Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Review

Mining location from social media: A systematic review

Kristin Stock

Massey University, Private Bag 102904, North Shore, Auckland 0745, New Zealand

ARTICLE INFO

Keywords:
Social media
Geographic information

ABSTRACT

During the last ten years, a large body of research extracting and analysing geographic data from social media has developed. We analyse 690 papers across 20 social media platforms, focussing particularly on the method used for extraction of location information. We discuss and compare extraction methods, and consider their accuracy and coverage. While much work has adopted location information in the form of coordinates in message metadata, this approach has very limited coverage in most platforms and reports on posting location rather than message location or the location that the message refers to (geofocus). In contrast, a wide array of other approaches have been developed, with methods that extract place names from message text providing the highest accuracy. Methods that use social media connections also provide good results, but all of the methods have limitations. We also present analysis of the range and frequency of use of different social media platforms, and the wide range of application areas that have been addressed. Drawing on this analysis we present a number of future areas of research that warrant attention in order for this field of research to mature.

1. Introduction

The potential for social media to provide useful geographic information to either replace or augment traditional methods of data collection has been recognised for some years. In that time, a large number of research efforts have explored this potential with applications including health, disaster management, tourism and recreation, environmental monitoring, crime, civil unrest and marketing.

In this paper we provide a systematic literature review of papers across the field, identifying 690 papers within scope, analysing their content in order to compare different aspects of the research and identifying gaps and future research potential, particularly focussing on three aspects. Firstly, we review the different social media platforms that have been used for extracting geodata in the published literature. There is a clear preference for Twitter over other platforms, and we discuss the reasons for this and the potential for the increased use of other platforms to extract data that is not currently being used. Secondly, we explore the methods used to extract location information from social media. While use of metadata geotagging is the most common method, it has a number of limitations, and other methods including text mining, user profiling and different kinds of inference have been developed. We discuss these methods, their use and advantages and disadvantages, analysing accuracy and coverage achieved by each method. Thirdly, we review the impressive array of applications that have been addressed with data extracted from social media, and discuss the dominance of different application areas. Finally, we propose future research directions to cover gaps in the current work, and to enable this research field to reach maturity.

The organisation of the paper is as follows. Section 2 discussed previous reviews that have been conducted on social media location data and related areas. Section 3 describes the methodology used for the systematic literature review and presents the research questions. Section 4 provides analysis of the social media platforms used to extract geographic data. Section 5 discusses and compares specific methods of location extraction, providing detailed discussion about the alternative approaches. Section 6 analyses the application domains used in the research papers surveyed and Section 7 discusses future research directions.

2. Literature review

A number of reviews have previously been completed in the field of social media location extraction, exploring different aspects. In this Section, we discuss firstly those reviews that address social media generally (not specifically spatial data), then those addressing location-based services and image analysis. We then consider those that focus more directly on extraction and analysis of geographic information, including reviews that focus on analysis focussed reviews, those that address VGI and those that review papers that address the use of specific application areas to extract geographic information. This summary then indicates the gap that we intend to fill with this paper.

Reviews that have looked at social media generally (not specifically

E-mail address: k.stock@massey.ac.nz.

https://doi.org/10.1016/j.compenvurbsys.2018.05.007
Received 16 March 2018; Received in revised form 17 May 2018; Accepted 18 May 2018

at spatial aspects) include Batrinca and Treleaven (2015), who provide a survey of social media technologies, methods for analysis and APIs as a tool for social scientists, but differ from our work in that they do not provide any detail about location, and Hua et al. (2012), who focus on approaches to extraction of content from Twitter, providing a section on detection of current location, briefly reviewing a few papers.

In the area of mobile and location-based social media, Kaplan (2012) surveys mobile marketing, and distinguishes between the different ways that marketing strategies consider location and time, but does not describe any aspects of location extraction in detail, and Bao et al. (2015) review recommendation systems for location-based social networks. Their review is mainly focussed around the approaches and methods used for making recommendations, and on location-based social networks (LBSN) in which places (e.g. venues) are first class citizens, enabling user check in. Although they refer to the ways in which geotags and other forms of location are used, they do not discuss this in detail.

Reviews that focus on image analysis include Liu (2011), who reviews a set of papers that use geographic information to analyse images, sometimes in combination with other information (e.g. image tags); Yanai (2015) provides a summary of analysis of web images, touching on the use of image analysis to infer location in the context of Flickr and Luo et al. (2011) discuss some approaches to determine the location of photos, reviewing several interesting papers in this area. Zheng et al. (2011) also discuss georeferencing from an image point of view, reviewing approaches to location landmarks and more general locations from photos.

Moving towards a greater emphasis on geographic information exclusively, Stieger et al. (2015) conduct a systematic literature review on the use of Twitter for geospatial analysis. They investigate 92 papers, examining discipline of authors, application domain, time of publication, type of data extracted and broad category of paper, and then look in more detail at the kind of analyses performed. They address the ways in which location information is extracted to some degree, but focus on the ways in which the location information has been analysed.

Senaratne et al. (2017) comprehensive review into quality issues in VGI points out some of the issues involved in using both active and passive (i.e. social media) VGI, and some of the approaches that have been developed to deal with these issues; and Yap et al. (2012) describe some of the requirements for a successful VGI-based LBS, including privacy, trust and information classification functions. They do not discuss details of location representation or extraction. Goodchild and Li (2012) propose three different approaches to dealing with quality: crowdsourcing, in which other contributors correct the errors of their peers; social, in which moderators police or verify contributions and geographic, in which spatial patterns can be used to identify unlikely or inconsistent contributions. Although this latter work applies more generally to all VGI, these approaches could be applied to data extracted from social media.

A number of studies review literature on the use social media in particular application areas, including Guy et al. (2011), who address the use of social media in disease surveillance. They do not discuss methods for extraction of location from social media in this context, but they acknowledge the need to "…determine the effectiveness of geolocation in garnering real-time estimates of ILI (influenza-like illness)" (p.5). Similarly, Velasco et al. (2014) explore the use of social media type approaches in disease surveillance, but do not discuss methods to extract location. Horita et al. (2013) discuss the use of VGI in disaster events and provide an overview of disaster types and the phase in which VGI is used. Their study is wider than social media, also including more active methods of crowdsourcing, with 6 of the 21 papers they summarise using social media. Imran et al. (2015) also focus on disaster events, providing a review of methods for processing social media messages in disaster situations. They only address location briefly. Klonner et al. (2016) review papers looking at the use of VGI in the preparedness and mitigation phases of a disaster, but do not discuss extraction of location information. Leung et al. (2013) review papers on the use of social media for tourism and hospitality, but do not address location. Yue et al. (2014) describe data collection options to study trajectory-based travel behaviour, one of which is social media, and identify several studies.

Most relevant to our work, Ajao et al. (2015) conduct a review of location extraction approaches that have been used in Twitter, identifying seven different types of location indicator (tweet content, geotag, social networks; user profile; geotag; third party sources (for geocoding and reverse geocoding); time zones and web snippets. They then discuss the way Natural Language Processing (NLP) (specifically Named Entity Recognition [NER]) and gazetteers have been used to extract location. Our work is very closely related to this previous survey, and builds on it.

Our work differs from the previous work in that we consider a wider view, looking across social media rather than only focussing on Twitter; identifying the differences and gaps across social media platforms and studying location extraction approaches in detail. We also provide a systematic (quantitative) review which offers figures regarding the use of different approaches. Finally, we summarise the range of applications to which this approach has been applied. We also go beyond much of the previous work in identifying future research directions required in order to make use of a broader range of available data and fully realize the potential of this research field.

3. Methodology

Our systematic literature review follows the methodology described in Kitchenham and Charters (2007) and Kitchenham et al. (2009). In addition, even though we do not consider our work to constitute a scoping study, our review shares some goals in common with Arksey and O′Malley's (2005) four reasons for conducting a scoping study: namely that we aim to "examine the extent, range and nature of research activity" (p.21) on the use of spatial data in social media and that we aim to identify gaps in the existing literature, specifically in relation to potential ways to exploit social media that have not yet been considered.

We address the following broad research questions:

RQ1. : From which social media platforms has geographic data been extracted?

RQ2. : What methods have been used to extract location information from social media?

RQ3. : Which domains, sub-domains and research questions has geographic data extracted from social media been used to addressed?

The selection of these research questions was motivated firstly by the goal of maximising the opportunities offered by social media data for geographic mapping and analysis. Our intention was to determine whether there were social media platforms that are popular among users, and that contain significant amounts of data, but that have been neglected in the literature, and similarly to determine whether there were obvious gaps or under-presentation in particular methods for extracting location. A scan of the existing literature suggested the dominance of geotagged data from Twitter, but we wanted to confirm whether this was the case, and to highlight opportunities to exploit, compare and evaluate other data sources and location extraction methods. Similarly, we were interested in gaps in the research in particular application areas, and whether there were opportunities for new investigations that had not yet been addressed.

Our initial selection of candidate papers was achieved through

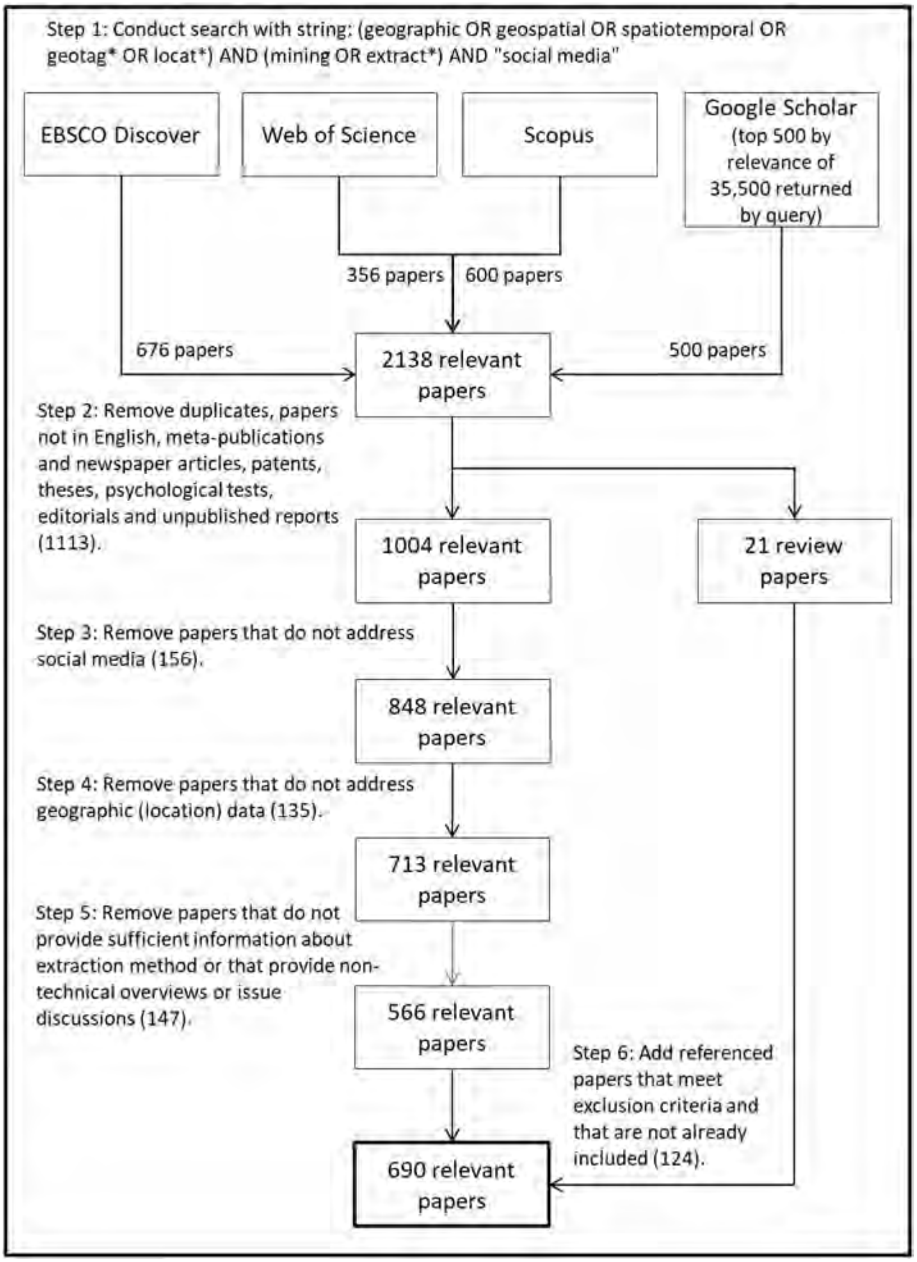
Step 1: Conduct search with string: (geographic OR geospatial OR spatiotemporal OR geotag* OR locat*) AND (mining OR extract*) AND "social media"
EBSCO Discover
Web of Science
Scopus
Google Scholar (top 500 by relevance of 35,500 returned by query)
356 papers
600 papers
676 papers
500 papers
2138 relevant papers
Step 2: Remove duplicates, papers not in English, meta-publications and newspaper articles, patents, theses, psychological tests, editorials and unpublished reports (1113).
1004 relevant papers
21 review papers
Step 3: Remove papers that do not address social media (156).
848 relevant papers
Step 4: Remove papers that do not address geographic (location) data (135).
713 relevant papers
Step 5: Remove papers that do not provide sufficient information about extraction method or that provide non-technical overviews or issue discussions (147).
566 relevant papers
Step 6: Add referenced papers that meet exclusion criteria and that are not already included (124).
690 relevant papers

Fig. 1. Paper selection process.

searches in four databases. Fig. 1 illustrates the process and the steps that were used to filter the results, and the quantity of papers returned at each step. The search string was defined widely, to include a wide range of options, but to exclude papers that were not relevant. Additional terms like “place” and “detect*” might have been included, but these are also used in a wider sense, and the cost in terms of irrelevant inclusions was thought to outweigh the benefits.

The scope of our study was confined to papers entirely written in English (not just the abstract), and that were published in the peer-reviewed literature. Unpublished reports, patents, theses, volume editorials (that summarise volume contents) and newspaper articles were excluded. The search also returned a number of meta-publications (volumes containing a collection of papers). These were excluded on the basis that relevant papers within that volume would be returned by the search process.

We confine our scope to only papers that extract content from publicly-accessible social media, and that address social media. While wider scopes may be of interest, our focus here was on geographic information from social media. Our scope was further restricted to social media in which geographic data is available as a by-product of use of the platform as part of everyday life, rather than as an end in itself. We excluded tools like OpenStreetMap,[1] which have the express purpose of collecting geographic information, which users do actively and with awareness. We are interested in data that people contribute passively, without necessarily being aware that the data is being put to this use. This is because this is the data that has the potential to be harvested without requiring extra effort on the contributor's part. Since many active crowdsourcing approaches present participation challenges, this is not an issue with passive approaches, and we are thus interested in the options and opportunities for harvesting of data as a less labour-intensive method for capturing geographic information. In the typology of social media provided by Wendling et al. (2013) listed as follows, our scope is largely confined to the types 1, 2 and 4 of these,

[1] https://www.openstreetmap.org.

Table 1
Social media use by surveyed papers.

Social media platform	Total	Percentage	Years active[a]	Estimated Number of Users (millions)[1]
Twitter	445	54.2%	2006 to present	319 (monthly active users)
Flickr	167	20.3%	2004 to present	87 (registered users)
FourSquare	73	8.9%	2009 to present[b]	45 (registered users)
Instagram	27	3.3%	2010 to present	700 (registered users)
Sina Weibo	24	2.9%	2009 to present	361 (monthly active users)
Facebook	17	2.1%	2004 to present	2000 (monthly active users)
YouTube	15	1.8%	2005 to present	800 (monthly active users)
Gowalla	13	1.6%	2007 to 2012	0.6 (registered users)
Panoramio	10	1.2%	2005 to 2016[c]	4 (registered users)
Brightkite	8	1.0%	2007 to 2011	
Tencent Weibo	4	0.5%	2010 to present	469 (registered users)
Tumblr	4	0.5%	2007 to present	555 (monthly active users)
Google +	3	0.4%	2011 to present	540 (monthly active users)
LinkedIn	3	0.4%	2003 to present	106 (active users)
Whrrl	3	0.4%	2007 to 2011	0.3 (registered users)
MySpace	1	0.1%	2003 to present	50 (monthly active users)
Viddy/Supernova	1	0.1%	2011 to 2014	50 (registered users)
Vkontakte	1	0.1%	2006 to present	410 (registered users)
WeChat Moments	1	0.1%	2012 to present	963 (monthly active users for WeChat)
YikYak	1	0.1%	2013 to 2017	4 (monthly active users)
Total	**821**			

[a] Sources: Wikipedia; https://www.lifewire.com/viddy-app-for-iphone-3486482; https://techcrunch.com/2012/04/18/viddy-tops-app-store/; http://mashable.com/2010/09/09/emerging-social-business-platforms/#rnZZ5IVmGaqj.

[b] Check in functionality moved to Swarm in 2014, after which the app used multiple methods to try to track location automatically.

[c] Uploaded photo will remain on display until 12 months after closure in November 2016.

and includes any social media platforms within those categories, other than the above exclusions:

1. Social networking media, focussed around connecting people (e.g. Facebook, MySpace).
2. Content sharing media (e.g. YouTube, Flickr).
3. Collaborating knowledge sharing media like wiki's and podcasts.
4. Blogging social media for information sharing (e.g. Twitter).
5. Volunteer technology communities for mapping and crisis situations like Open Street Map, Ushahidi[2] and Sahana.[3]

We also include analysis of 21 review papers that were returned by the search (or, in 2 cases, that were referenced by other review papers), and papers included in the reference lists of all of the relevant papers, subject to the same exclusion criteria as the rest of the papers. We did not search for 'cited by' papers, only for those that were cited in each paper.

The beginning of the temporal period within scope was unlimited, as social media only became popular (at least under that term) in around 2005. The earliest paper uncovered in the search was published in 2006. The collection period ended on 19 July 2017. The full list of papers analysed is contained in a supplementary file to this paper.

4. Analysis - RQ1: Social media platform

4.1. Comparison of social media platform use

Firstly, we explore the distribution of research activities among different social media platforms. Table 1 shows the percentage of papers that used data from each of the social media platforms that were within scope. Some papers used more than one social media platform, with 821 social media uses by 690 papers. The percentages shown in this table thus reflect the proportion of overall uses of social media, rather than the proportion of papers using each specific social media. Table 1 also shows the years in which each social media platform was active, and the estimated number of users.

The most evident observation is that Twitter is by far the most dominant social media platform used for geographic data extraction in the papers surveyed, even though there are several other platforms that have more users globally. Papers do not typically give a reason for their selection of Twitter over other options. While Twitter can be accessed via convenient APIs, it presents particular challenges for textual information extraction, including the short format of Twitter messages (tweets). Hua et al. (2012) suggest that tweets are often written hurriedly, which leads them to be noisy, ungrammatical, and to contain errors in spelling, abbreviations and other complexities, all of which require adjustments to traditional NLP approaches like part-of-speech (POS) tagging and named entity recognition (NER). Many papers focus on demonstrating an effective method (for example, identifying relevant tweets by keyword search) rather than dealing with some of the more complex issues of tweet processing like abbreviations and spelling mistakes, and thus are likely to suffer from low recall (missing many messages that may be relevant).

On the other hand, the short format of tweets makes information more immediate, which is useful for analysis. Also, Twitter has unidirectional links that do not require approval by the person being followed, unlike many other social media (Hua et al., 2012), allowing a different kind of interaction analysis. The limitations of Twitter's streaming API, which provides a 1% sample of tweets, are also evident (Imran et al., 2015). One advantage of Twitter is that messages are geotagged, and have been for some time (in contrast to Facebook, for example, which enabled location-based tagging in 2010), but estimates of the rate of geotagged messages out of all messages vary from 0.8% (Musaev et al., 2015) to 6% (Chen et al., 2014b), or 14.7% in the context of earthquakes (Crooks et al., 2013). Craglia et al. (2012) also point out the potential unreliability of Twitter coordinates, given their dependence on hardware, software and user settings. The problem of determining whether the geotag indicates the location at which the event or phenomena described occurred (referred to as the geofocus in this paper), given that it may not be the same as the location from which it was posted, as well as the problems arising from messages that are sent after an event is observed (introducing a time lag) are also

[2] https://www.ushahidi.com.

[3] https://sahanafoundation.org.

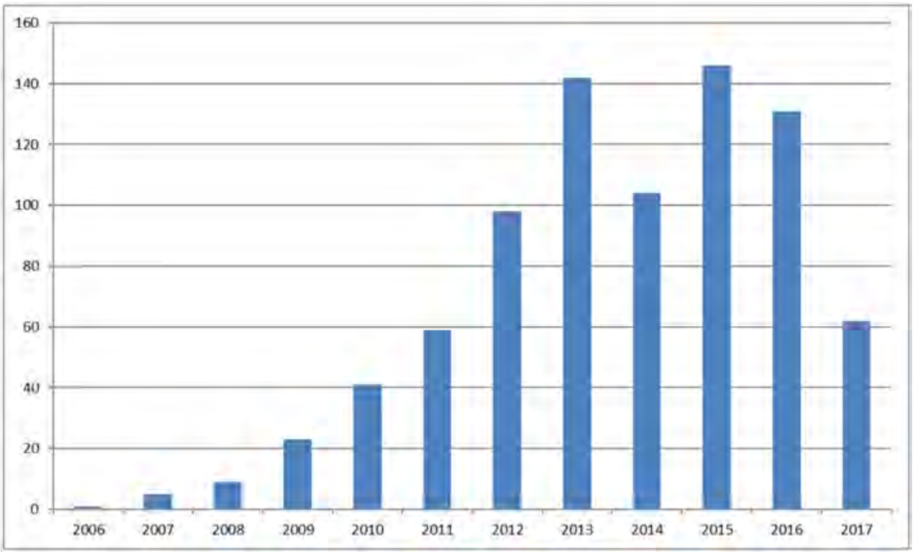

Fig. 2. Papers surveyed by year.

potential sources of inaccuracy. Finally, "From recent studies there is consensus that spam bots are prominent in Twitter data and can heavily skew analysis" (Benevenuto et al., 2010, no page numbers given).

The image-based content sharing platforms typically have higher rates of geo-tagging. For example, 80% of images in the now-defunct Google Panoramio were geotagged, and most Flickr photos are geotagged (Bae and Yun, 2017), although there is some disagreement regarding the latter, with Liu et al. (2014) calculating a geotag rate of 7.8% and Craglia et al. (2012) putting the figure closer to 20%. Figures for Instagram have ranged between 16% (Musaev et al., 2015) and 25% (Chandra et al., 2011). A geotagging rate of 6.4% has been suggested for YouTube (Musaev et al., 2015). While rates of geotagging for some of the content sharing social media platforms are higher than for Twitter, extraction of useful content from images and videos is easier for some application areas than others, and may require image processing, which can be a significant overhead. The extraction of information from tags from content sharing web sites allows the higher rates of geotagging in these sites to be taken advantage of.

Location-based sites (also known as location-based social networks or LBSNs) like Foursquare, Gowalla and Brightkite clearly focus on geotagging, but have smaller numbers of users and are restricted in scope and purpose (focussed around identified venues). They have been used frequently in proportion to their user numbers, presumably because of the ready access to location information due to their consideration of location as a first class citizen in such platforms.

The uneven use of social media platforms in extraction and analysis of geospatial information combined with their difference in nature suggests that current approaches may not be creating the most reliable data sets. For example, different platforms may be used for different purposes (e.g. Twitter for news vs. Instagram for daily life) (Xia et al., 2015) or have different levels of immediacy in reporting (Hyvärinen and Saltikoff, 2010). Silva et al. (2013) show that while Foursquare and Instagram provide similar results for some urban characteristics (e.g. population), they also show differences, with Foursquare being better at providing user route information, and Instagram providing a better picture of cultural behaviour of users. Simon et al. (2014) compare the use of Twitter and Facebook during the Westgate Mall Terror Attack in Kenya, showing that authorities used Twitter much more heavily than Facebook to provide updates to citizens, and had more followers in Twitter than in Facebook. Additional research is needed to identify the characteristics, strengths and weaknesses of the different social media platforms in particular situations and application areas.

It is also interesting to note that the research is focussed around a relatively small number of dominant platforms, and while the platforms studied include those that focus on particular geographic regions (e.g. Vkontakte and Sina Weibo) and particular media types (e.g. the YouTube and Viddy video sharing platforms), platforms focussing on particular segments of the population or interest groups were absent. A small volume of research has been conducted on such platforms, including Jack'd (Zhao et al., 2017) and Grindr (Roth, 2016) for the gay community, and while this work was not identified in our survey due to the specific search terms used and its relative sparsity at this point in time, it is an area that warrants increased attention in the research.

4.2. Comparison of social media platform use over time

Fig. 2 illustrates the spread of papers surveyed since 2006 (papers before that were in scope, but none were found). Note that the data was collected in July 2017, so the number of papers for 2017 might be expected to approximately double during the remainder of the year. This graph indicates that research in this field is no longer increasing in volume, and that there has been a slight decrease during the last two years. However, the quantity of research in this field is still substantial, and relatively stable since 2013, despite a noticeable dip in 2014.

Fig. 3 illustrates the use of the top seven social media platforms by year. We focus on the most frequently used platforms as the remainder were only addressed by 1–3 papers at most per year (and none in many years), so the influence of a single paper is much more significant and patterns are therefore less clear. The red lines on the horizontal axis show the years during which each platform was active. Flickr has been used consistently throughout the period, and although this may be due to its earlier launch, Flickr and Facebook were both launched in 2004 and YouTube in 2005 and the latter two did not show such early uptake. Only Foursquare showed similarly immediate use, while the other platforms took some years to be embraced by the research community studied in this paper. While Fig. 3 does not show any of the platforms that have ceased operation, Gowalla, Brightkite and Panoramio continue to be used at frequent levels, despite being officially disbanded. However, the numbers are relatively low (one or two papers per year), and are mainly due to researchers continuing to use previously extracted data.

The figures do not show a strong pattern of ebb and flow to reflect the popularity of particular platforms over time, with research attention being relatively consistent. Work on Sina Weibo has increased notably (given that the graph only includes data from the first half of 2017), but since it and also Vkontakte are largely used in non-English speaking

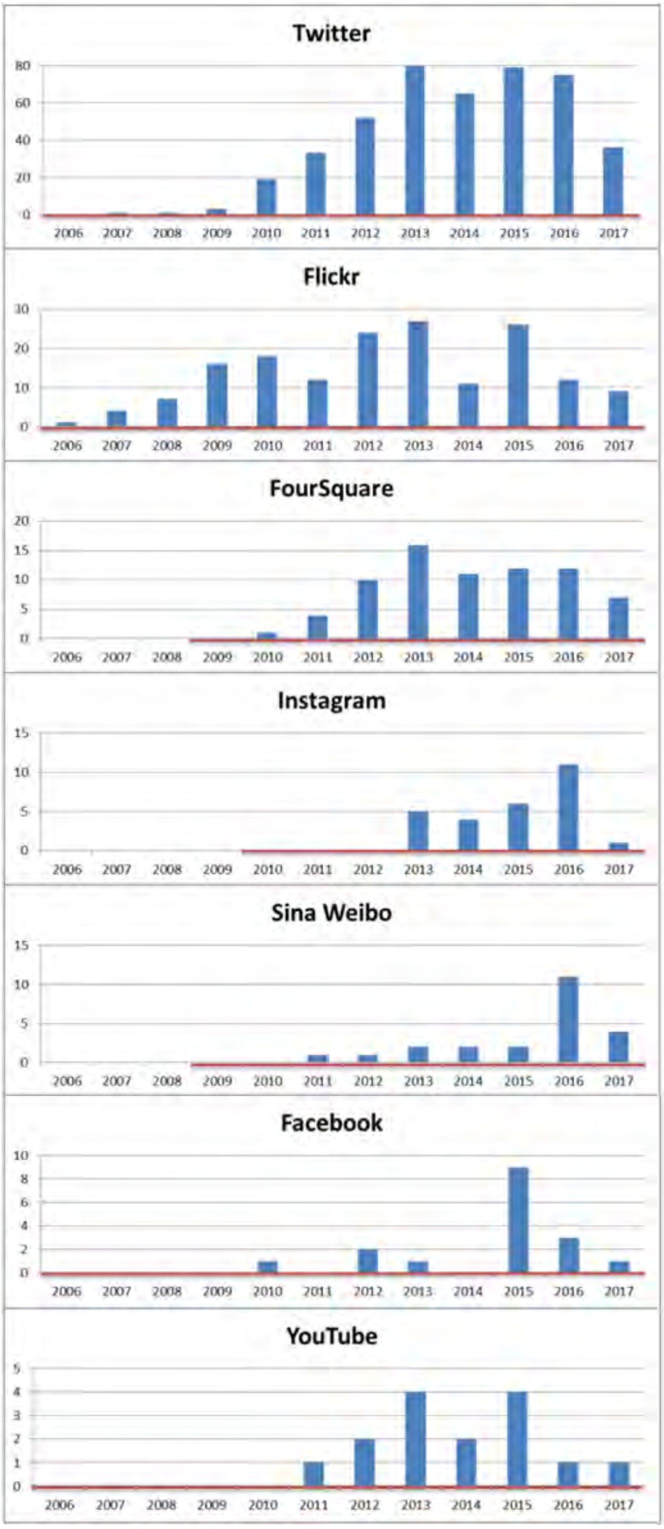

Fig. 3. Social Media platforms by year.

countries, these results may be influenced by variation rates of publication in English journals and conferences in those countries.

5. Analysis - RQ2: location extraction method

Table 2 provides a summary of the methods that have been used to extract location from social media, as well as the number of papers that have used each method across all social media platforms, using a typology that extends the classification provided by Ajao et al. (2015). The table shows that across the 690 papers that were surveyed, 1015 uses of a method were employed, since many papers used more than one method. As can be seen from Table 2 and the associated Fig. 4, use of geotags is by far the most popular approach, followed by extraction of place name from the user profile (UPC); extraction of place name from message content using Named Entity Recognition (NER) and use of location from Location Based Social Networks like Foursquare (LCC).

Figs. 5 and 6 break down the use of methods by year and social media platform respectively, with cells shaded according to the normalised quantity of papers using each method (number of papers using each platform or year are divided by the total across all methods), with darker shades indicating greater use. Both figures exclude platforms and years with fewer than 9 papers across all methods, as their inclusion would allow methods that were used for only one or two papers to appear to be more important than they are. The threshold of 9 was chosen because it was the point at which individual methods had very small numbers, and it represented a natural break in the data. In the case of social media platforms, 5 platforms had a total of 9–12 paper uses of different methods, after which the least frequently used 6 platforms had 4 or less papers across all methods. In the case of years, 2006 and 2007 amassed 1 and 7 uses of a method by a paper respectively, after which the number increased noticeably, so those two years were excluded.

These figures again emphasise the dominance of the four methods mentioned previously. MML (latitude and longitude geotags) is used across all social media platforms, particularly the content sharing methods, which is not surprising as alternatives like MCN (Named Entity Recognition of message content) and LLC (the use of LBSN) are more difficult for those platforms. Various approaches that analyse message content are used by many of the social media platforms (again, less so by content sharing and LBSN platforms), while methods that use the user profile mostly rely on the declaration of location by place name.

Turning now to the variations in usage by year shown in Fig. 6, greater use of a variety of methods can be seen in the earlier years of social media research, after which some approaches were abandoned or less frequently applied, with greater use of approaches that draw location from user profiles, and approaches that look at patterns among a combination of messages sent by a user. The pattern has changed little in the last few years, with clear dominance of the top four approaches and various methods of extraction from message content being employed.

Accuracy and coverage are particularly important aspects of the different methods, as they determine the scope within which each approach can be used and allow methods to be compared and evaluated against the requirements of a particular project. Appendix C summarises the accuracy and coverage figures provided across the approaches. However, many papers do not provide these figures and give little indication of accuracy achieved. Among those that do report accuracy figures, a number of challenges arise in interpreting and comparing results. Firstly, there are wide variations in the way that accuracy is reported. Many researchers report percentages of accuracy achieved at particular distances, others report Median Error Distance and others Average Error Distance. These three approaches are difficult to compare, and the first approach (percentage within a distance) often varies in the scale reported. While this makes sense because accuracy levels vary, it makes computational comparisons difficult. Furthermore, many authors report using distance figures that are not absolute, but from some specified area (e.g. distance from the user's home city, rather from their actual home location), or distance from a grid cell (e.g. Cheng et al., 2010; Chandra et al., 2011). In other cases, researchers report precision and recall values for a match in the top k values (k = 2, 3 or 5) rather than only the top value, and others only report a specific subset of the results (e.g. top 100 points of interest, users with a specified message frequency). It is also often difficult to identify the

Table 2
Summary of location extraction methods.

Source[a]	Source sub-type	Code	Total
Message metadata	Latitude and longitude (geotag)	MML	448
	Place name	MMP	30
	Time zone	MMT	3
	Description or title with Named Entity Recognition to select place name	MMD	2
User (or channel) profile	Location text field (often city name)	UPC	114
	Address	UPA	6
	Time zone	UPT	7
	Latitude and longitude	UPL	2
	IP address	UPP	4
	URL - location extracted from web page content	UPW	1
	URL - location of IP of URL	UPI	1
	URL - country of domain name in URL	UPD	2
	Manually set for known individual	UPM	1
	Previous towns of residence	UPN	1
	Place of previous or current employment	UPE	2
	Place of previous or current study	UPS	2
Social network	Inference of location from location of connections/mentions given in metadata or user profile	SNC	25
Message content	NER through POS tagging, gazetteers or combination (including manually)	MCN	122
	NER through lexico-syntactic rules (patterns, regular expressions etc.).	MLS	13
	Mining of place names using manually annotated training set	MCM	3
	Inference of location from interests and their mapping to POIs	MIN	3
	Inference of location from use of words that describe some located object (e.g. project, facility).	MPR	5
	Mining of spatial relation information from text.	MCS	6
	Inference from language, character set or language style.	MCL	1
	Machine learning from word use (including language models)	MCW	27
	Place name from URL in message	MUP	1
Message tags	Named Entity Recognition through POS tagging, gazetteers or a combination	MTN	20
	Machine learning from tag use (including language models)	MTW	15
Links to LBS	Inference of location from LBS	LLC	85
	Inference through similarity in image properties.	IMP	18
	Image-embedded geotag/geocode (including direction)	IMG	5
Video	Object recognition using geographic data	AVO	1
	ML from similarity in audio-visual features	AVS	1
Social media web site	Summary information on social media site (e.g. trends in cities).	SMW	1
Relationships among messages	Spatio-temporal distribution of messages (usually to infer home location etc.)	STD	27
	Distribution of messages that have similar image features.	STI	2
	Contextual information from other messages.	STO	1
	Orientation/direction of messages (images)	STP	1
	Learning from other messages by the same user that are geotagged	STL	2
	Inference of location from messages that are related in topic, event or time.	STR	1
	Sequence of messages in time	STS	2
Links to other Social media	Location adopted from repost or referencing of other social media.	LSM	1
Total			1015

[a] Adapted/extended from (Ajao et al., 2015).

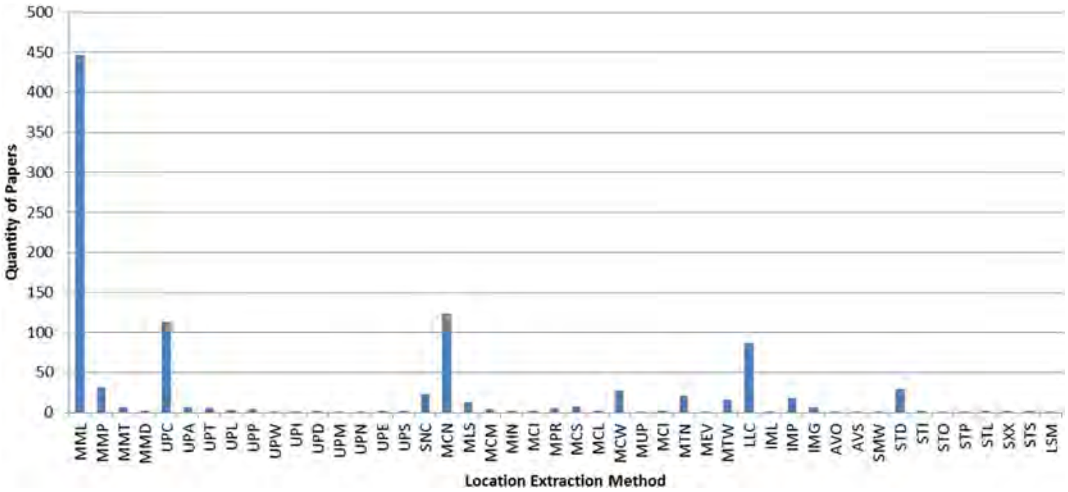

Fig. 4. Quantity of papers by location extraction method.

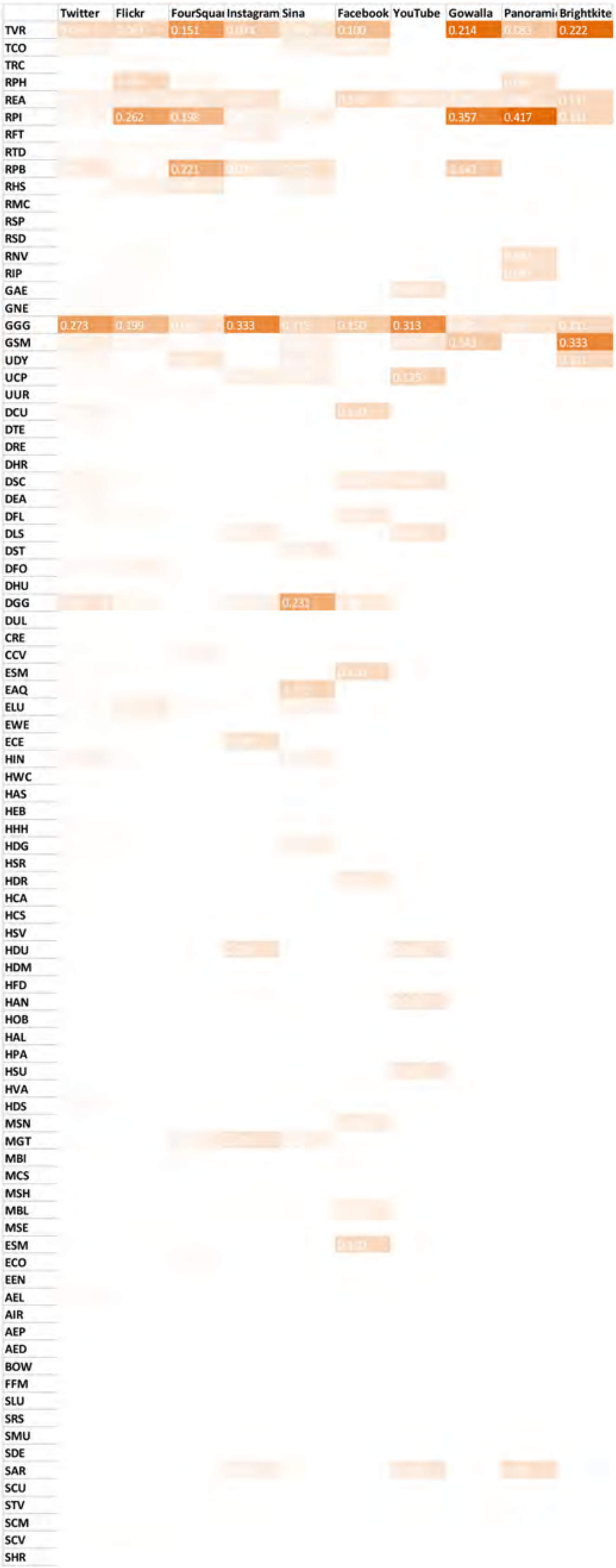

Fig. 5. Use of Location Extraction Methods by Social Media Platform.

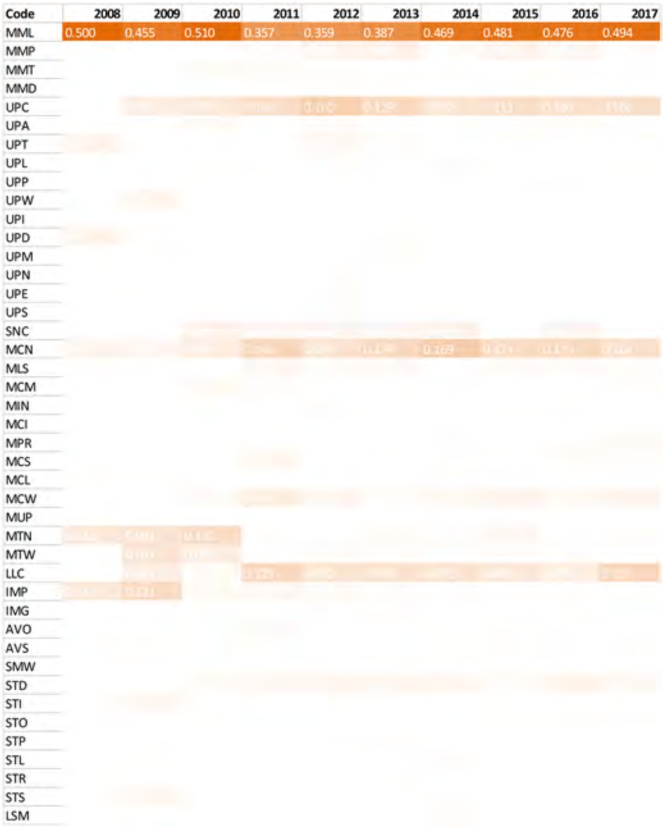

Fig. 6. Use of Location Extraction Methods by Year.

distance unit that is being reported (miles, metres, km), with this information often hidden in the text and excluded from tables and graphs.

Considering the accuracy of specific methods, accuracy information for the MML method is very limited, despite its widespread use. It is often used as a ground truth for other methods, but the accuracy of message GPS coordinates is far from assured, as Zhu et al. (2015) demonstrate in their study of Sina Weibo, in which an average error distance of 122 m was found. While this may be insignificant for some applications (e.g. thematic mapping of states or countries), it is important for many others. Papers that describe approaches that use the MCN method often report success rates for place names generally. However, place names vary widely in scale from country, state, city, neighbourhood down to individual point of interest at a very local scale). For this reason, in Appendix C, we only include papers that specify kinds of place names (e.g. country, building, street, city). Furthermore, in both MCN and MLS methods, a lot more attention has been given to identifying place names (e.g. through NER) than on actually geocoding those place names (which requires disambiguation), so accuracy figures often report the precision with which a place name can be successfully identified as a place name, rather than how frequently it can be successfully matched to the correct place on the ground.

Fig. 7 displays the accuracy in % within distance terms for a selection of methods that provided figures in that form. The vertical axis shows the percentage of messages that were within the distance specified on the horizontal axis. Each point (if only one percentage value is given) or line (if more than one percentage value is given) is labelled with the author, year, method code and first two letters of the social media platform used (e.g. Tw = Twitter; Fl = Flickr). Lines are always labelled at the end of the line. The points and lines are colour coded by method group/source (Column 1 in Table 2) (e.g. all of the methods that use message metadata are shown in blue), and the specific method can be determined using the method code, labelled on the group.

The most accurate methods appear in the top left of Fig. 7, and the

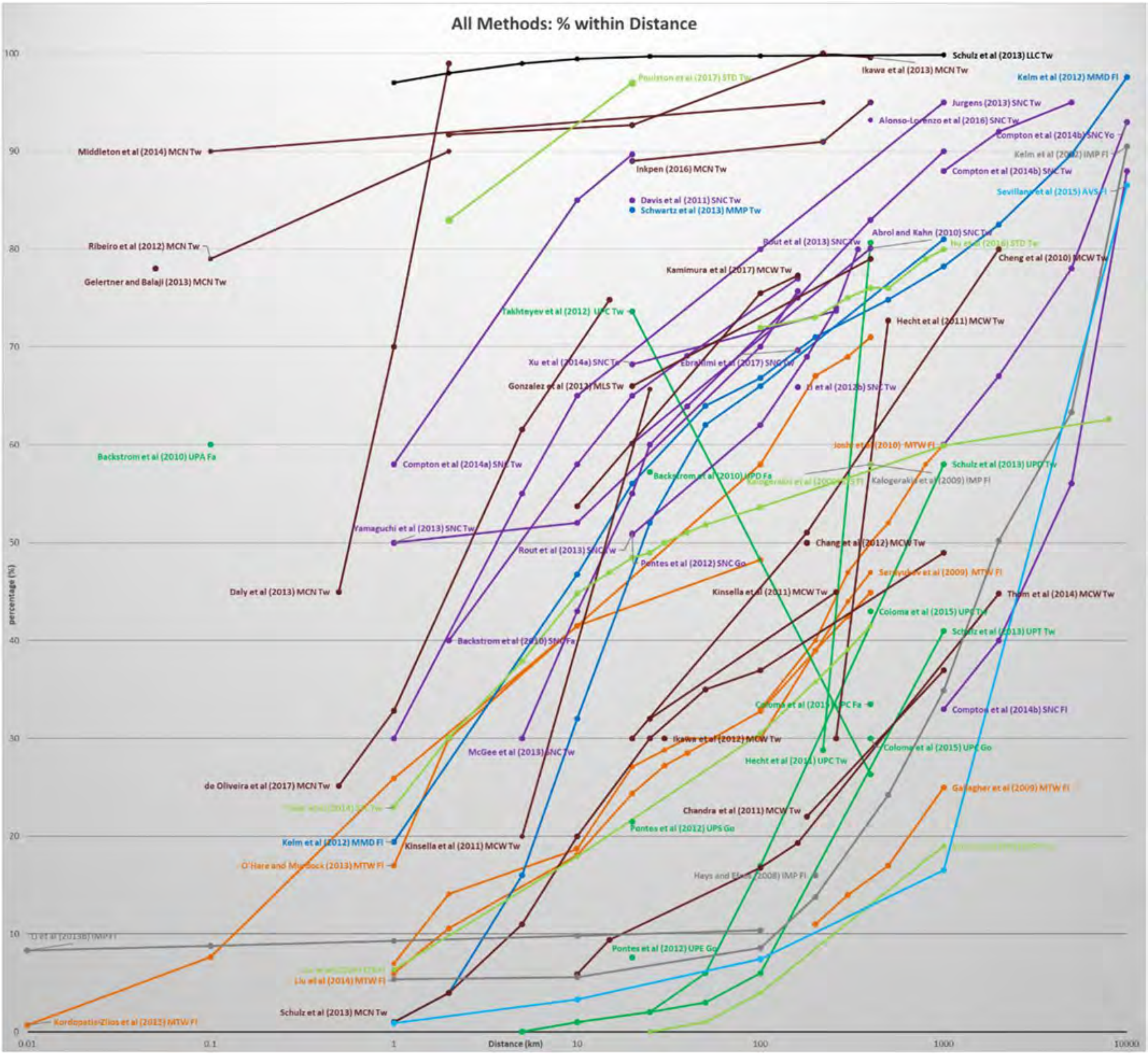

Fig. 7. Method accuracy.
Full references that correspond to the citations contained in Fig. 8 are contained in the supplementary file to this paper, containing references and abstracts for all of the 690 papers reviewed.

least accurate in the bottom right.[4] Clusters of method types can be seen on the graph. The most accurate work uses the MCN method, with Middleton et al. (2014) achieving 90% within 0.1 km, and several others with slightly lower accuracy. A number of other methods that use message content appear at other positions on the graph, representing lower accuracy, but nearly all of them use different methods, mostly MCW (language model approaches). The LLC approach described by Schulz et al. (2013) also has high accuracy, relative to other methods. The methods that use social networks (SNC) mostly appear around the middle of the graph, providing moderate accuracy. While lower accuracy is achieved by Compton, Keegan, and Xu (2014) using the SNC method, this is with Youtube and Flickr, rather than Twitter. A group of work using the MTW method (using tags) appears next in order of accuracy, and mostly uses data from the Flickr platform. Towards the bottom right of the graph, the methods that use user profile information are dominant, along with methods that use image properties (although the latter are few in number).

While an inverse relationship between coverage and accuracy might be expected, the figures are not sufficiently complete or consistent to support this. Coverage figures for the MCN method, which provides high accuracy relative to the other methods, range from 5.1% to 83%, but authors do not all report on the same thing. For a full picture of coverage of this method, we need to know how many messages that describe a location include a location name; of those, how many can have names successfully extracted; and of those, how many can be successfully matched to a location. Reporting of these aspects varies widely, and research evaluating the coverage levels at each of these stages is limited, although Schulz et al. (2013) provides recall figures

[4] The counter-intuitive line for Takhteyev et al. (2012) UPC Tw adopts two different strategies, one at country level and one at city level, hence the apparent reduction in accuracy.

for a range of MCN approaches. The SNC method, with moderate accuracy, reports coverage in the region of 71–82%, but often the highest accuracy is achieved for users who send frequent messages, or who have a lot of connections with other users. Coverage of the methods that use user profile is also variable. UPC (place name in user profile), which is the most common method, ranges in coverage from 58 to 66%. Levels of coverage and the relationship between coverage and accuracy is an under-researched area, and a more detailed examination, particularly for methods like MCN with good accuracy, is warranted. In contrast to some other approaches, the coverage of the MML method has been widely reported, mainly because it is so low, and is used to justify research into other methods.

The following Sections explain each of the methods and the ways they have been employed in different papers in more detail. We use the classification scheme presented in Table 2, and within each category of approaches, explore the individual approach, reviewing relevant literature and the main trends.

5.1. Approaches that extract location information from message metadata

5.1.1. Latitude and longitude geotagging (MML)

The use of geotagging to extract location is clearly the dominant approach, although a number of the papers that used this approach use it as ground truth to compare alternative methods for identifying location, rather than as a method in its own right. The approach has the advantage that location is easy to extract, involving simply accessing coordinates via APIs for most social media platforms. Some researchers then perform reverse geocoding to identify place names that correspond to the location using tools such as the Yahoo Place Finder and Google Maps API (e.g. Ardon et al., 2013; AlBanna et al., 2016; Becker et al., 2010; Castro et al., 2017).

However, the use of geotagging as a location method is not without problems. The low rate of geotagging of messages on platforms like Twitter has already been discussed in Section 3. Users are often justifiably concerned about privacy issues in making their accurate location known, which can lead to bias, particular as there is evidence that males share their location more frequently than females (O'Hare and Murdock, 2012). Another problem for some applications is that the geotag shows the location where the message was written, which is not necessarily the location that the message is about (the geofocus) for a number of reasons. This may be particularly problematic in applications such as disaster management, in which users commonly report locations that they observe (in the present or the past) or hear about. There is also evidence that some users geotag at low resolution (e.g. city or country level) and deliberately introduce incorrect information (Senaratne et al., 2017).

It is difficult to judge the location accuracy of the MML approach to location extraction, as it depends in part of the approach used by the social media platform. While coordinates may come from GPS on the user's phone or content embedded within the metadata of an uploaded photo, some APIs populate coordinate information from other sources, including geocoding of place names in the user profile. Zhu et al. (2015) establish mean average error of 122.14 m in Sina Weibo, while accuracy measures in Flickr range from 1 (world) to 16 (street), with suggestions that the mean accuracy is 14.58 (Straumann et al., 2014). Approaches to address the lack of accuracy exhibited by the MML method are limited in the literature, and many researchers assume that it provides high accuracy, often using it as a ground truth for other methods. The weight of evidence of many different message locations about the same topic is often used to improve reliability (for example, messages clustered around a point of interest in tourism applications) (e.g. Bui et al., 2016).

5.1.2. Place name (MMP)

Some of the platform APIs provide a message location using place name, rather than or as well as coordinates, including Twitter (Eo et al., 2016). The content of this metadata may come from various sources, sometimes explicitly entered by the user through typing in a place name or marking their location on a map, and sometimes automatically determined using device GPS, IP address with reverse geocoding or triangulation from the cellular network.

MMP is almost always used in conjunction with another method, like MML or UPC, in order to address the low level of provision of more precise coordinates. Some processing is then required to tokenise and match the provided string to a known place name, and remove the non-place name contents that are sometimes provided (users sometimes provide location descriptions such as ‘at home’). Once the place name has been extracted, it is frequently verified or geolocated via look up in a gazetteer, or using a third party service like Google Map's geocoding service (e.g. Lee et al., 2013a).

Attempts to establish the quality of MMP location information are rare, and place name is most frequently provided at city level. Schwartz et al. (2013) compare the accuracy of the values provided to coordinates for Twitter messages that contained both, gaining a correspondence of 87–93%. Coverage is a concern, ranging from 1% to 65% (Bouillot et al., 2012; Dredze et al., 2013; Croitoru et al., 2013).

5.1.3. Time zone (MMT) and message description or title (MMD)

Message time zone is very infrequently used as a location mechanism, and provides very low accuracy as time zones are very large (Mendoza et al., 2010; Lee, 2012). Approaches that use message description or title include Kelm et al. (2012), who study data from Flickr, extracting description, title and keywords, then performing Named Entity Recognition (NER), and Krauss et al. (2015), who extract location from description metadata for YouTube videos.

5.2. Approaches that extract location information from user profile

5.2.1. Location (UPC)

The location specified in the user profile is frequently used to establish the location of messages, particularly in Twitter, where it is often combined with MML to deal with the low rate of full geocoding of messages on that platform, and the lower accuracy of the UPC method. As with MMD, UPC is often combined with other services such as Google Maps geocoding service to generate coordinates from a place name or venue (e.g. Baucom et al., 2013). The use of non-locatable locations (e.g. the universe, the Internet) and slang was also common (Hecht et al., 2011).

The user profile location typically has low accuracy, usually providing information at city or state level (for example, Conover et al., 2013), and coverage of genuine entries (as many users provide invalid place names) in Twitter ranges from 46% to 77% (Hecht et al., 2011; Takhteyev et al., 2012; Kanta et al., 2012; Daume, 2016). The approach has also been used for Facebook and Google +, with coverage ranging from 17% to 50% with very small samples (Coloma et al., 2015), and mostly at country or global region level (Fire and Puzis, 2016; Oksanen et al., 2015).

5.2.2. Approaches that use other information from the user profile

The address field contained in a user profile (UPA method) can potentially provide relatively accurate location information for the user, and while this does not give the location of the post or the topic of the post (the geofocus), it may be useful for some applications. Full addresses were provided in 6% of Facebook user profiles, and of those, 60% could be parsed and matched to latitude and longitude, with more males than females providing this detail (Backstrom et al., 2010). Researchers have also used address from Twitter (Odlum and Yoon, 2015; Mirani and Sasi, 2016), Google+ (0.07% of users provided an address) and Foursquare (for venue locations) (Pontes et al., 2012).

User time zone (established from the user profile) (UPT method) has only been used for the Twitter platform in the papers surveyed. Burton et al. (2012) identifies time zone as the most widely used of all location methods in Twitter, with 77% of messages having a connected user profile that specified a time zone. Kulshrestha et al. (2012) found time zone location in 37.3% of users, with no indication of the proportion of messages to which this corresponds. Bouillot et al. (2012) point out that in Twitter, time zone is shown as the main city in the country from which the message is sent, and therefore provides more location information than just the time zone. Time zone is also commonly used to disambiguate place names, either in the message (MMP) or user profile (UPC) (Schulz et al., 2013; Ghosh and Guha, 2013; Tang et al., 2015).

Extraction of location via user latitude and longitude (UPL method) was also confined to the Twitter platform in the papers included in this study. Examples include Nagar et al. (2014), who follow extraction of message latitude and longitude with user profile latitude and longitude and Takhteyev et al. (2012), who find latitude and longitude in 7.5% of user profiles, apparently in many cases through automated provision of certain applications. IP address (UPP method) has been used as a source of location for data from Twitter (Li, Qian, et al., 2013), Flickr (Odlum and Yoon, 2015; Van Zwol, 2007) and WeChat Moments platforms (Jiang et al., 2016), with Twitter adopting this approach to reverse geocode to populate the place name.

The URL contained in a user profile has been used to identify location in various ways, including via a location extracted from text in the web page at the URL (Cheong and Lee, 2009) (UPW method); the IP address of the URL (Schulz et al., 2013) (UPI method) and the country code included in the URL domain name (Schulz et al., 2013) (UPD method), all in Twitter. All of these methods offer poor spatial accuracy, not unexpectedly, but Krishnamurthy et al. (2008) finds a general correspondence between time zone and URL, with around two thirds of users providing a URL.

Methods that derive location from information about previous and current home (UPN method), employment (UPE method) or study (UPS method) have been employed on Google+ and LinkedIn. For example, Pontes et al. (2012) extract previous residences, addresses, employment and study locations from Google+, along with other information, to infer home location, and while coverage is high, spatial accuracy is poor. Yoon et al. (2015) use data from LinkedIn to infer frequency of relocations in a person's lifetime, and match it to medical records extracted from medical networks and blogs to establish a correlation between lung cancer risk and relocation (providing a good illustration of the privacy risks of social media).

5.3. Approaches that use social networks to extract location (SNC)

While the most accurate of the metadata-based methods (MML) provide high accuracy and low coverage, and the best of the user profile based methods (UPC) provide low accuracy and moderate coverage, the SNC method has the benefit of providing moderate accuracy and potentially complete coverage. The SNC approaches infer a user or message location from the locations of connections provided by the social media platforms (e.g. friends, connections, mentions, retweets depending on the platform). This approach was first used in 2010, and has continued to be developed by researchers since then. Use of the approach is spread across many of the social media platforms.

Appendix A summarises some of the key surveyed papers, the methods used and the accuracy achieved. A range of methods have been used, including label propagation (Xu, Cui, et al., 2014; Yuan et al., 2016; Ebrahimi et al., 2017); clustering (Williams, 2016; Ebrahimi et al., 2017) and optimisation (Li, Wang, et al., 2012; Compton et al., 2014), among others. Work with Twitter most frequently relies on followed and following connections (both unidirectional and reciprocal), although sometimes mention connections are used (e.g. Compton et al., 2014; Ebrahimi et al., 2017). McGee et al. (2013) evaluates the importance of different kinds of connections in Twitter, finding that followed/following connections are more closely correlated with spatial proximity than other types of connections, with mentions being worst, but Di Rocco et al. (2016) suggest that mentions connections may be more indicative of message location, while followed/following connections better reflect home location. Other platforms (e.g. Facebook and Google+) focus on friend connections.

A number of additional factors are incorporated into the SNC models in order to improve results, including degree of friendship overlap and triadic friendships (investigating the hypothesis that the location of common friends might be a better indication of location than individual friends) (e.g. Abrol and Khan, 2010); presence of recent communication (e.g. Backstrom et al., 2010); geographic distance between friends (e.g. Backstrom et al., 2010) or similarity in tweet vocabulary (Sadilek et al., 2012a). Attention has been given to the exclusion of celebrity and news connections as poor indicators of location, and the idea that users with particularly low or high numbers of connections might distort the model.

While most researchers are able to achieve 100% coverage with their approaches, the accuracy achieved depends on the quantity and quality of the initial data in the seed models on which the propagation or optimisation depends. Researchers typically begin with data populated from MML, UPC and MCN approaches, and then use SNC to fill in gaps to improve the quality of the data. Furthermore, the large difference between median and average error distances in most cases suggests a significant number of outliers.

5.4. Approaches that use message content

A number of methods have been developed that use various components of message content in different ways to locate a message. They have the advantage of being able to establish the geofocus of a message rather than the location at which it was posted, and of having good coverage, and some of the methods have relatively good accuracy.

5.4.1. Place names in message text (NER with gazetteers) (MCN)

MCN is the second most common method for extracting location from social media, after MML (message geotagging). Other than one or two earlier uses, it has mainly been used since 2011 (see Fig. 5). It is used across most platforms other than the LBSNs (Foursquare, Gowalla, Brightkite), and is particularly popular for those platforms that do not include built-in geolocation methods, like YouTube and Tumblr. MCN has the advantage that it can be used for any text-based message (and most of the content-sharing social media sites enable text descriptions as well as photos, videos, etc.), and can be used to determine geofocus. The issue of geofocus vs. message posting location vs. user home location has been investigated by de Oliveira et al. (2017), who find the median distance between MML and MCN locations to be 2.48 km, with 25% < 0.5 km. The distance between home and mentioned locations was larger (median 5.64 km). Like UPC, MCN is frequently used to augment the low rate of coverage of MML. However, extraction of place name from text is more challenging than simply reading coordinates, and is not always successful due to ambiguities in place name and difficulties in text processing. Furthermore, it cannot provide the same level of spatial accuracy as MML.

The broad approach can be divided into two stages, and many researchers address only one of the two. The first stage involves identifying place names from among other text using Named Entity Recognition (NER), which identifies proper nouns, often using a gazetteer for place name look up to determine whether a noun identified through part-of-speech tagging refers to a place (as opposed to a person, organisation, etc.). NER tools used include the Stanford NER tool (Bassi et al., 2016; Gelernter and Mushegian, 2011; Li et al., 2015), which comparisons show has a better success rate than many alternatives (Lingad et al., 2013); GATE (Jaiswal et al., 2013) and OpenCalais (Gelernter and Balaji, 2013). Some researchers create their own, handcrafted gazetteers (Jaiswal et al., 2013; Sultanik and Fink, 2012; Ribeiro et al., 2012) that may be optimised to improve identification of place names by adding street name indicators like 'road' and 'street' (Jaiswal et al., 2013) or excluding certain locations (Inkpen, 2016). Commonly used gazetteers include GeoNames (Inkpen, 2016; Zhang and Gelernter, 2014; Ikawa et al., 2013), OpenStreetMap (Daly et al., 2013; Di Rocco et al., 2016), and the USGS Gazetteer (Bassi et al., 2016). Methods used include Conditional Random Fields by Stanford's NER, artificial neural networks (Inkpen, 2016); fuzzy matching of text using a phonetic encoding algorithm (Sultanik and Fink, 2012) and the C4.5 decision tree algorithm to handle abbreviations (Gelertner and Balaji, 2013). Success rates are widely variable, reported in different ways and often restricted in scope, and are thus difficult to compare, but many approaches manage to achieve precision and recall for place names in the 80s and sometimes 90s, as shown in Appendix C and Fig. 7.

The second stage in the process of extracting location is less prolifically addressed than the first stage, and involves finding the coordinates for the place name mentioned, referred to as disambiguation or grounding, as many place names are duplicated both within countries and globally. A number of methods are used to try to disambiguate place names, including weighting by population, geographic feature types, geographical proximity and other place names that are found nearby in the text (Zhang and Gelertner, 2014; Inkpen, 2016). Success rates for this stage range from 39% accuracy and 78% recall (Li et al., 2015), 84% and 83% for precision and recall respectively (Zhang and Gelertner, 2014) to 98% (Inkpen, 2016). Some researchers working in a limited geographical area approach the problem from the other direction, starting with a list of place names and looking for matching text for those place names (e.g. Daly et al., 2013; Bahir and Peled, 2016), in which case disambiguation is not usually required. With this approach, Daly et al. (2013) achieve spatial accuracy of 500 m (median error distance), with 100% within 2 km.

5.4.2. Place names in message text (NER with Lexico-syntactic rules) (MLS)

Approaches that identify the location of a message from place names contained in the message by relying on lexico-syntactic rules, rather than (or in addition to) gazetteers, are much less common. Such approaches have been used predominantly with Twitter, and in two cases, with Sina Weibo. The first surveyed use was in 2011, and has continued steadily since, and this approach has been applied in a range of languages, including English, Chinese, Indonesian, Japanese and Turkish.

Rule-based approaches have been applied to identify street names in a number of cases (e.g. Gelernter and Balaji, 2013; Gu et al., 2016; Hennig et al., 2016), to take advantage of the common format adopted by street names. Other kinds of location information extracted include journey origin and destination (Endarnoto et al., 2011); points of interest (Gu et al., 2016) and spatial relations (Zhang et al., 2017).

In some cases, rules are hand-crafted (e.g. Endarnoto et al., 2011; Jaiswal et al., 2013), while in others they are automatically mined (e.g. Gonzalez et al., 2012). Rule-based approaches are often combined with other approaches, including gazetteers (Paradesi, 2011; Zhang et al., 2017) and as features in a conditional random field model (Rao et al., 2016; Sagcan and Karagoz, 2015). The approach is commonly employed to assist in identification of place names following NER and sometimes gazetteer use, by looking for patterns of language within which place names frequently occur. For example, Sakaki et al. (2012) identify verb-preposition pairs that are commonly followed by a place name; Paradesi (2011) looks for 'spatial indicators' that commonly precede a noun and Joseph et al. (2015) look for nouns followed by one of a certain set of spatial prepositions. Zhang et al. (2017) also use this approach to extract non-toponym location information in its own right, rather than as a mechanism to identify toponyms. They use regular expressions to extract spatial relations like "40 km south-east of" from Chinese messages in Sina Weibo.

Accuracy and coverage information is difficult to establish for this method, as it is often employed as part of a suite of other methods, and accuracy is not commonly reported for this approach alone, but figures in the 80s for successfully identifying place names (excluding the disambiguation stage) have been achieved (Gelernter and Balaji, 2013; Gonzalez et al., 2012). Sagcan and Karagoz (2015) compared several different approaches with the rule-based approach combined with others, with the best precision being achieved by the use of POS tags, suffixes and combinations of n-grams.

5.4.3. Mining of place names using a manually annotated training set (MCM)

Methods that fall into this category manually annotate locations in various ways, and then apply machine learning approaches. There are some overlaps with the MLS approach, in that some MLS papers apply Conditional Random Fields using features that are defined using lexico-syntactic rules. The approaches in the MCM category are simpler and based on manual annotations rather than rules, and include papers that explore the effectiveness of manual annotation as a method (Fersini et al., 2017). The best approaches for manual annotation of locations has also been investigated by Finin et al. (2010), who compare the use of Mechanical Turk and CrowdFlower, and Gelernter and Mushegian (2011), who compare adjudicated annotations to individual annotations.

5.4.4. Inference of location from additional information user interests

The survey included three papers that presented approaches to location extraction that were dependent on user interests (MIN), the first of which was published in 2012. Chen et al. (2013a) model users' interests from their messages in Sina Weibo, then connect those interests to functions that might indicate location, and from that, estimate location. Their method relies on finding a point of interest with a given function in close proximity to the user, and can achieve good accuracy (0.734 km average error distance) in areas close to points of interest. Yuan et al.'s (2016) method has some similarities, but combines the MIN approach with SNC and extracts user interests from tweets using Latent Dirichlet Allocation. Dalvi et al. (2012) take a different approach, inferring the user's location from the location of the places the user tweets about. Their approach combines the use of language models (see method MCW) and the distance to places that fall within the user's area of interest.

In a similar direction, the MPR approach involves searching for words that indicate a particular facility or project that can be used to infer location. Examples include Putri et al. (2016), who use a keyword-based method to search for messages that have a public facility in their content, and Lei and Hilton (2013), who search for keywords that indicate a specific project. Li and Sun (2014, 2017) use points of interest (POIs) to identify locations by searching for the POIs in Foursquare, and extracting location from there. Poulston et al. (2017) use a range of text patterns that might indicate that the user is at home (e.g. 'at home'), and infer home location by searching for other location information from those tweets.

5.4.5. Mining of spatial relation information from text (MCS)

Much of the work that extracts location from message content that has been discussed so far in this paper focusses on extracting location in the form of place names, whether directly or through an intermediate step like user interests, functions or projects. However, locations are often described in more complex ways, either by reference to another location (e.g. *near the Waikato River; beside St Andrews Church*). Spatial relations are a key element of these descriptions, and several of the surveyed papers address the problem of extracting spatial relations from message content, from languages including English, Thai and Chinese. The approaches used to explore spatial relations are mostly fairly simple, adopting regular expressions (e.g. Zhang et al., 2017), a very restricted set of spatial relations (e.g. Bahir and Peled, 2016) and/or a limited range of variations in language structure (e.g. Bassi et al., 2016; Paradesi, 2011; Wanichayapong et al., 2011). Dittrich et al. (2015) is an exception to these limited approaches, distinguishing spatial prepositions from non-spatial, beginning with a set of prepositions that can be used spatially, and identifying a number of rules that can be used to indicate a non-spatial use. Bahir and Peled (2016) also point out that spatial location information may be implicit in phrases such as 'I see…', 'I hear…' or 'I smell…'. Bassi et al.'s (2016) approach focusses mostly on the use of distances and cardinal directions, with which they adjust coordinates determined by place name.

5.4.6. Machine learning from word use (including language and language-derived topic models) (MCW)

The MCW approaches use language models (or topic models derived from language used in messages) to reflect the range and quantity of words used in messages, with the idea that this would vary by location, and could thus be used to extract location information. The reasons given for the expected variation range from the use of location-specific words (e.g. venue names) to differences in dialect or language style in particular locations. Textual variations like abbreviations and spelling mistakes may cause particular challenges for this approach. MCW is the second most common approach of those that use message content to extract location, after MCN. The papers vary in the algorithms used to infer location from the language models; the range of words included in the model, and the inclusion of other elements in addition to specific words, and are summarised in Appendix B. Several approaches select only local, hyperlocal or semi-local words to build their models (e.g. Cheng et al., 2010; Ryoo and Moon, 2014), while others focus on topics derived from terms, most commonly using Latent Dirichlet Allocation (LDA) (e.g. Yuan et al., 2016) with variations (e.g. Tigunova et al., 2015; Lozano et al., 2017). Other work uses n-grams and collocated terms (Flatow et al., 2015; Ishida, 2015). Various methods are applied to segmentation of geographic space, allowing geographically limited language models to be built, including regular grids, adaptive quadtrees and uniform geometric decomposition (Kamimura et al., 2017; Thom et al., 2014).

5.4.7. Other approaches that use message content

Far less frequently used approaches include MCL, which infers location from language style or character set (for example, Cheng and Chen (2014) used the ISO language setting used by Twitter to determine whether they were part of the Taiwanese (who use Traditional Chinese) or mainland Chinese (who use Simplified Chinese) communities) and MUP, which extracts place name from the web page at URLs included in the message, and is similar to the UPW methods, except that the UPW method uses the URL from the user profile, rather than the message content (Wang et al., 2015).

5.5. Approaches that use message tags

Message tags are another common source of location information, particularly in the absence of other alternatives, as they are less complex than message content. Two broad approaches were used in the papers surveyed.

5.5.1. Place names in message text (NER with gazetteers) (MTN)

The MTN approach is very similar to the MCN approach, except that it extracts place names from tags attached to the message, rather than the message content itself. The approach has been used regularly since 2007, and although it is used with Facebook, Tumblr, Twitter and Foursquare in the papers surveyed, its heaviest use is with Flickr (15 of the 21 papers that use this method). Several of the papers, and most of those that use the method with text-based platforms such as Twitter, combine the method with MCN, using similar approaches (e.g. McClendon and Robinson, 2013; Xu, Lu, et al., 2014).

The MTN approach is most commonly used as part of a larger task, and is often driven by place names rather than designed to work generically. That is, the researchers start with a list of place names (cities, states, etc.) in which they are interested, and search for messages with tags containing those place names (e.g. Abbasi et al., 2009; Cheong and Lee, 2010; Xu, Lu, et al., 2014; Mendoza et al., 2010), sometimes performing further filtering (e.g. De Choudhury et al., 2010). Spatial accuracy is not often reported, as this task is part of a larger chain of activities. Place name tags have also been combined with other location methods (most commonly MML) to delineate the extents of a place name using methods such as kernel density estimation, support vector machines, k-means clustering and Delauney triangulation (e.g. Grothe and Schaab, 2009; Hollenstein and Purves, 2010; Keßler et al., 2009; Lee et al., 2008). Some of this work models the distribution of tags to determine whether they are locally specific and therefore may be useful in language models (e.g. Liang et al., 2010; Wen et al., 2015; Wen et al., 2017).

5.5.2. Machine learning from word use (including language and language-derived topic models) (MTW)

The MTW approach is very similar to the MCW approach, except that tag contents are used instead of message contents. A similar usage pattern also emerges for this approach, which has been used since 2009 with Twitter, Foursquare and Instagram, but by far the most frequently, with Flickr. Language models that incorporate tags have used approaches including a nested Chinese Restaurant Franchise (nCRF) statistical model (Ahmed et al., 2013); support vector machines (Crandall et al., 2009); Naïve Bayes (Joshi et al., 2010); maximum likelihood estimation (O'Hare and Murdock, 2013) and multinomial distributions (Serdyukov et al., 2009; Zhang et al., 2016). A number of approaches apply weights to the tags using methods such as spatial entropy (Kordopatis-Zilos et al. (2015), and other measures such as smoothing (O'Hare and Murdock, 2013; Serdyukov et al., 2009) and boosting by increasing weights for location-specific names (Serdyukov et al., 2009) are also applied. Clustering approaches are also frequently used, in some cases to cluster together photos with similar tag sets to identify landmarks (e.g. Gao et al., 2010) or events (e.g. Ranneries et al., 2016). As with MCW, these approaches often adopt a grid structure in order to handle the variable distribution of tags in geographic space (e.g.

Serdyukov et al., 2009; Kordopatis-Zilos et al., 2015; O'Hare and Murdock, 2013). Median error distances achieved range from 60 to 200 km. Much better accuracy has been achieved by combining MML and MTW approaches (e.g. median error distance of 0.6 km by Van Laere et al., 2010).

5.6. Approaches that use location based services (LLC)

Location-based services such as Foursquare, Gowalla and Brightkite are regularly used to infer location for messages in social media through the location of venues that is a core part of those platforms. This approach is used both directly, in which messages are sent from a venue using LBSN platforms, or indirectly, in which LBSN venue check-ins are included in other social media platforms like Twitter. The LLC approach has been used regularly since 2009 and most frequently in Foursquare, but also by Twitter and many other platforms.

LBSN are used in several different ways by other social media platforms. The first of these directly uses LBSN location from user check-ins in the LBSN platform (e.g. Cheng et al., 2012; Chiang et al., 2014; Cho et al., 2011; Fang and Dai, 2016; Gao et al., 2013; Le et al., 2014), often using this information to make inferences or predictions about user behaviour. A second approach exploits links between LBSN and other platforms, in which another platform like Twitter can be used to post LBSN check ins, and this may be an easier way to access check in information due to Twitter's APIs. For example, a number of papers use this approach to study user behaviour (e.g. Gao et al., 2012). Check ins have also been used to identify the user's home location (e.g. Cheng et al., 2011) and to identify missing data in travel trajectories (e.g. Hasan and Ukkusuri, 2017). A third approach involves using the LBSN effectively as a gazetteer to identify coordinates for a place name included in a message or tag in another platform. For example, Li and Sun (2014, 2017) create a point of interest inventory from Foursquare, including both official names and colloquial and abbreviated names, and processing of tweets involves searching the POI inventory to identify venue names. A fourth approach does the reverse of the third, associating venues with a coordinate location determined using another method, like MML. For example, Chen et al. (2014b) identify clusters of Twitter geotags around a given Foursquare venue. A fifth broad approach is used less frequently, but there is some work that looks for text in a social media platform like Twitter that conforms to a particular syntactic format (e.g. “I'm at…”) and thus can be assumed to include a place name reference to a venue or location. The coordinates of the venue can then be retrieved from the LBSN or using other NER approaches (e.g. Sanborn et al., 2015). This approach has also been used to extract richer information. For example, Grinberg et al. (2013) study the correlation between Foursquare check ins and Twitter messages to identify language that is used to describe certain kinds of activities (e.g. shopping, nightlife).

Information on the spatial accuracy of the LLC approach is very limited, and depends largely on the method used to locate a given venue. For example, Foursquare venues may be created by dropping a pin on a map, and while current versions of Foursquare eliminate check-ins and adopt a reverse approach of determining user location from location based services on the user devices (Heath, 2016), previous versions have used user check in locations to determine venue location (Jeffries, 2012).

5.7. Approaches that use images

5.7.1. Inference through similarity in image properties (IMP)

A number of papers use image content to assist in the process of location determination. The most common approach compares image properties to look for similar images and infer locations from those that are already geotagged (IMP). It has been used regularly since 2007, mainly with Flickr, but less frequently with Panoramio and Tencent Weibo. The task of estimating location of photos and/or videos using their content and metadata has been the subject of the MediaEval Placing task over several years, and many of the surveyed papers were a response to this challenge.

A range of approaches are used to determine image similarity, including SIFT (Crandall et al., 2009; Cristani et al., 2008; Kawakubo and Yanai, 2011; Kennedy et al., 2007), tiny images, colour histograms, GIST (Ji et al., 2011), bags of textons (Gallagher et al., 2009; Hays and Efros, 2008; Joshi et al., 2012); texton histograms and straight line statistics Kalogerakis et al., 2009), with many researchers adopting several approaches. Clustering and ranking are also commonly used (e.g. Li et al., 2009; Liang et al., 2010), for example, to give greater weight to more salient image features (Li, Larson, and Hanjalic, 2013, 2015). Variations on the approach include region specific matching of images (Cristani et al., 2008; Kawakubo and Yanai, 2011), or matching of images within cells (Kalogerakis et al., 2009), the incorporation of temporality, so that other images by the same user or in a sequence can assist in image matching (Kalogerakis et al., 2009; Li et al., 2009), the inclusion of neighbouring images (Li, Qian, et al., 2013) and matching via place name tags, rather than directly to coordinates (e.g. Ivanov et al., 2012). It is common to combine the IMP approach with other methods (e.g. Xu, Cui, et al., 2014; Zhang et al., 2016), most commonly the use of tags (MTW) (e.g. Joshi et al., 2012; Kelm et al., 2012; Kennedy et al., 2007), due to its common use in content-sharing social media platforms.

5.7.2. Image-embedded geotag/geocode (including direction) (IMG)

A small number of papers access location information from the metadata of images, rather than the metadata of the social media platform. In some cases, the metadata of a social media platform is automatically populated by metadata from the photograph (e.g. Flickr offers this option currently), so the two may be the same, but there may also be differences, and with this approach, the image metadata (usually in EXIF format) is used. All five of the surveyed approaches that used the IMG method extracted content from Flickr, which offers an API that provides direct access to EXIF metadata for photographs (e.g McDougall and Temple-Watts, 2012; Sun et al., 2013). The EXIF metadata also includes angle of view, which is used by Panteras et al. (2015) and Shirai et al. (2013) to determine photo orientation. The method is used to analyse and derive new information (e.g. delineating fire or flood area boundaries, as in McDougall and Temple-Watts (2012) and Daly and Thom (2016)), and the accuracy reported usually relates to the method for analysis or derivation, rather than the accuracy of location of the photos themselves. For example, Daly and Thom (2016) report average accuracy of 132 km in identifying the location of fires (from derived centroid to ground truth centroid). In regard to coverage, Daly and Thom (2016) report that 2.5% of Flickr photos have EXIF header information.

5.8. Approaches that use videos (AVO, AVS)

We identify two approaches to the use of videos to determine location. The first of these (AVO) identifies location using object recognition. For example, Shen et al. (2011) identify landmarks from videos in YouTube by extracting camera metadata (including location, direction and viewing angle) and using this information to identify geographic features in the view, also incorporating data from OpenStreetMap. They then generate tags for the scenes in the video that identify the specific landmarks. The second broad approach (AVS) uses machine learning on audio visual features. For example, Sevillano et al. (2015) create a training set of geotagged videos, and then extract key frames from the videos, comparing audio and visual features and then using k-NN to select the best match from the training set.

5.9. Approaches that use information from the social media platform web site (SMW)

We also identify an approach (SMW) used by one paper, that identifies location from the web site of the social media platform, which provides summary statistics. Ferrara et al. (2013) study the geofocus of topics that are trending on Twitter (and thus the geofocus of messages) by extracting them from the Twitter home page which lists the top 10 trending topics, and by monitoring trends in 63 locations in the US.

5.10. Approaches that use the Spatio-temporal distribution of messages

Approaches in this category exclude those that use relationships between users (which are covered by the SNC approach) rather than messages and those that extract language models from messages (which are covered in the dedicated MCW approach in the message content category). This Section discusses approaches that infer location from the relationships between a number of messages, rather than a single message. These relationships may be spatial, temporal or contextual.

5.10.1. Inference of location from the Spatio-temporal distribution of messages (STD)

The STD approach infers location from the distribution of multiple messages, and relies on another approach to geolocate the individual messages, with MML and MCN being the most commonly used. Several papers use this approach to identify home location. Methods to infer home location include the geometric median (Compton et al., 2013) of message locations; location with highest post frequency or longest stay (Li, Wang, et al., 2012; 2014); last check-in of day; the use of particular words and phrases (e.g. home) (Hu et al., 2016); first message (Poulston et al., 2017) or the most frequent night-time message location (Luo et al., 2016). Cheng et al. (2011) use a recursive grid search, in which they find the grid cell of progressively smaller sizes that contains the most located tweets. Combined approaches have also been used, incorporating location of historical messages, places mentioned in tweets (MCN), LBSN check-ins and friend's locations (SNC) (Li, Zhao, et al., 2012; Mahmud et al., 2012, Mahmud et al., 2014).

The approach has been used for a number of purposes, including identifying the location of activities in which users are involved (e.g. office, education, shopping). Clustering combined with temporal analysis is a common methods for achieving this (Huang et al., 2014; Luo et al., 2016; Maeda et al., 2016). Yuan et al. (2016) focus on identifying periodic visiting behaviour, identifying places that people visit using the Chinese Restaurant Process and then determining the period of visits to enable location prediction. A second purpose is identification of the location of topics or bursts of frequently used words using density-based spatio-temporal clustering (Tamura and Ichimura, 2013) and multimodal location dependent probabilistic latent semantic analysis, the latter combining words and visual properties of images from Flickr (Zhou and Luo, 2012). Event location is another common purpose of this method, with several approaches identifying clusters of messages with particular keywords (Ranneries et al., 2016; Shirai et al., 2013). Van Canneyt et al. (2014, 2016) use meanshift clustering to cater for spatially disjoint events; Watanabe et al. (2011) use the geohash algorithm and Sakaki et al. (2010, 2013) incorporate trajectories as well as location, applying Kalman filtering and particle filtering. Finally, the approach has also been used to locate points of interest (POIs), usually in a tourism context. As with event location, clustering is a common approach (Memon et al., 2015), with additional factors such as the likelihood of a photo being a close up image of the POI being used to weight the clustering process (Popescu and Shabou, 2013).

5.10.2. Other approaches that use the Spatio-temporal distribution of messages (STI, STO, STP, STL, STR, STS)

A number of less common approaches have been used to establish location from the spatio-temporal distribution of messages, including the inference of location from the distribution of messages that have similar image features (STI method), with Zhou and Luo (2012) combining text and visual factors from Flickr images to identify topic regions with probabilistic latent semantic analysis, and Zheng et al. (2009) performing location-based agglomerative hierarchical clustering followed by visual clustering to identify landmarks.

The STO approach incorporates contextual information from other sources, and is shown in Salfinger et al. (2016), in which the MCN message is augmented with contextual information to assist in disambiguation of the place names included in the message. The direction or orientation of images (STP method) may also be used to infer location. For example, Shirai et al. (2013) extract orientation from photo metadata from Flickr, and analyse the orientations of all of the photos in a hotspot. They use this information to weight photos within the hotspot. The STL approach learns a user's location from the spatial distribution of his or her historical messages (in contrast to the SNC approach, in which inference is from the location of a user's connections). For example, Thom et al. (2014) propose a method in which clusters are formed from a user's historical messages from Twitter, and the cluster with the most messages is considered the home base. Similarly, Tran and Lee (2016) estimate the home location of a Twitter user by calculating the mean latitude and longitude of their 200 most recent tweets.

The STR approach infers location from other messages that are related in topic, event or time. For example, Yuan et al. (2013) develop an approach that incorporates both the language model approach (MCW) as the distribution of messages in time. They find the most likely location of a message by minimising variations in words, day and time to incorporate the user's daily and weekly patterns of behaviour. The STS approach also incorporates time, but considers temporal sequences. Kalogerakis et al. (2009) use the time differences in sequences of photographs to determine their locations at the scale of 400 km × 400 km cells. Liu et al. (2014) also consider the time differences between photographs, building a location profile for each user with a language model from tags (MTW) and weighting historical images by difference in time taken (the location tags of photos taken more recently are given a greater weight).

5.11. Approaches that use links to other social media platforms (LSM)

The final method considered involves the cross-use of different social media platforms by the same user. Xu, Lu, et al. (2014) attempt to locate messages in Tumblr by aligning users with their Twitter profiles and adopting locations from Twitter in order to geolocate their Tumblr accounts. To achieve this, they first have to align Twitter and Tumblr accounts, for which they achieve a 96% success rate. They then use SNC to determine user location in Twitter.

As can be seen from this analysis, the available methods vary widely in their accuracy, coverage and the type of situation in which they are most useful. Many researchers combat the shortcomings of one method in any one area by combining several methods. In this way, deficiencies in coverage for higher accuracy methods can be reduced by the use of a less accurate method. A number of researchers demonstrate improvements in accuracy, coverage or both through this combined approach (e.g. Daume, 2016; Bhatt et al., 2014; Bouillot et al., 2012; Cao et al., 2012; Fang and Dai, 2016). Another approach that researchers take to address issues with accuracy when compared in absolute terms, is to

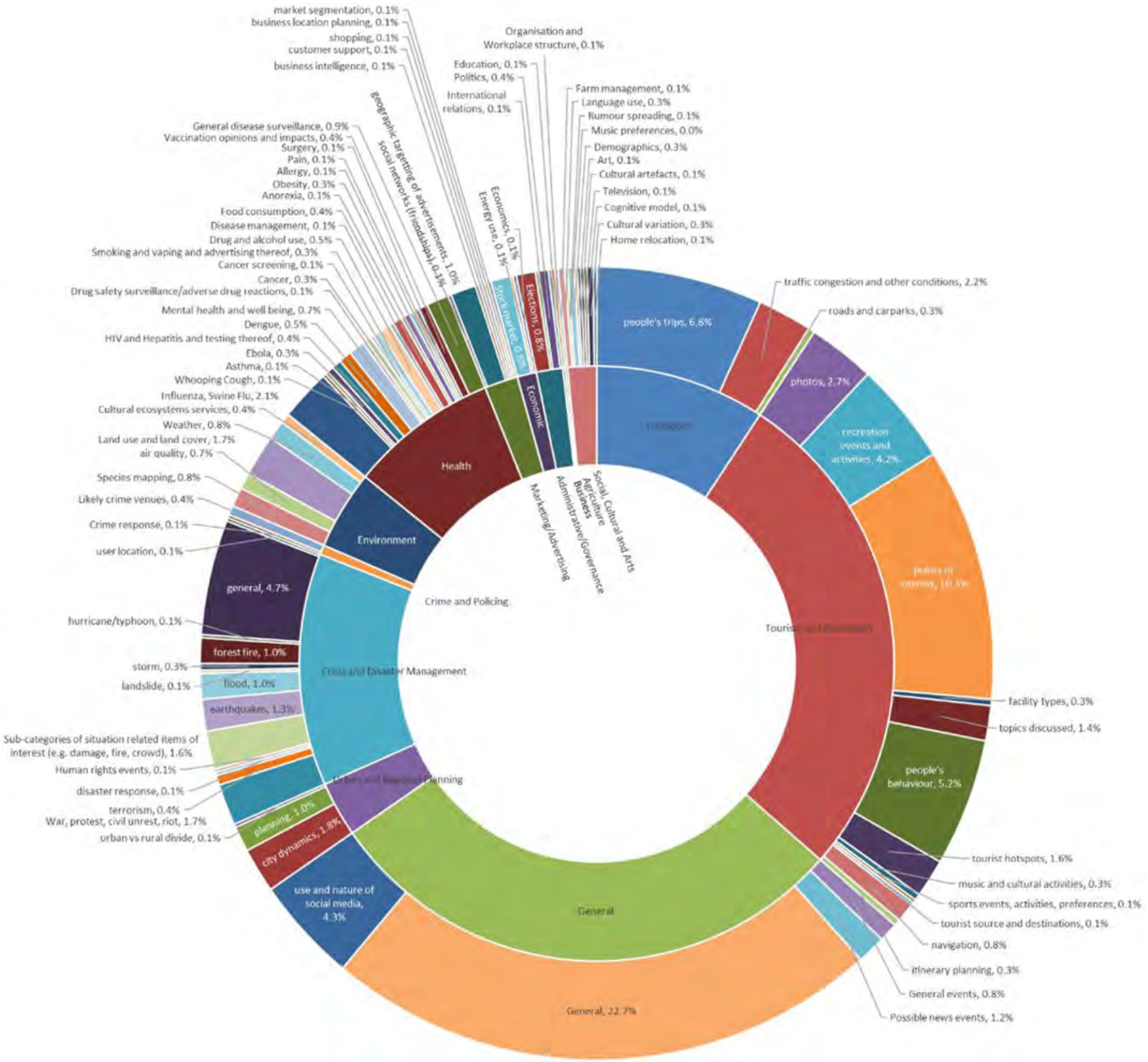

Fig. 8. Papers by application domain.

calibrate the data against a ground truth for the analysis being performed, often using some external source for the same data. If it can be demonstrated that the approach being used can successfully measure the phenomena of interest to the required accuracy, the absolute accuracy of individual measures is less important. For example, data on disease may be compared to data collected from hospitals or other official sources, and even if the accuracy or coverage is not high, it may still exceed the alternatives (e.g. Allen et al., 2016).

6. Analysis - RQ3: application domain

In this Section, we review the range of application domains in which social media information including location has been extracted. We developed a broad typology that grouped specific topics into themes, driven by the data itself rather than externally imposed. While other typologies and classification systems have been developed for geographic information application areas (for example, some are described in Maguire, 1991), most are too general for the purposes of this review, as social media has focussed on specific areas of interest. In this analysis, we distinguish between application domain and purpose. Steiger et al. (2015) combine the two by first using four broad categories focussed around the purpose of the extraction: event detection; location inference; social network analysis and no specific context of application, followed by a subdivision of event detection into three categories: disaster management; disease/health management and traffic management. We consider application domain independently of purpose.

Fig. 8 shows the distribution of papers by application domain. The inner ring contains broad categories, while the outer ring shows more specific categories within the broad areas. The largest broad category is General (28.9%), which mostly includes work that develops general methods for extracting location information from social media, and also includes studies of the use and nature of social media, news and general events (for reference, 0.8% = 6 papers). Many general papers used specific application areas as an illustration of their method, rather than a raison d'etre, and we used the mention of an application area in the title as an indicator of the general nature of the work, and thus a guideline for encoding the application area of a paper. Moving away from the general papers, important areas include tourism and recreation (27.2%), crisis and disaster management (12.6%), transport (9.2%) and health (8.1%). The broad range of application areas to which social media location extraction has been applied is notable, ranging from art and music, politics, crime and economics. The range of the specific application areas within each category is also interesting, with a broad spread of health topics ranging from infections (flu, whooping cough, HIV, dengue) to mental health and views about vaccinations, vaping and drug use and drug reactions. While the ways in which the more detailed categories of papers are aggregated into broader categories may be open to interpretation, the breadth of studies is clear.

7. Discussion

Edwards et al. (2013) discuss the role of social media research: as a surrogate for traditional methods; as a re-orientation of research around new objects, populations and techniques of analysis, or as augmentation, and evidence of all of these can be seen in the results of this study. In some cases, social media has been used as a replacement/alternative for data collection that would previously have been done manually. For example, studies of flu and other diseases are often compared to data collected by more traditional means (e.g. hospital admissions) to determine whether it is a true reflection (e.g. Allen et al., 2016). There is also evidence of research into areas for which it has previously been difficult to obtain data, with examples such as terrorism (e.g. Mirani and Sasi, 2016; Simon et al., 2014). The disaster domain is an example of an area for which social media has augmented existing data sources, and in some cases provided data that was not previous available.

Although Fig. 2 suggests that research into the use of geospatial data from social media is no longer increasing, there is still much potential for future research, and in this Section we discuss some of the areas in which future efforts might be directed.

7.1. Social media platforms

As discussed in Section 4.1, the dominance of Twitter in the research is clear. However, Statista's January 2018 ranking of web sites by number of users[5] ranks Twitter 11th globally, with Facebook, YouTube, Instagram and Sina Weibo all used more frequently, along with several messaging platforms. Facebook, by far the most popular social media platform, with 2.167 million users, only accounted for 2.1% of the usage reported in our survey, and YouTube, with 1.5 million users only 1.8%. Originally, Twitter was popular among researchers of social media location data extraction because of the high frequency of open accounts (not password protected or requiring membership), and because of its geolocation functionality. However, many other platforms have large quantities of open data. For example, Facebook now has a large quantity of public pages to which users cluster and add comments (for example, civil defence organisation Facebook pages, which are often full of posts before, during and after disaster events). Bird et al. (2012) show how Facebook has been used by members of the public to get useful and accurate information about a disaster event, including location-based information. Platforms such as Instagram and Youtube also have large quantities of open data. Furthermore, our research shows that the low proportion of tweets that are location tagged and the accuracy of alternative methods of geolocating tweets (e.g. MCN) make other platforms suitable alternatives for research focus. From a practical point of view, the Twitter API also presents a number of hurdles for automatic harvesting, with limited quotas and cost implications, making harvesting from other platforms easier in many cases. There is also much potential for increased research into specialist social media platforms, including those that address particular segments of the population (e.g. gay social media platforms), and those that are focussed around particularly topics (e.g. Strava for runners and cyclists).

We also highlight the need for research that investigates the characteristics, strengths and weaknesses of different platforms for different situations and application areas, continuing work started by Silva et al. (2013) and Simon et al. (2014). Current research shows little analysis of alternative platforms in selecting a strategy for social media harvesting, beyond high level considerations like whether text or images would be most useful, and researchers rarely discuss the reason for selecting a particular platform. In this paper, we have discussed the current usage of social media platforms, but more research is required on the suitability of particular platforms for different purposes.

Another related area for future research is the identification of appropriate search strategies to find useful content in different social media platforms. For social media that provide access via a stream (e.g. Twitter), this may involve identifying more efficient strategies than random selection or blanket coverage for finding relevant content, while for social media that provides access through particular entry points (e.g. Facebook pages), this may involve developing strategies to identify relevant pages and other entry points for a particular purpose. Social media platforms contain vast amounts of data that is not useful, and finding the useful geospatial content within it can be difficult. As more advanced strategies for extracting content that go beyond simply mapping specific words and topics become more developed, the ability to locate specific data will be more important.

7.2. Location extraction

The absence of accurate measures of coverage for many of the location extraction methods is an area that warrants future attention. Some of the methods described achieve high accuracy among a certain subset of messages or users, and this is not always clear when accuracy figures are examined. The definition of coverage is also not clear. Of most interest is the ability to determine for how many messages (or users, in the case of a base location determination), a location can be determined, but many researchers report the variations of this, including the percentage of messages that have a location in a particular format that can be located, or the percentage of place names that can be successfully extracted. While these are also interesting measures, they obscure the overall usefulness of the method for mapping or spatial analysis that can adequately represent a population.

In terms of prioritisation of methods for further research attention, Fig. 6 shows that a number of methods were given attention in the early years of social media location extraction research, but have not continued to be popular, including the tag-based methods (MTW, MTN), IMP (image-properties) and many of the methods that extract location from different aspects of the user profile. The lack of continued interest in these areas is generally supported by the accuracy figures in Fig. 7, with those methods generally falling towards the middle and lower right of the graph. The more successful methods (MTN and SNC) continue to receive attention, although less so in the latter case. The fact remains, however, that significant improvements are still needed in order to make social media location extraction effective at the sub 100 m level. Middleton et al. (2014) achieve 90% accuracy when identifying streets (by extracting street names), which is an easier task that identifying locations more generally. They do also achieve similar results for places, varying from buildings to rivers. Gelernter and

[5] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

Balaji's (2013) similar approach identifies buildings and place names with high success rates. These methods rely on a creation of local data sets that contain a reliable set of building and street names, and on the user employing those names, rather than more colloquial or flexible descriptions. Further research to make these approaches scalable would be appropriate, investigating methods for creation of reliable and complete gazetteers that include local and vernacular place names at a global level.

Another area that has been given very limited attention in the social media research is that of identifying location descriptions that are not confined to toponyms (including local place and street names). Many location descriptions that are used colloquially consist of relative location references or describe parts of wider areas (e.g. *the area opposite the train station was flooded; riots in downtown Nottingham; cannabis seized in raids in south-eastern suburbs in Melbourne)*, and there is very little research investigating the nature of this language, whether it differs from other natural language location descriptions, and the best methods to extract and interpret it. A few papers extract a subset of spatial relations (e.g. Bahir and Peled, 2016; Dittrich et al., 2015; Zhang et al., 2017) and descriptions that follow a specific template (e.g. *a traffic accident 10 km north of Wellsford on SH1*) (Bassi et al., 2016), but more extensive investigation could result in an increased ability to geolocate social media posts that are not adequately addressed by other location extraction methods.

An area that has been referred to but not studied in any detail is the influence of incorrect location data that is deliberately introduced. This has been identified as an issue by Sanaratne et al. (2017), who consider it a particular issue in image sharing sites like Flickr and Instagram, and Benvenuto et al. (2010) claim that spambots in Twitter skew the results of analysis. Issues such as location spoofing and fake check-ins are also recognised as an issue that may affect data quality with several different types and motivations for spoofing being identified by Zhao and Sui (2017). They detect up to 1.36% of geo-tags as spurious. It is difficult to determine how much this affects the results of location-based social media analysis, and it was not addressed systematically by the papers in our review. In some cases of social media mining of location data, it is less likely because users are unaware of the possibility that their data is being used, or of the strategy used to extract their location. For example, Benevento et al.'s (2010) work studied spam that tags advertising or pornographic URLs with unrelated tags, and the impact of this on the kinds of extractions and analysis reviewed in this paper depends very much on the level of analysis adopted. These strategies would have more impact on the simpler location extraction strategies (e.g. mapping a specific tag using MML) than on those that adopt more sophisticated or targeted extraction. In cases in which users are aware that their data is being used (e.g. when posting photos of species to a dedicated Facebook page, as in Deng et al., 2012), deliberate introduction of incorrect material may be more likely, but in these situations, the posts are visible to all users, so peer verification of the kind described by Goodchild and Li (2012) is possible. Future research to further investigate the scale of this problem in terms of its actual influence on analysis of location data extracted from social media is necessary, as is work to test and apply methods for data validation that incorporate the possibility of deliberately spurious data appearing in the data set. There has been very little of this work done to date.

7.3. Transferability and scalability

While research into geographic data extracted from social media has been conducted for some years and may provide a useful tool for researchers in a number of domains, the range of location extraction approaches are not yet sufficiently mature to allow easy application by researchers who are not experts in the use of social media, beyond the most basic tools offered by software applications for social media analysis (e.g. Microsoft's Social Engagement and SAS's Information Retrieval Studio). Such tools provide limited geographic extraction capabilities, and limited transparency in terms of how location is extracted and from where. This survey shows that it is currently very difficult to evaluate alternative approaches to collection of geographic data, and particularly to compare accuracy and coverage across methods, given the range of different reporting measures that are used. Some approaches are effective in a limited area, or within specified contexts, and it would be useful to adopt a more standardised approach to the reporting of success measures, in order to enable researchers from other domains to make more informed decisions about method suitability in a particular context, and the effort required to employ a particular method. Many researchers from other application areas fall back on the MML method with Twitter, and while this can provide useful analysis, it only provides access to a very limited amount of the geographic data that is potentially available.

8. Conclusions

The potential of social media as a source of geographic data that is not currently available, or as an alternative to more conventional data collection methods has been recognised by researchers since social media platforms became popular. A wide range of research has been generated that attempts to extract data of this kind, and in particular, to determine an accurate location of messages, users or topics discussed in messages. The most common method (MML), which uses coordinates that are contained in the message metadata, is very limited in coverage, and in some cases (if automatically extracted from the device) contains the location at which the message was posted, which may not necessarily be the geofocus of the message. Other methods that have received some attention and that give the best accuracy include MCN (which extracts place names from the message) and SNC (which uses social contacts to other users to establish location). MCN in particular can be effective in determining location using street names and building names, but relies on a well-developed gazetteer, and on users referring to relevant toponyms in their messages.

Twitter is by far the most frequently used social media platform for geospatial research, despite being only 11th in global rankings by number of users, and research on more popular platforms (e.g. Facebook) is much more limited. There is a need for research into some of these less frequently used platforms, including the analysis of the location of content of particular kinds across and within the platforms.

Geographic data has been extracted from social media across a vast range of application areas, from health to travel to politics, demonstrating the significant potential of the approach in collecting geographic data. The time has come to develop more comparable measures of quality of the methods that have been developed to extract location content, including a more standardised approach to the reporting of accuracy and coverage, that can enable researchers who are not experts in social media to better evaluate and employ the methods for location extraction that have been developed. It is currently very difficult for researchers who wish to apply social media data to a specific research question from some application domain to determine the best approach to use to extract geographic data, to evaluate the limitations of alternative approaches, and then to use the methods for their own research, and we assert that geographic data from social media could be used much more widely if this situation were addressed.

Appendix A. Summary of Selected SNC approaches

Reference	Social media platform	Method used	Type of connection	Factors considered	Results achieved AED = average error distance MED = median error distance
Abrol and Kahn (2010)	Twitter	Frequency of place names included in messages of self, followers and friends recursively. Probabilistic model; agglomerative hierarchical clustering.	Followed and following.	● Degree of friend overlap. ● Closeness of locations.	Correct city identified for 60.1% of messages.
Backstrom et al. (2010)	Facebook	Maximum likelihood.	Friends (always reciprocal in Facebook)	● Geographic distance between friends. ● Presence of recent communication or profile viewing (90 days) between friends. ● Number of friends (accounts with low numbers excluded).	69.1% of users with 16 or more located friends located within 25 miles.
Davis et al. (2011)	Twitter	Most frequent location of connected users.	Reciprocal followed and following.	● Number of friends (accounts with low and high numbers excluded).	Precision of 85% at city level, very low recall.
Li, Wang, et al. (2012)	Twitter	Optimisation.			63% accuracy but no detail.
Li, Zhao, et al. (2012)	Twitter	Unified discriminative influence model using bivariate Gaussian distribution.	Followed and following; locations mentioned in tweets.	● Set of edges under consideration (local vs global models).	AED 421.3 miles, 65.9% of tweets within 100 miles.
Pontes et al. (2012)	Google+[a]		Friends	● Number of friends (accounts with low and high numbers excluded).	MED ~7000 km, correct city in 50.89% of cases.
Sakaki et al. (2012)	Twitter	Dynamic Bayesian network		● Similarity in tweet vocabulary. ● Spatial and temporal collocation of tweets. ● Degree of friend overlap.	The correct location is predicted for 84.3% of time slices. Spatial accuracy is not clear.
McGee et al. (2013)	Twitter	Decision tree	Reciprocal followed and following.	● Number of friends (low numbers better than high). ● Frequency of communication through mentions. ● Account privacy (private is better than public). ● Geographical closeness of overlapping friends. ● Locality of user.	AED 364 miles, MED 10 miles, 63.9% within 25 miles.

Roth (2016)	Twitter	SVM Classifier	Reciprocated followed and following.	● City size. ● Degree of friend overlaps.	Correct city for 50.8% of users. MED 0 miles. 80% within 200 miles.
Yamaguchi et al. (2013)	Twitter	Landmark mixture model	Followed and following.	● Centrality and dispersion of user's connections.	AED 293 km, MED 2.7 km. 75.7% within 160 km.
Jurgens (2013)	Twitter	Label propagation by spatial arrangement of neighbours (nearest neighbour, geometric median, Oja's simplex median, geometric median of overlapping friends.		● Foursquare data to improve coverage.	MED 4 km (nearest neighbour)
Compton, Jurgens, and Allen (2014)	Twitter	Optimisation with parallel coordinate descent	Reciprocating mentions.	● Dispersion of user's connections (highly dispersed are excluded).	AED 289 km, MED 6.38 km, city level precision for 89.7% of users. Coverage: 81.9% of tweets, corresponding to 28.3% of volume, as the method requires high tweet volume.
Compton, Keegan, and Xu (2014)	Twitter, Flickr, YouTube, Tumblr	As above	As above		Twitter: AED 110 km, MED 6.7 km YouTube: AED 1001 km, MED 22.8 km Flickr: AED 2475 km, MED 372 km Tumblr: AED 3974 km, MED 1371 km.
Xu, Cui, et al. (2014)	Tencent Weibo	Social-content joint location propagation		● Degree of friend overlaps. ● Similarity in textual and visual content of messages.	AED 783 kms, MED 51 km, 68.2% accuracy at city level.
Di Rocco et al. (2016)	Twitter		Followed and following (for home location) Mentions and retweets (for message location)	● Ratio of followed to following (to exclude celebrities).	
Alonso-Lorenzo et al. (2016)	Twitter	Most frequent location of connected users.		● Type of account (organisation, celebrity and personal accounts are given different weights). ● Specificity of mentioned locations.	AED 658 km, MED 141 km.
Williams (2016)	Twitter	DBSCAN clustering	Followed and following		
Yuan et al. (2016)	Twitter, Foursquare	Jurgens algorithm (spatial label propagation) and Clauset-Newman-Moore algorithm for community detection		● User interests ● Historical tweets	AED 13.15 km.
Ebrahimi et al. (2017)	Twitter	label propagation (modified absorption), DBSCAN.	Mentions	● Local and global celebrities (the latter excluded).	AED and MED 476 and 32; 438 and 56 and 491 and 0 respectively for each data set (all in kms).

Full references that correspond to the citations contained in Appendix A are contained in the supplementary file to this paper, containing references and abstracts for all of the 690 papers reviewed.

[a] While other platforms are considered, the SNC approach is only optimal for Google+.

Appendix B. Summary of Selected MCW Approaches

Reference	Social media platform	Method used to create model	Location inferred	Results achieved
				AED = average error distance
				MED = median error distance

Cheng et al. (2010)	Twitter	Maximum likelihood estimation using only local words, with smoothing function.	City level user location	AED 535 miles MED (51%) 100 miles
Chandra et al. (2011)	Twitter	Reply-based Probability Distribution Model	City level user location	AED 1044.28 miles Around 50% of the results are within 500–1000 miles. 22% of the results are within 100 miles
Hecht et al. (2011)	Twitter	Term frequency multinomial Naïve Bayes model	Home location	Country (out of 4) 72.7% State (US) 30%
Kinsella et al. (2011)	Twitter	Bayesian inversion and Dirichlet smoothing to build the language model, then rank probability using the Kullback-Leibler (KL) divergence.	Tweet location, user location.	Tweet location: City 65.7% Neighbourhood 20% User location: State 45% Town 32% Postcode 15%
Li et al. (2011)	Twitter	Build language models for each POI and for the tweet, and using KL divergence between tweet and POI model to rank POIs.	Message location	POI 58%
Cheng et al. (2012)	Twitter	Guassian Mixture Model, incorporating the level of locality of words.	Home location	AED 509 miles MED 100 miles
Hua et al. (2012)	Twitter	Unigrams, local and semi-local words with word frequency-inverse city frequency (tf-icf), information gain and entropy models.	City level user location	AED 1383 km, MED 453 km
Hong et al. (2012)	Twitter	Topic model based on the Additive Generative model (SAGE), incorporating regional variations in language use with a Bayesian approach.	Message location	AED 120 km
Ikawa et al. (2013)	Twitter	Cosine similarity of new message against keyword lists for each location.	Message location	60% precision and 22% recall for 30 km for a single user 30% precision and 25% recall for all users
Ahmed et al. (2013)	Twitter	nested Chinese Restaurant Franchise (nCRF) model to model both topics and locations.	Message location	AED 194.81 km

Ryoo and Moon (2014)	Twitter	Build a model from local words, identified with measures of focus and dispersion and manually filtered.	Home location	AED 26.9 km MED 9 km
Thom et al. (2014)	Twitter	Term density estimations with an adaptive quadtree grid based on tweet density, then combined term density map used to select best match.	Message location	19.3% within 100 km, 44.8 within 1000 km (so MED is > 1000 km)
Xu, Cui, et al. (2014)	Tencent Weibo	Similarity in message content (using word vectors and cosine distance), combined with other approaches.	Message location	MED approximately 50 km (combined with IMP approach)
Flatow et al. (2015)	Twitter	Gaussian model with hyper-local n-grams and progressive elimination.	Message location	AED 2 km, < 10% coverage.
Ishida (2015)	Twitter	Calculate score for a term and for co-occurring terms in each administrative area based on frequency and dispersion.	User location	Generally < 50 km, but no precise figures given.
McClanahan and Gokhale (2015)	Twitter	Logistic regression classifier to identify sub-regions (defined through k-means clustering using message density), followed by incorporation of hyper local words with chi squared test; information gain ratio and geographic density.	Message location	MED 17 km (average across 3 data sets)
Musaev et al. (2015)	Twitter, Instagram, YouTube	Cluster analysis using both semantic and Euclidean distance between the place names.	Event location	
Tigunova et al. (2015)	Twitter	Location-aware behavioural topic model using Latent Dirichlet Allocation incorporating user, location and behaviour (message type).	Message location	
Williams (2016) and Williams (2016)	Twitter	Geotopical clustering using the topical similarity of tweets.	Message location	
Yuan et al. (2016)	Twitter, Foursquare	Latent Dirichlet Allocation to establish user interests, combined with social networks (SNC approach).	Next message location	AED 13 km (combined with SNC approach)
Ebrahimi et al. (2017)	Twitter	Logistic regression classifier over regions defined by a k-d tree. Combined with SNC using label propagation.	User location	
Kamimura et al. (2017)	Twitter	Kernel density estimation to develop a probability distribution for each word within a spatio-temporal region. The regions are defined with uniform geometric decomposition by number of messages.	Message location	52.4% within 5 km 77% within 100 km.
Lozano et al. (2017)	Twitter	Topic modelling using Streaming (to incorporate dynamics) Latent Dirichlet Allocation.	Message location	Election location 40%

Full references that correspond to the citations contained in Appendix B are contained in the supplementary file to this paper, containing references and abstracts for all of the 690 papers reviewed.

Appendix C. Accuracy and coverage of methods

Method	Reference	Platform	Type of location	0.01	Distance in km															
					0.05	0.1	0.2	0.5	1	2	5	10	15	20	25	30	40	50	100	160
					bldg	street				village/ neighbourhood				city						
MML	Zhu et al. (2015)	Sina Weibo	Message																	
	Straumann et al. (2014)	Flickr	Message																	
	Liu et al. (2014)	Flickr	Message																	
	Craglia et al. (2012)	Flickr	Message																	
	Musaev et al. (2015)	Twitter	Message																	
	Chen et al. (2014)	Twitter	Message																	
	Burton et al. (2012)	Twitter	Message																	
	Schulz et al. (2013)	Twitter	Message																	
	Bae and Yun (2017)	Google Panoramio	Message																	
	Musaev et al. (2015)	Instagram	Message																	
	Chandra et al. (2011)	Instagram	Message																	
	Musaev et al. (2015)	YouTube	Message																	
MMP	Schwartz et al. (2013)	Twitter	Message											84						
	Bouillot et al. (2012)	Twitter	Message																	
	Dredze et al. (2013)	Twitter	Message																	
	Burton et al. (2012)	Twitter	Message																	
	Schulz et al. (2013)	Twitter	Message						1	4	16	32			52			62	66	
	Croitoru et al. (2013)	Twitter	Message																	
MMD	Kelm et al. (2012)	Flickr	Message						19.4			46.8		56				64	66.8	
	Krauss et al. (2015)	YouTube	Message																	
UPC	Hecht et al. (2011)	Twitter	Base																	
	Takhteyev et al. (2012)	Twitter	Base											73.6						
	Daume (2016)	Twitter	Base																	
	Coloma et al. (2015)	Facebook	Base																	
	Coloma et al. (2015)	Twitter	Base																	
	Coloma et al. (2015)	Google+	Base																	
UPA	Backstrom et al. (2010)	Facebook	Base			*60*														
	Pontes et al. (2012)	Google+	Base																	
UPT	Burton et al. (2012)	Twitter	Base																	
	Schulz et al. (2013)	Twitter	Message								0	1			2			3	6	
UPL	Takhteyev et al. (2012)	Twitter	Base																	
UPP	Jiang et al. (2016)	WeChat Moments	Base																	
UPI	Schulz et al. (2013)	Twitter	Message												0			1	4	
UPD	Schulz et al. (2013)	Twitter	Message								0	1			2			6	17	
	Backstrom et al. (2010)	Facebook	Base												57.2					

UPN	Pontes et al. (2012)	Google+	Base														
UPE	Pontes et al. (2012)	Google+	Base									7.6					
UPS	Pontes et al. (2012)	Google+	Base									21.5					
SNC	Abrol and Kahn (2010)	Twitter	Base									60.1					
	Backstrom et al. (2010)	Facebook	Base					40		58		65		69.1			77
	Davis et al. (2011)	Twitter	Base									85					
	Li et al. (2012b)	Twitter	Base														65.9
	Pontes et al. (2012)	Google+	Base									50.9					
	McGee et al. (2013)	Twitter	Base						30	43		55	60	63.9			75
	Rout et al. (2013)	Twitter	Base									50.8				62	
	Yamaguchi et al. (2013)	Twitter	Base				50			52						70	75.7
	Jurgens (2013)	Twitter	Base				30		55	65						80	
	Compton et al. (2014a)	Twitter	Base				58			85		89.7					
	Compton et al. (2014b)	Twitter	Base														
	Compton et al. (2014b)	YouTube	Base														
	Compton et al. (2014b)	Flickr	Base														
	Compton et al. (2014b)	Tumblr	Base														
	Xu et al. (2014a)	Tencent Weibo	Base									68.2					
	Alonso-Lorenzo et al. (2016)	Twitter	Base														
	Yuan et al. (2016)	Twitter	Base														
	Ebrahimi et al. (2017)	Twitter	Base														69.7
MCN	Schulz et al. (2013)	Twitter	Geofocus				1	4	11	20			30		35	37	
	Bassi et al. (2016)	Twitter	Geofocus														
	Daly et al. (2013)	Twitter	Geofocus			45	70	*99*									
	Daume (2016)	Twitter	Geofocus														
	de Oliveira et al. (2017)	Twitter	Geofocus			25.2	32.9		61.5		74.8						
	Gelertner and Balaji (2013)	Twitter	Geofocus	*78*													
	Ikawa et al. (2013)	Twitter	Geofocus					91.7				92.7					
	Inkpen (2016)	Twitter	Geofocus									89					
	Middleton et al. (2014)	Twitter	Geofocus		90												
	Ribeiro et al. (2012)	Twitter	Geofocus		79			90									
	Sultanik and Fink (2012)	Twitter	Geofocus														
MLS	Gonzalez et al. (2012)	Twitter	Geofocus									66					
MIN	Chen et al. (2013)	Sina Weibo	Geofocus														
	Yuan et al. (2016)	Twitter	Geofocus														
MPR	Poulston et al. (2017)	Twitter	Geofocus														
	Li and Sun (2014, 2017)	Twitter	Geofocus														
MCS	Bassi et al. (2016)	Twitter	Geofocus														
MCW	Cheng et al. (2010)	Twitter	Base								30						51
	Chandra et al. (2011)	Twitter	Base														22
	Hecht et al. (2011)	Twitter	Base														
	Kinsella et al. (2011)	Twitter	Message					20				65.7					
	Kinsella et al. (2011)	Twitter	Base									32					
	Chang et al. (2012)	Twitter	Base														50
	Han et al. (2012)	Twitter	Base														
	Hong et al. (2012)	Twitter	Message														

	Ikawa et al. (2012)	Twitter	Message													30				
	Ahmed et al. (2013)	Twitter	Message																	
	Ryoo and Moon (2014)	Twitter	Base																	
	Thom et al. (2014)	Twitter	Message									5.9	9.4						16.8	19.3
	Flatow et al. (2015)	Twitter	Message																	
	McClanahan and Gokhale (2015)	Twitter	Message																	
	Kamimura et al. (2017)	Twitter	Message									53.7							75.5	77.3
MTW	Ahmed et al. (2013)	Twitter	Message																	
	Gallagher et al. (2009)	Flickr	Message																	
	Joshi et al. (2010)	Flickr	Message																	33
	Kordopatis-Zilos et al. (2015)	Flickr	Message		0.67		7.65		25.9				41.5							48.3
	Liu et al. (2014)	Flickr	Message						6	10.6			18.1		24.4		27.2	28.5		32.8
	O'Hare and Murdock (2013)	Flickr	Message						17	30										58
	Serdyukov et al. (2009)	Flickr	Message						7	14.1			18.7		27.1		28.8	30		30
LLC	Schulz et al. (2013)	Twitter	Message						97	98		99	99.5			99.7				99.8
IMP	Li et al. (2013b)	Flickr	Message		8.3		8.8		9.3				9.8							10.4
	Hays and Efros (2008)	Flickr	Message																	
	Kalogerakis et al. (2009)	Flickr	Message																	
	Zhang et al. (2016)	Flickr	Message																	
	Kelm et al. (2012)	Flickr	Message						5.4				5.6							8.6
IMG	Daly and Thom (2016)	Flickr	Geofocus																	
AVS	Sevillano et al. (2015)	Flickr	Message						0.9				3.3							7.42
STD	Hu et al. (2016)	Twitter	Base																	72
	Mahmud et al. (2012)	Twitter	Base																	
	Popescu and Shabou (2013)	Flickr	Geofocus																	
	Poulston et al. (2017)	Twitter	Base							83					97					
	Sakaki et al. (2010, 2013)	Twitter	Geofocus																	
	Van Canneyt et al. (2014, 2016)	Flickr	Geofocus																	
	Yuan et al. (2017)	Twitter, Gowalla	Base																	
STL	Thom et al. (2014)	Twitter	Base						23	30.1		37.9	44.8	47	48.5	49	50	51	51.8	53.6
STR	Yuan et al. (2013)	Twitter	Base																	
STS	Kalogerakis et al. (2009)	Flickr	Message																	
	Liu et al. (2014)	Flickr	Message						6.34				17.9							30.5

Method	Reference	Platform	Type of location	Distance in km														MED	AED	Coverage (%)
				180	200	220	260	300	340	400	500	800	1000	2000	5000	8000	####			
						state	province			country										
MML	Zhu et al. (2015)	Sina Weibo	Message																0.122	
	Straumann et al. (2014)	Flickr	Message																*4.5*	
	Liu et al. (2014)	Flickr	Message																	7.8
	Craglia et al. (2012)	Flickr	Message																	20
	Musaev et al. (2015)	Twitter	Message																	0.8
	Chen et al. (2014)	Twitter	Message																	6

	Burton et al. (2012)	Twitter	Message														0.91
	Schulz et al. (2013)	Twitter	Message												9.25	349	7.73
	Bae and Yun (2017)	Google Panoramio	Message														80
	Musaev et al. (2015)	Instagram	Message														16
	Chandra et al. (2011)	Instagram	Message														25
	Musaev et al. (2015)	YouTube	Message														6.4
MMP	Schwartz et al. (2013)	Twitter	Message														
	Bouillot et al. (2012)	Twitter	Message														37
	Dredze et al. (2013)	Twitter	Message														1
	Burton et al. (2012)	Twitter	Message														1.92
	Schulz et al. (2013)	Twitter	Message								81				23.3	1354	63.55
	Croitoru et al. (2013)	Twitter	Message														42.5
MMD	Kelm et al. (2012)	Flickr	Message		71					74.8	78.2	82.5	89.7	97.6			
	Krauss et al. (2015)	YouTube	Message														50
UPC	Hecht et al. (2011)	Twitter	Base			29			81								66
	Takhteyev et al. (2012)	Twitter	Base						26.4								77.4
	Daume (2016)	Twitter	Base														58.3
	Coloma et al. (2015)	Facebook	Base						33.5								
	Coloma et al. (2015)	Twitter	Base						43								
	Coloma et al. (2015)	Google +	Base						30								
UPA	Backstrom et al. (2010)	Facebook	Base														6
	Pontes et al. (2012)	Google +	Base														0.01
UPT	Burton et al. (2012)	Twitter	Base														77
	Schulz et al. (2013)	Twitter	Message								41				1543	2600	81.2
UPL	Takhteyev et al. (2012)	Twitter	Base														7.5
UPP	Jiang et al. (2016)	WeChat Moments	Base														
UPI	Schulz et al. (2013)	Twitter	Message								19				3287	5529	34.4
UPD	Schulz et al. (2013)	Twitter	Message								58				494	2618	6.46
	Backstrom et al. (2010)	Facebook	Base														
UPN	Pontes et al. (2012)	Google +	Base														61.8
UPE	Pontes et al. (2012)	Google +	Base														34.5
UPS	Pontes et al. (2012)	Google +	Base														53
SNC	Abrol and Kahn (2010)	Twitter	Base						80.1								
	Backstrom et al. (2010)	Facebook	Base														
	Davis et al. (2011)	Twitter	Base														
	Li et al. (2012b)	Twitter	Base													677.5	
	Pontes et al. (2012)	Google +	Base												7000		
	McGee et al. (2013)	Twitter	Base						83		90				16	585	
	Rout et al. (2013)	Twitter	Base	69			74	80									
	Yamaguchi et al. (2013)	Twitter	Base												2.7	293	
	Jurgens (2013)	Twitter	Base								95				4		
	Compton et al. (2014a)	Twitter	Base												6.38	289	81.9
	Compton et al. (2014b)	Twitter	Base								88	92	95		6.7	110	

	Compton et al. (2014b)	YouTube	Base								60	67	78	93	22.8	1001	
	Compton et al. (2014b)	Flickr	Base								33	40	56	88	372	2475	
	Compton et al. (2014b)	Tumblr	Base								25	40	57	88	1371	3974	
	Xu et al. (2014a)	Tencent Weibo	Base			73.7									51	783	
	Alonso-Lorenzo et al. (2016)	Twitter	Base					93.2							141	658	
	Yuan et al. (2016)	Twitter	Base													13.15	
	Ebrahimi et al. (2017)	Twitter	Base												29.3	468.3	70.67
MCN	Schulz et al. (2013)	Twitter	Geofocus								49				1100	3689	5.13
	Bassi et al. (2016)	Twitter	Geofocus												0.5		83
	Daly et al. (2013)	Twitter	Geofocus														
	Daume (2016)	Twitter	Geofocus														30.33
	de Oliveira et al. (2017)	Twitter	Geofocus												2.48		
	Gelertner and Balaji (2013)	Twitter	Geofocus														
	Ikawa et al. (2013)	Twitter	Geofocus		100			99.6									
	Inkpen (2016)	Twitter	Geofocus		91			95									
	Middleton et al. (2014)	Twitter	Geofocus		95												
	Ribeiro et al. (2012)	Twitter	Geofocus														
	Sultanik and Fink (2012)	Twitter	Geofocus												80		
MLS	Gonzalez et al. (2012)	Twitter	Geofocus					79								28.8	
MIN	Chen et al. (2013)	Sina Weibo	Geofocus													10.3	
	Yuan et al. (2016)	Twitter	Geofocus													13.15	
MPR	Poulston et al. (2017)	Twitter	Geofocus												0.01	1.55	
	Li and Sun (2014, 2017)	Twitter	Geofocus														35.7
MCS	Bassi et al. (2016)	Twitter	Geofocus												1.06	54.2	
MCW	Cheng et al. (2010)	Twitter	Base								80					861	
	Chandra et al. (2011)	Twitter	Base							37						1681	
	Hecht et al. (2011)	Twitter	Base		30			72.7									
	Kinsella et al. (2011)	Twitter	Message														
	Kinsella et al. (2011)	Twitter	Base		45												
	Chang et al. (2012)	Twitter	Base												160	820	
	Han et al. (2012)	Twitter	Base												453	1383	
	Hong et al. (2012)	Twitter	Message													120	
	Ikawa et al. (2012)	Twitter	Message														
	Ahmed et al. (2013)	Twitter	Message													194.8	
	Ryoo and Moon (2014)	Twitter	Base												9	26.9	
	Thom et al. (2014)	Twitter	Message								44.8						
	Flatow et al. (2015)	Twitter	Message													2	10
	McClanahan and Gokhale (2015)	Twitter	Message												17		
	Kamimura et al. (2017)	Twitter	Message														
MTW	Ahmed et al. (2013)	Twitter	Message													194.8	
	Gallagher et al. (2009)	Flickr	Message	11			14		17		25						

	Joshi et al. (2010)	Flickr	Message	40	47	50	52	58	60					400		
	Kordopatis-Zilos et al. (2015)	Flickr	Message											160		
	Liu et al. (2014)	Flickr	Message	39	42.5	44.9										
	O'Hare and Murdock (2013)	Flickr	Message	67	69	71										
	Serdyukov et al. (2009)	Flickr	Message	39	44	47										
LLC	Schulz et al. (2013)	Twitter	Message						99.8							
IMP	Li et al. (2013b)	Flickr	Message													100
	Hays and Efros (2008)	Flickr	Message	16												100
	Kalogerakis et al. (2009)	Flickr	Message			58										100
	Zhang et al. (2016)	Flickr	Message												40	100
	Kelm et al. (2012)	Flickr	Message	13.8			24.2		34.9	50.2	63.3		90.5			100
IMG	Daly and Thom (2016)	Flickr	Geofocus												132.3	2.5
AVS	Sevillano et al. (2015)	Flickr	Message						16.5				86.6			100
STD	Hu et al. (2016)	Twitter	Base	73	75	76	76	79	80							80
	Mahmud et al. (2012)	Twitter	Base													6.6
	Popescu and Shabou (2013)	Flickr	Geofocus												65.6	
	Poulston et al. (2017)	Twitter	Base												1.55	
	Sakaki et al. (2010, 2013)	Twitter	Geofocus												3.295	
	Van Canneyt et al. (2014, 2016)	Flickr	Geofocus													35
	Yuan et al. (2017)	Twitter, Gowalla	Base												4	
STL	Thom et al. (2014)	Twitter	Base						59.9			62.6				
STR	Yuan et al. (2013)	Twitter	Base												100	
STS	Kalogerakis et al. (2009)	Flickr	Message			58										
	Liu et al. (2014)	Flickr	Message	35.8	39.1	41.5										60

Full references that correspond to the citations contained in Appendix C are contained in the supplementary file to this paper, containing references and abstracts for all of the 690 papers reviewed.

Appendix D. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compenvurbsys.2018.05.007.

References

Abbasi, R., Chernov, S., Nejdl, W., Paju, R., & Staab, S. (2009). Exploiting Flickr tags and groups for finding landmark photos. In M. Boughanem, C. Berrut, J. Mothe, & C. Soule-Dupuy (Eds.). *Advances in Information Retrieval: Proceedings of the Thirty-First European Conference on IR research, ECIR 2009, Toulouse, France* (pp. 654–661). Berlin: Springer.

Abrol, S., & Khan, L. (2010). Tweethood: Agglomerative clustering on fuzzy k-closest friends with variable depth for location mining. *Social computing (SocialCom), 2010 IEEE second international conference on* (pp. 153–160). IEEE.

Ahmed, A., Hong, L., & Smola, A. J. (2013). Hierarchical geographical modeling of user locations from social media posts. *Proceedings of the 22nd international conference on World Wide Web* (pp. 25–36). ACM.

Ajao, O., Hong, J., & Liu, W. (2015). A survey of location inference techniques on Twitter. *Journal of Information Science, 41*(6), 855–864.

AlBanna, B., Sakr, M., Moussa, S., & Moawad, I. (2016). Interest Aware Location-Based Recommender System Using Geo-Tagged Social Media. *ISPRS International Journal of Geo-Information, 5*(12), 1–19.

Allen, C., Tsou, M. H., Aslam, A., Nagel, A., & Gawron, J. M. (2016). Applying GIS and Machine Learning Methods to Twitter Data for Multiscale Surveillance of Influenza. *PloS One, 11*(7), e0157734.

Ardon, S., Bagchi, A., Mahanti, A., Ruhela, A., Seth, A., Tripathy, R. M., & Triukose, S. (2013). Spatio-temporal and events based analysis of topic popularity in twitter. *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (pp. 219–228). ACM.

Arksey, H., & O'Malley, L. (2005). Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology, 8*(1), 19–32.

Backstrom, L., Sun, E., & Marlow, C. (2010). Find me if you can: improving geographical prediction with social and spatial proximity. *Proceedings of the 19th international conference on World wide web* (pp. 61–70). ACM.

Bae, S. H., & Yun, H. J. (2017). Spatiotemporal distribution of visitors' geotagged landscape photos in rural areas. *Tourism Planning & Development, 14*(2), 167–180.

Bahir, E., & Peled, A. (2016). Geospatial extreme event establishing using social network's text analytics. *GeoJournal, 81*(3), 337–350.

Bao, J., Zheng, Y., Wilkie, D., & Mokbel, M. (2015). Recommendations in location-based social networks: a survey. *GeoInformatica, 19*(3), 525–565.

Bassi, J., Manna, S., & Sun, Y. (2016). Construction of a geo-location service utilizing microblogging platforms. *Semantic Computing (ICSC), 2016 IEEE Tenth International Conference on* (pp. 162–165). IEEE.

Batrinca, B., & Treleaven, P. C. (2015). Social media analytics: a survey of techniques, tools and platforms. *AI & Society, 30*(1), 89–116.

Baucom, E., Sanjari, A., Liu, X., & Chen, M. (2013). Mirroring the real world in social media: twitter, geolocation, and sentiment analysis. *Proceedings of the 2013 international workshop on Mining unstructured big data using natural language processing* (pp. 61–68). ACM.

Becker, H., Naaman, M., & Gravano, L. (2010). Learning similarity metrics for event identification in social media. *Proceedings of the third ACM international conference on Web search and data mining* (pp. 291–300). ACM.

Benevenuto, F., Magno, G., Rodrigues, T., & Almeida, V. (2010). Detecting spammers on twitter. *Presented at the Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*.

Bhatt, S. P., Purohit, H., Hampton, A., Shalin, V., Sheth, A., & Flach, J. (2014). Assisting coordination during crisis: a domain ontology based approach to infer resource needs from tweets. *Proceedings of the 2014 ACM conference on Web science* (pp. 297–298). ACM.

Bird, D., Ling, M., & Haynes, K. (2012). Flooding Facebook-the use of social media during the Queensland and Victorian floods. *Australian Journal of Emergency Management, 27*(1), 27.

Bouillot, F., Poncelet, P., & Roche, M. (2012). How and why exploit tweet's location information? *AGILE' 2012: 15th international conference on geographic information science*.

Bui, T. H., Kim, A. Y., Park, S. B., & Lee, S. J. (2016). *Generating Point of Interest Description with Geo-tagged Web Photos. In Information Science and Applications (ICISA) 2016.* Singapore: Springer1013–1023.

Burton, S. H., Tanner, K. W., Giraud-Carrier, C. G., West, J. H., & Barnes, M. D. (2012). "Right Time, Right Place" Health Communication on Twitter: Value and Accuracy of Location Information. *Journal of Medical Internet Research, 14*(6), 366–376.

Cao, N., Lin, Y.-R., Sun, X., Lazer, D., Liu, S., & Qu, H. (2012). Whisper: tracing the spatiotemporal process of information diffusion in real time. *IEEE Transactions on Visualization and Computer Graphics, 18*(12), 2649–2658.

Castro, R., Kuffó, L., & Vaca, C. (2017). Back to# 6D: Predicting Venezuelan states political election results through Twitter. *eDemocracy & eGovernment (ICEDEG), 2017 Fourth International Conference on* (pp. 148–153). IEEE.

Chandra, S., Khan, L., & Muhaya, F. B. (2011). Estimating twitter user location using social interactions—a content based approach. *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on* (pp. 838–843). IEEE.

Chen, F., Joshi, D., Miura, Y., & Ohkuma, T. (2014). Social media-based profiling of business locations. *Proceedings of the 3rd ACM multimedia workshop on geotagging and its applications in multimedia* (pp. 1–6). ACM.

Cheng, C., Yang, H., King, I., & Lyu, M. R. (2012). Fused matrix factorization with geographical and social influence in location-based social networks. *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI'12), Toronto, ON, Canada, 22–26 July 2012. Vol. 12. Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI'12), Toronto, ON, Canada, 22–26 July 2012* (pp. 17–23).

Cheng, Y.-C., & Chen, P.-L. (2014). Global social media, local context: A case study of Chinese-language tweets about the 2012 presidential election in Taiwan. *Aslib Journal of Information Management, 66*(3), 342–356.

Cheng, Z., Caverlee, J., & Lee, K. (2010). You are where you tweet: a content-based approach to geo-locating twitter users. *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 759–768). ACM.

Cheng, Z., Caverlee, J., Lee, K., & Sui, D. Z. (2011). Exploring millions of footprints in location sharing services. *International AAAI Conference on Weblogs and Social Media 2011* (pp. 81–88).

Cheong, M., & Lee, V. (2009). Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base. *Proceedings of the 2nd ACM workshop on Social web search and mining* (pp. 1–8). ACM.

Cheong, M., & Lee, V. (2010). Twittering for earth: A study on the impact of microblogging activism on Earth Hour 2009 in Australia. *Intelligent Information and Database Systems,* 114–123.

Chiang, M. F., Chen, C. C., Peng, W. C., & Philip, S. Y. (2014). Mining mobility evolution from check-in datasets. *Mobile Data Management (MDM), 2014 IEEE 15th International Conference on. Vol. 1. Mobile Data Management (MDM), 2014 IEEE 15th International Conference on* (pp. 195–204). IEEE.

Cho, E., Myers, S. A., & Leskovec, J. (2011). Friendship and mobility: user movement in location-based social networks. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1082–1090). ACM.

Coloma, P. M., Becker, B., Sturkenboom, M. C., Van Mulligen, E. M., & Kors, J. A. (2015). Evaluating social media networks in medicines safety surveillance: Two case studies. *Drug Safety, 38*(10), 921–930.

Compton, R., Lee, C., Lu, T. C., De Silva, L., & Macy, M. (2013). Detecting future social unrest in unprocessed twitter data:"emerging phenomena and big data". *Intelligence and Security Informatics (ISI), 2013 IEEE International Conference On* (pp. 56–60). IEEE.

Compton, R., Jurgens, D., & Allen, D. (2014). Geotagging one hundred million twitter accounts with total variation minimization. *Big Data (Big Data), 2014 IEEE International Conference on* (pp. 393–401). IEEE.

Compton, R., Keegan, M. S., & Xu, J. (2014). Inferring the geographic focus of online documents from social media. *Computational Approaches to Social Modeling (ChASM) Workshop, WebSci 2014, Bloomington, Indiana-June 24–26 2014*.

Conover, M. D., Davis, C., Ferrara, E., McKelvey, K., Menczer, F., & Flammini, A. (2013). The geospatial characteristics of a social movement communication network. *PLoS One, 8*(3), 1–8.

Craglia, M., Ostermann, F., & Spinsanti, L. (2012). Digital earth from vision to practice: Making sense of citizen-generated content. *International Journal of Digital Earth, 5*, 398–416.

Crandall, D., Backstrom, L., Huttenlocher, D., & Kleinberg, J. (2009). Mapping the World's Photos. *Proceedings of the 18th International World Wide Web Conference* (pp. 761–770).

Cristani, M., Perina, A., Castellani, U., & Murino, V. (2008). Geolocated image analysis using latent representations. In: *Proc. Of IEEE Computer Vision and Pattern Recognition*.

Croitoru, A., Crooks, A., Radzikowski, J., & Stefanidis, A. (2013). Geosocial gauge: a system prototype for knowledge discovery from social media. *International Journal of Geographical Information Science, 27*(12), 2483–2508.

Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). # Earthquake: Twitter as a distributed sensor system. *Transactions in GIS, 17*(1), 124–147.

Dalvi, N., Kumar, R., & Pang, B. (2012). Object matching in tweets with spatial models. *Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 43–52). ACM.

Daly, E. M., Lecue, F., & Bicer, V. (2013). Westland row why so slow?: fusing social media and linked data sources for understanding real-time traffic conditions. *Proceedings of the 2013 international conference on Intelligent user interfaces* (pp. 203–212). ACM.

Daly, S., & Thom, J. A. (2016). Mining and classifying image posts on social media to analyse fires. *Proceedings of the International ISCRAM Conference. CONF, RMIT University: Information Systems for Crisis Response and Management.* ISCRAM.

Daume, S. (2016). Mining Twitter to monitor invasive alien species - An analytical framework and sample information topologies. *Ecological Informatics, 31*, 70–82.

De Choudhury, M., Feldman, M., Amer-Yahia, S., Golbandi, N., Lempel, R., & Yu, C. (2010). Automatic construction of travel itineraries using social breadcrumbs. *Proceedings of the 21st ACM conference on Hypertext and hypermedia* (pp. 35–44). ACM.

de Oliveira, M. G., de Souza Baptista, C., Campelo, C. E., & Bertolotto, M. (2017). A gold-standard social media corpus for urban issues. *Proceedings of the Symposium on Applied Computing* (pp. 1011–1016). ACM.

Di Rocco, L., Bertolotto, M., Catania, B., Guerrini, G., & Cosso, T. (2016). Extracting fine-grained implicit georeferencing information from microblogs exploiting crowdsourced gazetteers and social interactions. *2016 19th AGILE international conference on geographic information science*.

Dittrich, A., Vasardani, M., Winter, S., Baldwin, T., & Liu, F. (2015). A classification schema for fast disambiguation of spatial prepositions. *Proceedings of the 6th Acm Sigspatial International Workshop on Geostreaming (Iwgs) 2015* (pp. 78–86).

Dredze, M., Paul, M. J., Bergsma, S., & Tran, H. (2013). Carmen: A twitter geolocation system with applications to public health. *AAAI workshop on expanding the boundaries of health informatics using AI (HIAI)* (pp. 20–24).

Ebrahimi, M., ShafieiBavani, E., Wong, R., & Chen, F. (2017). Exploring celebrities on inferring user geolocation in Twitter. *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 395–406). Cham: Springer.

Edwards, A., Housley, W., Williams, M., Sloan, L., & Williams, M. (2013). Digital social research, social media and the sociological imagination: Surrogacy, augmentation and re-orientation. *International Journal of Social Research Methodology, 16*(3), 245–260.

Endarnoto, S. K., Pradipta, S., Nugroho, A. S., & Purnama, J. (2011). Traffic condition information extraction & visualization from social media twitter for android mobile application. *Electrical Engineering and Informatics (ICEEI), 2011 International Conference on* (pp. 1–4). IEEE.

Eo, S.-W., Lee, Y., Yu, K., & Park, W. (2016). Establishing the process of spatial informatization using data from social network services. *Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography, 34*(2), 111–120.

Fang, M.-Y., & Dai, B.-R. (2016). Power of bosom friends, POI recommendation by learning preference of close friends and similar users. In S. Madria, & T. Hara (Vol. Eds.), *Big data analytics and knowledge discovery, Dawak 2016*. *Vol. 9829*. *Big data analytics and knowledge discovery, Dawak 2016* (pp. 179–192).

Ferrara, E., Varol, O., Menczer, F., & Flammini, A. (2013). Traveling trends: social butterflies or frequent fliers? *Proceedings of the first ACM conference on Online social networks* (pp. 213–222). ACM.

Fersini, E., Messina, E., & Pozzi, F. (2017). Earthquake management: a decision support system based on natural language processing. *Journal of Ambient Intelligence & Humanized Computing, 8*(1), 37.

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

Fire, M., & Puzis, R. (2016). Organization mining using online social networks. *Networks and Spatial Economics, 16*(2), 545–578.

Flatow, D., Naaman, M., Xie, K. E., Volkovich, Y., & Kanza, Y. (2015). On the accuracy of hyper-local geotagging of social media content. *Web Search & Web Data Mining,* 127.

Gallagher, A., Joshi, D., Yu, J., & Luo, J. (2009). Geo-location inference from image content and user tags. *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on* (pp. 55–62). IEEE.

Gao, H., Tang, J., & Liu, H. (2012). Exploring social-historical ties on location-based social networks. *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*.

Gao, H., Tang, J., Hu, X., & Liu, H. (2013). Exploring temporal effects for location recommendation on location-based social networks. *Proceedings of the 7th ACM conference on Recommender systems* (pp. 93–100). ACM.

Gao, Y., Tang, J., Hong, R., Dai, Q., Chua, T.-S., & Jain, R. (2010). W2Go: A travel guidance system by automatic landmark ranking. *Proceedings of the seventeen ACM international conference on Multimedia*. MM.

Gelernter, J., & Balaji, S. (2013). An algorithm for local geoparsing of microtext. *GeoInformatica, 17*(4), 635–667.

Gelernter, J., & Mushegian, N. (2011). Geo-parsing messages from Microtext. *Transactions in GIS, 15*(6), 753–773.

Ghosh, D., & 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.

Gonzalez, R., Figueroa, G., & Chen, Y. S. (2012). Tweolocator: a non-intrusive geographical locator system for twitter. *Proceedings of the 5th ACM SIGSPATIAL international workshop on location-based social networks* (pp. 24–31). ACM.

Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics, 1*, 110–120.

Grinberg, N., Naaman, M., Shaw, B., & Lotan, G. (2013). Extracting diurnal patterns of real world activity from social media. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*.

Grothe, C., & Schaab, J. (2009). Automated footprint generation from geotags with kernel density estimation and support vector machines. *Spatial Cognition & Computation, 9*(3), 195–211.

Gu, Y., Qian, Z.(. S.)., & Chen, F. (2016). From Twitter to detector: Real-time traffic incident detection using social media data. *Transportation Research Part C, 67*, 321–342.

Guy, S., Ratzki-Leewing, A., Bahati, R., & Gwadry-Sridhar, F. (2011). Social media: A systematic review to understand the evidence and application in infodemiology. *International conference on electronic healthcare* (pp. 1–8). Berlin, Heidelberg: Springer.

Hasan, S., & Ukkusuri, S. V. (2017). Reconstructing activity location sequences from incomplete check-in data: a semi-Markov continuous-time Bayesian network model. *IEEE Transactions on Intelligent Transportation Systems, 19*(3), 687–698.

Hays, J., & Efros, A. A. (2008). IM2GPS: estimating geographic information from a single image. *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (pp. 1–8). IEEE.

Heath, A. (2016). Foursquare has an amazing 'superpower' that wants to take over your phone. *Business Insider*. Jan. 1, 2016, 10:17 AM. Retrieved from http://www.businessinsider.com/inside-foursquares-pilgrim-technology-2015-12/?r=AU&IR=T (last accessed 19/12/17) .

Hecht, B., Hong, L., Suh, B., & Chi, E. H. (2011). Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 237–246). ACM.

Hennig, L., Thomas, P., Ai, R., Kirschnick, J., Wang, H., Pannier, J., ... Uszkoreit, H. (2016). *Real-time discovery and geospatial visualization of mobility and industry events from large-scale, heterogeneous data streams. ACL 2016.* 37.

Hollenstein, L., & Purves, R. (2010). Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science, 2010*(1), 21–48.

Horita, F. E. A., Degrossi, L. C., LFG, A., Zipf, A., & Albuquerque, J. P. (2013). The use of volunteered geographic information and crowdsourcing in disaster management: A systematic literature review. *Proceedings of the Nineteenth Americas Conference on Information Systems, Atlanta, Georgia* (pp. 1–10).

Hu, T. R., Luo, J. B., Kautz, H., & Sadilek, A. (2016). Home location inference from sparse and noisy data: models and applications. *Frontiers of Information Technology & Electronic Engineering, 17*(5), 389–402.

Hua, W., Huynh, D. T., Hosseini, S., Lu, J., & Zhou, X. (2012). Information Extraction From Microblogs: A Survey. *International Journal of Software and Informatics, 6*(4), 495–522.

Huang, Q., Cao, G., & Wang, C. (2014). From where do tweets originate?: a GIS approach for user location inference. *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks* (pp. 1–8). ACM.

Hyvärinen, O., & Saltikoff, E. (2010). Social media as a source of meteorological observations. *Monthly Weather Review, 138*(8), 3175–3184.

Ikawa, Y., Vukovic, M., Rogstadius, J., & Murakami, A. (2013). Location-based insights from the social web. *Proceedings of the 22nd international conference on World Wide Web* (pp. 1013–1016). ACM.

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR), 47*(4), 67.

Inkpen, D. (2016). Text mining in social media for security threats. *Recent Advances in Computational Intelligence in Defense and Security* (pp. 491–517). Springer International Publishing.

Ishida, K. (2015). Estimation of user location and local topics based on geo-tagged text data on social media. *Advanced Applied Informatics (IIAI-AAI), 2015 IIAI 4th International Congress on* (pp. 14–17). IEEE.

Ivanov, I., Vajda, P., Lee, J.-S., Goldmann, L., & Ebrahimi, T. (2012). Geotag propagation in social networks based on user trust model. *Multimedia Tools and Applications, 56*(1), 155–177.

Jaiswal, A., Peng, W., & Sun, T. (2013). Predicting time-sensitive user locations from social media. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 870–877). ACM.

Jeffries, A. (2012). Why Apple Maps needs Foursquare's 50 million venues. *The Verge*. Dec 18, 2012, 4:50pm EST. Retrived from https://www.theverge.com/2012/12/18/3781336/foursquare-and-apple-maps-problem-joke-venues (last accessed 19/12/17) .

Ji, R., Duan, L. Y., Chen, J., Yang, S., Yao, H., Huang, T., & Gao, W. (2011). Learning the trip suggestion from landmark photos on the web. *Image Processing (ICIP), 2011 18th IEEE International Conference on* (pp. 2485–2488). IEEE.

Jiang, W., Chen, B., He, L., Bai, Y., & Qiu, X. (2016). Features of rumor spreading on WeChat moments. In A. Morishima, L. Chang, T. Z. J. Fu, K. Liu, X. Yang, J. Zhu, R. Zhang, W. Zhang, & Z. Zhang (Vol. Eds.), *Web Technologies and Applications: APWeb 2016 Workshops, WDMA, GAP, and SDMA*. *Vol. 9865*. *Web Technologies and Applications: APWeb 2016 Workshops, WDMA, GAP, and SDMA* (pp. 217–227).

Joseph, S. L., Xiao, J., Zhang, X., Chawda, B., Narang, K., Rajput, N., ... Subramaniam, L. V. (2015). Being aware of the world: Toward using social media to support the blind with navigation. *IEEE Transactions on human-machine systems, 45*(3), 399–405.

Joshi, D., Gallagher, A., Yu, J., & Luo, J. (2010). Exploring user image tags for geo-location inference. *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on* (pp. 5598–5601). IEEE.

Joshi, D., Gallagher, A., Yu, J., & Luo, J. (2012). Inferring photographic location using geotagged web images. *Multimedia Tools and Applications, 56*(1), 131–153.

Jurgens, D. (2013). That's what friends are for: Inferring location in online social media platforms based on social relationships. *ICWSM, 13*(13), 273–282.

Kalogerakis, E., Vesselova, O., Hays, J., Efros, A. A., & Hertzmann, A. (2009). Image sequence geolocation with human travel priors. *Computer Vision, 2009 IEEE 12th International Conference on* (pp. 253–260). IEEE.

Kamimura, T., Nitta, N., Nakamura, K., & Babaguchi, N. (2017). On-line geospatial term extraction from streaming geotagged tweets. *Multimedia Big Data (BigMM), 2017 IEEE Third International Conference on* (pp. 322–329). IEEE.

Kanta, M., Simko, M., & Bieliková, M. (2012). Trend-aware user modeling with location-aware trends on Twitter. *Semantic and Social Media Adaptation and Personalization (SMAP), 2012 Seventh International Workshop on* (pp. 23–28). IEEE.

Kaplan, A. M. (2012). If you love something, let it go mobile: Mobile marketing and mobile social media 4x4. *Business Horizons, 55*(2), 129–139.

Kawakubo, H., & Yanai, K. (2011). Geovisualrank: a ranking method of geotagged imagesconsidering visual similarity and geo-location proximity. *Proceedings of the 20th international conference companion on World wide web* (pp. 69–70). ACM.

Kelm, P., Schmiedeke, S., & Sikora, T. (2012). Multimodal geo-tagging in social media websites using hierarchical spatial segmentation. *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Location-Based Social Networks* (pp. 32–39). ACM.

Kennedy, L., Naaman, M., Ahern, S., Nair, R., & Rattenbury, T. (2007). How flickr helps us make sense of the world: context and content in community-contributed media collections. *Proceedings of the 15th ACM international conference on Multimedia* (pp. 631–640). ACM.

Keßler, C., Maué, P., Heuer, J., & Bartoschek, T. (2009). Bottom-up gazetteers: Learning from the implicit semantics of geotags. In: *GeoS '09: Proc. Third International Conference on GeoSpatial Semantics (Berlin, 2009)* (pp. 83–102). Springer.

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews* *in software engineering.* Keele, UK: Keele University and Durham University Joint Report.

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering: A systematic literature review. *Information and Software Technology, 51*, 7–15.

Klonner, C., Marx, S., Usón, T., Porto de Albuquerque, J., & Höfle, B. (2016). Volunteered geographic information in natural hazard analysis: a systematic literature review of current approaches with a focus on preparedness and mitigation. *ISPRS International Journal of Geo-Information, 5*(7), 103.

Kordopatis-Zilos, G., Papadopoulos, S., & Kompatsiaris, Y. (2015). Geotagging social media content with a refined language modelling approach. *Pacific-Asia Workshop on Intelligence and Security Informatics* (pp. 21–40). Cham: Springer.

Krishnamurthy, B., Gill, P., & Arlitt, M. (2008). A few chirps about twitter. *Proceedings of the first workshop on Online social networks* (pp. 19–24). ACM.

Kulshrestha, J., Kooti, F., Nikravesh, A., & Gummadi, P. K. (2012). Geographic dissection of the Twitter network. *Proceedings of the sixth international AAAI conference on weblogs and social media, Dublin, Ireland*.

Le, A., Pelechrinis, K., & Krishnamurthy, P. (2014). Country-level spatial dynamics of user activity: a case study in location-based social networks. *Proceedings of the 2014 ACM conference on Web science* (pp. 71–80). ACM.

Lee, C. H. (2012). Mining spatio-temporal information on microblogging streams using a density-based online clustering method. *Expert Systems with Applications, 39*(10), 9623–9641.

Lee, R., Wakamiya, S., & Sumiya, K. (2013). Urban area characterization based on crowd behavioral lifelogs over Twitter. *Personal and Ubiquitous Computing, 17*(4), 605–620.

Lee, S., Won, D., & McLeod, D. (2008). Tag-geotag correlation in social network. *SSM `08 Proceeding of the 2008 ACM workshop on Search in social media*.

Lei, L., & Hilton, B. (2013). A spatially intelligent public participation system for the environmental impact assessment process. *ISPRS International Journal of Geo-Information, 2*(2), 480–506.

Leung, D., Law, R., Van Hoof, H., & Buhalis, D. (2013). Social media in tourism and hospitality: A literature review. *Journal of Travel & Tourism Marketing, 30*(1–2), 3–22.

Li, C. and Sun, A., 2014. Fine-grained location extraction from tweets with temporal awareness. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval (pp. 43–52). ACM.

Li, C., & Sun, A. (2017). Extracting fine-grained location with temporal awareness in tweets: A two-stage approach. *Journal of the Association for Information Science & Technology, 68*(7), 1652–1670.

Li, C., Zhao, Z., Liu, S., Yin, L., & Luo, J. (2012). Relationships between geographical cluster and cyberspace community: A case study on microblog. *2012 20th international conference on geoinformatics (GEOINFORMATICS)* (pp. 1–5). IEEE.

Li, C., Zhao, Z., Luo, J., Yin, L., & Zhou, Q. (2014). A spatial-temporal analysis of users' geographical patterns in social media: A case study on microblogs. In W. S. Han, M. L. Lee, A. Muliantara, N. A. Sanjaya, B. Thalheim, & S. Zhou (Vol. Eds.), *Database systems for advanced applications, Dasfaa 2014. Vol. 8505. Database systems for advanced applications, Dasfaa 2014* (pp. 296–307).

Li, J., Qian, X., Tang, Y. Y., Yang, L., & Mei, T. (2013). GPS estimation for places of interest from social users' uploaded photos. *IEEE Transactions on Multimedia, 15*(8), 2058–2071.

Li, J., Qian, X., Lan, K., Qi, P., & Sharma, A. (2015). Improved image GPS location estimation by mining salient features. *Signal Processing: Image Communication, 38*, 141–150. (Recent Advances in Saliency Models, Applications and Evaluations). JOUR. Retrieved from http://10.1016/j.image.2015.07.007.

Li, R., Wang, S., Deng, H., Wang, R., & Chang, K. C. C. (2012). Towards social user profiling: unified and discriminative influence model for inferring home locations. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1023–1031). ACM.

Li, X., Larson, M., & Hanjalic, A. (2013). Geo-visual ranking for location prediction of social images. *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval* (pp. 81–88). ACM.

Li, Y., Crandall, D. J., & Huttenlocher, D. P. (2009). Landmark classification in large-scale image collections. *Computer vision, 2009 IEEE 12th international conference on* (pp. 1957–1964). IEEE.

Liang, C.-K., Hsieh, Y.-T., Chuang, T.-J., Wang, Y., Weng, M.-F., & Chuang, Y.-Y. (2010). Learning landmarks by exploiting social media. In S. Boll, Q. Tian, L. Zhang, Z. Zhang, & Y. P. P. Chen (Vol. Eds.), *Advances in multimedia modeling, proceedings. Vol. 5916. Advances in multimedia modeling, proceedings* (pp. 207–217).

Lingad, J., Karimi, S., & Yin, J. (2013). Location extraction from disaster-related microblogs. *Proceedings of the 22nd international conference on world wide web* (pp. 1017–1020). ACM.

Liu, B., Yuan, Q., Cong, G., & Xu, D. (2014). Where your photo is taken: Geolocation prediction for social images. *Journal of the Association for Information Science & Technology, 65*(6), 1232–1243.

Liu, Z. (2011). A survey on social image mining. *Intelligent Computing and Information Science, 134*, 662–667.

Lozano, M. G., Schreiber, J., & Brynielsson, J. (2017). *Tracking geographical locations using a geo-aware topic model for analyzing social media data.* Decision Support Systems.

Luo, F., Cao, G., Mulligan, K., & Li, X. (2016). Explore spatiotemporal and demographic characteristics of human mobility via Twitter: A case study of Chicago. *Applied Geography, 70*, 11–25.

Luo, J. D. J., Yu, J., & Gallagher, A. (2011). Geotagging in multimedia and computer vision: A survey. *Multimedia Tools and Applications, 51*(1), 187–211.

Maeda, T. N., Yoshida, M., Toriumi, F., & Ohashi, H. (2016). Decision tree analysis of Tourists' preferences regarding tourist attractions using Geotag data from social media. *Proceedings of the Second International Conference on IoT in Urban Space* (pp. 61–64). ACM.

Maguire, D. J. (1991). An overview and definition of GIS. *Geographical information systems: Principles and applications, 1*, 9–20.

Mahmud, J., Nichols, J., & Drews, C. (2012). Where is this tweet from? Inferring home locations of twitter users. *ICWSM, 12*, 511–514.

Mahmud, J., Nichols, J., & Drews, C. (2014). Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology (TIST), 5*(3), 47.

McClendon, S., & Robinson, A. C. (2013). Leveraging geospatially-oriented social media communications in disaster response. *International Journal of Information Systems for Crisis Response and Management (IJISCRAM), 5*(1), 22–40.

McDougall, K., & Temple-Watts, P. (2012). The use of LIDAR and volunteered geographic information to map flood extents and inundation. ISPRS annals of the photogrammetry. *Remote Sensing and Spatial Information Sciences, 1*, 251–256.

McGee, J., Caverlee, J., & Cheng, Z. (2013). Location prediction in social media based on tie strength. *International Conference on Information and Knowledge Management, Proceedings* (pp. 459–468).

Memon, I., Chen, L., Majid, A., Lv, M., Hussain, I., & Chen, G. (2015). Travel recommendation using geo-tagged photos in social media for tourist. *Wireless Personal Communications, 80*(4), 1347–1362.

Mendoza, M., Poblete, B., & Castillo, C. (2010). Twitter Under Crisis: Can we trust what we RT? In July (Ed.). *Proceedings of the first workshop on social media analytics* (pp. 71–79). ACM.

Middleton, S. E., Middleton, L., & Modafferi, S. (2014). Real-time crisis mapping of natural disasters using social media. *Intelligent Systems, IEEE, 29*(2), 9–17.

Mirani, T. B., & Sasi, S. (2016). Sentiment analysis of ISIS related tweets using absolute location. *Computational Science and Computational Intelligence (CSCI), 2016 International Conference on* (pp. 1140–1145). IEEE.

Musaev, A., Wang, D., Shridhar, S., Lai, C.-A., Pu, C., & Zhu, H. (2015). Toward a real-time service for landslide detection: Augmented explicit semantic analysis and clustering composition approaches. *2015 IEEE International Conference on Web Services (Icws)*, 511–518. http://dx.doi.org/10.1109/icws.2015.74.

Nagar, R., Yuan, Q., Freifeld, C. C., Santillana, M., Nojima, A., Chunara, R., & Brownstein, J. S. (2014). A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives. *Journal of Medical Internet Research, 16*(10), e236–e236.

Odlum, M., & Yoon, S. (2015). What can we learn about the Ebola outbreak from tweets? *American Journal of Infection Control, 43*(6), 563–571.

O'Hare, N., & Murdock, V. (2012). Gender-based models of location from flickr. *Proceedings of the ACM multimedia 2012 workshop on geotagging and its applications in multimedia* (pp. 33–38). ACM.

O'Hare, N., & Murdock, V. (2013). Modeling locations with social media. *Information Retrieval, 16*(1), 30–62.

Oksanen, A., Garcia, D., Sirola, A., Näsi, M., Kaakinen, M., Keipi, T., & Räsänen, P. (2015). Pro-anorexia and anti-pro-anorexia videos on YouTube: Sentiment analysis of user responses. *Journal of Medical Internet Research, 17*(11), e256–e256. http://dx.doi.org/10.2196/jmir.5007 (JOUR).

Panteras, G., Wise, S., Lu, X., Croitoru, A., Crooks, A., & Stefanidis, A. (2015). Triangulating social multimedia content for event localization using Flickr and Twitter. *Transactions in GIS, 19*(5), 694–715.

Paradesi, S. M. (2011). Geotagging tweets using their content. *FLAIRS conference*.

Pontes, T., Magno, G., Vasconcelos, M., Gupta, A., Almeida, J., Kumaraguru, P., & Almeida, V. (2012). Beware of what you share: Inferring home location in social networks. *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on* (pp. 571–578). IEEE.

Popescu, A., & Shabou, A. (2013). Towards precise POI localization with social media. *MM 2013 - Proceedings of the 2013 ACM Multimedia Conference* (pp. 573–576).

Poulston, A., Stevenson, M., & Bontcheva, K. (2017). Hyperlocal home location identification of Twitter Profiles. *Proceedings of the 28th ACM Conference on Hypertext and Social Media* (pp. 45–54). ACM.

Putri, A. N., Akbar, S., & Danar Sunindyo, W. (2016). Public facilities recommendation system based on structured and unstructured data extraction from multi-channel data sources. *Proceedings of 2015 International Conference on Data and Software Engineering, ICODSE 2015* (pp. 185–190).

Ranneries, S. B., Kalør, M. E., Nielsen, S. A., Dalgaard, L. N., Christensen, L. D., & Kanhabua, N. (2016). Wisdom of the local crowd: Detecting local events using social media data. *Proceedings of the 8th ACM Conference on Web Science* (pp. 352–354). ACM.

Rao, W., Du, Y., & Tang, M. (2016). A novel Chinese organization name extraction approach using CCRFs in micro-blog. *Computer and Communications (ICCC), 2016 2nd IEEE International Conference on* (pp. 2531–2535). IEEE.

Ribeiro, S. S., Jr., Davis, C. A., Jr., Oliveira, D. R. R., Meira, W., Jr., Gonçalves, T. S., & Pappa, G. L. (2012). Traffic observatory: A system to detect and locate traffic events and conditions using Twitter. *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Location-Based Social Networks* (pp. 5–11). ACM.

Roth, Y. (2016). Zero feet away: The digital geography of gay social media. *Journal of Homosexuality, 63*(3), 437–442.

Ryoo, K., & Moon, S. (2014). Inferring twitter user locations with 10 km accuracy. *Proceedings of the 23rd International Conference on World Wide Web* (pp. 643–648). ACM.

Sagcan, M., & Karagoz, P. (2015). Toponym recognition in social media for estimating the location of events. *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on* (pp. 33–39). IEEE.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. *Proceedings of the 19th international conference on World wide web* (pp. 851–860). Raleigh, North Carolina: ACM.

Sakaki, T., Matsuo, Y., Yanagihara, T., Chandrasiri, N. P., & Nawa, K. (2012). Real-time event extraction for driving information from social sensors. In May (Ed.). *Cyber*

Technology in Automation, Control, and Intelligent Systems (CYBER), 2012 IEEE International Conference on (pp. 221–226). IEEE.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2013). Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Transactions on Knowledge and Data Engineering, 25*(4), 919–931.

Salfinger, A., Schwinger, W., Retschitzegger, W., & Pröll, B. (2016). Mining the disaster hotspots-situation-adaptive crowd knowledge extraction for crisis management. *Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2016 IEEE International Multi-Disciplinary Conference on* (pp. 212–218). IEEE.

Sanborn, R., Farmer, M., & Banerjee, S. (2015). Assigning geo-relevance of sentiments mined from location-based social media posts. In E. Fromont, T. DeBie, & M. VanLeeuwen (Vol. Eds.), *Advances in intelligent data analysis XIV*. *Vol. 9385*. *Advances in intelligent data analysis XIV* (pp. 253–263).

Schulz, A., Hadjakos, A., Paulheim, H., Nachtwey, J., & Mühlhäuser, M. (2013). A multi-indicator approach for geolocalization of tweets. *ICWSM* (pp. 573–582).

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Lucas, R. E., Agrawal, M., ... Ungar, L. H. (2013). Characterizing geographic variation in well-being using tweets. *ICWSM*.

Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., & Haklay, M. (2017). A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science, 31*(1), 139–167.

Serdyukov, P., Murdock, V., & Van Zwol, R. (2009). Placing flickr photos on a map. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (pp. 484–491). ACM.

Sevillano, X., Valero, X., & Alías, F. (2015). Look, listen and find: A purely audiovisual approach to online videos geotagging. *Information Sciences, 295*, 558–572.

Shen, Z., Arslan Ay, S., Kim, S. H., & Zimmermann, R. (2011). Automatic tag generation and ranking for sensor-rich outdoor videos. *Proceedings of the 19th ACM international conference on Multimedia* (pp. 93–102). ACM.

Shirai, M., Hirota, M., Ishikawa, H., & Yokoyama, S. (2013). A method of area of interest and shooting spot detection using geo-tagged photographs. *Comp@SIGSPATIAL* (pp. 34–41).

Silva, T. H., Vaz de Melo, P. O., Almeida, J. M., Salles, J., & Loureiro, A. A. (2013). A comparison of foursquare and instagram to the study of city dynamics and urban social behavior. *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing* (pp. 4). ACM.

Simon, T., Goldberg, A., Aharonson-Daniel, L., Leykin, D., & Adini, B. (2014). Twitter in the cross fire—The use of social Media in the Westgate Mall Terror Attack in Kenya. *PLoS One, 9*(8), 1–11.

Steiger, E., Albuquerque, J. P., & Zipf, A. (2015). An advanced systematic literature review on spatiotemporal analyses of Twitter data. *Transactions in GIS, 19*(6), 809–834.

Straumann, R. K., Coeltekin, A., & Andrienko, G. (2014). Towards (Re)constructing narratives from georeferenced photographs through visual analytics. *The Cartographic Journal, 51*(2), 152–165.

Sultanik, E. A., & Fink, C. (2012). Rapid geotagging and disambiguation of social media text via an indexed gazetteer. *Proceedings of ISCRAM, 12*, 1–10.

Sun, Y., Fan, H., Helbich, M., & Zipf, A. (2013). Analyzing human activities through volunteered geographic information: Using Flickr to analyze spatial and temporal pattern of tourist accommodation. *Progress in location-based services* (pp. 57–69). Berlin Heidelberg: Springer.

Takhteyev, Y., Gruzd, A., & Wellman, B. (2012). Geography of Twitter networks. *Social Networks, 34*(1), 73–81.

Tamura, K., & Ichimura, T. (2013). Density-based spatiotemporal clustering algorithm for extracting bursty areas from georeferenced documents. *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on* (pp. 2079–2084). IEEE.

Tang, L., Ni, Z., Xiong, H., & Zhu, H. (2015). Locating targets through mention in Twitter. *World Wide Web, 18*(4), 1019–1049.

Thom, D., Bosch, H., Krueger, R., & Ertl, T. (2014). Using large scale aggregated knowledge for social media location discovery. In R. H. Sprague (Ed.). *2014 47th Hawaii international conference on system sciences* (pp. 1464–1473).

Tigunova, A., Lee, J., & Nobari, S. (2015). Location prediction via social contents and behaviors: Location-aware behavioral LDA. *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on* (pp. 1131–1135). IEEE.

Tran, T., & Lee, K. (2016). Understanding citizen reactions and Ebola-related information propagation on social media. *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on* (pp. 106–111). IEEE.

Van Canneyt, S., Schockaert, S., & Dhoedt, B. (2014). Estimating the semantic type of events using location features from Flickr. *Proceedings of the 8th Workshop on Geographic Information Retrieval* (pp. 11). ACM.

Van Canneyt, S., Schockaert, S., & Dhoedt, B. (2016). Categorizing events using spatio-temporal and user features from Flickr. *Information Sciences, 328*, 76–96. http://dx.doi.org/10.1016/j.ins.2015.08.032.

Van Laere, O., Schockaert, S., & Dhoedt, B. (2010). Towards automated georeferencing of flickr photos. *Proceedings of the 6th workshop on geographic information retrieval* (pp. 1–7). New York: ACM.

Van Zwol, R. (2007). Flickr: Who is looking? In November (Ed.). *Proceedings of the IEEE/WIC/ACM international Conference on Web intelligence* (pp. 184–190). IEEE Computer Society.

Velasco, E., Agheneza, T., Denecke, K., Kirchner, G., & Eckmanns, T. (2014). Social media and Internet-Based data in global systems for public health surveillance: A systematic review. *The Milbank Quarterly, 92*(1), 7–33.

Wang, Y., Fink, D., & Agichtein, E. (2015). SEEFT: Planned social event discovery and attribute extraction by fusing twitter and web content. *ICWSM* (pp. 483–492).

Wanichayapong, N., Pruthipunyaskul, W., Pattara-Atikom, W., & Chaovalit, P. (2011). Social-based traffic information extraction and classification. In August (Ed.). *ITS Telecommunications (ITST), 2011 11th International Conference on* (pp. 107–112). IEEE.

Watanabe, K., Ochi, M., Okabe, M., & Onai, R. (2011). Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs. *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 2541–2544). ACM.

Wen, Y.-T., Cho, K.-J., Peng, W.-C., Yeo, J., & Hwang, S.-W. (2015). KSTR: Keyword-aware skyline travel route recommendation. In C. Aggarwal, Z. H. Zhou, A. Tuzhilin, H. Xiong, & X. Wu (Eds.). *2015 IEEE international conference on data mining* (pp. 449–458).

Wen, Y.-T., Yeo, J., Peng, W.-C., & Hwang, S.-W. (2017). Efficient keyword-aware representative travel route recommendation. *IEEE Transactions on Knowledge & Data Engineering, 29*(8), 1639–1652.

Wendling, C., Radisch, J., & Jacobzone, S. (2013). *The use of social Media in Risk and Crisis Communication.* OECD Working Papers on Public Governance, No. 24Paris: OECD Publishing.

Williams, E. (2016). GeoContext: Discovering geographical topics from social media. *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on* (pp. 1342–1346). IEEE.

Xia, C., Hu, J., Zhu, Y., & Naaman, M. (2015). What is new in our city? a framework for event extraction using social media posts. *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 16–32). Cham: Springer.

Xu, D., Cui, P., Zhu, W., & Yang, S. (2014). Graph-based residence location inference for social media users. *IEEE Multimedia, 21*(4), 76–83.

Xu, J., Lu, T.-C., Compton, R., & Allen, D. (2014). Civil unrest prediction: A tumblr-based exploration. *Social Computing, Behavioral-Cultural Modeling & Prediction: 7th International Conference, SBP 2014, Washington, DC, USA, April 1–4, 2014. Proceedings* (pp. 403).

Yamaguchi, Y., Amagasa, T., & Kitagawa, H. (2013). Landmark-based user location inference in social media. *Proceedings of the first ACM conference on Online social networks* (pp. 223–234). ACM.

Yanai, K. (2015). [Invited Paper] A review of web image mining. *ITE Transactions on Media Technology and Applications, 3*(3), 156–169.

Yap, L. F., Bessho, M., Koshizuka, N., & Sakamura, K. (2012). User-generated content for location-based services: a review. *Virtual communities, social networks and collaboration* (pp. 163–179). New York: Springer.

Yoon, H. J., Tourassi, G., & Xu, S. (2015). Residential mobility and lung cancer risk: Data-driven exploration using internet sources. *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction* (pp. 464–469). Cham: Springer.

Yuan, G., Murukannaiah, P.K. and Singh, M.P., 2016. Percimo: A personalized community model for location estimation in social media. In Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on (pp. 271–278). IEEE.

Yuan, Q., Cong, G., Ma, Z., Sun, A., & Thalmann, N. M. (2013). Who, where, when and what: discover spatio-temporal topics for twitter users. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 605–613). ACM.

Yue, Y., Lan, T., Yeh, A. G., & Li, Q. Q. (2014). Zooming into individuals to understand the collective: A review of trajectory-based travel behaviour studies. *Travel Behaviour and Society, 1*(2), 69–78.

Zhang, W., & Gelernter, J. (2014). Geocoding location expressions in Twitter messages: A preference learning method. *Journal of Spatial Information Science, 9*, 37–70.

Zhang, X., Ji, S., Wang, S., Li, Z., & Lv, X. (2016). Geographical topics learning of geo-tagged social images. *IEEE Transactions on Cybernetics, 46*(3), 744–755.

Zhang, Y., Wu, W., Wang, Q., & Su, F. (2017). A geo-event-based geospatial information service: A case study of typhoon hazard. *Sustainability (2071–1050), 9*(4), 1.

Zhao, B., & Sui, D. Z. (2017). True lies in geospatial big data: Detecting location spoofing in social media. *Annals of GIS, 23*(1), 1–14.

Zhao, B., Sui, D. Z., & Li, Z. (2017). Visualizing the gay community in Beijing with location-based social media. *Environment and Planning A, 49*(5), 977–979.

Zheng, Y., Zha, Z., & Chua, T. (2011). Research and applications on georeferenced multimedia: A survey. *Multimedia Tools and Applications, 51*(1), 77–98.

Zheng, Y. T., Zhao, M., Song, Y., Adam, H., Buddemeier, U., Bissacco, A., ... Neven, H. (2009). Tour the world: building a web-scale landmark recognition engine. *Computer vision and pattern recognition, 2009. CVPR 2009* (pp. 1085–1092). IEEE.

Zhou, Y., & Luo, J. (2012). Geo-location inference on news articles via multimodal pLSA. *Proceedings of the 20th ACM international conference on Multimedia* (pp. 741–744). ACM.

Zhu, J.-Q., Lu, L., & Ma, C.-M. (2015). From interest to location: Neighbor-based friend recommendation in social media. *Journal of Computer Science & Technology, 30*(6), 1188–1200.