USER GEOLOCATED CONTENT ANALYSIS FOR URBAN STUDIES: INVESTIGATING MOBILITY PERCEPTION AND HUBS USING TWITTER

M. E. Molinari, D. Oxoli, C. E. Kilsedar, M. A. Brovelli*

Department of Civil and Environmental Engineering, Politecnico di Milano, Milan, Italy -
(moniaelisa.molinari, daniele.oxoli, candaneylul.kilsedar, maria.brovelli)@polimi.it

Commission IV, WG IV/4

KEY WORDS: Twitter, Mobility, User Perception, Urban Management, Python, Geo-crowdsourcing

ABSTRACT:

The availability of content constantly generated within the Web has resulted in an incredibly rich virtual social environment from which it is possible to retrieve almost any sort of information. Since the advent of the social media connection with location-based services, this information has attracted the interest of manifold disciplines connected to the spatial data science. In this context, we introduce the URBAN-GEO BIG DATA (URBAN GEOmatics for Bulk Information Generation, Data Assessment and Technology Awareness), a Project of National Interest funded by the Italian Ministry of Education that aims at contributing to the exploitation of heterogeneous geodata sources such as VGI, geo-crowdsourcing, earth observation, etc. for a better understanding of urban dynamics. The presented work tackles one of the tasks requested by the project, which is connected to an investigation of the use of Twitter as a geodata source for retrieving valuable insights on the citizens' interaction with mobility services and hubs. The study refers to five Italian cities, namely Milan, Turin, Padua, Rome, and Naples. Data collection is performed through the use of the Twitter streaming application programming interface. Collected data is analyzed by means of natural language processing techniques with Python. Results include a) extractions of mobility-related tweets presented by means of maps enabling the exploration of their spatial distribution within the cities, and b) a classification of the mobility-related tweets by means of sentiment analysis, allowing to investigate citizens' perceptions of mobility services. A light and reproducible procedure to achieve these results is also outlined. In general terms, the results are intended for providing snapshots of the citizen interaction with both mobility infrastructure and services enabling a better description of mobility patterns and habits within the studied cities. The work leverages the geo-crowdsourced data within the traditional urban management practices in Italy and investigates the benefits, drawbacks, limitations connected to these data sources, which is the ultimate goal of the URBAN-GEO BIG DATA project.

1. INTRODUCTION

In the last decade, the diffusion of mobile devices has promoted a rich virtual social environment and a growing interaction with location-based services and social media for everyday tasks such as accessing local news, consulting mobility information, share opinions, etc. (Bothorel et al., 2018). Besides the private use of these technologies, the opportunity of sharing content in such a direct and informal way has definitely attracted the attention of many actors involved in both the business sector (Jussila et al., 2014) as well as in the public administration (Mergel and Bretschneider, 2013). One of the main reasons behind this interest is the underlying capability of this phenomenon to provide a valuable digital footprint in a geographical context. This can be exploited to analyze users' interaction with - and perception of - the physical territory on which they live or move through (Graells-Garrido et al., 2018).

According to this, we introduce here the URBAN-GEO BIG DATA (URBAN GEOmatics for Bulk Information Generation, Data Assessment and Technology Awareness), a Project of National Interest (PRIN) funded by the Italian Ministry of Education that aims at contributing to the exploitation of the passive geo-crowdsourced data, the Volunteer Geographic Information (VGI), and the user-generated content to better understand a number of urban dynamics. In this work, we use Twitter (https://twitter.com) as a geo-crowdsourced data source for gathering valuable insights and assessing citizens' relationships with a specific part of the urban environment, namely the mobility services and hubs. The study refers to five Italian cities, namely Milan, Turin, Padua, Rome, and Naples. We collected and preprocessed tweets using natural language processing (NLP) techniques to both detect the most significant city mobility hubs as well as to investigate the citizens' perception of the mobility services.

Despite the extensive use of Twitter data in the social and urban sciences literature (Bao et al., 2017, Frias-Martinez and Frias-Martinez, 2014, Kovacs-Gyori et al., 2018) for the study areas few extensive mobility-related analyses based on social media data are available. Therefore, this work aims also at contributing to the assessment of benefits and drawbacks deriving from the adoption of such data sources within comprehensive urban analysis frameworks in Italy.

The paper continues as follows. In Section 2, we provide a more detailed introduction of the URBAN-GEO BIG DATA project in order to describe both the context and the expectations for this analysis. In Section 3, we report the collection and the preprocessing strategy adopted for the Twitter data. We include in Section 4 a description of both methods and technologies we used for the Twitter data analysis together with an overview of the obtained results. Finally, in Section 5 we disclose both conclusions and future directions of the presented work.

*Corresponding author

2. THE URBAN-GEO BIG DATA PROJECT

The overall mission of the URBAN-GEO BIG DATA project is to contribute to the improvement of the exploitation of the new geospatial data, such as modern Earth observation (EO) data, data generated using mobile sensors, and VGI for a better understanding of a number of urban dynamics (http://urbangeobigdata.it). The project is funded by the Italian Ministry of Education within the PRIN programme 2017-2020. The project team is composed of four Italian universities that are Politecnico di Milano (project leader), Politecnico di Torino, Università degli Studi di Roma “La Sapienza”, and Università degli Studi di Padova. The academic partners are accompanied by two national research centres, namely the Institute for Electromagnetic Sensing of the Environment of the National Research Council (IREA-CNR) and The Italian Institute for Environmental Protection and Research (ISPRA).

The research work mainly focuses on two emerging issues affecting the urban environment, which have been identified as soil consumption and mobility management. These are also selected as project case studies. The Italian cities of Rome, Milan, Turin, Naples and Padua have been chosen as study areas for the project. Each research unit has contributed according to its expertise which includes remote sensing, geographic information systems (GIS), digital and Web mapping, transportation management, and software development. The expected outcomes of the research work are assessing the contribution of the new geospatial data and technologies to the analysis of the soil consumption and mobility phenomena in the urban context and deploying new data-driven strategies for urban management. The work we present in this paper tackles one of the tasks of the URBAN-GEO BIG DATA project. This is connected with the exploitation of the geo-crowdsourced data from the Twitter platform as complementary information to be considered in the analysis and management of both mobility patterns and citizens’ perception of mobility services. The final goal of this project task is to boost both the awareness and the comprehension of the urban mobility within the study areas by integrating multiple geospatial data sources. We report the analysis strategy and the outputs of this specific project task in the following.

3. TWITTER DATA COLLECTION

We performed the Twitter data collection by taking advantage of the Twitter streaming application programming interface (API) (https://developer.twitter.com). We designed a tweets collector using JavaScript and the NodeJS framework (https://nodejs.org) running on a server machine as shown in Figure 1. The collector relies on a MongoDB NoSQL database installation (https://mongodb.com) for storing the collected data streaming. The stored data includes the whole tweet text, tags and metadata such as coordinates, user ID, publication timestamp, etc. The collected data can be retrieved from the database in JSON format.

We ran two separate collections in order to produce two datasets. The first one includes a selective collection of tweets obtained by bounding the search on the neighbourhood of selected transportation terminals. This is performed by using the point and radius search option, enabled by the API, setting a 500 m radius from each selected transportation terminal, such as main train and metro stations, toll roads, etc. This provides a focused collection of content on transportation places to be used as test data for the development of the analysis procedure. The second dataset includes the whole Twitter data stream over the five cities of the project, collected using the bounding box search option by setting the search area extent equal to the minimum rectangle containing each city administrative boundary. We filtered this latter collection to obtain the subset of georeferenced tweets. This means the tweets that include the exact locations of the users retrieved from the GPS sensor of their devices. The georeferenced subset is intended to prevent us from processing the mobility-related information which cannot be pinned to a specific location in space. This allows for an assessment of the portion of information which can be actually integrated into an urban analysis framework based on geospatial data, as requested by the URBAN-GEO BIG DATA project.

Figure 1. Software architecture of the Twitter data collector

Moreover, we performed a preliminary exploration of the collected data by processing the JSON exports with Python. This includes operations such as global data counts and languages detection to better targeting the analysis procedure we explain in the next section. Collected data refers to the period May 2017 - May 2018. We report a summary of the collected data in Table 1.

Collection	N. of tweets (size in memory)	Language [%]
test dataset	91893 (35.5 MB)	IT ∼46% ENG ∼32%
whole dataset	1450813 (5 GB)	IT ∼58% ENG ∼16%
whole dataset georeferenced	287521 (114.5 MB)	IT ∼48% ENG ∼31%

Table 1. Collected data summary

4. TWITTER DATA FOR MOBILITY HUBS EXTRACTION AND PERCEPTION ANALYSIS

According to the outcomes of the collection procedure, and considering this as a preliminary test for assessing the benefits of adoption of the Twitter data within the URBAN-GEO BIG DATA project context, we consider for the analysis the whole tweets georeferenced in Italian language (∼138000 tweets). This corresponds to the 48% of the all available georeferenced information and approximately to the 10% of the whole collected data (see Table 1). The shares of tweets among the cities are 49% for Rome, 33% for Milan, 8.5% for Naples, 7.5% for Turin, and 2% for Padua. These are aligned with the population size of the cities. We use the test dataset, collected around transportation terminals only, as an independent information to be used in the a priori selection of keywords to extract mobility-related tweets from the whole georeferenced dataset in Italian language and map them. After this, we attempt to classify the extracted tweets in order to sense the users’ perception on mobility as well as to assess the actual amount of valuable information that can be used in the mobility management. We report the details on both procedures and outcomes in the following.

4.1 Methods

We performed the analysis by taking advantage of natural language processing (NLP), a well-known artificial intelligence technique that enables the investigation of free text, the analysis of its contents, and the extraction of information from it. In particular, we accomplished this task through the use of Natural Language ToolKit (NLTK), a Python library providing tools (e.g. tokenization, stemming, part of speech tagging, chunking, lemmatizing, Naive Bayes classifier, etc.) for statistical NLP.

First, we created a list of mobility-related keywords, including commonly used Italian nouns, adjectives, and verbs describing infrastructures, services, and actions connected to the mobility. We performed a manual check of the test dataset in order to extend and correct the arbitrarily selected keywords. Using the keywords list, we extracted mobility-related tweets from the whole dataset of tweets georeferenced in Italian language, after cleaning and organizing the data by means of tokenization, stemming, and stop words removal operations. Then, we mapped the extracted mobility-related tweets using QGIS to explore also their spatial distribution.

Lastly, we trained a Naive Bayes classifier (Schütze et al., 2008) by manually tagging a randomly selected subset (∼20%) of the extracted mobility-related tweets text using two classes, namely positive/neutral and negative/complain messages. The main purpose of this classification is to quantify the amount of information which can be used in the context of mobility management, with a specific focus on negative feedback on the mobility services and infrastructures. We implemented the classification procedure in Python by taking advantage of the NLTK - Naive Bayes classifier module. We report the results of the classification in Section 4.3.

4.2 Filter on mobility-related keywords

The extraction of mobility-related tweets by means of keywords filtering produced a subset of ∼2300 tweets unevenly distributed among the 5 cities. In decreasing order, the tweets shares are 42% in Milan, 40% in Rome, 8% in Naples, 7% in Turin, and 3% in Padua. This partially reflects the population size of the cities where just Milan (the second largest city in terms of population after Rome) shows higher activity if scaled to its population size. We include an overview map of the mobility-related tweets in Figure 2.

We computed raster heatmaps from the mobility-related tweets for each city by using the quadratic kernel density estimator with a 2000 m fixed bandwidth (Anderson, 2009) at a pixel resolution of 100 m. This enables an intuitive visualization of density hotspots in the spatial distribution of the mobility-related tweets within the cities, as shown in Figure 3.

According to the heatmaps, the mobility-related tweets cluster mainly on the main train stations such as the central stations in Milan, Padua and Naples, the Termini Station in Rome, and the Porta Nuova Station in Turin as well as on the city centres as in the case of Naples. Therefore, we identify as the main urban mobility hubs the above locations which nevertheless could be expected by considering both the nature and the context of the performed analysis.

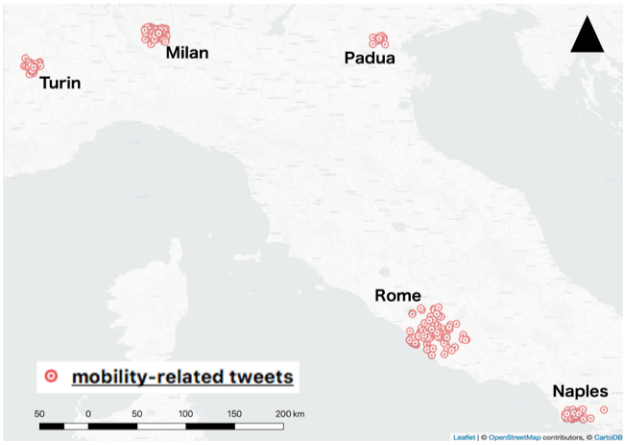

Figure 2. Map of mobility-related tweets extracted from the whole dataset of tweets georeferenced in Italian language

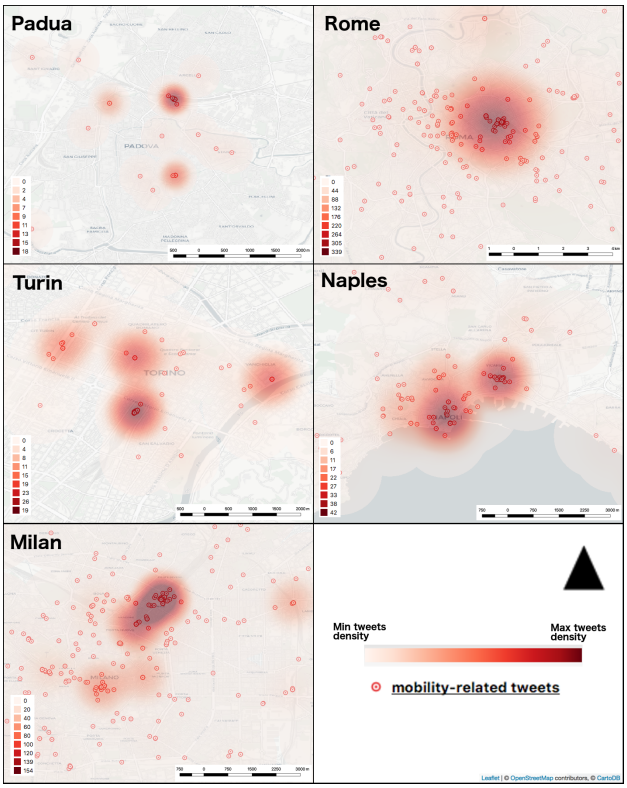

Figure 3. Heatmaps computed from mobility-related tweets for 5 city centres

4.3 Tweets classification

Through the classification of mobility-related tweet texts, we sought to retrieve an indication for the amount of information containing valuable user feedback on mobility services and infrastructures, as we mentioned in Section 4.1. We focused mainly on negative feedback which is key to the improvement of mobility systems (Taniguchi and Fujii, 2007). Out of the ∼2300 mobility-related tweets, we found the negative feedback or complains to be 7% (∼150), which follows proportionally the spatial distribution as well as the share among cities of the considered tweets dataset. This gives a measure of the amount of actual information which can be retrieved, classified and potentially reused within the con-

text of urban mobility studies. The numbers we reported do not consider for the accuracy of the classifier. We did not tackle this aspect within this preliminary test. Due to the small amount of information fitting our purpose, we are reconsidering the potential use as well as applications of such data within the URBAN-GEO BIG DATA project. We include considerations on this latter, together with conclusions on the most significant outcomes of this work as well as its future directions in the following section.

5. CONCLUSIONS AND FUTURE WORK

In this work, we presented a procedure to collect and analyze Twitter data focusing on the investigation of urban mobility patterns. The results for the presented case study indicate that the amount of mobility-related data, extracted with the proposed analysis framework, represents only a marginal portion of the whole dataset. On one hand, the spatial distribution of the mobility-related tweets turned out to be informative for the identification of mobility hubs even if these depict mainly locations that can be perhaps figured out by using a number of alternative data sources as well as common sense. On the other hand, the text analysis, as performed in the case study, did not produce significant information to justify its further consideration within the URBAN-GEO BIG DATA project. According to this, future work will redesign the analysis procedure to test additional application of Twitter data for mobility-related study in the urban environment. This will include e.g. the automatic detection of movements of people within the cities during weekdays, weekends, social events as well as the delineation of urban routes that tourists prefer.

ACKNOWLEDGEMENTS

This work was supported by the URBAN-GEO BIG DATA, a Project of National Interest (PRIN) funded by the Italian Ministry of Education, University and Research (MIUR) - id. 20159CNLW8.

REFERENCES

Anderson, T. K., 2009. Kernel density estimation and k-means clustering to profile road accident hotspots. *Accident Analysis & Prevention* 41(3), pp. 359–364.

Bao, J., Liu, P., Yu, H. and Xu, C., 2017. Incorporating twitter-based human activity information in spatial analysis of crashes in urban areas. *Accident Analysis & Prevention* 106, pp. 358–369.

Bothorel, C., Lathia, N., Picot-Clemente, R. and Noulas, A., 2018. Location recommendation with social media data. In: *Social Information Access*, Springer, pp. 624–653.

Frias-Martinez, V. and Frias-Martinez, E., 2014. Spectral clustering for sensing urban land use using twitter activity. *Engineering Applications of Artificial Intelligence* 35, pp. 237–245.

Graells-Garrido, E., Caro, D., Miranda, O., Schifanella, R. and Peredo, O. F., 2018. The www (and an h) of mobile application usage in the city: The what, where, when, and how. In: *Companion of the The Web Conference 2018 on The Web Conference 2018*, International World Wide Web Conferences Steering Committee, pp. 1221–1229.

Jussila, J. J., Kärkkäinen, H. and Aramo-Immonen, H., 2014. Social media utilization in business-to-business relationships of technology industry firms. *Computers in Human Behavior* 30, pp. 606–613.

Kovacs-Gyori, A., Ristea, A., Havas, C., Resch, B. and Cabrera-Barona, P., 2018. #london2012: Towards citizen-contributed urban planning through sentiment analysis of twitter data. *Urban Planning* 3(1), pp. 75–100.

Mergel, I. and Bretschneider, S. I., 2013. A three-stage adoption process for social media use in government. *Public Administration Review* 73(3), pp. 390–400.

Schütze, H., Manning, C. D. and Raghavan, P., 2008. *Introduction to information retrieval*. Vol. 39, Cambridge University Press.

Taniguchi, A. and Fujii, S., 2007. Promoting public transport using marketing techniques in mobility management and verifying their quantitative effects. *Transportation* 34(1), pp. 37.