Extracting Social Media data for the use of transportation stakeholders

Itay Shoor
Computer Science Dep., University of Haifa

Tsvi Kuflik
Information Systems Dep., University of Haifa

Einat Minkov
Information Systems Dep., University of Haifa

Susan M. Grant-Muller
Institute for Transport Studies, University of Leeds

Ayelet Gal-Tzur
Transportation Research Institute, Technion - Israel Institute of Technology

Silvio Nocera
Architecture and arts Dep., IUAV University of Venice

The decisions of transportation planners and policy makers affect transportation systems, which in turn, affect the community in terms of wellbeing and growth (both economical and commercial) (Mihyeon, Amekudzi & Vanegas, 2006; Sinha & Labi, 2011). In order to achieve educated decisions, policy makers need to analyze the needs of the users' community. This is done by analyzing current transportation system's performance (by collecting information such as - travel times, waiting times, delays, fluctuations, congestions, public transport usage, etc.) and the customer's perceived quality – allowing location of root dissatisfaction causes (Sinha & Labi, 2011). The most common methods of gathering such data nowadays are questionnaires, focus groups and conducting surveys (Richardson, Ampt & Meyburg, 1995). The bloom of user generated content (UGC) in the social media (SM), where users express their opinion on every matter, suggests that SM may be a source for the needed transport related data. Although SM (and specifically Twitter) has been used in the industry and academia for various data collection tasks, for example, using tweets to replace election polls (O'Connor et al. 2010), to detect a large-scale event (such as an earthquake or a traffic jam) in real-time (Sakaki, Okazaki & Matsuo, 2010) or even to predict stock market indicators (Zhang, Fuehres, & Gloor, 2010), limited use has been made in the transportation related domain.

This work explores the potential of SM as a source for transportation related data collection. Specifically, we used Twitter that in many cases is used for distributing short, immediate messages about events happening in real-time. We applied supervised machine learning techniques for automatic filtering and classification of transportation related content which included information about transportation needs, opinions about transportation infrastructure and updates about transportation events from private users as well as from public authorities.

With the help of transport domain experts, an initial lexicon of transport related terms was created and used for initial information harvesting. The data collected was then annotated by the domain experts and then used to train and evaluate classifiers that were applied sequentially as a pipeline for the task of identifying transport related tweets, that were written by individuals and express personal opinion about different modes of transportation. Other identified tweets include tweets that report on transportation related events (for

example congestions and accidents) and express a need for transportation (planning on reaching a destination or event). We evaluated the classifiers using 10-fold cross validation (examining precision, recall and F_1) and in addition, we measured the classifiers success using precision@k based on the classification confidence (Manning, Raghavan, & Schütze, 2008).

Results of the cross validation for the detection of tweets written by private users and tweets that contain information related to transportation proved to be high (with F_1=0.905 and F_1=0.965 respectively). In other cases, where we trained classifiers to detect the purpose of the message, we achieved moderate success. Specifically, classifiers that detect opinions and reports on transport related events achieved F_1 performance of 0.735 and 0.618, respectively. The results of the classifier detecting whether a tweet expresses a need for travel showed that this task is more challenging (F_1=0.558). In general, we find the results to be encouraging. We believe that using enhanced feature schemes, including the representation of entities that are mentioned in the tweets, e.g., location names, will improve performance substantially. Increasing the volume of data that is used for training is another direction for future work, which can further improve the classification results.

The evaluated performance of the classifiers shows the high potential of using SM for transportation information elicitation for a variety of tasks, where use of information extraction, classification and sentiment analysis can enrich, supplement and in some cases even replace legacy data collection techniques, such as surveys. Further investigation on the results of methodology used in this work show that the use of a heuristic grading method, based on domain specific lexicon, contributed to our results, thus strengthening the notion that use of domain ontology can improve classification results.

References

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval (Vol. 1). Cambridge: Cambridge University Press.

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. Journal of urban planning and development, 132(3), 172-186.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. ICWSM, 11, 122-129.

Richardson, A. J., Ampt, E. S., & Meyburg, A. H. (1995). Survey methods for transport planning. Melbourne: Eucalyptus Press

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

Sinha, K. C., & Labi, S. (2011). Transportation decision making: Principles of project evaluation and programming. John Wiley & Sons.

Zhang, X., Fuehres, H., & Gloor, P. (2010). Predicting stock market indicators through Twitter--“I hope it is not as bad as I fear”., (pp. 1-8).