Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 33 (2018) 291–298

Transportation Research Procedia

www.elsevier.com/locate/procedia

XIII Conference on Transport Engineering, CIT2018

TRANSPORT ANALYSIS APPROACH BASED ON BIG DATA AND TEXT MINING ANALYSIS FROM SOCIAL MEDIA

Ainhoa Serna[a,*] Slaven Gasparovic[b]

[a]*Computer Science Department, Faculty of Engineering, Mondragon Unibertsitatea, Arrasate-Mondragón 20500, Spain, aserna@mondragon.edu*

[b]*Department of Geography, Faculty of Science, University of Zagreb, Zagreb 10 000, Croatia, slaveng@geog.pmf.hr*

Abstract

The goal of the study of the paper is to propose a dashboard with dynamic graphics using a qualitatively and quantitatively approach to investigate the tourists' satisfaction according by transport mode used. The methodology implemented in the research includes data collection from TripAdvisor.com with geographic locations and their integration with statistical territorial data. Text mining techniques are applied in order to assess tourists' perceptions on success factors, which may be used as planning support tools. The case study concerns Croatia country and shows the value and complementarity of Social Media-related data with official statistics for transport and tourism planning.

© 2018 The Authors. Published by Elsevier Ltd.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0/)
Selection and peer-review under responsibility of the scientific committee of the XIII Conference on Transport Engineering, CIT2018.

Keywords: transport; social media; text mining; natural language processing; user generated content

1. Introduction

Tourism is a complex socio-economic phenomenon. Its effects are manifested at every level, from global to local. Tourism is in correlation with other economic activities, including transportation (Pasalic, 2001; Jugovic, Kovacic, Saftic, 2010). Transportation and transportation systems represent the basis of the modern economy of any space since successfully developed business activities, including tourism, could not be present without developed transportation and transportation system (Gasparovic, 2011).

2352-1465 © 2018 The Authors. Published by Elsevier Ltd.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0/)
Selection and peer-review under responsibility of the scientific committee of the XIII Conference on Transport Engineering, CIT2018.
10.1016/j.trpro.2018.10.105

Tourism has a direct impact on transportation since it requires the development of transportation systems used by tourists. Tourism is impossible without transportation because transportation always precedes tourism. Tourism is characterized by temporary migration of tourists. Therefore, tourism implies movement, the change of living location and the use of transport modes (Viducic, Viducic, 2004).

Transport and its infrastructure are important factors of tourism (Mrnjavac, Marsanic, Krpan, 2008). Transport has a multiple impact on tourism. It is the basic component of tourism whose main aim is connecting emissive markets and tourist destinations. Transport also facilitates mobility and accessibility within a tourist destination and can also be combined with some physical or social factor, or be a tourist attraction by itself (Knowles, Shaw and Docherty, 2008).

For that reason, the aim of this research is to determine insights regarding to transportation bearing in mind the geographical perspective, through the automatic identification of the different modes of transportations in TripAdvisor comments. The methodology is a quantitative and qualitative content analysis using text mining techniques taking as reference a categorization framework based on WordNet lexical database and SUMO ontology. This paper demonstrates empirically the feasibility of the automatic identification of transportation assessments in the discourses generated by the user generated content, through a powerful ad-hoc software combining Natural Language Processing field tools.

The structure of the paper is as follows: section 2 presents an overview of research in the field of Transport through Social Media, as well as the contributions of this article. After that, section 3 explains the methodology followed in this study. Section 4 outlines the case study, section 5 shows the results, and in the last section, the conclusions and future lines are explained.

2. Related work

It is noteworthy that are recent investigations of Social Media analysis in the field of urban transport (Serna, Gerrikagoitia, Bernabe & Ruiz, 2017a; Kuflik, Minkov, Nocera et al., 2017) to explore from another perspective urban mobility. For example, on-line social networks are being used to conducting transport surveys (Efthymiou & Antoniou, 2012). Lately, substantial increase in research in this area is appreciated.

Rashidi et al. (2017) presents results of a quality survey form travel demand-modelling experts around the world on applicability of Social Media data for modelling daily travel behavior. The results of the survey reveal positive view of experts about usefulness of such data sources. Furthermore, Pereira (2017) designed and developed a framework for collection, processing and mining of geo-located Tweets. More specifically, it provides functionalities for parallel collection of geo-located tweets from multiple pre-defined bounding boxes (cities or regions), including filtering of non-complying tweets, text pre-processing for Portuguese and English language, topic modelling, and transportation-specific text classifiers, as well as, aggregation and data visualization.

According to Ruiz, Mars, Arroyo and Serna (2016) there are different potential sources of transport data, which is characterized by the huge amount of information available, the velocity with it is obtained and the variety of format in which is presented. This sort of information is commonly known as Big Data. To use this data on Transport Planning application is a challenge, which require employing complex data mining techniques. They identified potential sources of social network related big data that can be used in Transport Planning, discussing their advantages and limitations. Then, a review of current applications in Transport Planning is presented. Finally, some future prospects of using social network related big data that are included in the MINERVA project are highlighted.

Gal-Tzur et al. (2014) finds possible practices of Social Media for transport service suppliers and transport policy makers indicating that transport policy significant data can be collected from user-generated content. Grant-Muller (2014) give the details of technical defiances associated with mining Social Media data in transport. A text mining procedure to obtain significant data from transport sector corpus is presented including taxonomies, polarity analysis and measuring exactness, presenting the basis for new research in this field. Grant-Muller (2015) validates that data gathered from user generated content can supplement, enhance (or even substitute) traditional data collection, emphasizing the importance of improving automatic approaches to collecting and analysing Social Media information related to transportation.

Moreover, Bregman (2012) explores the utilization of user-generated content between transit agencies and documents successful modus operandi in the United States and Canada. An examination of significant literature was

merged with discoveries from a survey of designated transit agencies. Based on survey results, various case studies were elaborated to define pioneering and effective practices more specifically. Serna et al. (2017b) empirically prove the viability of the programmed identification of the Sustainable Urban Mobility difficulties in the comments produced by the user generated content. Their methodology improves the information of the traditional surveys, increases traditional analysis with Big Data approaches, through Sentiment Analysis methods. Gu et al (2016) suggest mining posts from Twitter to find incident data on both highways and arterials as an effective and cost-efficient option to existing data sources. In addition, they expose an approach to crawl, process and filter tweets that are available by the public for free.

3. Methodology

The methodology follows a quantitative and qualitative content analysis approach (Walle, 1997) using text mining techniques and takes into account user - generated content in TripAdvisor Social Media sources. TripAdvisor bills itself as the world's largest travel site "with 435 million reviews and opinions covering 6.8 million accommodations, restaurants and attractions, and a wide variety of travel choices and planning features,…reaching 390 million average monthly unique visitors" (TripAdvisor, 2017). TripAdvisor relies mostly on UGC, in the form of individual reviews, to provide advice and booking options for visitors. It is remarkable that TripAdvisor has the highest ranking in search engines and supplies the biggest amount of travel reviews in the tourism sector (De Ascaniis and Gretzel, 2013).

The process consists on the following steps:

- Source Identification. A manual process has been carried out to reveal sources of relevant data to be analysed. On one hand, TripAdvisor comments refer to experiences that include information about mobility as one of the categories. The data collection has been performed selecting posts from the different categories related to transportation, such as Transport, Walking Tours in Croatia, Walking & Biking Tours in Croatia, Bike & Mountain Bike Tours in Croatia, Day Trips & Excursions in Croatia, Private & Custom Tours in Croatia, Walking Tour of Dubrovnik,... In turn, the category Transport is broken down into 4 subcategories: Taxis and others, Ferries, Public transport systems, and railway. Taxis and others category include transfer, shuttle and taxi transport mode. All are individual public transport. Ferries category include ferries and taxi boat. Public transport systems category comprise Zagreb electric Tram and funicular railway. Funicular railway and tram are 'guided' modes of transport (with cables the first, and lanes the second). That's why TripAdvisor groups them as a railway and collective public system.

- Data acquisition. It is the process of gathering, filtering and cleaning the unstructured data before making them persistent in a storage solution on which data analysis can be carried out. In this step, the information is extracted automatically from TripAdvisor with scraping techniques.

- Data preparation for analysis. This process is concerned with making the raw data amenable to use in decision - making. First, the comments are loaded one by one and after that, the language about the comments from Social Media is detected with Shuyo language detector (Shuyo, 2010). Later, the texts are corrected using Aspell (Atkinson, 2003), a spell checker that is customized with localism and abbreviation. Abbreviations are usually used instead of full names, so it is necessary the normalization of the comment. The proper matching between the abbreviations and the right word is a critical process. Once the text is corrected, each word is morph syntactically noted, using Freeling methodology (Padró & Stanilovsky, 2012). After full parsing process through Freeling language analyser tool, the Named Entity classification and WordNet sense annotation using UKB disambiguation (Agirre and Soroa, 2009), is obtained. UKB is a collection of programs for performing graph - based Word Sense Disambiguation (WSD) and lexical similarity/relatedness using a pre - existing knowledge base.

- Data Curation. This phase categorizes, groups and analyses concepts (terms or common nouns) and the adjectives and adverbs with which the different modes of transport are qualified. The concepts (syntactically

identified such as nouns / substantive) that appear into the comments are grouped into categories. These categories are defined in standard taxonomies /ontologies /vocabularies. Thus, categories are identified automatically. For this purpose, WordNet lexical database (Miller, 1995) aligned with Suggested Upper Merged Ontology (SUMO) ontology (Niles, Pease, 2003), and ad-hoc software is used. WordNet is a lexical database that relates hyponyms/hypernyms with sets of synonyms called synsets, which can be interpreted as specialization relations between conceptual categories. SUMO is the formal ontology that has been mapped to all of the WordNet lexicon. In this way, the concepts are classified into categories. Among all identified categories by SUMO, only those related to 'Transportation' or 'TransportationDevice' (travel mode) are selected.

- Data Storage. This process is the responsible of storing and managing data in a scalable way satisfying the needs of the applications that require access to the data and it is optimized for indexing large volumes of data in real time. For this, Apache Solr (Smiley, Pugh, Parisa, & Mitchell, 2015) is used.

- Data Visualization, powered by Kibana dashboard (Gupta, 2015) to create rich and flexible user interfaces, enabling users leverage the power of Apache Solr time series (see Fig. 1).

4. Case Study

According to the surface Croatia belongs to the group of smaller European countries (an area of 56,594 km^2). It is located at the contact of three geographical units: the Mediterranean, the Dinaric Mountains and the Pannonian Plain. Although small in size, it is characterized by great geographical diversity. This diversity manifests in geomorphological, climatic, hydrogeographic and biogeographical features. Different political and cultural influences throughout history have left a large number of cultural attractions. Because of all this, Croatia is attractive European country (Curic, Glamuzina, Opacic, 2013). With regard to tourist regionalization, the Republic of Croatia can be divided into five geographical regions (Curic, Glamuzina, Opacic, 2012; Curic et al., 2013): Northern Littoral, Southern Littoral, Mountainous, Peripannonian and Pannonian. These regions can also be divided into regions of lower (second) order: Northern Littoral region is divided into Istria and Kvarner and covers the entire island of Pag. Southern Littoral region is divided into North, Central and South Dalmatia. Mountainous region consists of two regions of Gorski Kotar and Lika, and Pannonian and Peripannonian are not divided into lower-order regions (Curic, Glamuzina, Opacic, 2013).

Depending on the data of the Croatian Bureau of Statistics (2017) in 2016 the number of tourists in the Republic of Croatia was 15,594,000 (of which 13,809,000 foreign and 1,786,000 domestic). They made 78,050,000 overnight stays (of which 72,193,000 foreign and 5,857,000 domestic). Marked predomination of foreign tourists (89% of foreign tourists made 93% of overnight stays) is obvious. Tourism in Croatia is characterized by spatial disparity. The difference is expressed between the coastal and continental parts. The North and South Adriatic region are most important Croatian tourist regions. In 2016, there were 95% of all tourist beds in the Republic of Croatia, and 87% of tourist arrivals and 95% of tourist nights were realized. Foreign tourists predominate in these regions of Croatia (in 2016, 90% of foreign tourists realized 97 % of overnight stays). Tourism is an extremely important economic branch in the Republic of Croatia since it accounts for 18.1% of GDP compared to the average of EU (4.7%) (Eurostat, 2017).

The study is based on comments from TripAdvisor about transportation in Croatia from 12/12/2007 to 06/10/2017. In this period there are 12,928 reviews and 11,924 of them are written in English. The five top languages are: English, Spanish, Italian, Portuguese and French. It is noteworthy that approximately 92% of the comments are written in English. For this reason, English was chosen for transportation analysis and the rest of languages were discarded.

5. Results

We have developed a dashboard platform with dynamic graphics that analyses Social Media data from TripAdvisor with opinions of visitors, identifying positive and negative factors and their potential impact on

sustainable tourism and transport. Furthermore, the different modes of transport are grouped and filtered by date, location, rating, transport mode and language. Also, the original comment and its corresponding title are acquired.

This dashboard facilitates the interpretation of the results through different types of graphs (tables, pie chart, histogram, word cloud). Starting at the top left of the Fig. 1 (top section), there is a first block that allows to filter by range of dates. The following block allows to search and filter by concept. Then, there is a cloud tag that shows the different modes of transport found. Next, there is a pie chart with the five top languages (English, Spanish, Italian, Portuguese and French). Finally, the last block of this section that is upper part shows a summary of the information about dates and dates format.

In the middle section of the Fig. 1, the first results are described in a temporal evolution line chart with the number of observations (comments) corresponding to TripAdvisor from 2007 to 2017. Years are shown on the x axis and the number of comments on the y axis. Also, Fig. 1 shows that the research topic is gaining relevance and presence in social media year by year. Note that 2017 is an incomplete dataset yet, with approximately only 9 months of observations gathered. Moreover, this histogram is a dynamic graph, in consequence, it is possible to select different frames of time and the graphic is updated automatically.

In the bottom section, there is a first block that allows filtering by term, the second block by adjective and adverb, the third block by cities and the last block by star rating.

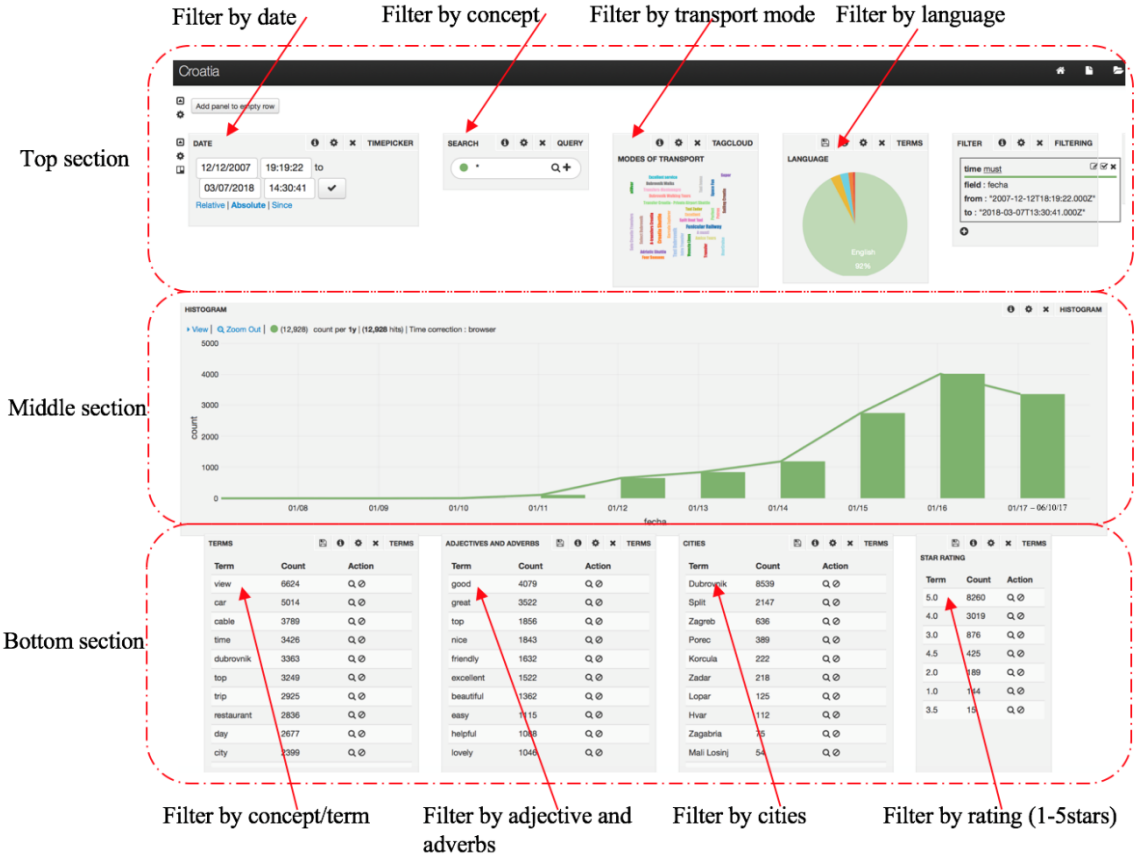

Fig. 1. Dashboard of the Transport Modes from Croatia in TripAdvisor

Grouping the different transport modes and filtering by English language, the next distribution (Table 1) is obtained. Almost two thirds of the transport modes correspond to the *tram*. The second place with 35% correspond to *Taxi and others*. With only a 3%, the *railway* is represented, and with 2% the *ferry*. *Private Tours*, *boat*, *bus* and *car* is almost non-existent.

Table 1. Transport modes of the comments by rating in English.

	# stars							
Transport Mode	5	4.5	4	3.5	3	2	1	Total
Tram	3,872	0	2,312	0	703	146	61	7,094
Taxi and others	3,548	388	145	15	22	8	43	4,169
Railway	124	0	98	0	70	10	11	313
Ferry	191	0	47	0	27	15	15	295

Zagreb is the only city that has presence from Peripannonian region. It is noteworthy, that *Zagreb* is the third more commented city. 75.4% of the comments are about the *tram*, 22.4% *taxi and others* and 2.2% bus transport mode. Also, 51% are rated with 5 stars and 26% with 4 stars, in sum 77% are very positive rated. Only 5% are rated with 1-2 stars. In concrete, railway is rated 40% with 5 star, 31% with 4 stars and 7% with 1-2 stars. *Bus* 100% with 4 stars rated. *Taxi and others* 95% with 5 stars and 5% with 4-4.5 stars.

Corresponding to the 'Northern Littoral' region, the following cities are present in TripAdvisor: *Porec, Rovinj, Mali Losinj* and *Pula*. 85.5% of the comments about *Porec* transport are rated with 4-5 stars, 9.5% with 3 stars and only 5% are rated with 1 star that is the worst punctuation. More than a half of the comments (52%) are about the *ferry* transport mode and 48% about *taxi and others*. Moreover, *Rovinj*, *Mali Losinj* and *Pula* have a low presence in TripAdvisor and all the comments are very positive. Exactly, in *Rovinj's* comments the distribution is 100% *taxi and others*. Besides, all the comments are rated with the maximum punctuation, 5 stars. Also, total comments referring to *Mali Losinj* are about *taxi and others* category. Moreover, 76% are rated with 4.5 stars and 24% with 3.5 stars. Finally, in *Pula* transport modes, comments about *ferry* and *taxi and others* categories appear with a distribution of 80% and 20% respectively. It should be noted that the rating are very high, 5 stars and 4 stars rated corresponding to the previous distribution.

There is a greater presence in the 'Southern Littoral' region, with comments about 7 cities such as *Dubrovnik*, *Split*, *Korcula*, *Zadar*, *Hvar*, *Makarska* and *Trogir*. *Dubrovnik* has a very prominent presence with respect to the rest of cities in Croatia. The detailed analysis is shown in Table 1 and described below. Regarding *Split*, *taxi and others* (97%), *ferry* (2%) and *private tours* (1%) are found. 95% are 5 stars rated, 97% are about *Taxi and others*, and 2% about ferry and 1% private tours mode. In *Korcula* and *Makarska*, taxi and others are 100% of the comments. In *Zadar*, 75% *taxi and others* and 25% *ferry*. In *Hvar*, 55% are *taxi and others*, *ferry* is 26% and *boat* is 19%. Last, *Trogir* has the following distribution: 71% *ferry* and 29% *taxi and others*. Majority of the comments about 'Southern Littoral' region are about *taxi and others* and *ferry* except transport modes in *Dubrovnik*.

Almost the 86% of the comments are about the transport mode *Tram*, which is 55% rated with 5 stars, 33% with 4 stars and only 3% with 1-2 star. Furthermore, 24% are about '*Taxi and others*' transport mode with 88% five stars, 5% four stars, 6% three stars and 1% two stars rated.

About the tram, the most highlighted concepts are the views that are qualified as *amazing*, *spectacular*, *fantastic*, *incredible*, *panoramic*, *wonderful*, *stunning*, *great*, *priceless*. The tram is described like a *great experience*, *highly recommend*, *reasonable price, very clean*, *quick* and *efficient*, *modern*. With negative attributes are qualified the tickets sellers *extremely rude*, *queue badly* organised, *expensive*, *small*, *bad*...

Most of the time, 'Taxi and others' transport mode is qualified with positive attributes such as *great*, *good*, *friendly*, *beautiful*, *excellent*, *wonderful*, *fantastic*. Only 1% are negative attributes, such as *dirty*, *expensive*, *bad*.

6. Conclusions

The platform and its visual analytics capabilities with dynamic graphics and text mining techniques facilitates the interpretation of the results gathered from Social Media. In this way, this platform is a powerful tool to identify positive and negative factors and their potential impact on sustainable tourism and transport.

87% of the comments about transport are positive. Furthermore, most of the comments are about the southern littoral region, highlighting the city of *Dubrovnik*.

Due to the comments found about the *cable car* in *Dubrovnik*, it would be advisable that they should take more care of the service, both in the treatment to the public and in the organization of the sale of tickets. On the other hand, cable car's maintenance and cleaning are very well qualified.

In the Peripannoian region, three quarters of the comments are about a public transport mode, the tram. In the regions of the northern and southern littoral except *Dubrovnik*, the greatest presence is the modes of individual public transport, such as taxi, shuttle or transfer. In addition, maritime modes such as the ferry appear.

A larger Social Media dataset will make it possible to use disaggregate opinions to study the use of any travel mode. The usefulness of Social Media data for modelling individual choice behaviour increase because it is cheaper to obtain than the information collected using surveys. Additionally, it is available in real time and for large periods, which provide the possibility of carrying out a dynamic analysis.

The results of the research are potentially useful for policy and decision making related to the sustainable mobility of people in urban areas and urban transport. The current study proves that text mining of Social Media data can be used as a complementary approach to the conventional methods to study travel behaviour, obtaining a richer and more complete picture for urban transport planning. The availability of continuous Social Media data would allow to update demand prediction models, and to monitor the quality of the different transport modes. Thereby, public and private investments in transport could be more efficient.

For future work, the research will provide data for predictive analytics. Through the creation and sharing of a 'Sentiment Labelled Sentences Data Set' as training dataset for supervised learning algorithm. In fact, it will be completed with the creation of a training machine learning models through the created corpus (positive and negative comments) of this research. Furthermore, it will be measure the accuracy of the different algorithms such as maximum entropy, SVM, SLDA, BAGGING, RF, decision TREE model. In this way, content of Social Media in the transport domain that is not rated can be evaluated automatically.

Acknowledgements

Ministry of Economy and competitiveness in Spain. Project reference: TRA2015-71184-C2-2-R (MINECO/FEDER, UE). MINERVA- Innovative Travel Data Collection Methods for Transport Planning coordinated by Universitat Politècnica de Valencia and Mondragon Unibertsitatea.

Cost Action TU1305 Social Networks and Travel Behaviour. COST is supported by the EU Framework Programme Horizon 2020.

References

Agirre, E., & Soroa, A., 2009. Personalizing pagerank for word sense disambiguation. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (pp. 33-41). Association for Computational Linguistics.

Atkinson K., 2003. GNU Aspell. Retrieved from http://aspell.sourceforge.net/

Bregman, S., 2012. Uses of Social Media in public transportation (Vol. 99). Transportation Research Board.

Croatian Bureau Of Statistics, 2017. Tourism in 2016. Statistical Reports, No. 1594. Croatian Bureau of Statistics, Zagreb.

Curic, Z., Glamuzina, N., Opacic, V. T., 2012. Contemporary Issues in the Regional Development of Tourism in Croatia. Croatian Geographical Bulletin, 74(1).pp. 19-40.

Curic, Z., Glamuzina, N., Opacic, V. T., 2013. Geografija turizma - regionalni pregled (Geography of Tourism - Regional Overview, In Croatian), Naklada Ljevak, Zagreb, Croatia, pp. 280.

De Ascaniis, S., & Gretzel, U., 2013. Communicative functions of online travel review titles. A pragmatic and linguistic investigation of destination and attraction OTR titles. Studies in Communication Sciences, 13(2), pp. 156–165.

Efthymiou, D., & Antoniou, C., 2012. Use of Social Media for transport data collection. Procedia-Social and Behavioral Sciences, 48, pp. 775-785.

Eurostat Tourism Statistics, http://ec.europa.eu/eurostat/statistics-explained/index.php/Tourism_statistics, accessed 12/12/2017.

Gal-Tzur, A., Grant-Muller, S. M., Kuflik, T., Minkov, E., Nocera, S., & Shoor, I., 2014. The potential of Social Media in delivering transport policy goals. Transport Policy, 32, pp. 115-123.

Gasparovic, S., 2011. Air Transportation and Tourism in the Croatian Littoral, Geoadria, 16 (2), pp. 155-187.

Grant-Muller, S. M., Gal-Tzur, A., Minkov, E., Kuflik, T., Nocera, S., & Shoor, I., 2015. Transport Policy: Social Media and User-Generated Content in a Changing Information Paradigm. In Social Media for Government Services (pp. 325-366). Springer International Publishing.

Grant-Muller, S. M., Gal-Tzur, A., Minkov, E., Nocera, S., Kuflik, T., & Shoor, I., 2014. Enhancing transport data collection through Social Media sources: methods, challenges and opportunities for textual data. IET Intelligent Transport Systems, 9(4), pp. 407-417.

Gu, Y., Qian, Z. S., & Chen, F., 2016. From Twitter to detector: Real-time traffic incident detection using Social Media data. Transportation research part C: emerging technologies, 67, pp. 321-342.

Gupta, Y., 2015. Kibana Essentials. Packt Publishing Ltd.

Jugovic, A., Kovacic, M., Saftic, D., 2010. Choice of destination, accomodation and transportation in times of economic crisis, Tourism and Hospitality Management, 16 (2), pp. 165-180.

Kennedy, A., & Inkpen, D., 2006. Sentiment classification of movie reviews using contextual valence shifters. Computational intelligence, 22(2), pp. 110-125.

Knowles, R., Shaw, J., Docherty, I., 2008. Transport geographies: mobilities, flows and spaces, Blackwell Publishing Ltd., Oxford, UK, pp. 293.

Kuflik, T., Minkov, E., Nocera, S., Grant-Muller, S., Gal-Tzur, A., & Shoor, I., 2017. Automating a framework to extract and analyse transport related Social Media content: The potential and the challenges. Transportation Research Part C: Emerging Technologies, 77, pp. 275-291. ISSN 0968-090X.

Miller, G. A., 1995. WordNet: a lexical database for English. Communications of the ACM, 38(11), pp. 39-41.

Mrnjavac, E., Marsanic, R., Krpan, Lj., 2008. Influence of traffic on the development of tourism in the town of Opatija. In International Scientific Symposium Transport Systems (3-4), pp. 264-267.

Niles, I., & Pease, A., 2003. Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In Ike (pp. 412-416).

Padró, L., & Stanilovsky, E., 2012. Freeling 3.0: Towards wider multilinguality. In LREC2012.

Pasalic, Z., 2001. Razvojna međuovisnost i konfliktnost prometa i turizma (Interdependence and Conflict of Transport and Tourism, In Croatian), Modern Traffic, 21 (3-4), pp. 155-160.

Pereira, J. F. F., 2017. Social Media Text Processing and Semantic Analysis for Smart Cities. arXiv preprint arXiv:1709.03406.

Rashidi, T. H., Abbasi, A., Maghrebi, M., Hasan, S., & Waller, T. S., 2017. Exploring the capacity of Social Media data for modelling travel behaviour: Opportunities and challenges. Transportation Research Part C: Emerging Technologies, 75, pp. 197-211.

Ruiz, T., Mars, L., Arroyo, R., & Serna, A., 2016. Social Networks, Big Data and Transport Planning. Transportation research procedia, 18, pp. 446-452. ISSN 2352-1465.

Serna, A., Gerrikagoitia, J. K., Bernabé, U., & Ruiz, T., 2017a. Sustainability analysis on Urban Mobility based on Social Media content. Transportation Research Procedia, 24, pp. 1-8.

Serna, A., Gerrikagoitia, J. K., Bernabe, U., & Ruiz, T., 2017b. A Method to Assess Sustainable Mobility for Sustainable Tourism: The Case of the Public Bike Systems. In Information and Communication Technologies in Tourism 2017 (pp. 727-739). Springer, Cham.

Shuyo, N., 2010. Language detection library for java. http://code.google.com/p/language-detection/. Last access 2015-07-09.

Smiley, D., Pugh, E., Parisa, K., & Mitchell, M., 2015. Apache Solr enterprise search server. Packt Publishing Ltd.

Tripadvisor, 2017. Retrieved from https://www.tripadvisor.com/PressCenter-c6-About_Us.html

Viducic, Lj., Viducic, V., 2004. Uloga prometa i morskoga putničkog brodarstva u razvitku hrvatskog turizma (The Role of Traffic and Maritime Passenger Shipping in the Development of Croatian Tourism, In Croatian), Modern Traffic, 24 (1-2), pp. 141-145.

Walle, A. H., 1997. Quantitative versus qualitative tourism research. Annals of Tourism Research, 24(3), pp. 524-536.