Transportation Research Part C 77 (2017) 275–291

Contents lists available at ScienceDirect

Transportation Research Part C

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

Automating a framework to extract and analyse transport related social media content: The potential and the challenges ☆

Tsvi Kuflik [a,*], Einat Minkov [a], Silvio Nocera [b], Susan Grant-Muller [c], Ayelet Gal-Tzur [d], Itay Shoor [a]

[a] The University of Haifa, Mount Carmel, Haifa 31905, Israel
[b] IUAV University of Venice, Santa Croce 191, I-30135 Venice, Italy
[c] University of Leeds, Woodhouse Ln, Leeds LS2 9JT, United Kingdom
[d] Technion - Israel Institute of Technology, Technion City, Haifa 32000, Israel

ARTICLE INFO

Article history:
Received 30 May 2016
Received in revised form 17 January 2017
Accepted 2 February 2017

Keywords:
Mining Twitter for transport information
Social media
Text mining
Opinion mining
Twitter

ABSTRACT

Harnessing the potential of new generation transport data and increasing public participation are high on the agenda for transport stakeholders and the broader community. The initial phase in the program of research reported here proposed a framework for mining transport-related information from social media, demonstrated and evaluated it using transport-related tweets associated with three football matches as case studies. The goal of this paper is to extend and complement the previous published studies. It reports an extended analysis of the research results, highlighting and elaborating the challenges that need to be addressed before a large-scale application of the framework can take place. The focus is specifically on the automatic harvesting of relevant, valuable information from Twitter. The results from automatically mining transport related messages in two scenarios are presented i.e. with a small-scale labelled dataset and with a large-scale dataset of 3.7 m tweets. Tweets authored by individuals that mention a need for transport, express an opinion about transport services or report an event, with respect to different transport modes, were mined. The challenges faced in automatically analysing Twitter messages, written in Twitter's specific language, are illustrated. The results presented show a strong degree of success in the identification of transport related tweets, with similar success in identifying tweets that expressed an opinion about transport services. The identification of tweets that expressed a need for transport services or reported an event was more challenging, a finding mirrored during the human based message annotation process. Overall, the results demonstrate the potential of automatic extraction of valuable information from tweets while pointing to areas where challenges were encountered and additional research is needed. The impact of a successful solution to these challenges (thereby creating efficient harvesting systems) would be to enable travellers to participate more effectively in the improvement of transport services.

☆ This article belongs to the Virtual Special Issue on: Social Network.

* Corresponding author.
E-mail address: tsvikak@is.haifa.ac.il (T. Kuflik).

http://dx.doi.org/10.1016/j.trc.2017.02.003
0968-090X/

1. Introduction

Transport systems support economic growth alongside influencing the wellbeing of people, who need access to employment, services and social interaction (Mihyeon Jeon et al., 2006). Transport stakeholders (such as transport planners, operators and policy makers), need to not only predict changes which will occur in the transport system (for example, in terms of demand), but also to understand the extent to which the system is meeting customers' expectations and needs (Sinha and Labi, 2007). In order to achieve these goals, these stakeholders need to analyse objectively the performance of the transport system, as well as customer's perceived quality of service – identifying the root causes of any dissatisfaction (Sinha and Labi, 2007). Surveys have historically been an established source of information for such analysis, complemented by additional sources of information such as travellers' feedback (in writing, by phone and email and online forms). Objective data on performance has been collected using largely embedded technologies, generating information on delays and congestion for example. The design and implementation of surveys for this purpose is however costly, time consuming and their results may also be surprisingly inaccurate (Flyvbjerg et al., 2005). Moreover, some aspects of short term transport planning (for example responding to road accidents, malfunctioning systems and major community events causing congestion) require constant monitoring, which can be resource intensive. However, many parts of the network are highly instrumented with embedded technologies collecting information for a number of purposes.

The advent of Web 2.0,[1] has resulted in a large volume of User Generated Content (UGC) on a variety of Websites and services, which are collectively called Social Media (SM) (Kaplan and Haenlein, 2010) or Social Web. The high availability of variable UGC allows the application of opinion mining[2] techniques to harvest and analyse opinions and product trends (Tuarob and Tucker, 2015), political events and political orientations (Maynard and Funk, 2011; Tumasjan et al., 2010), entertainment (Pang et al., 2002), online news (Kim and Hovy, 2006) and more. Other work in opinion and SM mining uses SM data to predict economic indicators (Zhang et al., 2010), within recommender systems (Geyer et al., 2010; Tiroshi et al., 2011) and in creating user-opinion search engines (Macdonald et al., 2007; Liu, 2009).

While SM data has been used in many contexts, to date the use of SM in the transport sector is growing, but is still far from reaching its full potential (Gal-Tzur et al., 2014b). SM provides new channels for the expression of users' views and experiences on transport services (Schweitzer, 2012; Collins et al., 2013; Cornwell et al., 2015). Users tend to share participation in particular events (Rattenbury et al., 2007; Java et al., 2007) and future plans, as well as reporting specific events such as heavy traffic (Endarnoto et al., 2011; Cornwell et al., 2015) and car accidents (Gao and Wu, 2013; Mai and Hranac, 2013; Gu et al., 2016). This information is increasingly available in real time, is authentic, is generated at no cost for the transport stakeholders, and, with automatic archiving, is available for off-line analysis. New sources of information can improve the reliability of performance indicators (Cottrill and Derrible, 2015), thereby supporting the achievement of transport policy goals and ultimately reducing transport impacts (Nocera and Cavallaro, 2014; Nocera et al., 2015). Transport service suppliers have identified the potential value of transport-related UGC and the use of SM in connecting with their customers (Gal-Tzur et al., 2014a, 2014b; Bregman, 2012). The potential that SM holds for the transport sector is that the information harvested can complement, enrich, or even replace traditional data collection. It is worth noting that although the data is freely available, harvesting and analysing it does have costs and challenges, some of which are described in this paper.

In this study, we consider MicroBlogs - a specific type of SM that have proven to be a valuable source for real time updates during prominent and critical events, such as natural disasters (e.g. earthquakes) or internal state affairs (e.g. terror attacks, large protests, etc.). This is due to the "instant messaging" nature of the MicroBlog's small posts, facilitating rapid dissemination of news and opinions (Mai and Hranac, 2013). The most well-known MicroBlogging site is probably Twitter,[3] created in 2006. It enables users to send and read text-based posts of up to 140 characters (known as "*tweets*"). Due to its high popularity (generating more than 340 million tweets per day and it has been described as "the SMS of the Internet" (Wikipedia, Twitter, 2016). Although tweets can be restricted to be visible by followers only (users that are subscribed to posts from certain other users and receive constant updates) they are publicly visible by default. This fact has enabled the creation of many third party applications that gather and analyse Twitter posts for various purposes, from adverse drugs reaction (Nikfarjam et al., 2015) to forest monitoring (Daume et al., 2014).

Unlike previous studies that have focused on specific aspects, the overall aim of our research was to propose a generic framework for mining a wide range of transport-related tweets for the purposes of informing transport stakeholders on the status of the transport system and capturing public opinion about it. The first phase of this research has already shown the potential of SM and specifically Twitter, to be a valuable source of information for transport policy makers (Grant-Muller et al., 2014, 2015a, 2015b; Gal-Tzur et al., 2014a, 2014b). However, the process needs to be automated in order to cope with the volume of SM information available and to generate timely, actionable information. Hence, an important outcome from that initial research was a proposed framework to automate the process by applying text mining techniques to extract relevant information from SM. The framework was implemented and demonstrated using messages extracted from Twitter, and highlighted some challenges in automating the process.

[1] The term Web 2.0 is associated to the transformation of the Web into a true collaborative and social platform (Chi, 2008).
[2] Opinion Mining and Sentiment Analysis is a research area that aims at understanding opinions and sentiment expressed in text.
[3] https://twitter.com/?lang=en.

As a case study, transport related tweets that were posted in relation to three football games were collected and analysed. The research hypotheses for the first phase of the research were:

Hypothesis I – SM contains valuable information for transport planning and management, both in terms of content and quantity.
Hypothesis II – This information can be harvested automatically or semi automatically.

These hypotheses were confirmed. In particular, in Gal-Tzur et al. (2014a) we analysed the use of SM by transport stakeholders and evaluated passenger perceptions expressed through SM in an exploratory study. The main conclusions of this first paper were that service providers disseminate information and encourage the public to express opinions on specific topics of interest to them from a "top-down" perspective, and that mining social network data may enable stakeholders to better understand public views and needs on a range of issues and then form future policies and strategies accordingly. Further investigation into integrating social media with transport planning, management and operational activities, while addressing the socio-technical factors that play a role in this operation, was still needed.

In Gal-Tzur et al. (2014b), we included a formulation of the goals for harvesting transport-related information from SM, the hypotheses to be tested to demonstrate that such information can provide valuable input to transport policy and the challenges this involves. We conducted a small scale exploratory empirical evaluation using authentic Twitter data, with the goal of associating tweets with the categories defined. Our results supported the first hypothesis i.e. that valuable information for transport policy makers existed on SM and that such information can be effectively harvested, while the second hypothesis remained unproved.

The goal of Grant-Muller et al. (2014) was to address three research questions related to data requirements in the transport sector. In brief, whether SM data may be used either alongside or potentially instead of current transport data, what the technical challenges are in text mining SM for high quality transport data and whether institutional barriers to harnessing the potential of SM data in transport sit alongside technical issues. For the first we concluded that SM can be a cost-effective data source that includes the potential to capture the whole trip, preserve elements of the associated context and/or the individual socio-characteristics and garner qualitative data on large scale. For the second, the challenges that arise from the dynamic, location dependent and informal nature of transport textual content were elaborated. For the final research question, a literature review suggested that the need to be constantly active when handling SM, the resource requirement and concerns to safeguard corporate image are all potential barriers that should be carefully addressed. Further consideration of these last two points may be found in Grant-Muller et al. (2015b). Evidence within the 'grey' literature revealed that an increasing number of authorities appreciate the advantages from overcoming the barriers and routinely engage with SM. However, the potential of this engagement did not seem fully exploited, especially with regards to the use of aggregated SM information to improve decision processes associated with transport planning and management, performance measurement and quality evaluation.

The goals of this paper are to briefly summarize the outcomes of the previous research for context, to report an extended set of experimental results and finally to describe in some depth the challenges that need to be addressed before a large-scale application of the framework can take place. The focus is specifically on the automatic harvesting of relevant and valuable information from Twitter and the second of the original research hypothesis, which remained unproven. The results of an automatic mining process and transport related messages from two scenarios are presented i.e. with a small-scale labelled dataset and with a large-scale data set of 3.7 m tweets. The challenges faced in automatically analysing Twitter messages, written in natural language and in Twitter's specific language, are also illustrated and discussed in some detail.

2. Mining of social media in the transport domain

There has been a recent surge of initiatives concerning the use of SM UGC for transport-related purposes. Cottrill and Derrible (2015) highlight the potential of social media (among other new technologies) as an information source for transport stakeholders to better understand users' attitudes towards different transport modes and responses to travel disruptions. This type of information could potentially be used to improve the quality of transport sustainability indicators.

Incident detection is one of the most common goals set by researchers focusing on SM as an information source. Mai and Hranac (2013) collected over 5 million tweets and compared incident records with tweets related to roadway events occurring in the same time period. The two datasets (tweets and traffic incident records) were compared to evaluate whether the tweets could potentially complement the incident records. The work showed a correlation between accidents reported in Twitter and those in the California Highway Patrol dataset. The authors conclude that "*Twitter offers a low-cost, readily available data source for agencies interested in uncovering trends in incidents or in gaining information on incidents on their roadways and their effects on the population*". Grosenick (2012) used 352 tweets mined from Twitter, combined with sensor data to detect traffic incidents. It was shown that the predictions generated by sensor and social data combined were more accurate than the predictions generated by sensor data alone. Likewise, Schulz et al. (2013) tried to identify traffic incidents from microblogs in real-time. Using 6 million tweets, selected based on spatial and temporal filtering. 10,000 of these were classified as relevant or not, yielding a balanced dataset of 1986 tweets. For testing, a further 1.5 million tweets were collected and from these, a balanced set of 640 tweets was obtained. The results (i.e. incidents discovered) were compared with infor-

mation from the Linked Open Government Data. Success was reported in detecting all incidents in the tested time span and location. In recent work by Gu et al. (2016), several methodologies were combined with the goal of mining tweets to extract incident information, including incident categories. The scope covered both highways and arterials and the process was intended as an efficient and cost-effective alternative to existing incident data sources. Gu et al. collected over 22,000 tweets and manually labelled them, finding over 8000 tweets that were related to traffic incidents. The data was then split into a training set of 17,200 tweets and a test set of 5000 tweets in order to explore the potential of automatic identification of traffic incidents. To summarise their approach, first, an iterative process of querying Twitter, using adaptive data acquisition to improve the word dictionary is applied. This forms the basis for building a classifier to identify incident-related tweets posted both by authorities and individuals. A geo-parser is used to focus on tweets relating to a specific area. Five categories of incident are then identified using a classifier. While the geocoding process produced good results, the incident-category classifier produced results of lower quality. In summary, all the studies described above focused on using Twitter data for the identification of traffic incidents, a specific and highly important task that, as demonstrated, proved to be quite successful. Our study takes a broader view. We aim not only to collect traffic incidents related tweets but any traffic related information reported by individuals and analyse it, so it can be used later on by stakeholders. Hence our view is broader and with a different and long term goal of providing information to stakeholders rather than identification of incidents in real time.

Several researchers focused on dedicated accounts managed by transport authorities as information source. Endarnoto et al. (2011) created an application using NLP techniques to process information from a Twitter account used to update traffic conditions in Jakarta. They collected 100 tweets that were analysed and the traffic conditions they described were mapped. D'Andrea et al. (2015) used tweets to identify traffic congestion. They compared the quality of classifiers developed using several text-mining algorithms both for a 2-class classification (traffic-related or not, using 1330 tweets) and for a 3-class classification (traffic due to an external event, traffic congestion or crash, and non-traffic, using 999 tweets). In both cases SVM was found to be the best classifier according to performance indicators, and the quality of results obtained under all criteria was high (only one result was lower than 86%). Pathak et al. (2015) demonstrated how UGC posted on dedicated accounts managed by transport authorities can serve as a basis for calculating traffic flow performance indicators and forming a control dashboard. They used 500 tweets and 600 Facebook posts. The analysis of content extracted from Twitter and Facebook differed slightly, both in terms of classification categories and in terms of the algorithms found to be effective. Nevertheless, data from both sources was jointly analysed to create information regarding traffic flow conditions. In general, the researchers have managed to achieve higher quality classifiers when mining tweets than when mining Facebook posts, the latter being longer and thus more complicated to analyse. Zhang et al. (2016) reported a one-year study, in which over 500,000 tweets with geo location were extracted and used to explore its potential in improving transport management and control. They investigated the correlation between tweets, traffic surge and accidents and concluded that "*The results prove the potentials of using tweets to detect the traffic surge within a given scale of space and time*". However, they also concluded that there is still quite some research work left to be done and that fusing SM data with other sources of information is a promising path for the future.

These four studies also focused on using Twitter data for the identification of traffic conditions while, as already noted, our study takes a broader view.

Chaniotakis and Antoniou (2015) attempted to link quantitative attributes of both geotagged and non-geotagged tweets to transport-related activities. Based on data retrieved during one month around Athens (more than 2.5 million tweets), the researchers found some correlation between the volume of tweets posted and spatial and temporal characteristics, such as destinations attracting travellers and time periods associated with non-work activities. Collins et al. (2013) used 557 tweets that had been manually collected to conduct sentiment analysis regarding the Chicago Transit Agency's rail service. The research challenges included the selection of relevant tweets from the stream of (mostly non transport-related) tweets available through the "fire hose" API supplied by Twitter.

These works have similarities to ours in their attempt to exploit Twitter data for purposes beyond identification of traffic situations, however, in a way our work complements and extends these studies that focused on specific aspects, in taking a broader view of a wider range of transport-related information.

Kocatepe et al. (2015) investigated the effectiveness of Twitter as a tool for disseminating traffic information, and in particular, focused on the question "Can the Florida Department of Transportation (FDOT) twitter accounts disseminate information as efficient as they are meant to". The research focused on the number of followers of the 7 Twitter accounts of the FDOT districts, their level of activity (retweets) and spatial reach (based on geotagged tweets) as the main criteria in answering this question. The five most active users were found to be either large official entities or agencies (i.e. other transport-related accounts or accounts of journalists or journals) that act as a disseminating channel in themselves. However, the majority of the followers of the FDOT account were found to be rarely active. The spatial reach was difficult to assess, as the ratio of geotagged tweets in the stream is, in general, very low. Given the data available, it seemed that reaching the population in urban areas is easier than reaching the rural population. The research concluded that although Twitter is a valuable tool for DOTs, there is a need for further steps to increase its effectiveness as a dissemination channel.

In summary, there is a growing body of evidence that Twitter is considered a valuable, inexpensive and reliable source of transport-related information. Each of the studies outlined above demonstrate the potential of Twitter as a source of transport-related information. These studies however focused on addressing specific information needs using social media, traffic incidents in most cases, users opinions, traffic congestions/flow in others, some of them addressed specific aspects that we have also addressed, but we did that in a broader view. In contrast, we take an exploratory approach, seeking to iden-

tify all transport-related issues that are discussed on social media, in order to bring this information to the attention of policy makers. The research reported in this paper therefore complements and extends previous work by adopting a broad view of transport-related issues, and by extending current knowledge and understanding regarding the appropriate technical methodology and the challenges associated with it.

3. Tools and methods

This study was fundamentally multidisciplinary, at the interfaces of text mining and social informatics, with the goal of an automated approach for information elicitation from SM. Twitter was selected for experimentation for the reasons described above and every tweet was considered as a separate document. The first task was to define a machine readable representation of domain knowledge. This was followed by the application of classical text mining techniques using this knowledge for information extraction from social media texts. A major research effort was the preparation of the knowledge base and dataset as described in Gal-Tzur et al. (2014b), Grant-Muller et al. (2014, 2015a, 2015b). Here we briefly recall these results to inform the discussion that follows.

3.1. Mining framework for transport–related tweets

In the first phase of this study, a framework for automatic mining of transport-related information from Twitter was proposed, consisting of the following steps:

- Extraction of transport-related information from SM.
- Association of the extracted information to a set of predefined domain categories.
- Aggregation, summary and visualization of the data retrieved.

Fig. 1 outlines a flow of text mining process, applied to SM for mining transport-related information. Let us consider the process using also an illustrative example: "***@lpoolcouncil Bus stops and pavements cleaned and being rid of dog s∗∗∗. Why not look after the Anfield area like this all year round?*** "

3.1.1. Initial message filtering

The data stream on SM relates to a large variety of topics, where for computational reasons, possibly relevant messages must first be extracted from the general messages stream. A reasonable strategy would be to define a set of keywords that are typical of the transport sector. However, such lexicon may include many general words, which are highly ambiguous. For example, consider the simple text "*a match made in heaven*", where the term "*match*", which is associated with football and may therefore be of interest to the study, is used with irrelevant sense and context. Filtering messages by keywords may therefore yield highly noisy results. Having identified candidate messages using initial, but inaccurate criteria, a better assessment of relevance and a more detailed interpretation of contents is needed using additional text mining steps. Our example above may be defined as relevant and pass the initial filtering step given the fact that it is linked to Anfield area, which is an area of interest.

3.1.2. Message relevancy

Once possibly relevant messages are extracted, their association with transport is further assessed using supervised machine learning approaches. Such approaches rely on "*labelled examples*", implying that a *dataset* (a collection of messages) needs to be constructed containing example texts with their correct labels (categories/classes). To learn a classification model that fits labelled examples and generalizes to new examples, example texts are abstracted into pre-defined *feature* values. In the popular "*bag-of-words*" approach (Manning et al., 2008) a document is represented as an unordered set of word *unigrams*, i.e., individual words. 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 such as "car accident" "(a more elaborate discussion of this representation scheme is included in Section 4.1). Furthermore, the identification of the location can be a relevant pertinence index for a given message. Once classified, messages that are identified as irrelevant to transport can be discarded. Our illustrative example should be identified as a relevant message since it contains the terms Bus and Bus Stop.

3.1.3. Semantic processing

Messages judged as relevant can then be classified into finer categories within the transport domain, as the definition of "relevant" is too broad. Specifically, we consider two classification tasks. One is classifying a message according one of three purposes: reporting transport events, expressing a wish to travel to some known destination or expressing an opinion about a transport-related issue (Gal-Tzur et al., 2014b). The second is classifying messages according to the transport mode referred to. The second half of the example message expresses a personal opinion in the form of a question, hence in the semantic processing we expect it to be defined as an individually authored and possibly opinionated message.

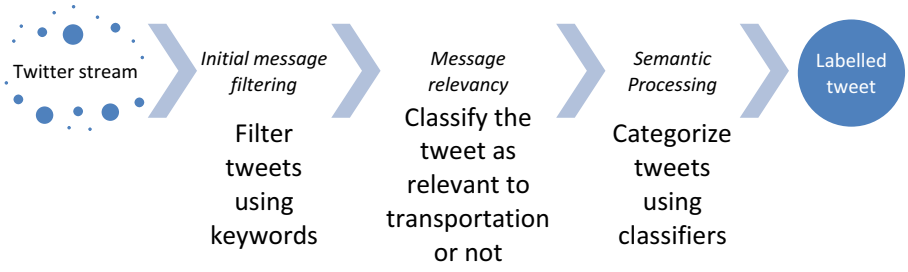

Fig. 1. Tweet processing general flow.

In summary, the proposed framework includes a pipeline of classifiers (illustrated by Fig. 1), determining firstly the relevance of the message to transport (so our example message is labelled "relevant"), then assessing whether it was written by an individual and whether it represents a personal opinion (in this step our example message is labelled "individual" and then "opinionated") and finally identifying the specific transport category of the message (the mode of transport is "bus"). The infinite nature of the message stream is, however, a challenge from several perspectives. As far as the filtering performances are concerned, a periodic check of the state of the system may be required, given that SM contents evolves rapidly over time. It is also evident that all relevant messages of a given data drift cannot be identified because of their large volume. This feature, which belongs intrinsically to the nature of social media, is not an unsurmountable hurdle however. As SM information generally has a high degree of redundancy, many posts have in fact a different form but similar content. This potentially allows the retrieval of some relevant content that may be initially disregarded from a sub-optimal harvesting process (D'Andrea et al., 2015; Kaya and Conley, 2016).

3.2. Dataset creation and challenges encountered

In order to empirically evaluate the potential of harvesting transport-related information from SM, data were extracted from Twitter and the ability of trained classifiers to identify relevant transport-related information from it was verified. To focus the problem of harvesting transport-related information from SM, a sub-problem was defined i.e. extracting transport-related information in the context of specific events. Mass sporting events were the focus: such events being well defined by date and location, and providing a convenient environment for seeking meaningful information regarding transport. The Liverpool Football Club team was selected in particular due to the frequent transport queries made by football fans trying to reach Liverpool football matches. Three games that took place during 2012 were selected (see Table 1 for details). Keywords were manually defined that included 'nicknames' for these teams, the name of the stadium and so forth. Tweets that were posted from the period up to three days prior to each game until up to three days following the game, and which contained one or more relevant keywords, were collected using the Twitter4j API. This accesses Twitter's native streaming API that offers samples of the public data flowing through Twitter, and enables filtering of tweets out of the general stream that match one or more filter predicates) (Yamamoto, 2007).

While event-specific keywords are targeted at extracting postings from the general message stream that related to the events, only a fraction of these messages are expected to be related to transport, hence this dataset appeared appropriate for the purposes of the research. The assumption was made that transport-related messages contain transport-related terms. However, a transport-specific dictionary or ontology that could be used for this purpose was unavailable at the time of the research and to the authors' best knowledge is still unavailable (Grant-Muller et al., 2014). To this end, a dictionary of transport-related terms was constructed semi-automatically and used to rate tweets by their presumed association with the transport domain. In order to construct the dictionary, a pool of 35 transport-related documents was considered, including stakeholders' Web sites (e.g. of taxi services, transport magazines, etc.); academic research papers and white papers, transport Web forums, blogs and SM accounts (for example, the Twitter account of the Department for Transport (UK), etc.). These documents were analysed and the most informative terms were extracted as representing transport-related documents. The terms 'transport' and 'traffic' are among the top-ranking terms as may be expected. However, the list of terms was highly noisy, including outliers such as 'tfltrafficnews' (the name of a Twitter profile, appearing in one document only). As a result, numbers, non-English terms, non-words (for example, hyperlinks) etc. were automatically eliminated from the list of terms. This resulted in a list of 840 terms[4] that included a mix of transport-related terms (see examples in Table 2) and general words (e.g., 'reply'; that appears frequently on discussion boards, which were included in the documents).

In order to obtain a high-quality dictionary, the list of most frequent terms was manually assessed by domain experts, with the goal of approving the relevance of the term to transport, having each term assigned a score between 1 and 5. Using the resultant dictionary of weighted terms, the messages were filtered and only those that appeared to be transport-related were selected. In order to assess message relevancy to transport, we applied a heuristic scoring, by which each message was

[4] The complete list is available at: https://www.dropbox.com/s/oxtiy0idn0owvwn/Transport%20related%20dictionary.pdf?dl=0

Table 1
Tweets collected per match.

Match	Date	Number of tweets collected
Liverpool vs. Westbrom	22 April 2012	847,665
Liverpool vs. Chelsea	8 May 2012	1,905,124
Swansea vs. Liverpool	13 May 2012	964,369
Total tweets collected		3,717,158

Table 2
Example of terms extracted from the transport corpus.[a]

Term	Term frequency	Document frequency
Transport	4474	24
Traffic	1930	20
Reply	1780	4
tfltrafficnews	1706	1
Road	1494	26
B	1169	17
Function	1069	25
1	943	27
May	939	27
bit.ly	407	5

[a] The complete list can be found at: https://www.dropbox.com/s/oxtiy0idn0owvwn/Transport%20related%20dictionary.pdf?dl=0.

assigned the aggregate scores of the transport-related terms it contained. For example, the sentence "the congestion is due to a car accident", includes three terms that appear in our dictionary: 'congestion', 'car' and 'accident'. If these terms are assigned the scores of 5, 5, and 2, respectively, the total sentence score is 12. Following manual testing, we incorporated a threshold (3 in this case) where terms assigned a score lower than the threshold were ignored. The heuristic for message scoring described in Algorithm 1 - transport related tweet grading method, served mainly to select transport-related messages to be labelled by experts, as described below. The algorithm takes a tweet (T) and a threshold (L, set to 3 after some experimentation) as an input. It evaluates the tweet according to the cumulative score of the transport terms it contains.

Algorithm 1. Transport related tweet grading method

```
gradeTweet(T,L)
grade ← 0
for word W from T
   // return the grade of word W in the graded dictionary
   // only if the grade is equal or larger than L, else return 0
   grade ← grade + getWordGrade(W,L)
return (grade)
```

3.3. Message annotation and construction of domain taxonomy

An initial step of the analysis process was the creation of transport taxonomy (Gal-Tzur et al., 2014b), which was needed in order to enable the classification of tweets. It was created based on typical characteristics of SM content, i.e. content referring to an experience or action that is of some importance to the individual posting the message, and has occurred shortly before or shortly after the time at which the content was created. When analysing transport-related content, three main categories, reflecting three main purposes of a message, were identified. These purposes constitute the first hierarchy of the transport taxonomy, as illustrated in Fig. 2. It is important to note that a message can be related to more than one purpose.

Each of the three categories was then expanded to reflect a more detailed description of the content of the post. The taxonomy was developed in a manner that was as comprehensive as possible, independently of the specific tweets analysed in the current case study. As an example, Fig. 3 illustrates the expansion of the "Opinion regarding a transport service" category. The middle level describes the possible transport modes, and is almost identical for all three purposes. The bottom layer specifies details that relate to the purpose e.g., the specific characteristics of the transport service that the message addresses. Naturally, the volume of examples decreases with the increase in the depth of the hierarchical classification.

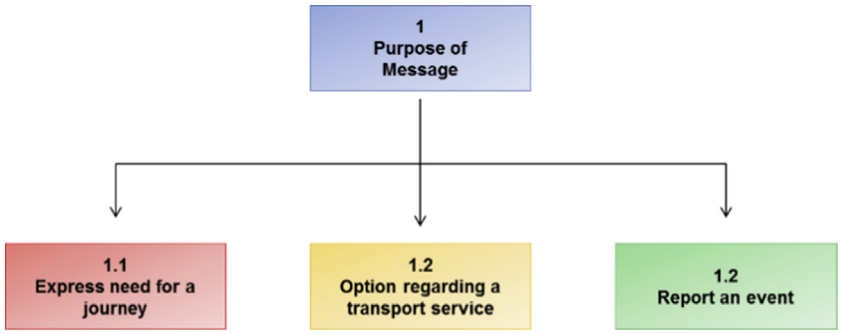

Fig. 2. Top level of annotation hierarchy - purpose of message. Source: Gal-Tzur et al., 2014b.

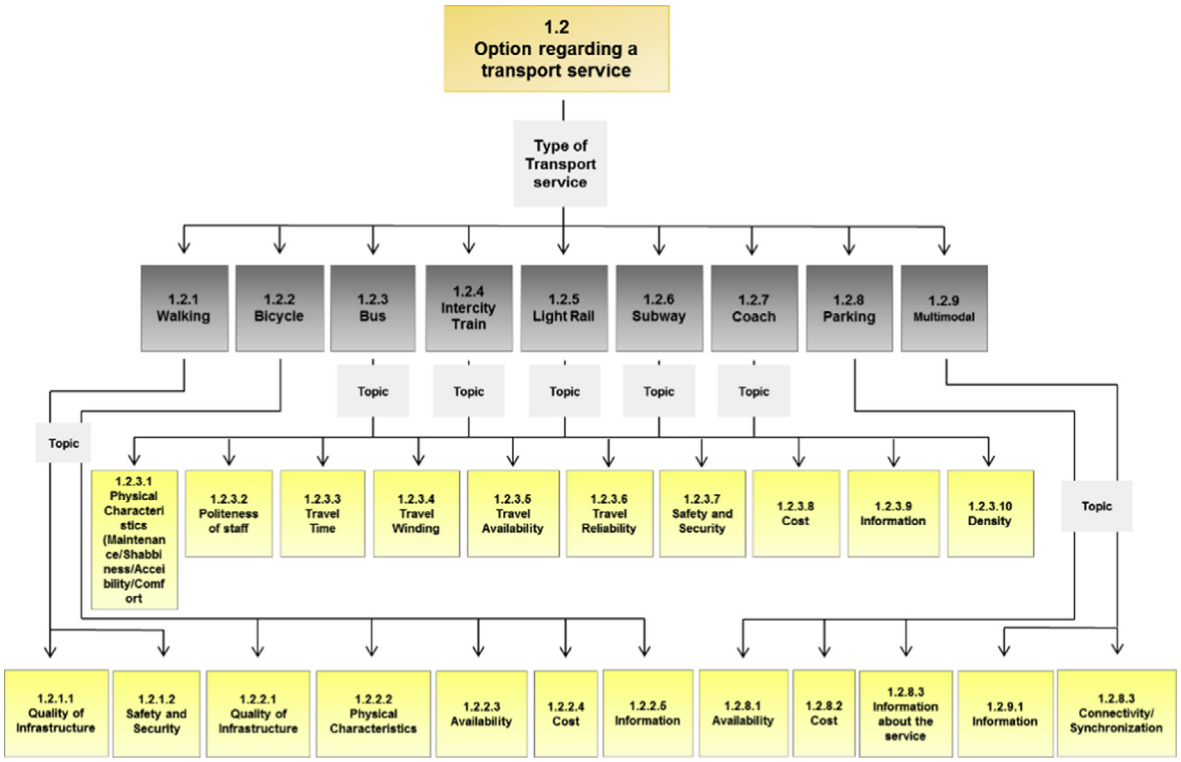

Fig. 3. Expansion levels of annotation hierarchy purpose 1.2 (Opinion about service). Source: Gal-Tzur et al., 2014b.

Hence, only the first two levels of the taxonomy were eventually used in this research for manual labelling and for the creation of classifiers. It might be interesting to note that in Gal-Tzur et al. (2014b) where a highways-based taxonomy was developed, we found the volume to decrease non-uniformly with some high-volume categories and some sparse categories, although the sparse categories were still very relevant to the purpose of the taxonomy

Once the relevant transport taxonomy was defined, labelled datasets had to be generated in order to enable the training and testing of classifiers that can automatically identify transport-related information, following the process described above. Manual classification is a resource intensive approach that is not practical for large scale implementation. The solution for reducing the workload was finding potential candidate tweets automatically. Hence, initially, the grading heuristic was used to select example messages that were likely to be related to transport and these were then assessed by two domain experts. Messages that the experts deemed genuinely transport-related were then further labelled with the specific purpose of the message (i.e. expressing a need/opinion or reporting an event). In order to evaluate inter-annotator agreement rates, 100 messages per task[5] were co-annotated by the two experts. Inter-annotator agreement was calculated

[5] Please note that for the first two tasks there were only 92 and 93 messages.

Table 3
Kappa rates for co-annotated tweet categories.

Category	Annotator decisions				Pr(a)	Pr(e)	Kappa	Agreement
	Y/Y	Y/N	N/N	N/Y				
Authored by individual	89	1	1	1	0.98	0.96	0.49	Moderate
Transport-related	88	0	2	3	0.97	0.93	0.56	Moderate
Purpose of message:								
Opinionated	37	6	47	10	0.84	0.50	0.68	Substantial
Report an event	12	5	82	1	0.94	0.74	0.77	Substantial
Expressing a need	7	20	71	2	0.78	0.69	0.29	Fair
Mode of transport:								
Bike	1	0	99	0	1	0.98	1	Almost Perfect
Bus	11	3	84	2	0.95	0.77	0.79	Substantial
Car	9	1	89	1	0.98	0.82	0.89	Almost Perfect
Coach	1	1	97	1	0.98	0.96	0.49	Moderate
Parking	1	0	99	0	1	0.98	1	Almost Perfect
Taxi	2	0	98	0	1	0.96	1	Almost Perfect
Train	32	9	53	6	0.85	0.52	0.69	Substantial
Tube	9	0	90	1	0.99	0.83	0.94	Almost Perfect
Walking	1	2	97	0	0.98	0.96	0.49	Moderate

using Cohen's kappa (Carletta, 1996). Table 3 presents the agreement levels for the "authored by an individual", "transport-relatedness" and "purpose of message" categories, together with the mode of transport. Annotator decisions refers to the classification of a tweet to the specific category, where Y/Y means an agreement on a message being classified as relevant to the category (Y/N, N/Y – disagreement and N/N agreement on the message being irrelevant), Pr(a) is the probability of agreement and Pr(e) is the probability of agreement by chance. Note that an individual message may be positively associated with multiple categories; e.g., one message may express both a need for travel and an opinion. For example, the message "Gotta catch a train to Birmingham and the brakes are stuck on the train to Liverpool, not happy" both expresses a transport need and an opinion.

The labelling process yielded approximately 4000 labelled tweets. Table 4 presents the distribution of the labelled tweets, including the sub-categories for mode of transport. As illustrated in Figs. 2 and 3, the domain taxonomy is quite detailed. A large variety of specific sub categories represent different types of information that may be useful for decision makers. From Table 4, it is quite clear that the 4000 tweets that were manually labelled provide a relatively small number of examples when considering the lowest level of the hierarchy (the number of "positive" labels refers to the number of tweets found to be related to the specific category out of the set of candidate tweets, the remainder were used as negative examples for training a classifier). As a result, the analysis was limited to the top-level category of the "purpose-of-message" category and to the mode of transport, given every tweet referred to a specific mode.

3.4. Tools and procedure

Using WEKA (Hall et al., 2009), several classification algorithms were explored, including: Support Vector Machines (SVM) with a linear kernel,[6] Decision Trees (specifically a C4.5 implementation called J48 - Quinlan, 1993) and Naïve Bayes (NB) (Witten and Frank, 2005; Liu, 2009). All classifiers were used with WEKA default parameters. The evaluation metrics used in this work were *precision*, *recall* and F_1 (Manning et al., 2008), where precision is the ratio of the correctly extracted items to the total number of extracted items, recall is the ratio of the relevant information found by automatic text mining to the total number of relevant information items, and F_1 is the harmonic mean of precision and recall. *Precision@K* (precision at K the percentage of correctly classified items out of the top K items in a list (Manning et al., 2008)), that ia considered the most intuitive measure of Web search results (Agichtein et al., 2006) was also calculated.

Experiments were conducted based on two scenarios. The first involved a small scale, labelled dataset of the 4000 labelled tweets, while the second included the full dataset of over 3 M tweets. Due to the relatively small size of the labelled datasets, 10-fold cross validation experiments (Manning et al., 2008) were conducted. For every task, multiple combinations of classifier and feature representation were evaluated. We experimented with different classifiers and with different representations of tweets – hence in every task different combinations of classifier and representations were used, as illustrated by Table 5. The detailed results of the cross validation experiments for each of the classification tasks are reported below in terms of precision, recall and F_1. For the large-scale experiment precision@K is reported, as it was impractical to label all 3.7 M tweets and hence only the top 100 were evaluated. As a general note, evaluating the precision@K is a common practice in information retrieval where the top 1, 5, 10, 50 and 100 are commonly used. In the evaluation, we vary K, reporting precision at K = 10, 50 and 100.

[6] We used default parameters (C = 1). Results using SVM with linear kernel were comparable to SVM with RBF kernel in our experiments.

Table 4
Distribution of labelled tweets concerning the mode of transport. Positive examples are tweets that were identified by experts as belonging to the category from the dataset that was examined.

Category (label)	Dataset size	Number of positive labels	Ratio of positive labels
Authored by individual	479	220	45.9%
Transport-related	2687	1101	40.9%
Purpose-of -message	821		
Transport need		231	28.1%
Opinionated		483	58.8%
Report an event		182	22.1%
Mode of transport	821		
Train		396	48.2%
Multi-Modal		129	15.7%
Bus		123	14.9%
Car		79	9.6%
Tube		28	3.4%
Unknown		22	2.6%
Bike		12	1.4%
Walk		11	1.3%
Taxi		9	1.09%
Coach		7	0.85%
Tram		2	0.2%
Park		1	0.1%
Boat		1	0.1%
Rideshare		1	0.1%

Table 5
Summary of evaluation results for the mode of transport classifier.

Mode	Vector	Classifier	Precision	Recall	F_1
Multi-modal	Unigrams	Naïve Bayes	0.344	0.426	0.381
Bike	Unigrams	Naïve Bayes	1.000	0.250	0.400
Boat*	Unigrams	Naïve Bayes	0	0	0
Bus	Unigrams	Dec. Tree J48	0.739	0.805	0.770
Car	Multigrams	Naïve Bayes	0.551	0.684	0.610
Coach	Unigrams	Dec. Tree J48	0.500	0.429	0.462
Park*	Unigrams	Naïve Bayes	0	0	0
Rideshare*	Unigrams	Naïve Bayes	0	0	0
Taxi	Unigrams	Dec. Tree J48	0.778	0.778	0.778
Train	Unigrams	Dec. Tree J48	0.786	0.912	0.844
Tram*	Unigrams	Naïve Bayes	0	0	0
Tube	Multigrams	Dec. Tree J48	0.524	0.393	0.449
Unknown	Multigrams	Dec. Tree J48	0.556	0.250	0.345
Walk	Unigrams	Dec. Tree J48	0.455	0.455	0.455

Since SVM performance was the lowest in all of the categories, it is absent from the table. Due to an insufficient number of positively labelled training examples (under 0.5% of the annotations, see Table 4), several categories did not have enough examples to train a classifier. These categories are marked with an asterisk (*). Transport modes for which there was high number of positive examples in the dataset, however, showed adequate performance. Specifically, precision and recall rates of 0.79 and 0.91 were obtained for 'Train', and similar rates of 0.74 and 0.81 were obtained for 'Bus'. This implies that better performance can be achieved by increasing the number of positive examples per mode of transport.

4. Experimental results

For experimental purposes 3.7 m tweets were extracted from Twitter. Out of these, approximately 4000 were manually labelled and used for training and testing specific classifiers. The remainder of the dataset was then analysed as a second step of testing. Here we report the experimental results and challenges encountered during the process.

4.1. Classification of transport-related tweets using a labelled dataset

4.1.1. Message representation schemes

Two types of tweet representations were used: *unigram* and *multigram*. Using *unigram* representation, each message is represented by the individual words that it contains. The *multigram* representation includes word sequences that compose the message. Specifically, we considered unigrams, bigrams and trigrams. For example, the phrase "this is a text" includes 4 unigrams ("this", "is", "a" and "text"), 3 bigrams ("this is", "is a" and "a text") and 2 trigrams ("this is a" and "is a text"). The multigram scheme allows learning of meaningful word sequences. To illustrate this, consider the meaning of the term 'walk' in the following texts:

i. "I'm going to walk home after the concert"
ii. "I'm going to walk the dog after the concert"

The word 'walk' is often associated with transport, but may be used in irrelevant context of 'accompany', as in the second sentence. Modelling the collocation 'walk the' or 'walk home' can effectively distinguish between the word senses. Using both types of representations, tweets were represented as an unordered set of multigrams or unigrams (this is also known as the 'bag-of-words' approach). We note that words may also be represented as vectors in a semantic space. While such semantic representations have been proven useful for sentence classification tasks (Kim, 2014), SM includes many irregular word forms, for which the quality of the representations learned from general text may be low. We leave the exploration of such word representation models for the tasks at hand to future research.

4.2. Individually authored messages

The first task was to identify tweets created by individuals (as distinct from tweets created by organizations/authorities). The SVM classifier yielded the best performance (in terms of F_1) using both unigrams and multigram representations. Overall, SVM with multigram features resulted in the best performance: $F_1 = 0.91$, *recall* = 0.93 and *precision* = 0.88. This high level of performance aligns with the high inter-annotator agreement rate observed during the construction of the dataset, indicating that this classification task is clear. Manual examination of the results indicated possible reasons for classification errors. Consider the following tweet: "*RT @VelataFun: Transport your chocolate easily or conveniently store leftover fondue with warmer lids to match your warmer #velata #genius*". While this tweet is a commercial one, it is phrased as a personal message. Careful inspection is needed in order to understand the context and classify it properly.

4.3. Transport-related messages

The best classifier for the identification of transport-related individual messages (the second step in the framework depicted by Fig. 1) was Naïve Bayes using the multigram features. This yielded F1 of 0.965, precision rate of 0.973 and recall rate of 0.956. Naïve Bayes using unigram features was the second best classifier, giving comparable performance. It is worth noting that the SVM performed comparably well while the decision tree was less successful. These good results align with the relatively high inter-annotator agreement rates on this task. An interesting example for classification errors in the "trans port-relatedness" task is the following tweet: "Mancini may just turn into Frank Drebin post match Oh sure maybe not as much as landing on a bicycle with the seat missing but it hurts!". In this message, a transport-related term ("bicycle") is used as part of an idiom in a non-transport-related context. Various similar phrases exist. As another example, consider the popular phrase "drive me crazy". It is interesting to note, that in the case of the latter, using a trigram feature (which is included in the multigram scheme) may resolve the issue given enough training examples.

4.4. Purpose of messages

Three sets of cross validation experiments were performed with respect to the purpose of the message, given it included three categories: expression of an opinion ("opinionated"), transport need, or reporting an event. Classifiers were trained in a pairwise fashion, with the goal of distinguishing between messages that were associated with the target class (positively labelled with the target category) and the other messages. Overall, results were moderate. The best performing classifier in identifying "opinionated" messages was SVM using unigram features, yielding precision, recall and F_1 scores of 0.68, 0.80 and 0.74, respectively. Sentiment or subjectivity analyses are considered to be challenging tasks (Pang and Lee, 2008). The results obtained correspond with the results reported by studies such as that of Pang et al. (2002). Manual inspection of erroneously classified messages reveals the challenges involved in subjectivity analysis. Consider the following subjective tweet, which was classified as not subjective: "*so the trains have ALL been cancelled from southend victoria to Liverpool street. . .looks like I might be staying with bezzie longer! #score*". The word "like" here is not used in its subjective sense. However, the author of the message makes an atypical subjective use of the word *score* to provide a positive opinion about the scenario. In order to classify this message correctly, it is necessary to decode the meaning of the word 'score' in the given context, which requires deep semantic understanding and world knowledge.

Mediocre results are observed in the category of "reporting an event". While the respective inter-annotator agreement rates were high, this appears as a challenging classification task. The best performing classifier, SVM using unigram features, yielded precision, recall and F_1 scores of 0.63, 0.61 and 0.62, respectively. Notable features in terms of mutual information include "*traffic*", "*fire*", "*bridge*", "*stuck*", and "*failure*". Although all of these words are related to transport events, the classifiers' performance is somewhat low. One possible explanation for the low performance (despite the high inter-annotator agreement rate) is the small size of the training data in addition to an unbalanced dataset. Only 182 examples out of a total of 821 labelled tweets (22.1%) were classed as reporting an event. Consider the following misclassified example: "*UK tweeties. Is Liverpool too crowded to drive/park near the city center? Or should we take to train in?*". This tweet was erroneously classified as reporting an event. Although it does contain a word related to transport events ("crowded"), this tweet merely inquires (rather than reports) whether a transport event occurred.

The best classifier in the third category, expressing a "transport need" was Naïve Bayes using multigram word features. This yielded precision, recall and F_1 of 0.54, 0.58 and 0.56, respectively. Again, this relatively poor performance can be partly attributed to the small number of positive training examples, as well as to the unbalanced training set that included 231 relevant tweets out of 821 (28.1%). Inter-annotator agreement rates for this task were the lowest, implying that this task may be challenging for humans also.

4.5. Transport mode

The third step in the filtering process was to identify the specific mode of transport discussed in the transport-related messages. Table 5 summarizes the experiments, showing the results that were obtained using the best performing classifier for each mode (or class).

4.6. Experimenting with a large scale, unseen dataset

In another set of experiments, the entire corpus (3.7 M tweets) was classified according to the pipeline described above (Section 3.1, Fig. 1), using the best performing classifiers having been trained using the labelled datasets. That is, the classifiers were trained using the labelled data, and then applied to the full corpus. This approach was used as the number of labelled examples was relatively small and it allowed an evaluation of the performance of the classifiers on a larger data set. The relevancy of the top-k ranked tweets produced for each classification task and model evaluated were manually assessed. A summary of the performance of the classifiers precision@K for the top 10, 50 and 100 tweets with highest prediction confidence[7] is shown below in Table 6. From Table 6, in general, it appears that the results for this experiment are very good at identifying individual tweets, individual tweets that are related to transport and express a specific need, and also the identification of the transport mode was good. However, unlike our results in the small-scale experiment, the identification of tweets that are related to transport was not good at the top 10, but then improved dramatically for the top 50 and top 100. The identification of tweets that report on an event was substantially worse than the performance of other classifiers and also weak compared with the results of the first experiment.

Given these results (based on ranking the tweets according to the confidence level of the classifier), the results were re-ordered according to the original heuristic scores computed per message. This was also based on the domain related dictionary previously constructed, rather than relying on the classifier alone. The results are presented in Table 7, which compares the two ranking approaches. As can be seen from Table 7, the use of the heuristics-based ranking scheme significantly improved performance with respect to identifying transport-related tweets and the identification of the mode of transport. This approach also performs comparably well at the identification of an event and transport need.

5. Discussion

Two hypotheses have been investigated and proved true within a first phase of this research (see Gal-Tzur et al., 2014b; Grant-Muller et al., 2014).

1. **Hypothesis I** – SM contains valuable information for transport planning and management, both in terms of content and quantity.
2. **Hypothesis II** – This information can be harvested automatically or semi automatically.

However, researching these hypotheses triggered two main challenges that need to be discussed further. The first was the issue of inter-annotator agreement and the second was the misclassification of tweets.

5.1. The need for domain ontology

The accurate identification of transport-related tweets is very important, as this is the entry point to the classification process. For this task, comprehensive domain ontology is needed that can be used to extract relevant tweets from Twitter. For the purpose of our exploratory study, we extracted transport-related terms from a corpus of transport-related documents, ordered them by their frequency (combining term frequency in the document and inverse document frequency in the corpus, the classical TF∗IDF method (Manning et al., 2008)), and then let domain expert review the list. This was a good approach for extracting tweets for our exploratory study, but was insufficient for a systematic mining process. While in the context of our study this was not a limitation, for broader application, further research is needed in order to define a comprehensive set of transport-related terms to be used for filtering relevant tweets from Twitter.

[7] Weka computes class membership probabilities for a given test instance which is the confidence value. These probabilities are computed differently for each classifier. For the Naïve Bayes classifier, this is straightforward, as Naïve Bayes outputs posterior class probabilities. For SVM, these probabilities are computed based on the distance to the separating hyperplanes.

Table 6
Precision@K for the top confidence classified tweets.

Tweet category	Top-10	Top-50	Top-100	Classifier	Representation
Individual tweets	1.00	1.00	1.00	SVM	Multigram
Related to transport	0.40	0.78	0.84	Naïve Bayes (SVM)	Multigram
Express a Need	1.00	0.96	0.59	Naïve Bayes (SVM)	Multigram
Report on Event	0.60	0.40	0.31	SVM	Unigram
Opinionated	1.00	0.98	0.96	SVM	Unigram
Mode of transport	1.00	0.88	0.77	SVM	Unigram

Table 7
Precision@K of top graded classified tweets.

Category	Pass	Top-10	Top-50	Top-100	Classifier	Representation
Related to Transport	Conf.	0.40	0.78	0.84	Naïve Bayes (SVM)	Multigram
Related to Transport	H-grade	1.00	0.98	0.98	Naïve Bayes (SVM)	Multigram
Express a Need	Conf.	1.00	0.96	0.59	Naïve Bayes (SVM)	Multigram
Express a Need	H-grade	0.90	0.76	0.70	Naïve Bayes	Multigram
Report on Event	Conf.	0.60	0.40	0.31	SVM	Unigram
Report on Event	H-grade	0.50	0.56	0.57	SVM	Unigram
Mode of transport	Conf.	1.00	0.88	0.77	SVM	Unigram
Mode of transport	H-grade	1.00	0.94	0.90	SVM	Unigram

5.2. Inter annotator agreement and classification success

In order to train and test classifiers for text classification, labelled data is needed to serve as a "gold standard". For that purpose, two domain experts labelled tweets according to the hierarchical classification of transport created in the earlier research. The labelled tweets could then be used to train the classifiers to automatically classify tweets. However, classification of text in natural language is not easy even for humans and this is a well-known problem in information retrieval. In our case two transport experts independently labelled approximately 100 tweets per task. It is interesting to discuss the results in more detail while also looking at the inter annotator agreement of these experts (Table 3), given the results may be impacted by disagreement between annotators. It is worth noting that the automatic classification of messages written by individuals and those that were defined as transport-related was extremely successful in each case, while the inter annotator agreement was only moderate. This can be explained as a general phenomenon related to a known limitation of the Kappa measure when the dataset is biased (as in our case) and hence the values do not accurately reflect the real rate of inter-annotator agreement (Feinstein and Cicchetti, 1990). It is interesting to note that in these cases, even though humans had some disagreements, the textual features extracted from the examples enabled high quality automatic classification (Sections 4.2 and Tables 6 and 7).

Concerning the more specific transport aspects (Section 4.4 and Tables 3 and 6), it appears that the identification of an opinion expressed by users had substantial agreement between annotators and the automatic classification was also good. The identification of messages that express a need had only fair agreement between annotators and also a relatively low F1 value, implying that both humans and machines had difficulties in identifying such messages. The most interesting aspect was the identification of messages that report an event – while human annotators easily identified such messages, the trained classifiers had only moderate success.

It should be noted that one possible reason for the relatively low level of success here may be the small number of messages representing a need (231) and reporting an event (182). This may have impacted on the training of the classifiers, as the training set was highly unbalanced.

The disagreement between annotators needs further analysis and there may be many reasons why this occurs, one being differences in the cultural backgrounds of the experts. These reasons need to be identified and addressed in order to improve the annotation process that is a pre-condition for training classifiers

It is also worth noting that while in this study the focus was on assigning a single label to a message, in real life (and also in some cases in this study) messages may be related to several categories.

Finally, this issue is similar to that of Raykar et al. (2010) who proposed a probabilistic framework for supervised learning with multiple annotators providing labels but no absolute gold standard. Future work may consider exploring the potential of the methods suggested by those researchers in order to achieve a better gold standard.

An interesting alternative to annotation by experts may be the use of crowdsourcing. This may solve the cost issue and the need to resolve disagreements; however, the applicability of crowdsourcing for creating a gold standard needs to be validated first.

5.3. Misclassification and classification challenges

The ultimate goal of any text classifier is to achieve a value for F1 of 1.00, however, this is quite challenging given the fact that the problem involves natural language and its ambiguities. This limitation is amplified for Twitter as, with the limited number of characters in a tweet, a specific jargon has evolved for tweets. Examples of misclassified tweets were given in Section 4, hence this issue is only briefly discussed here. Considerable research effort is still needed to address this, but a possible solution is to apply natural language processing tools and to consider phrases in addition to unigram, bigrams etc. This challenge is not unique to transport but to any text classification task aimed at classifying documents written in natural language. Twitter, with its own language, just adds another level of difficulty.

Moreover, from Table 4, it is clear that some of the datasets we used were unbalanced, and as already noted, this may have impacted the results. The real challenge in this research was to obtain enough labelled tweets. For future work, it is suggested that the experiment is repeated with a larger number of positive examples, together with application of balancing techniques.

5.4. Additional challenges

Given the hierarchy of categories defined by domain experts for this study (which has over 60 categories at the level of leaves in the hierarchy tree) it is clear that a much larger labelled dataset is needed to achieve satisfactory results at this level. This has been illustrated by the use of our simple grading heuristics compared with the confidence level of the classifier and would require a major annotation effort. A possible solution may be a joint community effort in creating a dataset that could be commonly used for research, following the example of the domain of information retrieval (see for instance TREC[8]). It is worth noting that there are cases when even a small number of examples is sufficient, for example the transport mode where terms including "bus", "train" etc. provide a clear and unambiguous identification.

Another option may be to explore the potential of crowdsourcing for message annotation. This needs to be carefully examined, as a major challenge may be the ability of non-experts to accurately label transport-related tweets following a detailed domain ontology.

More importantly, further analysis of the textual content of the messages may reveal the origin and destination of the journey (which in some cases are not explicitly mentioned) and other relevant aspects. Transport messages have temporal characteristics as traffic conditions change. Hence, continuous monitoring of tweets, considering the spatiotemporal aspects, may enable authorities to identify patterns over time such as the build-up of congestion and unexpected disruptions. This may have the potential to enable timely reaction to extreme events and support improved planning to address recurrent patterns.

Another aspect to consider is the removal of "stop words" commonly used terms that are known to add noise to text, however, this should be carefully considered, as for example the word "to", a common stop word, may point to a destination which is important for transport related tweets.

Finally, Twitter has some inherent limitations that need to be taken into account. Its limited size leads users to use hashtags, short forms of words, slang etc, hence it greatly differs from common language. Given the work of Hassan and Menezes (2013) that showed modest improvement in the translation of tweets when normalization was used, this option may be considered while automating the tweet analysis process.

6. Conclusion and future work

The main goal of this paper has been to present a detailed study of the challenges faced when investigating hypotheses concerned with the presence of relevant information in SM and the ability to successfully harvest it, for the benefit of transport stakeholders. To address the issues, the detailed outcomes from a large-scale exercise are presented for the first time. The paper extends current knowledge and understanding for those concerned with SM methodological development, specifically those with an interest in the potential of so-called 'new generation' transport data, and transport practitioners. It illustrates some of the challenges that need to be addressed in developing automatic analysis of SM data and when aggregating micro level SM data concerning transport experiences. In general, customer comments found in SM that concern transport systems tend towards negative sentiments. The automatic analysis of texts in order to discover whether the opinions were broadly positive or negative builds on a long history of development in the capability of machines to understand the content and tone of documents. If successful, such analysis may enable better understanding of public opinions and needs, as a step towards improving transport services.

The previous phase of the research showed that SM contains information of relevance to the transport sector and the further work presented here (Section 1) confirms the hypothesis that Twitter is a valuable and important source of transport-related information. Overall, the results achieved in the experiments show moderate efficacy (i.e. when examining the opinionated, expressing a need and reporting an event classifiers) to high efficacy (i.e. when examining the individual or authority and related to transport classifiers) in harvesting information valuable to the transport sector. Relevance in this

[8] http://trec.nist.gov/.

context of this study was determined by transport experts. This information supports the notion raised in hypothesis II (SM information can be practically harvested automatically or semi automatically). Using the hierarchical classification tree and the text mining framework previously proposed, we evaluated whether the classification performance is useful for practical purposes. The results, though positive and encouraging, pointed towards some of the classic challenges of text mining Twitter messages that require a substantial body of further research.

Several major specific challenges faced when extracting useful transport-related information from Twitter have been elaborated and illustrated. In summary, these are the need for domain ontology in order to enable the training of the classifiers and the achievement of agreement between human annotators. Other possible research directions include the use of crowdsourcing for the manual labelling of examples – this may reduce the cost of labelling and help overcoming inter-annotator disagreement. Moreover, it is suggested that parts of the experiment are repeated with a larger number of positive examples, together with the application of balancing techniques. Finally, normalization of the tweets as well as well as the removal of stop words may be considered.

While it is clear that Twitter contains valuable information, the automatic harvesting of this information is challenging. The impact of a successful solution to these challenges (thereby creating efficient harvesting systems) would be to enable travellers to participate more effectively in the improvement of transport services. The work presented here therefore has interest to those beyond the immediate transport domain who are concerned with the research and policy practice of increasing public participation.

Funding

The research was partially funded through the Worldwide Universities Network (WUN) scheme and by the University of Haifa.

Acknowledgements

This paper also draws on work conducted by Ebithal Sheety (Technion) and Frances Hodgson (University of Leeds).

References

Agichtein, E., Brill, E., Dumais, S., 2006. Improving web search ranking by incorporating user behavior information. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, Seattle, pp. 19–26. Available at: <http://dl.acm.org/citation.cfm?id=1148177> (accessed 14 Jan. 2017).

Bregman, S., 2012. Uses of Social Media in Public Transport. Transit Cooperative Research Program (TCRP) Synthesis 99. Transportation Research Board, Washington.

Carletta, J., 1996. Assessing agreement on classification tasks: the kappa statistic. Comput. Linguist. 22 (2), 249–254. Available at: >http://dl.acm.org/citation.cfm?id=230390> (accessed 14 Jan. 2017).

Chaniotakis, E., Antoniou, C., 2015. Use of geotagged social media in urban settings: empirical evidence on its potential from Twitter. In: IEEE 18th International Conference on Intelligent Transportation Systems (ITSC). IEEE, Canary Islands, pp. 214–219. Available at: <http://ieeexplore.ieee.org/document/7313136/> (accessed 14 Jan. 2017).

Chi, E., 2008. The social Web: research and opportunities. Computer 41 (9), 88–91. Available online at: <http://ieeexplore.ieee.org/document/4623229/> (accessed 14 Jan. 2017).

Collins, C., Hasan, S., Ukkusuri, S., 2013. A novel transit riders' satisfaction metric: riders' sentiments measured from online social media data. J. Public Transp. 16 (2), 21–45. Available at: <http://scholarcommons.usf.edu/jpt/vol16/iss2/2/> (accessed 14 Jan. 2017).

Cornwell, I., Grant-Muller, S., Cross, P., Clarke, M., Heinrich3, Daniel Elias, T., Catchesides, B., 2015. Increasing understanding of the quality of new sources of traffic data. In: *22nd ITS World Congress*, Bordeaux, 12p.

Cottrill, C.D., Derrible, S., 2015. Leveraging big data for the development of transport sustainability indicators. J. Urban Technol. 22 (1), 45–64. Available at: <http://www.tandfonline.com/doi/abs/10.1080/10630732.2014.942094> (accessed 14 Jan. 2017).

D'Andrea, E., Ducange, P., Lazzerini, B., Marcelloni, F., 2015. Real-time detection of traffic from twitter stream analysis. IEEE Trans. Intell. Transp. Syst. 16 (4), 2269–2283. Available at: <http://ieeexplore.ieee.org/document/7057672/> (accessed 14 Jan. 2017).

Daume, S., Albert, M., von Gadow, K., 2014. Forest monitoring and social media–complementary data sources for ecosystem surveillance? For. Ecol. Manage. 316, 9–20. Available at: <http://www.sciencedirect.com/science/article/pii/S037811271300618X> (accessed 15 Jan. 2017).

Endarnoto, S.K., Pradipta, S., Nugroho, A.S., Purnama, J., 2011. Traffic condition information extraction and visualization from social media Twitter for android mobile application. In: 2011 International Conference on Electrical Engineering and Informatics (ICEEI). IEEE, Bandung, pp. 1–4. Available at: <http://ieeexplore.ieee.org/document/6021743/> (accessed 14 Jan. 2017).

Feinstein, A.R., Cicchetti, D.V., 1990. High agreement but low kappa: I. The problems of two paradoxes. J. Clin. Epidemiol. 43, 543–549. Available online at: <http://www.sciencedirect.com/science/article/pii/089543569090158L> (accessed 14 Jan. 2017).

Flyvbjerg, B., Skamris Holm, M.K., Buhl, S.L., 2005. How (in) accurate are demand forecasts in public works projects? – The case of transportation. J. Am. Plan. Assoc. 71 (2), 131–146. Available at: <http://www.tandfonline.com/doi/abs/10.1080/01944360508976688> (accessed 14 Jan. 2017).

Gal-Tzur, A., Grant-Muller, S.M., Kuflik, T., Minkov, E., Nocera, S., 2014a. The impact of social media usage on transport policy: issues, challenges and recommendations. Proc. Soc. Behav. Sci. 111, 937–946. Available at: <http://www.sciencedirect.com/science/article/pii/S1877042814001293> (accessed 14 Jan. 2017).

Gal-Tzur, A., Grant-Muller, S.M., Kuflik, T., Minkov, E., Nocera, S., Shoor, I., 2014b. The potential of social media in delivering transport policy goals. Transp. Policy 32, 115–123. Available at: <http://www.sciencedirect.com/science/article/pii/S0967070X14000225> (accessed 14 Jan. 2017).

Gao, L., Wu, H., 2013. Verb-based text mining of road crash report. In: Transportation Research Board 92nd Annual Meeting. TRB, Washington. Available at: <https://trid.trb.org/view.aspx?id=1241434> (accessed 14 Jan. 2017).

Geyer, W., Freyne, J., Anand, S., Dugan, C., Mobasher, B., 2010. Recommender systems and the social web. In: 2nd Workshop on Recommender Systems and the Social Web, Proceedings of the Fourth ACM Conference on Recommender Systems (RecSys, 10). ACM, Barcelona, pp. 379–380. Available online at: <http://dl.acm.org/citation.cfm?id=1864798> (accessed 14 Jan. 2017).

Grant-Muller, S.M., Gal-Tzur, A., Minkov, E., Nocera, S., Kuflik, T., Shoor, I., 2014. The efficacy of mining social media data for transport policy and practice. In: Transportation Research Board 93rd Annual Meeting. TRB, Washington. Available at: <http://amonline.trb.org/14-1716-1.2494063?qr=1> (accessed 14 Jan. 2017). 14-1716.

Grant-Muller, S.M., Galtzur, A., Minkov, E., Kuflik, T., Nocera, S., Shoor, I., 2015a. Transport policy: social media and user-generated content in a changing information paradigm. Soc. Media Gov. Serv., 325–366 Available at: [accessed 14 Jan. 2017].

Grant-Muller, S.M., Gal-Tzur, A., Minkov, E., Nocera, S., Kuflik, T., Shoor, I., 2015b. Enhancing transport data collection through social media sources: methods, challenges and opportunities for textual data. IET Intel. Transport Syst. 9 (4), 407–417. Available at: <http://ieeexplore.ieee.org/document/7108354/> (accessed 14 Jan. 2017).

Grosenick, S., 2012. Real-time traffic prediction improvement through semantic mining of social networks. MSc dissertation. Available at: <https://digital.lib.washington.edu/researchworks/handle/1773/20911> (accessed 14 Jan. 2017).

Gu, Y., Qian, Z.S., Chen, F., 2016. From Twitter to detector: real-time traffic incident detection using social media data. Transp. Res. Part C: Emerg. Technol. 67, 321–342. Available at: <http://www.sciencedirect.com/science/article/pii/S0968090X16000644> (accessed 14 Jan. 2017).

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software: an update. ACM SIGKDD Explor. Newsl. 11 (1), 10–18. Available at: <http://dl.acm.org/citation.cfm?id=1656278> (accessed 14 Jan. 2017).

Hassan, H., Menezes, A., 2013. Social text normalization using contextual graph random walks. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. ACL, Sofia, pp. 1577–1586. Available online at: <http://www.anthology.aclweb.org/P/P13/P13-1155.pdf> (accessed 14 Jan. 2017).

Java, A., Song, X., Finin, T., Tseng, B., 2007. Why we twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis. ACM, San Jose, pp. 56–65. Available at: <http://dl.acm.org/citation.cfm?id=1348556> (accessed 14 Jan. 2017).

Kaya, M., Conley, S., 2016. Comparison of sentiment lexicon development techniques for event prediction. Soc. Netw. Anal. Min. 6 (1), 1–13. Available online at: <http://link.springer.com/article/10.1007/s13278-015-0315-8> (accessed 14 Jan. 2017).

Kaplan, A.M., Haenlein, M., 2010. Users of the world, unite! The challenges and opportunities of Social Media. Bus. Horiz. 53 (1), 59–68. Available at: <http://www.sciencedirect.com/science/article/pii/S0007681309001232> (accessed 14 Jan. 2017).

Kim, S.-M., Hovy, E., 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In: Proceeding of SST '06 Proceedings of the Workshop on Sentiment and Subjectivity in Text. ACL, Stroudsburg, pp. 1–8. Available online at: <http://dl.acm.org/citation.cfm?id=1654642> (accessed 14 Jan. 2017).

Kim, Y., 2014. Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). ACL, Doha, pp. 1746–1751. Available online at: <http://www.aclweb.org/website/old_anthology/D/D14/D14-1181.pdf> (accessed 14 Jan. 2017).

Kocatepe, A., Lores, J., Ozguven, E.E., Yazici, A., 2015. The reach and influence of DOT Twitter accounts: a case study in Florida. In: IEEE 18th International Conference on Intelligent Transportation Systems (ITSC). IEEE, Canary Islands, pp. 330–335. Available at: <http://ieeexplore.ieee.org/document/7313155/?reload=true&arnumber=7313155>.

Liu, B., 2009. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. [PDF] Springer-Verlag, Berlin, Heidelberg. Available at: <http://www.springer.com/us/book/9783642194597#aboutBook> (accessed 14 Jan. 2014).

Macdonald, C., Ounis, I., Soboroff, I., 2007. Overview of the TREC 2007 blog track. NIST, 13p. Available online at: <http://trec.nist.gov/pubs/trec16/papers/BLOG.OVERVIEW16.pdf> (accessed 14 Jan. 2017).

Mai, E., Hranac, R., 2013. Twitter interactions as a data source for transportation incidents. In: Transportation Research Board 92nd Annual Meeting. TRB, Washington. Available at: <https://trid.trb.org/view.aspx?id=1241097> (accessed 14 Jan. 2017). 13-1636.

Manning, C.D., Raghavan, P., Schütze, H., 2008. Introduction to Information Retrieval. Cambridge University Press, Cambridge. 496p.

Maynard, D., Funk, A., 2011. Automatic detection of political opinions in tweets. In: Proceeding of the 8th International Conference on The Semantic Web. ACM, Heraklion, pp. 88–99. Available online at: <http://dl.acm.org/citation.cfm?id=2186817> (accessed 14 Jan. 2017).

Mihyeon Jeon, C., Amekudzi, A.A., Vanegas, J., 2006. Transportation system sustainability issues in high-, middle-, and low-income economies: case studies from Georgia (US), South Korea, Colombia, and Ghana. J. Urban Plan. Dev. 132 (3), 172–186. http://dx.doi.org/10.1061/(ASCE)0733-9488(2006)132:3(172). Available online at: (accessed 14 Jan. 2017).

Nikfarjam, A., Sarker, A., O'Connor, K., Ginn, R., Gonzalez, G., 2015. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. J. Am. Med. Inform. Assoc. ocu041. Available at: <http://jamia.oxfordjournals.org/content/early/2015/03/08/jamia.ocu041.abstract> (accessed 15 Jan. 2017).

Nocera, S., Cavallaro, F., 2014. The ancillary role of CO_2 reduction in urban transport plans. Transp. Res. Proc. 3, 760–769. Available online at: <http://www.sciencedirect.com/science/article/pii/S235214651400218X> (accessed 14 Jan. 2017).

Nocera, S., Tonin, S., Cavallaro, F., 2015. The economic impact of greenhouse gas abatement through a meta analysis: valuation, consequences and implications in terms of transport policy. Transp. Policy 37, 31–43. Available at: <http://www.sciencedirect.com/science/article/pii/S0967070X14002030> (accessed 14 Jan. 2017).

Pang, B., Lee, L., 2008. Opinion mining and sentiment analysis. Found. Trends Inf. Retr. 2 (1–2), 1–135. Available online at: <http://dl.acm.org/citation.cfm?id=1454712> (accessed 14 Jan. 2014).

Pang, B., Lee, L., Vaithyanathan, S., 2002. Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. ACL, Stroudsburg, pp. 79–86. Available online at: <http://dl.acm.org/citation.cfm?id=1118704> (accessed 14 Jan. 2014).

Pathak, A., Patra, B.K., Chakraborty, A., Agarwal, A., 2015. A city traffic dashboard using social network data. In: Proceedings of the 2nd IKDD Conference on Data Sciences. Bangalore: ACM paper 8. Available online at: <http://dl.acm.org/citation.cfm?id=2778873> (accessed 14 Jan 2017).

Quinlan, J.R., 1993. C4. 5: Programs for Machine Learning. Morgan Kaufmann, San Francisco, CA, USA. 302p.

Rattenbury, T., Good, N., Naaman, M., 2007. Towards automatic extraction of event and place semantics from flickr tags. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, Amsterdam, pp. 103–110. Available online at: <http://dl.acm.org/citation.cfm?id=1277762> (accessed 14 Jan. 2017).

Raykar, V.C., Yu, S., Zhao, L.H., Valadez, G.H., Florin, C., Bogoni, L., Moy, L., 2010. Learning from crowds. J. Mach. Learn. Res. 11, 1297–1322. Available online at: <http://www.jmlr.org/papers/v11/raykar10a.html> (accessed 14 Jan. 2017).

Schulz, A., Ristoski, P., Paulheim, H., 2013. I see a car crash: real-time detection of small scale incidents in microblogs. In: The Semantic Web: ESWC 2013 Satellite Events. Springer, Montpellier, pp. 22–33. Available online at: <http://link.springer.com/chapter/10.1007/978-3-642-41242-4_3> [Accessed 14 Jan. 2017].

Schweitzer, L., 2012. How are we doing? Opinion mining customer sentiment in US Transit Agencies and Airlines Via Twitter. In: Transportation Research Board 91st Annual Meeting. TRB, Washington 12-2659. Available at: <https://trid.trb.org/view.aspx?id=1129878> (accessed 14 Jan. 2017).

Sinha, K.C., Labi, S., 2007. Transportation Decision Making: Principles of Project Evaluation and Programming [PDF]. John Wiley and Sons, p. 576. Available online at: <http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0471747327.html> (accessed 14 Jan. 2014).

Tiroshi, A., Kuflik, T., Kay, J., Kummerfeld, B., 2011. Recommender systems and the social web. In: Ardissono, L., Kuflik, T. (Eds.), Advances in User Modeling, vol. 7138 of the series Lecture Notes in Computer Science, 1st ed. Springer, Heidelberg, Berlin, pp. 60–70.

Tuarob, S., Tucker, C.S., 2015. Quantifying product favorability and extracting notable product features using large scale social media data. J. Comput. Inf. Sci. Eng. 15 (3), 031003. 12p. Available online at: <http://computingengineering.asmedigitalcollection.asme.org/article.aspx?articleid=2090327> (accessed 14 Jan. 2017).

Tumasjan, A., Sprenger, T., Sandner, P., Welpe, I., 2010. Predicting elections with Twitter: what 140 characters reveal about political sentiment. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media ICWSM. AAAI, Washington, pp. 178–185. Available online at: <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1441> (accessed 14 Jan. 2017).

Wikipedia, 2016. Twitter. Retrieved May 2016, from Wikipedia. <http://en.wikipedia.org/wiki/Twitter>.

Witten, I., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco. 664p.

Yamamoto, Y., 2007. Twitter4J: A Java library for the Twitter API. Available online at: <http://twitter4j.org/en/> (accessed 14 Jan. 2017).

Zhang, X., Fuehres, H., Gloor, P., 2010. Predicting stock market indicators through Twitter–"I hope it is not as bad as I fear". Pro. Soc. Behav. Sci. 26, 55–62. Available online at: <http://ac.els-cdn.com/S1877042811023895/1-s2.0-S1877042811023895-main.pdf?_tid=2e830896-da7e-11e6-baf4-00000aacb362&acdnat=1484414796_8a8bd11801097959e68acfeac66739fa> (accessed 14 Jan. 2017).

Zhang, Z., Ni, M., He, Q., Still, S., Gao, J., 2016. Mining transportation information from social media for planned and unplanned events. Final report. Prepared for: Transportation Informatics Tier I University Transportation Center 204 Ketter Hall University at Buffalo, Buffalo, NY, 14260. <https: //www.buffalo.edu/content/www/transinfo/Research/socialmediaminingforevents/_jcr_content/par/download/file.res/MiningSocialMediaEvents_FinalReport.pdf>.