Audio paper 5 – MEDIATING SUSTAINABLE CITIES: data collection from social media

Welcome to the series of audio papers: Mediating Sustainable Cities!

I am Paola Monachesi, researcher at Utrecht University. In this audio paper, I will discuss a novel methodology to collect social media data from specific communities active on Twitter. The aim is to highlight an alternative smart city discourse and identify communities that shape a human-centred smart city with their communication and actions. The data collection is at the basis of the work presented in this series.

Jingle

Let's start by considering some reasons for why it can be useful to collect social media data for our research.

Social media platforms can offer massive, dynamic and diverse data that allow for an investigation of the behaviours, opinions and feelings of their users. These are spontaneous and authentic data that allow for the creation of textual corpora based on millions of users. Social media data can complement more traditional data collections, such as those emerging from focus groups, interviews and surveys.

Social media data make possible to identify voices that might not easily emerge through traditional approaches since these data emerge from real-world situations, from various people and they are not elicited by the prompt of researchers. Finally, it might be less time consuming to retrieve these data if the right methodology is in place and appropriate tools are available. However, these data present also challenges such as interactivity among users, content can be ephemeral, dynamic and massive, giving rise to large amounts of data or the so-called big data.

Big data availability is often associated to the use of computational methods in order to harvest data. APIs are used for this purpose. APIs are interfaces that the social media platforms make available for automatic extraction of data or alternatively web scraping in case the data are not publicly available. The availability of big amounts of data calls for quantitative analysis. Big data are investigated to find patters making use of word frequency, topic modelling or social network analysis. However, these approaches often fail to show the deep meaning of data.

Jingle

You might wonder what is different in my approach to data collection from social media.

In my research, I have integrated traditional approaches to data collection with social media content harvesting. I have not collected data containing specific terms or from a particular time frame, but I have looked into data produced by specific communities in order to analyse their behaviour. The identification of communities or group of users, however, can be challenging. I have focussed on two groups that could be interesting to analyse in the context of urban sustainability because they might highlight a different smart city discourse. They are the elderly and creative skilled migrants.

Jingle

You might wonder why I have chosen these two among other possible communities.

Well, my interest on elderly has emerged from the European project Grage, dealing with green and healthy living in urban context. In this project, I was leading a research group with focus on technology and elderly well-being in the city.

As for the skilled migrants, my interest in this group comes from the fact that I am a creative migrant myself, an Italian that has been living in the Netherlands for almost 30 years and I was particularly interested in the differences between skilled migrants and refugees in mediating urban spaces of innovation. While the former are usually invited to contribute to the development of smart cities the latter don't have the right to the urban space they often occupy illegally.

Jingle

The next step then is to identify these communities in social media.

In order to extract social media data from these two communities, Twitter has been used as social network platform. Twitter data are publicly available, making data collection easier and ethical issues less problematic.

So, how do we identify elderly?

It might be problematic because in Twitter, the age of users is not visible and detecting users of the desired age is quite challenging. In addition, defining age is not trivial. We have decided to make use of chronological age of users, but we also group them in three classes, that is users below 55 years old, users between 55 and 67 and above 67 years old. These groups correspond to three life stages that are related to the active working life of the individuals (i.e. below 55), the pre-retirement stage (between 55 and 67) and post retirement (above 67).

We carried out a manual step to identify elderly on Twitter. We started from a profile created for this purpose and we followed several organizations related to ageing and that one could expect to have old adults among its followers.

In this way, we have identified users for which it would be possible to classify their age in one of the three groups. The identification of the age was possible using various tactics: sometimes users display their date of birth in their Twitter name but we have considered also the picture as quite revealing and the information in their profile.

The users selected had to meet several requirements: their tweets should be publicly accessible and written in Dutch; they should have tweeted at least 400 tweets and they should show social activity, that is users with a number of followers higher than 300.

Once the users have been identified, the next step was to extract the relevant data, such as their tweets and retweets or profile of their relations, through web crawling, which is a way to extract content from a web source via the API available through Twitter. Extracted data has been archived in a database created to this end. Thanks to the database structure, it becomes quite easy to create corpora based on specific queries relevant for a given research question. For example, a corpus that includes all the tweets of the elderly above 67 or all the tweets that contain specific words.

The tweets from these selected users constitute a data set to train machine learning classifiers to detect age automatically. In our machine learning approach, we have used language features, social media specific metadata and in particular URL's and hashtags.

Jingle

One may wonder whether this methodology can be employed to identify different communities.

I have adapted the methodology to identify another group of users, that is skilled creative migrants. Identifying creative migrants is even more challenging than the elderly since they do not characterize themselves as such.

The first step was to identify occupations related to creativity. We have used three different resources to single out the relevant creative industries and derive the corresponding occupations. The result was a list of 164 creative professions that we have further categorized in 11 sectors, within the creative industries.

We have used this information to identify creative migrants. We have matched the professions identified with the profile description of Twitter users in order to select migrants that characterize themselves through their work. The output was filtered on the basis of the location of these users in order to retain only those ones living in the Netherlands. As in the case of the elderly, we started from a profile created for this purpose on Twitter and by following several organizations related to creative industries that one could expect to have skilled creative migrants among their followers.

The final step was to retain the users that satisfy a number of criteria: the user's screen name is not Dutch, the user clearly states in the profile description one of the creative professions we have selected, there is geo-location information attesting that the person lives in the Netherlands, the tweets are in more than one language other than English and Dutch and the person has a public account, that is, the account was not protected at the time of the data collection.

The final list was manually verified to satisfy these criteria but we have also carried out a content analysis of the profile description of our users to further verify the correctness of the selection. It is based on word frequency and it confirms the validity of the approach.

As in the case of the elderly, data was collected from these users and stored in a database in order to create corpora to analyse specific research questions, as will become clear in the next audio paper.

Jingle

Collecting users' data rises ethical questions. What about ethical issues related to our data collections?

Twitter data is publicly and freely available; no informed consent applies. If users do not want to make their data available can choose to do so in their privacy settings. Twitter explicitly mentions analyzing tweets and their metadata by universities as a valid and to be expected use of information shared via Twitter.

The Twitter dataset does not contain sensitive data. Data were collected following the principles of finality and proportionality, so data were processed only with the purpose of conducting research work.

The methodology developed can be easily adapted to deal with other languages and professional categories. It is consistent with the idea that citizens need to regain control over their data. The collected data are employed to promote and improve sustainable policies.

I hope the data collection methodology I have described will make people realize how easy it is to detect people on the basis of the data they produce. Information that big tech companies use to make profit. They don't even need to detect people the way I have done since they own the data.

In the next two audio papers, we will explore methods of analysis and interpretation of data. I will discuss quantitative and qualitative data analyses and show how they can be integrated to provide deeper insights into our cases.

Don't miss the next audio papers!