, with the ability to distinguish between granularities, using the Twitter data as representations of the population became more appropriate; therefore, this experiment also compares the cluster locations to population counts from the 2011 census. Most of this work is published in [Bennett et al. \(2018\)](#).

3.3.1 Methodology

The two most significant differences employed by this third experiment are to focus on tweet source and location resolution. Tweet source refers to the application that generated the tweet, such as 'Instagram', 'Facebook' or various news and sports websites. It became apparent that some of these sources, especially the latter, create geolocated tweets but without meaningful narrative contribution, such as a local Southampton newspaper geolocating all their tweets to their headquarters. While these types of tweets would be useful in identifying buildings that produce such data, they do not contribute towards a wider understanding of the social use of space and are thus outside the scope of the thesis, therefore it is important to remove these tweets from the dataset, an approach corroborated by ([Gilani et al., 2017](#)). Additionally, as this experiment seeks to identify tweets that share identical coordinate pairs, such as several Instagram posts

being shared at the same time, when these tweets are discovered they should also be removed from the primary dataset to form a secondary dataset for further analysis. These tweets share the same coordinate field either due to the user telling the Twitter API to only geolocate their tweets to a nearby neighbourhood or city centre, or Twitter’s geolocation parser assigning them the same coordinates. Clustering these tweets to identify neighbourhoods is a relatively straightforward task; however, difficulty arose in distinguishing between neighbourhoods and POIs. Therefore, this experiment also aimed to identify the differences between these two cluster types.

3.3.1.1 Pre-processing

Similar to the previous experiment, a 26-mile radius bounding circle was used covering Southampton and wider Hampshire between November 2016 and April 2017. However, the data collection had been continuing in order to collect more geolocated tweets to test the robustness of the framework. For this experiment, an accumulated 5,000,098 tweets were collected from 640,044 unique users within the time frame and bounding circle. Of the total tweets, 36,471 were geolocated. Once 1,914,672 retweets were removed, the dataset stood at 3,085,426 tweets by 54,589 unique users, of which 36,468 were geolocated, highlighting the lack of geolocation information attributed to retweets. While the large collection includes neighbouring towns and cities, for this experiment the spatial extent was restricted to the city of Southampton to aid with mapping and more nuanced analyses. The final geolocated tweet count that is used for this experiment is 28,393. This pre-processing step removed a great deal of irrelevant tweet data; however, clustered tweets and unwanted sources still remained. Unlike the previous experiments, the tweet cap was not set at 3,000 as the experiment spanned a much longer time period than in the work by [Longley and Adnan \(2016\)](#) and [Lloyd and Cheshire \(2017\)](#).

3.3.1.2 Source Analysis

There were 67 unique third-party sources that had at least one geolocated tweet within the 28,393 tweet dataset. To identify which of these sources produced an excessive amount of location data they were ranked by tweet count and compared against the number of users who used each source. The sources that produced a large ratio of tweets to users were removed. For instance, Table 3.7 shows the source “Southampton FC chat zone”, an automated source that shared news and betting information about the Southampton football team at a ratio of 58054:1. Similarly, in Table 3.8, the automated source ‘dlvr.it’ produced 4,313 geolocated tweets by 4 users, a ratio of 1078:1, creating a disproportionate contribution. In comparison, 21,065 ‘Instagram’ users produced 5,340 geolocated tweets at a ratio of 4:1. Issues with this method arose when the sources produced few tweets by fewer users, for instance ‘Wordpress.com’ produced 5 geolocated tweets by 2 different users. Arguably, sources with fewer than 5 users are unlikely

Source	Tweets	Users
Twitter for iPhone	1,245,506	28,765
Twitter for Android	487,752	11,252
Twitter Web Client	394,828	14,189
IFTTT	108,318	251
Twitter for iPad	74,394	3,362
Facebook	69,418	1,443
Hootsuite	67,186	800
TweetDeck	58,733	843
Instagram	58,223	7,649
Southampton FC chat zone	58,054	1

Table 3.7: Breakdown of top 10 Sources

Source	Tweets	Users
Instagram	21,065	5,340
dlvr.it	4,313	4
Blue Rhinos Web Services	3,142	2
Foursquare	1,719	214
Twitter for Android	806	104
Untappd	625	62
TweetMyJOBS	533	18
Tweetbot for iOS	512	42
Twitter for Windows Phone	299	16
Twitter for iPhone	251	52

Table 3.8: Breakdown of top 10 Geo Sources

to contribute significant narrative data; however, ideally these sources should not be dismissed automatically as it does not future-proof the framework against popular new sources or under-represented communities.

Despite the attempt to include all sources, these smaller sources (a sample of which is shown in Table 3.9) that had either 5 or fewer users or 10 or fewer tweets were analysed manually and found to contribute no meaningful narrative data and were removed. Many of these sources also included the undesirable accounts, such as ‘dlvr.it’ with its 4313:4 tweet to user ratio. Some of the sources shown in the table are from genuine users, but have such a small impact on the dataset that they do not contribute meaningful narrative information. It is perhaps a consequence of quantitatively analysing large-scale Twitter data that in order to remove noise and improve clarity these small online communities will not feature in narrative analyses.

Source	Users	Tweet Count	Example tweet
Crowdfire - Go Big	5	8	'I use Product A to get myself going!'
dlvr.it	4	4313	'#Boats: Youth Regionals, all smooth?'
Fenix for Android	3	5	'Chrome keeps crashing'
Talon (Plus)	2	19	'Got tickets to the Game'
TweetLogix	2	12	'Happy birthday Jane'
Blue Rhinos Web Services	2	3142	'Boat has just set sail for Flensburg'
WordPress.com	2	5	'Blogging is tough http://...'
automicv12demo	1	198	'enjoy 30% off all products all day today'
Ratlake Transponder	1	109	'ping Vespa has entered geofence Ratlake'
mtvan.com news feed	1	13	'Courier available Soton to Basingstoke.'

Table 3.9: Example of geo sources with 5 or fewer users (text from human users anonymised).

3.3.1.3 Stratifying Coordinate Spaces

To understand the impact of user privacy settings and third-party geolocation parsing on the level of tweet aggregation in the dataset, tweets were clustered by their coordinate field. In total, 1,888 clusters comprised of at least two or more identical coordinate fields with 3,290 unique tweets remaining. The mean cluster size was 12 and the median was 4, thus to stratify the tweet dataset the clusters of 4 or more were removed and added to a secondary dataset. As the coordinate field can be up to 8 decimal places in accuracy, it is highly unlikely that tweets would share the exact same coordinates unless they are stationary, such as originating from a desktop or laptop computer. Despite the low likelihood, these stationary tweets are still organically created thus clusters of fewer than 4 tweets were kept in the main dataset with clusters greater than 4 being saved to a secondary dataset. The primary and secondary datasets are mapped in figures 3.10 and 3.11.

3.3.2 Results

These maps reflect the clustering results, with the primary dataset containing more of a spread and the secondary dataset concentrating around particular areas. The first map shows the small clusters, including 13,465 tweets. The second map shows the remaining larger clusters and includes 96,518 tweets. The large central cluster is more pronounced in the second image, situated at the northern end of the High Street in figure 3.11. To determine whether these clusters were generated by the user checking in to a POI or genuinely from stationary devices, it was necessary to query a data source for information. For instance, if a particular cluster rests over a point of interest then it would be logical that those users who are part of the cluster have tagged themselves at that point of interest. The difference comes when they have stated they are at a particular location but their privacy settings diverts their coordinate field to represent a

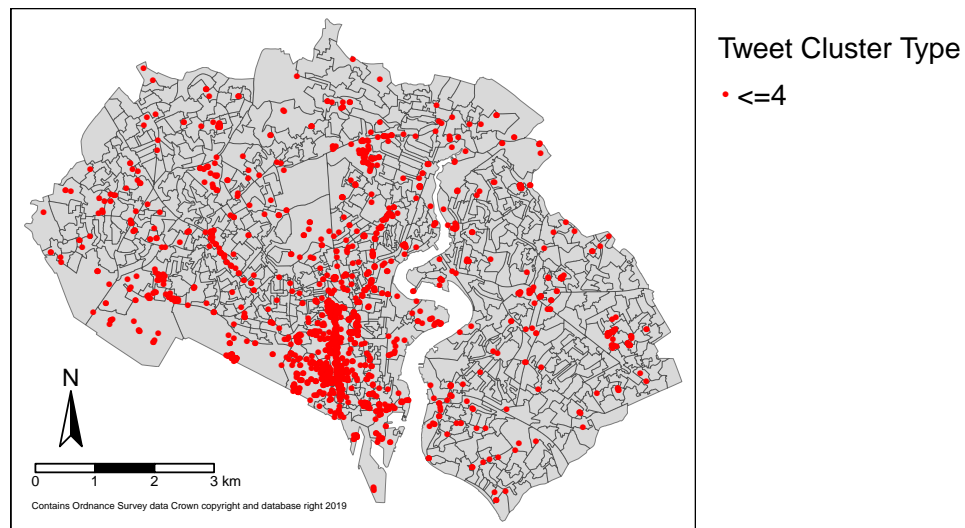


Figure 3.10: Map of Southampton showing the tweet clusters comprising of fewer than 4 tweets, a total of 13,465 tweets.

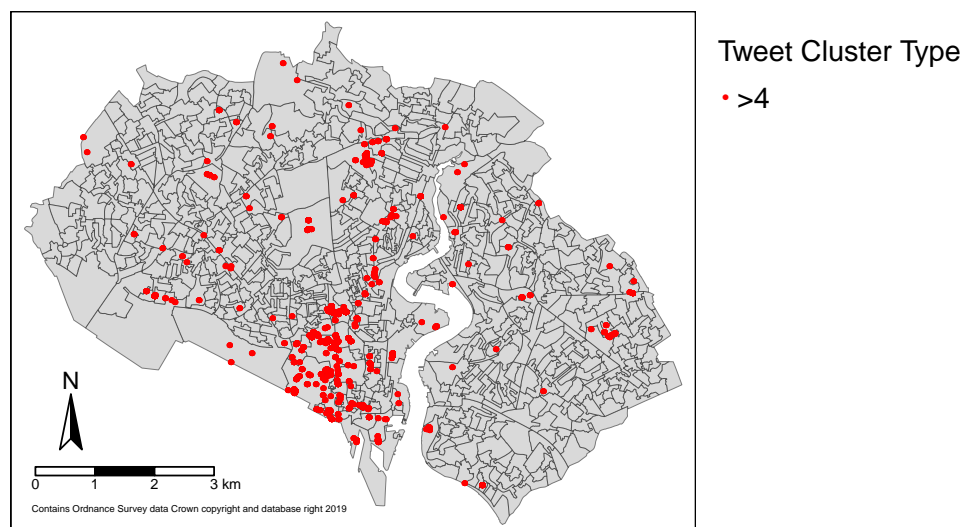


Figure 3.11: Map of Southampton showing the clusters comprising more than 4 tweets, a total of 96,518 tweets.

generalised neighbourhood area. As Twitter uses Foursquare's API to define their POI coordinates, it seemed reasonable that their neighbourhood clusters would share similar coordinates with Foursquare's database; therefore, the Foursquare API was queried with

the tweet clusters to distinguish POIs from neighbourhoods.

To distinguish between city-level and neighbourhood clusters, the clusters from the second dataset were ranked by how many unique users shared their coordinates, the theory being that clusters with more unique users in them were likely to be at city and neighbourhood level rather than genuinely produced clusters. In total there were 301 clusters with 228 comprising at least two unique users. The top cluster had 631 unique users with the second only containing 194, signalling a distinct difference. This distinct difference highlights the most popular one as the city-level cluster.

The neighbourhood clusters were then queried against Foursquare’s POI database. When defining the search terms within the API, a tolerance of 25m was allocated the results, allowing for parent POIs to be returned rather than individual locations; for example, St Mary’s football stadium in Southampton has a few component POIs, thus when tweets are returned without a tolerance they often get improperly classed as a fast food restaurant rather than ‘St Mary’s’, despite manual inspection clearly showing the tweet within the grounds. The tolerance also mitigated against potential GPS inaccuracies preventing a POI from being returned, such as the large metallic structure of St Mary’s interfering with signal. This querying resulted in 129 POI matches with 99 not returning a result, indicative of a neighbourhood cluster. These results are mapped in figures 3.12 and 3.13.

3.3.2.1 Applying Census Data

The ability for Twitter data to represent the population has been previously debated in the literature (Sloan and Morgan, 2015; Hamstead et al., 2018). To extend the usefulness of this experiment beyond cluster analysis, census data were included to compare cluster location to population and conclude upon its validity.

The inclusion of census data allows for a comparison of densely populated areas and the identified clusters, adding a visual analysis layer. To create a cartographic representation of population distribution across the city in relation to neighbourhood and POI clusters, Output Area census data was obtained from NOMIS⁹. As Output Areas are the smallest geographic representation of population, these provide useful units for comparing against mobile Twitter users. Figure 3.12 shows a fairly even spread of neighbourhood clusters, both in terms of geographic spread across the city and distance from each node, as to be expected. Figure 3.13 show strong clusters of POIs around the city centre and High Street, which is also to be expected as these match with shops and monuments which are more common in the city centre. This strengthens the argument for the differentiation between neighbourhood and POI clusters being an appropriate experiment and one that would return realistic results. However, in terms of POI and neighbourhood cluster

⁹<http://nomisweb.co.uk/census/2011> [Accessed 13 May 2018]

locations, these do not seem to match any pattern observable in the census data. A more useful map layer would perhaps have been one that classified Output Area types, such as residential or commercial districts.

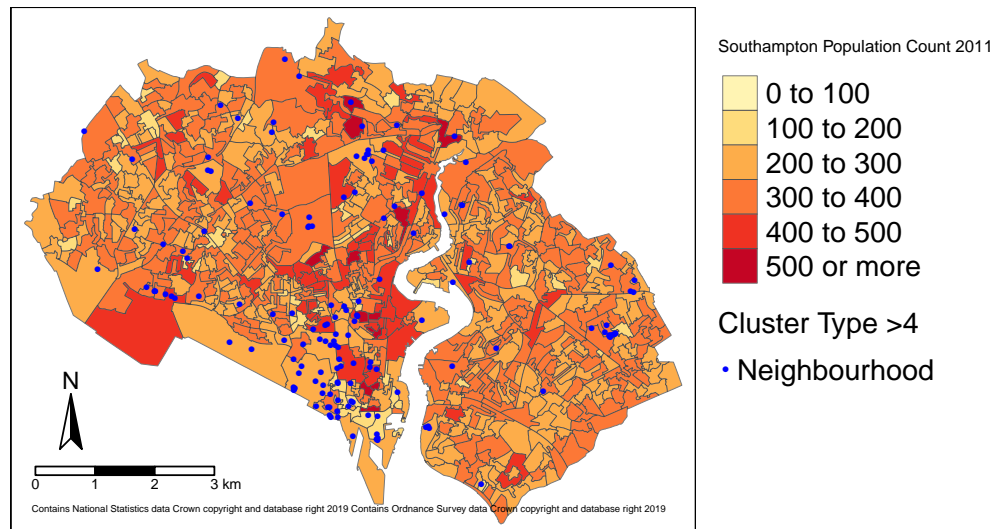


Figure 3.12: Map showing the clusters classified as neighbourhoods. The shading relates to the population count within the Output Area.

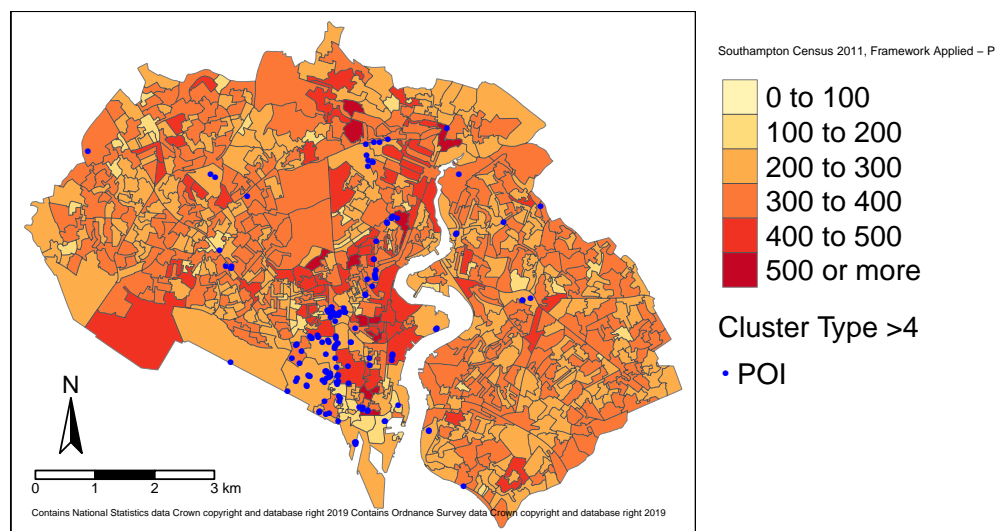


Figure 3.13: Map showing the clusters of POIs, emphasising how many there are in the dataset that would otherwise have been overlooked. The shading relates to the population count within the Output Area.

3.3.3 Discussion

This experiment aimed to discover and resolve geographic anomalies found in Twitter data. It also aimed to create an experiment capable of processing tweet sources and removing those deemed irrelevant, to then feed in to an overarching framework. To this extent, the experiment was successful in these two endeavours as it was able to resolve tweet geolocation resolution and identify problematic sources. It was able to distinguish between neighbourhood and POI clusters and identify and remove sources that contributed a disproportionate number of tweets. By manually inspecting the tweets from the removed sources it was clear that they were automatically generated, such as tweets from ‘Blue Rhino Web Services’ creating a great deal of geolocation information but only pertaining to the travel routes of a ferry, thus not contributing relevant narrative information, though would be of interest to those investigating mobility patterns. Similarly, analysing tweets from the remaining sources showed a high level of narrative information, such as tweets from ‘Instagram’ and ‘Foursquare’ contributing opinions, discussions and locations of individuals.

Future work would need to focus on identifying naturally generating clusters, such as desktop computers, and automating the removal of smaller third-party sources. Related to this is the removal of clusters generated by single users who repeat, or ‘spam’, the same tweet more than once, as this creates misleading spatial representations of activity. It would also be useful to generate statistics that relate population density to the spread of neighbourhood and POI clusters, enabling an understanding as to whether population density determines neighbourhood cluster locations or whether they are randomly generated as Twitter does not divulge this information. An issue that arose during the experiment is with coordinates with fewer than 5 decimal places, potentially caused by the tweet sources having differing levels of numeric precision. This resulted in the coordinates being up to 11m off its true location, causing issues with some API requests as many tweets from St Mary’s stadium were appearing at a local fast food outlet despite their text content referring to a current football match.

3.3.4 Concluding Remarks

This experiment enabled the differentiation of tweets at POI and neighbourhood resolution. Tweets were previously taken at face value, meaning that artificial clusters generated by third-party applications created misleading activity areas, as shown by the KDE analyses. However, there existed available data that could more precisely label a tweet as belonging to a POI or neighbourhood area that was not part of this experiment. When a tweet is sent from a third-party source like ‘Instagram’ or ‘Foursquare’, a URL is automatically appended to the tweet message that links back to the original post. This original post contains the full text, frequently cut short due to Twitter’s

character limitation, as well as often a venue with its associated information. Accessing this level of detail would greatly improve the quality of the location data, as knowing from which venue the tweet originated would further refine the POI and neighbourhood classifications. Therefore, the following experiment addressed this new source of data.

3.4 Experiment 4: Resolving Third-Party POIs

The previous experiments contributed to an overall framework for detecting location-based narratives by improving thematic extraction and location modelling. Keyword matching was shown as an effective method of extracting specific themes and their related tweets, while KDE was used to model their spatial relationships. A key behaviour discovered within the tweet dataset was the undocumented clustering of tweets from third-party sources around either points of interest, neighbourhoods or city-level centroids. It was therefore necessary to investigate this behaviour and extract more precise location information from the original sources in order to better represent the tweets' spatial resolution.

This was achieved by analysing tweets from 'Instagram' and 'Foursquare', two popular online social media platforms that allow users to create and share location-based content, to obtain the original post information. These two were selected as they were the most frequent non-automated third-party sources that produced geolocated Twitter data (as was shown in table 3.8). Through visual analyses of tweet clusters it became evident that tweets from these third-party sources were often placed erratically around what should have been a central POI node.

3.4.1 Methodology

To focus the experiment to a specific complex area, WestQuay Shopping Centre was chosen as the proof-of-concept location. WestQuay is a large shopping centre in Southampton that is the host to a multitude of shops and restaurants. In order to identify from which shop or restaurant (POIs) a tweet was sent, it was necessary to obtain a dataset of POIs within WestQuay as well as boundary data for the building. This was obtained from manually querying the online database Edina Digimap¹⁰, after which they sent the corresponding data within a few minutes. The boundary data came in the form of a geospatial database, while the POI data was a spreadsheet. Both were loaded into the industry-leading geospatial analysis software ArcGIS for analysis.

Tweets were collected from the 1st November 2016 to 4th August 2018 using the same 26-mile bounding circle as in the previous two experiments, though as this experiment was conducted long after the initial data collection was started, a considerably larger

¹⁰<http://www.digimap.edina.ac.uk> [Accessed November 2018]

tweet dataset had been obtained. In total, 24,834,206 tweets were collected, of which 176,315 were geolocated. When importing the tweets into ArcGIS, only those that fell within the WestQuay boundary were used for the experiment. After undesirable sources identified in the previous experiment were removed, the final dataset stood at 958 geolocated tweets from the 27th November 2016 until 3rd August 2018.

3.4.1.1 Analysing Third-Party URLs

To discover what extra information can be gathered from the original post it was important to understand what was currently available. Though other third-party applications existed, only tweets from the two largest third-party sources ‘Instagram’ and ‘Foursquare’ were included. The two social media platforms allow their users to express themselves in many more characters than is allowed on Twitter. When a user creates a post on these platforms and shares it to Twitter, Twitter will cut off the message if it exceeds the 140 or 280 character limit. For example, a common occurrence is to see a tweet text such as: “Having a great time, nice drinks and food at this place. Great atmosphere, fine people. - at @Ve...”, where the name of the location was cut short. Another similarity between these two platforms is a URL in the tweet text that links to the original post, as well as the original post often having a further link to the venue page, which includes its name and location. Other third-party sources also created tweets with a URL that linked to the original post; however, the nature of the link changed depending on the source. For example, tweets from sources promoting job offers often have time-sensitive URLs, thus when visited after a certain period the researcher will be presented with an error as the advert has been removed. This therefore promotes ‘Instagram’ and ‘Foursquare’ as two reliable sources of information as their URLs do not expire unless the user has deleted the original post.

3.4.1.2 Technical Approach

Within each tweet from ‘Instagram’ and ‘Foursquare’ is a URL that links the user back to the original post. As it is against Instagram’s terms of service to computationally collect information, no data were saved from the site; instead, the coordinates and text within the metadata were manually compared against the original tweet and if they differed then this was logged. Fortunately, Foursquare allows a reverse check-in lookup. This involves querying their API with the check-in ID from the tweet URL, which returns the metadata for the original post such as text, venue name and location. This was done using Python and therefore means that similar experiments using this method are possible with a small sample of Instagram posts, but if there are thousands then it would require a team to manually inspect the pages which makes it less desirable in larger volumes.

Tweet Source	Total	Error	Remaining	+Text	+Geo
Instagram	729	134	595	458	474
Foursquare	184	0	184	10	171
Total	913	134	779	468	645

Table 3.10: Tweet sources before and after resolving POI URLs. Some were private, had no venue data or venue coordinates and are grouped as “Error”. +Text means if the URL-scraped text was longer than the original tweet text and +Geo means if the scraped POI coordinates differed from the original ones.

One key advantage to this method is the ability to obtain the full text of the original message from Foursquare, rather than the often shorter tweet text. This longer text often had the full name of the mentioned venue which would otherwise have been cut off, as well as extra narrative information. To understand which venues were being discussed by shoppers, all mentions were extracted from the tweet text and the Foursquare metadata. Mentions are created in a uniform manner whenever users check in to a location using a third-party app, such as “Having a good time - at @Venue”. Only whole matches were counted, thus if a post originated from Instagram and had its text content shortened by Twitter, this could potentially have removed some matches.

These mentions were compared against the POI list from Digimap and resulted in 18 matches with 121 POIs not having a corresponding tweet and 54 venues originating in the tweet text not having a corresponding Digimap POI. These results are summarised in Table 3.11. This is due to a number of factors, primarily the time at which the POI dataset was collected may have resulted in some newer POIs that featured in the social media not being present in the POI database, as well as former POIs that have subsequently been closed down or moved being retained in the Digimap data. Another issue was the POI set from Digimap included the smaller boutiques, which are relatively ephemeral compared with the bricks-and-mortar shops and thus less likely to have a presence within the social media data. While this particular problem is specific to Digimap data, it highlights potential problems for other interested parties when analysing social media geospatial data, as existing government- or industry-produced POI data may be out of date.

Source	POIs	Digimap	Instagram	Foursquare
Digimap	137	-	9	9
Instagram	44	9	-	6
Foursquare	28	9	6	-

Table 3.11: Text-based POI matches between each data source.

3.4.2 Discussion

A key result from this experiment was the discovery of the usefulness of tweet URLs. It was apparent that some tweets originating from third-party sources had erroneous coordinates in the Twitter data; for example, if the tweet text said “Reading a book - at @Waterstones_Westquay”, the tweet coordinates would place the post several metres outside the building while the Foursquare API would return the proper location for the shop to which the tweet is geotagged. This is either due to Twitter not receiving adequate metadata from these third-party sources and thus attempting to resolve the tweet coordinates itself, or the original coordinates of the POI from the third-party source being incorrect¹¹. This inconsistency means all geolocated tweets from third-party sources need to be scrutinised. As shown in Table 3.10, the inaccuracy was strongly present in the ‘Instagram’ tweets as ‘Instagram’ uses its own location gazetteer (more recently Facebook’s¹²), but was even present in the ‘Foursquare’ tweets. This is quite surprising as Twitter bought the rights to use Foursquare’s database¹³, therefore one would think tweets from ‘Foursquare’ would be accurate. This has implications for maps generated without this knowledge, as mapping the unrefined coordinates produces looser clusters around POIs with some tweets being 10s or 100s of metres off their true location. Even when analysing tweets at the larger town or city geographic scale, if the user restricts the tweets to within a certain spatial extent they may be losing valuable information or unwittingly including tweets from neighbouring areas.

‘Instagram’ and ‘Foursquare’ were selected as the two sources as they comprised the majority of the third-party geolocated tweets. Other sources exist that produce geolocated tweets, such as ‘Untappd’ and ‘TweetMyJOBS’. These were not included in the experiment as including each geolocated source would have added significant time onto the processing stage and thus was out of scope for the thesis. These two sources in particular provide contrasting POI quality. ‘Untappd’ will list the desired metadata in a similar fashion to Foursquare, thus serves as a potential third source for future experiments. ‘TweetMyJOBS’ linked to the original job offer which included a location field for the place of work; however, the job offers are removed once the job is no longer available, thus depending on when the experiment is run these links may have expired, causing ‘TweetMyJOBS’ to be an unreliable source. Trying to ascertain the data quality of each third-party source that provides geolocated data is a valuable experiment but not one that is within the scope of this thesis.

A further discovery within the tweet metadata was the presence of native Twitter check-ins, such as those from ‘Twitter for iPhone’ or ‘Twitter for Android’. These do not

¹¹There is no official documentation about how Twitter resolves third-party POIs, therefore this is a conclusion based on various online discussions and largely deduced from conducting the experiment

¹²<https://developers.facebook.com/docs/instagram-api/changelog/> [Accessed 8th November 2018]

¹³<https://support.foursquare.com/hc/en-us/articles/115002413228-Foursquare-Places-on-Twitter> [Accessed 8th November 2018]

include a regular lat/long coordinate pair as with the other directly geolocated tweets, thus were not immediately evident. However, all tweets with location information attached will have a general bounding box attribute. This bounding box is comprised of four coordinate pairs that construct a square-based polygon over the respected area. Where a check-in has been created, these four pairs will be identical, thus constructing a singular point. The bounding boxes were not used in the analyses due to their lack of precision; however, future work could be done on modelling narratives at these lower granularities. Discovering these native check-ins added 13,271 geolocated tweets to the overall dataset.

3.4.3 Concluding Remarks

This experiment aimed to understand the behaviours of third-party tweets from ‘Instagram’ and ‘Foursquare’. In doing so, it highlighted the unknown relationship between the third-party sources and Twitter’s internal gazetteer, often leading to tweets from third-party sources appearing a great distance away from their intended POI.

While reverse-engineering Twitter’s gazetteer is not within the scope of this thesis, Experiment 3.3 did go some way to classifying clustered tweets as belonging to a point of interest, a neighbourhood or city-level centroid. Accessing tweets from ‘Foursquare’ showed that extra spatial information is available from third-party sources that allow its collection. Unfortunately, a lot of these sources including ‘Instagram’ deliberately block researchers from accessing this lucrative data. Therefore, obtaining information via the URLs in tweets would not be a suitable task to attempt with a larger tweet dataset. It was thus necessary to return to resolving coordinate pairs in order to obtain address-level information and classify tweets as relating to a POI, neighbourhood or city-level resolution. While the method used in Experiment 3.3 was successful in differentiating between POIs and neighbourhoods, there was no attempt to validate this method nor use several APIs to avoid selection bias. Therefore, the experiment outlined below addressed these issues.

3.5 Experiment 5: API Calling

From the previous experiments in sections 3.3 and 3.4, it was clear that not all geolocated tweets are created equal. Some are geolocated precisely, others to POIs, others still to neighbourhoods or entire cities. These results are further obfuscated by third-party applications that do not transfer granular location data across to Twitter. Therefore, there is a need to resolve the granularity of the coordinates and match them against any potential POI, neighbourhood or city-level centroids.

In order to achieve this, and to improve upon the previous approach, the tweet coordinates were compared against three open APIs that include building-level location information. There are many public APIs that can be accessed to achieve this, and for this experiment the Foursquare, Google and OpenStreetMap (OSM) APIs are used. These are chosen due to several factors; Foursquare is known to be used by Twitter for geolocation, Google is a well-known company with an accessible API and a long history of collecting geographic information, with OSM offering a competitive crowd-sourced alternative with quality that is comparable with established national mapping agencies such as Ordnance Survey (Mooney et al., 2016). At the time of conducting the experiment, Ordnance Survey did not have an equivalent free API¹⁴. This builds upon the previous experiment in section 3.3 by adding additional APIs to reduce the risk of bias introduced by only using one API, as well as validating the results by manually evaluating five random samples of 100 tweets. A bonus for interested parties is this experiment also returns street-level address data, and while resolving a specific address was out of scope of this thesis, the rich data could be used to append address information to tweet clusters.

3.5.1 Methodology

The same tweet dataset used in the previous experiment was applied to this one. The key difference for this experiment is the lack of further spatial restriction; the entire geolocated tweet dataset was used. This was to fully test the capabilities of the framework to handle tweets from both urban and rural areas. Therefore, for this experiment, after implementing the source cleaning portion of the framework 189,453 geolocated tweets formed the primary dataset covering the 1st November 2016 to 4th August 2018. These geolocated tweets were clustered, due to API limitations discussed below, identifying 7,039 clusters of at least two or more tweets.

3.5.2 Calling the APIs

To contact the Foursquare, Google and OSM APIs a unique identifier is often required. To obtain this, the user must register an account with Foursquare and Google (OSM does not require this, instead it limits the user to 1 request a second). The registration and use of all APIs is free; however, Foursquare requires a credit card authentication to allow the user to make more than 950 calls per day, a feature that is necessary with the volume of data collected during the thesis process. Google also requires a credit card and does charge for requests, but they offer \$200 a month of free usage¹⁵.

¹⁴<https://developer.ordnancesurvey.co.uk/os-places-api> [Accessed 10 September 2019]

¹⁵<https://cloud.google.com/maps-platform/pricing/sheet/> [Accessed 15 August 2019]

A script was written in Python to contact the APIs using the various credentials. The data were saved in JSON as well as a human-readable set saved to CSV. The data returned included the original tweet cluster coordinates and the response from the APIs such as closest venue, address and distance.

3.5.3 Formulating Accuracies

As previously investigated in Experiment 3.3, there are noticeable tweet clusters formed over city centroids. These clusters often contain the most tweets and users, and are considerably larger than the subsequent neighbourhood-level clusters. It can therefore be deduced that the largest cluster by tweet- and user-count in a dataset that covers a city will be the city centroid. Thus, if a city cluster coordinate pair is queried against the APIs, the resulting data must indicate that the cluster is at city-level granularity.

However, if a city-level or neighbourhood cluster forms over a known POI this can create false positives. Therefore, an argument should be set up between the APIs to return the most accurate result. There are metadata fields within the API responses that indicate classification accuracy. Foursquare has a distance metric and a venue ‘type’, Google has a descriptor such as “ROOFTOP” or “GEOMETRIC_CENTER” to differentiate between accuracies as well as a ‘type’ value such as “point_of_interest” or “premise”, while OSM also has a ‘type’ value of ‘node’ which denotes a venue and “place_rank” which denotes a certainty between 0 and 30, with 30 being certain.

In addition to these accuracy fields, a tolerance of 25m taken from Experiment 3.3 was added to the API queries, such that if a result were returned that was more than 25m away, it automatically was not considered a potential POI indicator. If all three APIs returned an address that was more than 25m away then the tweet cluster was classified as a neighbourhood; if at least one of the APIs returned a result within 25m it was classed as a POI unless one of the accuracy fields suggested otherwise.

To avoid creating unnecessary errors, no attempt was made to resolve an address for the POI cluster. Each API would have its own way of structuring an address, thus creating one was outside the scope of the thesis. For example, Google might return an address like “4 High Street, Southampton, Hampshire, SO14”, while OSM might describe the first part as “4 Hight St” and not return a postcode, while Foursquare may not include a street number at all.

In addition to the three APIs for reverse geolocation is an extra Foursquare API called “Checkin Resolve”, which takes a unique check-in ID from the tweet URL and returns specific user and venue information. This therefore meant that if each third-party source had such an API then accurate POI information could be obtained. This is not the case, and as ‘Instagram’ is the most popular source for geolocated information it necessitates an experiment like this. Nevertheless, the results from the “Checkin Resolve” API can

be compared to those from the other three APIs to see if the data can aid in improving classifications.

3.5.4 Results

	Google	OSM	Foursquare
Clusters	7,039	7,039	7,033
Errors	0	0	6
Within 25m	3,607	3,804	2,099
25m & POI	1,569	1,827	2,089
Above 25m (N)	3,432	3,235	4,946

Table 3.12: Table showing the results of matching the 7,039 tweet clusters to the three APIs, along with how many responses were within 25 metres of the original tweet and of those responses how many were POIs. Tweets above 25m are automatically classified as neighbourhoods (N).

Foursquare Checkins	Matching Tweet Coords	Matching Venue Coords
8,065	16	7

Table 3.13: Table showing the total number of Foursquare checkins within the dataset, along with how many non-Foursquare tweets shared the same tweet coordinates and how many Foursquare venue coordinates matched existing tweets.

Location	Checkins
Pub	465
Coffee Shop	418
Train Station	349
Office	239
Cafe	228

Table 3.14: Table showing the top 5 checkin types from Foursquare’s Checkin Resolve API.

As can be seen from table 3.12, all three APIs reliably return information from each request with Foursquare having a few errors where it did not return a valid venue. There appears to be a fairly even split between results from Google and OSM in terms of within 25m and above, whereas Foursquare returns over twice the number of venues above 25m than within. This is due to Foursquare returning at least one venue no matter how far away it is, while Google and OSM can return non-traditional venues such as blocks of flats or industrial complexes that may be within the 25m radius.

The POI distinction is made at the metadata level. Google will return results with a variety of POI-related attributes such as the obvious “point_of_interest” but also “premise” or “park”. “Premise” can refer to venues but also named flat complexes, and while the

latter are not specific POIs such as monuments or restaurants, they do produce hubs of activity and are therefore included in the POI classification. OSM has a more nuanced approach due to the crowd-sourced nature of the data creation. This creates matching metadata pairs such as “node” and “historic”, but there are also “relation” and “historic” pairs as well as “way” and “amenity”, all of which return valid POIs. However, there are also “way” and “highway” pairs as well as both “node” and “relation” and “building” pairs, which often return mundane roads or buildings that are not POIs. The OSM metadata therefore needed a stricter filtering system imposed to extract the relevant “nodes” and “ways” while omitting the irrelevant ones.

3.5.5 Foursquare Checkin Resolve

Within the tweet dataset are 8,065 Foursquare checkins, which were queried against the Foursquare Checkin Resolve API. In total, there were 16 tweet coordinates matching other non-Foursquare checkins and 7 Foursquare venue coordinates matching existing tweets, summarised in table 3.13. This is a very interesting discovery and shows how Foursquare’s internal gazetteer is largely unique to Foursquare. Despite Twitter purchasing the rights to use their data, only a tiny number of non-Foursquare-originated tweets exactly matched any coordinates relating to the platform. This further shows how the coordinates obtained by Twitter from third-party sources require intense scrutiny to extract accurate geospatial granularities, as Twitter’s internal gazetteer must differ from Foursquare’s despite having access to it. This also reinforces the notion to test distances from tweet coordinates to nearby locations as there may have been GPS interference or, more likely, the third-party gazetteer had different coordinates than Twitter for the same POI. Lastly, the lack of significant matches between the resolved Foursquare checkins and the rest of the dataset shows that there is potential for Foursquare tweets to be resolved separately and removed from the main dataset. Other third-party sources could also develop this feature, thus if this option becomes more widely available in the future then a Twitter dataset can be subdivided by source and each source independently analysed to obtain more specific results.

3.5.6 API Evaluation

A limitation with Experiment 3.3 was the lack of evaluation. In the experiment, the Foursquare API was queried and the tweet coordinates classified as POI or neighbourhood depending on what was returned. However, there was no evaluation carried out. Therefore, as this latest experiment builds upon the aforementioned one, an evaluation was conducted to verify the precision and recall abilities of the experiment.

To check the results of the API querying, 5 random samples of 100 tweets were extracted and manually inspected. The original coordinates were entered into Google Maps for

Sample	Correct	POI Pr.	POI R.	Adjusted Pr.	Adjusted R.
1	83.7%	85.7%	80.0%	100%	83.5%
2	84.5%	76.6%	78.3%	100%	87.8%
3	81.7%	85.4%	72.9%	97.9%	80.0%
4	83.2%	80.4%	86.0%	100%	87.5%
5	78.7%	67.4%	88.5%	100%	89.7%

Table 3.15: The results of the API classification evaluation, including adjustment.

a visual inspection. If either Google, OSM or Foursquare had classified the tweet as a POI and Google Maps showed a POI either at or within 25m of the original point, the classification was deemed correct. If none of the APIs returned a POI within 25m and Google Maps also did not show any nearby POIs, the classification was also deemed correct. If the classification was POI but the tweet text suggested a city-level checkin, such as ‘@ Southampton’, then this was tagged as a false positive. Similarly, if the tweet text mentioned a venue such as ‘@ Winchester Cathedral’ but the APIs did not return anything within 25m and classed the cluster as a neighbourhood, this was tagged as a false negative. However, the latter issue was more nuanced. For example, large venues such as Winchester Cathedral may have several centroids based on where and from which source the tweet originated; Twitter may place the centroid at one location, but Google, OSM and Foursquare may have it in another. In these cases, were the radius to expand beyond 25m then the proper venue centroid would be captured. However, expanding the radius risks increasing the number of false positives.

A potential solution to the false negatives is to compare the erroneous classification to the list of tweet clusters. This list contains all 7,039 clusters alongside the number of unique users within each cluster. Querying this list with the Winchester Cathedral tweet coordinates matched a cluster containing 217 unique users and 311 tweets, a ratio of 1.43. Similarly, querying the list with the ‘@ Southampton’ tweet coordinates matched a cluster containing 1,960 unique users and 7,080 tweets, a ratio of 1.93. Despite the differing ratios, classifying tweets as POIs or neighbourhoods purely on the ratio value would not solve the problem; many smaller POIs could create lots of tweets while less popular neighbourhoods could produce few tweets, thus creating misleading ratios. While creating an automated solution to this particular issue is outside the scope of the thesis, a manual inspection showed that removing clusters of 1 user and comparing tweets to clusters containing many unique users resolved some of the false positive and negative results, summarised in table 3.15 within the ‘Adjusted’ columns. These adjustments resulted in a near 100% precision score and a similarly strong recall value, validating the approach. The 97.9% value was due to a POI having mislabelled coordinates, resulting in a point that was more than 25m away from the building to which it belonged.

3.5.7 Discussion

Understanding location-based narratives relies on analyses of both text and location data. There is a risk when dealing with noisy location data to create misleading interpretations due to the lack of clarity. The purpose of this experiment was to generate more geographical data about the tweet coordinates and refine the noise. As previously discussed, third-party tweets are produced at different geographic resolutions ranging from precisely geocoded to city and country levels. The Twitter API does not differentiate between these levels, instead attributing most tweets as ‘city’ or ‘admin’ without any further precision such as ‘neighbourhood’ or ‘street’. Additionally, two tweets can have the same place type of ‘Southampton’ but each classified as ‘city’ and ‘admin’ respectively within the metadata, thus Twitter’s own classification method is unreliable. Furthermore, tweets from third-party sources have been shown to have differing coordinates in the tweet metadata to those in the original posts. Therefore, it is important to obtain as much extra information from the tweet coordinates as possible to build a complete geographic picture of the tweet origin.

To that end, this experiment provided significantly more geographic information to the tweet. All three APIs will return addresses for the tweet clusters, ranging from precise street-level to city or country resolution. While this experiment does not derive a precise address for each tweet or tweet cluster, the address information is saved for interested parties. Instead, the three APIs are queried and if they return a POI value then the tweet is classified as such, otherwise it is classified as a neighbourhood. This helps to separate all tweets, both native and third-party, into different geographic resolutions for further analysis. For instance, a potential application could be to identify a popular street within a city that requires attention, or to understand the activity of Twitter users within a political boundary.

A limitation of this approach is only clusters that included 2 or more tweets were analysed. This was because tweets sharing the exact same coordinate pairs are unlikely to have been generated by a mobile device; instead the cluster was formed either through a third-party, or via a stationary device such as a shop computer. This leaves the assumption that unique tweets represent mobile users, but these may also be check-ins to lesser-known POIs. However, querying each tweet coordinate pair is a time consuming process and would exhaust the rate limits and potentially being very expensive, thus may be appropriate for a future experiment.

The radius of 25m from the original tweet within which to discover a venue was derived from Experiment 3.3. While this value suited a built-up environment where there are many shops and monuments, it was not suitable for rural areas and resulted in a high proportion of clusters classified as ‘neighbourhoods’ due to the lack of venues within a 25m radius. Despite the initial motivation to classify tweets as belonging to POIs or neighbourhoods, perhaps it was wiser to classify the tweets as ‘POI’ and ‘non-POI’

due to the inability to easily distinguish between an area with a lack of POIs and a true neighbourhood cluster. Further work into comparing ‘non-POI’ clusters with the list of unique users and tweets per cluster could aid in identifying ‘real’ neighbourhood clusters, as well as creating a variable API tolerance that increased from 25m depending on if the cluster were within a rural area.

3.5.8 Concluding Remarks

It can therefore be concluded that while querying additional APIs does return more address information than Foursquare alone, the classification of a tweet cluster as a POI or neighbourhood is not as straightforward as previously thought. Where a formal API is available, such as the Foursquare Checkin Resolve API, this provided sufficient information to classify a tweet cluster as at POI, village, town, or city-level granularity. Were this ability available for other popular social media platforms such as Instagram then this would be a great advantage. Querying APIs is currently the best method to differentiate between POIs and neighbourhoods, though the results require further analyses in order to automate the entire process. However, as shown by the high precision and recall values in table 3.15, this experiment showed that a large portion of detailed locative thematic narrative analysis can be automated, answering RQ2.

3.6 The Proposed Framework

In conclusion, this chapter has outlined five experiments that approached location-based narratives from different angles. The first one emphasised the need for an ambient narrative approach rather than an ephemeral, event-centric one. The second explored the use of term expansion to create these ambient location-based narratives, but at the time was unaware of anomalies within the tweet metadata. The third explored these anomalies and created an experiment to systematically address them. The fourth explored third-party coordinate resolutions in detail and developed a method to improve the available information within a small dataset. The fifth and final experiment built upon previous location-based ones by further examining whether obtaining additional address information for each tweet cluster was beneficial, creating a method capable of classifying a tweet cluster as a POI or neighbourhood with an average of 79.1% precision and 81.1% recall. Drawing from the conclusions from each experiment, it is clear that creating ambient location-based narratives is a complex process, especially regarding spatial resolutions, but one that is achievable if the lessons from each experiment are followed. To summarise, figure 3.14 shows the refined framework and table 3.16 explains each stage.

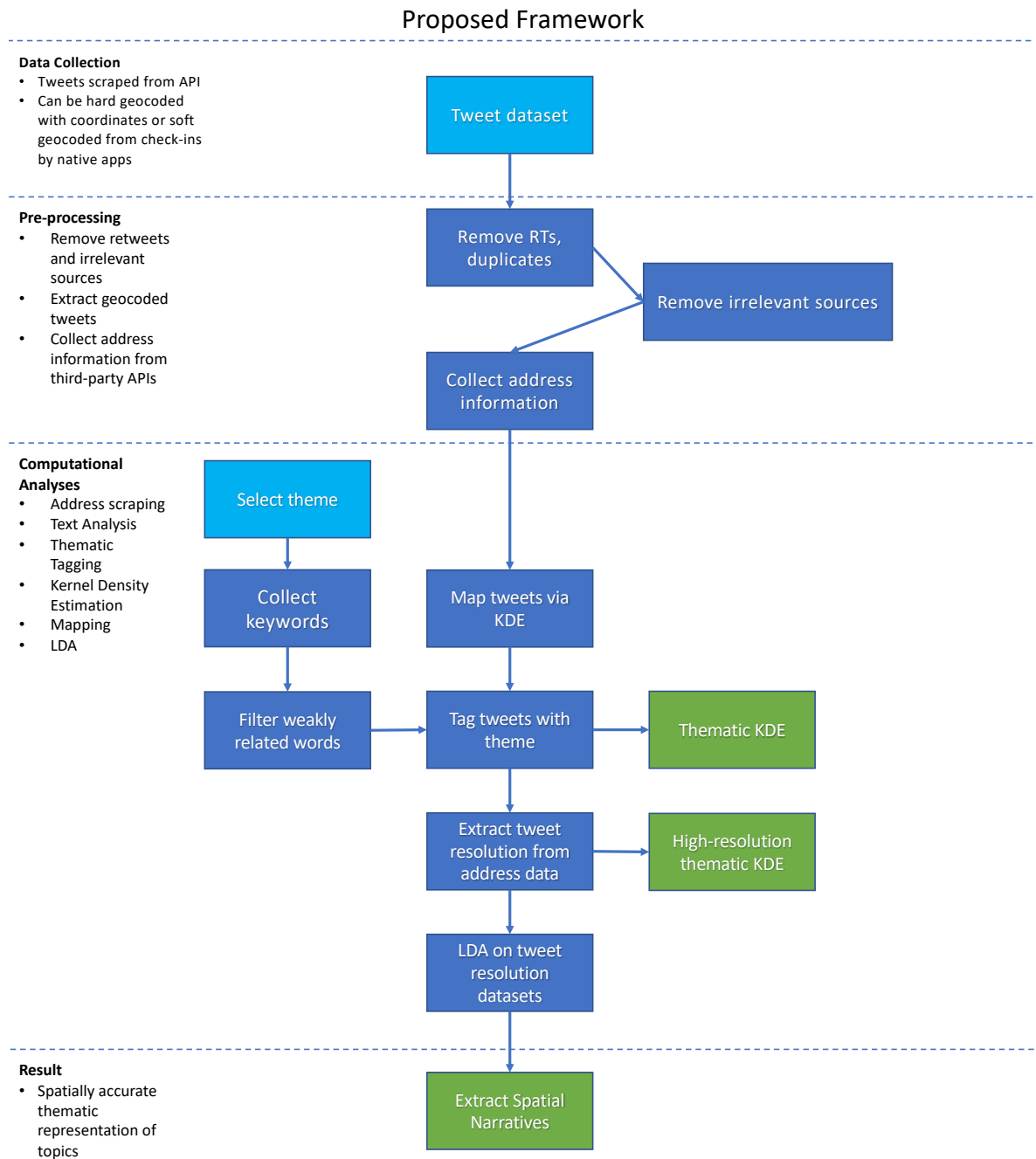


Figure 3.14: Visual workflow of methodology chapter.

Step	Process	Reason
Tweet dataset	Obtain the tweet dataset via the Twitter API.	Obtaining a rich dataset of location-based narrative information for analysis.
Remove RTs, duplicates	Identify retweets and duplicated tweets, remove them.	Retweets offer no new narrative information, and duplicated tweets will bias topic modelling and spatial analyses.
Remove irrelevant sources	Identify third-party sources that create undesirable tweets and remove them.	Noisy or automated third-party sources can produce a great deal of spam, making textual and spatial analyses challenging unless removed.
Collect address information	Contact geo-APIs like OSM to match geolocated tweets to addresses.	This step identifies whether a tweet is over a POI or not, refining the tweet granularity that is often missing from the metadata. It also creates a rich address dataset of interest to companies. Calls to the APIs can be greatly reduced by following the previous two steps.
Map tweets via KDE	Apply kernel density estimation to the tweets.	KDE will visually analyse the tweets and show hotspots of activity. These can prompt further investigation or support existing theories of activity.
Select theme	Identify a thematic approach to analysing the tweets.	General analysis of tweets can produce amorphous results. Targeted analysis will produce more relevant results.
Collect key-words	Once identified, obtain keywords from experts and/or an online thesaurus.	These keywords are important for identifying specific tweets of interest and refining a noisy dataset. Comparisons between keyword sets are useful for robustness checks and thematic nuances.
Filter weakly related words	Remove these words to refine the keyword datasets.	A refine keyword set will identify the tweets of interest and minimise false positives.
Tag tweets with theme	Apply keyword sets to the tweets.	Refines the tweet dataset to one specific to the research question.
Thematic KDE	Apply KDE to the newly created tweet dataset.	Produces a comparison map showing the effect of keyword tagging. Also highlights areas of activity specifically related to the research question.
Extract tweet resolution from address data	Analyse the API results to identify tweet granularity.	This step is key in stratifying the resolutions as third-party tweets will appear as precise when they may be aligned to POIs. This step increases the accuracy of the dataset and improves utility for interested parties.
High-resolution thematic KDE	Apply KDE to the stratified tweets.	This step further refines the spatial accuracy. KDE can be produced at precise, POI, neighbourhood or city levels from tweets previously assumed to be precise only.
LDA on tweet resolution datasets	Apply LDA to the different tweet resolutions.	With the refined datasets, LDA will be much more concise. The different granularities can be analysed to see if different topics are discussed.
Extract Spatial Narratives	Combine LDA and KDE to produce a holistic result.	With clearly refined and defined topic and spatial results, these can be combined to produce accurate models of activity.

Table 3.16: A detailed explanation of the framework.

Chapter 4

Case Study: HM Treasury

4.1 Introduction

In the previous chapter, the methodology for the thesis was outlined. This involved several different novel cleaning methods that were not previously explored within the literature. Through five experiments the behaviour of Twitter data was explored and a framework for extracting meaningful content was created.

An application of the framework to a real-world research problem was an appropriate test of its functionality, as well as a useful medium through which to focus and refine the novel cleaning methods. This chapter presents a case study application undertaken in collaboration with HM Treasury to discover economic indicators within Twitter data.

HM Treasury, the government’s economic department responsible for controlling public spending and economic growth, uses conventional methods like Gross Domestic Product (GDP) and Gross Value Added (GVA) to estimate the strength of the economy (discussed further in Section 4.1.1). These methods, however, are often substantially slower than real-time, delaying reports by at least a quarter due to the release schedule of GDP and GVA data. Furthermore, as processing tax returns is required to obtain accurate GDP readings and as a financial quarter’s GDP readings are only finalised when this value is obtained, the GDP publications for the same tax period are out of sync by a quarter and only accurate after a review following the next quarter. The experiment presented in this chapter could offer a substantial advantage over these conventional methods as location-based thematic tweets can be obtained significantly faster than quarterly. In applying the framework, well-defined areas in which economic concerns are discussed were highlighted, contributing both spatial and temporal signatures to a subset of the population who are engaging in economic activities, a source of data that the Treasury is not currently collecting.

To evaluate the thematic tweets, kernel density estimation is used to identify from where the tweets are originating and how closely they align to economic centres such as the High Street. Further, LDA is implemented to analyse the themes and topics within the tweets to confirm that only economic-related conversations are extracted.

4.1.1 Existing Methods

It is important to understand existing methods and datasets to fully appreciate the improvement that Twitter data can bring to economic indicator analyses. Once the strengths and weaknesses of these methods have been discussed, the need for hyper-local real-time economic indicators will become clear.

4.1.1.1 Gross Domestic Product

Gross Domestic Product (GDP) is the measure of the size and strength of the economy (Callen, 2008) and is usually published every financial quarter (three months) or annually¹. The Office for National Statistics (ONS) is responsible for calculating this value in the UK and takes into account a variety of data sources, including the money spent on goods and services produced in the UK, minus the cost of imports and plus the profit of exports; wages earned and profit made by the companies selling them; the value of the goods and services produced; i.e. the expenditure, income and output of the economy. The ONS often revises this value as new data arrives or new methods of calculating it are employed.

These values are produced every quarter in the UK; however, as VAT returns are only completed quarterly, the most recent GDP values published are estimates based on growth from previous data and are only accurate after the ONS has obtained and analysed all available data; in other words, the GDP and VAT returns are out of sync and thus the most recent GDP values are published without having analysed the VAT returns for the same time period. Therefore, while GDP is a useful value for indicating the strength of the economy, the most recent values cannot be used confidently by anyone wishing to compare against other forms of data from the same time period unless they also employ the same revisionary technique.

While GDP shows the overall performance of the economy in terms of the buying and selling of goods and services, it aggregates the services into broad categories like manufacturing, transport or education² and does not make specific details public, such as the type of manufacturing. So while it is known which categories makes up the GDP

¹<https://www.ons.gov.uk/economy/grossdomesticproductgdp/articles/whatisgdp/2016-11-21> [Accessed 25 September 2018].

²<https://www.ons.gov.uk/economy/grossdomesticproductgdp/datasets/quarterlycountryandregionalgdp> [Accessed 25 September 2018]

value, knowing precisely which goods and services are prominent in the public eye is not possible to extract from public information. Fortunately, text analysis on tweets can highlight which goods and services feature within a society and, if the tweets have location information, where these topics are discussed.

4.1.1.2 Gross Value Added

Gross Value Added (GVA) is the subset of GDP concerning profits³. It is the total profits produced by an industry, minus its expenses, for all industries across the UK and is often used to measure the regional output of UK counties. However, the datasets are either only at national level, or if the dataset is at city-level granularity it only publishes annual results or has ceased to produce them at all⁴.

GVA is therefore more useful than GDP in terms of understanding the industries and sub-divisions that contribute to economic performance. The ability to break down each large industry, such as manufacturing, into its component parts including the manufacturing of wood, coke, furniture or cars⁵, creates a more specific description of economic activity and produces a long list of potential economic indicators that could be reflected in tweets by users discussing the various categories that make up GVA and thus GDP.

4.1.1.3 Other Datasets

While GDP and GVA are the main two related methods for analysing economic activity, there are other sources of data that are pertinent. The Consumer Data Research Centre⁶ (CDRC) publishes market research surveys, but as these contain potentially identifiable data they are kept behind several security protocols and require user accreditation. Similarly, VAT returns housed by ONS provide highly granular economic performance statistics for companies but are also strongly protected and require accreditation, the obtaining of which would have taken longer than the time available during the thesis though is an interesting avenue for future research. There are a plethora of other datasets like net savings⁷, interest rates and government deficit publications that form proxies of economic activity and performance. While these latter datasets provide either monthly or annual summaries, they are not at the same regional granularity as GDP or GVA as they do not include any geographic indicators. The CDRC contains a large number of economic-related datasets such as the housing market performance, transport spending

³<https://www.ons.gov.uk/economy/grossvalueaddedgva/bulletins/regionalgrossvalueaddedbalanceduk/1998to2017> [Accessed 25 September 2018]

⁴<https://www.ons.gov.uk/economy/grossvalueaddedgva/timeseries/c4i7> [Accessed 26 September 2018]

⁵<https://www.ons.gov.uk/economy/economicoutputandproductivity/output/bulletins/indexofproduction/june2018> [Accessed 28 September 2018]

⁶<https://www.cdrc.ac.uk/> [Accessed 20 September 2018]

⁷<https://data.worldbank.org/topic/economy-and-growth> [Accessed 25th July 2018]

and energy consumption, but they either only release data after several years' lag or said data are unofficial statistics.

4.1.2 Real-Time Economic Indicators

Real-time economic indicators are aspects of the economy that indicate economic performance and can be discovered and analysed in real-time or close to real-time. Existing sources include public transport data, shipping routes and purchases⁸.

The key advantage that tweet analysis has over existing methods is it generates an understanding of human activity that the Treasury does not or cannot currently measure. For example, addresses of businesses are fixed entities thus when VAT returns are submitted the economic performance is restricted to their buildings of operation. However, for companies that produce take-away coffees or other removable goods that can be consumed elsewhere, they and the Treasury currently cannot measure the displaced impact. For instance, if a group of people purchase coffees then sit in a nearby park, the displaced location of consumption is a value that's hard to transfer over to economic performance. With the rising popularity of social media, this VGI will reflect the consumption of coffees at the location.

As an example of direct impact, the tweets can directly reflect economic activity. A person will purchase their goods, tweet about it and move on. There are likely certain activities that will generate more social media output than others, such as eating a meal and taking a photo of it. Purchasing a new piece of technology may be reflected too, but buying something like a pair of socks is unlikely to be reported. It is therefore important to retain the notion that different economic activities, or even the same economic activity but with different items, may be reflected in a disproportionate way. To this extent, the experiment outlined in this chapter will discover and analyse economic activities present in the dataset.

4.1.2.1 Price Indices

Alongside GDP and GVA is the Retail Price Index (RPI) and the Consumer Price Index (CPI). These two indices measure a range of regularly consumed products and are used as a measure of inflation (Levell, 2015); however, the CPI does not take into account house prices and is therefore more stable. In 2013 the ONS switched to using CPI to measure inflation⁹ as this did not take into account house prices, which fluctuate at a greater frequency than regular goods. From 2017 owner occupier's housing costs (CPIH)

⁸<https://www.ons.gov.uk/economy/grossdomesticproductgdp> [Accessed 14 October 2018]

⁹<https://www.ons.gov.uk/economy/inflationandpriceindices/articles/cpihcompendium/2016-10-13> [Accessed 14 October 2018]

were included in the inflation equation to better model an individual's expenditure¹⁰. The impact of inflation raising costs of certain products, such as petrol and food, is likely to affect the day-to-day activities of the population and therefore feature in a number of tweets. So, while the effects of the CPIH will be felt by consumers and reflected on Twitter, related tweets do not directly measure the economy but do reflect consumer behaviour, thus are useful as economic indicators.

4.2 Previous Work

The literature on the use of social media to discover and extract economic indicators at a finer level is fairly sparse. Prior work on discovering and analysing economic data on social media covered topics such as public mood (Bollen et al., 2009, 2011) and job losses (Antenucci et al., 2014).

The focus of these experiments was to extract temporally synchronous social media and economic data to understand whether social media data reflected economic variations and if they had predictive capabilities. In their work, Bollen et al. (2011) compared the values of the Dow Jones Industrial Average¹¹ to the creation of tweets within the same period, then used a sentiment scoring methodology to match sentiment change to economic stability. Their results show that only 'Calm' and 'Happy' sentiment scores are statistically significant ($p = 0.05$) in matching fluctuations in the Dow Jones. While obtaining significant matches is promising, their Twitter data was from February to December 2008, long before Twitter became a mainstream service and is therefore unlikely to represent the sentiment of the population. A similar criticism can be directed at their other Bollen et al. (2009), which used a dataset from the same period.

4.2.1 Economic Representivity of Twitter

The argument about older dataset being poor representations of the population is shared by Antenucci et al. (2014), who studied a later and longer period of Twitter data from July 2011 to November 2013 and explicitly criticised work that used older datasets. Their significant study matched fluctuations in tweet topics about and official publications on job losses, arguing that tweets do have a strong match with official statistics and have certain predictive abilities, especially when considering the time lag between official statistics being released and their true values being refined (a nice parallel to GDP and GVA). They further claim that their results are supplementary rather than substitutive to official statistics, highlighting that social media analysis is not to be taken as fact, but can provide elements of economic activity that are not covered by official methods,

¹⁰<https://www.ons.gov.uk/economy/inflationandpriceindices/methodologies/consumerpriceinflation-includesall3indicescpihcpiandrpqmi> [Accessed 14 October 2018]

¹¹The Dow Jones is a summary of the top 30 companies' share value and trading behaviours.

such as public reactions to economic crashes which can be obtained faster than through conventional means. A key methodological approach in the work by [Antenucci et al. \(2014\)](#) is to match tweets to a curated list of economic keywords created with domain knowledge. This approach is also argued as preferential to machine learning by [Bollen et al. \(2009\)](#), as complex computational analyses on short texts has proven difficult ([Maynard and Hare, 2015](#)).

Both articles conclude that with Twitter data that has been properly cleaned, the patterns observed in tweets relating to economic issues do match trends in official statistics. However, neither source reveals any cleaning methods aside from stopword removal and stemming. [Antenucci et al. \(2014\)](#) admits to including retweets in their work, a likely explanation for some of their n-grams having a strong signal in the data - for instance, a popular user tweeting about a job loss which gets retweeted by thousands of other users. As retweets have been shown to not provide any significant narrative information ([McMinn et al., 2013](#)), these should have been removed before keyword matching. Furthermore, neither source discussed geolocated tweets nor any third-party source cleaning. This is likely due to this metadata not being available in the work by [Bollen et al. \(2011\)](#), as well as third-party source not being as popular during the study period analysed by [Antenucci et al. \(2014\)](#) as today. Additionally, [Bollen et al. \(2009\)](#) published their final results online, including some that were not shown in their article, but since the article's publication the site has been removed. These are key issues for older studies: popular services change over time and therefore require constant research, as well as secondary data published to a location either expiring or migrating and thus readers wishing to find this data will be greeted with an error. In summary, older work in this area has now become relatively obsolete, with their work relying on an outdated 140 character limit and incomplete metadata.

Therefore, though these papers do statistically show a strong link between tweets and economic indicators, the datasets used are now between 6-10 years out of date. The continuing strength of this relationship can be improved by implementing the framework outlined in this thesis, which takes advantage of an increased character limit and understands that metadata cleaning is a vital component in creating reliable and relevant results.

4.3 Methodology

This section outlines the methods used to assess the construction of economic indicators from Twitter data to inform HM Treasury. It takes the framework outlined in the previous chapter and applies it to a real-world case study, the instance of which is shown in figure [4.1](#).

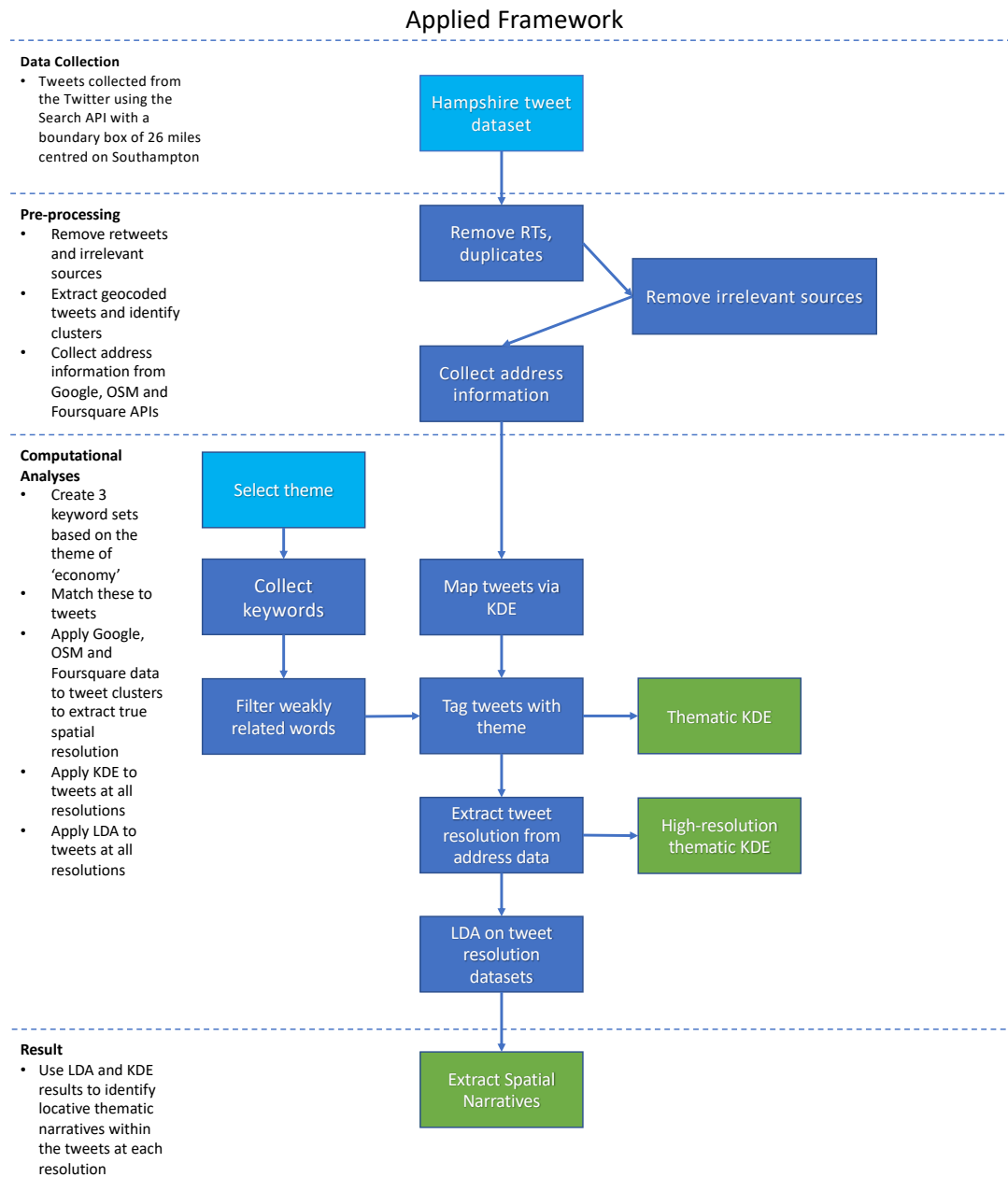


Figure 4.1: Flowchart depicting the application of the framework to this specific case study to extract locative economy-themed narratives.

4.3.1 Data Collection

To assess the ability of the framework to extract thematic narratives from a range of population densities, the 26-mile radius bounding box centred on Southampton, UK, was maintained from the previous experiment to cover several villages, towns and cities in the South of England, UK. The tweets were collected between 1st November 2016 and 4th August 2018, spanning 21 months. In total, the dataset consisted of 24,834,206 tweets by 2,465,515 users, of which 8,402,051 were retweets and 2,212,316 had some form of geolocation information, with 176,315 having precise coordinates. These results are

shown in greater detail in table 4.2. All data collection and cleaning was conducted in Python with subsequent visualisations produced in R.

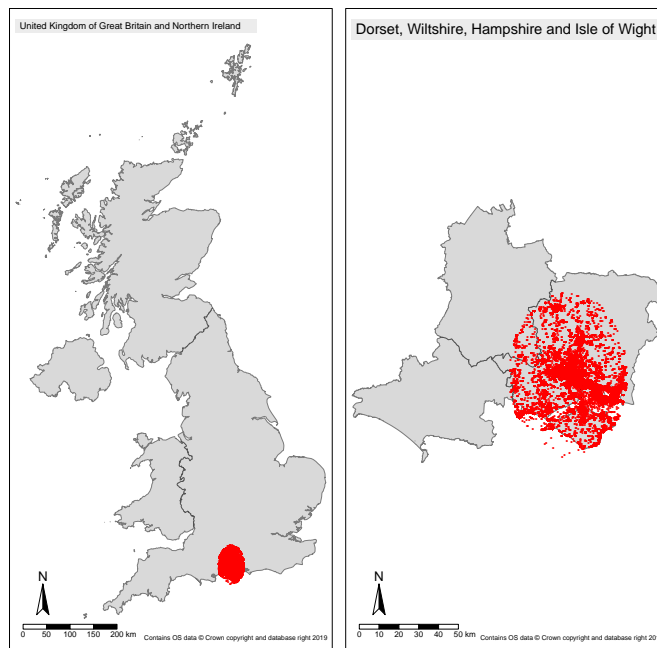


Figure 4.2: Map of the UK showing the study area in red. The counties in (b) from bottom-left clockwise: Dorset, Wiltshire, Hampshire, Isle of Wight.

Three sets of keywords with which to compare against tweets were obtained from several different sources. The first by searching an online thesaurus¹² for terms relating to ‘economy’ and using a similar relation-mapping exercise as outlined in Experiment 3.2 to create a semantically-linked term-expansion based keyword set; the second from the gold-standard Harvard University database used for machine learning content analysis¹³, from which a subset of economy-related terms denoted by ‘ECON’ were extracted; the third obtained in May 2018 from domain experts at HM Treasury. In total, the thesaurus-generated dataset had 756 terms, the Harvard set contained 743 terms, and the Treasury set had just 23 terms. These three sets were used to mitigate against subjective selection bias as well as providing three different keyword sources, affording an understanding of whether web-sourced, expert-sourced or gold-standard datasets construct the most useful and informative thematic narratives, an approach not seen in the analysed literature.

4.3.2 Applying the Framework

The thesis framework was outlined in Chapter 3. To summarise, the framework takes in raw Twitter data, removes noisy tweets and maps the outcome. As part of the framework retweets, irrelevant sources and users are removed and the geolocation resolution of

¹²<http://www.thesaurus.com> [Accessed 12 June 2018]

¹³http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm [Accessed 12 June 2018]

each tweet is scrutinised. This produces a clean dataset of relevant tweets from informative users that is better suited for policy makers, government organisations or industry companies to extract useful data from which to make decisions. Existing frameworks typically focus on removing retweets or spam users, but do not include either source or cluster analyses, thus this framework contributes these methods to the field. One such example of this is in aiding the Treasury in analysing real-time economic indicators. With the framework complete, it can be applied to extract economy-related tweets from the data.

4.3.2.1 Tweet Cleaning

The tweets were cleaned following the same processes as outlined in Chapter 3, namely removing retweets and irrelevant sources as well as expanding the geolocation precision by obtaining extra information from Foursquare, Google Maps and OSM.

4.3.2.2 Cluster Analysis

The rationale behind analysing tweet clusters is twofold: firstly, it identifies clusters created by just one user, therefore not truly representative of a larger geography; secondly, it identifies unique clusters that may arise due to events, new POIs that Twitter or others do not yet know about, or privacy settings that relocate a user's tweet based on their preferences (see Experiment 3.3 in the previous chapter or [Bennett et al. \(2018\)](#) for a practical application). Furthermore, identifying computationally-generated clusters (as opposed to natural points generated via a moving user) prevents them from artificially biasing any spatial analyses of otherwise presumed genuine users, such as seen in [Patel et al. \(2017\)](#) who, at the time, did not realise this difference. Analysing the entire tweet dataset with KDE without the knowledge that some clusters are not representative of activity at that location, for instance, would create a misleading impression for the policy maker. This logic is further reinforced by Twitter's plan to remove precise geolocation¹⁴, though as of January 2020 the company has neither announced a date nor confirmed the process is happening. Nevertheless, if this change were implemented, the ability for the framework to differentiate between neighbourhood and POI clusters will become even more valuable.

Thus, the tweets in the cleaned dataset were analysed for their coordinate pairs, and any tweets that shared coordinates were clustered. 7,039 clusters were generated that had at least two tweets with identical coordinate pairs and building on the classification method outlined in Experiment 3.5, these clusters were tagged as either precise, POI or non-POI. The splitting of the cluster types allowed for the precise dataset to represent mobile and ephemeral users, with the larger two datasets representing venues, POIs,

¹⁴<https://twitter.com/TwitterSupport/status/1141039841993355264> [Accessed 18 June 2019]

neighbourhoods or cities. A fourth tag was created called ‘ClusterOfOne’, indicating that while many tweets may have originated from within that cluster, they were all generated by a single user and would bias the dataset due to their uncertain spatial applicability, as well as their automated clustering behaviour skewing the KDE results. Therefore, any tweet tagged as ‘ClusterOfOne’ was removed. The impact of this can be seen in figure 4.3 with fewer red dots visible in the right-hand map.

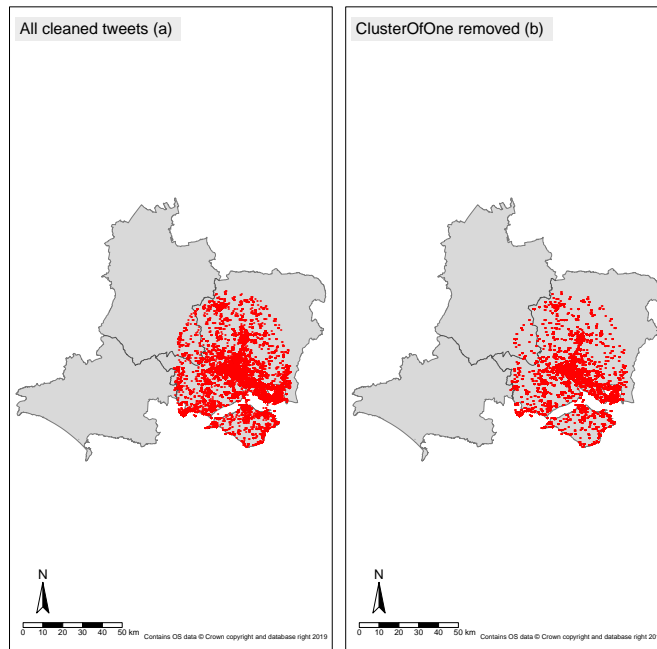


Figure 4.3: Comparison map of the cleaned Twitter dataset collected from the 1st November 2016 until 4th August 2018 (a) and the same dataset with ‘ClusterOfOne’ tweets removed (b).

Distinguishing between area resolutions is useful for policy makers when allocating resources, as some resources will be better suited to hubs rather than low-activity areas. Alternatively, low-activity areas may require additional resources to help increase engagement. The ability for the framework to output results in this manner is therefore appropriate and useful. Results from this part are shown in table 4.4.

4.3.2.3 Keyword Matching

With the cleaned and classified datasets created, the three keyword sets - Harvard, Thesaurus and Treasury - were imported and each tweet within the dataset compared to them, with any matches being logged and a tag applied to the tweet with the relevant keyword. To discover the matches, the keyword sets are converted into a list of unique terms and compared against each tweet, which was also reduced to its unique terms to avoid over-matching. If a term is present in both lists then it is tagged appropriately. The conversion and tagging is done in Python using set comprehension, the process of

converting two (or more) strings to lists of unique terms and comparing them together. It is to be expected that the Harvard and Thesaurus keyword sets will return the most matches as they are the largest, with the Treasury set being substantially smaller. However, a smaller dataset should create a more nuanced and specific result. The results from this tagging process are shown in table 4.1.

Keyword Set	Terms	Total Matches	Cleaned Matches	Cleaned Southampton
Harvard	675	41,923	31,133 (74.3%)	8,291
Thesaurus	428	6,622	4,953 (74.8%)	1,230
Treasury	23	716	469 (65.5%)	121

Table 4.1: List of keyword sets, their unique terms and subsequent matches in the raw, cleaned and Southampton area datasets.

4.4 Results

Tweet Type	Tweets	Users
Raw	24,834,206	2,465,515
RT	8,402,051	2,365,288
Geo	2,212,369	71,527
Boundary Box	2,022,765	44,776
Precise geo	176,333	19,903
Native check-ins	13,271	6,848

Table 4.2: Characteristics of the final dataset, each row a subset of the previous one. Here ‘Boundary Box’ refers to any user who has only allowed for their tweets to be at neighbourhood or city level, therefore do not produce any coordinate data aside from the overarching polygon. ‘Native check-ins’ refers to tweets from native Twitter apps that do not have precise coordinates but do have matching bounding box coordinates.

Table 4.2 shows the sample description of the final tweet dataset. With the addition of bounding box metadata first discovered in the previous chapter, the proportion of geolocated tweets greatly increases from the assumed 1-2% [Morstatter et al. \(2013, 2014\)](#) to around 10%. This 10% does include a wide variety of resolutions from the metre-precise to country-level, but is a promising indication of Twitter data being more useful for location analyses than previously thought. The further addition of matching bounding box coordinate pairs, indicating a native Twitter check-in to a POI, also increased the amount of precise data available for analysis, something that has not so far been mentioned in the literature. Therefore, this table exemplifies one of the contributions of this thesis: a greater understanding and appreciation for metadata analysis.

4.4.1 Source Analysis

As previously mentioned in Chapter 3, analysing sources provides useful insights into the behaviours of users. Tweets that are primarily used to advertise or are created by fully automated accounts will often use an identifiable source, such as ‘Blue Rhino Web Services’, rather than the native sources such as ‘Twitter for iPhone’ or ‘Twitter for Android’. These sources will have a high tweet to user ratio, i.e. thousands of tweets but only one or two users, so are easily identifiable. Once these sources are removed, the resulting dataset of geolocated tweets will be substantially less noisy. As shown in table 4.3, the sources were ranked by their tweet to user ratio and it can be clearly seen how many tweets were produced by these automated sources.

Source	Tweets	Users	Ratio
Blue Rhino Web Services	10,372	2	5186
dlvr.it	19,953	7	2850.43
Sandaysoft Cumulus	4,127	2	2063.5
auotmicv12demo	491	1	497
Out On A Shout Bot	421	1	421
chopperspot	358	1	358
Landmark Manager Web	351	1	351
Ratlake Transponder	252	1	252
local.angle	225	1	225
MOT Test Centre Finder	194	1	194

Table 4.3: Top 10 sources that contain geolocated tweets ranked by tweet to user ratio. After removing sources with fewer than 5 users and 10 tweets, 26 total sources remained.

The aim of removing these tweets is to reduce the false-positives when tagging or mapping the tweets. If a source is producing tweets that have ‘buying’ or ‘selling’ in the text but are merely advertising their website, the impression of economic activity would be biased by their tweets. Similarly with location data, sources can frequently tweet with geolocation information. The third-party source ‘Blue Rhino Web Services’ in particular tweets every time a ferry leaves or arrives at Southampton docks, creating a very detailed set of geolocated tweets that could be of interest to a study of travel or infrastructure patterns but does not contribute to a population-centric thematic narrative, as well as primarily creating a lot of location noise. Once these and others like it are removed, the behaviours and movements of ‘real’ people are clearer to see. The last stage in the source analysis section was to remove sources with fewer than 5 users or 10 tweets, which often contained tweets from websites.

Before cleaning there were 302 sources that contained at least one geolocated tweet. After 276 sources deemed irrelevant were removed, such as those with unfavourable tweet to user ratios highlighted in table 4.3, 26 remained that contained useful information.

The remaining tweet dataset comprised of 110,896 geolocated tweets by 19,492 users, including all precise tweets and native check-ins. This cleaning process is visualised in figure 4.4. While it might be difficult to see the differences, the density of tweets has been reduced as well as sporadic tweets made by shipping containers have been removed, shown by the lack of tweets off the coast of the Isle of Wight.

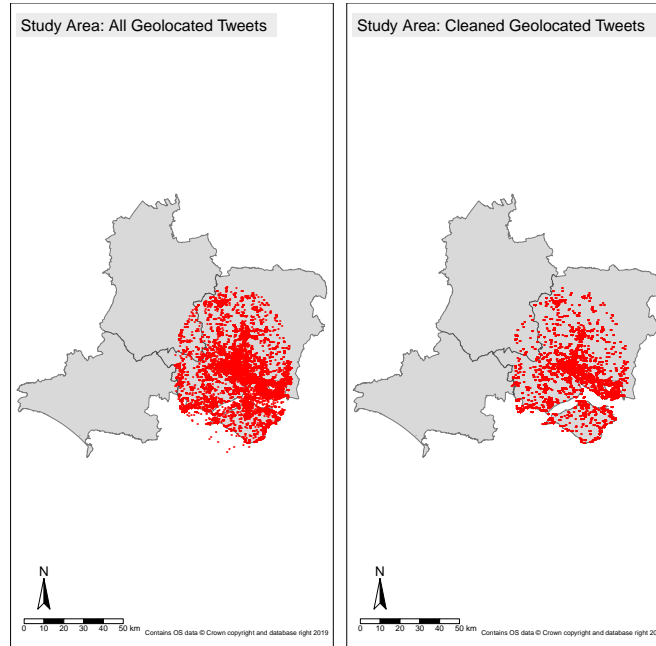


Figure 4.4: Map of all the raw tweets compared with the cleaned tweets within the southern UK counties related to the 26-mile radius Twitter search collected from the 1st November 2016 until the 4th August 2018.

From these remaining tweets, their geolocation resolution was tested to highlight that not all geolocated tweets share the same level of detail, ensuring that conclusions derived from the results are accurate to their respective spatial granularities.

4.4.2 Cluster Analysis

Cluster Rank	Tweets	Label
1	7,080	Southampton Centroid
2	3,349	Twyford Centroid
3	2,275	Portsmouth Centroid
4	1,080	New Forest Centroid
5	1,055	St Mary's Stadium

Table 4.4: Top five sources ranked by tweet count. Location labels obtained by manually searching Google Maps and OpenStreetMaps.

As shown table 4.4, the top five clusters contained both POI, city and area clusters, thus classifying clusters based on their tweet ranking is unwise. However, now that these results have been collected and manually labelled, future work can use the labelled set as training data during the classification process. In total there were 2,792 POI clusters and 1,875 non-POI clusters, comprised of 53,682 and 43,271 tweets respectively. There were also 2,372 ‘ClusterOfOne’ clusters comprised of 20,784 tweets, which were removed. While this is a substantial number of tweets to remove, the majority of the tweets are automated advertisements (verified by manual inspection) and therefore do not contribute genuine location-based narrative information. For easier visual comparison, the following maps will all use the Southampton area as clusters are more identifiable than within the counties version.

From the maps shows in figure 4.5 it is clear that the framework is impacting the KDE results. Map (a) shows the raw tweets, with noticeable clusters around the city centre in the south and University of Southampton to the north. However, these are fairly dominant clusters masking the signal generated by surrounding clusters. Map (b) highlights a greater activity in the city centre, mainly around the commercial districts comprising food and shopping. Map (c) removes a lot of the background signal to create much tighter clusters around the POIs identified by the framework. Similarly, map (d) also has tighter clusters but around the non-POIs, which is exemplified by the strong signal in the north-west over Southampton Common park previously not shown by the preceding two maps. With these extracted datasets of tweets at their respective geographic resolutions, further NLP approaches can be made with the confidence that topics are much less likely to get distorted by activities happening at different geographic scales.

4.4.3 Keyword Matches

With the fully cleaned and classified datasets obtained, the economy-related keywords were applied (shown previously in Table 4.1). It was not surprising that the Harvard dataset created the most matches, with the Thesaurus set second and the Treasury set third. What was surprising was the large difference between the Harvard and Thesaurus sets, despite them both having a large number of terms. The Harvard set contained a long list of economy-related terms that had been curated and published as a specific gold-standard. Though the online thesaurus created a similarly large dataset, the term expansion method applied by the framework created a more semantically similar set based on objective features such as semantic closeness. This therefore may have diluted the possible matches by creating a more abstract than literal keyword dataset. The Treasury set created the smallest number of matches due to its much smaller size. However, this set was created by HM Treasury and is therefore very specific, unlike the other two which are more generalised. The advantage with applying a smaller but bespoke

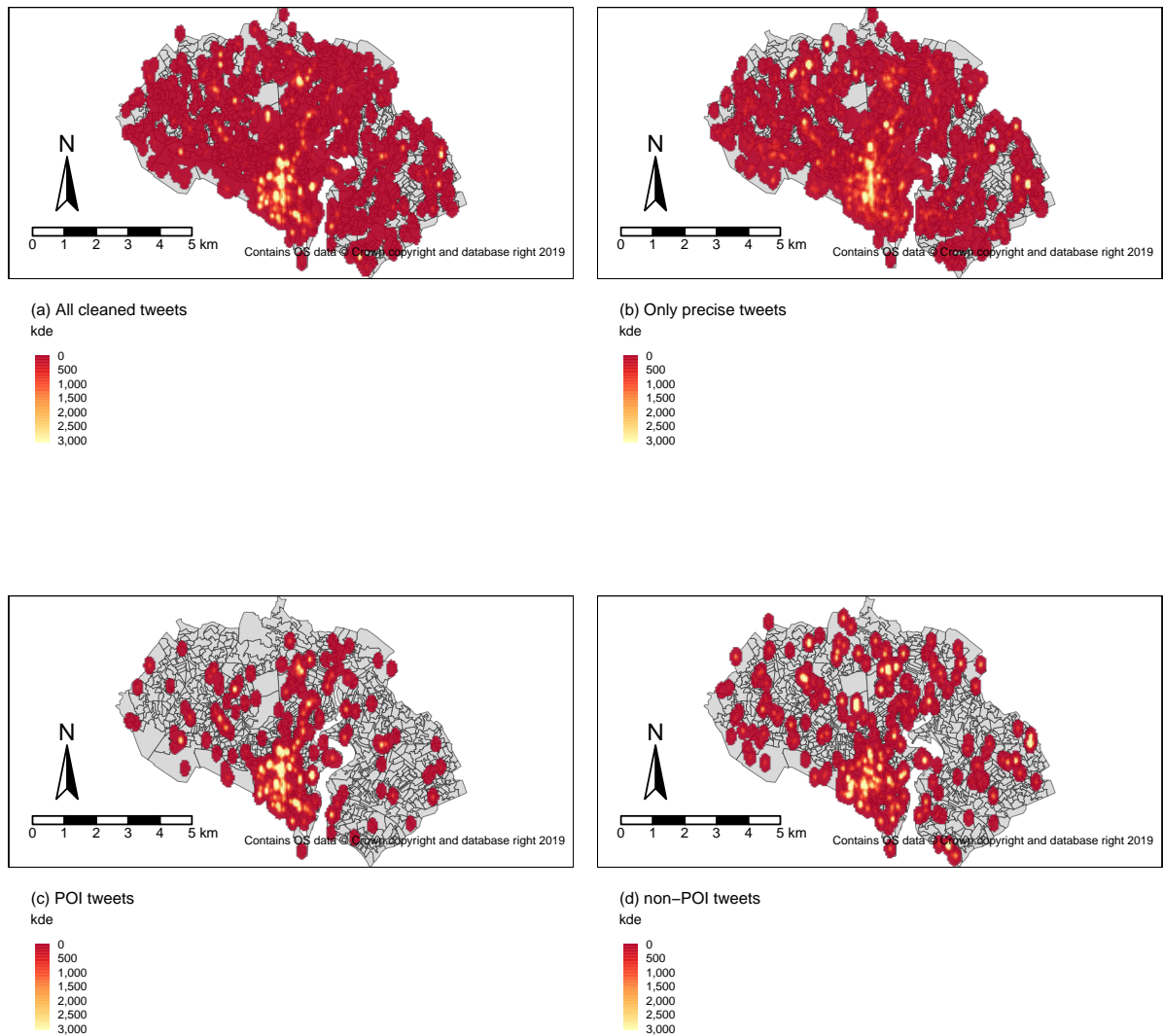


Figure 4.5: Map of tweets within Southampton collected from 1st November 2016 to 4th August 2018, with KDE applied. (a) shows all the tweets, (b) shows the tweets that do not belong to any cluster, (c) shows tweets at POI resolution, (d) shows tweets at non-POI resolution.

keyword set is the results it creates are likely to be more aligned with the investigator's research question. The disadvantage is it risks creating a dataset that is too small for any meaningful analysis; however, as shown in the next section, the Treasury keyword set produced insightful KDE maps. Nevertheless, the goal of the framework was to produce datasets and maps that are useful to policy makers, specifically HM Treasury for this case study. The ability for the Treasury tags to produce meaningful results is analysed in Chapter 5.

4.4.3.1 Keyword Set Maps

The maps in figure 4.6 show the tweets tagged with each keyword set with a kernel density estimation (KDE) overlay modelled using a 200m kernel, cropped to Southampton for easier comparison than at county-level.

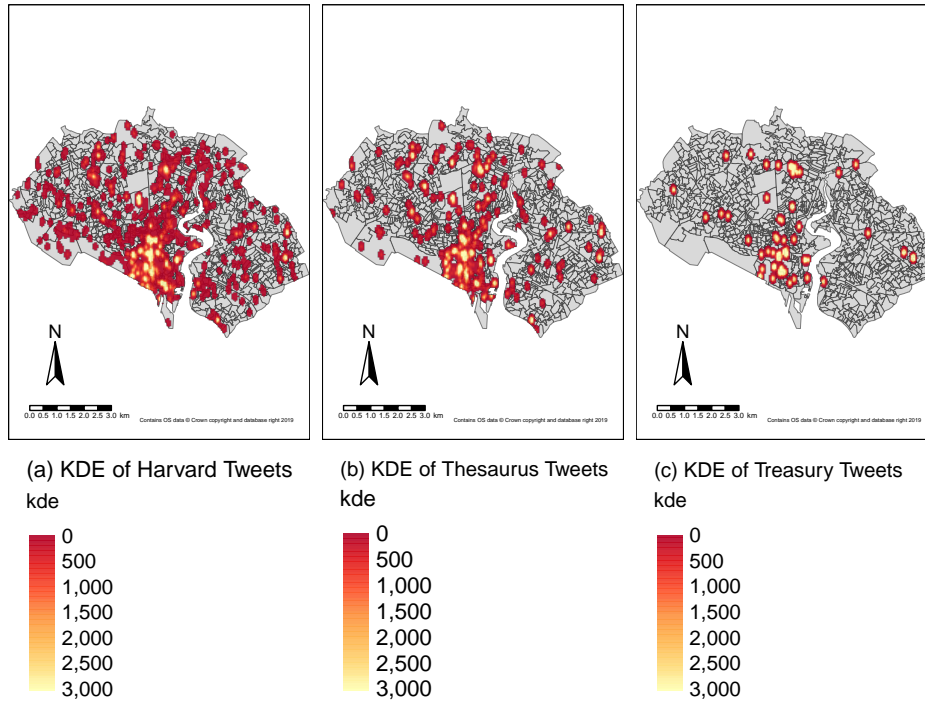


Figure 4.6: Map showing all the cleaned geolocated tweets within Southampton from 1st November 2016 to 4th August 2018 tagged with the Harvard (a), Thesaurus (b) and Treasury (c) keyword tags.

It is clear that the size of the keyword set influences the subsequent spread of the KDE overlay. Map 4.6a shows the cleaned geolocated tweets that were tagged with the Harvard keyword set. There is a larger KDE signature across Southampton than in the other two maps, which was to be expected due to the Harvard keyword set having the largest number of unique terms and therefore more tweet matches. All three maps show a strong signal over Southampton's High Street, which contains a plethora of shops and restaurants. A similar but weaker pattern can be seen in Shirley High Street, an area to the west of the maps. Map 4.6b shows a similar but smaller KDE activity, with fewer overall clusters but ones that are more spread out within the commercial areas, indicative of the more semantically-linked keyword set matching a broader spectrum of tweets. Map 4.6c shows few but very specific areas of economic activity due to the

significantly smaller keyword set. The cluster over the University of Southampton to the north-east is larger in this map compared with the others. This is likely due to a construction being undertaken during the study period, a topic that features prominently in the Treasury keyword set. This is also mirrored over the southern WestQuay area as this was a large construction site during the start of the study period¹⁵.

The KDE images highlight areas of interest and spatially match the keyword sets. To better understand the extent to which the keyword matches are relevant, LDA was applied to the results to extract salient topics. Theoretically, as the results have already been refined by keyword matching the LDA output should show clear topics, indicative of a successful cleaning framework.

4.4.4 LDA

In the previous section the keyword datasets and maps were produced. From looking at maps and numbers it is clear that the framework has refined the location data, but it may be challenging to understand the impact on the resulting topics. Therefore, this section outlines the application of LDA to understand the semantic similarities and differences present in the cleaned datasets. It is expected that, due to their sizes, the Harvard and Thesaurus sets will show the most variance, with the Treasury set having the most focused topics. It is also expected that the Thesaurus set will create the broadest results, with the Harvard keyword set refining this to more specific economy-related topics with the very specific Treasury set refining this still further.

As discussed in Chapter 2, the topic value k is often subjectively chosen, unless complex algorithms are added to derive a value. In this case, deriving a value is outside the scope of the thesis, therefore a best-fit value of 10 was chosen which seemed to produce the most distinct topics, with little overlap in the LDAVis output shown below. LDAVis is a Python library for interactive LDA topic modelling. It was originally created in R and ported over to Python. The library visualises each topic and its terms, allowing the user to navigate the topic-term relationships. Each of the LDA results show a tendency towards an economic topic, a promising result for the effectiveness of the keyword tagging process.

4.4.4.1 LDA Results

For the purposes of communication, LDA results are represented in LDAVis format. A table would not appropriately model the strength of association between terms in topics. Figure 4.7 shows the output of the LDA analysis for the tweets within Southampton collected from the 1st November 2016 until the 4th August 2018 tagged with the Thesaurus

¹⁵<https://www.bbc.co.uk/news/uk-england-hampshire-38155685> [Accessed 6th September 2018]

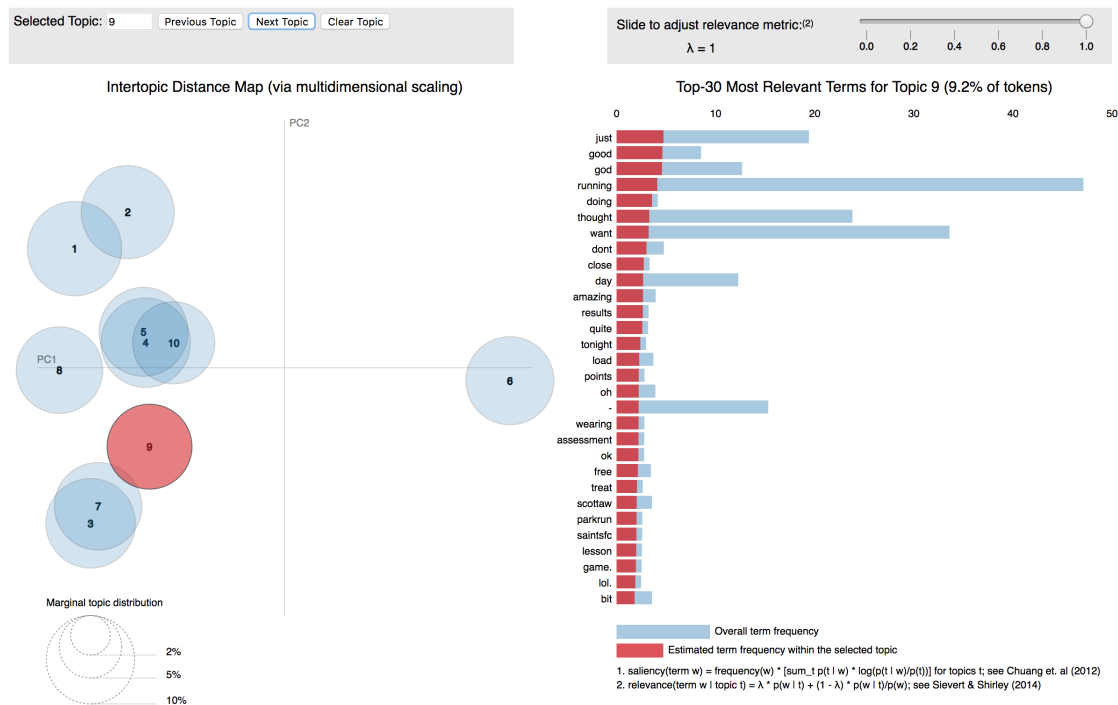


Figure 4.7: The output of the LDA analysis on the Thesaurus keyword set, visualised using the Python module LDAvis and with the 9th topic highlighted.

keyword set. The Python module LDAvis was used to transform the text-based LDA output into an interactive visualisation. The topics are transformed into the circles on the left, with their positions on the axes reflecting the extent to which the terms within each topic are related, and the size of the circles denoting the frequency of terms within the topic. The features that comprise each topic are also shown on the right-hand side, with the red bar indicating the frequency of features within that topic and the blue portion indicating the extent to which the term appears in other topics.

There is a clear topic about running and activity (topic 6), propagated by the third-party Twitter source Endomondo, a running app that posts run times and distances. While this might seem like an oversight in the framework, 4 other topics also discuss running including the term ‘parkrun’ (topic 9), a volunteer-based running group. As most of the topics are about running or physical activities, this indicates a strong narrative of fitness and outdoors events. This relates to the large cluster of activity shown in Figure 4.6b within the Southampton Common park area where parkrun is conducted. However, the topics are quite loosely defined, as shown by the lack of significant differentiation between the length of the red bars. This seems to indicate that term expansion via semantic linking has produced a synonym set for ‘economy’ to the extent that tweets containing topics about exercise (i.e. work) are returned. While this dataset was collected objectively and the approach backed by literature (Antenucci et al., 2014; Ghermandi and Sinclair, 2019), the resulting keyword set was diluted by too many weakly-related terms to be of equivalent use in this specific case study to the Harvard or Treasury keyword

sets. Future work could revisit the weights, as well as review each word within the set for relevance. This would remove the objectivity, but the resulting keyword set could be used as training data for future experiments.

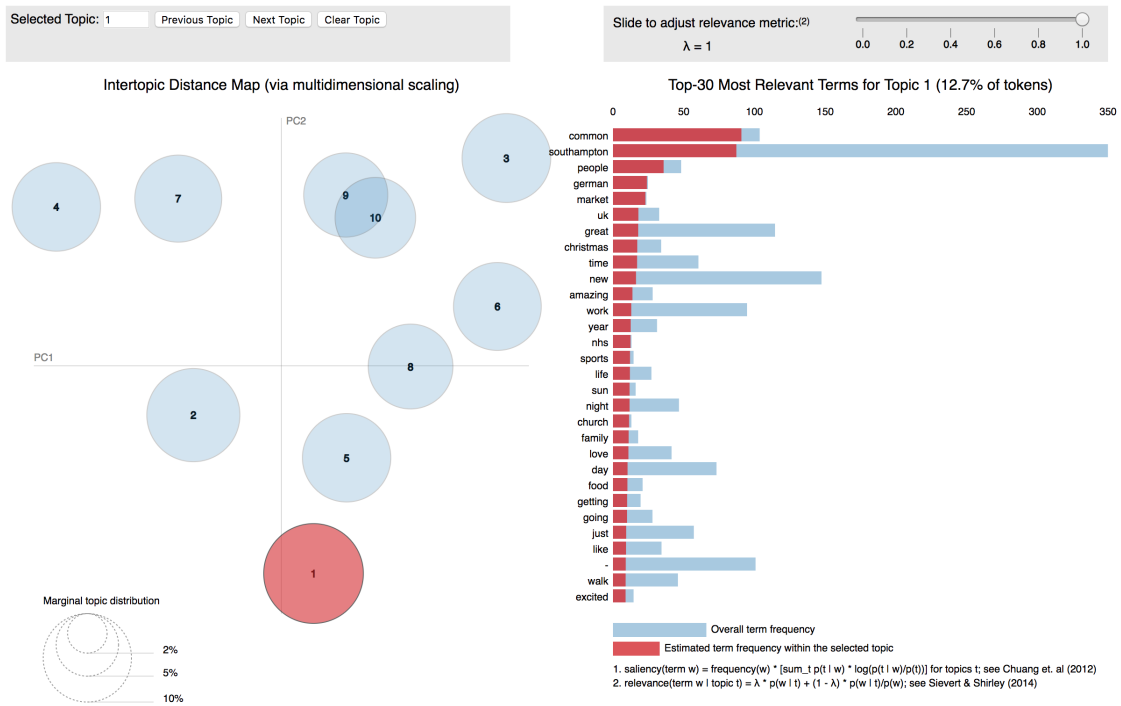


Figure 4.8: The output of the LDA analysis on the Harvard keyword set, visualised using the Python module LDAVis and with the first topic highlighted.

Figure 4.8 shows Harvard's 10 topics were fairly interrelated, with one clear topic about the Common People music festival hosted each year in Southampton Common park alongside a German market (topic 1). This also relates to the large cluster present in Southampton Common park. Southampton railway was also present in the topics (topic 3), including tweets aimed at National Rail due to delays and other issues. The other topics coalesced around venues, such as the University of Southampton (topics 5, 6) and events at these venues such as football matches at St Mary's Stadium (topic 7), with corresponding KDE clusters at these locations (see Figure 3.5 for these landmarks). Topic 10 also mentioned the new John Hansard Gallery which opened in the city centre on the 12th May 2018¹⁶. All of these topics are strongly related to the economy, mainly due to people attending the events to which the topics are related, or to the activity of commuters which gives insight into the transport network. Unsurprisingly, there is a strong emphasis on location, therefore this LDA analysis and visualisation also highlights specific areas of activity which match and validate the areas seen in the KDE analyses.

Finally, the Treasury topics visualised in figure 4.9 show a much more promising outcome. There is an independent topic about building and projects (topic 1), one encompassing

¹⁶<https://www.southampton.ac.uk/news/2017/12/hansard-gallery-opening.page> [Accessed 2 October 2018]

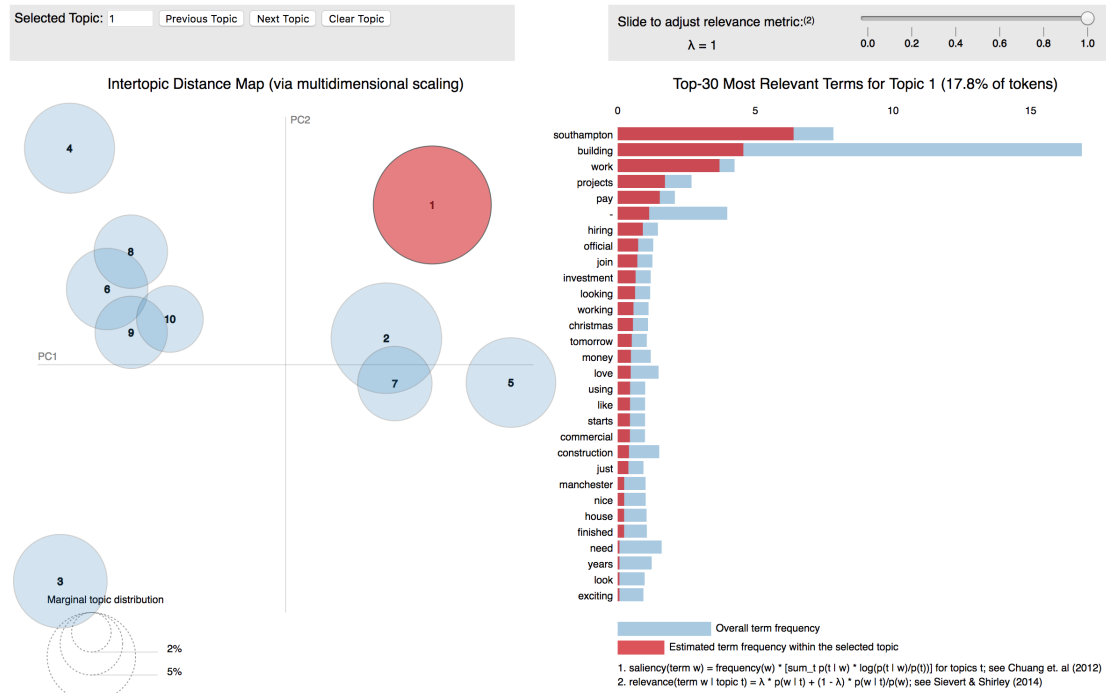


Figure 4.9: The output of the LDA analysis on the Treasury keyword set, visualised using the Python module LDAVis and with the first topic highlighted.

prices, income and trade (topic 2), another featuring contracts and employment (topic 3) along with a similar cluster of topics relating to jobs and construction (topics 6 and 9). The smaller keyword set has identified a much stronger narrative of economic activity, which is also reflected in the tighter clusters in the KDE maps. While the number of tweets tagged with the Treasury keywords is significantly smaller than the Harvard and Thesaurus sets, the topic relevance is much stronger. This indicates that when the framework is supplied with a specific keyword set, the resulting location-based thematic narratives are highly relevant.

4.4.4.2 LDA Resolutions

LDA has worked very well on the tweet datasets as a whole, extracting clear topics that require minimal amount of subjective analysis. However, when LDA was applied to the keyword sets at the different spatial resolutions produced by Experiment 3.5, the ability for LDA to extract meaningful topics varied based on the size of the underlying dataset.

Though the figures in the previous section were generated using a k value of 10, when the size of the datasets was reduced by the resolution classifications the 10 topics became indistinguishable due to LDA being unable to concisely identify 10 topics within the smaller dataset. When k was reduced to 5, the topics became much clearer, showing that LDA can successfully identify clear location-based topics within the dataset when the correct value for k is used. This further validated the spatial analysis conducted in

Experiment 3.5 as it isolated the correct tweets. The images below will represent tweets from each of the keyword sets visualised at one resolution each. The remaining images are in Appendix C.

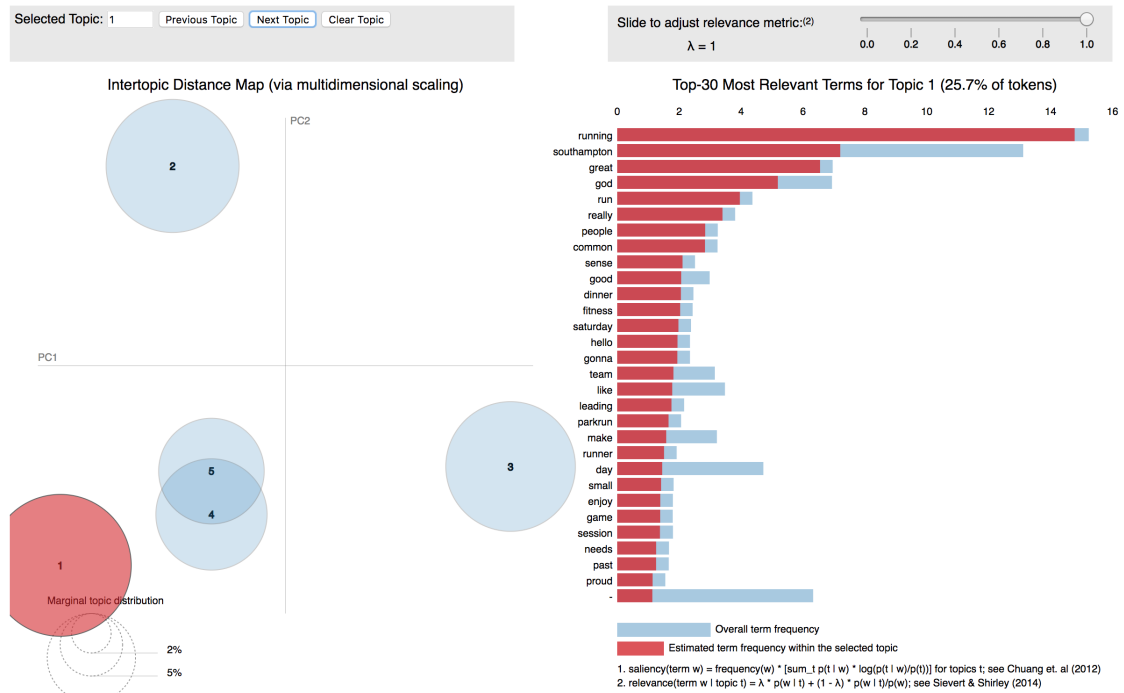


Figure 4.10: The output of the LDA analysis on the Thesaurus keyword set classified as originating from a non-POI, visualised using the Python module LDAvis and with the first topic highlighted. Image represents 294 tweets.

Similar to figure 4.7, topic 1 shown in figure 4.10 contains predominantly features representing the theme of running or exercise, shown by ‘running’, ‘run’ and ‘fitness’. Additionally, the ‘non-POI’ classification is validated by the presence of ‘parkrun’, an activity that takes place in the non-POI-classified Southampton Common park (shown in figure 4.5), which also appears in the topic.

The tweets that matched the Harvard keyword set and defined as relating to POIs, shown in figure 4.11, identify a strong relationship between tweets and known locations. Topic 5 includes features such as ‘railway’, ‘station’ and ‘nationalrailenq’, suggesting a theme of travel and highlighting potential issues, hence the users referencing the company’s official Twitter handle. The features also include ‘guildhall’, ‘mayflower’ (theatre) and ‘harbour’, three known locations within Southampton city centre, validating that the tweets were correctly classified as relating to POIs. The mixture of locations might imply that users were travelling to these locations, thus the result is useful for policy makers in understanding travel patterns.

Due to the smaller dataset from which to apply LDA, the topics identified from the Treasury keyword set were harder to define as there were fewer initial features. Some topics were prominent, as exemplified in figure 4.12, though the term frequency was

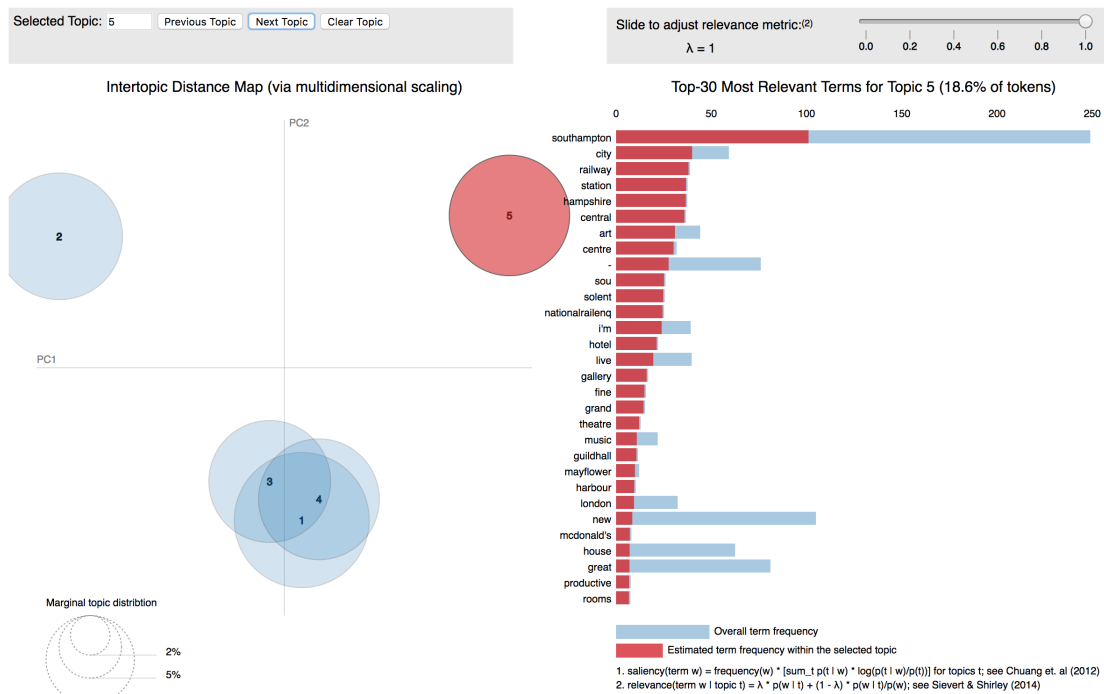


Figure 4.11: The output of the LDA analysis on the Harvard keyword set classified as originating from a POI, visualised using the Python module LDAVis and with the fifth topic highlighted. Image represents 5,618 tweets.

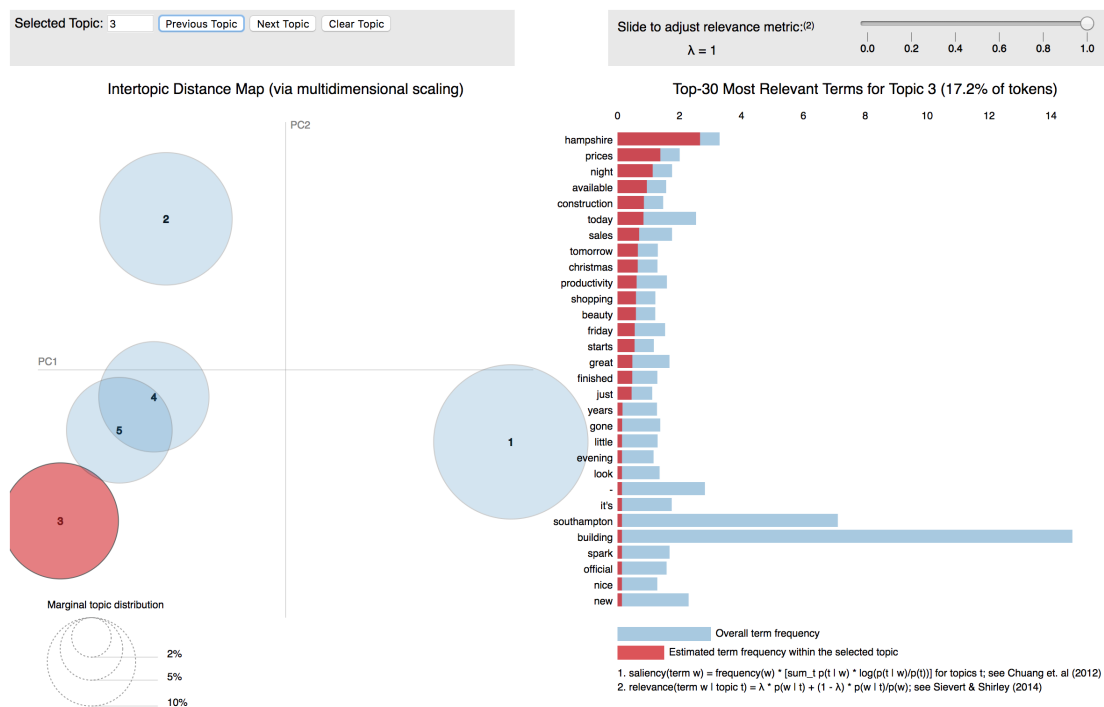


Figure 4.12: The output of the LDA analysis on the Treasury keyword set classified as originating from a POI, visualised using the Python module LDAVis and with the third topic highlighted. Image represents 77 tweets.

lower for all topics than in the previous LDA results, thus making topic definition more of a subjective endeavour. Topic 1 could be attributed to a narrative of the cost of construction within Southampton, but due to the very small number of associated terms, it becomes a challenging task to extract anything meaningful from the results.

It can therefore be concluded that extracting the tweets at their different resolutions is an appropriate exercise in understanding location-based narratives. The topics extracted from tweets at POI resolution show clear topics about transport and construction, two subjects that occur at identifiable locations. Topics from non-POIs show a similarly strong relationship with larger areas such as parks. This confirms that the method proposed in Experiment 3.5 to extract these spatial resolutions was a success.

When applying LDA to tweets classified as precise the ability to extract meaningful narratives differ. The larger datasets show clearer topics (shown in Appendix C), but when LDA is applied to a smaller dataset such as the Treasury, shown in figure C.6, it is challenging to extract clear topics as the term frequency and saliency is equally reduced. Further work is thus needed for understanding location-based narratives from smaller datasets. LDA on datasets with fewer than 100 tweets is therefore not currently recommended for policy makers without further investigation.

4.5 Summary

This chapter applied the framework outlined in Chapter 3 to an investigation into how economic indicators are represented on Twitter. This chapter presented KDE maps and LDA results that show the usefulness of three different keyword sets in discovering economy-related tweets, concluding that the set created by domain experts created the most specific results. However, the Thesaurus set highlighted different but relevant tweets about outdoor activities while the Harvard keyword set identified more specific thematic narratives about economy-related topics including festivals, transport and gallery openings. The latter two keyword sets are more indicative of consumer-level economic activity while the former is more aligned with traditional economic indicators such as construction and trade.

These results were very promising across all three keyword sets, highlighting the strong relationship between location and thematic narratives and thus validating the extraction of tweets at their proper resolutions. When applying LDA to smaller tweet datasets the results required more subjective analysis, thus these may not be as useful to policy makers as the larger LDA results.

4.5.1 Key Outcomes

This chapter has highlighted several key outcomes for the interest of policy makers. The three different keyword sets produced different LDA results, as was expected, but it emphasised the importance of keywords when discovering topics of interest. Careful consideration is required when selecting these keywords. If the keyword set is too small and specific then only a handful of tweets will be selected; these tweets will be highly relevant to the theme, but LDA will struggle to extract salient topics due to the small corpus. If the tweet set is sufficiently small enough it may be beneficial to analyse them manually. If the dataset is large, LDA will be able to extract several meaningful topics of interest, but domain expertise is required in both instances to extract the most benefit. The inclusion of $k = 5$ and $k = 10$ LDA results also show how the k value impacts on the output, something of which a policy maker may not be aware. Training the policy maker in how to interpret LDA results as well as change the k value would be required to gain the most benefit. Interpreting the raw LDA results is not advised as this is machine-readable only; however, the interactive LDAVis results should offer the policy maker sufficient interaction to understand the different topics and their related terms.

When considering KDE, the smaller the tweet set the more relevant the output. As the maps showed the exact locations of the users who tweeted about the topic, this allows policy makers a quick and easy overview of relevant areas within the city. This is particularly useful if the topic is more pressing, such as public opinion relating to development work, or concerns over job losses. With larger KDE results, the policy maker can still make generalisations about areas, such as the city centre being a commercial hub, or hotspots within the city centre relating to economic activity such as sales or new jobs. Even with the larger KDE results, isolated areas are easily identifiable and dense areas show intense areas of activity which could prompt further investigation by the policy maker. The framework does not produce interactive KDE results so the policy maker is unable to navigate around the map. In general, the KDE outputs help the policy maker as they visualise the Twitter data, something that is not natively possible using either the Twitter platform itself or institutionally used software such as Microsoft Excel. ArcGIS is an alternative approach to visualising these results, but their algorithms are not auditable or easily modifiable and thus the policy maker, providing they are able to manipulate R code, benefits from this framework's open approach to producing KDE maps.

The framework's key contribution to the policy maker is the combination of both LDA and KDE. The policy maker is able to see the LDA results and compare them to the KDE output, quickly and easily relating their theme of interest to locations within which it is being discussed. As a proof-of-concept model, the policy maker is equipped with two outputs that inform them of what is going on within a space. Future work would

enable the interactivity of both LDA and KDE, granting the policy maker a greater level of insight into their area of interest.

In the next chapter, the domain experts involved with the real-time economic indicators project are interviewed to evaluate the framework.

Chapter 5

Evaluative Interviews

In the previous chapter, the framework for discovering location-based narratives was tested with a real-world scenario. It successfully discovered, analysed and visualised real-time economic indicators through keyword matching and topic modelling. The three keyword sets produced different levels of detail; the Harvard and Thesaurus sets matched the most tweets due to their size, while the Treasury set matched the least but was more specific. These three outcomes are all relevant to the case study and are of use for different research questions. To evaluate these outcomes, and to answer RQ3, relevant staff at HM Treasury and ONS were interviewed and their feedback contributed to an understanding of the strengths and weaknesses of the framework constructed within the thesis.

This chapter outlines the interview design, the questions asked and the analytical codes applied. Key employees of the Open Innovation team within HM Treasury as well as leading statisticians at ONS are interviewed in a semi-structured manner, with questions emphasising the results of the framework and answers evaluating its impact.

5.1 Interview Design

There are many theoretical perspectives on qualitative interviews, including prescriptive or semi-structured techniques ([Campbell et al., 2013](#); [Garrison et al., 2006](#); [Kallio et al., 2016](#); [Smith and McGannon, 2018](#); [Tracy, 2010](#)). As the goal is to evaluate a multifaceted framework, it would be prudent to allow for flexibility during the interview to develop questions that naturally emerge from the participants' answers, as argued by [Kallio et al. \(2016\)](#). Therefore, a semi-structured approach was taken.

5.1.1 Coding Theory

The term ‘coding’ refers to classifying parts of speech within interviews that relate to certain themes, such as positive or negative opinion, general themes that run through each interview, or topics about future work. These can either be carried out inductively or deductively, i.e. using codes that emerge naturally from the interviewee’s answers versus establishing codes from the philosophical approach before the interview ([Fereday and Muir-Cochrane, 2006](#)).

[Campbell et al. \(2013\)](#) argue for an inter-rator approach, meaning two participants code a partial, whole or series of documents and compare their codes. If their codes match then they are applied to the rest of the documents; if they differ then they should adopt the approach proposed by [Garrison et al. \(2006\)](#) whereby they attempt to negotiate a settlement for these disagreements. They do not need to agree completely, but achieving a proportion of over 70% of agreed codes is deemed a success ([Campbell et al., 2013](#)).

A distinction [Campbell et al. \(2013\)](#) makes about the coding process is the involvement of the Principal Investigator (PI) or equivalent expert in the field, as this way the codes are more relevant and reliable. Having more than one source of codes is beneficial, but if an expert is present and the budget does not allow for a team of coders, then intra-rating (the coding of interviews by one person) is an acceptable approach. This opinion is shared with [Smith and McGannon \(2018\)](#) who argues for a “Critical Friend” who is knowledgeable in the area to review the codes. This theory differs from the argument proposed by [Tracy \(2010\)](#) who argue a universal system of coding is more efficient. However, as semi-structured interviews are partly spontaneous and one cannot predict all possible answers by the interviewee, using a rigid coding structure would be inappropriate. This view is reinforced by [Burke \(2016\)](#) who argues against the use of an existing coding structure as this risks creating stagnant research.

Therefore, for this case study, eight codes were initially created from the interview questions, discussed later in subsection 5.2.1. Extra codes emerged from the interview process, including several from each participant due to the semi-structured nature. Once the interviews had concluded, all the codes were used to analyse each interview. The codes are shown in table 5.1. Regarding the coding method, I was the sole person responsible for analysing the interviews. To conform as much as possible to accepted practise, each interview was coded then a week later was re-coded. This mimics the inter-rator approach by [Campbell et al. \(2013\)](#) by affording each interview a second perspective. In doing so, there were three conflicts out of 82 total coding sections where I had first coded three sections as Improvements but upon later viewing I decided they better related to Future Work. These disputes were resolved by accepting the second code, the argument for this being the second read-through was more critical due to the retention of the knowledge from the first coding process while still applying a fresh perspective. This represented a 96% agreement between the two coding processes.

While using one coder was necessary due to time and budget constraints it introduced subjectivity and researcher bias ([Chenail, 2011](#)), thus if the interviews were seen by a second party they might be coded differently. If this process were repeated then resources would be set aside for a second coder.

5.2 Interview Method

Ethical approval was obtained to contact the interviewees directly. This initially involved the HM Treasury staff, but was later amended to include the ONS. The initial application took just one day to process, with the amendment taking an additional two (ID 45418.A1, see [Appendix B](#)). The interviewees were selected by their involvement in the real-time economic indicators project. The initial phase involved directly contacting the staff at HM Treasury who created and worked on the project, as they were the domain experts. The group included a current employee, a former employee and the head of the team. The second target area was the Office for National Statistics who were in partnership with HM Treasury on the project. An interested party were three employees who worked on the GDELT project, a proof-of-concept investigation into extracting location-based information from global news sources ([GDELT, 2018](#)). In total, 3 employees from HM Treasury and 3 from ONS responded positively to the call for interviews. As these interviews were for evaluation purposes, this number was deemed acceptable. The interviewees were all told that the location-based framework aimed to clean noisy tweets by focusing on location and source analyses, and were shown before and after maps to emphasise the impact this had on the results.

The interviews were organised over e-mail and conducted via Google Hangouts with the HM Treasury participants and over the phone with the ONS. This method was preferable due to the busy nature of the interviewees. Once the interview time was arranged, a chatroom link was created by HM Treasury and a secure line was established by ONS. Open Broadcasting Software (OBS) was used to record the audio, with the interviewee's permission - see [Appendix B](#) for the ethics application. Video was used during the HM Treasury interviews but not during the ONS phone interview. At no point was video recorded. The interviews lasted around 30-60 minutes then were transcribed and the audio files deleted as per the ethics application.

As argued above, a semi-structured interview technique was the most appropriate method to extract evaluative information from the interviewees. There were questions developed beforehand, listed below, while organic questions arose during the interview process. This allowed for the different participants to develop their answers, promoting lengthy conversations about particular aspects that were of interest to them.

5.2.1 Questions and Codes

The questions were:

1. This real-time economic indicator project has been ongoing since January 2018 - I'd like to ask what is your plan for the project? What do you hope to get out of it?
2. Is location precision a key factor in it?
3. Is timeliness also a key factor?
4. This mapping process is part of a wider project that also covers topic modelling. What are your thoughts on this method allowing for scalable passive tracking of users including their location and topics of conversation?
5. The framework takes in noisy Twitter data and outputs clean datasets, tables and maps of activity. Is this something you would potentially include in future projects? Do you see this method as an improvement over existing methods?

Though these questions were prepared beforehand, due to the semi-structured nature of the interviews some questions were reworded, reordered or answered by the interviewee in relation to other questions.

The first five codes were generated from the pre-prepared interview questions: project aims, location, timeliness, scalable tracking and usefulness of framework. Once the first interview had been transcribed, more codes became prevalent. The first one was the participant's interests, as these often matched or expanded upon the existing project. The second code focused on the application of the framework and how it can handle data streams at different spatial resolutions, such as city versus country granularity. The third and fourth were suggestions from the participants on how to improve the thesis and potential future applications.

5.2.2 Transcribing and Coding

The audio files were played back using VLC media player. This was the preferred software as it allowed for a variable playback speed, useful for typing out high-paced conversations. The interviews were played back at 0.6x speed and the discussion transcribed manually using Microsoft Word. At several points of the recording it was necessary to skip back a few second to catch up on typing or to replay some quieter or distorted speech. The distortion arose largely due to the interviews being conducted over Google Hangouts, with the interviewees often using WiFi which can be throttled if in a busy office or patchy if the user is far away from a connection point or the connection point

Code	From Questions	From Interviews
Project Aims	X	
Project Interests		X
Importance of Location	X	
Importance of Timeliness	X	
Scalable Tracking	X	
Scalable Application		X
Usefulness of Framework	X	
Improvements		X
Future Work		X

Table 5.1: Table showing the codes derived from the questions and interviews.

is obstructed by walls. At times the connection was poor and distorted noise high, necessitating the use of the skip back function in VLC.

To code the interviews, the Microsoft Word document was imported into NVivo. NVivo is a widely-used commercial information storage and analysis tool that allows for the labelling of textual data using nodes. These nodes are manually created by the user and can be manipulated into thematic categories, such as ‘Project Aims’, ‘Usefulness of Framework’ or ‘Improvements’, which can then be applied to the text to label sections of the interview. In the literature, nodes are used to give structure to unstructured text or images (Bazeley and Jackson, 2014) and are popular amongst qualitative work (Fereday and Muir-Cochrane, 2006; Clifford et al., 2016; Jeske et al., 2017). When the text matched a code, the text was highlighted and dragged over the appropriate node. As a semi-structured interview technique allows for natural questions and answers to arise, both a deductive and inductive approach was taken to generate the codes.

5.2.2.1 Interview Overview

The first three interviews were carried out with current or former employees of the Open Innovation team at HM Treasury, include a current and former technical lead and the head of the team. These interviews were highly focused on how the Treasury could benefit from the thesis framework, with very positive suggestions as to its inclusion in future projects. An issue shared with all participants was the trustworthiness of the results, with them agreeing that a team working with the framework would be required to understand how it aligns with HM Treasury’s project goals.

The last three interviews were conducted over the phone with employees of ONS who work with GDELT¹ data. This was more of a ‘Big Picture’ approach, as these participants were knowledgeable about other existing research projects. These interviews were

¹GDELT is a research project led by ONS to investigate global news publications and extract location-based events.

very positive, with a possible inclusion of this thesis framework in a GDELT spin-off. Similar concerns arose about trustworthiness, but as GDELT is also a proof-of-concept model these concerns were not as paramount as with HM Treasury.

5.3 Evaluating the Framework

All interviews started off with brief introductions and job roles. When asked about project aims and interests (Q1), the Treasury team had very specific answers.

Treasury lead: *[T]he rationale for doing this stream of work was to meet a need expressed by the Treasury to generate better quality real-time economic indicators to inform their thinking and provide better quality advice to ministers ... there are currently very limited good quality real-time economic indicators and therefore Treasury and other officials feel like they're not able to give ministers particularly good quality picture ... so they want to try to improve that situation and it seems like the best way to do that is to encourage and raise awareness of the experimental work on economic indicators that is going on.*

This quotation from the team lead highlights their need for clear communication. The thesis addresses this concern by its focus on cleaning the raw data, therefore producing maps and datasets that have significantly reduced noise.

When asked the same question, the former and current technical leads expressed a desire to obtain information faster than current existing methods.

Former Treasury employee: *[The real-time economic indicator project] is an ongoing project ... to try and promote connections between academics and policy makers. The project specifically I was working on was looking at experimental economic indicators. People at Treasury and ONS are interested in new ways of measuring how the economy is doing at a particular moment in time. Part of that is trying to capture parts of the economy which our traditional surveys that the ONS do don't capture particularly well, like something like an online labour market, we don't really have good data on that.*

Current Treasury employee: *So my understanding is the ONS collects official data that is used and trusted by government and by private sector. I think there's appetite to see whether other types of indicators can be useful, especially in real-time, in helping decision makers make policies to understand*

what is happening in the economy. Having indicators that are available immediately - ones you don't have to wait a year for - could be of super value to the policy-making process.

All three Treasury participants agreed that the real-time economic indicators project aimed to obtain economic data faster than established methods, primarily to create and advise on policies faster than their existing timeframe of months or years. They therefore require succinct information that can be requested and returned as quickly as possible.

The ONS participants worked with GDELT data with an economic focus, rather than directly as with the Treasury employees. Therefore, their answers were less policy focused and quite similar, and can be summed up by the lead's answer. The ONS participants also afforded an insight into the wider applications of the thesis framework outside of economic indicators.

ONS lead: So I'm at the moment working on the day-to-day global database for language emotion and tone (GDELT) and they scan news articles across the globe every 15 minutes and build a big database based on a lot of the information that they extract from these news articles. Another project that is still in the planning phase is about looking at articles that are about the economy and trying to see what kind of sentiment colours the article over time and also trying to see which of those articles are about locations within the UK or people, public figures from the UK to try and see articles about the UK economic issues and seeing how sentiment changes over time.

As the ONS were more concerned about sentiment and topics, the LDA sections of the thesis framework will particularly apply to their needs. To summarise, both the Treasury and ONS are interested in location-based economic indicators that can be detected, analysed and visualised in a timely manner for the purposes of informing policy makers or understanding the sentiment of the population.

5.3.1 Importance of Location Analysis

For all participants, location was an important factor (Q2). For the Treasury, knowing where an economic indicator originated was key for informing on policy; for the ONS, location was a more abstract notion and could be at country level, but they also expressed a desire to generate more specific results, an aspect which the thesis framework can improve.

Treasury lead: So for the regional policy agenda of which there are many subdivisions including things like city deals or various other kind of deals that

go on, to be able to analyse what is happening in localities in more detail with greater confidence and in a more timely way is really really valuable. Whatever kind of light can be shed on those investments and what is actually happening in the economy and therefore what kind of investment we might want to make to improve the situation is extremely valuable. A lot more needs to be gathered to improve it, there is a lot of appetite for that.

The lead focused on the ability of location data enable policy making to be more efficient and specific. The former Treasury employee did not have as much to say about location information, but tied it back to the Treasury's goals and emphasised the predictive capabilities.

Former Treasury employee: *What they're [HM Treasury] concerned with is economic output, economic activity. So if there were potential for this to give an indicator as a predictor or as a certain nowcasting or forecasting economic activity that certainly could be of interest.*

The current Treasury employee was more concerned with the ability for the thesis framework to clean up potential location noise.

Current Treasury employee: *Yeah I think it's interesting. So it shows geolocation of people talking about subjects as tags right? But it depends on what you're tagging - what do you mean they're talking about the economy exactly?*

This answer directly questioned the cleaning process in the thesis framework. However, once the participant had seen the before and after cleaning images, they were confident that the framework was capable of producing usable data. This answer also points towards potential future work.

Current Treasury employee: *You can overlay [the tweets] with police data where knife crime is taking place. That could show up some really interesting things to investigate and to you know, if there is a correlation between what people talk about knife crime and the location of knife crime attacks for example, then that's what you'd expect to see. But if there isn't then investigating that and taking a closer look at that is very interesting.*

The ONS lead was initially uncertain as to the extent of Twitter geolocation granularity, but after some discussion they saw the benefits of using such fine-grained data.

ONS lead: *And also maybe looking for things outside the UK eventually. I don't know if you have the same level of granularities in different areas in the world. Our idea is to go beyond the UK ideally.*

After some discussion: *The greater granularity in terms of location would be something that would be very useful, especially thinking about the disaster use case for example, one thing we are trying to do is to look at economic damage for example in a certain area. So if we actually knew where the disaster hit an area we can narrow it down and look specifically for mentions of damage to buildings or power cuts or things like that.*

As previously mentioned, the ONS participants were more focused on country or global issues. One employee was similarly concerned that the Twitter data was too fine-grained but once assured that the data can be aggregated to suit their needs they gave a positive response.

ONS employee: *I don't think we'll be looking for the Southampton fish market or something, which is how your maps are going down to that level to know what is going on within a city. I'm thinking for GDELT, the obvious thing to start with is city-level.*

After some discussion: *You've got that definite geolocation, it's much clearer as to what you're looking at, whereas the GDELT is processed data. The data we're looking at is heavily processed with text analysis. There's very little control over the raw data going in.*

5.3.2 Importance of Timeliness

In line with location granularity, being able to distinguish changes over time is an important factor (Q3). All tweets come with a timestamp, therefore the framework is capable of creating outputs between time frames of interest. The Treasury lead revealed a change in their project semantics from a focus on real-time to purely something that is faster than current methods without sacrificing quality.

Treasury lead: *We now don't call it real-time economic indicators we call it experimental economic indicators, because they don't necessarily need to be available or represent what is going on in the economy right in this moment, they just need to be an improvement of what we have now. It's not about the timeliness of the work it's more about how you can produce the indicator in a better way that tells us more about the economy, that's more valid. Our work in this space is no longer limited to real-time indicators, it's an improvement in timeliness and also other types of economic indicators that are improvements, whether they are improvements in timeliness or not.*

When told the thesis framework analyses tweets in a timely manner but is likely to output results that are at least a day lagged, the lead elaborated.

Treasury lead: *Yeah that's absolutely fine. That's why we ditched real-time as it is misleading what our interests are. If it was an improvement in the status quo then that's alright. That's all we can hope for right?*

The former Treasury employee also reinforced the goal to achieve results faster than waiting for official statistics, emphasising the lack of data currently available to the Treasury.

Former Treasury employee: *In terms of work I was doing, the main focus was on essentially GDP at a more regional level is great, GDP more accurate is great, GDP in real-time, or with slower lags is great. That is definitely a lot of appetite for that. The project is trying to solve the lack of regional data that comes out in a timely way. ... There can be some encouragement there that there's a hole in what's currently available, we don't have particularly good data on something like economic activity at a regional level at real-time.*

As has already been mentioned above, the current Treasury employee expressed a keen interest in obtaining actionable data faster than existing methods. When interviewing the ONS employees, they did not have such a focus on timeliness as the GDELT data were known to be lagged. However, they did express an interest in incorporating the thesis into a GDELT spin-off that focused much more on real-time analysis.

ONS employee: *GDELT is quite good at political things like social signs and protests and demonstrations are picked up quite nicely. Disasters that are not quite directly social science are a bit trickier. ... I think what you just explained is pretty much aligned with one of the potential spin-offs that we're looking at which is the economic-sentiment idea based on GDELT articles that are about things in the UK in real-time.*

5.3.3 Spatial and Temporal Scalability

All participants were interesting in how the thesis framework can be applied at different scales (Q4), both temporally and spatially. The Treasury team were more concerned with the trustworthiness of the data and less so about the finer details, while the ONS team were very interested in the ability for the framework to be applied at different geographic granularities. Due to scalability having an overlap with location, a large section of these answers have already been discussed above, therefore only new content is displayed.

Treasury lead: *To answer your question directly about the geographies its really important to be able to compare this geography with other types of geography. Great stuff.*

The current employee raised policy making advantages relating to a city- or country-level analysis.

Current Treasury employee: *Yeah it'd be interesting to see, you know, why not? Why not look at this city, look at how people behave in big cities, travel to work areas, how do you track people across the city? Movement? I think there's value in looking at cities and other cities and I think in terms of subject areas it could be anything from like what are people talking about? What are the key things that people do in their free time, in the evenings? What do people care about? There could be political aspects here. People talk about their concerns, what they're not happy about, this has political implications. Getting a sense of what is the mood of the country?*

ONS employee: *How do you account for Twitter research pattern where it comes to an area where you get a lot of people who use Twitter compared with areas without the social media coverage?*

The point from the ONS employee was a key criticism that emerged from the interviews. As this thesis framework is a proof-of-concept model, it does not claim to replace existing methods but aims to supplement them. This framework is capable of examining tweets and creating maps showing hotspots of activity. Therefore, the thesis provides details about topics and locations represented through this medium that are not currently used by the Treasury or ONS. Areas without this data are outside the scope of the thesis.

5.3.4 Usefulness of the Framework

Perhaps the most crucial question of all, at the end of the interviews the participants were asked their opinions of the framework and whether they find it useful. All participants responded positively. The Treasury lead, having the broadest understanding about the context of the real-time economic project, expressed the greatest appreciation for the location-based and topic modelling approaches of the framework.

Treasury lead: *We don't have good quality up-to-date information about what's happening in the economy or society. We're very interested in gathering this type of data to inform our decisions. This seems to be a promising approach to filling in some of those blanks.*

The former employee was reluctant to adopt the framework without further testing, expressing the need for the Treasury to work with proven models. Despite this, the former employee did reinforce the Treasury's interest in collecting the level of detail that is extracted and presented by the framework.

Former Treasury employee: *Experimental work certainly has value along the way to inform the approach and to help shed a bit of light on things we don't really understand. I think obviously the stakes are quite high for the Treasury, they don't want to make mistakes and will only use data they trust. Caveating it with that, I am very happy for you to say that this is the sort of work that could be of interest to policy makers.*

The current Treasury employee expressed a desire to understand the emotions felt by the public, thus reinforcing the idea to only use LDA on the larger datasets to produce topics with stronger term relevance (discussed previously in Chapter 4) or to carry out sentiment analysis, the machine learning approach of classifying text as expression either positive, negative or neutral emotion, often with accompanying strength scores (Maynard et al., 2012; Cambria and White, 2014).

Current Treasury employee: *The applications and use of this is quite extensive, it is a really interesting project. This is one of the benefits of social media that you know people generate their own content. That makes it quite unique because people can really say what they feel while a lot of other statistics is measuring what people have reported which may not be the same as what's actually happened. When people know when individuals are being audited or asked something certain they tend to say different things, so in terms of economic indicators is really useful. It's not just about real-time, you can reveal what people really feel rather than what they just tell people. I think combining that information with other information is very useful, in terms of not just what people are saying or where they are saying but in terms of other data, so I mentioned knife crime for example, you could really lead to some interesting and innovative ways of looking at an issue.*

There is an understandable reluctance for the Treasury employees to be instantly accepting of the location-based narrative framework due to their reliance on trustworthy sources that have had years of proven accurate results. Despite this, their keen interest in developing proof-of-concept applications speaks positively to this thesis. The analytical techniques used by the framework could be applied to entirely other areas, such as discovering topics of knife crime, rather than anything related to the economy. The framework is sufficiently flexible to be attributed to any topic of interest, a feature desired by those interviewed.

The ONS were similarly positive about the strengths of the framework, and as the GDELT team are used to proof-of-concept models, they were more enthusiastic to adopt the thesis framework into their work.

ONS lead: *I think that could be quite a nice addition in a way, because GDELT and Twitter look at different levels of closeness, with Twitter being embodied by the people directly and GDELT looking at how news can influence what people think about the economy. So having levels of this in terms of experimental economic indicators would be great. I'm looking at something along these lines and I see that in terms of being experimental and especially because of the timeliness there will be a use for it.*

The ONS employee further corroborated the idea of the location-based narrative framework being a useful tool to incorporate into the larger GDELT project.

ONS employee: *I think your work is incredibly useful, it looks very similar to things we've been thinking about with the GDELT data. I can imagine a more useful application would be for GDELT to tell you that a disaster is happening ... [then] you can see [tweet] hotspots in Newcastle, maybe there's a storm happening. Twitter goes to a much more fine detail. The fact that there have been some papers that compare GDELT with Twitter ending up saying Twitter is better.*

These positive interviews show that location-based narrative analysis, while still proof-of-concept, is highly valuable to government organisations. While the topic of real-time economic indicators was used for the case study, the techniques used in the framework can be applied to any topic of interest, a key factor in the interest expressed by both the Treasury and ONS interviewees.

5.4 Suggested Improvements and Future Work

As previously discussed, a large proportion of the improvements suggested by the Treasury team related to trust. The team offered several suggestions for future work, including creating a monitoring system that covered the UK.

Treasury lead: *I can imagine having a map of the entire country where this sort of thing is monitored constantly and where you are able to kind of show hotspots across the country of positive and negative related economic words and then see how that correlates with well known and more trusted sources of*

indicators of economic performance. So that sort of thing would be the way to I imagine an obvious way to further develop it and refine it, and make it more valuable over time. On its own, having Southampton just like this is probably interesting but not that valuable as it needs the wider context to understand the meaning of it. But I mean, and I'm sure you weren't expecting to completely solve it straight away, but I imagine that's the sort of thing that would be useful.

They also suggested using existing datasets to compare against the Twitter data, offering a means for comparison between the framework output and established economy-related methods which could form part of a future work experiment.

Treasury lead: Try to look at claimant data ... the claimant count is more up to date. So people who are claiming jobseekers' benefits - it's available and accurate at a local level and is updated very frequently.

The former employee suggested producing simpler results, as policy makers are interested in the broader aspects. For the thesis, if the target audience were policy makers this would mean still applying the cleaning framework but only producing maps and topics encompassing all spatial resolutions.

Former Treasury employee: The simple solutions are always going to be the preferable ones to policy makers, because they're time constrained, they're generalists, you're not going to have someone there who is an expert in scraping social media data and text analysis, for good reason there's no point in them being there in that role, they have to do a lot of different things and have an overview.

The current employee expressed an interest in creating an informative dashboard that would be of use to policy makers, reinforcing the former employee's suggestion of focusing the framework results into a simpler form while emphasising the need to collect a wider geographical area for better context and potential applications.

Current Treasury employee: Could you make this into a dashboard that tells you, you know, on any given day this is what people are talking about at any given time of day on social media. You could even have an economy or politics dashboard that tells us what people are talking about at any given time. This could be applied to decision making processes. The applications and use of this is quite extensive right?

The ONS team employees had a more statistical approach, suggesting a different way of highlighting topics rather than using KDE.

ONS lead: *Have you considered instead of going for absolute numbers of tweets in an area about a specific topic, going for the percentage of tweets about a certain area that you're interested in? I think that goes in the same direction. If you have like few tweets in an area but they are about a specific topic and in another area and you have loads of tweets but relatively few about this specific topic, because of the sheer number that is a lot.*

They also were keen to adopt the framework within future projects.

ONS employee: *This phase of exploration is what we're trying to wrap ... and then we've got to sit down and think about whether we've got ideas about what the next project is. Then we'll have to get that prioritised. We can certainly think about Twitter within that and the work you've been doing.*

ONS employee: *I think that is a really valuable piece. With all of this big data type of work everyone is trying to find that validation. It's not just "oh you have something that looks plausible". If people build up that credible link there, I think this is something that we might be able to continue with GDELT.*

5.5 Summary

The participants in the interview phase were shown before and after maps of the tweets that had been tagged with the Harvard, Thesaurus or Treasury keyword sets. They agreed that the 'after' output was significantly different and a preferred outcome than the noisy 'before' images. They were all asked the five main questions outlined in subsection 5.2.1 and the semi-structured interviews evolved from there. The interviews were coded with a mixture of pre-prepared codes and ones that emerged naturally from the interviews. These codes were summarised into five main topics to evaluate the framework: aims of their projects, importance of location, importance of timeliness, scalability and usefulness.

Suggestions were made about future projects, mostly focusing the keyword and topic modelling analysis down to specific themes, such as knife crime or natural disasters, as well as expanding the area of interest from Southampton and surrounding Hampshire to the whole of the United Kingdom. This increased dataset could then be visualised with an interactive dashboard that would be of significant value to policy makers.

The positive reactions to the framework were driven by its ability to remove noise and increase the accuracy of textual and spatial results. Though still as a proof-of-concept, all interviewees agreed it showed a promising new avenue of research and the ONS were keen to adopt the framework into their ongoing GDELT experiments. There were several areas for improvement, some common and some unique to the teams, and along with the overview and limitations of the thesis framework shall be discussed in the next chapter. The positive evaluation answers RQ3, proving that a locative thematic narrative analysis framework can be applied to a real-world scenario to produce actionable results.

5.5.1 Outcomes

The original aim of the project was to analyse Twitter data to extract economic activity, and present this in such a way as to be useful for policy makers. While the goal was noble, the broad category of ‘economic’ indicators caused several issues with specificity. While policy makers are interested in generalisable advice, the lack of a defined economic activity prevented the framework from creating meaningful economic results for HM Treasury. For instance, focusing the framework to analyse unemployment would have been preferable for them as that would target a specific issue, thus the main outcome of these interviews was to better understand HM Treasury’s perspective once faced with a proof-of-concept model, and to highlight the need for more actionable results. Nevertheless, the GDELT team were very pleased with the proof-of-concept model and could see an immediate uptake into their projects. Furthermore, the mapping aspect of the framework was very interesting for other projects currently undertaken at HM Treasury. Therefore, as will be discussed further in the next chapter, future adaptation and applications of the framework will cover specific economic topics.

A second, indirect outcome was the clarification of what HM Treasury meant by ‘real-time’. As already mentioned, the project was relabelled to ‘experimental economic indicators’, removing the emphasis on ‘real-time’ as the results only needed to be an improvement on their existing methods. To borrow from military terminology, this indicates a shift from ‘tactical’ to ‘strategic’, which means a change from ‘right here right now’ to a ‘longer, more thought out’ approach. While this shift might sound self-explanatory, it represents a paradigm shift in how the data are collected and processed. The emphasis is now no longer on obtaining as much data as quickly as possible and immediately verifying its validity to generate accurate results, which is very resource intensive and requires rigorous processes to be in place; instead, a strategic approach implies the collection of data over a longer period of time, experimenting with different analyses and presentations, as well as any result forming part of a wider portfolio of information for policy makers.

5.5.2 Advice and Improvements

While the interviews were at a more conceptual level, some brief advice was conveyed as to how to improve the framework. As already mentioned, applying the framework to specific economic problems such as unemployment would create a smaller but more actionable set of results for HM Treasury. The GDELT project focuses on analysing world media, thus some ability to analyse newspapers or online news websites to match against any results from the framework would be an advantage. A key piece of advice from HM Treasury was to create a dashboard. This dashboard would show a summary of the results in an interactive fashion to allow policy makers to audit the results before advising on policies. It could also pull in existing data from HM Treasury to support the outcomes, such as knife crime statistics being pulled in next to a map of tweets relating to knife crime. This combination of information sources is of interest to both HM Treasury and the GDELT team. The auditing and pulling in of other data sources, such as lagged knife crime statistics, allows for a robustness check of the dashboard, an ability enabled by taking the more strategic approach.

Chapter 6

Conclusions

This thesis sought to combine narratology and computational geography to create a framework for extracting specific location-based topics. It achieved this by identifying methodological gaps in the research and conducting progressive experiments to address them. The thesis was exploratory and tried several different approaches before concluding upon a validated framework. This chapter summarises the flow and contribution of the thesis before discussing limitations and future work.

6.1 Overall Thesis Summary

Chapter 1 introduced the relevant academic and technical themes that were involved in this study, covering an overview of narrative approaches, location analyses and technical aspects of collecting Twitter data.

Chapter 2 critically analysed previous work into the three main aspects of studying this medium: narratology, natural language processing and spatial analytics. The chapter discussed the gaps in the research created by authors who either did not appropriately clean the data or used methods that have since become outdated. The main conclusion from the existing literature was that sufficient work into analysing third-party sources and location clustering had not been conducted, thus results based on this limited knowledge were not adequately refined.

Chapter 3 comprised the creation of the framework that forms the main contribution of this thesis. The framework arose through a series of data-driven experiments. The first experiment attempted to model the multivariate nature of events, drawing from literature to create a parallel coordinate model and validate the approach through examples. While this was a useful experiment to identify these natures, it could not properly model the rich narratives inherent in these events. The second experiment rectified this by focusing on two themes, entertainment and commerce, and used term expansion to map

out related tweets over Southampton, UK. While the results were more informative, inconsistencies in how third-party tweets were generated created anomalous clusters. These clusters were explored in more depth in the third experiment, which queried the Foursquare API to reclaim the missing POI information, revealing how previously perceived tweet clusters were in reality artificially generated. The fourth experiment traced the tweets back to their origin to collect more data, concluding that third-party tweets are often clustered around POIs but lose this metadata when shared, as well as reclaiming some of the longer third-party posts that had been truncated to conform to Twitter's character limitations. The fifth and final experiment greatly expanded on the location analyses by querying the Foursquare, Google and OSM APIs to obtain an extensive dataset of tweet-related location information. These queries allowed the framework to distinguish between native and third-party tweets at street, neighbourhood and city levels with 79.1% accuracy and 81.1% recall, providing rich and useful information for policy makers.

Chapter 4 took this framework further by applying it to a real-world problem. HM Treasury and the Cabinet Office are interested in discovering and analysing experimental economic indicators. The experiment used three sets of keywords to generate different related themes. The first originated from a gold-standard set of economic terms, the second from an online thesaurus and the third from experts within the Treasury. The framework was applied to the three resulting datasets, with maps illustrating the impact of each keyword set and the cleaning process on the raw tweet dataset. The results showed areas of economic activity that matched expected areas, such as high streets. It also identified localised areas of economic activity, such as offices or groups of users discussing the economy. The application of the thesis' rigorous location cleaning allowed for more representative maps of activity rather than the misleading ones that could have been produced had tweet clusters not been analysed. Creating trustworthy results is key for policy makers and the Treasury, thus the results produced by the framework were shown to members of the Treasury and the Office for National Statistics during evaluative interviews.

Chapter 5 covers these interviews. The participants are asked relevant questions that relate to what they require from the data as well as their professional opinions about the framework. They all agreed the project had merit and could see the benefit of adopting the framework, with some suggesting future work into more specific applications such as analysing tweets mentioning knife crime.

6.1.1 Addressing the Research Questions

As mentioned at the end of Chapter 2, the research questions addressed by this thesis were:

- RQ1: To what extent does a thematic approach afford a richer understanding of location-based activity than topic modelling?
- RQ2: Can location-based thematic modelling be automated?
- RQ3: Can this approach be applied to a real-world situation with trustworthy results?

The first research question addressed the issue of topic modelling via Latent Dirichlet Allocation (LDA) and keyword extraction. This thesis showed that LDA alone produced one-dimensional results; the extracted topics only reflected those that were popular within the dataset while simultaneously removing their context. Some of these popular topics were shown to originate from music, historical or sporting events, thus general topics that were pervasive but obscure were hidden. Summarising the conversations in this way produced ephemeral topics and ones that were not useful in understanding ambient narratives. Term expansion and keyword tagging produced broader and often noisier results, but with sufficient cleaning the data were more reflective of themes of conversations rather than summarised topics. These themes are more interesting and useful than summarised topics as they are more widely applicable and encompass a larger audience, thus enabling the understanding of a larger proportion of the population to better create policies. Had LDA been applied to the tweets without first filtering, the results would be skewed by noisy users or popular topics, as was shown in the first experiment in Chapter 3. Keyword matching allowed specific themes to first be extracted before applying LDA, resulting in an LDA output that reflected topics of interest rather than only what was prevalent, shown in the second experiment in Chapter 3 and the case study in Chapter 4. This more nuanced approach created more relevant outputs for policy makers, though to be actionable would require more evaluation and validation in-line with government auditing and robustness checks.

The second research question addressed the automation of the entire process. While the question would suggest a binary answer, there are nuances that need to be considered. The majority of the framework elements have been fully automated, such as the collection of tweets using Python and the Twitter API (with human-defined search parameters), synonym collection, tweet cleaning and mapping scripts, LDA generation and geospatial API calling. Nuances arose in several areas of the framework, such as how to clean the synonym sets, at which point to cut off the term expansion, how to determine which third-party sources to remove, or at what distance from a cluster to accept a POI classification. These aspects were configured in the framework to fit a particular theme and an urban environment, thus if the framework were to be applied to a rural area then these aspects would need adjusting in the code. It emerged during the interviews that HM Treasury does not desire a fully automated system; they prefer a human, such as a policy maker, to have control over the system and configure the framework as they see fit, such as through a dashboard. While a dashboard was not created as part of

the framework, it was clear that HM Treasury wished for human intervention in the automation process for auditing and validation of the results.

The data collection, cleaning and analysis code was written in Python 2.7.10 with visualisations created in R 3.6.0. These versions are not mandatory as all the libraries used were ‘mainstream’, and as such later versions of Python and R will still create the same results¹. Some level of coding experience is needed to make modifications to the scripts, but as all knowledge of programming was gained during the PhD process, this level is not high. The software suites Anaconda² and RStudio³ were used to create and run all scripts used in the thesis on a standalone Macbook Pro 2015 with 256GB hard drive and 16GB RAM, with an external 2TB hard drive for data storage and a consistent Internet connection. If an unfamiliar user were to take on this framework, they would need a working knowledge of Python and R, as well as Microsoft Excel and JSON for data storage and a reliable Internet connection for tweet and synonym collection. All software was platform agnostic, so it will work on Windows or Mac (Linux was not tested). If the user wished to collect tweets over a larger spatial or temporal extent this would increase the size of the JSON files quite quickly, thus data storage management would also be required. A rule of thumb is to assume a million tweets is 1GB of space, thus careful consideration is required when collecting several million tweets as this figure is just for JSON file sizes and not any further files saved during the analysis process.

The final research question involved applying the framework to a real-world case study and measuring its success. Chapter 3 initially outlined four experiments that created and refined the framework, culminating with a fifth experiment that evaluated the framework by measuring its precision, recall and overall accuracy. Chapter 3 therefore made significant progress towards measuring the trustworthiness of the framework. Chapter 4 outlined the real-world case study with the Treasury and Cabinet Office to discover and model tweets that related to the economy. The experiment analysed the topics present in tweets tagged with three keyword sets and validated these topics by matching them against their geographical profiles, proving that real-world topics were being extracted rather than false positives. For this experiment to be successful, it would need external validation by HM Treasury. As shown by the qualitative interviews in Chapter 5, members of HM Treasury were in favour of the framework and saw its validity and usefulness in future applications. There was also little risk of the interviews being biased as HM Treasury were highly critical of the framework’s trustworthiness. As with all experiments, they require years of robustness checks and evaluations before a framework is adopted into mainstream policy advice. They were keen to develop the framework as a source of supplementary information for policy makers in the form of a dashboard. Their willingness to continue work on the framework could ultimately be considered a measure of the success of the thesis and positively answered RQ3. Additionally, the

¹As of June 2020.

²<https://www.anaconda.com/>

³<https://rstudio.com/>

GDELT team were very keen to adopt the framework in its current form, further confirming its validity and trustworthiness. While ‘trust’ is a hard value to measure, the framework proved to be statistically reliable and of interest to HM Treasury and ONS; therefore, the thesis has been successful in creating a framework of sufficient robustness and trustworthiness to be of interest to central government.

6.2 Contribution

This thesis contributes a text cleaning, modelling and mapping framework. Parts of this work were published in [Bennett et al. \(2016, 2017a,b, 2018\)](#) and helped to advance the literature by critically analysing the metadata fields present in tweets to an extent that had not been previously researched, allowing a much more geographically accurate analysis. This work highlighted key methodological changes that need to be implemented to make Twitter analysis more robust, exemplifying these changes with a real-world case study.

The thesis advances the literature in several areas. Firstly, it highlighted the risks of one-dimensional analyses of LDA, arguing for and proving that a thematic approach generated more narrative-rich results. This was achieved by first filtering the tweets by keywords, creating a more specific LDA result, then validating these results by comparing the textual and spatial relationships. Secondly, through Chapter 3 it discovered major issues with Twitter metadata that were either ignored or not addressed in previous work. In particular it brought to attention the clustering of third-party tweets that would go unnoticed by researchers if they did not investigate further, as seen in [Patel et al. \(2017\)](#) who generated what they thought were accurate results using clustered tweets, and [Lin and Cromley \(2018\)](#) who assumed that a single tweet could be a cluster, resulting in their classifier of home locations often being under 50% accurate. Thirdly, it synthesised methods from previous LDA, Twitter and spatial works into a real-world case study, transforming academic examples into what could be used by government. This example is perhaps the most impactful but also least explored, as translating academic work into actionable government policy takes many years of validation. However, as shown in Chapter 5, this thesis made a solid contribution towards this goal.

The tweets and additional data were collected and stored as JSON, the JavaScript Object Notation format that is particularly common when querying APIs. From this the Python code translated the JSON metadata into a spreadsheet (CSV) for further analyses in Python and R. Therefore, if data are obtainable in JSON or CSV format, the framework can be applied to them. The vital elements for LDA and KDE are the text and coordinate fields, but with these obtained the framework can process additional data sources such as gazetteers.

If this framework were applied to other sources with longer text data it would improve performance, thus it is scalable to conventionally larger text-based sources such as newspaper articles. As LDA relies on having several different texts to analyse, it would struggle if presented with datasets containing fewer than 100 texts, as seen in Experiment 4. KDE does not have such limitations, though the ability to extract meaningful spatial narratives is similarly hindered by small datasets.

While the initial source of this data came from Twitter, additional data were collected from Foursquare, Google and OSM, all of which were processed by the framework. The framework is thus not restricted by only being applicable to Twitter data, and interesting future work would be to apply the framework to other media sources. With the overall contribution outlined, the component parts are similarly discussed.

6.2.1 Twitter Cleaning Methodology

The idea to analyse location data in the way presented by the thesis had not been previously addressed in the literature, but made a significant impact on the representation of subsequent maps. Previous work that did not take into account source cleaning or cluster analysis risked producing results that were misleading at best or completely wrong at worst. For example, city centroids that were generated by Twitter or its third-party providers would create a strong KDE signal, indicative of high activity. Work that did not delve more deeply into these clusters risked misinterpreting the signals. This thesis therefore sheds much needed light onto this problem and provides several solutions both for cleaning the data and validating its spatial granularity.

6.2.1.1 Source Analysis

This thesis provides a novel methodology for identifying and removing irrelevant third-party sources, cleaning artificially clustered tweets and extracting topics of interest. While this work was based on Twitter data, it can be applied to any textual data that originates from more than one source. The sources and unique users are ranked and a ratio for source-to-user is created. This ratio identifies noisy sources that are only populated by a few users but create hundreds or thousands of posts. Were these sources not removed from the overall data, subsequent analyses and results would be based on unnecessarily noisy data.

6.2.1.2 Cluster Analysis

While the majority of noisy clusters are resolved by removing irrelevant sources, some that remain (including native sources) will create artificial clusters. These are due to

a stationary user, a popular point of interest, or users attributing their posts to just the relevant neighbourhood or city. Again, sources that do not respect these technical behaviours and take all tweets at face value risk misrepresenting their results. However, these clusters can be analysed to properly classify them at their true spatial granularity. The experiment outlined in Chapter 3 section 3.5 compared these clusters to three publicly available APIs and generated address information to validate their location. By querying these APIs the clusters were classified as at uniquely clustered, POI, neighbourhood or city granularity with around 80% default accuracy and up to 100% accuracy when adjusted with a manual inspection. With the cluster granularity identified and validated, the investigator obtains a more accurate understanding of the nature of the tweet posts, thus subsequent results better reflect the users than before. The tweets can now be reliably mapped at precise, POI, neighbourhood or city-level granularities. Therefore, this thesis has produced a powerful analytical framework that makes a significant improvement upon previous work.

6.2.1.3 Topic Analysis

Both machine learning and simplistic approaches to topic modelling were used as part of the framework creation process outlined in Chapter 3. LDA creates a probabilistic model of common threads of conversation within the dataset, while keyword matching extracts specific examples of the desired topics. While LDA has its strengths within known datasets, such as those generated during disasters, to use LDA over a general dataset extracts all the possible topics and thus the results can be hard to interpret. If the investigators know which topics are of interest, it is better to use keyword matching to extract all those that are relevant. This was demonstrated in Chapter 4 where keyword matching was used to extract specific topics about the economy. LDA was used to validate the existence of related topics, but was not used as the primary analytical method. First using keywords then LDA is a strong approach, as the keywords extracted the relevant tweets then the LDA confirmed the presence of topics from the smaller and more specific subset.

6.2.2 Limitations

This thesis successfully modelled the location-based narratives in tweets. It did this by analysing topics and location data inherent in the tweet metadata to produce maps of economy-related activity. It was then validated and evaluated by interviewing key employees of government departments.

As with any project involving social media data, the prominent caveat is the data only reflect a subset of the population that uses social media. This is a small subdivision of the population, though is very much a proportion that is well represented and their

behaviours can be used to inform policy makers. As previously mentioned, though the data were initially obtained via Twitter, extra sources were used to support location-based narrative extraction. Due to Twitter's character limit of 140 then 280 in early 2018, the likelihood for each tweet to have a word that matches the three keyword sets used is smaller than if all the data came from platforms that do not restrict the text. Twitter was chosen for its open and accessible API, though if the thesis experiments were repeated, more emphasis would be added on collecting data directly from other platforms such as Foursquare or its partner service Swarm. Instagram, while by far the largest source of geolocated posts, is actively restricting the use of its APIs by researchers, thus the framework is needed to analyse such sources. With the framework itself, there were several decisions that had to be made along the way. Choosing to use KDA and LDA will have strongly influenced the results.

6.2.2.1 Spatial Analyses

Using KDE was one of many choices. There are other ways to cluster spatial data which would have produced slightly different maps, such as k-means or random forest. Future work would investigate the use of spatial statistics such as Local Moran's I and Getis Ord to identify anomalous clusters, creating additional understanding of the nuances of tweet activity. These models differentiate between local clusters, comparing them against neighbouring clusters, to identify anomalous hot- or cold-spots within larger clusters. This is useful for more intricate spatial analyses as it identifies the nuances within larger unit areas. Similarly, adaptive KDE (as discussed in [Lloyd and Cheshire \(2017\)](#)) could be applied to more intricately map the hotspots by adjusting the kernel size used to generate the signal and potentially improve the KDE accuracy. These techniques would create different heatmaps, thus potentially identifying other areas of interest; however, to fully understand and implement the methods was outside the scope of the thesis.

6.2.2.2 Topic Modelling

A similar discussion can be made of topic modelling algorithms. LDA is one such algorithm that allows words to belong to several topics thus allowing ambiguous terms to fit within their (restricted) context. Allowing ambiguity within topic modelling better fits the complex natural language of everyday life, thus the results will be more reflective. However, the topics are therefore less well-defined than they could have been and are thus open to interpretation. K-means, an algorithm mentioned before as a potential one for spatial clustering, also works for topic modelling. It applies the same approach by converting the text into a vector matrix and measuring distances. While k-means creates more definitive topics, it only allows each word to have one corresponding vector, thus a word can only belong to one topic which does not truly reflect natural language.

While keyword matching has the potential to better reflect natural language, discussed in Chapter 2, it is a simplistic approach that only extracts topics that are exact matches. While this produces specific results that are strongly related to the desired topic, it will miss the more nuanced matters indirectly discussed by the users. Using a thesaurus keyword set mitigates against this, though produces less specific results. If the desire is to extract all potential topics then this thesis framework achieves this; however, if the desire is to extract the topics alongside the manner in which they are discussed, then a more refined approach is needed such as sentiment analysis or an LDA model trained with an existing set of relevant keywords. This thesis has created three sets of economy-related keywords, so there is potential for these to be converted into a training set for future use.

6.3 Future Work

There are several areas of the framework that could be improved to make it more powerful and robust. Privacy is a key issue for social media users and this impacts their willingness to attribute location information to their posts. As seen in [Middleton et al. \(2014\)](#), their geo-parsing of textual messages enabled the location tagging of non-geolocated tweets. This method is complex and has a tendency for false positives, but is a promising avenue of future work to enable the framework to analyse tweets that were not directly geolocated upon creation. In the final tweet set used in this thesis there were over 16 million non-geolocated tweets collected. If even 1% of these tweets were enriched with a location tag then this would double the existing geolocated dataset. Increasing the applicable dataset would enable a better understanding of how reflective Twitter data are of the population. For example, tweets about coffee shops could be compared to coffee shop sales to see if there is a statistical relationship. This would also test the framework's ability to model other thematic narratives.

Related to this linguistic analysis is part-of-speech tagging. This involves classifying a word as a noun, verb, adjective and so on. There were several instances during Chapter 3 where the same word was in both Commerce and Entertainment classifications, such as 'work'. In the context of commerce, 'work' was used as a noun, for instance 'going to work', whereas in entertainment it was a verb, such as 'this act works wonders'. The keyword tagging process did not take part-of-speech into account, so a few tweets about jobs appeared in Entertainment, and some tweets commending an act appeared in Commerce. Part-of-speech tagging would thus differentiate between the two instances of 'work' and refine this tagging process.

The various ways to apply kernel density estimation and related tests like Getis Ord and Local Moran's have the potential to improve the statistical analysis of the geolocated tweets. As explained in Experiment 2 in Chapter 3 the inclusion of all density values

would create diluted plots, as these were significantly greater than the median density values. Adopting an adaptive KDE model like in [Lloyd and Cheshire \(2017\)](#) would mitigate against very dense areas skewing the results. Also, applying Local Moran's would identify local areas of high density activity within already dense urban areas. This would expand the statistical impact of the maps and contribute an estimation of the expected density within a city like Southampton.

In later stages of the framework, clusters of one user were removed. These clusters had the exact same coordinate pairs, so at first glance could be a neighbourhood or POI cluster, but only contained a single user. Through manual analysis the clusters were frequently identified as sales spam, where a user had set up a bot to promote their company or shop, which would match some or all of the real-time economic indicator keywords and create false positives. These tweets were removed to mitigate against the users biasing location results, as well as most of the textual content being irrelevant to the task. However, a blanket approach to removing these clusters also removed genuine clusters by people using a desktop computer. While these were definitely in the minority, their contributions were being overlooked. To analyse these clusters, keyword matching and LDA would be used to identify and summarise these users' tweets. The results from the LDA would then be converted into a pseudo-tweet and the number of tweets reduced to 1. This is an important reduction to make as even if the cluster were useful and generated by a genuine user, they would still be biasing location-based analyses. This reduction method would enable the user and their tweets to be appropriately analysed by the framework. As shown in Chapter 4, 2,372 of these clusters were identified, thus a future analytical stage to extract genuine users would add a substantial new data source. Similarly, over 1.8 million tweets by almost 25,000 users were tagged to a larger bounding box such as a city or county. While only precisely geolocated tweets were considered by this framework, expanding its remit to cover other types of geolocation would significantly increase the number of available tweets and users. These bounding box tweets would not be used for any street-level analyses, but existing precise tweets could be incorporated into city- or county-level analyses.

The most significant future work piece would be to transform the input and output of the framework into an interactive dashboard. This was a key suggestion from HM Treasury as it would simplify the results, enable policy makers to interact directly with the process and support robustness checks and auditing. The dashboard would be written in R to best interact with existing processes. R has a popular library called Shiny which enables and supports website creation. The user would then be able to upload raw tweet data to the website, which would be hosted on a powerful server to enable data processing on demand. The raw data would then be analysed by the framework, summarised and displayed on the dashboard. Additionally, the user would have the ability to filter the results as they see fit. The framework would highlight and remove sources that it deemed spam, as was done in this thesis, but the user would be able to see this and potentially

undo this stage if they wished the source to remain. Similarly, if they are not interested in keeping precise granularity, they could keep clusters of one user or incorporate the city- or county-level tweets. They would be able to upload their own keyword sets and run LDA over the tagged tweets with their preference for k , though the framework would also produce its own at $k = 5$ and $k = 10$ for comparison. Importantly, they would be able to adjust their parameters to suit their research questions and be able to re-run the analytical processes as many times as necessary. Creation of this dashboard would be a very long process which would require learning all about client-server communications. However, it would drastically improve the impact of the framework and enable a faster uptake by interested parties.

6.3.1 Future Applications

This section outlines the different ways the thesis framework can be applied. Due to its focus on text and location data, and its ability to include other forms of data such as census data, there are a wide range of potential and useful research studies that could be conducted.

From the Treasury interviews emerged several different applications of this thesis into more specific experiments. These all focused on government policy making and how the framework can be adapted to fit the new topics. Additionally, the thesis used data sourced from Twitter, though the methods were deliberately aligned to an objective approach, thus can be easily transferred to other sources of text and location data.

Explicitly mentioned in the first interview was using tweets tagged as relating to crime, specifically knife crime, and comparing that with police data of the same topic. The overlapping heatmaps would show areas of reported knife crime as well as areas in which knife crime is discussed. Where the heatmaps do not overlap is potentially even more interesting as these areas would highlight areas where knife crime is apparent but not reported. The text and location data do not have to only come from Twitter either; newspaper reports, online journals, Facebook or other social media platforms can provide the data and the framework will still follow the same methodology to produce the desired results. The framework would not necessarily claim to identify previously unreported knife crime incidents, but it would help policy makers and law enforcement to better distribute resources based on areas where knife crime is actively discussed.

This framework can be applied to a more specific economic project investigating how negative tweets about the economy is reflected in existing economic datasets. A keyword set generated by an economist in combination with a semantic one would identify any geolocated tweet related to unemployment and map it. A linegraph of tweets over time related to unemployment would then be generated and compared against existing

government unemployment data, which is published at monthly intervals and is accurate to city level⁴.

In this example, the negative sentiment in tweets could be compared with unemployment claims (job seekers' allowance) on a monthly basis, as well as mapping job centres and analysing tweet density over these areas. If the sentiment matches an increase or decrease in unemployment rates then a predictive model could be created to theorise future claim rates and thus allow the Treasury to be prepared for the potential impact on their budget.

Using POI data it is possible to match tweets about healthcare to locations of GP practices and hospitals. The ability to analyse the spatial and narrative attributes of tweets can shed light on how efficiently the facilities are placed and whether more needs to be done to meet the healthcare needs of the population. This can also cover physical as well as mental health.

Similarly, homeless shelters form part of the POI dataset. The tweets can be searched for mentions of homelessness then cross-referenced with the locations of shelters or kitchens to see if additional places are needed. This could potentially reflect the performance of existing locations, and suggest new locations that could better serve those in need within their area.

6.4 Final Thoughts

The research goal of this thesis was to produce a location-based narrative framework that analysed textual and locative data to produce thematic maps and topic models to describe an area or topic of interest. To this extent, the thesis successfully produced a framework that achieved this goal, created by the experiments outlined in Chapter 3 and evaluated by the positive feedback from HM Treasury and ONS shown in Chapter 5.

The results produced by the framework are indicative of activity both spatially and thematically. Topic modelling is notoriously difficult on short texts (Maynard and Hare, 2015) but the topics outlined in Chapter 4 show that it is possible to both spatially extract different tweet resolutions, as well as model topics at each of them. This became more challenging with smaller datasets due to the topic modelling algorithm being unable to adequately construct topics, at which point manual inspection would produce better insights than the computational approach.

As shown in Experiment 3.5 and Chapter 5, the thesis has created valuable insights into the ability for sophisticated location-based cleaning methods to create more accurate results. This thesis has therefore contributed knowledge about and solutions for how

⁴https://www.nomisweb.co.uk/reports/lmp/la/1946157287/subreports/cc_time_series/report.aspx points to the current Southampton unemployment dataset.

a dataset of seemingly precise texts can in fact be an amalgamation created at various geographic resolutions, a vital addition to the wider field of social media analytics.

Appendix A

Appendix A - Tweet Ethics Application

FPSE Ethics Committee FPSE EC Application Form

Ver 6.6e

Refer to the *Instructions* and to the *Guide* documents for a glossary of the key phrases in **bold** and for an explanation of the information required in each section. The *Templates* document provides some text that may be helpful in preparing some of the required appendices.

Replace the highlighted text with the appropriate information.

Note that the size of the text entry boxes provided on this form does **not** indicate the expected amount of information; instead, refer to the *Instructions* and to the *Guide* documents in providing the complete information required in each section. Do **not** duplicate information from one text box to another. Do not otherwise edit this form.

Reference number: ERGO/FPSE/20230	Submission version: 1	Date: 2016-04-20
Name of investigator(s) : Nicholas Bennett		
Name of supervisor(s) (if student investigator(s)): David Millard, David Martin		
Title of study: Towards a classification of event types		
Expected study start date: 2016-05-02	Expected study end date: 2016-09-01	
<p>Note that the dates requested on the “IRGA” form refer to the start and end of <i>data collection</i>. These are <i>not</i> the same as the start and end dates of the study, above, for which approval is sought. (A study may be considered to end when its final report is submitted.)</p> <p>Note that ethics approval must be obtained before the expected study start date as given above; retrospective approval cannot be given.</p> <p>Note that failure to follow the University’s policy on Ethics may lead to disciplinary action concerning Misconduct or a breach of Academic Integrity.</p> <p>By submitting this application, the investigator(s) undertake to:</p> <ul style="list-style-type: none"> Conduct the study in accordance with University policies governing: Ethics (http://www.southampton.ac.uk/ris/policies/ethics.html); Data management (http://www.southampton.ac.uk/library/research/researchdata/); Health and Safety (http://www.southampton.ac.uk/healthandsafety/); Academic Integrity (http://www.calendar.soton.ac.uk/sectionIV/academic-integrity-statement.html). Ensure the study Reference number ERGO/FPSE/xxxx is prominently displayed on all advertising and study materials, and is reported on all media and in all publications; Conduct the study in accordance with the information provided in the application, its appendices, and any other documents submitted; Submit the study for re-review (as an amendment through ERGO) or seek FPSE EC advice if any changes, circumstances, or outcomes materially affect the study or the information given; Promptly advise an appropriate authority (Research Governance Office) of any adverse study outcomes (via an adverse event notification through ERGO); Submit an end-of-study form if required to do so. 		

REFER TO THE INSTRUCTIONS AND GUIDE DOCUMENTS WHEN COMPLETING THIS FORM AND THE TEMPLATES DOCUMENT WHEN PREPARING THE REQUIRED APPENDICES.

PRE-STUDY

Characterise the proposed participants

Members of the UK public who discuss events, specifically those who use social media, and specifically those who use Twitter as this social medium. Neither their race, gender nor age is a factor. They will have no relationship to the investigator.

Describe how participants will be approached

The participants will not be approached directly. They will be chosen based on the content of their tweets. They will be discoverable by their Twitter presence.

Describe how inclusion and/or exclusion criteria will be applied (if any)

There will be no direct inclusion or exclusion of the participants. Their tweet messages will either be accepted or discarded based on their content. The criteria will include indicative markers, such as "there's a group of people running in the park", or "a new shop opened down my road" along with accompanying geo-referenced data.

Describe how participants will decide whether to take part

The participants, along with their tweet text and geo-reference, have already contributed their information within the public domain.

Participant Information (Appendix (i))

Provide the **Participant Information** in the form that it will be given to **participants** as Appendix (i). All studies must provide **participant information**.

Consent Form/Information (Appendix (iii))

FPSE EC Application Form

Provide the **Consent Form** (or the request for consent) in the form that it will be given to **participants** as Appendix (iii). All studies must obtain **participant** consent. Some studies may obtain verbal consent (and only present consent information), other studies will require written consent, as explained in the *Instructions, Guide, and Templates* documents.

DURING THE STUDY

Describe the study procedures as they will be experienced by the **participant**

The study aims to understand how event data, such as geolocation and description, can be detected and tracked through Twitter. All data is actively obtained through the public API and thus the participant will not experience any part of the study.

Identify how, when, where, and what kind of data will be recorded (not just the formal research data, but including all other study data such as e-mail addresses and signed consent forms)

The full content of the tweet will be recorded. This includes username, handle, tweet timestamp, tweet text, retweet count, language, time zone, follower count and geo-reference. This also complies with the Twitter terms and conditions. The data will be stored on a secure drive within the University (OneDrive).

Participant questionnaire/data gathering methods (Appendix (ii))

As Appendix (ii), reproduce any and all **participant** questionnaires or data gathering instruments in the exact forms that they will be given to or experienced by **participants**. If conducting less formal data collection, or data collection that does not involve direct questioning or observation of participants (eg secondary data or “big data”), provide specific information concerning the methods that will be used to obtain the data of the study.

POST-STUDY

Identify how, when, and where data will be stored, processed, and destroyed

FPSE EC Application Form

If Study Characteristic M.1 applies, provide this information in the **DPA Plan** as Appendix (iv) instead and do *not* provide explanation or information on this matter here.

STUDY CHARACTERISTICS

(L.1) The study is funded by a commercial organisation: **Yes** (delete one)

If 'Yes', provide details of the funder or funding agency *here*.

The Ordnance Survey sponsors the PhD as a whole, but not this individual study directly.

(L.2) There are **restrictions** upon the study: **No** (delete one)

If 'Yes', explain the nature and necessity of the **restrictions** *here*.

(L.3) Access to **participants** is through a third party: **Yes** (delete one)

If 'Yes', provide evidence of your permission to contact them as Appendix (v). Do *not* provide explanation or information on this matter here.

(M.1) **Personal data** is or *may be collected or processed: **Yes** (delete one)

Data will be processed outside the UK: **No** (delete one)

If 'Yes' to either question, provide the **DPA Plan** as Appendix (iv). Do *not* provide information or explanation on this matter here. Note that using or recording e-mail addresses, telephone numbers, signed consent forms, or similar study-related **personal data** requires M.1 to be "Yes".

(* Secondary data / "big data" may be *de*-anonymised, or may contain **personal data**. If so, answer 'Yes'.)

(M.2) There is **inducement** to **participants**: **No** (delete one)

If 'Yes', explain the nature and necessity of the inducement *here*.

(M.3) The study is **intrusive**: **No** (delete one)

If 'Yes', provide the **Risk Management Plan**, the **Debrief Plan**, and Technical Details as Appendices (vi), (vii), and (ix), and explain *here* the nature and necessity of the intrusion(s).

(M.4) There is **risk of harm** during the study: **No** (delete one)

If 'Yes', provide the **Risk Management Plan**, the **Contact Information**, the **Debrief Plan**, and Technical Details as Appendices (vi), (vii), (viii), and (ix), and explain *here* the necessity of the risks.

FPSE EC Application Form

(M.5) The true purpose of the study will be hidden from **participants**: **No** (delete one)
 The study involves **deception of participants**: **No** (delete one)
 If 'Yes' to either question, provide the **Debrief Plan** and Technical Details as Appendices (vii) and (ix), and explain *here* the necessity of the deception.

(M.6) **Participants** may be minors or otherwise have **diminished capacity**: **No** (delete one)
 If 'Yes', AND if one or more Study Characteristics in categories M or H applies, provide the **Risk Management Plan**, the **Contact Information**, and Technical Details as Appendices (vi), (vii), & (ix), and explain *here* the special arrangements that will ensure informed consent.

(M.7) **Sensitive data** is collected or processed: **No** (delete one)
 If 'Yes', provide the **DPA Plan** and Technical Details as Appendices (iv) and (ix). Do *not* provide explanation or information on this matter here.

(H.1) The study involves: **invasive** equipment, material(s), or process(es); or **participants** who are not able to withdraw at any time and for any reason; or animals; or human tissue; or biological samples: **No** (delete one)
 If 'Yes', provide Technical Details and further justifications as Appendices (ix) and (x). Do *not* provide explanation or information on these matters here. Note that the study will require separate approval by the Research Governance Office.

Technical details

If one or more Study Characteristics in categories M.3 to M.7 or H applies, provide the description of the technical details of the experimental or study design, the power calculation(s) which yield the required sample size(s), and how the data will be analysed, as separate appendices.

APPENDICES (AS REQUIRED)

While it is *preferred* that this information is included here in the application form, it may be provided as separate document files. If provided separately, *name the files precisely* as "Participant Information", "Questionnaire", "Consent Form", "DPA Plan", "Permission to contact", "Risk Management Plan", "Debrief Plan", "Contact Information", and/or "Technical details" as appropriate. Each appendix or document must specify the reference number in the form ERGO/FPSE/xxxx, the document version number, and its date of last edit.

Appendix (i): **Participant Information** in the form that it will be given to **participants**.

Appendix (ii): Data collection method (eg for secondary data or "big data") / **Participant Questionnaire** in the form that it will be given to **participants**.

Appendix (iii): **Consent Form** (or consent information if no **personal data** is collected) in the form that it will be given to **participants**.

Appendix (iv): **DPA Plan**.

Appendix (v): Evidence of permission to contact (prospective) **participants** through any third party.

Appendix (vi): **Risk Management Plan.**

Appendix (vii): **Debrief Plan.**

Appendix (viii): **Contact Information.**

Appendix (ix): Technical details of the experimental or study design, the power calculation(s) for the required sample size(s), and how the data will be analysed.

Appendix (x): Further details and justifications in the case of: **invasive** equipment, material(s), or process(es); **participants** who are not able to withdraw at any time and for any reason; animals; human tissue; or biological samples.

Appendix (iv) DPA Plan

DPA Plan

Ethics reference number: ERGO/FPSE/20230	Version: 1	Date: 2016-04-20
Study Title: Towards a classification of event types		
Investigator: Nicholas Bennett		

The following is an exhaustive and complete list of all the data that will be collected (through questionnaires, interviews, extraction from records, etc)

1. Twitter username, handle, tweet location, time, textual content, time zone, retweet count, follower count, language.

The data is relevant to the study purposes because as the study aims to detect and track event data through Twitter, the tweet content will inform the investigator about events in the participant's location. The data is adequate because it contains all the required information to extract a location and event description, and the data is not excessive because the participant's username is only their real name if they choose to publicise it, and their gender, age, ethnicity and other personal details are not part of the investigation. The username shall be anonymised before publication.

The data will be processed fairly because the participants intentionally made the data public.

The data's accuracy is ensured because of subsequent algorithmic processing. The data may not be accurate as it is voluntarily created; if this is discovered it will be discarded.

Data will be stored on the University's OneDrive server(s). The data will be held in accordance with University policy on data retention.

Data files will be protected by the University guidelines; laptops will be protected by the investigator; desktops will not be used; there will be no physical data.

The data will be destroyed by the investigator at the end of the study through deleting the non-aggregated data.

The data will be processed in accordance with the rights of the participants because they will have the right to access, correct, and/or withdraw their data at any time and for any reason. Participants will be able to exercise their rights by contacting the investigator (e-mail: n.bennett@soton.ac.uk) or the project supervisor (e-mail: dem@ecs.soton.ac.uk).

The data will be anonymised by removing usernames/handles. No consent form required.

No data will be transferred outside the European Economic Area (EEA).

Appendix (v) Access to Participants

Through the Twitter API, a public platform where the participants have already posted their data.

Ethics Application Form for SECONDARY DATA ANALYSIS

Please consult the guidance at the end of this form before completing and submitting your application.

1. **Name(s):** Nicholas Bennett
2. **Current Position:** PhD Researcher
3. **Contact Details:**
Division: ECS
Email n.bennett@soton.ac.uk
Phone
4. **Is your research being conducted as part of an education qualification?**
 Yes ☒ No ☐
5. **If Yes, please give the name of your supervisor**
 Dr David Millard, Professor David Martin
6. **Title of your research project / study:**
 Data Mining Narratives of Social Space
7. **Briefly describe the rationale, aims, design and research questions of your research**
Please indicate clearly whether you are applying for ethics approval for a specific piece of research, or for overarching ethics approval to use certain datasets for a range of research activities. Approval for the latter will only cover the datasets specified here, for a maximum of 3 years and then subject to renewal.
 The popularity of social media has created a wealth of information useful to academia, business and the wider world. The aim is to detect and track events through Twitter to contribute towards the Ordnance Survey's understanding of how social events are created and discussed as well as how the land use changes based on the type of event. The research question is: how can event detection and tracking of events through Twitter create an understanding of the social use of space? The aggregated data will be used for the overall purpose of the PhD but this specific experiment will run for a 5-month period.
8. **Describe the data you wish to analyse**
Please give details of the title of the dataset, nature of data subjects (e.g. individuals or organisations), thematic focus and country/countries covered. Indicate whether the data are qualitative or quantitative, survey data, administrative data or other types of data. Identify the source from where you will be obtaining the data (including a web address where appropriate).
 The data is public data from the Twitter API (<http://www.twitter.com>) including usernames, tweet content, location and time zone. The data will focus on geo-tagged tweets (either directly or by mentioning place names), and are quantitative by nature but will involve natural language processing and some qualitative analyses.

9. **What are the terms and conditions around the use of the data? Did data subjects give consent for their data to be re-used? If not, on what basis is re-use of the data justified?**
Please state what (if any) conditions the data archive imposes (e.g. registration, signing of confidentiality agreement, specific training etc.). In many cases the data controller will have given explicit permission for data re-use. Please explain how you justify the use of data if approval and consents for the original data collection and re-use are not in place. This may be the case where, for example, the original data collection predated requirements for ethics review or occurred in a jurisdiction where explicit consent and approval are not required.
 The participants have not directly consented but published their data on an open, public platform. The data archive does not impose any conditions.
10. **Do you intend to use personal data** (https://ico.org.uk/media/1549/determining_what_is_personal_data_quick_reference_guide.pdf) **or sensitive personal data** (<http://www.legislation.gov.uk/ukpga/1998/29/section/2>) **as defined by the Data Protection Act (even if the data are publicly available)?**
 Yes ☒ No ☐
 If YES, please specify what personal data will be included and why.
 Personal data to be included are usernames and location, both of which will be aggregated and thus anonymised. These data are key to understanding how events are discussed, such as where they are and which groups are affected by them.
11. **Do you intend to link two or more datasets?**
Data linkage refers to merging of information from two or more sources of data to consolidate facts concerning an individual or an event that are not available in any separate record. Please note that for the purposes of research ethics we are not interested in the merging of different waves of a particular survey, or the merging of data from different countries for the same survey.
 Yes ☐ No ☒
 If YES, please give details of which datasets will be linked and for what purposes.
12. **How will you store and manage the data before and during the analysis? What will happen with the data at the end of the project?**
Please consult the University of Southampton's Research Data Management Policy (<http://library.soton.ac.uk/researchdata/storage> and <http://www.calendar.soton.ac.uk/sectionIV/research-data-management.html>), and indicate how you will abide by it.
 Data will be stored on the University's secure data servers and on OneDrive. Upon obtaining the data, it will be automatically stored on the OneDrive account. During the study, the usernames will not be anonymised to aid in tracking conversations about events. After the analytical period and before any publication or presentation, the usernames will be anonymised and the location data aggregated. The unprocessed data will be destroyed at the end of the research period.
13. **How will you minimise the risk that data subjects (individuals or organisations) could be identified in your presentation of results?**
Please consider whether disclosive ID codes have been used (e.g. date of birth) and whether it is theoretically possible to identify individuals by combining characteristics (e.g. widow in

FPSE EC Application Form

Hampshire with 14 children) or by combining datasets. How will you protect individuals' anonymity in your analysis and dissemination?

The data obtained from Twitter will include usernames. Before any public viewing of the data, the usernames will be anonymised. No datasets will be combined and any tweet text published will be reworded.

14. What other ethical risks are raised by your research, and how do you intend to manage these?

Issues may arise due to the nature of the research you intend to undertake and/or the subject matter of the data. Examples include: data or analysis that are culturally or socially sensitive; data relating to criminal activity, including terrorism, and security sensitive issues.

To mitigate against these risks, keywords will be used to obtain the data, therefore a strong relationship between the text and the keyword is assumed, thus the risk of obtaining sensitive material is minimised as the controller is only searching for public event information. Furthermore, as per Twitter's terms and conditions, many tweets relating to untoward topics are automatically removed.

15. Please outline any other information that you feel may be relevant to this submission.

For example, will you be using the services or facilities of ONS, ADNR, or HSCIC and/or are you obtaining ethical review from NRES (through IRAS) or other? Please confirm whether the data being used are already in the public domain.

The data are already in the public domain.

16. Please indicate if you, your supervisor or a member of the study team/research group are a data controller and/or data processor in relation to the data you intend to use as defined by the Data Protection Act, and confirm that you/they understand your/their respective responsibilities <https://ico.org.uk/for-organisations/guide-to-data-protection/key-definitions/>.

No additional supervisor shall partake in the study. I understand my responsibility.

Note: This Ethics Application Form is currently being piloted. If you have any comments on any of the questions, it would be helpful if you could email them to rgoinfo@soton.ac.uk with "Secondary Data Analysis Form" in the subject line.

Guidance on applying for ethics approval for secondary data analysis

If your research PURELY involves the following, you do not need to apply for ethics approval:

- analysis of **aggregated** data on individuals or organisations (e.g. GDP, labour force participation rates, fertility rates);
- meta-analyses (i.e. the analysis of studies);
- literature reviews or reviews/analyses of reports, policies, documents, meeting minutes, newspaper articles, films.

Filling in the online IRGA Form:

- Please answer the questions about dates of 'data collection' to refer to the dates of your proposed study.
- Please answer NO to the question 'Will your study involve humans?' UNLESS you are applying for a mixed method study which also includes a data collection component.

Additional Forms:

If your study PURELY involves secondary analysis of data, you only need to fill in the 'Ethics Application Form for Secondary Data Analysis'. You do not need a Risk Assessment Form.

If your study is a mixed-method study involving secondary data analysis AND some component of data collection (e.g. interviews, online survey), or the analysis of non-anonymised data (e.g. social media data), then you need to fill in additional forms:

- Ethics Application Form (for studies other than secondary data analysis)
- Risk Assessment Form
- Participant Information Sheet
- Consent Form
- Draft research instrument

Please note:

- You must not begin data analysis until ethical approval has been obtained.
- It is your responsibility to follow the University of Southampton's Ethics Policy and any relevant academic or professional guidelines in the conduct of your research. This includes ensuring confidentiality in the storage and use of data.
- It is your responsibility to provide full and accurate information in completing this form.

Appendix B

Appendix B - Treasury Interview Ethics Application

FEPS Ethics Committee
FEPS Ethics Application Form Ver 1

Refer to the *Instructions* and to the *Guide* documents for a glossary of the key phrases in **bold** and for an explanation of the information required in each section. The *Templates* document provides some text that may be helpful in preparing some of the required appendices.

Replace the **highlighted text** with the appropriate information.

Note that the size of the text entry boxes provided on this form does **not** indicate the expected amount of information; instead, refer to the *Instructions* and to the *Guide* documents in providing the complete information required in each section. Do **not** duplicate information from one text box to another. Do not otherwise edit this form.

Reference number: ERGO/FPSE/45418	Submission version: 1	Date: 2019-02-19
Name of investigator(s) : Nicholas Christopher Bennett		
Name of supervisor(s) (if student investigator(s)): David E. Millard		
Title of study: Evaluative Interview with Members of HM Treasury		
Expected study start date: 06-03-19	Expected study end date: 31-10-19	
<p>Note that the dates requested on the "IRGA" form refer to the start and end of <i>data collection</i>. These are <i>not</i> the same as the start and end dates of the study, above, for which approval is sought. (A study may be considered to end when its final report is submitted.)</p> <p>Note that ethics approval must be obtained before the expected study start date as given above; retrospective approval cannot be given.</p> <p>Note that failure to follow the University's policy on Ethics may lead to disciplinary action concerning Misconduct or a breach of Academic Integrity.</p> <p>By submitting this application, the investigator(s) undertake to:</p> <ul style="list-style-type: none"> Conduct the study in accordance with University policies governing: Ethics (http://www.southampton.ac.uk/ris/policies/ethics.html); Data management (http://www.southampton.ac.uk/library/research/researchdata/); Health and Safety (http://www.southampton.ac.uk/healthandsafety); Academic Integrity (http://www.calendar.soton.ac.uk/sectionIV/academic-integrity-statement.html). Ensure the study Reference number ERGO/FPSE/xxxx is prominently displayed on all advertising and study materials, and is reported on all media and in all publications; Conduct the study in accordance with the information provided in the application, its appendices, and any other documents submitted; Submit the study for re-review (as an amendment through ERGO) or seek FPSE EC advice if any changes, circumstances, or outcomes materially affect the study or the information given; Promptly advise an appropriate authority (Research Governance Office) of any adverse study outcomes (via an adverse event notification through ERGO); Submit an end-of-study form if required to do so. 		

08082018

FPSE EC Application Form v1

REFER TO THE INSTRUCTIONS AND GUIDE DOCUMENTS WHEN COMPLETING THIS FORM AND THE TEMPLATES DOCUMENT WHEN PREPARING THE REQUIRED APPENDICES.

STUDY DETAILS

What are the aims and objectives of this study?

To evaluate results produced from tagging tweets with Treasury and non-Treasury supplied keywords.

Background of the study (*a brief rationale for conducting the study*)

This type of experiment has not been done in the past, therefore evaluating the results is extremely difficult. The members of HM Treasury and interested parties will form part of a panel of experts to review the results and offer any form of improvement.

Key research question (*Specify hypothesis if applicable*)

Hypothesis: the tweets tagged with Treasury terms will be economy-related and useful

Null hypothesis: the tweets tagged with Treasury terms will not be economy-related and useful

RQ: To what extent do these tagged tweets reflect expected or non-expected behaviours of public interest towards the economy.

Study design (*Give a brief outline of the study design and why it is being used*)

Interviewing and quizzing the panel of experts will impart their precise knowledge onto this experiment and conclude whether it has been useful. I will ask them questions based on the results of my experiment and inquire as to whether HM Treasury and interested parties will find these results useful. I will ask them how accurate these results seem compared to their existing methods or expected behaviours, and whether any improvements upon the methodology would improve the overall results.

PRE-STUDY

Characterise the proposed participants

A panel of experts who work at HM Treasury and at other interested parties, such as Office for National Statistics and Ordnance Survey.

Describe how participants will be approached

If any e-mail lists are used, including FEPS distribution lists, justify their use *here*

I have met several of them already during workshops and consultations. I will use known email addresses of these people to contact them and arrange interview times.

FPSE EC Application Form v1

Describe how inclusion / exclusion criteria will be applied (if any)

They must be members of HM Treasury working in Innovation for the Treasury participants, and data scientists, economists or government liaisons from other interested parties.

Describe how **participants** will decide whether or not to take part

They will accept or deny the interview request with full knowledge of what accepting it would entail.

Participant Information (Appendix (i))

Provide the **Participant Information** in the form that it will be given to **participants** as Appendix (i). All studies must provide **participant information**.

Consent Form/Information (Appendix (iii))

Provide the **Consent Form** (or the request for consent) in the form that it will be given to **participants** as Appendix (iii). All studies must obtain **participant** consent. Some studies may obtain verbal consent (and only present consent information), other studies will require written consent, as explained in the *Instructions*, *Guide*, and *Templates* documents.

DURING THE STUDYDescribe the study procedures as they will be experienced by the **participants**

First they will be approached by email and asked to be interviewed. In the same email will be a sample of the questions I will ask them. Once they have accepted and a time is arranged, I will ask them further questions about the results and whether they find it useful or have other comments. I will record the interviews, with their consent, and email them after the experiment has concluded to inform them of the final result.

Identify how, when, where, and what kind of data will be recorded (not just the formal research data, but including all other study data such as e-mail addresses and signed consent forms)

How: email already known, recorded using a recording device.
 When: sometime between February and September 2019 based on availability.
 Where: HM Treasury, London, ONS campus, Ordnance Survey, online.
 What: face-to-face interviews recorded (audio only).

Participant questionnaire/data gathering methods (Appendix (ii))

As Appendix (ii), reproduce any and all **participant** questionnaires or data gathering instruments in the exact forms that they will be given to or experienced by **participants**. If conducting less formal data collection, or data collection that does not involve direct questioning or observation of participants (eg secondary data or “big data”), provide specific information concerning the methods that will be used to obtain the data of the study.

POST-STUDY

Identify how, when, and where data will be stored, processed, and destroyed

If the Study Characteristic M.1 applies, provide this information in the **DPA Plan** as Appendix (iv) instead and do *not* provide explanation or information on this matter here

The data will be stored locally as a .wav file containing the audio recording of the interview. The data will be transcribed into a text document and the .wav file deleted. The transcription will be uploaded alongside the thesis. The consent forms will be kept secure alongside the password-protected laptop and kept in accordance with the University policies.

STUDY CHARACTERISTICS

(L.1) The study is funded by a commercial organisation: **Yes** (delete one)

If ‘Yes’, provide details of the funder or funding agency *here*.

Ordnance survey – sponsors the entire PhD.

(L.2) There are **restrictions** upon the study: **-No** (delete one)

If ‘Yes’, explain the nature and necessity of the **restrictions** *here*.

(L.3) Access to **participants** is through a third party: **No** (delete one)

If ‘Yes’, provide evidence of your permission to contact them as Appendix (v). Do *not* provide explanation or information on this matter here.

(M.1) **Personal data** is or *may be collected or processed: **Yes** (delete one)

Data will be processed outside the UK: **-No** (delete one)

If ‘Yes’ to either question, provide the **DPA Plan** as Appendix (iv). Do *not* provide information or explanation on this matter here. Note that using or recording e-mail addresses, telephone numbers, signed consent forms, or similar study-related **personal data** requires M.1 to be “Yes”.

(* Secondary data / “big data” may be *de*-anonymised, or may contain **personal data**. If so, answer ‘Yes’.)

(M.2) There is **inducement** to **participants**: **No** (delete one)

If ‘Yes’, explain the nature and necessity of the inducement *here*.

FPSE EC Application Form v1

(M.3) The study is **intrusive**: **No** (delete one)

If 'Yes', provide the **Risk Management Plan**, the **Debrief Plan**, and Technical Details as Appendices (vi), (vii), and (ix), and explain *here* the nature and necessity of the intrusion(s).

(M.4) There is **risk of harm** during the study: **No** (delete one)

If 'Yes', provide the **Risk Management Plan**, the **Contact Information**, the **Debrief Plan**, and Technical Details as Appendices (vi), (vii), (viii), and (ix), and explain *here* the necessity of the risks.

(M.5) The true purpose of the study will be hidden from **participants**: **No** (delete one)

The study involves **deception of participants**: **No** (delete one)

If 'Yes' to either question, provide the **Debrief Plan** and Technical Details as Appendices (vii) and (ix), and explain *here* the necessity of the deception.

(M.6) **Participants** may be minors or otherwise have **diminished capacity**: **No** (delete one)

If 'Yes', AND if one or more Study Characteristics in categories M or H applies, provide the **Risk Management Plan**, the **Contact Information**, and Technical Details as Appendices (vi), (vii), & (ix), and explain *here* the special arrangements that will ensure informed consent.

(M.7) **Sensitive data** is collected or processed: **No** (delete one)

If 'Yes', provide the **DPA Plan** and Technical Details as Appendices (iv) and (ix). Do *not* provide explanation or information on this matter here.

(H.1) The study involves: **invasive** equipment, material(s), or process(es); or **participants** who are not able to withdraw at any time and for any reason; or animals; or human tissue; or biological samples: **No** (delete one)

If 'Yes', provide Technical Details and further justifications as Appendices (ix) and (x). Do *not* provide explanation or information on these matters here. Note that the study will require separate approval by the Research Governance Office.

Technical details

If one or more Study Characteristics in categories M.3 to M.7 or H applies, provide the description of the technical details of the experimental or study design, the power calculation(s) which yield the required sample size(s), and how the data will be analysed, as separate appendices.

APPENDICES (AS REQUIRED)

While it is *preferred* that this information is included here in the application form, it may be provided as separate document files. If provided separately, *name the files precisely* as “Participant Information”, “Questionnaire”, “Consent Form”, “DPA Plan”, “Permission to contact”, “Risk Management Plan”, “Debrief Plan”, “Contact Information”, and/or “Technical details” as appropriate. Each appendix or document must specify the reference number in the form ERGO/FPSE/xxxx, the document version number, and its date of last edit.

Appendix (i): **Participant Information** in the form that it will be given to **participants**.

Appendix (ii): Data collection method (eg for secondary data or “big data”) / **Participant Questionnaire** in the form that it will be given to **participants**.

Appendix (iii): **Consent Form** (or consent information if no **personal data** is collected) in the form that it will be given to **participants**.

Appendix (iv): **DPA Plan**.

Appendix (v): Evidence of permission to contact (prospective) **participants** through any third party.

Appendix (vi): **Risk Management Plan**.

Appendix (vii): **Debrief Plan**.

Appendix (viii): **Contact Information**.

Appendix (ix): Technical details of the experimental or study design, the power calculation(s) for the required sample size(s), and how the data will be analysed.

Appendix (x): Further details and justifications in the case of: **invasive** equipment, material(s), or process(es); **participants** who are not able to withdraw at any time and for any reason; animals; human tissue; or biological samples.

Participant Information Sheet

Study Title: Evaluation of Economy-Tagged Tweets

Researcher: Nicholas Christopher Bennett

ERGO number: 45418

You are being invited to take part in the above research study. To help you decide whether you would like to take part or not, it is important that you understand why the research is being done and what it will involve. Please read the information below carefully and ask questions if anything is not clear or you would like more information before you decide to take part in this research. You may like to discuss it with others but it is up to you to decide whether or not to take part. If you are happy to participate you will be asked to sign a consent form.

What is the research about?

My name is Nick Bennett, a third-year PhD researcher at the University of Southampton sponsored by Ordnance Survey. I have collected a large quantity of tweets and, with HM Treasury, have tagged them as relating to the economy. I am interested in your opinion on these results and whether you have any further comments that would improve the experiment.

Why have I been asked to participate?

You are working in HM Treasury and have an interest in new ways of collecting economic data.

What will happen to me if I take part?

You will take part in an interview, either as a group or individually depending on availability, and shown results from my experiment on which to comment. Your responses will be recorded on an audio device and later transcribed with the audio recording being deleted. None of your personal data will be part of this study and nothing about you will be published aside from acknowledging you work at HM Treasury.

Are there any benefits in my taking part?

You will be contributing towards a PhD thesis studying real-time economic indicators through Twitter, taking the role of an expert in the field. Your feedback will help to evaluate the results.

Are there any risks involved?

There are no risks. The interview will be in your workplace and the only equipment in the room will be a laptop and an audio recording device.

What data will be collected?

Audio recordings of the discussion will be collected using an audio recording device, by me, and later transcribed. No personal information will be collected and the only reference to you will be as part of a panel of experts. The collection will take place in a closed room, analysed on my local laptop and the recording deleted once transcribed. The only personal data collected will be the consent forms.

Will my participation be confidential?

08082018

Your participation and the information we collect about you during the course of the research will be kept strictly confidential.

Only members of the research team and responsible members of the University of Southampton may be given access to data about you for monitoring purposes and/or to carry out an audit of the study to ensure that the research is complying with applicable regulations. Individuals from regulatory authorities (people who check that we are carrying out the study correctly) may require access to your data. All of these people have a duty to keep your information, as a research participant, strictly confidential.

In keeping with previous sections, the only data collected will be an audio file. This file will be transcribed and deleted. All transcription will be done by me, using headphones, on my private laptop. Consent forms are required, but this information does not feature within any of the transcriptions.

Data Protection Privacy Notice

The University of Southampton conducts research to the highest standards of research integrity. As a publicly-funded organisation, the University has to ensure that it is in the public interest when we use personally-identifiable information about people who have agreed to take part in research. This means that when you agree to take part in a research study, we will use information about you in the ways needed, and for the purposes specified, to conduct and complete the research project. Under data protection law, 'Personal data' means any information that relates to and is capable of identifying a living individual. The University's data protection policy governing the use of personal data by the University can be found on its website (<https://www.southampton.ac.uk/legalservices/what-we-do/data-protection-and-foi.page>).

This Participant Information Sheet tells you what data will be collected for this project and whether this includes any personal data. Please ask the research team if you have any questions or are unclear what data is being collected about you.

Our privacy notice for research participants provides more information on how the University of Southampton collects and uses your personal data when you take part in one of our research projects and can be found at <http://www.southampton.ac.uk/assets/sharepoint/intranet/Is/Public/Research%20and%20Integrity%20Privacy%20Notice/Privacy%20Notice%20for%20Research%20Participants.pdf>

Any personal data we collect in this study will be used only for the purposes of carrying out our research and will be handled according to the University's policies in line with data protection law. If any personal data is used from which you can be identified directly, it will not be disclosed to anyone else without your consent unless the University of Southampton is required by law to disclose it.

Data protection law requires us to have a valid legal reason ('lawful basis') to process and use your Personal data. The lawful basis for processing personal information in this research study is for the performance of a task carried out in the public interest. Personal data collected for research will not be used for any other purpose.

For the purposes of data protection law, the University of Southampton is the 'Data Controller' for this study, which means that we are responsible for looking after your information and using it properly. The University of Southampton will keep identifiable information about you for 0 years after the study has finished after which time any link between you and your information will be removed.

To safeguard your rights, we will use the minimum personal data necessary to achieve our research study objectives. Your data protection rights – such as to access, change, or transfer such information – may be limited, however, in order for the research output to be

reliable and accurate. The University will not do anything with your personal data that you would not reasonably expect.

If you have any questions about how your personal data is used, or wish to exercise any of your rights, please consult the University's data protection webpage (<https://www.southampton.ac.uk/legalservices/what-we-do/data-protection-and-foi.page>) where you can make a request using our online form. If you need further assistance, please contact the University's Data Protection Officer (data.protection@soton.ac.uk).

Do I have to take part?

No, it is entirely up to you to decide whether or not to take part. If you decide you want to take part, you will need to sign a consent form to show you have agreed to take part.

What happens if I change my mind?

You have the right to change your mind and withdraw at any time without giving a reason and without your participant rights being affected. If you wish to withdraw from the study, contact n.bennett@soton.ac.uk

What will happen to the results of the research?

Your personal details will remain strictly confidential. Research findings made available in any reports or publications will not include information that can directly identify you without your specific consent. The results of the study will be kept by the University as part of the thesis submission process. The aggregated evaluation may be published in future academic conference or journal papers. As no personal data is retained, nothing aside from the evaluation will be published.

Where can I get more information?

For more information, email n.bennett@soton.ac.uk

What happens if there is a problem?

If you have a concern about any aspect of this study, you should speak to the researchers who will do their best to answer your questions.

If you remain unhappy or have a complaint about any aspect of this study, please contact the University of Southampton Research Integrity and Governance Manager (023 8059 5058, rgoinfo@soton.ac.uk).

The lead researcher is Nick Bennett (n.bennett@soton.ac.uk) supervised by David Millard (dem@ecs.soton.ac.uk)

Thank you for reading this information sheet and participating in the study.

Appendix iii – ERGO/FPSE(FEPS)/45418

CONSENT FORM**Study title:** Evaluation of Economy-Tagged Tweets**Researcher name:** Nicholas Christopher Bennett**ERGO number:** 45418

Participant Identification Number (if applicable):


Please initial the box(es) if you agree with the statement(s):

I have read and understood the information sheet Participant Information and have had the opportunity to ask questions about the study.	
I agree to take part in this research project and agree for my data to be used for the purpose of this study.	
I understand my participation is voluntary and I may withdraw (at any time) for any reason without my participation rights being affected.	
I understand the only data used will be a transcription of my voice, no other data, aside from "Works at the Treasury", will be used.	

Name of participant (print name).....

Signature of participant.....

Date.....

Name of researcher (print name)...Nicholas Christopher
Bennett.....Signature of researcher Date...15th August 2018.....

08082018

DPA Plan

Ethics reference number: ERGO/FEPS/45418	Version: 1	Date: 2019-02-19
Study Title: Evaluative Interview with Members of HM Treasury and Interested Parties		
Investigator: Nicholas Christopher Bennett		

The following is an exhaustive and complete list of all the data that will be collected (through questionnaires, interviews, extraction from records, etc)

1. Consent forms
2. Audio recording of interviews
3. Answers to the questions in the Questionnaire/Survey appendix

The data is relevant to the study purposes because the interviews will help to evaluate the results of the thesis. The data is adequate because it comes from a panel of experts, and the data is not excessive because it is only dealing with their opinions of the results.

The data will be processed fairly because the participants will have given explicit consent.

The data's accuracy is ensured because it comes from the experts themselves.

Data will be stored on the Investigator's laptop. The data will be held in accordance with University policy on data retention.

Data files will be protected by being stored locally on a password-protected laptop; laptops will be protected by password protecting them and storing them securely at the university inside a locked drawer.

The data will be destroyed once the interviews have been transcribed and by the end of the study.

The data will be processed in accordance with the rights of the participants because they will have the right to access, correct, and/or withdraw their data at any time and for any reason. Participants will be able to exercise their rights by contacting the investigator (e-mail: n.bennett@soton.ac.uk) or the project supervisor (e-mail: dem@ecs.soton.ac.uk).

The data will be anonymised by not including any identifying names within the transcription.

Consent forms will be linked to the data by name, then anonymised.

Data Protection Act 2018 (DPA) best practice

If the study involves personal or sensitive data, explicitly explain how data will be collected, stored, analysed, held securely, and in turn destroyed. The DPA does not apply to anonymous data and a DPA Plan is not required in the case of such data.

The principles of the DPA are that personal data must be:

1. Processed fairly and lawfully.
2. Processed for specified purposes and in an appropriate way.
3. Adequate, relevant and not excessive for the purposes.
4. Accurate and up-to-date.
5. Not kept for longer than necessary.
6. Processed in accordance with the rights of data subjects (participants).
7. Protected by appropriate security, both practical and organisational.
8. Not transferred outside the European Economic Area (EEA) without adequate data protection controls.

Data is recorded information, whether stored electronically on computer or in paper-based filing systems. Personal data is information about an identifiable living individual. It can be factual, such as the date of a person's interview, or an opinion, such as someone's view on how the person has performed on a task. It obviously includes individuals' contact addresses or telephone numbers. (Less obviously, note that personal data is being processed where information is collected and analysed with the intention of distinguishing one individual from another and to take a particular action in respect of an individual. This can take place even if no obvious identifiers, such as names or addresses, are held.) Processing is any activity that involves data, including collecting, recording or retrieving, using, disclosing, organising, adapting, changing, updating, or destroying it.

The DPA identifies Sensitive Personal Data as:

- a) the racial or ethnic origin of the data subject (participant);
- b) his political opinions;
- c) his religious beliefs or other beliefs of a similar nature;
- d) whether he is a member of a trade union;
- e) his physical or mental health or condition;
- f) his sexual life;
- g) the commission or alleged commission by him of any offence or
- h) any proceedings for any offence committed or alleged to have been committed by him, the disposal of such proceedings and the sentence of court in such proceedings.

The processing of sensitive data must meet at least one of the 10 stricter conditions laid down in Schedule 3 of the DPA. It may be useful to know that condition 1 of this schedule permits processing of such data if the data subject has given his explicit consent, and condition 5 if the information has been made public as a result of steps deliberately taken by the data subject.

Keep in mind that the Police have a right of access to personal data held by the study for the purpose of safeguarding national security; preventing or detecting crime; prosecuting or apprehending offenders; assessing or collecting tax; or protecting the vital interests of the data subject or another.

Researchers are exempted: from the second data protection principle, meaning that personal data can be processed for purposes other than for which they were originally obtained; from the fifth data protection principle, meaning that personal data can be held indefinitely; and from the data subject's right of access to his personal data provided the data is processed for research purposes and the results do not identify data subjects. In addition, the Data Protection (Processing of Sensitive Personal Data) Order 2000 para.9 provides that processing in the course of maintaining archives for research purposes is permissible where the sensitive personal data are not used to take decisions about any person without their consent and no substantial damage or distress is caused to any person by the keeping of those data. These exemptions do NOT give a blanket exemption from all the Data Protection Principles to data provided and/or used for research purposes.

FPSE EC Application Form v1

Researchers wishing to use personal data should be aware that the Data Protection Principles still generally apply, notably the requirement to keep data secure¹.

A study may seek to anonymise the data it keeps. Anonymisation involves the removal of participants' personal information (names; e-mail address; whatever data it is that might permit identification; etc) from the data such that what remains cannot be used to identify them. Note that audio and video recordings (and often transcriptions too) cannot easily be anonymised, so they should normally be treated as non-anonymous data. Anonymised data can usually be kept without security and can easily be passed to other investigators for specialist analysis.

The DPA requires access to be granted to participants to all of their data, if any part of that data allows their identification. If the data has been anonymised, two issues arise.

1. If the personal information has been removed from the data AND DESTROYED, then the DPA is no longer applicable, and the data can be kept without security. However, investigators should note that they will be unable to follow up or subsequently contact participants in any way, or associate individuals with particular data, and should not attempt to suggest they might do so.

2. If the personal information has been removed from the bulk of the data, but NOT destroyed (ie, is kept separately), then the DPA remains applicable. In this situation, the personal information needs to be (a) kept both separately and securely from the anonymised data, and (b) to be linked or 'keyed' to the anonymised data, such keys to be similarly kept securely (and often kept with the personal information).

If personal data is collected, in the 'Participant Information', inform the participant of:

- the processes the study will take to ensure data security;
- their right to access and correct their data and their right to request removal of their data;
- the authority which will give them access to their data (provide the contact information).

If sensitive data is collected, or the study involves clinical studies, human tissue samples, invasive procedures, or young or vulnerable people, provide additional detail. In the 'Participant Information', inform the participant of:

- the separation of identifying data and the anonymisation process;
- the method of linking the consent form (if any) to the participant's data;
- the processes for the destruction of all study data (if appropriate).

The study should conform to the University policy on data management applicable:

<http://www.southampton.ac.uk/library/research/researchdata/>

Investigators may find the University's survey platform useful:

<https://www.isurvey.soton.ac.uk/>

CONTACTS

risethic@soton.ac.uk.

¹ http://www.jisc.ac.uk/publications/generalpublications/2001/pub_dpacop_0101.aspx

Appendix C

Appendix C - LDA Analysis on Treasury Tweets

This appendix includes the remaining LDA results for the tweets tagged with the Thesaurus, Harvard and Treasury keyword sets.

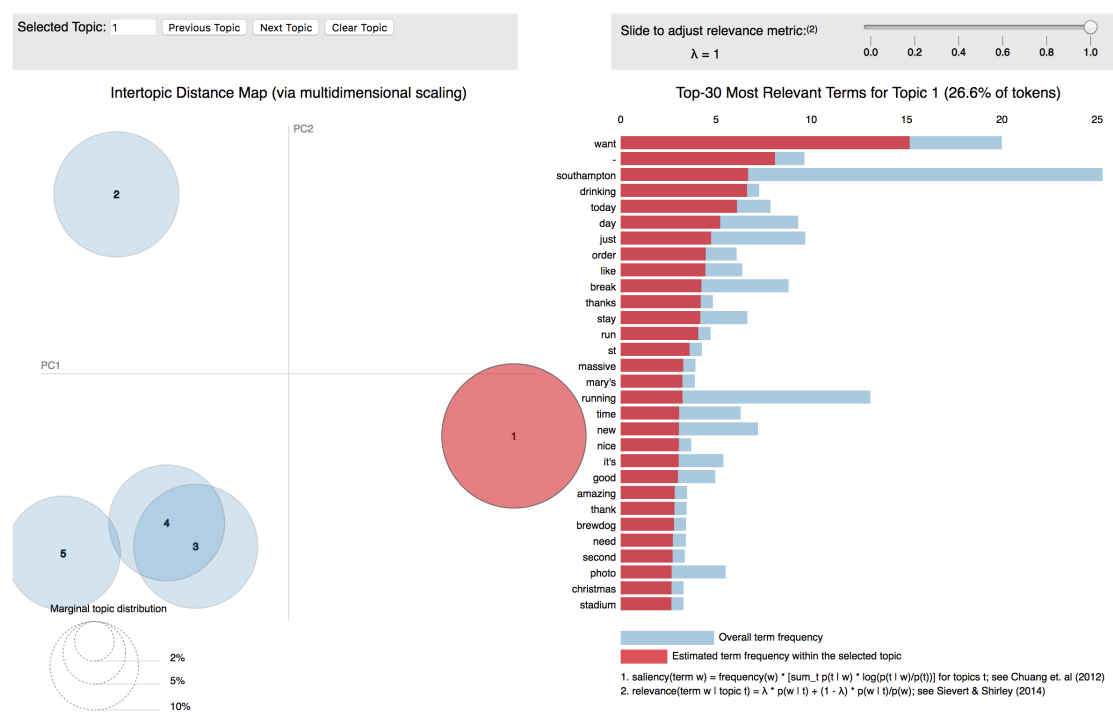


Figure C.1: The output of the LDA analysis on the Thesaurus keyword set classified as originating from a POI, visualised using the Python module LDAVis and with the forth topic highlighted. Image represents 725 tweets.

Figure C.1 shows a similar theme of running and exercise as the other spatial resolutions. Due to the broad semantic links between the keywords, the topics are quite hard to distinguish subjectively. Topic 1 includes features that connote a discussion of St Mary's

football stadium. Topic 2 also has feature such as ‘great’ and ‘save’ which could also contribute to a theme of football. Topics 3, 4 and 5 are harder to discern, with a mixture of features that focus around exercise.

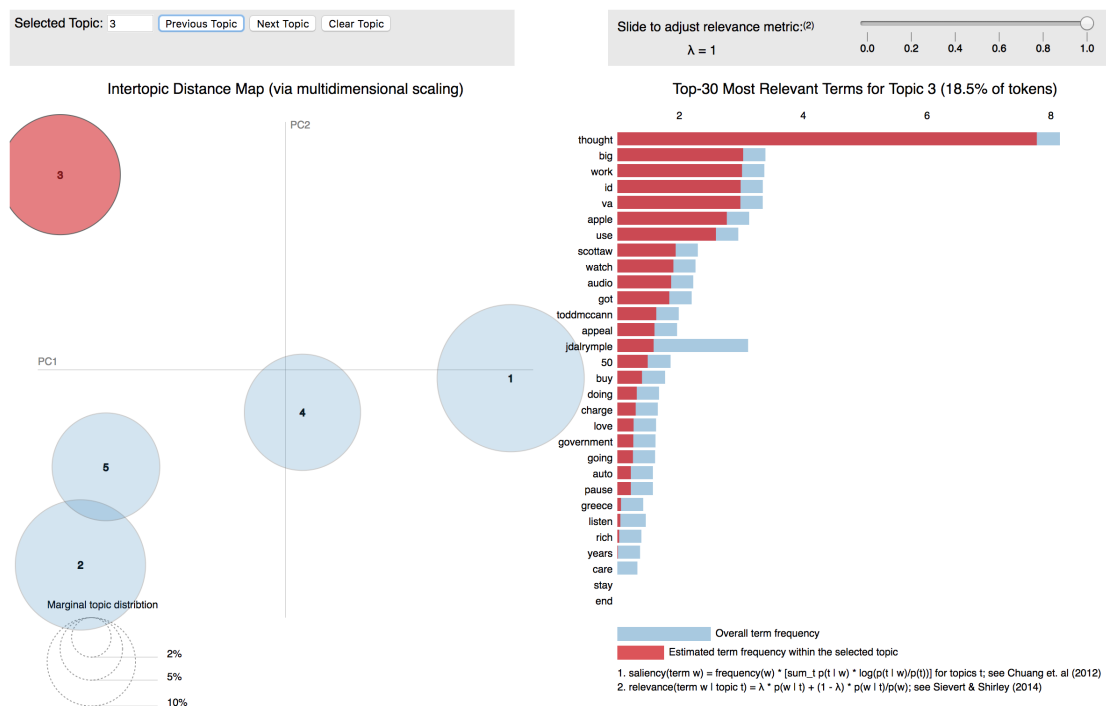


Figure C.2: The output of the LDA analysis on the Thesaurus keyword set classified as precise, visualised using the Python module LDAVis and with the forth topic highlighted. Image represents 1,019 tweets.

Figure C.2 includes more topics about running, a logical inclusion due to it reflecting users at precise geolocations. Topic 1 clusters tweets from the third-party exercise app ‘Endomondo’, with topic 4 also containing features such as ‘running’ and ‘endorphins’ that construct a narrative of exercise. Topic 3 (highlighted) shows a potential discussion of an Apple watch, a high-end consumer wearable that links with other Apple devices, though this is a tenuous topic. Topics 2 and 5 are hard to understand, though contain the term ‘cex’ so could be referring to technology, a similar theme as topic 3. More likely is that these topics are general discussions and a manual analysis of the tweets would be required to fully understand the topics. However, as 1,019 tweets are represented by these topics, they could provide interesting insights into the spatial activities of users discussing technology.

Topic 1 in figure C.3 shows a strong location-based narrative of the Common People music festival, hosted in Southampton Common, which was classified as a non-POI within the framework. Its absence in the POI version in figure 4.11 shows how the tweet resolution classification has successfully extracted relevant tweets. The rest of the topics are quite general, with some mentions of ‘hospital’, ‘golf’ and ‘farm’, though the term saliency is very low for the other topics thus it is hard to extrapolate topics.

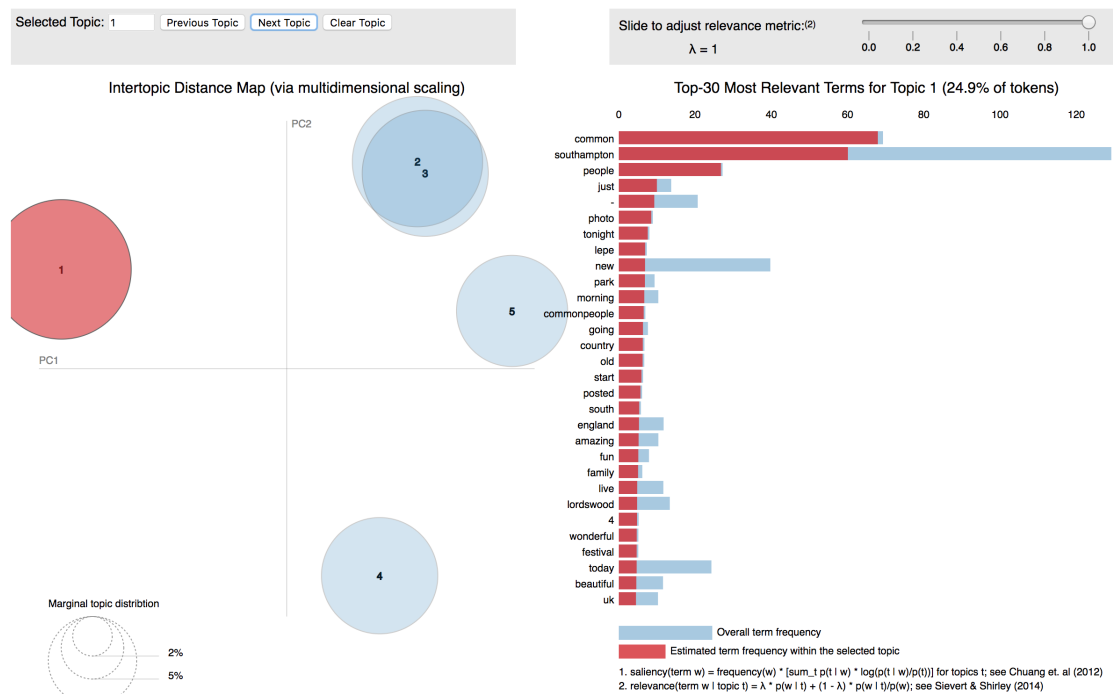


Figure C.3: The output of the LDA analysis on the Harvard keyword set classified as originating from a non-POI, visualised using the Python module LDAVis and with the forth topic highlighted. Image represents 1,782 tweets.

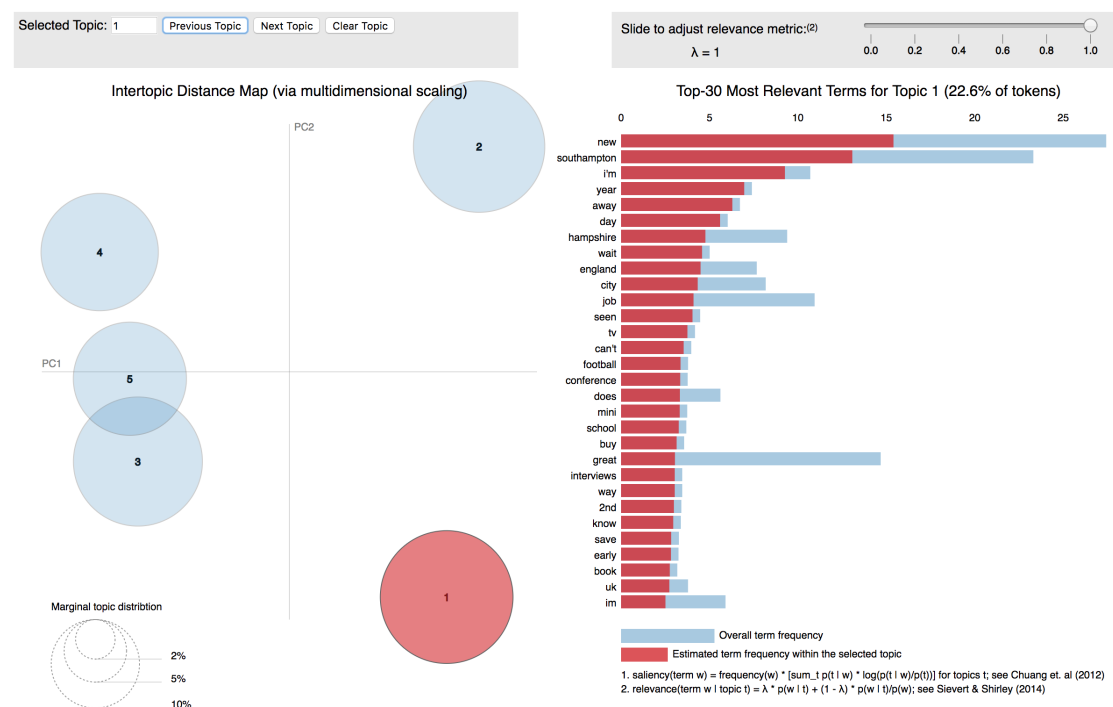


Figure C.4: The output of the LDA analysis on the Harvard keyword set classified as precise, visualised using the Python module LDAVis and with the forth topic highlighted. Image represents 7,400 tweets.

Figure C.4 shows 5 topics but all with a mixture of features. ‘Southampton’ features strongly in each of the topics, but corresponding terms are inconsistent; topic 1 shows potential discussions about ‘job’, while topic 4 mentions ‘apple’ and ‘watch’, a similar topic as shown in figure C.2 also depicting topics at the precise resolution. Topic 5 includes ‘park’ and ‘common’ as two prominent features, emphasising that Southampton Common is a hub of social activity.

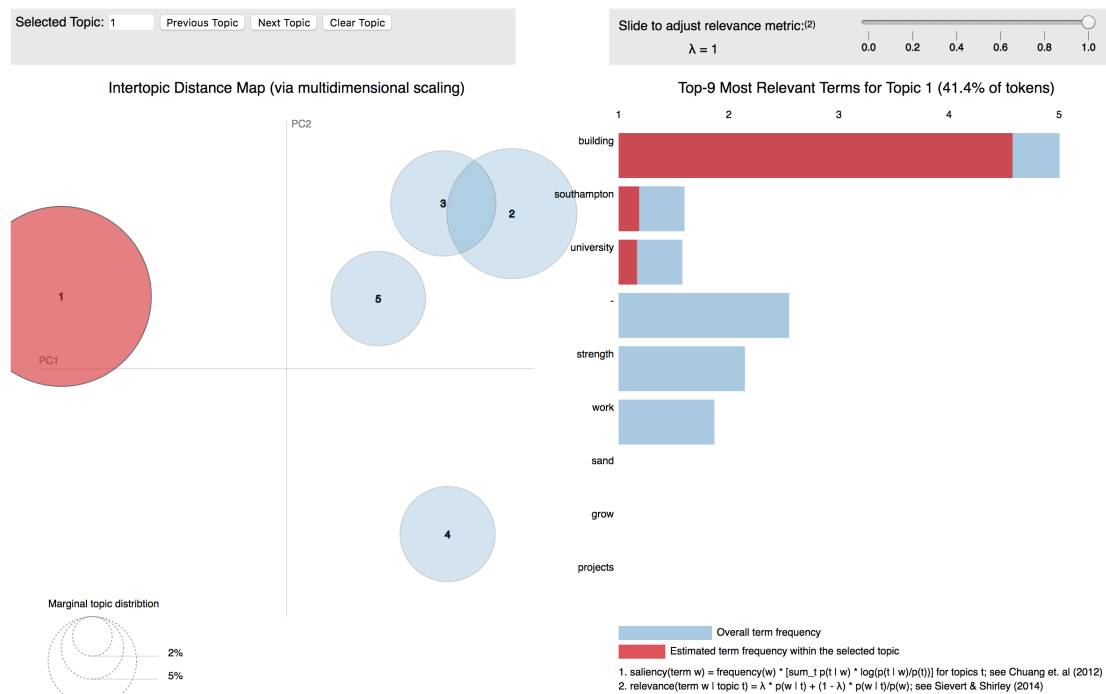


Figure C.5: The output of the LDA analysis on the Treasury keyword set classified as originating from a non-POI, visualised using the Python module LDAVis and with the forth topic highlighted. Image represents 15 tweets.

The Treasury dataset, at 23 terms, represents the smallest out of the three keyword sets. When this subset is further reduced to just tweets that were made at different spatial resolutions, LDA struggles to extract meaningful topics. This can be seen in figure C.5 that represents only 15 tweets. The term frequency and saliency are much lower than in the other results, and as such the topics are harder to distinguish. Despite this, ‘southampton’ and ‘university’ feature in the first topic alongside ‘building’, indicative of conversations about the campus and potentially about construction work. However, as can be seen in the figure, there are significantly fewer terms per topic as the LDA was only conducted on 15 tweets, thus making topic extraction challenging. The remaining topics do not include terms with high enough frequency to extract meaningful topics.

Figure C.6 includes more tweets than the previous figure, though issues arise over term frequency due to a small dataset. Topic 1 includes the similar terms of ‘southampton’ and ‘building’, while topic 2 (highlighted) includes different terms about ‘contract’ and ‘pay’. As with figure C.5, the remaining topics also do not have high enough frequencies

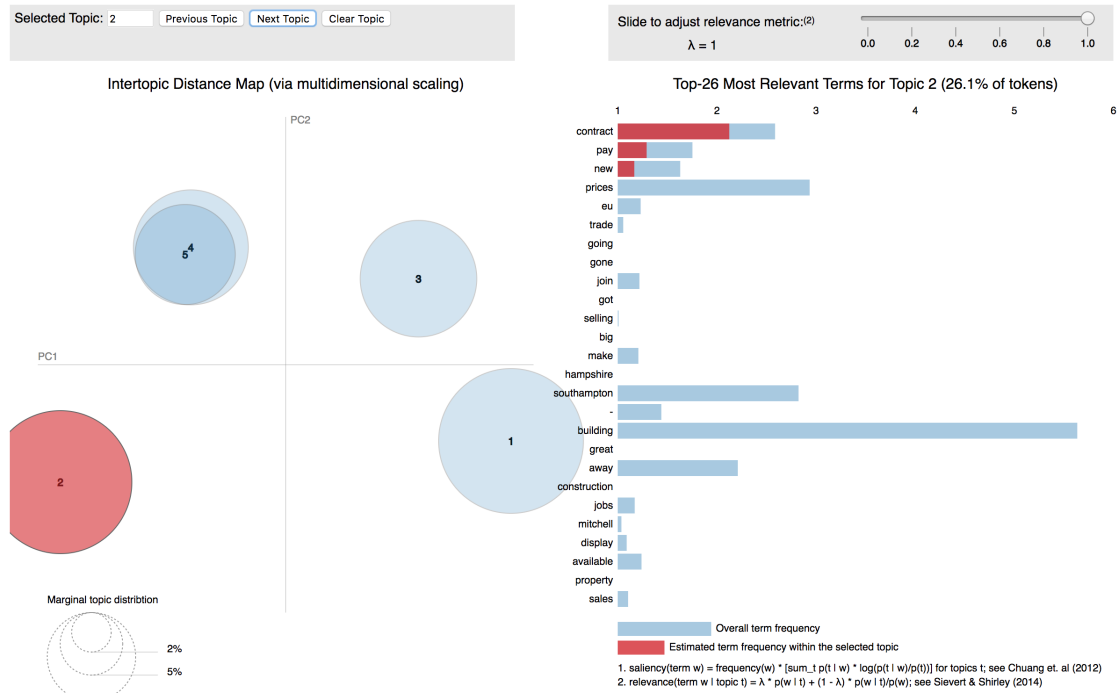


Figure C.6: The output of the LDA analysis on the Treasury keyword set classified as precise, visualised using the Python module LDAVis and with the forth topic highlighted. Image represents 92 tweets.

to extract meaningful topics. It is therefore unwise to conduct LDA on tweet sets with fewer than 100 tweets.

C.1 Concluding Remarks

All of the LDA results have shown a strong relationship between geographic resolution and thematic topics. For example, the tweets tagged at POIs all contain shop names or other location markers that indicate the POIs to which they belong. Similarly, some non-POI topics also contain location tags such as ‘southampton’ and ‘common’, highlighting the large park in the city. Extracting tweet resolutions is therefore an appropriate method with proven results.

Tweets classified as precise will have a wider range of topics due to POIs or other popular non-POIs not having an influence on the tweet creation. In terms of location analyses, however, precise tweets are much more informative about the location of the users and their mobility patterns than those clustered at POIs or non-POIs. With a large dataset like Harvard or Thesaurus, the precise topics still indicate a thematic narrative of retail, such as discussions about ‘apple’ and ‘watch’. However, these themes are harder to identify and would require further work to refine.

References

- Abilhoa, W. D. and De Castro, L. N. (2014). A keyword extraction method from twitter messages represented as graphs. *Applied Mathematics and Computation*, 240:308–325.
- Agnew, J. (2011). Space and place. In Agnew, J. and Livingstone, D., editors, *Handbook of Geographical Knowledge*, chapter 23, pages 316–330. Sage, London.
- Anselin, L. (1995). Local Indicators of Spatial Association - LISA. *Geographical Analysis*, 27(2):93–115.
- Antenucci, D., Cafarella, M., Levenstein, M. C., Ré, C., and Shapiro, M. D. (2014). Using Social Media to Measure Labor Market Flows.
- Arefi, M. (1999). Non-place and placelessness as narratives of loss: Rethinking the notion of place. *Journal of Urban Design*, 4(2):179–193.
- Ashktorab, Z., Brown, C., Nandi, M., and Culotta, A. (2014). Tweedr: Mining twitter to inform disaster response.
- Bailey, T. C. and Gatrell, A. C. (1995). *Interactive spatial data analysis*. Longman Scientific & Technical, Harlow.
- Barthes, R. (2002). *The Narrative Reader*. Routledge, London.
- Bassano, C., Barile, S., Piciocchi, P., Spohrer, J. C., Iandolo, F., and Fisk, R. (2019). Storytelling about places: Tourism marketing in the digital age. *Cities*, 87:10–20.
- Bauer, C. (2013). On the (In-)Accuracy of GPS Measures of Smartphones. In *Proceedings of International Conference on Advances in Mobile Computing & Multimedia - MoMM '13*, pages 335–341.
- Bazeley, P. and Jackson, K. (2014). *Qualitative data analysis with NVIVO*. Sage, Bristol, second edition.
- Beigi, G., Hu, X., Maciejewski, R., and Liu, H. (2016). An Overview of Sentiment Analysis in Social Media and Its Applications in Disaster Relief. *Sentiment Analysis and Ontology Engineering*, pages 313–340.

- Bennett, N. C., Millard, D., and Martin, D. (2016). Narrative Extraction through the Detection and Characterisation of National and Local Events. In Gartner, G. and Huang, H., editors, *Proceedings of the 13th International Conference on Location Based Services*, pages 196–200, Vienna, Austria. Vienna University of Technology.
- Bennett, N. C., Millard, D., Martin, D., and Amirian, P. (2017a). Spatial Narrative Construction using Thematic KDE. In *Proceedings of the 2017 International Conference on GeoComputation*, pages 1–8, Leeds, UK. Centre for Computational Geography, University of Leeds.
- Bennett, N. C., Millard, D., Martin, D., and Amirian, P. (2017b). Towards a Unified Narrative-Centric Spatial Clustering Model of Social Media Volunteered Geographic Information. In *Proceedings of the 25th GIS Research UK (GISRUK) 2017*, pages 1–7, Manchester, UK.
- Bennett, N. C., Millard, D. E., and Martin, D. (2018). Assessing Twitter Geolocation Resolution. In Akkermans, H., Fontaine, C., and Vermeulen, I., editors, *Proceedings of the 10th ACM Conference on Web Science*, pages 239–243, Amsterdam, Netherlands. ACM.
- Berners-Lee, T. and Fischetti, M. (1999). *Weaving the Web: The Past, Present and Future of the World Wide Web by its Inventor*. HarperInformation.
- Blei, D. M., Edu, B. B., Ng, A. Y., Edu, A. S., Jordan, M. I., and Edu, J. B. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3(1):993–1022.
- Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.
- Bollen, J., Pepe, A., and Mao, H. (2009). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 450–453, Barcelona. The AAAI Press.
- Boy, J. D. and Uitermark, J. (2017). Reassembling the city through Instagram. *Transactions of the Institute of British Geographers*, 42(4):612–624.
- Burke, S. (2016). Rethinking validity’ and trustworthiness’ in qualitative inquiry: How might we judge the quality of qualitative research in sport and exercise sciences? In *Routledge handbook of qualitative research in sport and exercise*, pages 330–339. Routledge.
- Callen, T. (2008). Back to Basics: What Is Gross Domestic Product? *Finance and Development*, 45(4):48–49.
- Cambria, E. and White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, 9(2):48–57.

- Campbell, J. L., Quincy, C., Osserman, J., and Pedersen, O. K. (2013). Coding In-depth Semistructured Interviews: Problems of Unitization and Inter-coder Reliability and Agreement. *Sociological Methods and Research*, 42(3):294–320.
- Casadei, P. and Lee, N. (2020). Global cities, creative industries and their representation on social media: A micro-data analysis of Twitter data on the fashion industry. *Environment and Planning A: Economy and Space*, 0(0):1–26.
- Casas, I. and Delmelle, E. C. (2017). Tweeting about public transit: Gleaning public perceptions from a social media microblog. *Case Studies on Transport Policy*, 5(4):634–642.
- Chae, J., Thom, D., Bosch, H., Jang, Y., Maciejewski, R., Ebert, D. S., and Ertl, T. (2012). Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *IEEE Conference on Visual Analytics Science and Technology 2012, VAST 2012 - Proceedings*, pages 143–152. IEEE.
- Chenail, R. J. (2011). Interviewing the investigator: Strategies for addressing instrumentation and researcher bias concerns in qualitative research. *Qualitative Report*, 16(1):255–262.
- Cheng, Z., Caverlee, J., and Lee, K. (2010). You are where you tweet. *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*, page 759.
- Clifford, N., French, S., and Valentine, G. (2016). *Key Methods in Geography*. Sage.
- Cranshaw, J., Monroy-Hernández, A., and Needham, S. (2016). Journeys and Notes: Designing Social Computing for Non-Places. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, pages 4722–4733.
- Deerwester, S., Dumais, S. T., Furnas, G. W., and Landauer, T. K. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407.
- Dou, W., Wang, X., Ribarsky, W., and Zhou, M. (2012). Event Detection in Social Media Data. *IBM Almaden Research Center*, pages 2–5.
- Eickhoff, M. and Wienieke, R. (2018). Understanding Topic Models in Context: A Mixed-Methods Approach to the Meaningful Analysis of Large Document Collections. *Proceedings of the 51st Hawaii International Conference on System Sciences*, pages 903–912.
- EMarketer (2015). TV Sports, Entertainment Get UK Twitterers Typing. [Online] <https://www.emarketer.com/Article/TV-Sports-Entertainment-UK-Twitterers-Typing/1012370>. [Accessed 20 September 2018].

- Fang, A., Ounis, I., Habel, P., Macdonald, C., and Limsopatham, N. (2015). Topic-centric Classification of Twitter User's Political Orientation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '15*, pages 791–794.
- Farrow, E., Dickinson, T., and Aylett, M. P. (2015). Generating narratives from personal digital data: Using sentiment, themes, and named entities to construct stories. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9299, pages 473–477. Springer International Publishing.
- Fereday, J. and Muir-Cochrane, E. (2006). Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. *International Journal of Qualitative Methods*, 5(1):80–92.
- Fludernik, M. (2009). *An Introduction to Narratology*. Routledge, New York, 1 edition.
- Friedman, T. L. T. (2005). *The world is flat: A brief history of the twenty-first century*. Macmillan.
- Fritz, S., See, L., and Brovelli, M. (2017). Motivating and Sustaining Participation in VGI. In Foody, G., See, L., Fritz, S., Mooney, P., Olteanu-Raimond, A.-M., Fonte, C. C., and Antoniou, V., editors, *Mapping and the Citizen Sensor*, chapter 5, pages 93–117. Ubiquity Press, London.
- Gao, S., Li, L., Li, W., Janowicz, K., and Zhang, Y. (2017). Constructing gazetteers from volunteered Big Geo-Data based on Hadoop. *Computers, Environment and Urban Systems*, 61:172–186.
- Gao, Y., Wang, S., Padmanabhan, A., Yin, J., and Cao, G. (2018). Mapping spatiotemporal patterns of events using social media: a case study of influenza trends. *International Journal of Geographical Information Science*, 32(3):425–449.
- García-Palomares, J. C., Salas-Olmedo, M. H., Moya-Gómez, B., Condeço-Melhorado, A., and Gutiérrez, J. (2018). City dynamics through Twitter: Relationships between land use and spatiotemporal demographics. *Cities*, 72:310–319.
- Garrison, D. R., Cleveland-Innes, M., Koole, M., and Kappelman, J. (2006). Revisiting methodological issues in transcript analysis: Negotiated coding and reliability. *Internet and Higher Education*, 9(1):1–8.
- Gattani, A., Doan, A., Lamba, D. S., Garera, N., Tiwari, M., Chai, X., Das, S., Subramaniam, S., Rajaraman, A., and Harinarayan, V. (2013). Entity extraction, linking, classification, and tagging for social media. *Proceedings of the VLDB Endowment*, 6(11):1126–1137.

- GDELT (2018). The GDELT Project. [Online] <https://www.gdeltproject.org/>. [Accessed 3 October 2018].
- Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, 61(1):115–125.
- Ghermandi, A. and Sinclair, M. (2019). Passive crowdsourcing of social media in environmental research: A systematic map. *Global Environmental Change*, 55:36–47.
- Ghosh, D. D. and Guha, R. (2013). What are we tweeting’ about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and Geographic Information Science*, 40(2):90–102.
- Gilani, Z., Farahbakhsh, R., Tyson, G., Wang, L., and Crowcroft, J. (2017). An in-depth characterisation of Bots and Humans on Twitter. *arXiv preprint arXiv:1704.01508*.
- Gong, Z. and Liu, Q. (2009). Improving keyword based web image search with visual feature distribution and term expansion. *Knowledge and Information Systems*, 21(1):113–132.
- Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, 69(4):211–221.
- Gu, Y., Qian, Z. S., and Chen, F. (2016). From Twitter to detector: Real-time traffic incident detection using social media data. *Transportation Research Part C: Emerging Technologies*, 67:321–342.
- Guo, Y., Barnes, S. J., and Jia, Q. (2017). Mining Meaning from Online Ratings and Reviews: Tourist Satisfaction Analysis Using Latent Dirichlet Allocation. *Tourism Management*, 59:467–483.
- Habernal, I., Ptáček, T., and Steinberger, J. (2013). Sentiment analysis in czech social media using supervised machine learning. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, number 1, pages 65–74, Atlanta, Georgia. The Association for Computational Linguistics.
- Halford, S., Weal, M., Tinati, R., Carr, L., and Pope, C. (2018). Understanding the production and circulation of social media data: Towards methodological principles and praxis. *New Media and Society*, 20(9):3341–3358.
- Hamstead, Z. A., Fisher, D., Ilieva, R. T., Wood, S. A., McPhearson, T., and Kremer, P. (2018). Geolocated social media as a rapid indicator of park visitation and equitable park access. *Computers, Environment and Urban Systems*, 72:38–50.
- Hargood, C. (2011). *Semiotic term expansion as the basis for thematic models in narrative systems*. PhD thesis, University of Southampton.

- Hasan, S., Zhan, X., and Ukkusuri, S. V. (2013). Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, pages 1–8.
- Hecht, B., Hong, L., Suh, B., and Chi, E. H. (2011). Tweets from Justin Bieber’s heart. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 237.
- Himelboim, I., Cameron, K., Sweetser, K. D., Danelo, M., and West, K. (2016). Valence-based homophily on Twitter: Network Analysis of Emotions and Political Talk in the 2012 Presidential Election. *New Media and Society*, 18(7):1382–1400.
- Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57.
- Ibrahim, R., Elbagoury, A., Kamel, M. S., Karray, F., Elbagoury, B. A., and Ca, K. (2017). Tools and approaches for topic detection from Twitter streams: survey. *Knowledge and Information Systems*, pages 1–29.
- Jeske, D., McNeill, A. R., Coventry, L., and Briggs, P. (2017). Security information sharing via Twitter: ‘Heartbleed’ as a case study. *International Journal of Web Based Communities*, 13(2):172.
- Juntao, L., Cheng, T., and Lansley, G. (2015). Spatio-Temporal Patterns of Passengers’ Interests at London Tube Stations. In *Proceedings of the 23rd Conference on GIS Research UK*, pages 1–5.
- Kallio, H., Pietilä, A. M., Johnson, M., and Kangasniemi, M. (2016). Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. *Journal of Advanced Nursing*, 72(12):2954–2965.
- Kaneko, T. and Yanai, K. (2015). Event photo mining from twitter using keyword bursts and image clustering. *Neurocomputing*, 172:143–158.
- Kim, Y., Huang, J., and Emery, S. (2016). Garbage in, garbage out: Data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection. *Journal of Medical Internet Research*, 18(2):41.
- Landau, M. (1984). Human Evolution as Narrative. *American Scientist*, 72(3):262–268.
- Lansley, G. and Longley, P. A. (2016). The geography of Twitter topics in London. *Computers, Environment and Urban Systems*, 58:85–96.

- Levell, P. (2015). Is the Carli index flawed?: Assessing the case for the new retail price index RPIJ. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 178(2):303–336.
- Li, X., Pham, T.-A. N., Cong, G., Yuan, Q., Li, X.-L., and Krishnaswamy, S. (2015). Where you Instagram? Associating Your Instagram Photos with Points of Interest. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management - CIKM '15*, pages 1231–1240.
- Lichman, M. and Smyth, P. (2014). Modeling human location data with mixtures of kernel densities. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 35–44, New York, New York, USA. ACM Press.
- Lin, J. and Cromley, R. G. (2018). Inferring the home locations of Twitter users based on the spatiotemporal clustering of Twitter data. *Transactions in GIS*, 22(1):82–97.
- Liu, C. and Fuhrmann, S. (2018). Enriching the GIScience research agenda: Fusing augmented reality and location-based social networks. *Transactions in GIS*, 22(3):775–788.
- Lloyd, A. and Cheshire, J. (2017). Deriving retail centre locations and catchments from geo-tagged Twitter data. *Computers, Environment and Urban Systems*, 61:108–118.
- Lloyd, C. D. (2010). *Spatial Data Analysis: An introduction for GIS users*. Oxford University Press.
- Longley, P. A. and Adnan, M. (2016). Geo-temporal Twitter demographics. *International Journal of Geographical Information Science*, 30(2):369–389.
- Mahmood, A. (2009). Literature Survey on Topic Modeling. Technical report, Delaware.
- Mahmud, J., Nichols, J., and Drews, C. (2012). Where Is this tweet from? Inferring home locations of Twitter users. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM-12)*, pages 511–514.
- Mahmud, J., Nichols, J., and Drews, C. (2014). Home Location Identification of Twitter Users. *ACM Transactions on Intelligent Systems and Technology*, 5(3):1–21.
- Martin, D., Cockings, S., and Harfoot, A. (2013). Development of a geographical framework for census workplace data. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 176(2):585–602.
- Martin, D., Gale, C., Cockings, S., and Harfoot, A. (2018). Origin-destination geodemographics for analysis of travel to work flows. *Computers, Environment and Urban Systems*, 67:68–79.

- Maynard, D., Bontcheva, K., and Rout, D. (2012). Challenges in developing opinion mining tools for social media. In *LREC 2012*, pages 15–22.
- Maynard, D. and Hare, J. (2015). Entity-based Opinion Mining from Text and Multimedia. In Gaber, M. M., Cocea, M., Wiratunga, N., and Goker, A., editors, *Advances in Social Media Analysis*, volume 602, chapter 4, pages 65–86. Springer International Publishing.
- McMinn, A. J., Moshfeghi, Y., and Jose, J. M. (2013). Building a large-scale corpus for evaluating event detection on twitter. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM 2013)*, pages 409–418.
- Middleton, S. E., Middleton, L., and Modafferi, S. (2014). Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, 29(2):9–17.
- Millard, D. E. and Hargood, C. (2015). A Research Framework for Engineering Location-Based Poetics. In *Narrative and Hypertext*, pages 13–16.
- Mislove, A., Lehmann, S., Ahn, Y.-y., Onnela, J.-p., and Rosenquist, J. N. (2011). Understanding the Demographics of Twitter Users. *Artificial Intelligence*, pages 554–557.
- Mooney, P. and Corcoran, P. (2014). Analysis of interaction and co-editing patterns amongst openstreetmap contributors. *Transactions in GIS*, 18(5):633–659.
- Mooney, P., Minghini, M., Laakso, M., Antoniou, V., Olteanu-Raimond, A. M., and Skopeliti, A. (2016). Towards a protocol for the collection of VGI vector data. *ISPRS International Journal of Geo-Information*, 5(11).
- Morstatter, F., Pfeffer, J., and Liu, H. (2014). When is it biased? Assessing the representativeness of twitter’s streaming API. *WWW 2014 Companion - Proceedings of the 23rd International Conference on World Wide Web*, pages 555–556.
- Morstatter, F., Pfeffer, J., Liu, H., and Carley, K. (2013). Is the sample good enough? Comparing data from Twitter’s streaming API with Twitter’s firehose. In *Proceedings of ICWSM*, pages 400–408, Cambridge, MA. AAAI Press.
- Narayanan, A. and Felten, E. W. (2014). No silver bullet: De-identification still doesn’t work. *White Paper*, pages 1–8.
- Office for National Statistics (2018a). Internet access households and individuals, Great Britain: 2017. [Online] <https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/bulletins/internetaccesshouseholdsandindividuals/2018>. [Accessed 20 September 2018].

- Office for National Statistics (2018b). Overview of the UK population. [Online] <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/overviewoftheukpopulation/july2017>. [Accessed 20 September 2018].
- Office for National Statistics (2018c). Social networking by age group, 2011 to 2017. [Online] <https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/adhocs/007401socialnetworkingbyagegroup2011to2017>. [Accessed 20 September 2018].
- Omnicoreagency (2018). Twitter by the Numbers (2018): Stats, Demographics & Fun Facts. [Online] <https://www.omnicoreagency.com/twitter-statistics/>.
- Patel, N. N., Stevens, F. R., Huang, Z., Gaughan, A. E., Elyazar, I., and Tatem, A. J. (2017). Improving Large Area Population Mapping Using Geotweet Densities. *Transactions in GIS*, 21(2):317–331.
- Pavel, T. G. (1988). Formalism in Narrative Semiotics. *Poetics Today*, 9(3):593–606.
- Pereira, J., Pasquali, A., Saleiro, P., Rossetti, R., and Cacho, N. (2017). Characterizing geo-located tweets in brazilian megacities. In *2017 International Smart Cities Conference, ISC2 2017*, pages 1–6. IEEE.
- Pohl, D., Bouchachia, A., and Hellwagner, H. (2012). Automatic sub-event detection in emergency management using social media. *Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion*, page 683.
- Quercia, D. and Saez, D. (2014). Mining urban deprivation from foursquare: Implicit crowdsourcing of city land use. *IEEE Pervasive Computing*, 13(2):30–36.
- Quesnot, T. and Roche, S. (2015). Platial or locational data? Toward the characterization of social location sharing. In *Proceedings of the Annual Hawaii International Conference on System Sciences*, volume 2015-March, pages 1973–1982. IEEE.
- Reigadinha, T., Godinho, P., and Dias, J. (2017). Portuguese food retailers Exploring three classic theories of retail location. *Journal of Retailing and Consumer Services*, 34:102–116.
- Relph, E. (1976). *Place and Placelessness*. Pion, London.
- Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. *WWW '10: Proceedings of the 19th international conference on World wide web*, page 851.
- Saker, M. (2017). Foursquare and identity: Checking-in and presenting the self through location. *New Media and Society*, 19(6):934–949.

- Saker, M. and Evans, L. (2016). Everyday life and locative play: an exploration of Foursquare and playful engagements with space and place. *Media, Culture and Society*, 38(8):1169–1183.
- Singh, K., Malik, D., and Sharma, N. (2011). Evolving limitations in K-means algorithm in data mining and their removal. *IJCEM International Journal of Computational Engineering & Management ISSN*, 12(April):2230–7893.
- Sloan, L. and Morgan, J. (2015). Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. *PLoS ONE*, 10(11):1–15.
- Smith, B. and McGannon, K. R. (2018). Developing rigor in qualitative research: problems and opportunities within sport and exercise psychology. *International Review of Sport and Exercise Psychology*, 11(1):101–121.
- Smith, L., Liang, Q., James, P., and Lin, W. (2017). Assessing the utility of social media as a data source for flood risk management using a real-time modelling framework. *Journal of Flood Risk Management*, 10(3):370–380.
- Tamburrini, N., Cinnirella, M., Jansen, V. A. A., and Bryden, J. (2015). Twitter users change word usage according to conversation-partner social identity. *Social Networks*, 40:84–89.
- Tomar, G., Singh, M., Rai, S., and Kumar, A. (2013). Probabilistic Latent Semantic Analysis for Unsupervised Word Sense Disambiguation. *IJCSI International Journal of Computer Science Issues*, 10(5):1694–0784.
- Tomashevsky, B. (1965). Russian Formalist Criticism: Four Essays. In Lemon, L. T. and Rees, R. J., editors, *Thematics*, pages 66–68. University of Nebraska Press.
- Toubia, O., Iyengar, G., Bunnell, R., and Lemaire, A. (2018). Extracting Features of Entertainment Products: A Guided Latent Dirichlet Allocation Approach Informed by the Psychology of Media Consumption. *Journal of Marketing Research*, 56(1):18–36.
- Tracy, S. J. (2010). Qualitative quality: Eight “big-tent” criteria for excellent qualitative research. *Qualitative Inquiry*, 16(10):837–851.
- Tremayne, M. (2014). Anatomy of Protest in the Digital Era: A Network Analysis of Twitter and Occupy Wall Street. *Social Movement Studies*, 13(1):110–126.
- Wing, M., Eklund, A., and Kellogg, L. (2005). Consumer-grade global positioning system (GPS) accuracy and reliability. *Journal of Forestry*, 103(4).
- Xu, Z., Liu, Y., Yen, N., Mei, L., Luo, X., Wei, X., and Hu, C. (2016). Crowdsourcing based Description of Urban Emergency Events using Social Media Big Data. *IEEE Transactions on Cloud Computing*, pages 1–11.

- Yalcinkaya, M. and Singh, V. (2015). Patterns and trends in Building Information Modeling (BIM) research: A Latent Semantic Analysis. *Automation in Construction*, 59:68–80.
- Yasunaga, M. and Lafferty, J. D. (2019). TopicEq: A Joint Topic and Mathematical Equation Model for Scientific Texts. *arXiv preprint arXiv:1902.06034*.
- Yu, D., Xu, D., Wang, D., and Ni, Z. (2019). Hierarchical Topic Modeling of Twitter Data for Online Analytical Processing. *IEEE Access*, 7:12373–12385.
- Zafeiropoulou, A. M., O’Hara, K., Millard, D. E., and Webber, C. (2012). Location data and privacy: a framework for analysis. *Réseaux Sociaux: Culture Politique et Ingénierie des Réseaux Sociaux*, pages 185–200.
- Zhang, Y. and Eick, C. F. (2019). Tracking Events in Twitter by Combining an LDA-Based Approach and a Density-Contour Clustering Approach. *International Journal of Semantic Computing*, 13(1):87–110.
- Zhao, L., Wang, J., Chen, F., Lu, C. T., and Ramakrishnan, N. (2017). Spatial Event Forecasting in Social Media with Geographically Hierarchical Regularization. *Proceedings of the IEEE*, 105(10):1953–1970.
- Zielinski, A., Middleton, S. E., Tokarchuk, L. N., and Wang, X. (2013). Social media text mining and network analysis for decision support in natural crisis management. In *Proceedings of the Tenth International Conference on Information Systems for Crisis Response and Management*, pages 840–845, Baden-Baden, Germany.