Mapping Energy Frames: an Innovative Approach Engaging with New Forms and Types of 'data' in Social Media Analysis to Frame Energy Demand

Mauri, Andrea; Bozzon, Alessandro; De Kok, Roos

DOI
10.17418/B.2019.9789491937439

Publication date
2019

Document Version
Final published version

Published in
From efficiency to reduction. Tackling energy consumption in a cross disciplinary perspective

6 MAPPING ENERGY FRAMES: AN INNOVATIVE APPROACH

Engaging with New Forms and Types of 'data' in Social Media Analysis to Frame Energy Demand

ANDREA MAURI, ALESSANDRO BOZZON AND ROOS DE KOK

INTRODUCTION

Understanding how individuals consume energy is considered to be a fundamental step in improving energy conservation and stimulating energy efficiency.

Multiple studies have shown how feedback loops encourage energy conservation and efficiency among policy makers and citizens. Fischer (2008), for instance, explores the ways in which a sense of competition, social comparison, and peer pressure impels people to adopt better energy consuming behavior. The online platform 'Social Electricity' allows citizens to compare energy footprints with friends, neighbors, or other users. This process, Kamilaris, Pitsillides, and Fidas (2016) argue, affects people's energy awareness, making them more sensitive to the environment and motivating them to behave more sustainably. Many other energy-saving applications have been developed (Albertarelli et al., 2018), which exploit gamification and social interaction to promote energy-efficient lifestyles.

All of these studies are only effective, however, insofar as we thoroughly understand which energy-consuming activities people perform and how they carry them out. By energy-consuming activity, here, we mean a practice that impacts energy consumption, whether directly or indirectly.

To date, researchers have used multiple methods to collect insights about energy-consuming activities. Smart meters and smart plugs provide aggregate data on domestic energy consumption. Looking at different current waveforms and voltage signatures makes it possible to isolate the signal of the single appliance (Froehilich et al., 2011; Weiss, Helfenstein, Mattern, & Staake, 2012; Parsa, Najafabadi, & Salmasi, 2017). For their part, surveys and interviews, if planned correctly, can potentially break down energy overall consumption data into detailed end uses (Vassileva, Wallin, & Dahlquist, 2012; Torriti, 2017).

In presenting composite readings, comprising both sensors and answers to ad-hoc questions, the aforementioned sources are especially reliable in terms of quantitative data and qualitative information. Nonetheless, they come with several drawbacks. Smart meters and plugs are costly to set-up. What is more, the data that they provide lacks of contextual information, and often not publicly accessible. Further still, their disaggregation of data is far from perfect (Froehilich, et al., 2011). Conducing and processing surveys and interviews is too time consuming to perform frequently.

However, hundreds of thousands of people use social media daily, sharing texts, videos, and pictures related to their activities. Unlike traditional data, social media offer semantically rich information, not to mention frequent, high-granular updates that cost little or nothing to extract. For these reasons, researchers have been using social media to study human practices (Zhu, Blanke, & Gerhard, 2016; Bodnar, Dering, Tucker, & Hopkinson, 2017) such as travel behavior (Bocconi, Bozzon, Psyllidis, Bolivar, & Houben, 2015; Rashidi, Abbasi, Maghrebi, Hasan, & Waller, 2017; Zhang, He, & Zhu, 2017), modes of transportation

(Zhang, He, & Zhu, 2017), and nutrition patterns (Fried, Surdeanu, Kobourov, Hingie, & Bell, 2014; Abbar, Mejova, & Weber, 2015; Fard, Hadadi, & Targhi, 2016).

Our hypothesis is as follows: since posts on social media refer to daily activities, they are either directly linked to energy-consuming activities or contain information about them in their semantic signature. Hence, by processing the content of a social media post, it should be possible to extract information about the energy consuming activity to which it refers. Through detailed analysis, we aim to answer the following research question: *how useful is social media as a complementary source of information in describing energy-consuming activities?*

In this chapter, we refer to four types of energy-consuming activity: (1) *dwelling*, (2) *food consumption,* (3) *leisure,* and (4) *mobility.* This typology – which is based on previous studies (Tukker et al., 2006; Backhaus, Breukers, Mont, Paukovic, & Mourik, 2013) – includes a wide range of practices impact upon lifestyles' energy footprints. Here we are interested in individuals' energy consumption; accordingly, we have set activities related to labor and industries aside. *Dwelling* encompasses the use of home appliances (e.g. washing machines and dryers); *mobility* refers to the energy required in moving among places; *food consumption* includes the use of resources for preparing, processing, and consuming food; and *leisure* refers to the energy used to perform recreational activities (e.g. watching tv, playing videogames, and socializing).

It should be remembered, though, that social media also come with disadvantages. They are biased toward a certain demographic (e.g. young people) and they can be noisy and ambiguous – this is only to be expected, since their original purpose was not to share energy-consuming activities. We will broach these drawbacks in the course of this chapter.

To address the challenges, we created an energy-consuming activities ontology, which has allowed us to identify important con-

cepts and see how they are interrelated. It provides a structured body of knowledge about how social media posts are connected with practices in the physical world. We designed a data processing pipeline, which extracts information about energy-consuming activities from social media posts. Composed by different modules, this pipeline collects social media posts from different sources (Twitter and Instagram); enriches the data (e.g. by annotating the images); classifies the posts according in the four categories of energy-consuming activity established above (using a dictionary and rule-based classification algorithm); and, finally, publishes the information obtained in the previous steps in the JSON-LD format, presenting them as instances of the ontology in question.

Starting from our work published in de Kok, Mauri, & Bozzon (2019), this chapter elaborates on the challenges addressed in developing an energy-consuming activities ontology and data-processing pipeline. In addition, we reflect in more depth on the potential and weaknesses of our approach, before proposing future applications. The remained of this chapter is structured as follows: first we summarize the design of the ontology and data-processing pipeline, highlighting the challenges that arose during their development. We report their use in two case studies, Amsterdam and Istanbul, unpacking both the strengths and weaknesses of our approach. Finally, we discuss our results, proposing possible future research directions that can be built upon our research.

THE SOCIAL SMART METER ONTOLOGY

In creating an ontology, we intend to understand and unambiguously conceptualize, the domain of energy-consuming activities. This, we hope, will facilitate interaction among different fields of study interested in energy consumption. To do so, we aim to identify significant concepts and showing how these are interrelated, establishing terms for describing and representing this do-

main in a structured way that can be read by machines. While it not may be useful for end users, the ontology has allowed us to design the data processing pipeline described in the next section and others might use it to develop an IT framework in this field. Moreover, the ontology makes it possible for external services to seamlessly integrate our pipeline's outputs into their process.

To build the ontology, we used the guidelines provided by the 'Methontology' approach (Fernandez-Lopez, Gomez-Perez, & Juristo, 1997), a well-structured process for building ontologies from scratch. The requirements were defined by competency questions as provided in Suárez-Figueroa, Gómez-Pérez, & Villazón-Terrazas (2009).

Table 7 shows the competency questions created to define the ontology. We are interested in associating specific people with different energy-consuming activities. We want to understand where and when these activities took place and what kind of appliances or tools were used. Furthermore, we want to aggregate this information by time and location.

Figure 23 shows the developed ontology. An Individual consumes Energy by performing an Activity in a given Location, which might either be a single Place (i.e. a point of interest) or Path (i.e. a sequence of places). An activity can be of different types: Mobility, Leisure, Dwelling, and Food Consumption. A Mobility activity includes a Mode of Transportation, which the Individual uses to move among places. Both Leisure and Dwelling activities may use Appliances, which are divided into White (big appliances) and Brown Goods (small appliances). Food Consumption is associated with the Food being eaten, where it is produced, and how it is processed.

Figure 24 shows those elements in our ontology that relate to social media. A User, with a social media Profile, can publish Posts containing text or links to Video, Images, Events, or other resources. A Post can also refer to a Location and mention other Users.

#	Competency Question
1	Does the individual perform an energy-consuming activity?
2	If so, what type of energy-consuming activity is performed?
3	At what place is the activity performed? 1. To what category does this place belong? 2. What are the coordinates of this place?
4	At what time is the activity performed?
5	What is the duration of the activity?
6	Does the individual use an object to perform the activity? 1. If so, what kind of object?
7	In case of mobility activity, what kind of mode transportation is used? 1. What path (i.e., set of places) was taken?
8	In case of leisure activity, what kind of artifacts are used? 1. In case the artifact is an appliance, what is its power consumption?
9	In case of dwelling activity, what appliances are used? 1. What is their power consumption?
10	In case of food consumption activity, what kind of food is consumed? 1. What are the ingredients? 2. How is the food processed? 3. What kind of appliances are used to process the food? 4. Where is the food processed?
11	How many energy-consuming activities are performed during a certain time-span and within an area?

Table 7 Competency questions that form the ontology's set of functional requirements

To avoid a proliferation of ontologies covering the same concepts, and facilitate our data models' integration with other systems, we looked at existing ontologies that bear upon energy consumption, food, travel, and social media.

Table 8 summarizes existing ontologies that partially overlap with concepts included in our own. 'The Suggested Upper Merged Ontology' (SUMO) (Niles & Pease, 2001) is the largest existing formal public ontology. Since it covers many of the con-

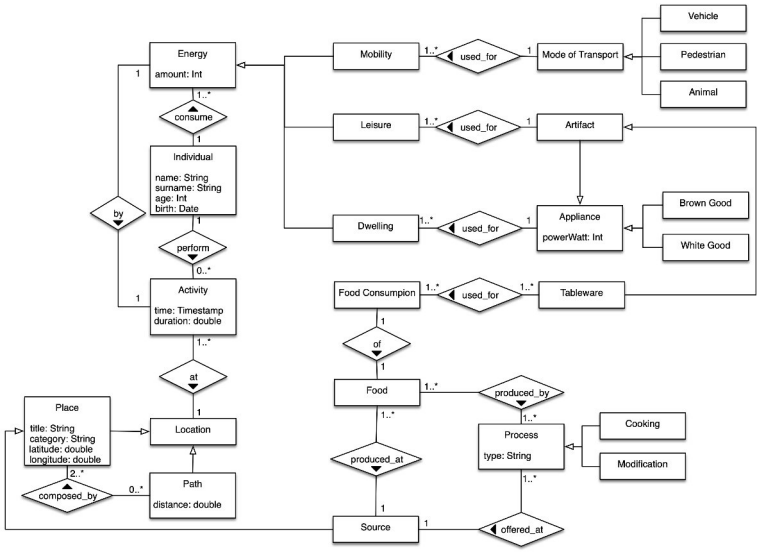

Figure 23 Conceptual model of the energy consumption ontology

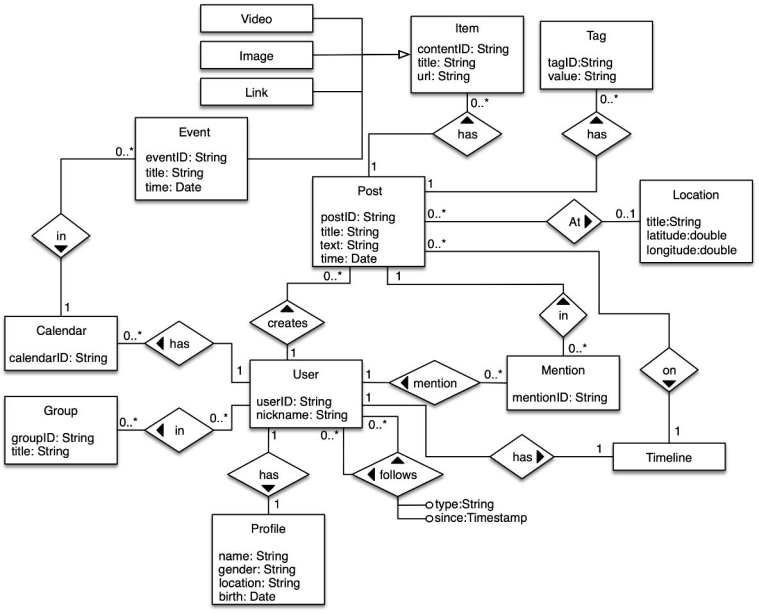

Figure 24 Social media conceptual model

cepts needed in our ontology, we used it as foundation for our model. 'The Semantic Tools for Carbon Reduction Energy Model' (SEMANCO) (Madrazo, Sicilia, & Gamboa, 2012) focuses on concepts related to energy consumption and CO2 emis-

	SUMO	SEMANCO	EU	FO	TO
Energy Activity					
• Energy units	+	+	+	-	-
• Consumption	+/-	+	+	-	-
• Individual	+	+	+	+	-
Location					
• Location	+	+	+	-	+
• Path	+	-	-	-	+
Dwelling					
• Activity	+	+	-	-	-
• Appliance	+	+	+		-
Food consumption					
• Activity	+/-	-	-	+	-
• Food	+	-	-	+	-
• Food chain	-	-	-	+	-
• Tableware	+	-	-	-	-
Leisure					
• Activity	+	+	-	-	+
• Artifact	+	-	-	-	+
Mobility					
• Activity	+	+	-	-	-
• Mode of transportation	+	-	-	-	+

Table 8 Overview of the current state-of-the-art in ontologies that focus on the domain of energy-consuming activities (+: included; +/−: covered to some extent; −: not included)

sion. We included it in our ontology to model energy consumption. In modelling the household, we also included aspects of an existing ontology: the EnergyUse platform (EU) (Burel, Piccolo, & Alani, 2016), which describes home appliances. We also included concepts related to food consumption from the BBC Food Ontology (FO) (BBC Food Ontology, 2014), which contains recipes and information ingredients and processing methods. Finally, the mobility domain was covered by drawing on the Travel Ontology (TO) (Stevens, 2009).

To cover the social media domain, we used the Friend-of-Friend (FOAF) and Semantically-Interlinked Online Community (SIOC) ontologies. They contain concepts relating to user accounts, posts, and the relations among users (i.e. friendship).

Whilst we could have built most of our ontology simply by reusing these ontologies, they were created for different purposes

and needs. Having been designed to model appliances' energy consumption; describe food and ingredients; or model modes of transportation and social networks, they were not intended to describe energy-consuming activities. It is important to remember, therefore, that while we chose to include aspects of these ontologies that refer to concepts of interest, their semantic meaning may only partially cover our concept or slightly differ from it.

For this reason, we had to find the right trade-off between reusing existing ontologies and creating new elements. Although focusing purely on the first strategy would ensure maximum interoperability with existing frameworks, the resulting ontology would lack the specificity required for the domain of energy-consumption. We address this challenge in two ways:

1. Where an existing entity already covers one of our concepts, but its actual meaning differs from that concept, we have created a new entity and drawn a relation of equivalence between them. This applies to the *Energy_Quantity_And_Emission* entity in the SEMANCO ontology. Although it refers to the concept of energy consumption, its precise meaning is that of direct energy consumption. Since in our ontology we want to include indirect consumption of energy too, we have created the *Energy* entity, which is *equivalent* to the SEMANCO one.

2. Where an existing entity partially covers one of our concepts, we have created a new entity and drawn what we call an 'is-a' relation between them. The class *Cooking* in the SUMO ontology, for instance, partially covers the food processing concept. Hence, we have created the *Process* entity in our ontology. *Cooking*, we propose, *is-a Process.*

The ontology was implemented using Web Ontology Language (OWL 2 Web Ontology Language Document Overview, 2012) with Protegè, which is available on the following companion website (http://social-glass.tudelft.nl/social-smart-meter/#ontology).

THE SOCIAL SMART METER DATA PROCESSING PIPELINE

Figure 25 shows the data processing pipeline. It is composed of four main modules: Data Collection, Data Enrichment, Classifier, and Linked Data publisher.

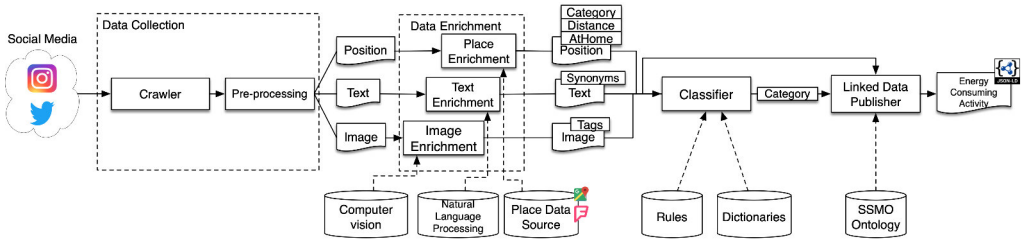

Figure 25 Data processing pipeline

First the pipeline retrieves users' posts from social media. It then enriches the data using state-of-the-art computer vision and natural language processing techniques, which respectively apply to the images and text in a given social media post. In addition, information about place is enriched by looking for its category in external data sources. The enriched information is then used to classify the social media post as corresponding to one or more types of energy-consuming activity. Finally, the pipeline publishes information about the energy-consuming activity as linked data[1] by combining the outputs of all the previous steps.

DATA COLLECTION

The Data Collection module retrieve posts using the API provided by social networks. These social networks are queried by providing the GPS coordinates as a bounding box, thus retrieving posts that are created within a specific area. The module also preprocesses the posts, removing stop-words, hashtags, and special characters from the text. It then proceeds to perform the stemming (i.e. reducing words to their root form) and tokenization (i.e. segmenting the words in a message).

DATA ENRICHMENT

The Data Enrichment module takes in the text and images included in a post, as well as its location. The text goes through a word disambiguation algorithm, in this case the Lesk algorithm (Lesk, 1986), which disambiguates a term's meaning by looking at the surrounding words. We use the 'Adapted Lesk' algorithm implementation (Banerjee & Pedersen, 2002), which incorporates the Word Net lexical database. The output of this step in our process are the words enriched with their definition and synonyms.

The modules apply the state-of-the-art techniques for visually detecting objects and scenes in images. We decided to use both since they provide complementary information about what is represented in the pictures. For object recognition, we use a convolutional neural network, 'Mask R-CNN' (He, Gkioxari, Dollar, & Girshick, 2017), which is trained on the Microsoft COCO dataset[2]. For scene recognition, we use the ResNet50 neural network[3], which is trained on the Places dataset[4]. The output of this step is a set of terms describing the objects contained in the images and the scenes that they depict.

Finally, the module uses geographical coordinates (or toponym) that locate the creation of a post in an attempt to retrieve additional information about the category of energy consumption with which it is associated. Our intuition is that the type of place where such an activity is performed will help us understand its type. There is a high chance that an activity performed in a restaurant, for instance, will belong to the food consumption category. We use Google Places, Foursquare, and their APIs to retrieve the category of the place indicated in either its name or geographical coordinates. Moreover, once we have the set of places that a user visits, we attempt to estimate his home location using a density-based spatial clustering algorithm (DBSCAN (Ester, Kriegel, Sander, & Xu, 1996)). This algorithm separates high-density clusters from low-density clusters. The home of a user, we assume, is in the high-density cluster.

CLASSIFIER

The Classifier module uses information retrieved in the previous steps to classify social media posts into one or more energy-consuming categories. State-of-the-art classifiers need a large set of manually annotated data. To the best of our knowledge, such a dataset does not exist and its creation was out of scope of this work. We have therefore chosen to take a hybrid approach to classification based on dictionaries and classification rules.

We define a dictionary as a set of terms related to a type of energy-consuming activity. Ingredients or cooking utensils, for example, are related to the category of food consumption. The underlying idea is to compare the terms extracted from the message, image, and location with the words contained in the dictionaries. We built a distinct dictionary for each of data type, because this helps to rule out ambiguity to some extent. The text token 'tram,' for instance, may refer to an activity related to mobility, while a tram present in the background of an image is not necessarily related to a user's activity.

The dictionaries for image and place tokens are predefined. They are composed respectively by the set of classes contained in the pre-trained models and the set of venue categories present in the data sources. We associate each of term with one or more type of energy-consuming activity.

The dictionary for the textual component of a social media post is more complex. Creating it manually is labor intensive and all of the relevant terms are unknown. Its creation is delegated to a

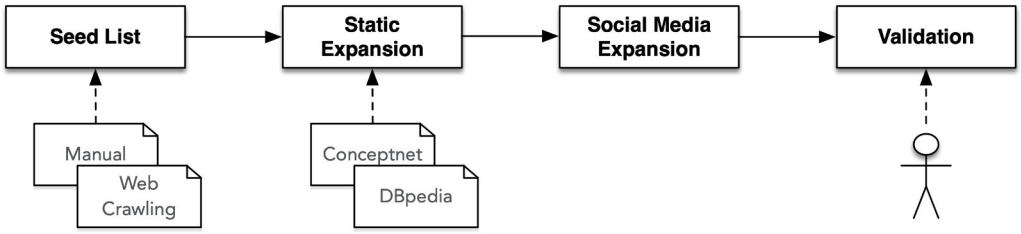

Figure 26 Overview of the process of creating the dictionary for textual components.

component shown in figure 26. This process has the following steps:

3. **Seed List**: a set of topical keywords is defined for each of type of activity being addressed (dwelling, mobility, food consumption, and leisure). The idea is to identify terms (expressed in the languages spoken in the targeted areas) that are associated with energy consumption or energy saving measures. This list is compiled both manually and by crawling web sources such as Oxford Food Reference, Wikipedia, and E-Commerce website.

4. **Static Expansion**: in enriching the set of domain-specific keywords, this semi-automatic step aims to increase the amount of relevant social media posts retrieved by the system. In this phase, each term detected in the previous step is searched in different sources, namely ConceptNet[5] and Wikipedia Category tree. ConceptNet provides a large multilingual knowledge graph that helps computers understand the meanings of the words that people use. ConceptNet expresses concepts (words and phrases) extracted from natural language text, as well as their relation to other concepts. These relations (e.g. synonym, antonym, Isa, PartOf, ExternalURL, etc) were derived from a wide variety of sources (e.g WordNet, Wikipedia, and Dbpedia etc.). We can filter out the noisy data based on its edge weight, type of relation, and number of hops pattern. The Wikipedia category tree presents the contents of each category name as a tree structure. As an example, the category name "meat" is articulated as following tree structure: Beef→ Beef dishes→ Hamburger Steak.

5. **Social media expansion** (Mauri, Psyllidis, & Bozzon, 2018): this step starts from a small set of social media posts. It retrieves a set of candidate posts, considering those that contain one of the dictionary terms established in the previous step. It then computes the similarity[6] among the candidates and the centroid of the starting set. Given a confidence

range [h, l], it labels tweets whose similarity is greater than h as positive examples, and twitters with similarity lower then l as negative, leaving the others unlabeled. A set of candidate words is then created by taking all of the terms contained in tweets labeled as positive. The process considers a word valid if it appears on a list of similar words obtained from the Word2Vec model on the Google News corpus. In this way, we are able to isolate those words that belong to the same dictionary context. The process is iterated until either all of the tweets are labeled or no more candidate tweets can be found.

6. **Validation**: We ask human validators, who are based in different countries, to validate the English terms and their corresponding translation in a given language[7]. The validators will be asked to check whether the terms are relevant to the category in which they have been placed (i.e. dwelling, mobility, food consumption, and leisure) and that the terms translation has been correctly assigned. The validators are also asked to validate the country-specific terms and add missing terms if possible.

Figure 27 illustrates the rule-based approach used to classify social media posts. We check to see whether each term obtained in the enrichment step appears in any of the dictionaries. In the case of food consumption and leisure activities, for example, we classify posts created at home also as dwelling. Furthermore, we look at the distances among users' posts. If a distance between two posts exceeds a threshold of 0.2km[8], we classify them also as a mobility activity. In this case, we try to infer the mode of transportation by looking at the duration elapsed between posts and the distance travelled.

To address the problem of the noisiness of social media data, we model the confidence of our classifier by using three parameters: (1) for each type of data, the ratio of relevant tokens; (2) for each term, its relevance to the category of energy-consuming activities; (3) a score indicating the extent to which the type of data is

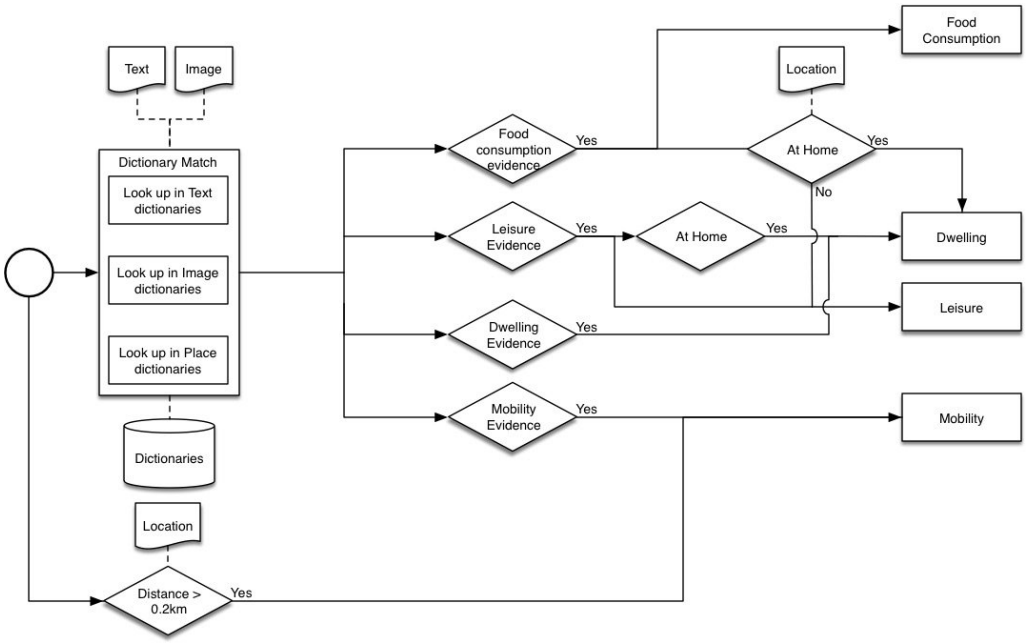

Figure 27 Illustration of the rule-based approach

informative about its category. While the presence of food in an image makes it easy to allot it to the category of food consumption activity, for example, a picture of a plane or train hardly identifies a mobility activity, for users rarely post photographs of mode of transportation while traveling.

Taking all the above into the account, we compute our classifier's confidence as follows:

$$\begin{aligned}\text{confidence}_x &= \sum_y \left(\frac{N_{relevant,x,y}}{N_{relevant,y}} \cdot w_{x,y} \cdot \frac{1}{N_{relevant,x,y}} \sum_x \text{scores}_{x,y} \right) \\ &= \sum_y \left(\frac{1}{N_{relevant,y}} \cdot w_{x,y} \cdot \sum_x \text{scores}_{x,y} \right)\end{aligned}$$

Where $N_{relevant}$ is the number of relevant terms, *w* is the informativeness score, *x* is the type of energy consuming activity, *y* is the type of data, and *score* is the vector of the scores of all the relevant terms.

The relevance scores are computed separately for each type of energy consuming activity. In the case of text token, the score is computed as the similarity between the term vector and word embeddings of the words contained in the dictionary obtained using Word2Vec. In the case of an image annotation, we use the score provided by the object or scene recognition model. For the place, we use a binary score, depending of the presence of the place category in the dictionary.

The values for the weight were defined by asking to a group of users to rank the data type according to their informativeness on a scale from 0 to 10 (*Not informative at all* to *Very Informative*). We used a sample of 100 posts, collecting 9 responses for each post. The final values were computed as an average, as shown in Table 9.

Table 9 Weight values obtained by asking user's opinions

	Text	Image	Place
Dwelling	0.35	0.40	0.25
Food	0.33	0.37	0.30
Leisure	0.35	0.32	0.33
Mobility	0.37	0.33	0.30

Finally, the classifier confidence for a category x is the average of the contribution of each data type.

LINKED DATA PUBLISHER

This component takes the output of the previous modules and combines them to create instances of the social smart meter ontology. We use an instance of TripleWave (Mauri et al., 2016), a reusable and generic tool for publishing linked data streams on the web in JSON-LD format.

The aim of this final step is to help establish our ontology's inter-

operability with other services by sharing a common understanding of the domain of energy-consuming activities. Now that we have published our data in JSON-LD, others can define custom queries in a standard language (e.g. SPARQL and RDF query language[9]) and perform ad-hoc aggregation to satisfy their own research needs.

EVALUATION

We conducted case studies in the cities of Amsterdam and Istanbul. Unfortunately, we were not able to perform a case study on the city of Graz since there was not enough social media activity to produce meaningful results, probably due to the size of the city.

We collected data between two periods: 22–27 June and 27– 28 July 2018. At first, to provide a first round of insights into 'the Social Smart Meter framework,' only social media posts created in Amsterdam were collected. After this, social media posts created in Istanbul were also collected, so as to compare results between the two cities. Whereas about 150k posts were collected in Amsterdam (130k from Instagram and 20k from Twitter), 120k were gathered in Istanbul (90k from Instagram and 30k from Twitter).

We collected posts regardless of language. In the case of Amsterdam, we considered terms in English and Dutch, while in that of Istanbul we considered terms in English and Turkish. It is worth noting that terms in different languages are needed only for the textual part of the social media posts, not for image labels and place categories.

For text processing, we used three pre-trained embeddings: for the English language we used the model trained on the Google News corpus[10]; for Dutch we used a model trained on the combined datasets of Wikipedia[11], Sonar500[12], and Roularta corpus[13];

while for the Turkish language we used a model trained on the Turkish Wikipedia dataset[14].

PERFORMANCE EVALUATION

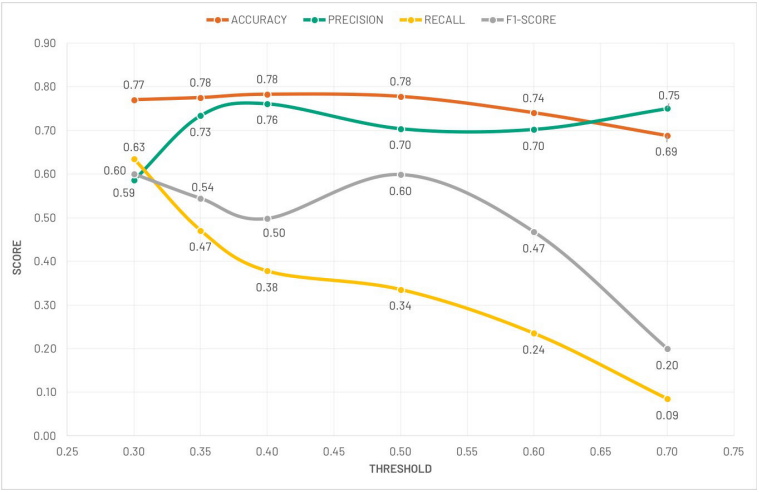

Figure 28 Accuracy, precision, recall and F1-Score at different threshold values

First, we evaluated the performance of our pipeline in terms of accuracy, precision, recall, and F1-score. Precision is the ratio between posts classified correctly in one of the categories and all of the classified posts. Recall is the ratio between posts classified correctly in one of the categories and the set of all relevant posts. Accuracy is the fraction of posts correctly classified, taking into the account the true negatives (i.e. the posts correctly classified as not belonging to any category). Finally, the F1-score is the harmonic average of the precision and recall.

Figure 28 shows the values of performance scores with respect to different threshold values. The recall scores decrease while increasing the threshold; less relevant social media posts have sufficiently high confidence scores to exceed the threshold. Increasing the threshold results in both less true and less false positives. However, the numbers of true and false positives do

not decrease proportionally. Based on the plot, a threshold of either 0.30 or 0.35 appears to result in the best performance.

USE CASES

Figure 29 shows the overall distribution of posts classified as corresponding to any one of the energy-consuming activities. In Amsterdam (figure 29a), most social media posts are created around the city center, Burgwallen-Nieuwe Zijde, the neighborhood with the highest density. In Istanbul (figure 29b), multiple districts exhibit a high volume of energy-consuming activities: Başakşehir and Beşiktaş in the European side of the city and Kadıköy in the Asian side.

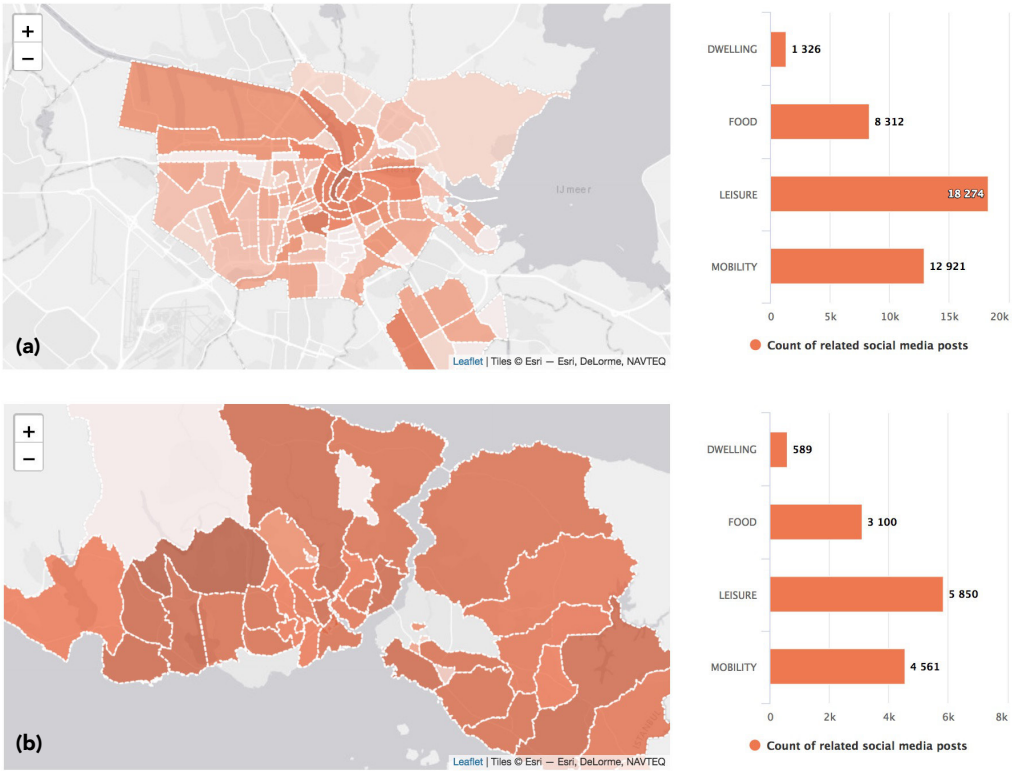

Figure 29 Overall distribution of energy-consuming activities in Amsterdam (a) and Istanbul (b). In the case of Istanbul, blank areas refer to district where no social media posts were found.

	Amsterdam	Istanbul
Dwelling	3.25%	4.18%
Food	20.36%	21.99%
Leisure	44.75%	41.49%
Mobility	31.64%	32.32%

Table 10 Percentage of posts classified in any of energy-consuming activity category

Table 10 shows the percentage of posts classified as falling into any of the energy-consuming categories. In both cities we found very few posts classified as dwelling. For both Amsterdam and Istanbul, the leisure category has the largest share (approximately 40%) of posts compared to the other categories. The mo-

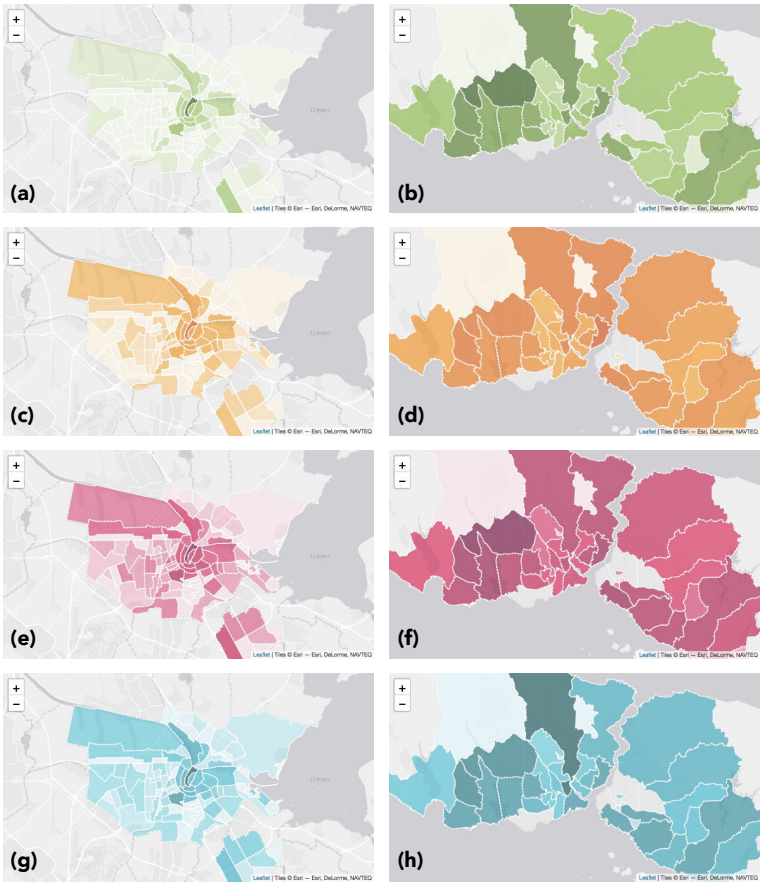

Figure 30 Map visualizing the distribution of social media posts; (a,b) refer to dwelling, (c,d) refer to food consumption, (e,f) refer to leisure, and (g,h) refer to mobility. In the case of Istanbul, blank areas refer to districts in which no social media posts were found.

bility category has the second largest (approximately 30%). The category of food consumption has a rather small share (approximately 20%).

In Amsterdam (figure 30a) dwelling activities are concentrated in the city center. In Istanbul (figure 30b) the posts are more evenly distributed, with a higher concentration in the European part of the city. As figure 31 shows, the most informative terms are *house*, *TV*, and *gaming*, while the most recognized objects are *tv*, *laptop*, and *keyboard*, indicating both recreational and work activities.

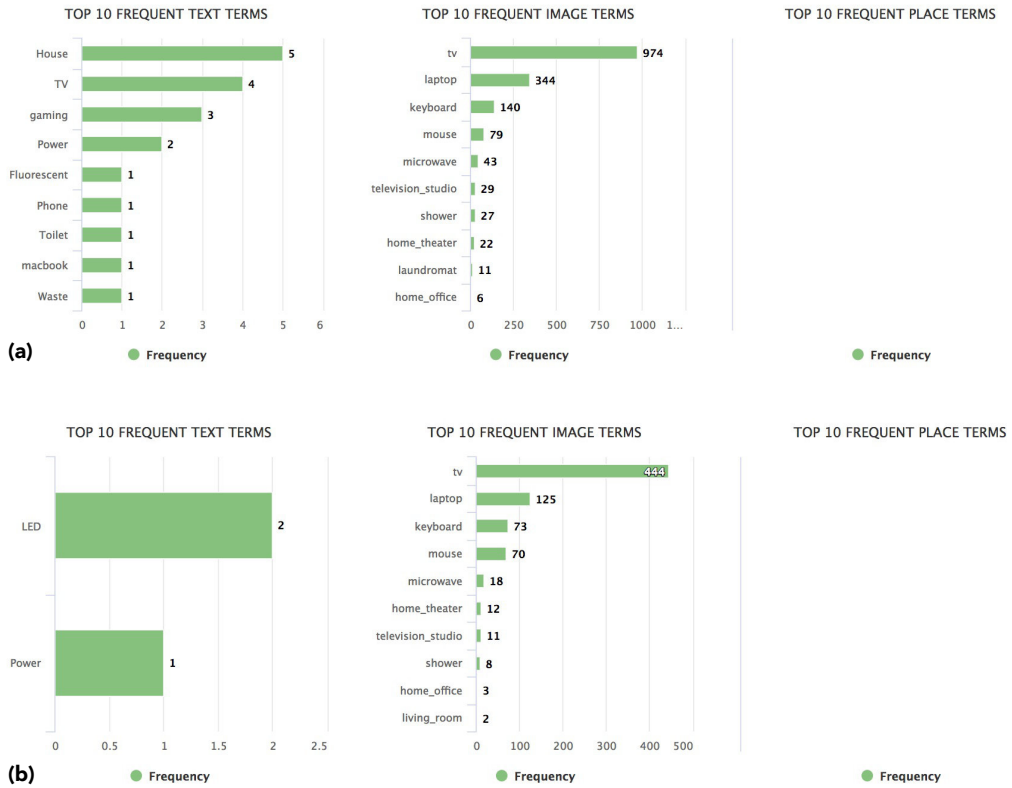

Figure 31 Bar charts visualizing the terms that occur most frequently in social media posts classified as dwelling activities in Amsterdam (a) and Istanbul (b). For the sake of legibility, the figures show English terms only.

As figure 30c shows, the highest concentration of food energy-consuming activities in Amsterdam is in the city center. Figure 30d, on the other hand, shows how in Istanbul these activities

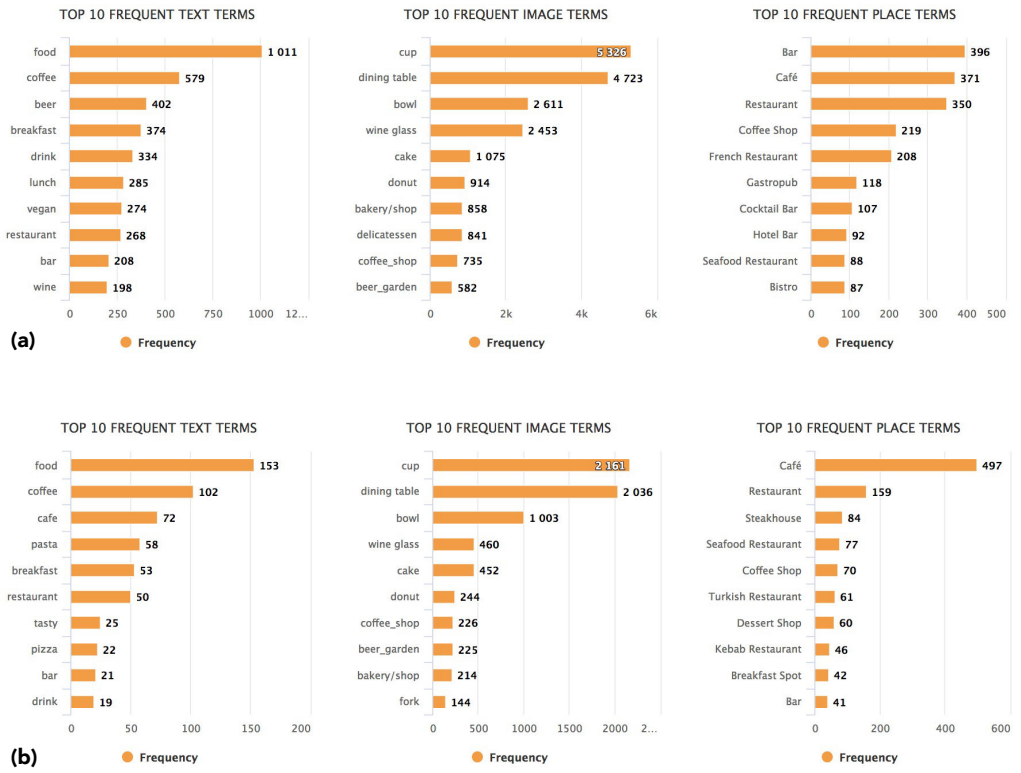

Figure 32 Bar charts visualizing the terms that occur most frequently in social media posts classified as food consumption activities in Amsterdam (a) and Istanbul (b). For the sake of legibility, the figures show English terms only.

peak in the Beşiktaş district and northern neighborhoods. Figure 32 shows that in both cities *food* and *coffee* were the most frequent text terms indicating a food consumption activity. Besides these, individuals appear to create posts related to food consumption most often while visiting a 'Bar' (Amsterdam), 'Café' (both cities), or 'Restaurant' (both cities).

In figure 30e, we notice that in Amsterdam the distribution of social media posts classified as corresponding to leisure activities seems to be broadly distributed over different neighborhoods. The city center (Burgwallen-Nieuwe Zijde), for instance, is frequented by numerous tourists, who socialize and drink; visit flower markets and museums; or enjoy the canals. This is reflected in the most frequent terms: *night*, *holiday*, *party* (text), *Flower Shop*, *Art Museum*, and *Hotel* (place). However, in the

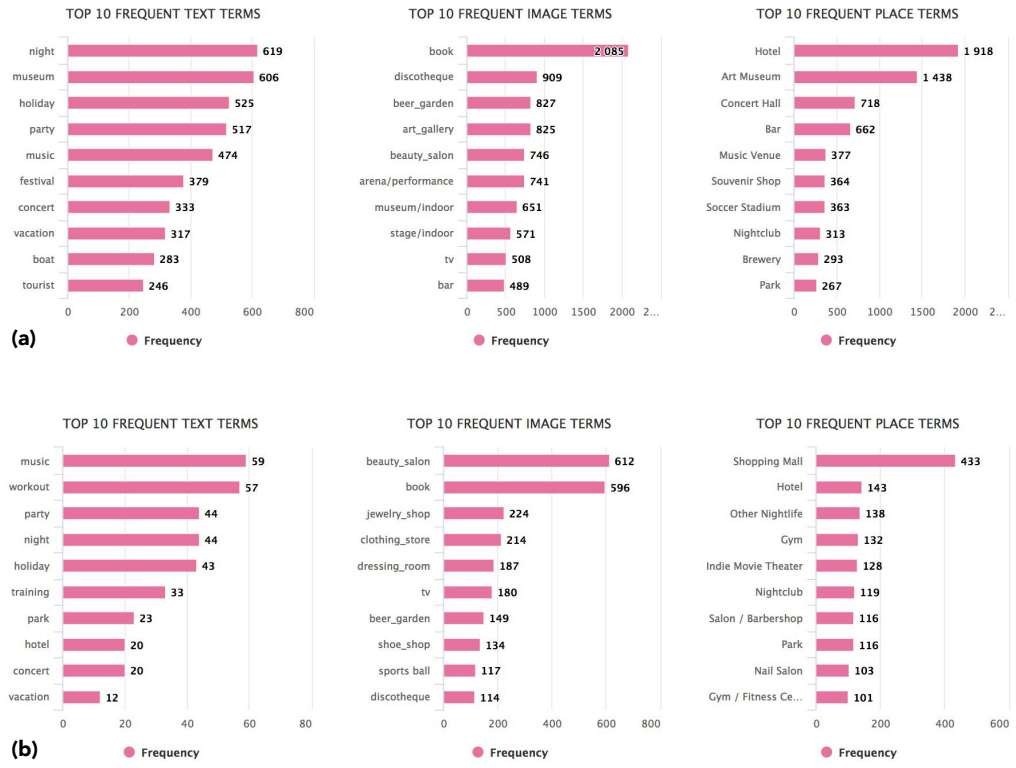

Figure 33 Bar charts visualizing the terms that occur most frequently in social media posts classified to leisure activities in Amsterdam (a) and Istanbul (b). For the sake of legibility, the figures show English terms only.

Museumkwartier, where the most famous museums are located, we found terms such as 'museum' (text), 'art_gallery' and 'museum/indoor' (image), and 'Art Museum' (place).

The distribution of the leisure-related social media posts over Istanbul's districts (figure 30f) is rather similar to that of food consumption-related posts. They are densest in the center and west (the Başakşehir district, which is also the location of the homonymous soccer team's stadium). Interestingly, figure 33 shows that in Istanbul the majority of leisure activities seem to take place in shopping malls.

Regarding mobility activities, in Amsterdam they are concentrated in the city center, where the central station is situated. Another reason this that people tend to post about canal trips, which dock there. In Istanbul, mobility activity is densest in two neigh-

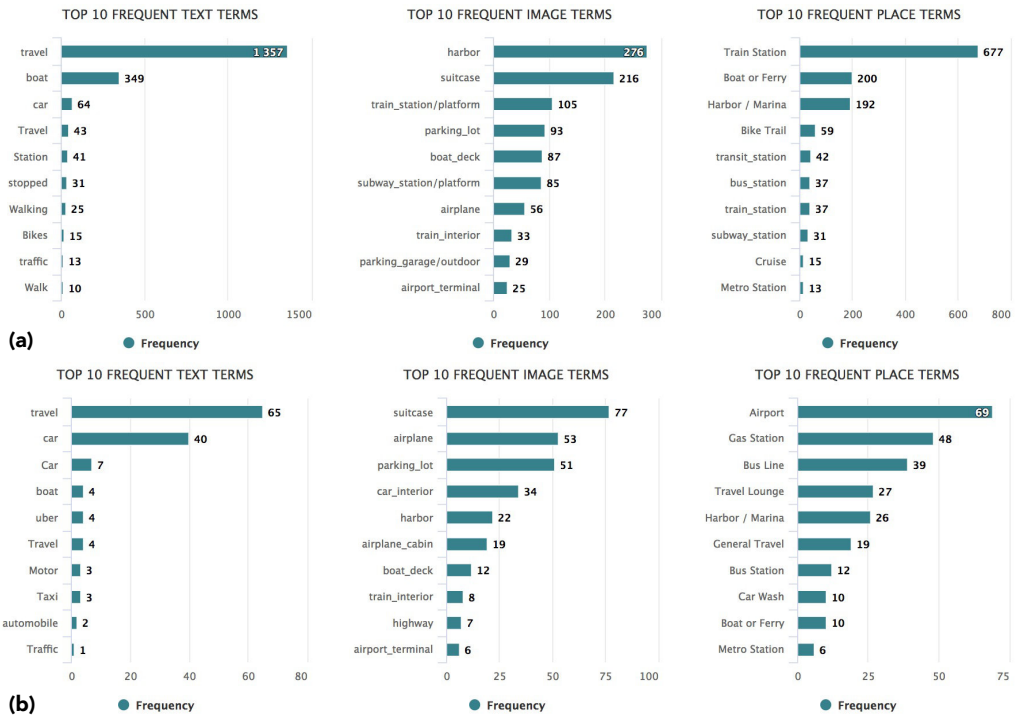

Figure 34 Bar charts visualizing the terms that occur most frequently in social media posts classified to mobility activities in Amsterdam (a) and Istanbul (b). For the sake of legibility, the figures show English terms only.

borhoods, Başakşehir and Eyüp. Multiple highways run through these districts (particularly Eyüp, which connects the Black Sea to the Golden Horn) as does a large highway junction. Looking at the terms (figure 34), we notice that Istanbul features more terms related to transportation by car (e.g. Gas Station, Car Wash, parking lot, car, etc.).

The framework captured few social media posts referring to dwelling activities in either city. This may be because social media users do not consider their regular domestic activities interesting enough to share with other social media users. Even if we look at posts related to food consumption, they appear to occur outside the home. As one might expect of social media, the majority of posts belong to the leisure energy-consuming activity. Moreover, typically people do not post directly content about their mobility activities, although we can use the distance be-

tween posts to detect whether a transportation activity was performed.

Amsterdam and Istanbul present similar ratio of energy-consuming activities, but across a different spatial distribution. This is probably due to the two cities' different features: Amsterdam has a well-defined center where the main venues are concentrated, while the main venues in Istanbul, given its different size, are scattered throughout the city. By looking at the terms that occur most frequently, we notice a small difference in how energy-consuming activities are characterized in the two cities. With regard to the food category, we see place categories that are more closely related to Turkish cuisine (e.g. Turkish restaurant and kebab restaurant) and that many leisure activities in Istanbul seem to take place in shopping malls. Finally, with regard to the mobility category, we notice a higher occurrence of terms related to transportation by car in Istanbul.

In sum, our pipeline can detect more practices that fall in the broad category of indirect energy-consuming activities. As we mentioned in the introduction, these activities are related to the production, transportation, and disposal of various consumer goods and services. As might be expected, people often post on social media when they are going out, whether to drink and dance or enjoy a special dinner. Only rarely do they share their domestic activities. This is not a flaw to our approach: rather it suggests that social media can indeed be used as a *complementary* source of information regarding energy-consuming activities, best used in combination with others. In fact, domestic activities are already partially captured by traditional data sources, while the indirect activities are either neglected or require costly methods, which have low temporal resolution (e.g. surveys).

In the case of the CODALoop project, we believe that our study has demonstrated that social media is a valid complementary source of information for understanding peoples' energy lifestyles. Social media can be used as additional source, along with

data gathered in Chapter 3. In the living lab described in Chapter 5 and the intervention described in Chapter 2, social media can be used to spark additional discussions about sustainable lifestyles. Indeed, they provide additional insights into the energy lifestyles of the community members involved. Social media could show unexpected aspects of a person's energy lifestyle, for instance.

We acknowledge that our approach has limitations. Social media are inherently biased: they are used by only part of a population (e.g. youngsters, tourists, etc.) for purposes quite different to that of sharing energy-consuming activities. This is reflected by the low volume of social media posts related to the dwelling activity and the prominence of leisure and food consumption categories. A study of demographic representation, however, exceeds the scope of this research. We leave that to future work.

Information shared on social media it is often ambiguous and noisy (e.g. a picture of a tram does not mean that the user is traveling). This is partially mitigated by our rule-based approach, which has proven very promising. Language can be an issue when applying our method to areas in which English is not the native language. However, this is addressed with multi-language dictionaries and by the use of embeddings trained on the main language spoken in the relevant area (e.g. Dutch for Amsterdam and Turkish for Istanbul). In addition, this problem only concerns the analysis of the text in social media posts, not images or locations.

CONCLUSION AND FUTURE WORKS

In this chapter we have demonstrated social media's potential to serve as a complementary source of information with which to describe energy-consuming activities. We foresee several possible research directions that might stem from this work.

ADDING OTHER SOCIAL MEDIA SOURCES

In our research, we focused on Twitter and Instagram. Although these are the most highly used social media to provide easy to access to their API, they are also biased and noisy. There are other, more specialized social media, on which users post about only specific types of activity. Here we briefly discuss how Spotify and Steam could be used as additional source for social media activity.

7. Steam[15] is a gaming social platform for PCs. Having started as a digital platform for distributing games, it has evolved into a social media platform. Users have their own profile pages, which contain statistics about that games they have played. Steam also provides an API through which to access this information. Hence, we might use this data to understand when a user is playing a videogame and classify the activity as *leisure* and *dwelling* (since gameplay is most probably performed at home).

8. Spotify is a streaming platform, providing music and podcasts from record labels and media companies. It provides an API, which makes it possible, upon previous authorization, to access information about users' devices and music. With this information we can understand when users are listening to music and classify the activity as *leisure*. Moreover, if they are using Spotify on a desktop PC, we can also classify the activity as *dwelling*, since it is an activity performed – most probably – at home.

ENRICHING ACTIVITY DESCRIPTIONS WITH ENERGY CONSUMPTION AND CO2 VALUES.

Though this work we obtained a qualitative description of energy consuming-activities. For instance, we established that a user had dinner at a restaurant or that another played videogames on a

console, and so forth. Although this information is useful in trying to understand how people consume energy, it would be more effective for both validation and feedback purposes if estimated values of energy consumptions and CO2 emissions could be attached to these descriptions. We envision that this enrichment to be performed in the following ways:

- Should an appliance or tool be found by our pipeline, we can directly attach energy consumption information to it by looking at manufacturers' websites or user generated databases.[16,17]

- In the case of a mobility activity being found, we can infer the mode of transportation by looking at the distance among posts or querying external services, such as GoogleMaps direction API.[18] Once we know the mode of transportation, we can attach information about energy consumption by looking at external databases.[19]

- Should a food consumption activity be found, it is possible to estimate the impact of the energy consumption by looking at the type food and where the activity takes place. Thanks to our ontology, we can consider not only how is cooked (e.g. fried, baked, etc.), but also grasp the chain of production as a whole. One study (Notarnicola, Tassielli, Renzulli, Castellani, & Sala, 2017), identified 17 groups of products and their impact on the environment. Using their findings, it is possible to associate a food consumption activity with its environmental impact. Should the type of food not be detected, we can use information about the location where the activity is registered. Extrapolating from the type of venue (e.g. Italian restaurant, vegan restaurant, bar, etc.), and the type of food that is typically served, we can infer the energy impact of the registered activity.

The main challenge presented by this step is that of handling uncertainty, for it builds on previous steps that yield estimates, not exact values (which is impossible). Moreover, there are many

variables that cannot be taken into the account by looking only at social media posts. These considerations include the model of the appliances being used, whether a car is electric or not, the exact duration of an activity, and so on.

We do not claim to be able to estimate energy consumption from social media with precision. Instead, we try to provide a more concrete value, which it would be presented alongside the information retrieved by the pipeline described in this chapter.

CLOSING THE LOOP, TOWARD A PERSONAL SOCIAL SMART METER.

In our use case studies, we analyzed social media posts at the scale of two cities, Amsterdam and Istanbul. By comparing these two cities, it is possible to highlight both similarities and differences in energy consuming activities. Moreover, citizens can compare energy-consuming activities at a neighborhood scale.

Our analysis can be shifted onto the user, who might focus on the activities performed by a single individual. In this way, the user can see the footprint of his or her own energy-consuming activities and compare it with that of his or her peers. The city-level aggregation performed in this chapter was initially chosen because testing our initial hypothesis required a large amount of data, which it is not possible to gather when focusing on a single person.

Given the conclusions established in this chapter, we now envision designing a platform that would implement a process comprising the following steps:

- Users provides access tokens to their social media accounts, logging in to our platform through the standard OAuth2.0 protocol.[20]

- With these tokens, we access various information about the users, depending on the social network under consideration. We retrieve their profiles, previous posts and other activities, and friendship network.

- We reuse the data processing pipeline described in this chapter to classify users' posts according to any of the four types of energy-consuming activity.

- By combining the information extracted, we build a profile of the users' based on their energy-consuming activities. On the basis of their profiles, users can visualize and compare their activities with those of other individuals (e.g. people living in the same area or friends, if they also provided their tokens). Users can also validate the results of our analysis, providing direct feedback as their posts are processed.

Such a platform could be integrated with that described in Chapter 3 and in the processes presented in Chapters 2 and chapter 4. Moreover, it can be also used in gamification approaches (Albertarelli et al., 2018) to provide users with feedback.

NOTES

1 http://linkeddata.org/

2 https://github.com/matterport/Mask_RCNN/releases

3 https://github.com/CSAILVision/places365

4 http://places2.csail.mit.edu/index.html

5 http://conceptnet.io/

6 We use the vector representation of the tweets obtained by training a Doc2Vec model on the tweet corpus.

7 The dictionaries are available on the companion website: http://social-glass.tudelft.nl/social-smart-meter/#dictionary.

8 This value provides the best trade-off between precision and recall in our context

9 https://www.w3.org/TR/rdf-sparql-query/

10 https://github.com/mmihaltz/word2vec-GoogleNews-vectors

11 https://dumps. wikimedia.org/nlwiki/20150703

12 http://hdl.handle.net/2066/151880

13 http://www.roularta.be/en

14 https://github.com/akoksal/Turkish-Word2Vec

15 https://store.steampowered.com

16 http://www.tpcdb.com/

17 https://www.energyefficiencydatabase.com/index.html

18 https://developers.google.com/maps/documentation/directions/start

19 https://www.bts.gov/content/energy-consumption-mode-transportation

20 https://en.wikipedia.org/wiki/OAuth

REFERENCES

Abbar, S., Mejova, Y., & Weber, I. (2015). You tweet what you eat: Studying food consumption through twitter. Paper presented at *Proceedings of the 33rd Annual ACM Conference*

on Human Factors in Computing Systems, 3197-3206. Seoul: ACM.

Albertarelli, S., Fraternali, P., Herrera, S., Melenhorst, M., Novak, J., Pasini, C., … Rottondi, C. (2018). A Survey on the Design of Gamified Systems for Energy and Water Sustainability. *Games, 9*(3), 38. https://doi.org/10.3390/g9030038

Backhaus, J., Breukers, S., Mont, O., Paukovic, M., & Mourik, R. (2013). *Sustainable Lifestyles. Today's Facts and Tomorrow's Trends. D1.1 Sustainable lifestyles baseline report.* Engery Amsterdam: Research Centre of the Netherlands.

Banerjee, S., & Pedersen, T. (2002). An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. Paper presented at *CICLing 2002: Computational Linguistics and Intelligent Text Processing*, 136-145. Mexico City: Springer.

BBC Food Ontology. (2014, March 18). Retrieved from https://www.bbc.co.uk/ontologies/fo

Bocconi, S., Bozzon, A., Psyllidis, A., Bolivar, C. T., & Houben, G. J. (2015). Social Glass: A Platform for Urban Analytics and Decision-making through Heterogeneous Social Data. Paper presented at *WWW '15 Companion Proceedings of the 24th International Conference on World Wide Web*, 175-178. Florence: ACM.

Bodnar, T., Dering, M. L., Tucker, C., & Hopkinson, K. M. (2017). Using large-scale social media networks as a scalable sensing system for modeling real-time energy use patterns. *IEEE Transactions on Systems, Man, and Cybernetycs: Systems*, *47*, 2627-2640.

Burel, G., Piccolo, L. S., & Alani, H. (2016). EnergyUse - A Collective Semantic Platform for Monitoring and Discussing Energy Consumption. Paper presented at *International Semantic Web Conference*, 257-272. Kobe: Springer.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. Paper presented at *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining Pages*, 226-231. Portland, OR: ACM.

Fard, M. A., Hadadi, H., & Targhi, A. T. (2016). Fruits and vegetables calorie counter using convolutional neural networks. Paper presented at *DH '16 Proceedings of the 6th International Conference on Digital Health Conference,* 121-122. Montreal: ACM.

Fernandez-Lopez, M., Gomez-Perez, A., & Juristo, N. (1997). METHONTOLOGY: From Ontological Art Towards Onto-

logical Engineering. Paper presented at *Proceedings of the Ontological Engineering AAAI-97 Spring Symposium Series,* 33-40. Stanford, CA: AAAI.

Fischer, C. (2008). Feedback on household electricity consumption: a tool for saving energy? *Energy Efficiency*, *1*(1), 79-104.

Fried, D., Surdeanu, M., Kobourov, S., Hingie, M., & Bell, D. (2014). Analyzing the language of food on social media. Paper presented at *IEEE International Conference on Big Data (Big Data),* 778-783. Washington, DC: IEEE.

Froehilich, J., Larson, E., Gupta, S., Cohn, G., Reynolds, M., & Patel, S. (2011). Disaggregated End-Use Energy Sensing for the Smart Grid. *IEEE Pervasive Computing*, *10*(1), 28-39.

Havasi, S. R. (2012). Representing General Relational Knowledge in ConceptNet 5. Paper presented at *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012),* 3679-3686. Paris: European Languages Resources Association.

He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask R-CNN. Paper presented at *2017 Proceedings of the IEEE International Conference on Computer Vision (ICCV),* 2961-2969. Venice: Springer.

Kamilaris, A., Pitsillides, A., & Fidas, C. (2016). Social Electricity: a case study on users perceptions in using green ICT social applications. *International Journal of Environment and Sustainable Development, 15*(1), 67-88. doi: 10.1504/IJESD.2016.073336

de Kok, R., Mauri, A., & Bozzon, A. (2019). Automatic Processing of User-Generated Content for the Description of Energy-Consuming Activities at Individual and Group Level. *Energies - Open Data and Energy Analytics*, *12*(1), 15.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone froman ice cream cone. Paper presented at *SIGDOC '86 Proceedings of the 5th annual international conference on Systems documentation*, 24-26. Toronto: ACM.

Madrazo, L., Sicilia, A., & Gamboa, G. (2012). SEMANCO: Semantic Tools for Carbon Reduction in Urban Planning. Paper presented at *9th European Conference on Product and Process Modelling,* 899-908. Reykjavik: CRC Press.

Mauri, A., Calbimonte, J.-P., Dell'Aglio, D., Balduini, M., Brambilla, M., & Della Valle, E. (2016). TripleWave: Spreading RDF Streams on the Web. In P. Groth, E. Simperl, A. Gray, M. Sabou, M. Krötzsch, F. Lecue, … Gil, Y. (Eds.), *The Se-*

mantic Web- ISWC 2016. ISWC 2016. Lecture Notes in Computer Science, 140-149. Kobe: Springer, Cham.
Mauri, A., Psyllidis, A., & Bozzon, A. (2018). Social Smart Meter: Identifying Energy Consumption Behavior in User-Generated Content. Paper presented at *Companion Proceedings of the The Web Conference 2018,* 195-198. Lyon: ACM.
Niles, I., & Pease, A. (2001). Towards a standard upper ontology. Paper presented at *Proceedings of the international conference on Formal Ontology in Information Systems*, 2-9. New York, NY: ACM.
Notarnicola, B., Tassielli, G., Renzulli, P. A., Castellani, V., & Sala, S. (2017). Environmental impacts of food consumption in Europe. *Journal of Cleaner Production*, *140*(2), 753-765.
OWL 2 Web Ontology Language Document Overview. (2012, December 11). Retrieved from https://www.w3.org/TR/2012/REC-owl2-overview-20121211/
Parsa, A., Najafabadi, A., & Salmasi, F. R. (2017). Implementation of smart optimal and automatic control of electrical home appliances (IoT). Paper presented at *2017 Smart Grid Conference (SGC),* 1-6. Teheran: IEEE.
Rashidi, T. H., Abbasi, A., Maghrebi, M., Hasan, S., & Waller, T. S. (2017). Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. *Transportation Research Part C: Emerging Technologies*, *75*, 197-211.
Stevens, R. (2009). Travel Ontology. Retrieved from http://www.cs.man.ac.uk/~stevensr/ontology/travel.zip
Suárez-Figueroa, M. C., Gómez-Pérez, A., & Villazón-Terrazas, B. (2009). How to Write and Use the Ontology Requirements Specification Document. Paper presented at *OTM 2009: On the Move to Meaningful Internet Systems: OTM 2009*, 966-982. Vilamoura: Springer.
Torriti, J. (2017). Understanding the timing of energy demand through time use data: Time of the day dependence of social practices. *Energy Research & Social Science*, *25*, 37-47.
Tukker, A., Huppes, G., Guinée, J., Heijungs, R., Koning, A. d., Oers, L. v., . . . Nielsen, P. (2006). *Environmental Impact of Products (EIPRO) Analysis of the life cycle environmental impacts related to the final consumption of the EU-25.* European Commission, Joint Research Centre, Institute for Prospective Technological Studies.
Vassileva, I. N., Wallin, F., & Dahlquist, E. (2012). Understanding energy consumption behavior for future demand response strategy development. *Energy*, *46*(1) 94-100.

Weiss, M., Helfenstein, A., Mattern, F., & Staake, T. (2012). Leveraging smart meter data to recognize home appliances. Paper presented at *2012 IEEE International Conference on Pervasive Computing and Communications,* 190-197. Lugano: IEEE.

Zhang, Z., He, Q., & Zhu, S. (2017). Potentials of using social media to infer the longitudinal travel behavior: A sequential model-based clustering method. *Transportation Research Part C: Emerging Technologies*, *85*, 396-414.

Zhu, Z., Blanke, U., & Gerhard, T. (2016). Recognizing composite daily activities from crowd-labelled social media data. *Pervasive and Mobile Computing*, *26*, 103-120.