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/© 2017 Elsevier Ltd. All rights reserved.

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.