# Analyses of Geo-Referenced Twitter Data for Understanding Spatial Distribution and Content Classification

Nikola Dzakovic, Nikola Dinkic, Jugoslav Jokovic, Leonid Stoimenov, Dejan Rancic
University of Niš, Faculty of Electronic
Engineering, Aleksandra Medvedeva 14
18000 Niš, Serbia

***ABSTRACT:*** *We intend to study the spatial patterns of the urban regions and design it for open spaces. To do so we have employed the user-generated twitter data that could support and improve the understanding. The features such as temporal and spatial distribution, content classification, language determination and sentiment analyses are studied using the data generated by Twitter social network. We have sued the Twitter search engine application to classify and process the geo-referenced tweets collected.*

**Keywords:** Twitter Data, Geo Reference Data, Text Analysis, Text Mining

**Received:** 24 October 2020, Revised 27 January 2021, Accepted 27 February 2021

**DOI:** 10.6025/jcl/2021/12/2/35-42



## 1. Introduction

Open public spaces are fundamental element of vibrant, inclusive and smart cities. Represented as active, social, attractive and secure, open public spaces play key role in revitalization of community, it promotes its sense of identity, culture and economic growth. Having methods to properly determine attractiveness of open public spaces has always been a challenge for urbanists, but also an important tool in fields like urban planning, transport, marketing, business, migration and tourism.

The use of Twitter data is very interesting to make analysis of how people use urban open spaces and what is the geographical pattern of their communications. Since the Twitter is a massive platform for online communication within extremely diverse social groups, with data generated by users of this network it is possible to research the spatiotemporal dynamics of location and different aspects of users' behaviour. Recently, Twitter has gained significant popularity among the social network services. Twitter contains an enormous number of text posts. Lots of users often use Twitter to express feelings or opinions about a variety of subjects. All this information can be obtained from micro-blogging services, as their users post their opinions on many aspects of their life regularly.

Therefore, it is possible to collect text posts of users from different social and interest groups. In this context, Twitter presents interesting challenges. Its short texts (tweets), widespread use of non-standard grammar, spelling and punctuation, as well as slang, abbreviations and neologisms, etc. make syntactic and semantic analysis difficult. Analyzing this kind of content can lead to useful information for fields such as personalized marketing or social profiling. However, analyzing Twitter data comes with its own bag of difficulties. Tweets are small in length, thus ambiguous. The informal style of writing, a distinct usage of

orthography, acronymization and a different set of elements like hashtags, user mentions demand a different approach to solve this problem.

The goal of this paper is to illustrate possibilities of methodologies based on user-generated data that could support and improve the understanding of spatial patterns for urban planning and design of open spaces in urban areas. The case study considered in this paper is a network of open public spaces in Barcelona, representing one of the most attractive and important urban ambient. The method that was used in analysis is the method of mapping users on the social maps (via social networks) based on a new software application Twitter search engine [1]. This Web application enables the collection, storage, processing and analysis of data from the social network Twitter. It is the micro-blogging platform that provides a rich collection of real-time commentaries on almost every aspect of life. Data collection is based on the Twitter REST API [2] that allows the collection of tweets in the space defined with geo-referenced points and the given radius. This API provides a wide range of information related to their own tweets and users who post them. In addition to basic information such as text, time, number of retweets, the number of likes and information about the application from which it was posted/sent, as well as the geographical location where the tweet was shared present the basis for the analysis and processing of geospatial data.

## 2. Related Work

Traditionally, the attractiveness of places has been calculated from survey data, geographical features, and population distribution. For instance, the attractiveness measure of points of interest proposed by Huang et al. [3] considers static factors (e.g., the size of commercial places, the distance to their customers' homes) and dynamic factors (e.g., restaurants are more attractive at mealtimes). Geoinformation is now used on a daily basis - photos can be stored with location information, users on social networks publish their location or require the shortest path to the desired object in the city. Geographic information attached tweets are used primarily as a mechanism for filtering [4]. Geotagging is the case when Twitter users make available their position, so others can see the exact place where the tweet was sent. Information can be analyzed based on location and profile generated by the user.

In this paper [5] Andre S. introduces so called M-Attract, novel method which goal is to assess how much places of interest are attractive, based on trajectory episodes that occur in their surroundings. This section describes the places, subregions and region of interest and the trajectory episodes considered by the method. Christoph Breser in his article [6] discussed about technical solutions for representing archival sources in urban areas. He strive to realise the interconnectedness of sources, its beholders and the concerning entity through the location where the information was recorded the first time. There to, they try to identify problems in the analogue world mainly dealing with the classification of archiving, semiotic systems, descriptions and assignments. They use existing mobile technologies and software applications from different application fields and test their suitability or our concern. Comparing and transferring analogue methods to the digital world is a real challenge they try to accept when it comes to solving identified problems that arise in the context of modes of practice in archives and we representations.

In this paper we describe how can social networks, such as Twitter, be used to quantify attractiveness of urban places. One of he ways is through semantic analyze [7, 8] of data, for locations of interest, that is generated through social net work. One of the most common ways for semantic analyze is described in [9]. V. Pandey and C. V. Krishnakumar Iyer classify messages through two classifiers. First classifiers is used to determine if the message is neutral or polar (positive or negative). Then if after first classifiers message appears to be polar, it goes through second classifiers so it can be decided if it is a positive or a negative message. In their paper they also apply sentimental analyzis to messages gathered through Twitter social network, and the approach for executing it is very similar to one describe in this paper.

## 3. Data Processing and Analysis

The analysis of geospatial data requires data to be in the specified format so that geospatial queries can be executed. However, since all information obtained from the Twitter REST API is in JSON format, before any analysis it is necessary to perform transformation of geo-information to specific format. This process of transformation of the original data to geospatial data types represents the pre-processing, and this is the first step in this analysis.

In order to illustrate possibilities of TSE application, this section examines data coming from the social network platform

of Twitter to provide a visual and scientific exploration of the resulting spatial pattern, specifically of locals, visitors and tourists which have used urban open spaces in Barcelona.

Figure 1 presents map with all georeferenced tweets posted in city of Barcelona in period of 13 – 20 January, 2017, while cumulative statistic data based on collected tweet are summarized in Table 1.

| **Type of analysis** | **All** | **Geo** |
|---|---|---|
| Number of tweets | 120600 | 16200 |
| Number of users | 10412 | 3955 |
| Number of retweets | 20208 | 720 |
| Number of likes | 51989 | 2894 |
| Number of applications | 45 | 40 |
| Number of languages | 120 | 91 |

Table 1. Cumulative data in period 13-20. January

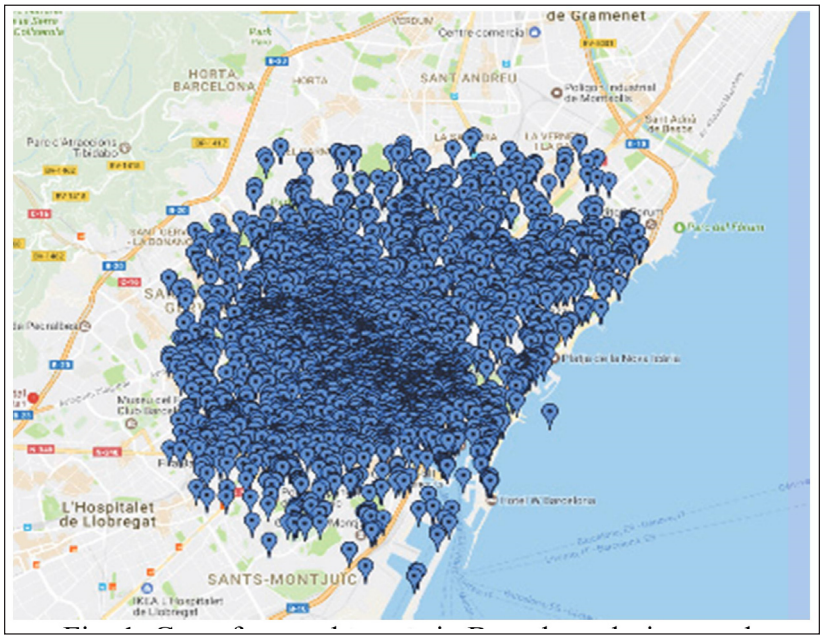


Figure 1. Georferenced tweets in Barcelona during week 13-20. January

Considering the Tweets according to the day of the week, it is possible to see if there are some sensible variations in the use. Figure 2 and Table 2, representing statistics in terms of number of tweets by days, show that users generally were most active on Thursday, posting 18115 (15 %) tweets, an d the least active on Friday, when it was posted 15681 (13%) tweets.

Splitting the Tweets on the time steps which cover the 24 hours per day, it is possible to better understand and some aspects of the city life. The radial diagram in the Figure 3 show how the number of Tweets increases during the evening, reaching a peak at 9.00-10.0 0 pm, with 8852 Tweets, while the users were the least active between 4.00-5.00 am , posting only 629 tweets.

To generate maps shown in the Figure 4, all the geo-referenced Tweets between 1 3-20 January have been used. It is particularly interesting to observe some different behaviors. For example, between noon and 3 p.m., the activity is constant and well

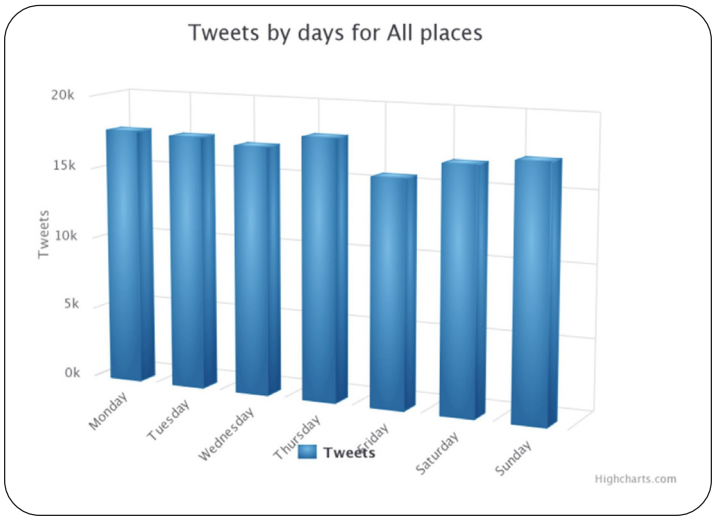


Figure 2. Distribution of all tweets in BCN by days

| ***Tweets by days*** | ***All*** | ***Geo*** |
|---|---|---|
| *Monday* | 17790 | 2157 |
| *Tuesday* | 175 66 | 2096 |
| *Wednesday* | 171 90 | 2358 |
| *Thursday* | 181 15 | 2962 |
| *Friday* | 156 81 | 1863 |
| *Saturday* | 168 92 | 2403 |
| *Sunday* | 173 66 | 2361 |
| ***TOTAL*** | 120600 | 16200 |

Table 2. Distribution of tweets by days

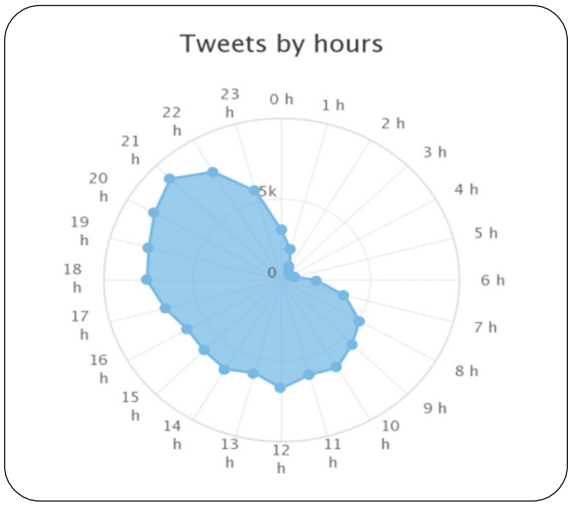


Figure 3. Distribution of all tweets by hours

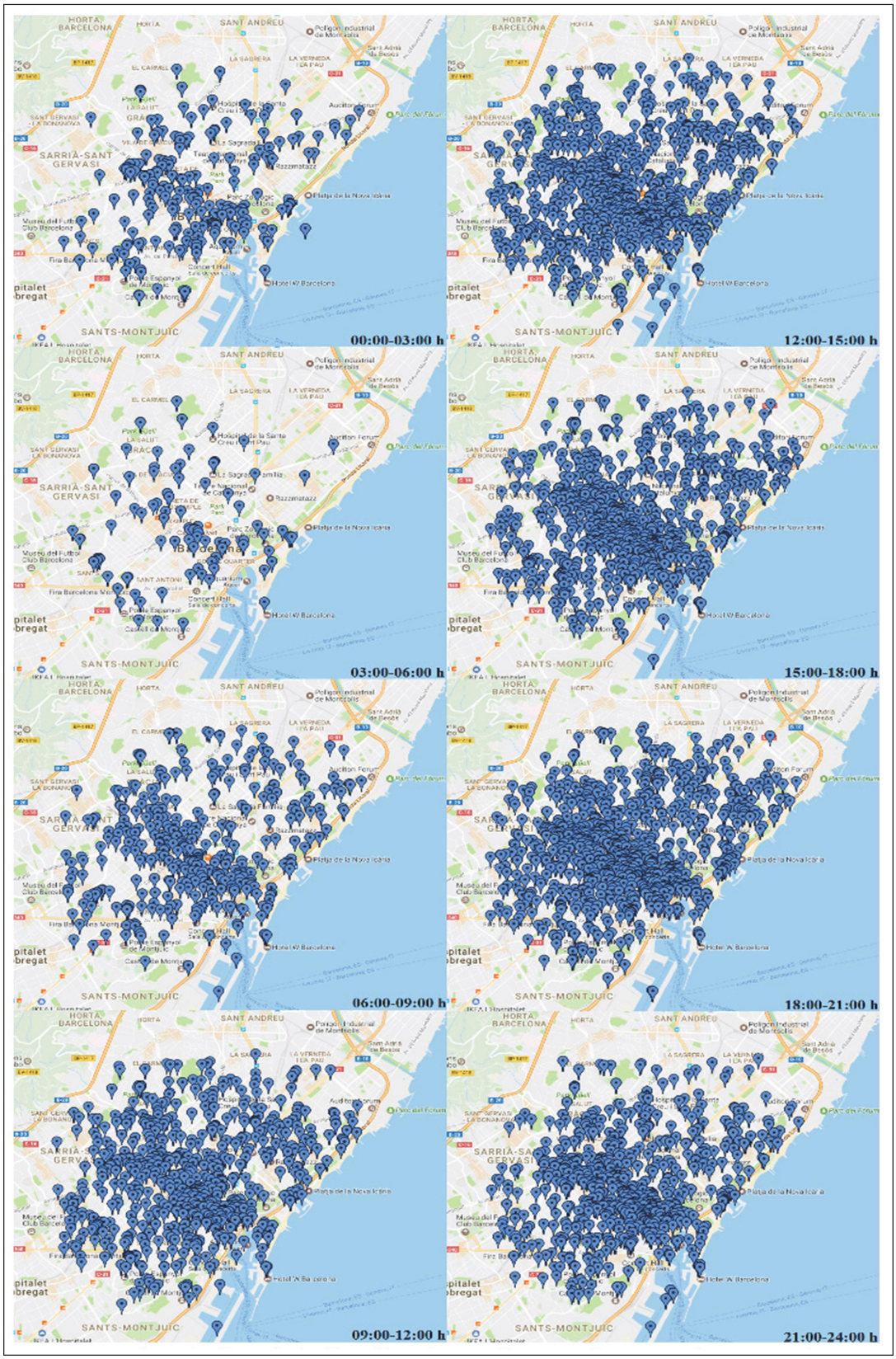


Figure 4. Spatial distribution of georeferenced tweets by time interval in day

over the city. Otherwise, during the nighttime there are more differences: the city center and some specific axes remain active, while other areas are practically abandoned. The visualization based on map allows the viewer to explore data and to choose the rang of hours to be visualized or to view the sequence of tweets on an hour by hour basis.

Analysis of all tweets using TSE application detected 120 different languages [10]: Spanish, English, Catalan, Portuguese, French, Italian, Danish, Arabic, Russian, of which the most common are Spanish with 44%, English with 22.7% and Catalan with 16.3% (Figure 5). The result is very particular, because it highlights the use of tourists from local people. Using the total number of geo -referenced Tweets collected in the selected period, the maps could be generated by differentiating users by the language chosen for their accounts. This analysis would provide more information on the origin of each user and would allow greater detail in the diversification between locals and tourists.

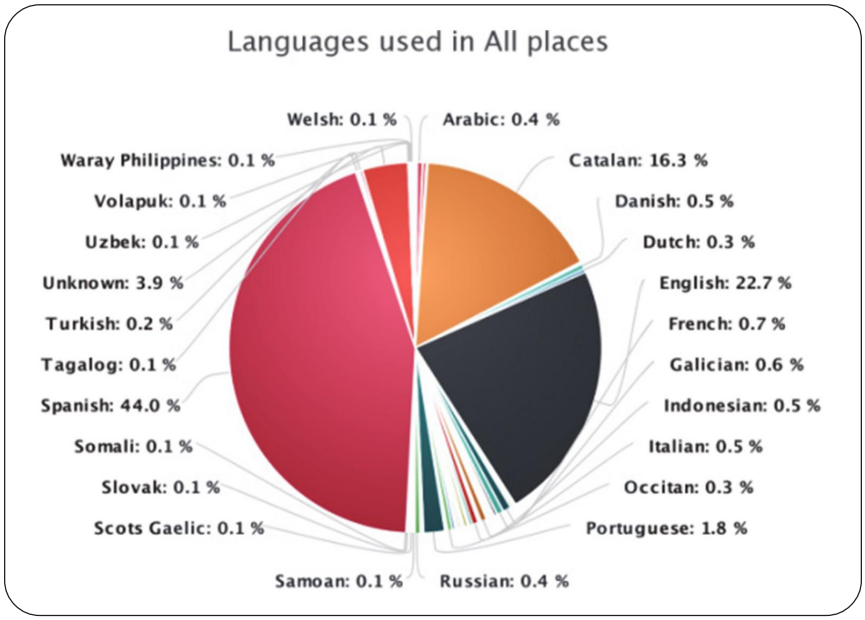


Figure 5. Distribution of tweets by language

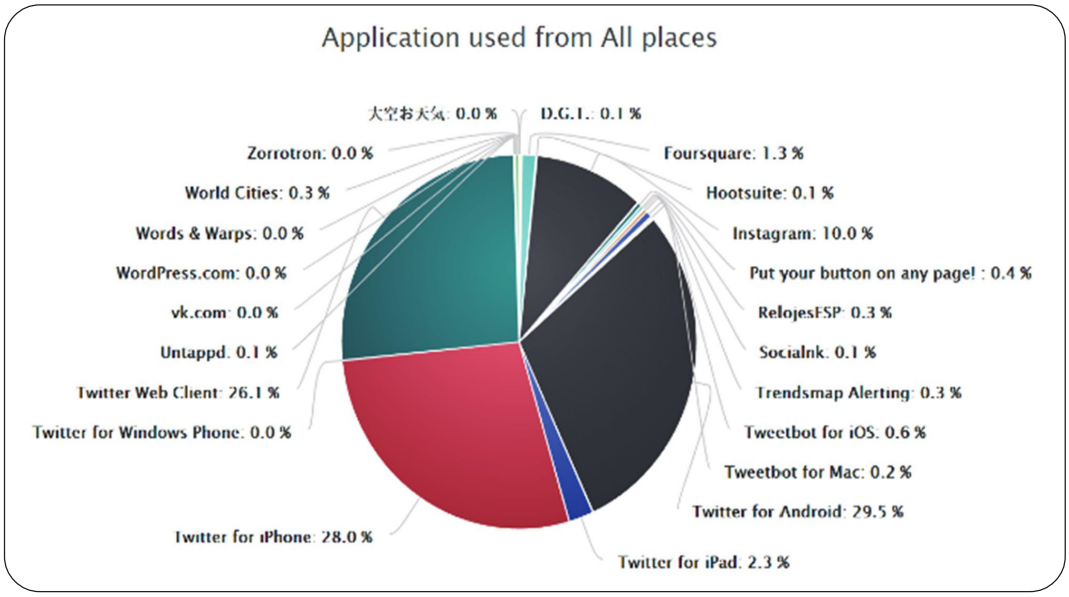


Figure 6. Distribution of tweets by application

Based on data gathered on Twitter for Barcelona, TSE application detected that the users of this social network posted content for location of interest from twenty two different applications. Figure 6 presents that most used applications are "Twitter for Android", "Twitter for iPhone", "Twitter Web Client" and "Instagram" and they combined make 93.6% of all tweets posted for this location, while the rest 18 applications make only 6.4% of total data.

In order to illustrate possibilities of TSE application regarding spatial filtering, Figure 7 presents georeferenced tweets in Enric Granados Street, posted by Instagram. Enric Granados Street is not highly touristic, but hosts several leisure activities that are mostly for neighborhood residents and other local people.

The street represent a pedestrian oriented public space, where open-air facilities and seating are populating the urban space. Enric Granados has also a larger open space, like a Mediterranean square, corresponding to the crossing with Aragò Street.

Figure 7. Spatial distribution of georeferenced tweets in Enric Granados Street – Instagram

| ***Type of analysis*** | ***Georeferenced*** | ***Instagram*** |
|---|---|---|
| *Number of tweets* | 1 70 | 554 |
| *Number of users* | 219 | 184 |
| *Number of retweets* | 0 | 40 |
| *Number of likes* | 183 | 148 |

Table 2. Cumulative data for enric granados street in period 13-20. January

Finally, in order to illustration the role of ICT tools in the design and use of urban open spaces, the data generated by Twitter social network in considered case studies in Barcelona are analyzed regarding sentiment analysis of tweets. Based on sentiment analysis of content posted in Barcelona, tweets were divided into three groups, tweets which contain positive wo ds, negative words and which contain both positive and negative. Since on Twitter it is a very popular to use hashtags (#), tweet content can be divided into two groups, text and hashtags. Sentiment analysis can be executed only on tweets that contain text. Application detected 12814 text tweets to belong to one of three groups (positive, negative, and complex). The results of semantic analysis of tweets that contain either positive or negative words are shown in Table 3.

## 4. Conclusion

The presented results of analysis illustrate possibilities of Twitter search engine application for analysis of urban open

| | ***Only text*** | ***only #*** | ***Both*** | ***%*** |
|---|---|---|---|---|
| *POSITIVE* | 6798 | 586 | 7292 | 59.20% |
| *N EGATIVE* | 4184 | 213 | 4338 | 28.32% |
| COMPLEX | 1141 | 15 | 1184 | 12.48% |
| *TOTAL* | 12123 | 814 | 12 14 | 100 |

Table 3. Semantic analysis of tweets

spaces using georeferenced data. In general, the social network Twitter is convenient for this type of research, since the platform thro ugh its REST API provides support for data analysis, primarily based on a large a mount of public information that is crucial to any successful analysis. Sentiment analysis also shows that attraction sits of this region leave a positive impression on tourists who come to visit them. The data generated by users of Twitter social network in considered case studies in Barcelona are analysed regarding temporal distribution, content classification, language determination and sentiment analysis of tweets.

Visualization that TSE application offers can very effective for the analysis of land use and, in general, for decision- and policy-making processes in the (re) design of public spaces. In particular, it can quickly express the concentration and sprawling of people over te urban area. The clusters of Tweets can help in identifying the density of activities within the city. These an as well be used to quantify popularity of locations of interest and public spaces in general, as well as to determine correlations between locations.

**References**

[1] Dzakovic, N., Dinkic, N., Jokovic, J., Stoimenov, L. (2016). Web Application for Mining, Storing, Processing and Geo-analysis Data from Twitter Social Network, *YU INFO*, Kopaonik, Serbia, (March).

[2] https://dev.Twitter.com/rest/public, accessed on March 15th, 2017.

[3] Huang, L., Li, Q., Yue, Y. (2010). Activity Identification from GPS Trajectories using Spatial Temporal POIs' Attractiveness, *ACM SIGSPATIAL Workshop on Location Based Social *Networks*, San Jose, USA, p. 27-30.

[4] Vukmirovic, M., Vanista Lazarevic, E. (2015). *Competitiveness Express though Digital Data*, in E. Vanista Lazarevic, M. Vukmirovic A. Krstic-Funurndzic, and A. Djukic (Eds.), *Keeping up with Technologies to Improve Places*, Newcastle upon Tyne: Cambridge cholras Publishing.

[5] Furtado, A. S., Fileto, R., Renso, C. (2013). Assessing the Attractiveness of Places with Movement Data, *Journal of Information and Data Management*, 4 (2) 124-133, (June).

[6] Breser, C., Zedlacher, S., Winkler, R. a. (2016). The Principle of Geotagging. Cross-linking Archival Sources with People and the City Through Digital Urban Places, *ICiTy Conference*, Valletta, Malta, 18-19 (April).

[7] Hu, M., Liu, B. (2004). Mining Opinion Features in Customer Reviews, *American Association for Artificial Intelligence*.

[8] Liu, B. (2012). *Sentiment Analysis an d Opinion Mining*.

[9] Pandey, V., Krishnakumar Iyer, C. V. (2009). Sentiment Analysis of Microblog, *CS* 229: *Machine Learning Final Projects*, 2009 - cs229.stanford.edu

[10] *Language Detection API* -https://detectlanguage.com/, last visited 15.03.2017.