Using Sentiment Analysis of Twitter Data for Determining Popularity of City Locations

Nikola Dinkić[1(✉)], Nikola Džaković[1], Jugoslav Joković[1], Leonid Stoimenov[1], and Aleksandra Đukić[2]

[1] Faculty of Electronic Engineering, University of Niš, Niš, Serbia
{dinkicnikola, ndzakovic}@elfak.rs, {Jugoslav.Jokovic, Leonid.Stoimenov}@elfak.ni.ac.rs
[2] Faculty of Architecture, University of Belgrade, Belgrade, Serbia
adjukic@rcub.bg.ac.rs

Abstract. The paper considers mining and analyzing data generated by Twitter social network, regarding content classification, language determination and sentiment analysis of tweets. Analyzes are based on geospatial tweets collected in timespan of four months within region Vračar in Belgrade, Serbia. All of collected data is first being preprocessed, filtered and classified by given criteria, by using "Twitter search engine" (TSE) application, that has been upgraded in order to detect tweet language and execute sentiment analysis of the tweets written in English. This type of analysis can be used for determining popularity of city locations of interest and public spaces in general.

Keywords: Natural language processing · Sentiment analysis and opinion mining · Geospatial data · Twitter social network

1 Introduction

The synergy between technology and built environment generate urban culture with digital streams. On the other hand, cities and their open public spaces are reflections of users changing needs. Furthermore, the future of urban space depends on the role of information and communication technologies (ICT) and importance of their networks should be reconsidered since they have become indispensable ingredients of urban life [1]. ICT provides the overlapping of real and virtual spaces and allows creative participation of users.

With the explosive growth of social media (e.g., reviews, forum discussions, blogs, micro-blogs, Twitter, comments, and postings in social network sites) on the Web, individuals and organizations are increasingly using the content in these media for decision making. Many big corporations have also built their own in-house decision making solutions, e.g., Microsoft, Google, Hewlett-Packard, SAP, and SAS. Although linguistics and natural language processing have a long history, research about people opinions and sentiments in social media is actual in last ten years.

This paper will present and analyze the connections which are established and intensified between users and open public spaces via Twitter. Twitter is one the most popular data sources for research [2] because of its open network allowing access to

G. Stojanov and A. Kulakov (eds.), *ICT Innovations 2016*,
Advances in Intelligent Systems and Computing 665,
https://doi.org/10.1007/978-3-319-68855-8_15

information published through the platform. Twitter is a novel microblogging service launched in 2006 with more than 310 million monthly active users. On Twitter, every user can publish short messages with up to 140 characters, so-called "tweets", which are visible on a public message board of the website or through third-party applications. The public timeline conveying the tweets of all users worldwide is an extensive real-time information stream of more than one million messages per hour. The original idea behind microblogging was to provide personal status updates. However, these days, postings cover every imaginable topic, ranging from political news to product information in a variety of formats, e.g., short sentences, links to websites, and direct messages to other users.

The method that was used in analysis is the method of mapping users on the social maps (via social networks). It was based on a software application TSE [3]. The aim was tracking and measuring the intensity of users in the monitored territory, testing the latest behavioral patterns of them as well as tracing the "positive routes". The obtained results have enabled the determination of the image of the open public spaces perceived by the users, as well as the potential of the analyzed area for the formation of transverse and longitudinal pedestrian flows that could help improving networking of open public spaces.

2 Related Work

Natural language processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages. NLP is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things. NLP researchers aim to gather knowledge on how human beings understand and use language so that appropriate tools and techniques can be developed to make computer systems understand and manipulate natural languages to perform desired tasks.

This paper is focused on sentiment analysis, also called opinion mining. It is the field of NLP that analyzes people opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. It represents a large problem space. While in industry, the term sentiment analysis is more commonly used, but in academia, both sentiment analysis and opinion mining are frequently employed. They basically represent the same field of study.

Sentiment analysis and opinion mining mainly focuses on opinions, which express or imply positive or negative sentiments. In paper [4], Hu and Liu proposed a lexicon-based algorithm for aspect level sentiment classification, but the method can determine the sentiment orientation of a sentence as well. It was based on a sentiment lexicon generated using a bootstrapping strategy with some given positive and negative sentiment word seeds and the synonyms and antonyms relations in WordNet.

The sentiment orientation of a sentence was determined by summing up the orientation scores of all sentiment words in the sentence. A positive word was given the sentiment score of +1 and a negative word was given the sentiment score of −1. In [5]

the relationships between the NFL betting line and public opinions in blogs and Twitter were studied. In [6] Twitter sentiment was linked with public opinion polls. In [7] Twitter sentiment was also applied to predict election results. In [8] Twitter data, movie reviews and blogs were used to predict box-office revenues for movies. In [9] Twitter moods were used to predict the stock market. In [10] they tracked opinions about movies on Twitter and predicted box-office revenues with very accurate results. They simply used their opinion parser system to analyze positive and negative opinions about each movie with no additional algorithms.

In general, sentiment analysis has been investigated mainly at three levels:

- **Document level:** The task at this level is to classify whether a whole opinion document expresses a positive or negative sentiment [11]. For example, given a product review, the system determines whether the review expresses an overall positive or negative opinion about the product. This task is commonly known as document-level sentiment classification. This level of analysis assumes that each document expresses opinions on a single entity (e.g., a single product). Thus, it is not applicable to documents which evaluate or compare multiple entities.
- **Sentence level:** The task at this level goes to the sentences and determines whether each sentence expressed a positive, negative, or neutral opinion. Neutral usually means no opinion. This level of analysis is closely related to subjectivity classification, which distinguishes sentences (called objective sentences) that express factual information from sentences (called subjective sentences) that express subjective views and opinions.
- **Entity and Aspect level:** Both the document level and the sentence level analyzes do not discover what exactly people liked and did not like. Instead of looking at language constructs (documents, paragraphs, sentences, clauses or phrases), aspect level directly looks at the opinion itself. It is based on the idea that an opinion consists of a sentiment (positive or negative) and a target (of opinion). An opinion without its target being identified is of limited use. Realizing the importance of opinion targets also helps us understand the sentiment analysis problem better. The goal of this level of analysis is to discover sentiments on entities and/or their aspects.

3 Classification of Geospatial Data

Twitter search engine [3] application (TSE) allows gathering, mining and storing of geospatial data produced on Twitter social network. This paper describes new features of application TSE and its ability to process and analyze data from this social network. Also in order to illustrate functionality of this application, this paper shows results of analysis of data collected from tweets for Vračar region in Belgrade, Serbia.

In addition to the collection and storage of data, TSE offers visualization and analysis, but also it can execute complex queries over stored data. This queries use special geospatial functions that are built within MySQL database. These functions represent correlation between two objects that are defined by geospatial points. This application also offers users to draw polygons on Google map, in order to define boundaries of their

analysis. Note, that all polygons must be within regions for which data is collected. Polygon drawing is done by using Google Maps JavaScript API. Web application TSE also has the ability to detect the language and perform sentiment analysis.

The most important indicators of sentiments are sentiment words, also called opinion words. These words are commonly used to express positive or negative sentiments. For example, good, wonderful, and amazing are positive sentiment words, and bad, poor, and terrible are negative sentiment words. Sentiment words and phrases are instrumental to sentiment analysis for obvious reasons. A list of such words and phrases is called a sentiment lexicon (or opinion lexicon). Although sentiment words and phrases are important for sentiment analysis, only using them is far from sufficient. The problem is much more complex. In other words, we can say that sentiment lexicon is necessary but not sufficient for sentiment analysis of complex texts. Since, tweets are generally short and informal, and use many Internet slangs and emoticons, they are easier to analyze due to the length limit because the authors are usually straight forward, and immediately get right to the point. Thus, it is often easier to achieve high sentiment analysis accuracy. Reviews are also easier because they are highly focused with little irrelevant information. Because of that for sentiment analysis, this paper uses lexicon-based algorithm to determine the sentiment orientation of a sentence.

Researchers have proposed many approaches to compile sentiment words. Three main approaches are: manual approach, dictionary-based approach, and corpus-based approach. [4] used a dictionary to compile sentiment words. This is an obvious approach because most dictionaries (e.g., WordNet) list synonyms and antonyms for each word. Thus, a simple technique in this approach is to use a few seed sentiment words to bootstrap based on the synonym and antonym structure of a dictionary. Specifically, this method works as follows: A small set of sentiment words (seeds) with known positive or negative orientations is first collected manually, which is very easy. The algorithm then grows this set by searching in the WordNet or another online dictionary for their synonyms and antonyms. The newly found words are added to the seed list. The next iteration begins. The iterative process ends when no more new words can be found. Their dictionary was compiled over many years starting from their first paper. Original dictionary consists of 4783 negative and 2006 positive words, but to the original dictionary were added emoticons, which are now very popular, and can also show if it is a positive or negative opinion. Tweets can be classified as a positive or negative depending on which group of words they contain. This gives similar results as simply counting positive and negative words, since Twitter messages are so short (about 11 words). Since the area of Serbia belongs to the world top by multilingualism, it is necessary to detect only the tweets that are in English. TSE uses web service "Language detection API" [12] for language detection. Language detection API has the ability to detect 160 different languages and offers 5000 requests for free on daily basis.

4 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, tweets collected over a period of four months (December 2015–March 2016) for region Vračar in Belgrade, were analyzed and results of this analysis are shown in this paper. This space is defined by the corresponding polygon on the map, as shown in Fig. 4. The execution of geospatial queries for the given polygon was obtained the cumulative data presented in Table 1.

Table 1. Cumulative figures for region Vračar

Type of analysis	Value
Number of tweets	3002
Number of users	603
Number of followers	1079721
Number of friends	421780
Number of retweets	35
Number of likes	490
Number of applications	8

In addition to the cumulative analysis, TSE allows filtering data based on the analysis in specified time intervals, in other words this means classification by months of the year, days of the week or otherwise defined time intervals. Figure 1 shows the distribution of tweets by month, which are shared in the studied area. It can be concluded that the Twitter users in the previous four-month period were the most active during the March 2016.

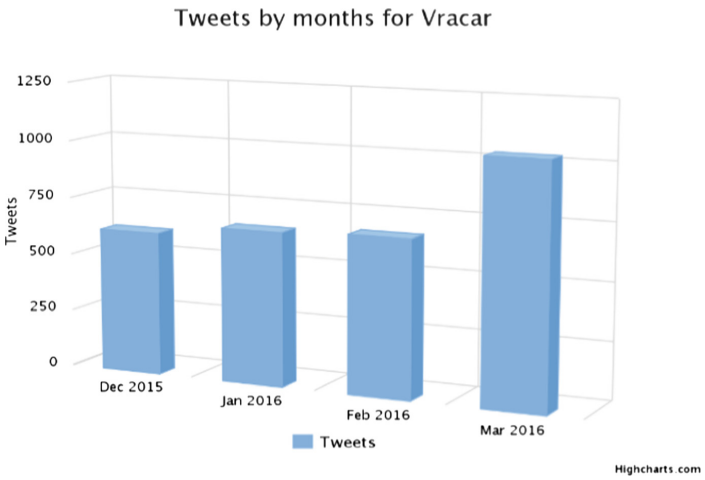

Fig. 1. Tweet count by months

The results of data classification by days of the week are shown in Fig. 2. Based on them, we can conclude that the users were most active on Saturdays.

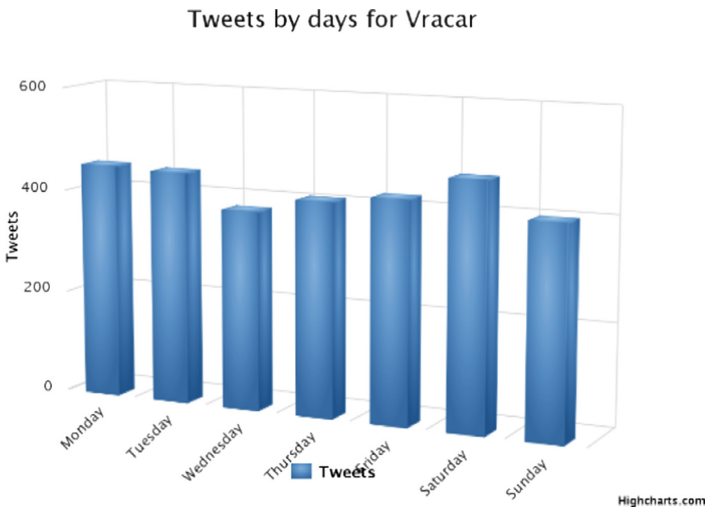

Fig. 2. Classification of tweets by days of the week

In addition to analysis of data for a given time, the type of content of tweet itself could classify the collected tweets. Figure 3 represents the percentage representation of content of analyzed tweets.

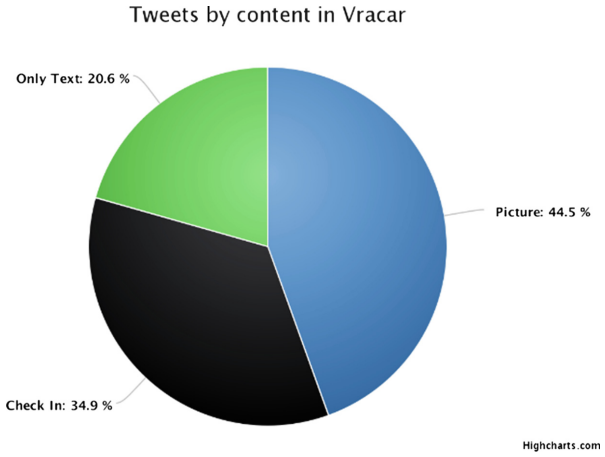

Fig. 3. Classification of tweets by content

Based on the distribution of tweets per application used to post it, it can be concluded, which content type has a tweet, without the need for a deeper analysis. For example, all tweets, which are sent by Instagram, must contain a picture. From the analysis results, it can be concluded that each Twitter user from community of Vračar on average posted 4.97 tweets, the users were most active during March, the largest number of tweets has been posted on Saturday and the most popular application is Instagram with 44.5%, following Foursquare with 34.9%, which indicates high attractiveness of the area.

The geospatial tweets classified by content are shown in Fig. 4. Tweets are displayed on the map with markers in different colors depending on the content of the tweet, picture (red), check in (green) and only text (blue).

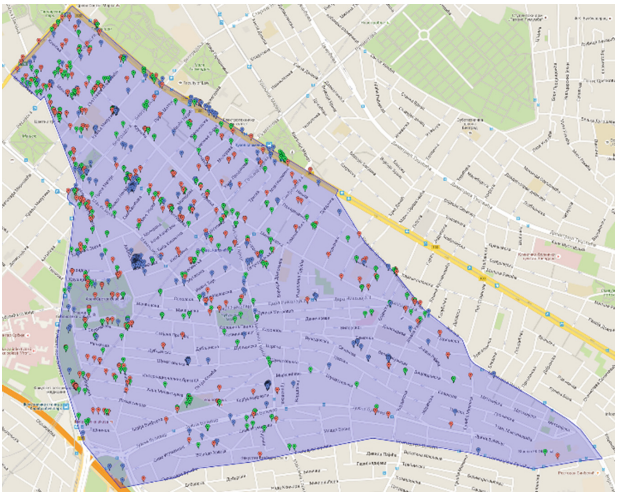

Fig. 4. Geospatial tweets for region Vračar classified by content

Analysis of all tweets detected 12 different languages: Serbian, English, German, French, Japanese, Thai, Russian, Turkish, Scottish, Bulgarian, Spanish and Portuguese, of which the most common are English with 67.4% and Serbian with 30.6%.

Based on sentiment analysis of content posted in region Vračar, tweets were divided into three groups, tweets which contain positive words, 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. The results of sentiment analysis of tweets that contain either positive or negative words are shown in Table 2.

Table 2. Sentiment analysis of tweets

	Hashtag(#)	Text	Both
Positive	55	155	200
Negative	9	43	52
Complex	5	10	14
$\sum$	69	208	266

Sentiment analysis can be executed only on tweets that contain text (including hashtags). Application detected 266 text tweets to belong to one of three groups (positive, negative, and complex) and this distribution chart is shown in Fig. 5. From 266 classified text tweets, 75.2% of them carried a positive message. The map of the tweets, which are classified as positive is shown in the Fig. 6.

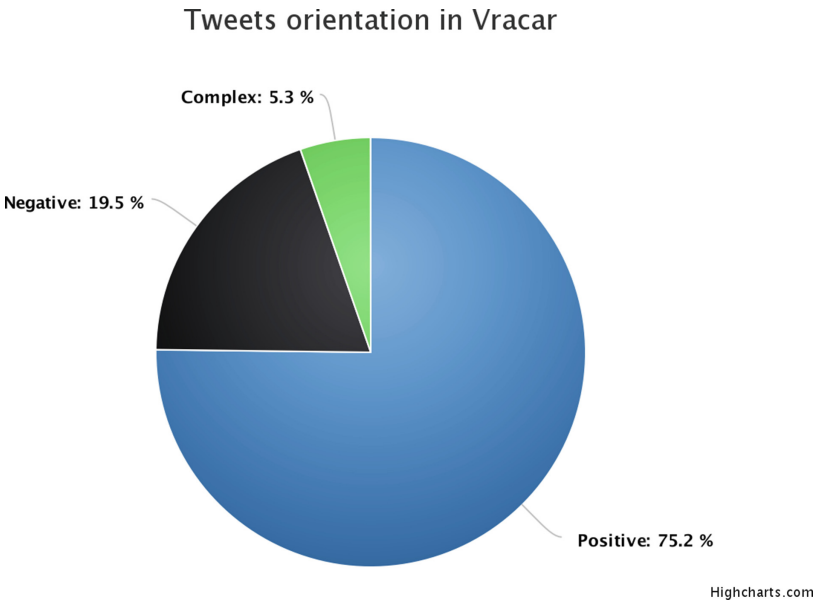

Fig. 5. Tweets orientation

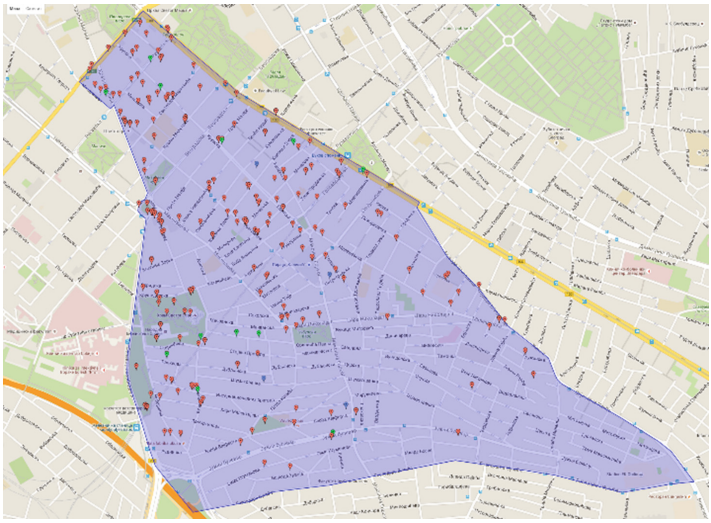

Fig. 6. Distribution of positive tweets

5 Conclusion

Generally, Twitter social network turned out to be a great basis for analysis of public space popularity. Its API provides a lot of publicly available information about tweets, but also about the users, which is the most important thing for every successful research. New feature of TSE application for language detection confirms the fact that Serbia belongs to the world top by multilingualism, which indicates that Vračar is very popular for foreign tourists. Sentiment analysis also shows that attraction sites of this region leave a positive impression on tourists who come to visit them. All analyses shown in this paper represent only a portion of possibilities that TSE application can offer and all of them can be used for creating better urban plans, in terms of (re)design of public spaces. These analyses can be used to quantify popularity of locations of interest and public spaces in general, as well as to determine correlations between locations.

References

1. Pigg, K.E., Crank, L.D.: Building community social capital the potential and promise of information and communications technologies. J. Commun. Inf. **1**(1), 58–73 (2004)
2. dos Santos, A.D.P., Wives, L.K., Alvares, L.O.: Location-Based Events Detection on Micro-Blogs (2012)
3. Nikola, D., Dinkić, N., Joković, J., Stoimenov, L.: Web application for mining, storing, processing and geo-analysis data from Twitter social network, YU INFO, Kopaonik, Srbija, March 2016
4. Hu, M., Liu, B.: Mining opinion features in customer reviews. In: American Association for Artificial Intelligence (2004)
5. Hong, Y., Skiena, S.: The wisdom of bookies? Sentiment analysis versus the NFL point spread. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (2010)
6. O'Connor, B.: From tweets to polls: linking text sentiment to public opinion time series. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (2010)
7. Tumasjan, A.: Predicting elections with Twitter: what 140 characters reveal about political sentiment. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (2010)
8. Asur, S., Huberman, B.A.: Predicting the future with social media. In: 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) (2010)
9. Bollena, J., Mao, H., Zeng, X.-J.: Twitter mood predicts the stock market. J. Comput. Sci. **2**, 1–8 (2011)
10. Liu, B.: Sentiment Analysis and Opinion Mining (2012)
11. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: Sentiment classification using machine learning techniques. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia (2002)
12. Language detection API. https://detectlanguage.com/. Accessed 15 Apr 2016