Published in IET Intelligent Transport Systems
Received on 30th December 2013
Revised on 6th September 2014
Accepted on 13th September 2014
doi: 10.1049/iet-its.2013.0214

ISSN 1751-956X

Enhancing transport data collection through social media sources: methods, challenges and opportunities for textual data

Susan M. Grant-Muller[1], *Ayelet Gal-Tzur*[2], *Einat Minkov*[3], *Silvio Nocera*[4], *Tsvi Kuflik*[3], *Itay Shoor*[5]

[1]Institute for Transport Studies, University of Leeds, Woodhouse Ln, Leeds LS1 3HE, UK
[2]Transportation Research Institute, Technion, Technion City, Haifa 32000, Israel
[3]Information Systems Department, University of Haifa, Mount Carmel, Haifa 31905, Israel
[4]IUAV University of Venice, Research Unit TTL - Transport, Territory and Logistics, Convento delle Terese - Dorsoduro 2206, I-30123 Venice, Italy
[5]Computer Science Department, University of Haifa, Mount Carmel, Haifa 31905, Israel
E-mail: S.M.Grant-Muller@its.leeds.ac.uk

Abstract: Social media data now enriches and supplements information flow in various sectors of society. The question addressed here is whether social media can act as a credible information source of sufficient quality to meet the needs of transport planners, operators, policy makers and the travelling public. A typology of primary transport data needs, current and new data sources is initially established, following which this study focuses on social media textual data in particular. Three sub-questions are investigated: the potential to use social media data alongside existing transport data, the technical challenges in extracting transport-relevant information from social media and the wider barriers to the uptake of this data. Following an overview of the text mining process to extract relevant information from the corpus, a review of the challenges this approach holds for the transport sector is given. These include ontologies, sentiment analysis, location names and measuring accuracy. Finally, institutional issues in the greater use of social media are highlighted, concluding that social media information has not yet been fully explored. The contribution of this study is in scoping the technical challenges in mining social media data within the transport context, laying the foundation for further research in this field.

1 Introduction

In recent years there has been a dramatic upsurge in the popularity of social media. Nowadays millions of people make use of a variety of online web platforms to express their opinions, thoughts and experiences. It has already been shown that social media can serve as a reliable resource for public opinions as well as factual information across several disciplines. Examples include the analysis of user opinions posted on social media relating to specific products in the marketing domain [1, 2], the assessment of political support rates as alternative to polls [3], aggregate financial measures [4–6] and the tracking of public health indicators to identify the outbreak of diseases [7, 8]. Social media has also created an opportunity for transport stakeholders and policy makers, where information flow has strategic importance both in long-term planning and short-term tactical system management. Specifically, social media may serve as a near real-time information source for tactical measures that require travel times, network demand or incident detection [9, 10] as well as supporting the development of strategic policies, such as those concerning levels of service quality [11, 12]. Carrasco *et al.* [13] mined social networks data in order to study activity-travel patterns. It is important to note that they used traditional tools as surveys and interviews for building social networks, while nowadays this information is freely available and easily accessible. Recently, Efthymiou and Antoniou [14] demonstrated the usefulness of today's social networks data for eliciting transport information in a study that analysed carsharing and bikesharing using a questionnaire distributed by email. They also noted and briefly demonstrated the potential of mining Twitter data for that purpose.

Compared with the traditional means of collecting user generated inputs (such as traffic counts, roadside and household surveys and focus groups), the main advantages of social media data are that very fresh and recent information can be obtained, plus it is a relatively low-cost method. The potential of a social media platform as a means to conduct transport surveys has already been recognised [15], and demonstrated [14]. However there are obstacles to overcome in order to extract useful information from social media effectively. First, there is the 'needle in a haystack' problem, whereby relevant data must be identified from a very large data mass. In addition, social media content is mostly in natural language form, which unlike other data forms, cannot be readily interpreted, queried or aggregated. For these reasons, relevant information must be 'harvested' using text mining techniques [16, 2]. A key question is whether social media information is of sufficient quality to meet the needs of the system operators/policy makers and travellers who, through traveller information systems, may also be end-users of the information.

To illustrate the type of transport-related content in social media and the challenges involved in the automatic processing of social media data, consider the following example short texts. Each corresponds to an authentic message posted on Twitter, a popular micro-blogging service.
Message 1: 'Just looked to get train to Liverpool to see giant exhibition. Two adults two children 358.20 could drive my 6cyl volvo up & back for 70'.
Categories: Transport-related? ('Yes') Subjective? ('Yes').
Message 2: 'Can you get a day rider that takes you to Liverpool on a bus other than the X1 obv an arriva one! Haven't got buses properly in ages :-('.
Categories: Transport-related? ('Yes') Subjective? ('No').
Message 3: A return ticket to Chester is 10 cheaper than a return ticket to Liverpool. Chester is further!!!!!'
Categories: Transport-related? ('Yes') Subjective? ('Yes').
Message 4: 'Time to walk the dawn - seeing what is going on in the football can wait a while'.
Categories: Transport-related? ('No') Subjective? ('No').

Messages 1–3 discuss transport-related topics, and may be of interest from several perspectives. For instance, Messages 1 and 2 indicate user intention to travel to a specified destination. If a large pool of similar messages were processed it may be possible to elicit popular trip routes and needs. Messages 1 and 3 also both include opinions on the quality of public transport (PT) services (specifically, relating to price). Although such information may be useful, the processing of textual messages into meaningful data is not straightforward. Crucially, only a proportion of social media messages concern genuine transport issues – Message 4, for example, contains the term `walk', however not in a transport context.

The task of automatically associating text with specified categories such as `transport' or 'subjectivity', is a well-researched field in text mining that is typically addressed using classification methods. Another important component of text mining is information extraction (IE). IE tasks include identifying names in the text (e.g. location names, underlined in the example above) and processing different references to the same entity unambiguously, for example, 'NYC' and 'New York'. Classifying message content, identifying location names and treating name variations are just three challenges that need to be addressed in order to obtain useful information on transport-related topics from textual content.

In this paper, the goal is therefore to review both the opportunity and challenges that are involved in harvesting large amount of social media in the area of transport. Specifically, we are interested in addressing the following questions:

(1) In which ways can social media data be used alongside or potentially instead of current transport data sources?
(2) What technical challenges in text mining social media data create difficulties in generating high-quality data for the transport sector?
(3) Are there wider institutional barriers in harnessing the potential of social media data for the transport sector in addition to the technical challenges?

The answers to these questions will enable policy makers to assess the challenges in obtaining information from social media text sources and subsequently either fuse the data with that from other sources or process it as an additional data stream.

This paper provides a review of the state-of-the-art in the text mining techniques needed to obtain transport-related data from social media text sources. A reflection on institutional issues is also made to assess whether technical or institutional issues are the greater obstacle, informing the direction for future research. This paper continues in Section 2 with a summary of uses of current transport data relationships with social media data. An overview of text mining methodology is given in Section 3, whereas Section 4 outlines text mining challenges, the state-of-the-art and a reflection of the implications for transport data requirements. In Section 5, the evidence concerning wider barriers to transport stakeholders (local government, practitioners and suppliers) using social media is outlined and Section 6 concludes.

2 Role of data in transport planning and policy

Information flow plays a central role in the decisions made by transport system users in how, when and whether to travel. It also supports recovery in cases of unexpected disruption [17, 18]. The question is therefore whether current data streams can be either integrated with (or replaced by) new sources, to provide cost effective and potentially more complete information. The many sources of new technology-enabled data include textual social media, geographic information systems and digital data from intelligent transport systems. The potential for information enrichment arises from data collection at various levels of aggregation and with some sources providing associated 'clues' to the socio-economic characteristics associated with individual data units.

A typology of primary data needs, data sources and the potential role of social media is given in Table 1, which draws on sources including the Common Highways Agency Rijkswaterstaat Model specifications [19] and other PT service performance specifications (for example [17]). Existing transport systems often comprise layers of technologies and monitoring equipment that have accumulated as technology has advanced [20]. However the distribution of instrumentation can be patchy, resulting in some geographic areas with dense data collection and others with sparse data (typically rural). This gives rise to two challenges: the first is whether new data forms can be integrated with other data sources to create new or better quality knowledge [13] and the second is whether there is the possibility to adopt various 'user generated data' where current data collection is sparse [21]. Uses may include monitoring system performance, informing new policies based on expected demand, providing cost effective, more detailed and potentially more complete information on context, improving understanding of behaviour and perceptions that underlie mode choice [22] and enriching the understanding of scheme impacts [23–27]. They may serve to improve the efficiency and effectiveness of current databases, for example, through reconciling data contradictions and reducing redundant data collection. To answer both challenges fully, a further significant tranche of research is needed. The remainder of this paper therefore focuses on one particular stream of new data only, that is, social media textual data.

doi: 10.1049/iet-its.2013.0214

Table 1 Overview of current and new transport data sources (source: authors)

Data purpose	Data issues	Potential role of social media or new technology data
speed monitoring inputs to estimated travel time for traffic management and advanced traveller information systems (ATIS) **current data sources:** (1) ANPR camera (2) loops (e.g. motorway incident detection and automatic signalling (MIDAS)) embedded in the highway	**(1) automatic number plate recognition (ANPR)** • automatic data collection following installation (fixed location) • can capture large proportions of population at location • costly to maintain • accuracy in plate reading and matching, for example, in poor weather • data are obtained without explicit consent • no user characteristics for each trip • applicable to motorised vehicles only	**GPS tracking** (e.g. on mobile phone) • can generate speeds • tracking has some errors in measurement • data are generated automatically • GPS data are given by consent • can be applied to non-motorised modes (such as bike riders and pedestrians)
	(2) loops (MIDAS) • automatic data collection following installation (fixed location or temporary loops) • can capture large proportions of population at location • faulty loops generate data gaps and errors, which can be substantive • loop maintenance costs and costs of downloading/processing data • speed is inferred and smoothed, missing spikes in actual vehicle speed • no user characteristics for each trip • applicable to vehicles only	**cellular-based monitoring** • less accurate than GPS based (accuracy is usually proportional to cell size) • however, it does not require activation of any GPS-based app • the number of samples is much higher (to compensate for the accuracy of identification of any single probe) • cellular operators provide the data to the companies calculating speeds/travel time • privacy is preserved by assigning each probe an ID which is different than the original ID of phone
constructing O–D movements input estimates of demand to policy decisions and traffic management strategies **current data sources:** (1) ANPR camera and (2) RP/SP questionnaires	**(1) ANPR** (issues as above) **(2) revealed preference (RP)/stated preference (SP)** • allows bespoke design and can collect user characteristics • resource intensive • potential sampling and other sources of bias, for example, in administration of complex design • response rate/participation may not be high	**social media text content** • content can contain O–D data, but coverage may be limited • depends on users choosing to contribute content • socio-economic and contextual data may be present alongside O–D **GPS tracking** • can generate individual O–D movements for full trip
link demand input estimates of demand to policy decisions and traffic management strategies **current data sources:** (1) roadside counts, (2) loops (e.g. MIDAS) embedded in the highway and (3) RP/SP questionnaires	**(1) roadside counts** • manual process, resource intensive therefore limited sample possible • human and other errors **(2) MIDAS** • inaccuracies in vehicle classification • other issues as above (3) **RP/SP** (issues as above)	**social media text content** • may generate additional data on demand, but unlikely to capture total demand • unlikely to identify specific link within a journey as part of content **GPS tracking** • can generate link demand, but would need large-scale monitoring to estimate total demand
PT mode demand inputs to policy making and/or commercial decision making by suppliers **Current data sources:** (1) in-mode counts, (2) patronage data (e.g. ticket sales) and (3) RP/SP questionnaires	**(1) in-mode counts** • can be targeted to areas/services of interest • manual, therefore resource intensive • limited sampling practical **(2) patronage data** • automatically collected • large proportions of population can potentially be captured • commercially sensitive • some inaccuracies, for example, in cross-mode (3) **RP/SP** (issues as above)	**social media text content** • useful for understanding mode choice rationale • unlikely to capture total demand • may capture evidence on demand for responsive services **GPS tracking** • can generate evidence on mode demand • large-scale monitoring needed to estimate mode demand with confidence
service quality and driver comfort inputs to policy planning and operations **current data sources:** (1) RP/SP questionnaires	(1) **RP/SP** (issues as above)	**social media text content** • analysis of text content effective in generating service quality data
public opinion (e.g. new schemes or services) inputs to long-term policy development and planning decision **current data sources:** (1) focus groups, committees and consultation meetings and (2) household questionnaires	**(1) groups and meetings** • resource intensive • limited samples possible • sources of bias, for example, in participation **(2) household questionnaires** • resource intensive • Some biases, for example, in response rates	**social media text content** • analysis of text content effective in supplementing or replacing public opinion data sources
detection of abnormal or undesirable event (various modes of transport) inputs to operations and ATIS. Includes incidents along the road network, train delays, packed bus, missing rental bikes etc. **current data sources:** (1) various types of physical devices (e.g. video cameras, loops, in-mode counters etc.) and (2) systems (such as patronage data, bike rental systems etc.)	**(1) physical devices** • continuous monitoring • level of accuracy is usually sufficient • high coverage is often costly **(2) management/operational/control systems** • systems often belong to private operators and the quality of data sharing is often a challenging issue • such systems do not necessarily enable real-time data processing which is required for event detection	**social media text content** • low cost for authority • even a small number of similar reports constitute a solid basis for verifying the event • many types of events can be detected in the same manner • depends on human reporting • time constraints require the use of very efficient text mining techniques

The premise that social media contains valuable transport information forms the rationale for the examination here of the mining task needed to extract the data for use. The characteristics of social media that are of particular value for the transport sector include: the potential for all users to contribute content, dematerialisation of data collection, community facilities (such as discussion boards/blogs, video sharing) and virtual meetings. Information harvesting may be: (i) dynamic, informing short-term decisions by system operators and users or (ii) off-line, supporting policy makers and stakeholders in forming improved policies. An overview of the text mining process follows as a background to a discussion of specific transport-related challenges in Section 3.

3 Mining transport data from social media text

Text mining uses a set of techniques and tools which need to be carefully adapted for the task in hand [28]. Fig. 1 outlines a general flow of the text mining process, as applied to social media, with the following main steps.

3.1 Initial message filtering

Owing to the computational intensity of the task, a set of potentially relevant messages must first be identified from the general social media message stream. Meta-data is often useful for this purpose if available. For example, Twitter's streaming application programming interface (an interface provided by Twitter for access to the real-time tweet stream) allows message filtering using criteria such as date and geographical meta-data. In addition, keyword specification allows the extraction of messages that contain a set of pre-specified words from the general stream, which may be highly effective in asserting message relevance in some contexts. For example, in the political arena, messages containing 'Obama' were analysed to assess presidential approval rates over time [6]. In the transport domain [12], a list of train names was used to filter social media message content, with the goal of eliciting user opinions on the transit system from social media. Further research by Mai and Hranac [9] involved the collection of incident statistics from social media. A set of word collocations was used to filter potentially relevant messages including the collocations 'traffic accident' and 'car crash'.

In this paper, we consider the scenario where a transport authority is interested in processing social media data about a wide variety of transport issues within its remit. Message relevancy in this case requires that texts are related to a broad range of transport issues. A reasonable strategy for keyword specification would be to use keywords that are typical for the transport sector, possibly deriving them from an existing transport lexicon or ontology. Although some contexts may involve keywords with unique meaning (such as 'Obama', 'influenza' or train names), typical words for the transport sector may be highly ambiguous. For example in the texts 'cross the bridge when we get there', or 'wash the car', the terms 'bridge' and 'car' are associated with transport, but are used with an irrelevant sense. Filtering messages by keywords may therefore yield very noisy results. However, having identified candidate messages using the initial criteria, an improved assessment of relevance and more detailed interpretation of content can be performed using further text mining steps, as described below.

3.2 Semantic annotation

The initial pool of filtered raw texts ('source texts', Fig. 1) can be further annotated with useful semantic information [29]. Specifically, named entity recognition (NER) techniques annotate the scope and types of entities of interest, including place names, facilities, organisation and person names. Recent NER models have been adapted to handle informal text such as social media [30, 31]. It is also useful to annotate transport-related concepts in the text, linking textual phrases to domain ontology as discussed in Section 4. This level of annotation can be used to assist in further decoding the meaning of the whole message, while place names provide evidence on the location orientation of the message (see Section 4.3).

3.3 Message relevancy

The relevance of annotated messages to the transport authority can then be more thoroughly evaluated. The automatic association of text with a topic – transport here – typically uses supervised machine learning approaches. In these approaches, a model is learned based on labelled examples, implying that a 'dataset' must be constructed containing example texts with their correct labels. Manual labelling may be costly, especially if domain expertise is required. To learn a classification model that fits labelled examples and generalises to new examples, example texts are abstracted into pre-defined 'feature' values. In the popular 'bag-of-words' feature schema, a document is represented as an unordered set of the words [16]. This simple representation can give good performance, for example, documents containing the terms 'train', 'bus', and 'ticket' are likely to be transport-related. Similarly, it may be useful to model word bigrams, or trigrams, capturing collocations like 'car accident'. Enhanced feature schemes encode additional semantic information rather than surface words; for example, features indicating whether location names or transport-related terms are observed in the text, as indicated by the semantic annotation [29]. Various classification paradigms are known to give good performance on text categorisation problems, including support vector machines, Bayesian models and more [16]. Similar classification techniques are well established in other fields of transport modelling, for example, [32]. Once classified, messages that are identified as irrelevant to transport are discarded at this stage. Finally, another aspect of relevancy is the location orientation of a message. We discuss methods for further identifying messages that are relevant with respect to location in Section 4.3.

3.4 Semantic processing

Messages judged as relevant can then be classified into finer categories within the transport domain; for example, identifying messages that report accidents, messages in which users express a wish to travel to some known destination etc. Similarly, messages may be automatically associated (using dedicated classifiers) with transport modes or sentiment analysis may be used (see Section 4.2, [30]).

3.5 Summarisation and presentation

The final stage is to aggregate and present the mining outcomes, to support decision making [33]. For example,

doi: 10.1049/iet-its.2013.0214

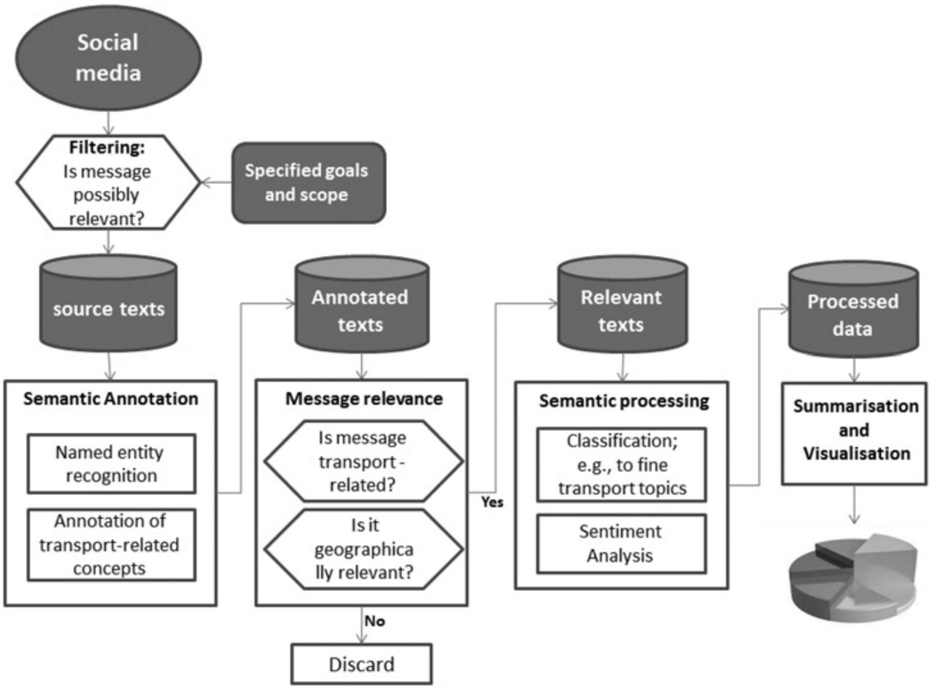
Social media
Filtering: Is message possibly relevant?
Specified goals and scope
source texts
Semantic Annotation
Named entity recognition
Annotation of transport-related concepts
Annotated texts
Message relevance
Is message transport-related?
Is it geographically relevant?
Yes
No
Discard
Relevant texts
Semantic processing
Classification; e.g., to fine transport topics
Sentiment Analysis
Processed data
Summarisation and Visualisation

Fig. 1 *Overview of text mining process (source: authors)*

graphical presentations of positive against negative public sentiment [34] towards service.

Automated text mining is inherently imperfect, however this should not imply the data cannot be used within the transport information cycle. To the contrary, an appreciation of where data quality is either strong or weak allows a more confident utilisation of the data. Performance considerations that are particularly relevant to the transport sector are discussed in Section 4.4.

4 Text mining challenges for transport data needs

In this section, some more detailed attention is given to four challenging areas of particular relevance to uses of social media in transport. The aim is to highlight the state-of-the-art in the technical process and reflect on the implications for increased uptake of social media as a transport information source.

4.1 Transport ontology

A main barrier to automating text processing in general, (and micro-blogs such as Twitter in particular) is the lack of accompanying context. By way of illustration, to infer that the text 'the 61C was late this morning' is relevant to transport, it must be known that '61c' is the name of a bus line and is being used intentionally in a transport context. One of the most effective ways to represent background (world) context is through ontologies. 'Ontologies' serve as a methodological framework for representing contextual information as a networked structure of objects or concepts, with related items linked by labelled relationships. Freebase [35] is a popular example of a general-purpose ontology. The mining process involves 'annotating' text, linking text segments to concepts in the ontology, rendering the text amenable to semantic search and processing. Following the example above, an ontology is needed that represents the term '61c' as an entity, connected with an 'is-a' (hyponymy) relation to the concept 'bus', where 'bus' in turn is mapped as a hyponym of the 'transport mode' concept etc. This allows an association of the text with transport categories at various granularities; for example, 'transport', 'transport mode', 'bus' etc.

A literature review of transport-related ontologies reveals two main categories. The first concerns the type of activity for which the ontology was created. Some work has focused on very specific tasks such as the transmission of communication between in-vehicle and external systems [36]. Others have targeted more general processes such as micro-simulation [37] or journey planning [38]. Generally, an ontology required for specific activities is narrower than one for more general processes. The second category concerns the transport mode the ontology covers, with some focusing on a single mode while others cover multi-modal travel. Combining both categories of ontology-related transport research means that all combinations appear in the literature:

- *Ontologies addressing a specific activity and a single mode:* Private vehicle context-aware services [36]; customer satisfaction of travellers of mass transit system [39]; situation awareness of city tunnel traffic [40].
- *Ontologies addressing a specific activity and multi-modal journey:* Military transport planning and scheduling [41].
- *Ontologies addressing a general activity and a single mode:* Personalised private vehicles route planning [42]; activity-based carpooling micro-simulation [37].
- *Ontologies addressing a general activity and multi-modal journey***:** Public transport query system [42]; journey planning [43].

Despite the substantial contribution of existing research, the work to date only provides a partial solution for the problem of creating overall comprehensive transportation ontology (see also [44]). Constructing such ontology would be resource intensive as it involves the abstraction and conceptualisation of the transport domain, typically conducted by domain experts. Fusing existing ontological resources may alleviate this effort and some attempts in this direction have already been made, for example, [45]. The dynamic and geography dependent nature of the transport-related social media content further contributes to the complexity in creating ontology. A full-scale ontology should, for example, capture the reality in which the underground system in London (UK) is called 'tube', whereas at the same time 'T' is commonly used for informally referring to the one in Boston (USA). To the best of our knowledge, this aspect has not yet been dealt with by researchers in this field. Ideally, a transport ontology would also be maintained using collaborative intelligence and drawing on contributions by non-experts, in a similar fashion to Wikipedia. Given the current lack of such a resource for the transport community, future research activities are likely to include modelling relevant semantic information given pre-specified tasks and consolidating dictionaries that are available in different formats.

4.2 Sentiment analysis

Sentiment analysis (or 'opinion mining') is the process of extracting opinions concerning an event or an entity from the text. This area generally has drawn a lot of recent attention [2], with the bloom of user generated content in social media boosting research efforts [46]. Addressing some of the challenges in sentiment analysis is important in order that social media plays an increasing role in the transport sector information loop. Opinion data includes bus, train or plane passenger views (e.g. on service quality) and in governance processes that include public participation, for example, consultations concerning new transport schemes [47].

Sentiment analysis often begins by creating of a lexicon of words marked with their prior polarity, that is, negative/positive [48], which is then used to analyse emotions in the text [49]. Sood *et al.* [49] suggested a methodology for detecting sentiments based on three steps: (i) collecting a training corpus of texts, often manually annotated according to the sentiments expressed in them; (ii) building a set of categories associated with positive, negative and neutral sentiments; and (iii) training a system to classify new texts automatically into the desired categories. Pang and Lee [2] claim that in order to perform a sentiment analysis task, labelled training data from within the same domain must be used. As with the issues surrounding ontology (see Section 4.1), some consideration to a mode-specific or task-specific lexicon may be appropriate. Given the nature of natural language, identifying negative/positive meaning is challenging [50]. For example, 'busy' may be positive in describing some transport contexts, for example, 'the road is busy and should qualify for upgrade', but negative in others 'the road is busy and unsuited for further housing development'. A text may say that a policy is 'not at all desirable' (negative sentiment) or a product is 'terribly good' (positive sentiment). Natural language may also include irony and sarcasm which add to the challenge [51]. Analysis of transport sentiment data [52] illustrated the difficulty with sarcasm in service quality related text. The message 'train service is just fantastic' needs the surrounding context for interpretation. In this case, clues in the preceding or subsequent content (e.g. relating to late running trains) may indicate whether it is genuine or sarcastic. Inferring sentiment can be posed as a text classification task [53], associating the text with the categories of positive, negative and neutral sentiments. Pang and Lee [2] claim that labelled training data from within the same domain must be used. As with ontology (see Section 4.1), some consideration of either a mode-specific or task-specific lexicon in the learning process may be appropriate. The social media platform from which the content is harvested might also influence sentiment analysis. For example, sites dedicated to complaints, such as Hellopeter.com, might be biased towards negative sentiments [47]. Despite the challenges and the fact that there are great differences in success of sentiment analysis in different domains, Pang and Lee [2] noted that machine learning techniques can achieve >80% accuracy in sentiment analysis.

In recent years, Twitter has been the target of intensive research with respect to content analysis and especially opinion mining, given its nature as a short and immediate response to events. Challenges in mining sentiments in Twitter's textual contents are exacerbated by the length of the text, its contextual nature and its lifespan. However, it has been widely used for estimating public mood [53], trends such as stock market behaviour [54], political elections results [55] and also in the transport sector. Collins *et al.* [12] used Twitter as an information source for evaluating transit rider satisfaction. Focusing on the rapid transit system of the Chicago Transit Authority as a case study, researchers have found a correlation between irregular events (such as extreme delays) and the volume of negative sentiments. This correlation supports the notion that Twitter is a valid source of information for inferring transport-related sentiments. The short message length creates difficulties in the transport sector as in others. Still, it has also the advantage that users have a lower payload in sending a message than for other types of social media, for example, Facebook. The need for contextual data is potentially more of an issue than for other domains because of the dynamic and spatial nature of travel and with journeys often involving more than one mode. The lifespan of transport sentiment data is possibly less of a problem given the links into both long-term planning and short-term responses (Section 1).

4.3 Location data

Most transport operators and managers are primarily concerned to identify transport-related information from social media that is closely associated with the transport services for which they have responsibility. It is a reasonable assumption that most messages posted on the formal websites for a transport authority (or supplier) will have relevance to the associated locality. However, the transport system inherently contains networks (e.g. of roads, PT services). As a result, both upstream and downstream transport activities may be of relevance to a particular geographic location. The governance of particular sections of the transport system that together form networks may be undertaken by different authorities with different websites. For example, complaints about connections between inter-city and local services may be posted on the web site of inter-city service operators, but be of interest to local providers seeking to improve connection services.

It is therefore necessary to identify those messages (from the very many that will be available) relevant to the location and/or specific transport services for the task. Two

doi: 10.1049/iet-its.2013.0214

possible location identification approaches are either (a) to identify the current location of the person posting the message and/or (b) to correctly identify locations from message content. Fig. 2 outlines the process involved for an example case of PT messages based on the fusion of information either within the message or attached to it.

A primary source of information on the location of the person posting the text message is voluntarily posted geo-meta-data associated with the social media-user ('user') account. In practice, many users do not provide this information [9] and even if a message is geo-tagged, it may be inaccurate. The message may also relate to transport in locations distinct from the users home town, for example, while travelling. Mobile device global positioning system (GPS) coordinates offer further implicit meta-data indicating the users' location, but is only a portion of all social media traffic and user consent is required to enable the functionality. Research continues to maximise the precision of location inference from pervasive devices [56]. Given current limitations in coverage of these types of meta-data [57, 58], other implicit information sources have been investigated for potential location inference. Social network structures can be used for this purpose as users tend to live in close geographic proximity to their social network peers [59]. An estimate of user location may be inferred based on the message content [60]. In particular, it has been shown that fine geographical distinctions are possible based on local language characteristics [61, 62].

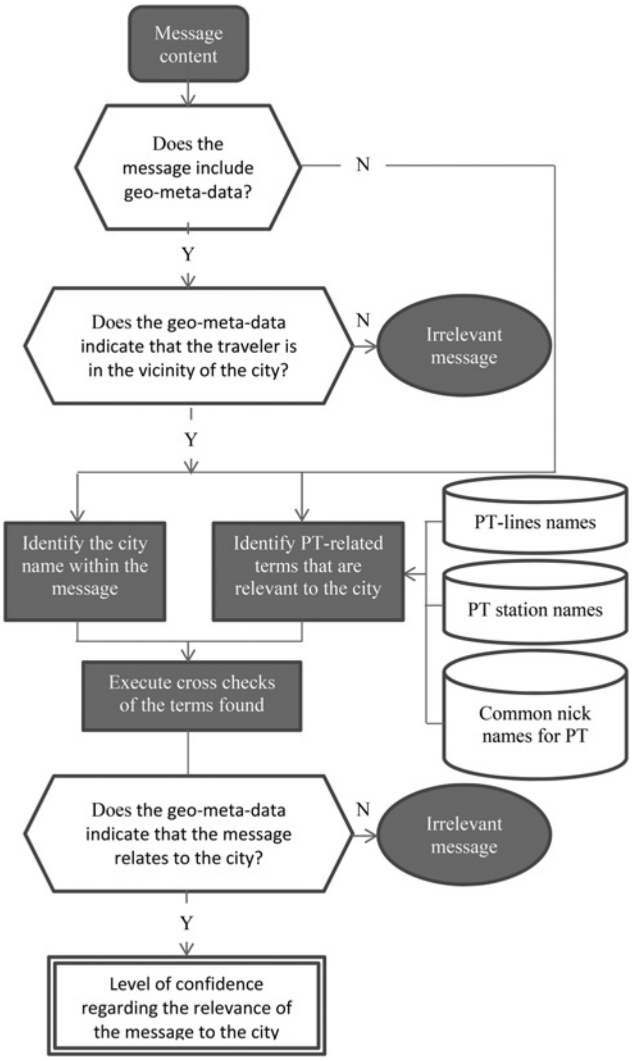
Message content
Does the message include geo-meta-data?
N
Y
Does the geo-meta-data indicate that the traveler is in the vicinity of the city?
N
Irrelevant message
Y
Identify the city name within the message
Identify PT-related terms that are relevant to the city
PT-lines names
PT station names
Common nick names for PT
Execute cross checks of the terms found
Does the geo-meta-data indicate that the message relates to the city?
N
Irrelevant message
Y
Level of confidence regarding the relevance of the message to the city

Fig. 2 *Analysis of geo-location data in social media transport data*

The second approach to identifying location data is from the contents of the message. This task is especially challenging when considering the high ambiguity of names of places. For example, 'Liverpool' is the name of a UK city, a London rail station (Liverpool Street), a city in the USA and an Australian suburb.

Several approaches have been proposed for identifying geo-location based on message content. NER techniques can automatically annotate the text with mentions of entity names. Having extracted candidate location names, disambiguation is needed to align inferred location with any other contextual information, in conjunction with relevant sources of location names.

Web-a-Where, a system for associating geography with web pages was one of the early works that have tackled this problem [63]. TwitterTagger [64] geotags tweets based on comparing their content with the United States Geological Survey (USGS (http://geonames.usgs.gov/)) database of locations. Another approach for content-based geo-location of multilingual tweets is based on collating contextual tweets into a document using a user-tweeting-frequency-based temporal window [65]. An approach based on 'local' words, for example, words that are typical to specific location such as 'Hoody' for Texas, was proposed by Cheng *et al.* [66] for estimating a Twitter-user city level location.

The availability of data sources containing transport-related entities (e.g. PT line identifications, station formal and informal names, names of parking facilities) may constitute a valuable asset for identifying locations for transport-related messages. Bry *et al.* [67] provide interesting examples of the use of such data sources in building a world model for geospatial data. The world model presented consists of concrete data (such as train connections) as well as logically formalised ontologies of transport networks. Following conjectures on the location of the message based on the different approaches, the data analyst can check for possible inconsistencies and choose whether to discard messages where there is low confidence in geographical orientation.

4.4 Measuring the accuracy of the text mining results

A quantitative evaluation of the text mining process (Section 3) is needed to tune the system and evaluate the degree of success. Evaluation generates common performance measures originating from the information retrieval domain [12], namely precision and recall. Precision measures the accuracy of the predictions made by the system (i.e. how many of the texts classified as relevant are indeed relevant). Recall corresponds to coverage ratio (how many of the total transport-relevant texts were classified as relevant). Given a dataset of examples associated with correct class labels on one hand, and automatically inferred labelled on the other hand, it is possible to evaluate the system's performance both in terms of precision and recall.

An important step in social media text mining is to classify an initial pool of texts as relevant or not (Section 3). For the potential uses here (see Table 1), a text is considered relevant if it contains either objective or subjective information about the transport system or a journey undertaken by an individual (s). This may include origin and/or destination information, mode-specific commentary, opinions and experiences about transport or transport-related activities, observations on the state of the system and information concerning system changes or interventions. In practice, however, relevance is a subjective notion, difficult to define deterministically [12] and subject to personal interpretation. To evaluate subjectivity, it is common practice to have a shared set of examples manually annotated with the target class. The inter-annotator agreement rate provides an indication of the level of subjectivity in the task. Automatic methods that learn a concept class based on annotated examples cannot be expected to outperform the inter-annotator agreement rate.

The infinite nature of the message stream in social media is challenging from several perspectives. From a performance perspective, as the content posted on social media changes rapidly over time, periodic monitoring and possibly re-tuning of the system is required. From an evaluation perspective, it is impossible to identify all relevant messages in the data stream and as a result one cannot compute recall precisely. The large mass of data on social media however also carries an important advantage. Social media information is characterised by a high degree of redundancy, having multiple messages that are phrased differently conveying similar content. This means that while some relevant messages may be overlooked by the text mining process this may not have a drastic effect on the output of the text analysis process. It has been shown that text analysis of social media can yield results that are consistent with formally conducted polls [28].

The individual components of the text processing pipeline, including the classifier and text annotators can be evaluated using labelled examples that were set aside for testing purposes. Each component is typically tuned until the output performance measures are considered satisfactory. In general, an important factor affecting the performance of learning systems is the size of the labelled data that is available to learn from. Rather than manually label large amounts of data, which are costly, automatic and semi-automatic methods for labelling examples can be applied. Using the pseudo relevance feedback approach, for example, texts that are classified with high confidence at early iterations are processed as labelled examples to retrain improved models [12]. Finally, semi-automatic settings may be preferred where the text processing outputs are further assessed by a human.

Summarising the findings overall, text mining provides a means for automatic identification of transport-relevant messages in a stream of incoming messages. Specific challenges remain and solutions are needed where the user must be in the loop for periodic monitoring and enhancement of the system.

5 Harnessing social media data in practice

This section seeks to address the final research question: are there wider institutional or other barriers in harnessing the potential of social media data in transport in addition to the technical issues? A review of institutional attitudes to social media use is followed by some findings on social media use by transport authorities in practice.

Although there is a growing tranche of literature concerned with the attitudes and perceptions of 'individuals' in social media use, rather less has been published on the formal stance of 'organisations' on social media use. This is particularly the case for those in a governmental (or public sector) role, which is often the case for transport sector. Given the important role of information in both operational activity and strategic planning [47], improved understanding of the barriers and enablers to accelerating effective uptake of social media in the transport sector is needed. Initial attempts to provide authorities with guidelines for effective use of social media have already been made [15, 68]. Based on more general social media literature, the following points arise as possible organisational stances.

A reluctance to engage with social media may arise from the need to be active as a 'key requirement of success' [58], potentially related to the need for resource input (see also [52, 69, 70, 11, 71]). There may well be concerns about safeguarding corporate image, given the dynamic nature of messaging. Less opportunity for 'lagged' responses may

Table 2 Summary of findings on the use of social media by authorities

Authority	Summary of findings
Buckinghamshire and City of Edinburgh [73]	using Twitter to give timely information to drivers about conditions on the road network
Transportation Safety Board of Canada [74]	mainly used as an alternative method for accessing the material shared through the Transportation Safety Board of Canada's really simple syndication feeds and websites
American Association of State Highway and Transportation Official, Third Annual State DOT Social Media Survey [72]	just <90% of states are using Twitter. More than three-quarters of states are using Facebook
Minnesota's Local Road Research Board [75]	25 cities and 25 counties were selected for closer examination. Among the 50 governments sampled, Facebook was found to be the most common social media outlet (used by 19 for any reason and by 10 for transport communications) followed by Twitter (used by 15 for any reason and by nine for transport). Across all social media channels, the most common transport-related topics for communication were planning and zoning road construction and street closures
New York region's major transport providers [76]	links to social media accounts from the homepage, in tandem with service alert tools. This feature allows riders to comprehend urgent information in the context of social media's engagement, showing the complementary nature of the different resources
various US authorities [77]	although it should be noted that social media could greatly enhance a transport organisation's public outreach, there are also some potential dangers. Some aspects of social media are beyond the authority's control: hackers are a threat, people will post old or false news and some leaks will occur
Virginia DOT [78]	the Virginia DOT is expanding its use of social media to communicate with the 7.5 million Virginians who depend on us to connect them with the things that are most important in their lives
local authorities in California [69]	cities are generally more interested in information-sharing through social media than constituent engagement
Bay Area Rapid Transit (BART), Oakland, CA [70]	Facebook page is mostly used to promote contests, highlight agency news and make followers aware of upcoming public hearings. The Twitter account mostly includes service alerts

doi: 10.1049/iet-its.2013.0214

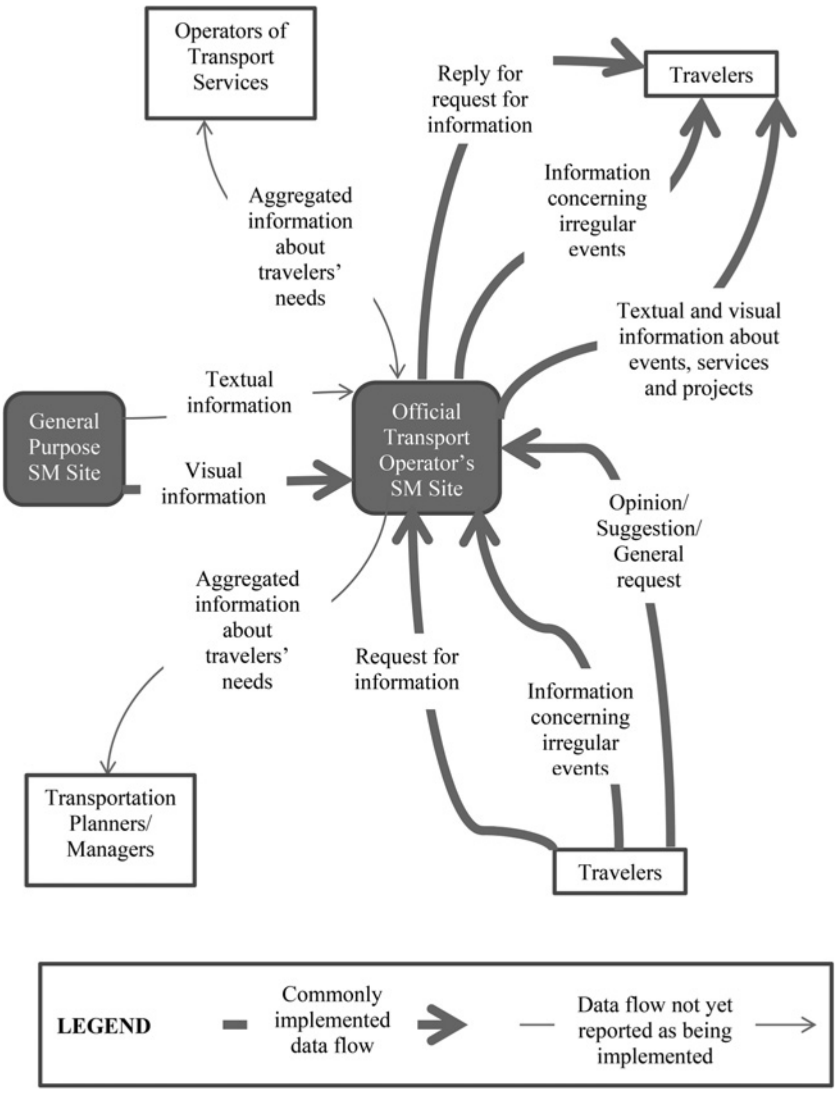
Operators of Transport Services
Reply for request for information
Travelers
Information concerning irregular events
Aggregated information about travelers' needs
Textual and visual information about events, services and projects
Textual information
General Purpose SM Site
Official Transport Operator's SM Site
Visual information
Opinion/ Suggestion/ General request
Aggregated information about travelers' needs
Request for information
Information concerning irregular events
Transportation Planners/ Managers
Travelers
LEGEND
Commonly implemented data flow
Data flow not yet reported as being implemented

Fig. 3 *Flow of information to and from transport authority social media sites*

give rise to fears around 'sending out the wrong messages'. [52] outline evidence of a 'code of conduct' having been established in the case of one heavily engaged transport supplier (Koninklijke Luchtvaart Maatschappij N.V.). Lack of formal evaluation and 'proof of concept' may be a further issue as a body of evidence on the benefits for the transport sector has yet to be established [71], although anecdotally they may be substantive. Further research to in-fill the formal evidence basis may be needed. A willingness to engage with social media may be a result of perceived advantages in closing the perceptual distance between public and governmental services, resulting in increased public satisfaction and trust [65, 70]. Social media can be used to create a positive image (e.g. for PT use and encouraging PT use through building a community of customers) and to support operational objectives (see Table 1). Finally, authorities may benefit through promoting and connecting related activities – social media can act in an integrative way for those who have a range of activities rather than just transport.

The willingness of transport authorities to engage with social media as a working tool is reflected by the interest of the state Departments of Transport (DOTs, USA) to improve the effectiveness of their social media programmes [72]. Not all agencies publish reports concerning social media related activity and therefore the evidence found is not complete. However descriptions within social media sites, articles, interviews with officials and surveys conducted by various organisations reveal a set of typical activities conducted either frequently or occasionally. Table 2 contains some examples of such activities.

This evidence of uptake supports previous findings [79] on both the volume and pertinence of the information contained in social media textual data. It also highlights potential uses of social media information that have not yet been explored by authorities. The most prominent concerns the potential to aggregate traveller's information. Aggregated information can serve as a basis for identifying major needs and perceived satisfaction that serve as a vital input for decision making in the medium and long term. Fig. 3 depicts the principle data flows to and from official transport authority social media sites, covering those reported as currently implemented and additional potential future flows.

6 Conclusions

The goal of this paper was to address three research questions, in brief, whether social media data may be used either alongside or potentially instead of current transport data, the technical challenges in text mining social media for high-quality transport data and whether institutional barriers to harnessing the potential of social media data in transport sit alongside technical issues.

For the first, it is clear that both established and new social media data have strengths and weaknesses; however, the advantages of social media sources include the ability to capture the whole trip, preserve elements of the associated context and/or the individual socio-characteristics, garner qualitative data on large scale, and finally cost effectiveness.

For the second, challenges arise from the dynamic, location dependent and informal nature of transport textual content. This contributes to the complexity in establishing ontology, but also to sentiment analysis, where contextual data are potentially more of an issue for transport than other domains. Efforts directed at creating transport-related ontologies are already emerging and future efforts are likely to set the basis for increasing the efficiency of the text mining task. The opportunities in combining geo-meta-data associated with the user account/location and place names included in the message are also promising to improve the quality of this task.

For the final research question, a literature review suggested that the need to be active for success with social media, the resource requirement and concerns to safeguard corporate image may be issues. Willingness to engage with social media may be based on the clear advantages in closing the perceptual distance between the public and governmental services. Image building (e.g. for PT) and the ability to support operational objectives may also be incentives to engage. Evidence within the 'grey' literature revealed that an increasing number of authorities appreciate the advantages from rising above the barriers and routinely engage with social media. However, the potential of this engagement has not yet been fully exploited, especially with regards to the use of aggregated social media information for transport planning and management, performance measurement and quality evaluation.

Overall, it is possible to conclude that social media has an increasingly important role in the transport sector, potentially filling some gaps and enriching other data sources. Although challenges in data quality remain, addressing institutional perspectives may yield many 'low hanging fruits' through greater uptake of the social media data already available.

7 References

1 Kushal, D., Lawrence, S., Pennock, D.M.: 'Mining the peanut gallery: opinion extraction and semantic classification of product reviews'. Proc. of the 12th Int. Conf. on World Wide Web, 2003, pp. 519–528

2 Pang, B., Lee, L.: 'Opinion mining and sentiment analysis', *Found. Trends Inf. Retr.*, 2008, **2**, (1–2), pp. 1–135

3 Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: 'Predicting elections with twitter: what 140 characters reveal about political sentiment'. Proc. of the Fourth Int. AAAI Conf. on Weblogs and Social Media, 2010

4 Antweiler, W., Frank, M.Z.: 'Is all that talk just noise? The information content of internet stock message boards', *J. Finance*, 2004, **59**, (3), pp. 1259–1294

5 Koppel, M., Shtrimberg, I.: 'Good news or bad news? Let the market decide'. AAAI Spring Symp. on Exploring Attitude and Affect in Text: Theories and Applications, 2004

6 O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: 'From tweets to polls: linking text sentiment to public opinion time series'. Proc. of the Fourth Int. AAAI Conf. on Weblogs and Social Media (ICWSM), Washington, DC, 2010, pp. 122–129

7 Grishman, R., Huttunen, S., Yangarber, R.: 'Information extraction for enhanced access to disease outbreak reports', *J. Biomed. Inform.*, 2002, **35**, (4), pp. 236–246

8 Corley, C., Cook, D., Mikler, A., Singh, K.: 'Text and structural data mining of influenza mentions in web and social media', *Int. J. Environ. Res. Public Health*, 2010, **7**, (2), pp. 596–615

9 Mai, E., Hranac, R.: 'Twitter interactions as a data source for transportation incidents'. TRB 92nd Annual Meeting Compendium of Papers, 2013

10 Pender, B., Currie, G., Delbosc, A., Shiwakoti, N.: 'Social media use in unplanned passenger rail disruptions – an international study'. TRB 93rd Annual Meeting, 2014

11 Schweitzer, L.: 'How are we doing? Opinion mining customer sentiment in US transit agencies and airlines via twitter'. Presented at the 91th Annual Meeting of the Transportation Research Board, Washington, DC, 2012

12 Collins, C., Hasan, S., Ukkusuri, S.V.: 'A novel transit rider satisfaction metric: rider sentiments measured from online social media data', *J. Public Transp.*, 2013, **16**, (2), pp. 21–45

13 Carrasco, J.A., Hogan, B., Wellman, B., Miller, E.J.: 'Collecting social network data to study social activity-travel behavior: an egocentric approach', *Environ. Plan. B: Plan. Des.*, 2008, **35**, (6), pp. 961–980

14 Efthymiou, D., Antoniou, C.: 'Use of social media for transport data collection, Procedia', Soc. Behav. Sci., 2012, 48, pp. 775–785, ISSN 1877–0428. Available at http://www.dx.doi.org/10.1016/j.sbspro.2012.06.1055

15 Barron, E., Peck, S., Venner, M., Malley, W.G.: 'Suggested Practices Guidance Resource', NCHRP 25–25 TASK 80, September 2013

16 Manning, C., Raghavan, P., Schtze, H.: 'Introduction to information retrieval' (Cambridge University Press, NY, USA, 2008)

17 Nocera, S.: 'An operational approach for quality evaluation in public transport services', *Ing. Ferrov.*, 2010, **65**, (4), pp. 363–383

18 Nocera, S.: 'The key role of quality assessment in public transport policy', *Traffic Eng. Control*, 2011, **52**, (9), pp. 394–398

19 Innovateuk.org.: 'Common Highways Agency Rijkswaterstaat Model (CHARM)', 2013. [online] Available at https://www.innovateuk.org/documents/1524978/1866952/CHARM+business+specification/b5f6281d-8701-4287-84e9-c00d266a15b3, accessed: 11 December 2013

20 Grant-Muller, S.M., Usher, M.: 'Intelligent transport systems: the propensity for environmental and economic benefits'. Technological Forecasting and Social Change, 2013, doi 10.1016/j.techfore.2013.06.010

21 Caceres, N., Romero, L.M., Benitez, F.G., del Castillo, J.M.: 'Traffic flow estimation models using cellular phone data', *IEEE Trans. Intell. Transp. Syst.*, 2012, **13**, (3), pp. 1430–1441

22 Libardo, A., Nocera, S.: 'Transportation elasticity for the analysis of Italian transportation demand on a regional scale', *Traffic Eng. Control*, 2008, **49**, (5), pp. 187–192

23 Nocera, S., Cavallaro, F.: 'Policy effectiveness for containing CO_2 emissions in transportation', *Procedia – Soc. Behav. Sci.*, 2011, **20**, pp. 703–713

24 Nocera, S., Cavallaro, F.: 'Economical evaluation of future carbon impacts on the Italian highways', *Procedia – Soc. Behav. Sci.*, 2012, **54**, pp. 1360–1369

25 Nocera, S., Cavallaro, F.: 'A methodological framework for the economic evaluation of CO_2 emissions from transport', *J. Adv. Transp.*, 2014, **45**, pp. 138–164

26 Nocera, S., Maino, F., Cavallaro, F.: 'A heuristic method for evaluating CO_2 efficiency in transport planning', *Eur. Transp. Res. Rev.*, 2012, **4**, pp. 91–106

27 Nocera, S., Tonin, S.: 'A joint probability density function for reducing the uncertainty of marginal social cost of carbon evaluation in transport planning'. Advances in Intelligent Systems and Computing, 2013, accepted for publication

28 Aggarwal, C.C., Zhai, C.-X.: 'Mining text data' (Springer, 2012)

29 Schulz, A., Ristoski, P., Paulheim, H.: 'I see a car crash: real-time detection of small scale incidents in microblogs'. 'The semantic web: ESWC 2013 satellite events', Berlin Heidelberg, New York, 2013 (*LNCS*, 7955), pp. 22–33

30 Li, C., Weng, J., He, Q., *et al.*: 'TwiNER: named entity recognition in targeted twitter stream'. Proc. of the Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2012

31 Ritter, A., Clark, S., Mausam, Etzioni, O.: 'Named entity recognition in tweets: an experimental study'. Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP), 2011

32 Oppenheim, N.: 'Urban travel demand modeling: from individual choices to general equilibrium' (John Wiley and Sons, Inc., New York, 1995)

33 Nugroho, A.S., Endarnoto, S.K., Pradipta, S., Purnama, J.: 'Traffic condition information extraction amp; visualization from social media twitter for android mobile application'. Proc. of the Int. Conf. on Electrical Engineering and Informatics (ICEEI), 2011

doi: 10.1049/iet-its.2013.0214

34 Kaur, A., Gupta, V.: 'A survey on sentiment analysis and opinion mining techniques', *J. Emerging Technol. Web Intell.*, 2013, **5**, (4) pp. 367–371

35 Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: 'Freebase: a collaboratively created graph database for structuring human knowledge'. Proc. of the ACM SIGMOD Int. Conf. on Management of Data, Vancouver, BC, Canada, 2008, pp. 1247–1250, ISBN 978-1-60558-102-6

36 Madkour, M., Maach, A.: 'Ontology-based context modeling for vehicle-aware services', *J. Theor. Appl. Inf. Technol.*, 2011, **34**, (2), pp. 158–166

37 Cho, S., Kang, J.Y., Yasar, A., *et al.* 'An activity-based carpooling microsimulation using ontology', *Procedia Comput. Sci.,* 2013 **19** pp. 48–55

38 Niaraki, A.S., Kim, K.: 'Ontology based personalized route planning system using a multi-criteria decision making approach', *Expert Syst. Appl.*, 2009, **36**, pp. 2250–2259

39 Trappey, C., Wu, H.Y., Liu, K.L.: 'Knowledge discovery of customer satisfaction and dissatisfaction using ontology-based text analysis of critical incident dialogues'. Proc. of the 2012 IEEE 16th Int. Conf. on Computer Supported Cooperative Work in Design, Wuhan, 2012, pp. 470–475

40 Li, L., Wu, W., Liu, N.: 'Ontology model for situation awareness of city tunnel traffic'. Proc of the Second Int. Symp. on Computer, Communication, Control and Automation (ISCCCA-13), Atlantis Press, Paris, France, 2013, pp. 601–603

41 Becker, M., Smith, S.F.: 'An ontology for multi-modal transportation planning and scheduling', Technical Report, CMU-RI-TR-98-15, Robotics Institute, Carnegie Mellon University, 1997

42 Wang, J., Ding, Z., Jiang, C.: 'An ontology-based public transport query system'. Proc. of the First Int. Conf. on Semantics and Grid', SKG, 2005

43 Houda, M., Khemaja, M., Oliveira, K., Abed, M.: 'A public transportation ontology to support user travel planning'. Proc. of the Fourth Int. Conf. on Research Challenges in Information Science (RCIS), Nice, France, 2010, pp. 127–136

44 Grosenick, S.: 'Real-Time Traffic Prediction Improvement through Semantic Mining of Social Networks'. Thesis (Master's), University of Washington, 2012. URI available at http://www.hdl.handle.net/1773/20911

45 Yang, W.D., Wang, T.: 'The fusion model of intelligent transportation systems based on the urban traffic ontology', *Phys. Procedia*, 2012, **25**, pp. 917–923

46 Pak, A., Paroubek, P.: 'Twitter as a corpus for sentiment analysis and opinion mining', *Computer*, 2010, **10**, pp. 1320–1326

47 Musakwa, W.: 'The use of social media in public transit systems: the case of the Gautrain, Gauteng province, South Africa: analysis and lessons learnt'. Proc. REAL CORP 2014 Tagungsband, Vienna, Austria, 21–23 May 2014. Available at http://www.corp.at

48 Wilson, T., Wiebe, J., Hoffmann, P.: 'Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis', *Comput. Linguist.*, 2009, **35**, (3), pp. 399–433

49 Sood, S., Owsley, S., Hammond, K., Birnbaum, L.: 'Reasoning through search: a novel approach to sentiment classification', WWW2007, North Western University, Electrical Engineering and Computer Science Department Technical Report, NWU-EECS-07–05, Banff, Canada, 21 July 2007, http://www.infolab.northwestern.edu/media/papers/paper10171.pdf, accessed 7th July 2013

50 Wiegand, M., Balahur, A., Roth, B., Klakow, D., Montoyo, A.: 'A survey on the role of negation in sentiment analysis'. Proc. of the Workshop on Negation and Speculation in Natural Language Processing (NeSp-NLP '10), Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 60–68

51 Davidov, D., Sur, O., Rappoport, A.: 'Semi-supervised recognition of sarcastic sentences in Twitter and Amazon'. Proc. of the Fourteenth Conf. on Computational Natural Language Learning, Uppsala, Sweden, 2010, pp. 107–116

52 Gal-Tzur, A., Grant-Muller, S.M., Minkov, E., Nocera, S.: 'The impact of social media usage on transport policy: issues, challenges and recommendations', *Procedia – Soc. Behav. Sci.*, 2014, **111**, pp. 937–946

53 Bollen, J., Pepe, A., Mao, H.: 'Modeling public mood and emotion: twitter sentiment and socio-economic phenomena'. Proc. of the Fifth Int. AAAI Conf. on Weblogs and Social Media (ICWSM), Barcelona, Spain, 17–21 July 2011, pp. 450–453

54 Bollen, J., Mao, H., Zeng, X.J.: 'Twitter mood predicts the stock market', *J. Comput. Sci.*, 2011, **2**, pp. 1–8

55 Chung, J., Mustafaraj, E.: 'Can collective sentiment expressed on twitter predict political elections?'. Proc. of the 25th AAAI Conf. on Artificial Intelligence, San Francisco, CA, USA, 2011, pp. 1770–1771

56 Bie, J., Bijlsma, M., Broll, G., *et al.*: 'Move better with tripzoom', *Int. J. Adv. Life Sci.*, 2012, **4**, pp. 125–135

57 Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., Shook, E.: 'Mapping the global twitter heartbeat: the geography of twitter', *First Monday*, 2013, **18**, (5), doi:10.5210/fm.v18i5.4366

58 Kaplan, A.M., Haenlein, M.: 'Users of the world, unite! The challenges and opportunities of social media', *Bus. Horiz.*, 2010, **53**, (1), pp. 59–68

59 Davis, C.A.Jr, Pappa, G.L., de Oliveira, D.R.R., de L Arcanjo, F.: 'Inferring the location of twitter messages based on user relationships', *Trans. GIS*, 2011, **15**, (6), pp. 735–751

60 Priedhorsky, R., Culotta, A., Del Valle, S.Y.: 'Inferring the origin locations of tweets with quantitative confidence'. Proc. of the 17th ACM Conf. on Computer Supportive Cooperative Work and Social Computing (CSCW), Baltimore, MD, 15–19 February 2014

61 Eisenstein, J., O'Connor, B., Smith, N.A., Xing, E.P.: 'A latent variable model for geographic lexical variation'. Proc. of Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA, 2010, pp. 1277–1287

62 Khanwalkar, S., Seldin, M., Srivastava, A., Kumar, A., Colbath, S.: 'Content-based geo-location detection for placing tweets pertaining to trending news on map'. Fourth Int. Workshop on Mining Ubiquitous and Social Environments (MUSE), Prague, Czech Republic, September 2013

63 Amitay, E., Har'El, N., Sivan, R., Soffer, A.: 'Web-a-where: 'Geotagging web content'. SIGIR'04 Proc. of the 27th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2004, pp. 273–280

64 Paradesi, S.: 'Geotagging tweets using their content'. Proc. of the 24th Int. Florida Artificial Intelligence Research Society Conf., 2011, pp. 335–356

65 Tapscott, D., Williams, A.D., Herman, D.: 'Government 2.0: transforming government and governance for the twenty-first century', New Paradigm, January 2008. Available at http://www.mobility.grchina.com/innovation/gov_transforminggovernment.pdf

66 Cheng, Z., Caverlee, J., Lee, K.: 'You are where you tweet: a content-based approach to geo-locating Twitter users'. Proc. of CIKM'10 Proc. of the 19th ACM Int. Conf. on Information and Knowledge Management, New York, 2010, pp. 759–768

67 Bry, F., Lorenz, B., Ohlbach, H.J., Rosner, M.: 'A geospatial world model for the semantic web'. Principles and Practice of Semantic Web Reasoning, Berlin, Heidelberg 2005 (*LNCS*, 3703), pp. 145–159

68 Gao, L., Zhang, Z., Wu, H.: 'Analyzing the use of Facebook page among state DOTs'. TRB 92nd Annual Meeting Compendium of Papers, 2013

69 Zimmer, C.G.: 'Social Media Use in Local Public Agencies: A Study of California's Cities', Master thesis, Department of Public Policy and Administration, California State University, Sacramento, 2012

70 Cotey, A.: 'Social media: transit agencies connect with riders in new ways', Progressive Railroading, January 2011. Available at http://www.progressiverailroading.com/passenger_rail/article/Social-media-Transit-agencies-connect-with-riders-in-new-ways–25447

71 Barron, E., Peck, S., Venner, M., Malley, W.G.: 'Potential Use of Social Media in the NEPA Process', NCHRP 25–25 TASK 80, September 2013

72 Third Annual State DOT Social Media Survey, AASHTO, September 2012. Available at http://www.communications.transportation.org/Documents/Social_Media_Survey_2012.pdf

73 Use of Social Networking to promote public transport and sustainable travel. Available at http://www.analytics.co.uk/resources/Use+of+Social+Media+to+promote+PT+$26+Sustainable+Travel.pdf, accessed 1 August 2013

74 Transportation Safety Board of Canada. Social media terms of use. Available at http://www.bst-tsb.gc.ca/eng/social, accessed 1 August 2013

75 Minnesota Department of Transportation, Office of Policy Analysis: 'Use of Social Media by Minnesota Cities and Counties', Transportation Research Synthesis, November 2011. Available at http:// www.lrrb.org/media/reports/TRS1104.pdf, accessed August 2013

76 Moss, M.L., Kaufman, S.: 'How Social Media Moves in New York – Final report'. Available at http://www.utrc2.org/sites/default/files/pubs/Final-Report-Social-Media-NYC.pdf, accessed 1 August 2013

77 Shepherd, P.A.: 'The Transportation World Should Embrace Social Media... Carefully', Eno Center of Transportation. Available at http://www.enotrans.org/ctp-blog/the-transportation-world-should-embrace-social-media-carefully, accessed 1 August 2013

78 Virginia Department of Transportation. VDOT on Social Media. Available at http://www.virginiadot.org/newsroom/social_media.asp, accessed 1 August 2013

79 Gal-Tzur, A., Grant-Muller, S.M., Kuflik, T., Minkov, E., Nocera, S., Shoor, I.: 'The potential of social media in delivering transport policy goals', *Transp. Policy*, 2014, **32**, pp. 115–123