Urban Data Mining with Natural Language Processing:
Social Media as Complementary Tool for Urban Decision Making

Nai Chun Chen [1], Yan Zhang [2], Marrisa Stephens [3], Takehiko Nagakura [4], Kent Larson [5]

[1, 2, 3, 4, 5] Massachusetts Institute of Technology
naichun@mit.edu, ryanz@mit.edu, marissa@mit.edu, takehiko@mit.edu, ekll@mit.edu

Abstract. The presence of web2.0 and traceable mobile devices creates new opportunities for urban designers to understand cities through an analysis of user-generated data. The emergence of "big data" has resulted in a large amount of information documenting daily events, perceptions, thoughts, and emotions of citizens, all annotated with the location and time that they were recorded. This data presents an unprecedented opportunity to gauge public opinion about the topic of interest. Natural language processing with social media is a novel tool complementary to traditional survey methods. In this paper, we validate these methods using tourism data from Trip-Advisor in Andorra.
"Natural language processing" (NLP) detects patterns within written languages, enabling researchers to infer sentiment by parsing sentences from social media. We applied sentiment analysis to reviews of tourist attractions and restaurants. We found that there were distinct geographic regions in Andorra where amenities were reviewed as either uniformly positive or negative. For example, correlating negative reviews of parking availability with land use data revealed a shortage of parking associated with a known traffic congestion issue, validating our methods. We believe that the application of NLP to social media data can be a complementary tool for urban decision making.

Keywords: Short Paper, Urban Design Decision Making, Social Media, Natural Language Processing

1 Introduction

Compelling arguments for the use of bottom-up social opinions to inspire urban designs can be found in influential books such as "The Image of the City" (Lynch, 1960), "Death and Life of Great American Cities" (Jacobs, 1964), and "City is not a Tree" (Alexander, 1966). A large scale survey of public opinion for this purpose, however, was difficult in the 1960s because it relied on time-consuming traditional ethnographic tools such as surveys and interviews. Presently, modern geo-located data mining techniques can be deployed.

Compared to the traditional methods, such as sampled survey, NLP with social media is not only more cost-effective, but furthermore highlights urban issues on a

microscopic scale. Instead of a random sampling, the comments from social media are issue-driven and spontaneous by nature. Compared to traditional methods, our approach provides a more objective perspective of subjective perception from mass. The result will not be limited by survey designers' knowledge or biased by their preconceived mindset.

This project aims to complement traditional ethnographic tools by mining social media data for the purpose of better understanding cities. These techniques are applied to analyze tourism data from Andorra, a small country between Spain and France. Our goal is to show how mining social media data for urban patterns may lead to useful recommendations to Andorran tourism authorities. We investigate how Andorra is perceived by tourists by analyzing social media. Specifically, we analyze a total of 68,500 Andorra-related tourist reviews obtained from Trip-Advisor (TripAdvisor, 2016).

We used "Natural language processing" (NLP) with Trip-Advisor to determine the sentiment of the users towards the hotels, restaurants, and attractions in the city of Andorra La Vella, the capital city of Andorra. By analyzing the words used in the review, this project extracts the topics most commonly associated with good and bad reviews, thus giving a greater sense of how the tourist experience can be improved for each Andorran attraction. This pinpoints problems and opportunities for Andorran urban designers and planners to focus on.

2 Previous Work

The origins of natural language processing sentiment analysis can be found in previous work such as "Sentiment Analysis and Opinion Mining" (Liu et al., 2012) which used sentiment analysis to examine the text of online movie reviews to automatically detect opinions about the various movies. Following this paper, others have begun to use sentiment analysis to examine other product reviews. Our approach is to apply it to urban design decision making.

3 Data and Methods

Our data analysis and urban decision making workflow involves 7 steps, summarized in Fig. 1. The steps are described below.

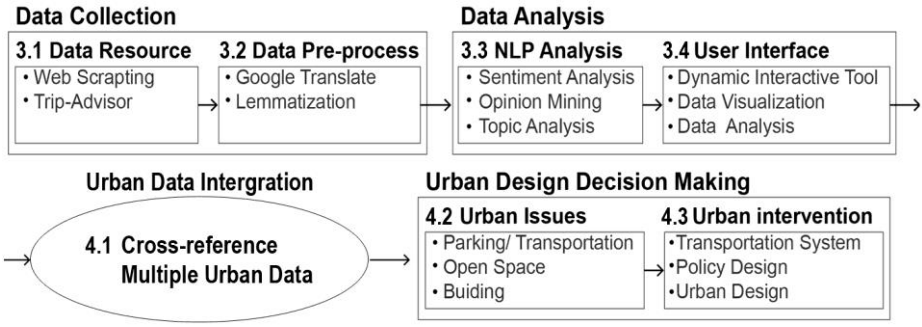

Fig. 1. Data analysis and urban decision making workflow

3.1 Data Resource: Web-Scripting Trip-Advisor

The first step of our analysis pipeline was to extract the relevant review information from each review for the Andorran destinations of interest. On Trip-Advisor we focused on three primary categories of tourist destinations: restaurants, hotels, and attractions. A web-scraper (a program which automatically obtains web content at large scale) was employed to extract the name, address, rating, and review attached to each of these destinations and record them in table format. From these re- views, we would identify sentences that describe urban issues, such as parking, restaurants or shopping. We used a script to parse the data from website based on the structured web format. Reviews were in multiple languages. Each review was translated into English, using Google Translate API, in order to enable subsequent language analysis. We are not using rating as part of the criteria, because it is not specific enough for the topic.

3.2 Data Pre-process: Translation and Lemmatization

Having standardized each TripAdvisor review page into English, the next step in our analysis pipeline was to simplify the page enough to enable NLP analysis by 1) extracting root words, and 2) removing superfluous words. This was done in several steps.

First, reviews had to be “tokenized”, that is, broken into single words using the Natural Language Processing Toolkit (NLTK) tokenizer (http://www.nltk.org). Sentences such as “the parking is awful” were broken into its individual words (“the”, “parking”, “is”, and “awful”).

Second, each word detected was “lemmatized”, by extracting their root word. The purpose of this step was to decrease the number of words that have to be analyzed by equating words like “run”, “runs”, “ran”, and “running” into their respective single root word “run”.

Next, the 200 most common words in the English language such as “the”, “is”, “are”, “a”, etc. were detected and deleted from the review pages, thus preventing these simple and frequent words from obscuring the more important and relevant words in the text that are actually important for understanding the topic matter. Noun phrases were next detected in the text using NLTK libraries.

3.3 NLP Analysis: Sentiment and Topic Analysis

This step helps to understand if a review is "positive" or "negative" for a related urban topic. We classified the reviews based on "sentiment". Based on NLTK libraries, we detected words that have a "positive" sentiment, and words that have a "negative" sentiment, within the reviews of each tourist destination. Each sentimental word was attached to its neighboring noun (which serves as the particular subject/topic), and the collection of sentiment-noun pairs were collected. For example, the sentiment-noun pair "beautiful lake" was collected.

Based on the sentiment analysis, an interactive visualization was created to enable user to have a comprehensive perspective of all Trip-Advisor destinations for La Vella. It allows for a heat map break down of reviews, sentiment, and relevance for all specific topics examined, as well as a breakdown by language of reviews to estimate the demographics of visitors to each destination.

3.4 User Interface: Dynamic Interactive Data Visualization and Analysis

The key technological development in this project was the production of a searchable visualization summary of all the Trip-Advisor locations to be used in the Andorra CityScope model. It includes a heat map of reviews, sentiment, and relevance to any searchable topic definable by a key word (e.g. "street", "parking", "shopping", etc.). It can also provide a breakdown by review language (Spanish, French, Russian, etc.), to further analyze the demographics of visitors to each specific attraction in Andorra. In addition to being searchable, each location has a popup card that displays information such as rating, summary of review, sentiment, and most popular keywords.

Fig. 2 shows an example of a search conducted in the user interface, searching for key word "street", in Spanish reviews. The degree of popularity of locations is indicated by regions colored from green (less popular) to red (most popular) with restaurants (red dot) and hotels (blue dot).

Fig. 2. HEAT map for search of key word "street" in the Spanish review

Fig. 3 shows the same search on key word “street”, but with the added condition to display the *sentiment* of the reviews, to show which locations were positively (green) and negatively (red) experienced by Spanish visitors.

Fig. 3. Map for Sentiment analysis search of key word “street” in the Spanish reviews

4 Result: Data driven Urban Design based on Sentiment Analysis

4.1 Urban Data Integration

We now demonstrate how our User Interface is a valuable tool for urban design decision making. Before we can identify the urban issues, we usually need to cross-reference and integrate multiple urban data, including quantitative and qualitative ones. In this case, we compare the Trip-advisor sentiment map to a land use map.

4.2 Urban Issues

There are two districts, the old city center area and the new pedestrian district, that have a particularly high concentration of negative reviews by searching key words which is most related to traffic issues - “street”, “traffic”, and “parking” (Fig. 4 below, red circled areas). We immediately note a strong correlation between the areas of the city that are most negatively reviewed, and areas that have a lack of parking (Fig. 5 below). This is a novel observation that is a direct result of our data mining.

In this way, sentiment analysis points out issues of importance to the urban designer. Most tourists in Andorra visit by car, and if the city supply of parking facilities does not meet the demand, it invariably has a negative effect on city tourism, and reduces the likelihood that people will want to visit or stay in Andorra.

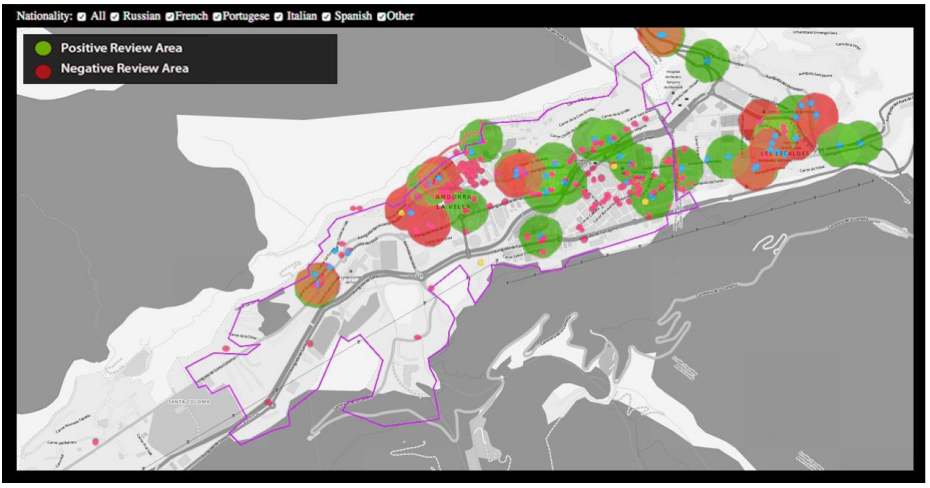

Fig. 4. Sentiment map for positively (green) and negatively (red) reviewed streets by searching key words "street", "traffic", and "parking", with areas that are concentrated with negative reviews circled in red

Fig. 5. Availability of parking (blue), plotted with negative review-concentrated areas of the city circled in red

4.3 Urban Intervention

Social media data may provide urban designers bottom up information to make decisions about where their future design improvements should focus on. In this case, urban designers may need to alter the parking facilities around the old shopping centers or renovate the area.

In parallel with the parking problem, we figure out that there is another more general problem regarding transportation systems. After searching the key word "transportation", we got both positive and negative entries on the heat-map. We then zoom into the areas that have more negative entries, and click on each of them to examine what the original review is about. The negative reviews include "too many cars", "do not have bus stop", "narrow road", etc. Through this method, we are able to identify the urban issues through users' point of view. In this case, we realize that with increasing car usage, the traditional road system in Andorra Le Vella does not meet road usage demand. One can address both of these problems simultaneously. One possible solution is to increase the public transportation system for intra-city network connections, which would help alleviate the overall traffic condition.

5 Conclusion

Our analysis of social media data revealed the most positively and negatively reviewed tourist locations. By comparing the regions to the land use map, we identified a prevalent issue of parking in the city. Urban decision makers may benefit from this new approach of the data source compared to the traditional methods, such as sampled survey. NLP with social media is not only more cost-effective, but also it provides an insight on the urban issues by examining spontaneous reviews, rather than survey answers guided by the survey designer. We suggest that our approach of combining social media "big data" with natural language processing to detect patterns of sentiment is a useful new methodology for the urban designer and planner, and can give data-driven insights that would have been hard to collect otherwise. We wish that our work will inspire more related research or applications

6 Future Work

In the current research, we successfully used NLP with social media to identify the parking problem in Andorra and to understand the causal reason. The same approach could be use by urban planners or designers in broader ways. Here are some possible applications:

- Land use – to understand if there is a good balance between residential and employment spaces in a certain area.
- Transportation – to analyze the composition and experience of different traffic modes: biking, working, driving and taking public transportation. It provides a great guide for improving the urban transportation infrastructure.

- Open space – to understand the perception of the public urban environment: if we need a park and, if so, where is the optimal location.
- Third places – if there is any complaint about a lack of restaurants, grocery stores, etc.; if there are enough cultural facilities, such as a museum or library.
- Healthcare – to examine the quantity and quality of healthcare resources, such as hospitals, pharmacies, and places to have physical exercises.
- Security – to understand how safe the neighborhood is.
- Education – if there are enough education resources like schools and day-care centers.

To enable the new applications mentioned above and to improve the precision and comprehensiveness of the methodology, future work will be conducted as follows:

- Trip-Advisor data analysis will be compared to government GIS data to validate our spatial observations.
- Analysis of other types of social media (Twitter, Flickr, Facebook, Instagram, etc.) will be conducted to reduce sampling bias in our data from analyzing only one type of social media.
- Trip-Advisor data analysis will be overlaid with Call Detail Record (CDR) data to understand the mobility patterns of tourists.
- Computer vision will be employed to detect the activities and the image of cities in social media pictures.
- Stakeholders will be able to conduct real-time monitoring or intervention using NLP with social media in circumstances such as during massive events or nature disaster.

Acknowledgements. We thank C. Sun for discussions and paper preparation, S. Chen for help in getting the user interface to run, C. Summit and J. Nawyn for helpful discussions and comments, members of Media Lab Changing Place group, and MIT Architecture Department for their encouragement. This work was part of the “Applied Machine Learning in Tour- ism of Andorra” project, which was supported by Andorra Government.

References

1. Agarwal, B.: Prominent Feature Extraction for Sentiment Analysis, Place of Publication Not Identified: Springer (2016)
2. Alexander, C.: A City Is Not a Tree, London (1966)
3. "CIA - The World Factbook -- Andorra." CIA - The World Factbook -- Andorra. Accessed April22,2016.http://people.uvawise.edu/pww8y/Supplement/ConceptsSup/Maps/CountryFactBook/NationsFactbook04/print/an.html (2016)
4. "Fact Sheet - TripAdvisor." Fact Sheet - TripAdvisor. Accessed January 22, 2016. https://www.tripadvisor.com/PressCenter-c4-Fact_Sheet.html (2016)
5. "Goslate: Free Google Translate API" Goslate: Free Google Translate API — Goslate 1.5.0 Documentation. Accessed January 22, 2016. http://pythonhosted.org/goslate/ (2016)

6. Guzman, E., Walid M.: How Do Users Like This Feature? A Fine Grained Sen- timent Analysis of App Reviews, 2014 IEEE 22nd International Requirements Engineering Conference (RE), 2014, doi:10.1109/re.2014.6912257 (2014)
7. Jacobs, J.: The Death and Life of Great American Cities, Pelican Books (1964)
8. Liu, B.: Sentiment Analysis and Opinion Mining, San Rafael, CA: Morgan & Claypool, (2012)
9. Lynch, K.: The Image of the City, Cambridge, MA: MIT Press (1960)
10. "Natural Language Toolkit" Natural Language Toolkit — NLTK 3.0 Documentation. Accessed January 22, 2016. http://www.nltk.org/ (2016)
11. "Text Mining Online | Text Analysis Online | Text Processing Online." Text Mining Online Text Analysis Online Text Processing Online. Accessed January 22, 2016. http://textminingonline.com/about (2016)
12. "TextBlob: Simplified Text Processing" TextBlob: Simplified Text Processing — TextBlob 0.11.1 Documentation. Accessed January 22, 2016. https://textblob.readthedocs.org/ (2016)