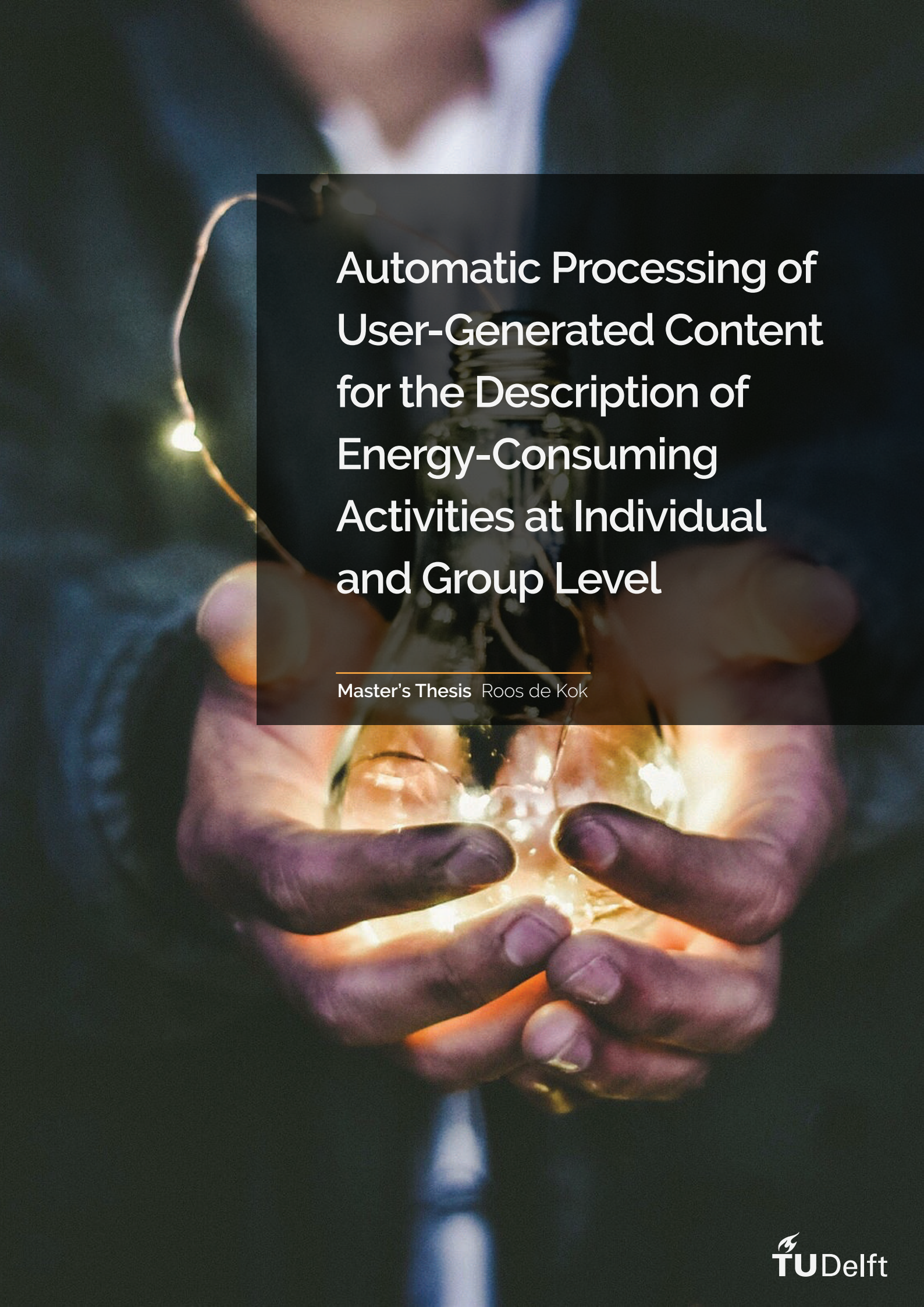


# Automatic Processing of User-Generated Content for the Description of Energy-Consuming Activities at Individual and Group Level

**Master's Thesis** Roos de Kok

TU Delft

# Automatic Processing of User-Generated Content for the Description of Energy-Consuming Activities at Individual and Group Level

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE
TRACK DATA SCIENCE & TECHNOLOGY

by

R.E. de Kok
born in Amsterdam, Netherlands

Web Information Systems
Department of Software Technology
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
http://wis.ewi.tudelft.nl

# Automatic Processing of User-Generated Content for the Description of Energy-Consuming Activities at Individual and Group Level


Author: R.E. de Kok
Student id: 4148320
Email: R.E.deKok@student.tudelft.nl


Thesis Committee:

| | |
|---|---|
| Chair: | Prof. dr.ir. G.J.P.M. Houben, Faculty EEMCS, TU Delft |
| University supervisor: | Dr. ir. A. Bozzon, Faculty EEMCS, TU Delft |
| University supervisor: | Dr. A. Mauri, Faculty EEMCS, TU Delft |
| Committee Member: | Prof. dr. K. Pfeffer, Faculty ITC, UTwente |

## Abstract

Our world population is increasing significantly resulting in a growing energy demand while the earth is running out if natural resources. This expanded energy demand directly affects climate change. Hence, there is an urgent need to move towards urban sustainability and to reduce our energy consumption [47]. This calls upon a behavioral change in energy consumption by the individuals (i.e., citizens) [31]. Social comparison, in the form of comparative feedback on energy usage with others, appears to be a more effective approach to stimulate energy conservation and efficiency than temporal self-comparisons [29, 70]. But before we can motivate people to change their energy consumption behavior, we need to have a thorough understanding of which energy-consuming activities they perform and how these are performed. Thus, insights into the individual's activities related to energy consumption should be gathered at a high-granular level.

Traditional sources of information about energy consumption, such as smart sensor devices and surveys, can be costly to set-up, may lack contextual information, have infrequent updates or are not publicly accessible. In this research, we propose to use user-generated content - and specifically, social media content - as a complementary source of information due to its rich and semantic nature. A huge amount of social media data is generated by hundreds of millions of people every day. These data sources are also publicly available and provide real-time data which is often tagged to space and time. Social media data also contains a lot of meta data, making it a good source for the recognition of energy-consuming activities performed by individuals [10, 91].

This thesis contributes the Social Smart Meter framework in order to meet the aim of automatically processing user-generated content for the description of energy-consuming activities, both at individual and group level. Four different categories of energy-consuming activities are distinguished: dwelling, food consumption, leisure, and mobility. To get a better understanding of the domain of energy-consuming activities, we contribute the Social Smart Meter Ontology (SSMO). This ontology forms the base for the data processing pipeline, which is developed in order to collect and enrich the data using several state-of-the-art techniques. Hereafter, the enriched data is classified to the different categories of energy-consuming activities using a dictionary- and rule-based approach, along with a classification confidence. To find ground truth and to evaluate the framework's performance, a user-based evaluation approach was used.

Furthermore, we contribute a Web-based application to support the analyses at group (i.e., city and neighborhood) level. Case studies are performed for the cities of Amsterdam and Istanbul, for which 275K social media posts are collected. The aggregated results are analyzed, providing more insights into the energy-consuming activities identified in the collected social media content. The majority of the classified social media posts refers to leisure activities. In addition, by examining for each post whether there exists a (significant) distance to the previous post created by this user, many mobility activities are inferred.

The case studies also contribute to the evaluation and discussion of the framework's performance; by analyzing its results, the framework's adherence to reality was discussed. Based on our preliminary results, it seems that using user-generated content has great potential as a complementary source of information for identifying and describing energy-consuming activities that are not yet captured by traditional data sources.

# Preface

Before you lies the Master's thesis "Automatic Processing of User-Generated Content for the Description of Energy-Consuming Activities at Individual and Group Level". It has been written to fulfill the graduation requirements of the MSc Computer Science at the Delft University of Technology. Writing this thesis engaged me from December 2017 to September 2018.

Climate change has been an emerging topic for several years now. I remember this already drew my interest during my second year of high school (more than 10 years ago), when we conducted a project on climate change. Nowadays, my generation is becoming more and more aware of the effects of climate change and our responsibility to take matters into our own hands. Yet, so much more awareness can still be created. Hence, I am really glad to have been working on such a topic for my thesis, in collaboration with the CODALoop project.

I would like to express my gratitude to my supervisors for their excellent guidance and support during this process. Fortunately, you were always available and willing to answer my questions. Thanks as well to the other members of the thesis committee for taking the time to read my report and attend the defence. I also wish to thank all of the respondents of my user-based evaluation surveys, without whose cooperation I would not have been able to conduct numerous analyses. And, of course, a special thanks to all my friends and family who supported me during this Master's thesis.

I hope you enjoy your reading.

R.E. de Kok
Delft, the Netherlands
August 21, 2018

# Contents

# List of Figures

# Chapter 1

# Introduction

The current world population of 7.6 billion keeps increasing; it is expected to reach 8.6 billion in 2030, 9.8 billion in 2050 and 11.2 billion in 2100.[1] The earth is gradually running out of (its limited amount of) natural resources while energy demand is still increasing. Between today and 2040, the global energy demand will expand by 30%[2], which directly affects greenhouse gas emissions and climate change. For instance, the global mean of surface temperature has increased by 0.6°C since the start of the 20th century. This asks for a move towards urban sustainability, and specifically a reduction in energy consumption [47].

Europe's 2030 Energy Strategy targets a 40% cut in greenhouse gas emissions compared to 1990 levels, at least a 27% share of renewable energy consumption, and at least 27% energy savings compared with the business-as-usual scenario.[3] In order to meet this targets, energy policies and programmes should be formed and individuals (i.e., citizens) should be motivated to change their energy consumption behavior [31], both in terms of energy conservation and energy efficiency. Energy efficiency involves using less energy to provide the same service; for instance, replacing a single pane window in the house with an energy-efficient one. On the other hand, energy conservation involves saving energy by reducing or omitting an activity; for instance, turning a light off or reducing the time one watches television. This calls upon a behavioral change by the individual. Energy consumption is significantly characterized by behavioral aspects [89], which is why understanding and changing the energy consumption behavior of individuals is considered as a powerful approach to improve energy conservation and stimulate energy efficiency at the individual level [70].

Multiple studies have examined how energy efficiency and conservation could be motivated among policy makers and citizens. In [29] the author explains how comparative feedback on energy usage with others can generate feelings of competition, social comparison or social pressure, which appears to be more effective in motivating energy conservation than temporal self-comparisons. The author of [44] endorses this in his Social Electricity case study, which "allows people to compare their energy footprint with other online peers or with the consumption at their

[1] https://www.un.org/development/desa/en/news/population/world-population-prospects-2017.html
[2] https://www.iea.org/weo2017
[3] https://ec.europa.eu/energy/en/topics/energy-strategy-and-energy-union/2030-energy-strategy

neighbourhood, village or town, to perceive if their own consumption is low, average or high". Multiple energy saving applications [31] have been developed yet, using visualized consumption feedback and gamified social interactions to motivate people to adopt energy-efficient lifestyles.

Before we can motivate individuals to change their energy consumption behavior, we need a thorough understanding of why and how (through which activities) they consume energy. In order to thoroughly understand this energy-consuming activities, insights into the individual's activities behind the energy consumption should be gathered at a high-granular level.

## 1.1 Traditional Data Sources for Understanding Energy-Consuming Activities

These days, multiple data sources are used to provide insights into energy-consuming activities, including (governmental) energy consumption surveys, smart meters, and smart plugs. Nevertheless, these mainly focus on residential energy-consuming activities (i.e., activities within the home) without taking external activity (activities outside the home) into account. Hence, most of the scientific research in energy-consuming activities has studied consumption at the household [70, 75, 89] or building [21, 73] level.

Smart sensor data (derived from smart meters and smart plugs) can be used to provide insights into domestic energy consumption. It focuses on aggregate energy consumption; however, by using specific techniques, insights into disaggregated end-use energy data (down to the individual appliance or device, which can be related to the individual's domestic activities) can be provided, which helps individuals understand how energy is consumed in the home. These techniques involve discriminating between appliances based on the total power consumed by each device, or differentiating by different features (in current wave forms or transient voltage noise signatures) during the device start-up [33, 67, 82]. However, since the energy usage is affected by a lot of variables, such as the number of appliances that are active simultaneously, it is questionable how well this approach reflects the actual energy usage [33].

In addition, survey or questionnaire research [11, 78, 81] is often used to gather insights into energy-consuming activities. Similarly to smart sensors, these data sources merely focus on the domestic energy consumption. Through a variety of questions, the energy-consuming activities is broken down into different end-uses. Yet, these data sources are not readily (i.e., real-time) available, since the surveys are conducted periodically (e.g., monthly or annually). This makes it very hard to model energy-consuming activities at a high-granular level, which is needed for a thorough understanding of the individual's energy-consuming activities.

Moreover, these numbers derived from smart sensor and survey data do only take direct energy usage into account. Indirect energy usage (i.e., energy usage that is "related to the production, transportation and disposal of a variety of consumer goods and services", such as "the availability of meat or cheese" [3]) is neglected [19]. Yet, conservation and efficiency of indirect energy consumption may also lead to a significant reduce in energy demand and should thereby also be integrated in the

energy-consuming activities models. Analysis of ecological (and carbon) footprint does take indirect energy usage into account though. For instance, carbon footprint calculators (e.g., the ones by WWF[4] or The Nature Conservancy[5]) estimate an individual's environmental footprint by assessing living habits in the domains of home, food, travel, and shopping. However, these footprints are not standardized yet and the calculators lack consistency. This leads to different results among the different calculators depending on the methodology, assumptions and data used by each carbon footprint calculator [25]. Besides the calculators, scientific models for ecological footprint analysis have been proposed. Yet, these only perform analyses at the city level and do not model individual footprints [34, 57].

Opposed to smart sensor and survey data sources, which focus on the domestic energy consumption, there are a lot of data sources that might provide relevant information about outdoor energy-consuming activities (such as supermarket data on daily groceries, financial transactions, etc.). However, these are not publicly accessible.

## 1.2 Social Media Data Sources as Complementary Sensors for Energy-Consuming Activities

Among different sources of user-generated content, social media data sources (including online social networks) arise as an alternative and novel approach, and typically do not face the above-mentioned issues that traditional data sources are coping with. Hundreds of millions of people frequently use social media to share, communicate, connect, interact, and create user-generated data, which makes social media an extraordinary source of big data [84]. Social media data sources are publicly available, and have lower setup and maintenance costs compared to certain physical sensors such as smart meters. Furthermore, real-time data can be obtained (i.e., they have a frequent update rate), and most social media data is tagged to space and time (hence, they include spatial and temporal dimensions) and allows for analysis at the individual (and group) level. In general, social media data contains a lot of meta data (about the post, user, checked-in location, etc.) and is thereby (semantically) very rich, which aids in extracting meaningful information out of it.

Because of its rich and dynamic data, social media has proven to be a good source for human (daily) activity recognition [10, 91]. If we consider the user to be a sensor, his/her corresponding social media data (in the form of textual or visual content, geolocation, and time) are signals that can be utilized for recognizing main activities. Since social media allows users to create posts about both domestic and outdoor activities, we might be able to not only capture information about the individual's domestic energy-consuming activities but also about his/her energy-consuming activities outside the home. Hence, social media data sources seem to be a good complementary source of information for describing energy-consuming activities.

However, social media data is generated by people with other intentions than creating information about their energy-consuming activities. Thus, the purpose of

[4] http://footprint.wwf.org.uk/

[5] https://www.nature.org/greenliving/carboncalculator/index.htm

social media data sources differs from the data sources that are designed to analyze energy-consuming activities, such as smart meters and plugs, and energy surveys. Moreover, daily life activities are rarely posted online at such high frequency, which makes it hard to get a complete overview of the individual's daily activities. In addition, users do not only create social media posts about their daily activities but also use social media to communicate with others, or express their feelings and interests. Thus, social media data often contains a lot of insignificant, irrelevant information, which makes it hard to separate signal (relevant content) from noise (irrelevant content) [12].

## 1.3 Problem Statement

A variety of data sources have been deployed to provide insights into individuals' energy-consuming activities. The traditional ones (smart sensor devices and energy surveys) merely focus on the energy-consuming activities at the domestic level; these only capture activity at the home. Outdoor activities related to energy consumption fall outside the scope of these data sources. The same applies for an individual's activity that involves indirect energy consumption (e.g., consuming a piece of meat during dinner); this activity can not be captured by traditional data sources either. In order to cope with those shortcomings, social media data sources could be used as a complementary source of information. Social media has yet been used to extract meaningful information about user behavioral patterns, such as food nutrition patterns [2, 32, 72], and user transport and activity patterns [10, 71, 91], which makes it a promising source to model outdoor energy-consuming activities at the individual level. Ultimately, a framework is desired in which both traditional and social media data sources are integrated, which allows us to gather insights into the individual's entire energy-consuming activities, at a high-granular level.

In order to gather those insights, we need a better understanding of the domain of energy-consuming activities, its main characteristics and all different instances (including the different types of activities) that are related to energy-consuming activities. We need to comprehend the meaning of those instances, and how they are mutually related. In addition, a definition is needed of how social media (instance) data relates to the physical world, and how it may reflect the individual's energy-consuming activities.

However, as mentioned before, multiple challenges will be faced when using social media as our data source for describing energy-consuming activities. One of the greatest challenges is to extract meaningful information from social media data. Social media data is often noisy; users might include shorthand or slang in their messages, words or expressions might vary in meaning depending on the context, and content might contain information that stretches over multiple locations [41]. Since social media data contains a lot of irrelevant information, it is hard to separate the signal (relevant information) from the noise (irrelevant information). Moreover, social media data is often biased, which makes it very important to understand the degree and nature of this bias. For instance, social media platforms are particularly used by the younger generations (i.e., selection bias) [41]. Besides that, a user's social media post might not always be a reflection of his/her daily activities in the physical

world, due to multiple semantic ambiguities and discrepancies that may be encountered. A user may create a post before or after an activity instead of during the activity itself - e.g., a user recalling an old memory of a concert from a while ago. In Figure 1.1 the user wrote a message about the tramway in Amsterdam, along with a picture of a tram; however, this does not necessarily mean that the user indeed travelled by (this) tram. He might have wanted to capture a typical Amsterdam street view instead. Also, in the case of Figure 1.1, it is clear that the mention of Amsterdam in the message refers to the capital of the Netherlands, since the user checked in there. Amsterdam is also a town in the state of New York (United States of America) though. Hence, it might also be the case that a user–when he/she has not checked into a location–refers to the town of Amsterdam in the state of New York instead of the Dutch capital (ambiguity in location). Besides that, words from the text message might be ambiguous as well. For instance, the word “park" has different meanings dependent on the context, as shown in Figure 1.2. It can either be used in the context of parking vehicles (related to mobility) or to indicate a recreational park (“a piece of ground in or near a city or town kept for ornament and recreation"[6], related to leisure). It is very important to be aware of these types of social media bias and semantic ambiguities and discrepancies when analyzing and evaluating the data.


Figure 1.1: Example of ambiguity in content

[6]https://www.merriam-webster.com/dictionary/park


(a)


(b)

Figure 1.2: Different meanings of the word “park"

## 1.4 Research Aim, Objectives and Scope

Given the challenges mentioned in the previous section, the aim of this work is to design a framework for the analysis, integration, and visualization of social media data to facilitate the understanding of individual energy-consuming activities.

### 1.4.1 Scope and Granularity of Energy-Consuming Activities

The scope of this work focuses on four categories of energy-consuming activities; based on previous work [7, 36, 64] - and in the context of the CODALoop[7] project - we define the following types of activities: dwelling, mobility, food consumption and leisure. Activities related to industry - e.g., the individual being at work - are not taken into account.

Furthermore, energy-consuming activities at the individual and group level are studied. Since the overarching goal in sustainability studies is to move towards urban sustainability (as mentioned in the introductory background section), we have mainly focused on urban individuals (and thereby, urban areas) within the target areas. Information gathered on the individual level has been aggregated to perform analyses at group level and provide insights into particular neighborhoods (or other (urban) areas).

With respect to the data sources, this work has been limited to data collection through Twitter and Instagram, mainly due to their public APIs. A more thorough analysis of the different data sources can be found in *4.3.1 Identifying Promising Data Sources*.

### 1.4.2 Research Questions

Our main research question follows from our research aim:

**MRQ: How can we automatically process user-generated content to describe energy-consuming activities at individual and group level?**

To answer this overarching research question, four sub-questions have been posed (in which “RQ" depicts a research sub-question and “C" a contribution):

**RQ1: How are energy-consuming activities studied by the state of the art?**

In order to determine the best methods and tools for describing and understanding user energy-consuming activities, a literature review (C1) is conducted to explore the state of the art in this field. Previous studies have yet explored several methods to provide insights into energy-consuming activities, though very generic, or with a focus on a single domain of energy-consuming activities [3, 19]. The strengths and weaknesses of the existing contributions were identified to lay the groundwork for this research in order to determine which methods and tools should be included in our framework.

[7] https://jpi-urbaneurope.eu/project/codaloop/

**RQ2: What are the main characteristics of energy-consuming activities?**

For a better comprehension of the domain of energy-consuming activities at the individual and group level, an ontological representation (based on a conceptual data model) is developed, including all four categories of activities (dwelling, mobility, food consumption, and leisure), that will facilitate the definition of the main characteristics of an individual's energy-consuming activities. On the other hand, it supports the definition of how social media data may refer to activities performed in the physical world. This results in the contribution of the Social Smart Meter Ontology (C2), an ontological representation of the domain of energy-consuming activities.

**RQ3: How can we extract the characteristics of energy-consuming activities from social media data?**

A data processing model (or pipeline) is developed that allows to extract the characteristics of energy-consuming activities from the social media data. This pipeline includes multiple components: (i) the data collection (and pre-processing) from the social media data sources, (ii) different steps of data enrichment, and (iii) a dictionary- and rule-based classification model that outputs to which categories of energy-consuming activities social media posts are classified. Here, we will also cope with many of the challenges that arise from social media data. In this work we contribute both an analysis of (social media and enrichment) data sources (C3A) and a data processing pipeline (C3B) to extract the characteristics of energy-consuming activities from social media data.

**RQ4: To what extent can social media be used as a complementary data source to describe energy-consuming activities?**

In order to understand to what extent the extracted characteristics of energy-consuming activities provide insights into an individual's pattern of energy-consuming activities, data exploration and visualization methods will be used to extract meaningful information from the data, e.g. in different dimensions such as space and time. The framework is evaluated through a case study performed for the social media activity in the cities of Amsterdam and Istanbul. The results are visualized through a Web application (C4) for the analysis of energy-consuming activities at group level. The framework's performance is evaluated through multiple evaluation metrics such as accuracy, precision and recall, and the F1-score.

## 1.5 Research Design: Approach and Methods

In Figure 1.3 our research design is illustrated using a schematic overview to visualize the research design structure and thesis outline. Each sub-question links to an action, which is represented by one of the numbered circles, comprising the approach and methods for that particular part. The number within the circle refers to the number of the corresponding chapter. The actions to be taken are (2) providing an overview of the *state of the art* in describing energy-consuming activities, (3)

defining the main *characteristics* of an individual's energy-consuming activities, (4) *extraction* of the characteristics of energy-consuming activities from social media data, and (5) *understanding* to what extent social media can be used as a complementary data source to provide insights into the individual's energy-consuming activities. The outcomes and findings of all parts together lead to the answer to the main research question.

In the first part of the research design a literature study (and review) has been conducted to explore the state of the art in energy-consuming activities. Key findings, strengths and weaknesses of existing studies are compared and analyzed in order to identify which methods and tools are most relevant and promising and should be included in our framework.

The second part introduces a conceptual data model of energy-consuming activities and its characteristics. At first, entity-relationship modeling has been the starting point for this conceptual data model. At a later stage this model was transformed into an ontological representation.

In the third part a variety of methods and tools has been used to design the different components of our data processing pipeline, which is designed to extract the characteristics of energy-consuming activities from social media data. The data has been collected trough the public APIs of Twitter and Instagram. These data sources have been selected based on a structured analysis of multiple (social media) data sources. A variety of state-of-the-art machine learning techniques was used to apply a multiple enrichment steps. Once the data was enriched, it was classified using a dictionary- and rule-based approach. In addition, a confidence was assigned to each classification. The performance of the framework was evaluated based on several metrics, such as precision and recall.

In the fourth part the framework has been evaluated through a case study. Test data from Amsterdam and Istanbul was acquired through several social media APIs. Then, our framework was evaluated by testing it on the acquired data, to find out to what extent social media data sources can serve as sensors to understand energy-consuming activities. Different dimensions were taken into account; temporal and spatial patterns were examined at group level along with a comparison between the different categories of energy-consuming activities (dwelling, mobility, food consumption, and leisure). Several data exploration and visualization techniques have been used to present the results.

## 1.6 Thesis Outline

Each of the six parts of the research design (Figure 1.3) is affiliated with a chapter of the thesis. Chapters 2 through 5 each address one of the four research sub-questions. The main research question is answered in Chapter 6, as well as a discussion following from the answers to each sub-question.

- Chapter 1 discusses the background (regarding energy-consuming activities) of the issue we aim to solve along with the aim, research questions (following up on the objectives), and scope of the research.
- Chapter 2 provides an overview of how the state of the art studies energy-consuming activities. Existing methods and tools, including its strengths and weaknesses, are reviewed.
- Chapter 3 introduces a conceptual data model of the domain of energy-consuming activities and its characteristics. It also sheds light on the ties between energy-consuming activities and social media activity.
- Chapter 4 proposes the design of a data processing pipeline, developed for the extraction of the characteristics of energy-consuming activities from social media data. Each of its components is discussed along with the framework's overall design. In addition, a structured analysis of relevant (social media and enrichment) data sources is provided.
- Chapter 5 explores and visualizes the extracted information for social media data from urban areas in Amsterdam and Istanbul in order to acquire meaningful information and insights into the energy-consuming activities at group level. In addition, the framework's performance is evaluated.
- Chapter 6 summarizes and discusses the findings for each sub-question, resulting in the explanation of our main research question. Subsequently, multiple possibilities for future research are suggested.

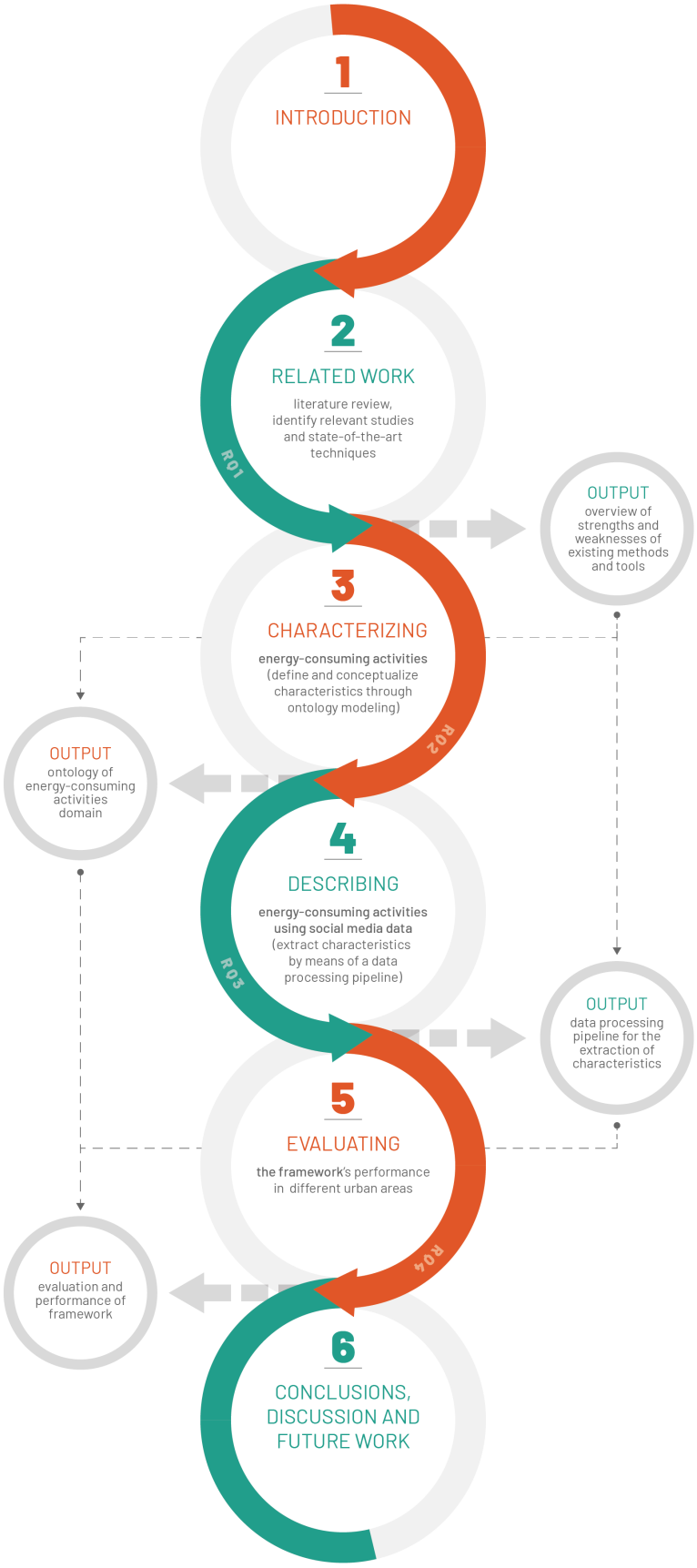





Figure 1.3: Schematic overview of the research design structure and thesis outline

# Chapter 2

# Related Work

In order to model the energy-consuming activities performed by individuals, we should have a clear understanding of what this actually entails. Energy consumption can be separated into direct and indirect consumption. Direct energy consumption is related to the use of gas, electricity and fuel, whereas indirect energy consumption is related to the production, transportation and disposal of a diversity of consumer goods and services [3]. Moreover, four categories of energy consumption can be distinguished based on previous work [7, 36, 64]:

**Dwelling** the energy consumption necessary for all activities performed within the home, peculiarly by using appliances available in the house.

**Food** the energy consumption necessary for all stages of the food chain (production, processing, distribution, consumption and waste).

**Leisure** the energy consumption necessary for citizens to achieve the "freedom provided by the cessation of activities"[1].

**Mobility** the energy necessary to bring one person from one place of activity to another place of activity.

Up to now, many studies have aimed to describe or model these energy-consuming activities. Most of the work relies on traditional data sources. Only a few studies include social media data sources to describe energy-consuming activities. However, many other studies have aimed to recognize user activities from social media data, though with a different purpose than providing insights into energy consumption. Therefore, we describe both how energy-consuming activities are studied in the state of the art, as well as how user activities are recognized from social media data.

## 2.1 Modeling Energy-Consuming Activities

As mentioned before, many studies rely on survey data, which is the basic collection method for information on energy consumption [77]. In [5] the authors proposes a

[1] www.merriam-webster.com

human-activity based residential energy modeling framework that can create power demand profiles considering the characteristics of household members, based on survey research data. The authors of [70] use home appliance energy use data for performing agent-based modelling (ABM) to represent the complexities of energy demand, such as social interactions and spatial constraints. In [11] the individual's environmental impact (assessed by different behavioral determinants of the ecological footprint) is compared with self-assessments of their own environmental impact, retrieved from Belgian survey data. Furthermore, the authors of [78] analyze national time use survey data to assess how dependent energy-related social practices in the household (preparing food, washing, cleaning, washing clothes, watching TV and using a computer) are in relation to the time of the day.

Besides surveys, smart sensor data (derived from smart devices such as smart meters and plugs) have also emerged as a valuable source of information. In [75] the authors demonstrate how household characteristics related to energy efficiency can be extracted from smart electricity meter data by using supervised-machine-learning-based techniques. The authors of [31] also use algorithms to extract activity from sensor data to profile different types of user behavior and infer activity context.

Nevertheless, these studies all rely on traditional data sources such as surveys and (smart) sensors and solely focus on the activity at the domestic level. Recently, some researchers have proposed the integration of social media data sources as an alternative to the traditional ones. However, these studies took only terms directly related to energy into account [74] or examined the correlation between external events (such as Christmas) discovered through social media and the actual energy consumption. Thus, both do not examine citizens' energy consumption *activity* discovered through social media. Furthermore, most studies perform top-down modeling of energy consumption: based on the total numbers, the energy consumption signals are disaggregated into different end uses [39, 52], whereas we aim for bottom-up modeling (i.e., aggregating the energy consumption based on the end uses discovered through social media posts by citizens).

## 2.2 User Activity Recognition from Social Media Data

Up to now, there have been rarely any studies that have focused on recognizing the actual user energy-consuming activities from social media data. In [12] Bodnar et al. propose a social media network-driven model that aims to approximate electricity utilization patterns from large-scale textual and geo-spatial social media data. Through a Bayesian process, topics are modeled for user posts that are compared to events and phenomena in the physical world; physical hardware systems are excluded here. Since we envision traditional and social media data sources as complementary sources of information, a framework that integrates both is desired. Furthermore, "real world energy utilization" is not defined prior to the case study and is only analyzed at the household level.

Moreover, there have been many studies that have examined and modeled user activity patterns in different fields than energy consumption though. For instance, previous studies have looked into nutrition patterns [2, 4, 32] and activity and

mobility behavior [26, 53, 71, 87, 90]. These topics, nutrition, mobility and activities, can all be related to energy-consuming activities in some way. Nutrition is affiliated with food consumption; thereby, it is indirectly also associated with energy consumption. The same reasoning holds for mobility and activities; it is not possible to travel or to perform another activity without consuming energy. For that reason–despite the fact that the purpose of these studies is different than our aim to describe user energy-consuming activities–the findings of these studies are relevant for this research as well. For each type of energy-consuming activities (dwelling, mobility, food consumption, and leisure) the most important and relevant findings are discussed below and summarized in Table 2.1.

### 2.2.1 Dwelling

Many traditional data sources focus on residential energy consumption, which is affiliated to dwelling. Nevertheless, many residential activities are not exclusively related to dwelling but could also be associated with leisure or food consumption. For instance, cooking (at home) is an activity that is related to both dwelling and food consumption. Existing literature [5, 12, 31] does not distinguish these different types of energy consumption. Moreover, many studies examine the total residential energy consumption and do not disaggregate the consumption into different end uses, which makes it hard to obtain insights into the corresponding energy-consuming activities. Additionally, existing work barely incorporates social media data in their models and does not focus on the individual level.

### 2.2.2 Mobility

Nowadays, location-based social networks (LBSNs) allow users to share geo-tagged information along with the content (text, images, videos, etc.) they post. A check-in is one of the ways a user can share such geo-tagged information. In previous studies users' check-in behavior (and corresponding geolocated tweets) are analyzed for different purposes: to model travel demand and behavior [71, 87], to detect traffic oddities [66], to predict turning points in migration trends [85], or to predict user activities [53, 90]. In [46] all algorithms related to LBSNs are elucidated by means of a survey.

For this research, we are interested in the purposes of both modeling travel demand and behavior, and predicting user activities. The latter one will be discussed in more detail in the paragraph on leisure below. Regarding the travel behavior modeling, not only spatial-temporal information but also semantic information is utilized. The check-in data is often enriched with activity information, such as the category of the location or point of interest (POI), as seen in the approaches of [26, 53, 90].

In [71] the author describes that the mode of transport should be determined using text mining and natural language processing approaches (by constructing a dictionary). As for duration of the activity, either text mining or natural language processing approaches should be utilized, or information for this travel attribute should be extracted by considering multiple tweets about the same trip joined together as a chain of tweets. Information for these two travel attributes, duration and mode of transport, have not been extracted and considered yet in the previously

mentioned work. In the case that the consumed energy per mobility activity should be determined, these attributes do actually have to be taken into account.

### 2.2.3 Food consumption

Recently, a lot of progress has been made regarding food detection and recognition in text and images. Food recognition in text is somewhat less common nowadays; natural language processing in combination with topic modeling [32] or Naive Bayes classification in combination with *n*-gram matching [2] approaches have been used in previous work. With regard to food recognition in images, the use of convolutional neural networks (CNNs) occurs to be one of the most frequently used approaches [4, 22, 23, 45, 56, 72].

In [22] the author describes an approach in which neural networks are utilized to jointly consider food recognition, ingredient recognition, and cooking method recognition. Verb-noun pairs are generated that indicate both the type of food and how it was cooked. Considering all those recognition factors could ease the process of determining the amount of energy consumed for preparing such a dish that is shared by a user through an image in its social network.

Furthermore, the author of [40] shows how the performance of food recognition can be improved by integrating multiple evidences (visual, location, and external knowledge, such as prior knowledge about the restaurant's menu). Besides improving the performance of the model, these context factors could also contribute to a better understanding of the user's food consumption related activity and provide more insights about the corresponding amount of consumed energy.

In [2] food-related tweets were determined using a unigram Naive Bayes classifier, after which the most popular food-related terms were enriched with nutritional information in terms of the amount of calories per serving. Longest *n*-gram matching was performed to detect the foods in the tweet text and aggregate their caloric content. Since the amount of calories is related to the amount of consumed energy, this enrichment and matching approach could also be applied in this research.

### 2.2.4 Leisure

In existing work, user activities are not necessarily separated into leisure activities. Nonetheless, most activities that are studied are implicitly affiliated with leisure. There have been multiple studies [26, 53, 90] that look into activity prediction based on travel behavior and patterns, as mentioned before. Check-in records contain semantic information (e.g. category of the POI) along with the spatial-temporal information, which makes a user's check-in behavior convenient input to infer activities. Opposed to the previous studies, [10] also takes the activity duration into account, which might be relevant when the amount of consumed energy will be linked to the activity. The authors follow up on the approach of [91], in which an SVM classification model is used to assign each tweet to an activity.

| **Type** | **Papers** | **Achievements** | **Limitations** |
|---|---|---|---|
| Dwelling | [5, 12, 31] | Analysis at the residential level | Often only traditional data sources incorporated in the models; no disaggregation of energy-consuming activities; no analysis at the individual level |
| Mobility | [26, 46, 53, 66, 71, 85, 87, 90] | Analysis of check-in data, semantic enrichment of data | Mode of transport and travel duration not taken into account; not aimed at identifying the corresponding energy consumption |
| Food | [2, 4, 22, 23, 32, 40, 45, 56, 72] | Food recognition in text and images, ingredient and cooking method recognition in images, nutritional enrichment of data, enrichment of information about food-related venues | Not aimed at identifying the corresponding energy consumption |
| Leisure | [10, 26, 53, 90, 91] | Activity prediction based on travel behavior and patterns, enrichment of venue information (e.g., category) | Not aimed at identifying the corresponding energy consumption |

Table 2.1: General overview of the state of the art in recognizing user behavior and activity from social media data

## 2.3 Conclusions

Multiple studies have been performed with regard to the four categories that were were distinguished before. In the field of dwelling activities, merely traditional data sources are incorporated in order to provide insights. Only a few studies have yet incorporated social media data but these studies do not distinguish different categories of energy-consuming activities. Furthermore, many studies have analyzed users' check-in data to predict activity and mobility patterns; the corresponding mode of transport and travel duration are not taken into account in these studies though. As for the categories of food and leisure activities, a lot of research has been done as well. The state-of-the-art studies are able to extract a lot of information about both food (as well as the cooking method and ingredients, nutritional information or corresponding venues) from text and images, and about activities from text and location-based check-ins. Yet, these studies are not aimed at identifying any information about the corresponding energy consumption.

In this work, a lot of state-of-the-art techniques can be (re-)used to extract information from social media data in the fields of dwelling, mobility, food, and leisure, though in a different context, namely the one of identifying and describing energy-consuming activities.

# Chapter 3

# Characterizing Energy-Consuming Activities

What are energy-consuming activities? There are plenty of answers to this question. One could define energy-consuming activities in many ways, depending on the need and background context, inducing a wide range of varying viewpoints and assumptions regarding the relevant concepts and relations in this subject matter. For instance, one might consider energy consumption behavior and energy-consuming activities as equivalent concepts; yet, one could also reason that an individual's behavior encompasses "the way in which someone conducts oneself or behaves"[1] and thereby differs from activities, which are "behavior or actions of a particular kind"[2]. These differences in jargon, overlapping or mismatched concepts, structures and methods may lead to a poor communication between people and organizations in this field of interest. In the context of developing an IT framework, this lack of a shared understanding could complicate the identification of requirements and definition of a specification of the system [79]. For that reason, a conceptual data model of the domain of energy-consuming activities is aspired, which will form the basis for our domain knowledge representation [20]. It lays the groundwork for a shared understanding of the field of energy-consuming activities and facilitates the corresponding communication between people, organizations, and (software) applications.

A conceptual data model helps to (i) understand the domain of energy-consuming activities and (ii) identify relevant and important concepts and how these are interrelated, by providing terms for describing and representing our knowledge about this domain in a structured manner [20]; in other words, it helps to determine what the main characteristics of energy-consuming activities are, which we eventually aim to extract from our social media sources. Since an ontological representation enables a shared understanding of energy-consuming activities, this knowledge could also be easily shared with others who have similar needs for a knowledge representation in this domain. This interoperability and potential for re-use prevents that others have to re-invent the wheel of this knowledge-analysis process [20, 79]. Moreover, it is easier for ontological-based application to organize information; e.g., having categories and

[1] https://www.merriam-webster.com/dictionary/behavior
[2] https://www.merriam-webster.com/dictionary/activities

subcategories helps organizing the search. In addition, ontologies can help identify semantic categories that are involved in understanding the energy-consuming activities, by acting as a concept dictionary [20].

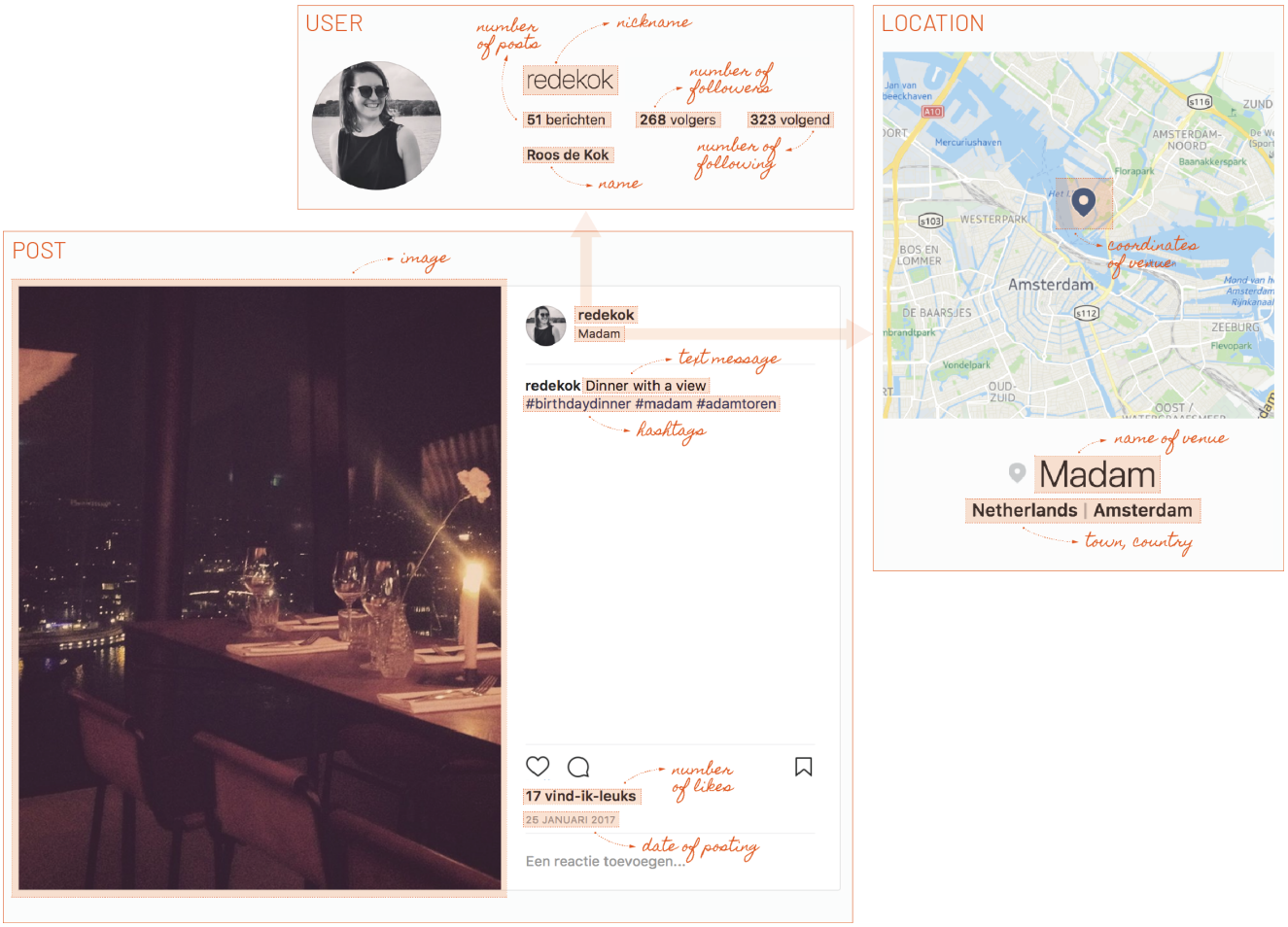



Figure 3.1: Example of social media post (Instagram)

Following up on our assumption that social media can be used as a sensor for an individual's daily activities [12], social media concepts should be included in the conceptual data model as well, by linking them to the relevant concepts of energy-consuming activities. Adding meaning to a user's social media data can help us understand to what extent these data sources reflect the individual's energy-consuming activities. But what kind of meaning should be added to social media data? What kind of information should be used to enrich the data? For instance, if we look at the social media post in Figure 3.1, the message (*Dinner with a view #birthdaydinner #madam #adamtoren*) indicates that the picture is taken by the user during dinner. In the image we can indeed identify a dinner table. Furthermore, the hash tags and the location, where the user has checked in, indicate that the dinner took place in Madam. By examining the meta data, we discover that Madam is a

venue, located in Amsterdam, the Netherlands. Moreover, we also have more information on the user, whose nickname (*redekok*) corresponds with the name Roos de Kok. Incorporating all this metadata, we know a whole lot more about this user's dinner activity than just the fact that she had a dinner.

Since a conceptual data model enables us to define these main characteristics of energy-consuming activities, it thereby facilitates the semantic enrichment of social media data. By linking a user's social media activity to the conceptual data model, we can describe if and how the social media data refers to the individual's energy-consuming activities. Moreover, a conceptual data model provides information to interpret the meaning (semantics) from the instances. For instance, Roos de Kok `<performs> a` dinner activity `<at>` Madam, `<located in>` Amsterdam, the Netherlands. In this way, the characteristics of an individual's energy-consuming activities can be defined.

## 3.1 Social Smart Meter Ontology

As mentioned before, social media data can be treated as a sensor that allows to recognize an individual's daily activity, including the activities that shape the energy-consuming activities. Hence, we can consider a social media user as an instance of an individual, who has a social media account. Thereby, a social media post by this user *may* reflect an energy consumption activity performed by this individual, at a certain location (Figure 3.2).

In order to get a better understanding of the domains of energy-consuming activities and social media activity, and to understand how the instances of both domains are related, conceptual data models of these two domains are required. By linking the two conceptual data models (as proposed in Figure 3.2), more insights could be provided into how social media could be used as a sensor for the performance of energy-consuming activities.

As conceptual data models and ontologies both consist of conceptual relations and rules, they are indeed quite similar. Hence, multiple researchers have yet proposed to (re)use those conceptual data models for ontology modeling, e.g. [8, 24, 30, 62]. Reusing conceptual modeling techniques for ontology engineering is beneficial for the following reasons: the large set of existing conceptual modeling methods, graphical notations, and tools can make ontologies better understandable, and makes it easier to adopt, construct, visualize, etc. [42].

### 3.1.1 Ontology Modeling

In this chapter we propose a novel ontology called SSMO (Social Smart Meter Ontology) able to encode and describe properties of social media activity reflecting energy-consuming activities. The design has been performed according to the *Methontology* guidelines presented in [28], including specification, knowledge acquisition, integration, conceptualization, implementation, documentation, and maintenance.

### 3.1.2 Specification

To scope the ontology specification, several elements are taken into account: the purpose, scope, intended uses, and actors of the ontology.

**Purpose** The Social Smart Meter Ontology aims to provide a better understanding of the domain of energy-consuming activities (categorized into different energy lifestyle domains) at the individual and group levels, and how this is related to individuals' social media activity. Hence, not only domestic but also outdoor activities are taken into account.

**Scope** Energy-consuming activities involve one or more of the four energy lifestyle domains (dwelling, food consumption, leisure, and mobility). Activities related to industrial activity (e.g., work) are not taken into account.

**Intended uses** Through a better understanding of an individual's energy-consuming activities, eventually a change in behavior (resulting in energy conversation and efficiency) can be pursued by governmental authorities and other global (energy) organizations. However, for such a change to be achieved, the ontology should be used as part of the process of understanding the domain of energy-consuming activities in order to extract the relevant information (in the form of entities) from social media, which should ultimately result in more insights into patterns of energy-consuming activities (at both individual and group level).

**Actors** Numerous stakeholders are involved, among which are: researchers, governments, municipalities, global (energy) organizations, and individuals (or citizens). Researchers are responsible for developing and enhancing ontological models such as this one. Governments, municipalities, and global (energy) organizations are the ones that are accountable for using ontological-based applications for developing energy policies and motivating a behavioral change among individuals.

**Requirements** To define what functionalities are required from the ontology, a set of requirements is included in the specification. Non-functional and functional requirements are distinguished. Non-functional ones are general requirements or aspects the ontology should accomplish, whereas functional ones are content specific requirements that the ontology should fulfil. Competency questions are used for this. We follow the methodological guidelines for specifying ontology requirements presented in [76] to compose a set of functional requirements for the SSMO ontology, which are presented in Table 3.1. Ultimately, these questions can be used to evaluate our ontology.

### 3.1.3 Knowledge acquisition

As a starting point for the knowledge acquisition on energy-consuming activities, we adopt the four energy lifestyle domains proposed by the CODALoop (short for

community data-loops) project[3], a platform that aims to raise energy consciousness through, both face-to-face and virtual, discussions about energy issues at the level of the neighborhood. These four domains include dwelling, mobility, food consumption and leisure. From there, brainstorming techniques and carbon footprint calculators[4][5] were used as an inspiration to further explore each domain.

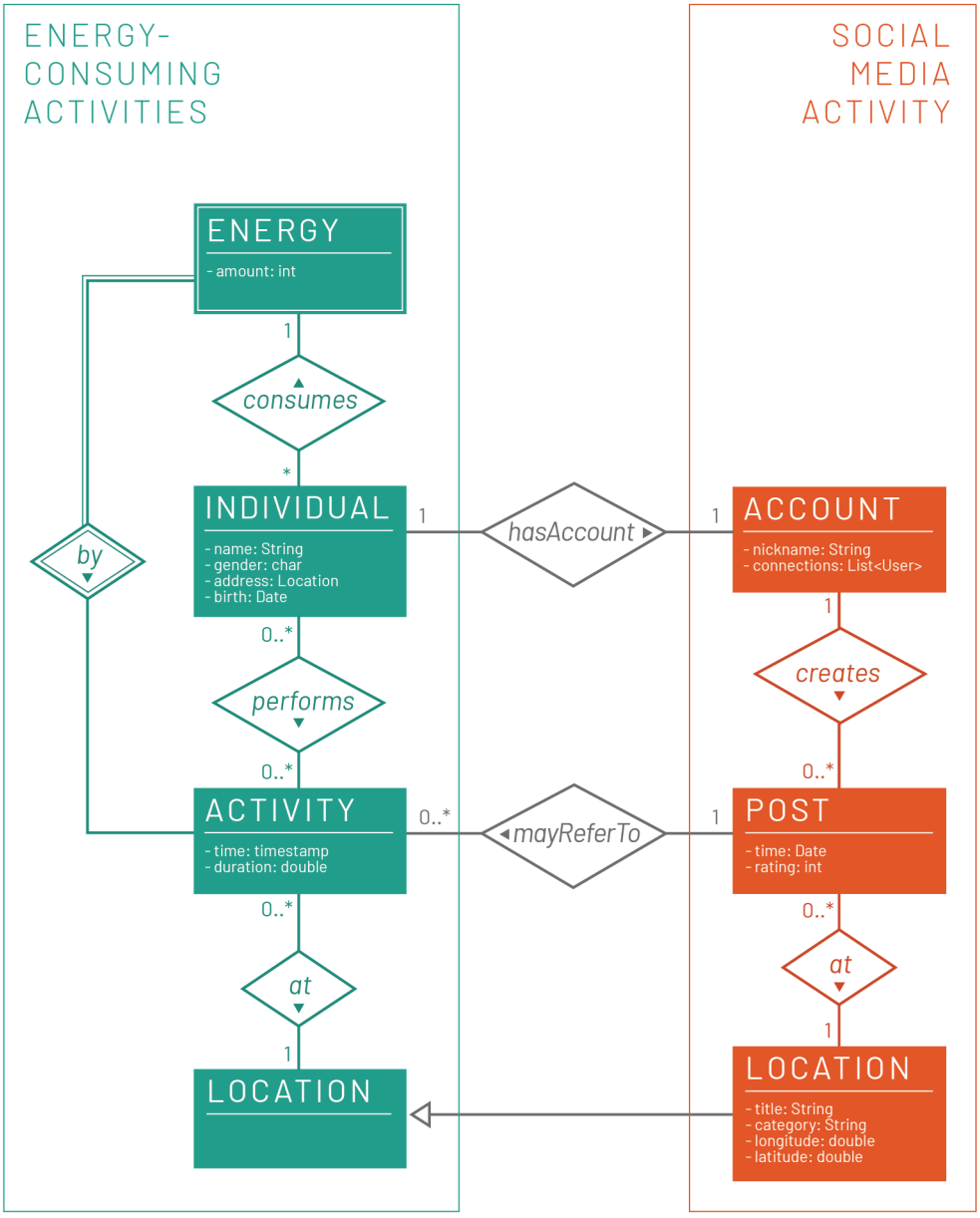



Figure 3.2: The social media's reflection of energy-consuming activities (high-level overview)

[3]https://jpi-urbaneurope.eu/project/codaloop
[4]http://footprint.wwf.org.uk
[5] https://www.nature.org/greenliving/carboncalculator/index.htm

| # | Competency Question (CQ) |
|---|---|
| **1** | Does the individual perform an energy-consuming activity? |
| **2** | If so, what type (or category) of energy-consuming activity is performed by the individual? |
| **3** | At what place is the activity performed by the individual?<br>(i) *To what type (or category) does this place belong?*<br>(ii) *What are the (sets of) coordinates of this place?* |
| **4** | At what time is the activity performed by the individual? |
| **5** | What is the duration of the activity? |
| **6** | Does the individual use an object to perform this activity?<br>(i) *If so, what kind of object?* |
| **7** | In case a mobility activity is performed, what kind of mode of transport is used?<br>(i) *What path (composed of different places, among which are the origin and destination) was taken?* |
| **8** | In case a leisure activity is performed, what kind of artifact(s) is (are) used?<br>(i) *In case the artifact is an appliance, what is its power?* |
| **9** | In case a dwelling activity is performed, what kind of appliance(s) is (are) used?<br>(i) *What is the power of this appliance?* |
| **10** | In case of a food consumption activity, what kind of food is consumed?<br>(i) *What ingredients are included in this food?*<br>(ii) *How (= through which process) is this food processed?*<br>(iii) *Does this process require an appliance? If so, what kind of appliance?*<br>(iv) *Where (= at what place) is this food processed?* |
| **11** | How many energy-consuming activities are performed at a certain (aggregation of) place(s) during a certain time span? |



Table 3.1: Competency Questions that form the set of functional requirements for the SSMO ontology

### 3.1.4 Conceptualization of Energy-Consuming Activities

Given our assumption that individuals consume energy by performing an activity, we consider the energy lifestyle domains as types of energy-consuming activities. Hence, an energy consumption activity can be of type dwelling, mobility, food consumption, and/or leisure, as depicted in Figure 3.3. These high-level concepts are further discussed below, and illustrated in figures 3.4 to 3.8:

**Energy-consuming activities** A user consumes energy by performing an activity at a certain location, at a certain time, and for a certain period of time. That activity can be of multiple types: dwelling, mobility, food consumption, and/or leisure.

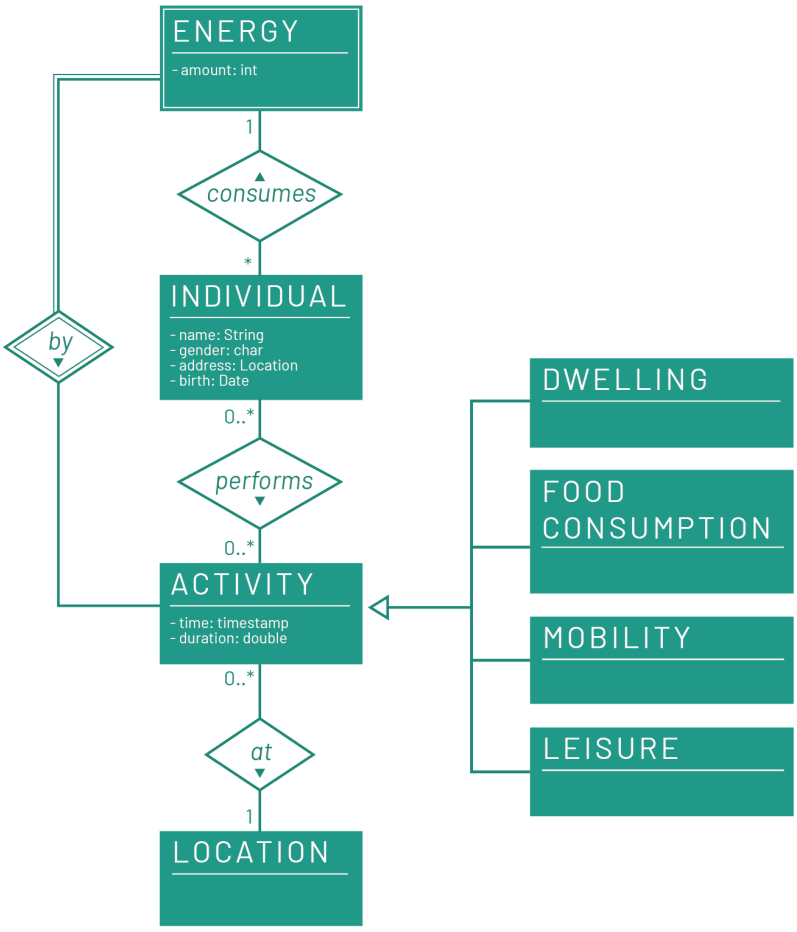



Figure 3.3: Different types of energy-consuming activities

**Location** A location can either be a path or place. A place can be a geographical location (e.g., a town or country) or a venue (e.g., a restaurant or airport) and is characterized by its corresponding coordinates and a category. A path is composed of multiple (at least two) places, among which the origin and destination.

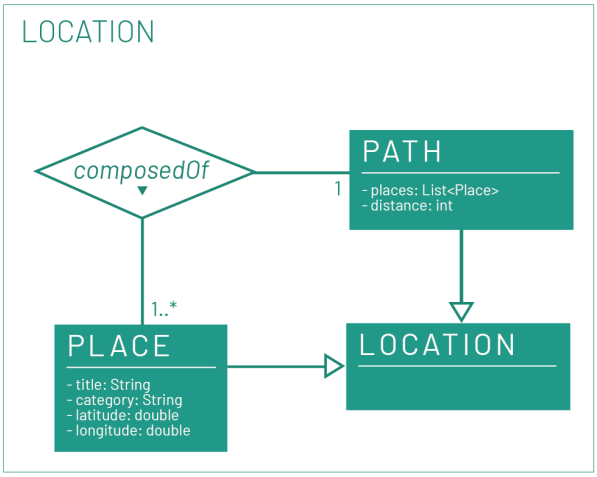



Figure 3.4: High-level concept of a location

**Dwelling** For a domestic activity, generally one or more appliances are used. Among appliances, brown goods (small household electrical entertainment appliances) and white goods (major household appliances) are distinguished [14].

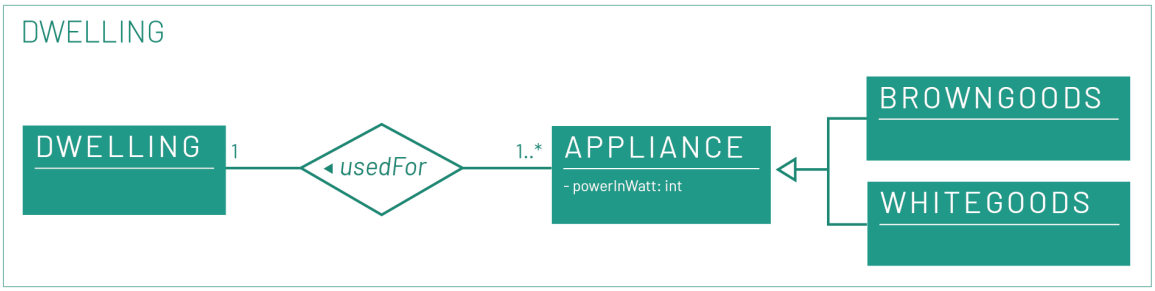



Figure 3.5: High-level concept of a dwelling activity

**Food consumption** In food consumption-related activities (having breakfast or lunch, dining, cooking, etc.), the food product itself and its ingredients, the tableware used for consumption, the food source, and the (cooking) process are relevant entities. Among processes, cooking and modification are distinguished. Modification involves a technique used to modify raw food into food that is ready for cooking.

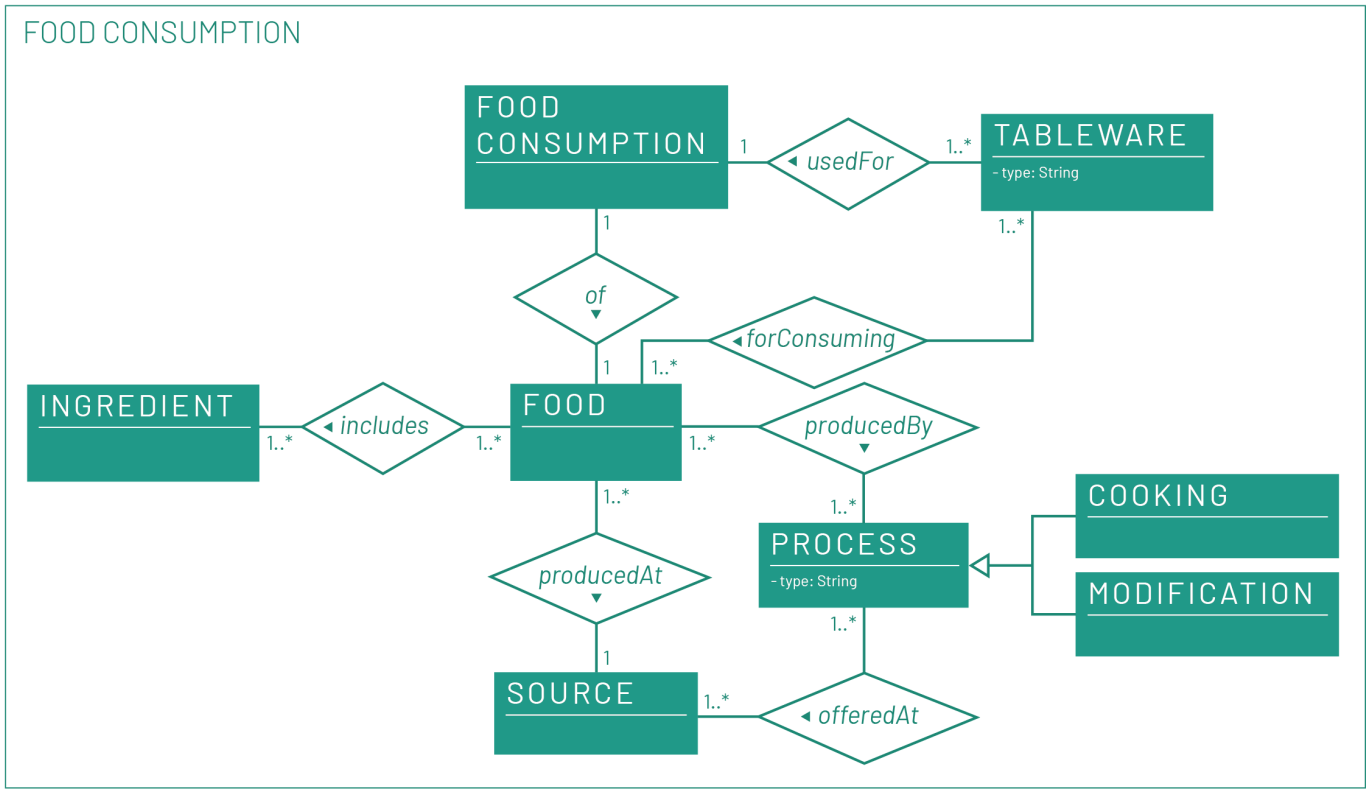



Figure 3.6: High-level concept of a food consumption activity

**Leisure** In leisure, several subcategories can be distinguished, among which: culture, event, gastronomy, playful, relaxation, social interaction, etc. In general leisure activities require the use of one or more artifacts, for instance an appliance.

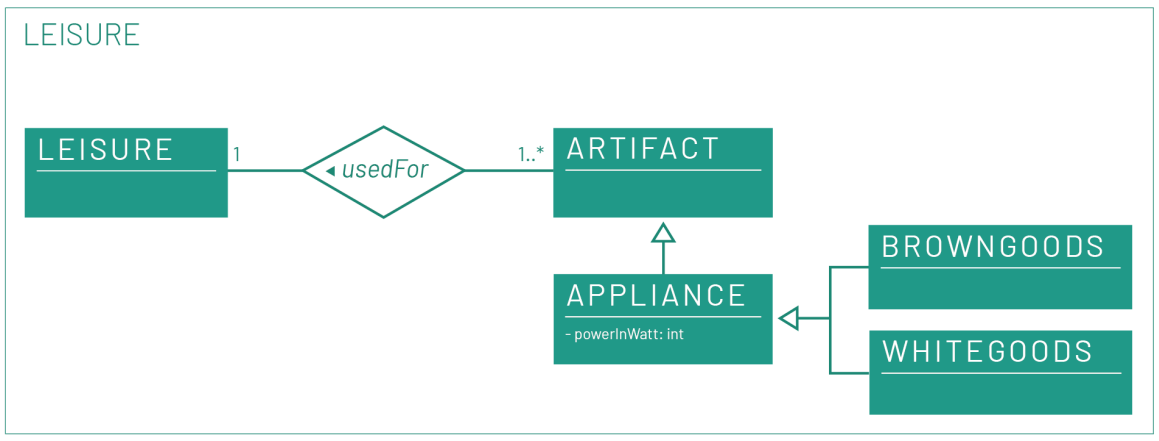



Figure 3.7: High-level concept of a leisure activity

**Mobility** An activity of this type is characterized by the transportation along a path composed of places, including the origin and destination. Furthermore, people travel by a certain mode of transport, for which the type indicates whether the mode of transport is public or private.

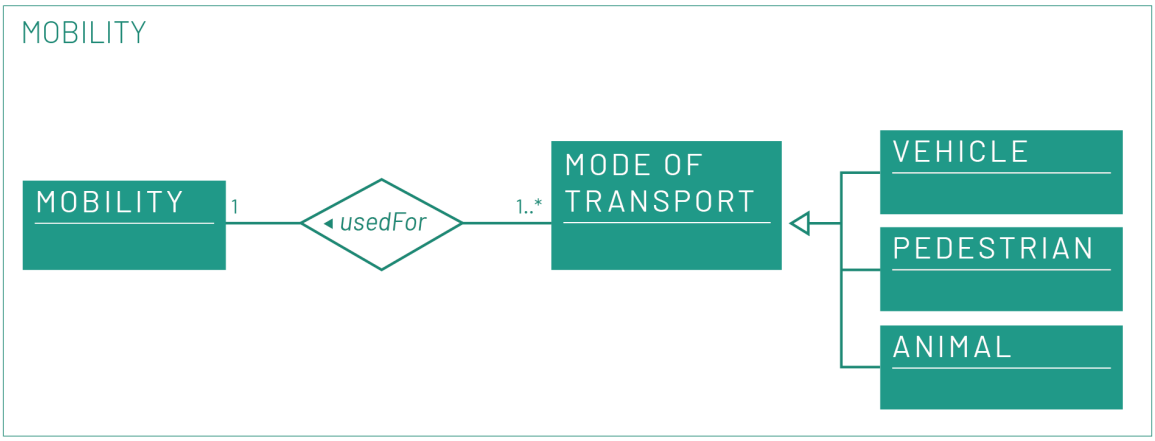



Figure 3.8: High-level concept of a mobility activity

Incorporating all these high-level concepts into one conceptual data model leads to the model depicted in Figure 3.9.

### 3.1.5 Conceptualization of Social Media Activity

Due to our claim that social media posts may reflect an individual's energy-consuming activities, it is important to also get a better understanding of the domain of social media activity. Yet, there have been ontologies developed to model the interaction (resulting in social media data) on social networks on the Web (e.g., [16] and [35]). The most important and relevant concepts within social media activity are as follows:

**User** A user has a social media user account, including a user profile. This user account includes a user profile, in which information about the user (such as name, gender, age, etc.) is stored.

**Post** A user can create one or more social media posts, which can be placed at a timeline or newsfeed in order to share those with other social media users.

**Item** A post contains one or more items, which can be of type message, image, video, link, etc.

**Mention** Within a post, a user can mention a concept, such as another user or a location. This mention provides a link to this concerning concept.

**Location** A user can check in at a location (point of interest); this can be a geographical location or a venue. Often, more information about the location is available, such as the corresponding coordinates or the location category.

Based on the existing ontologies, a conceptual data model was created, depicted in Figure 3.10.

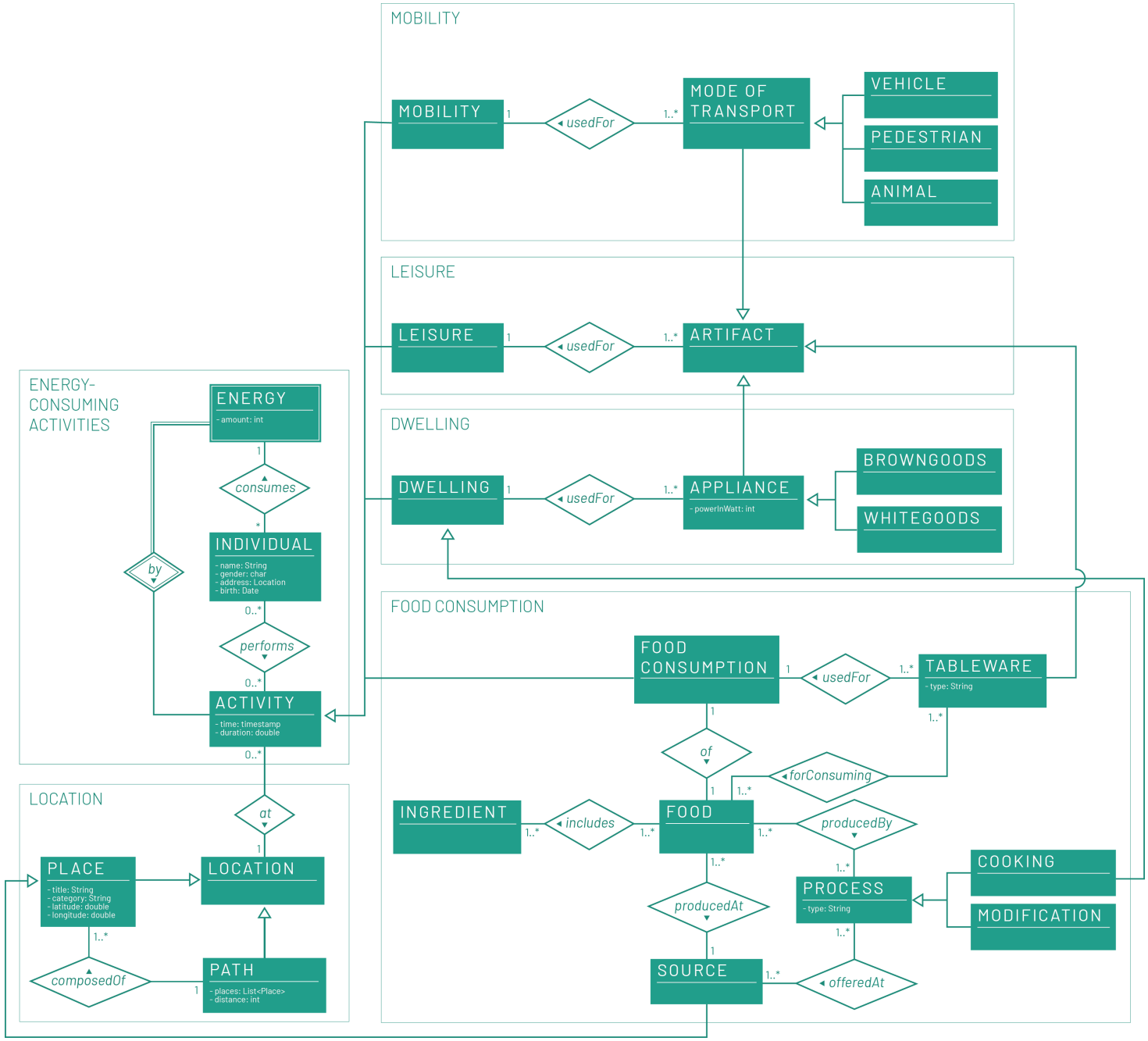



Figure 3.9: Conceptual data model of energy-consuming activities

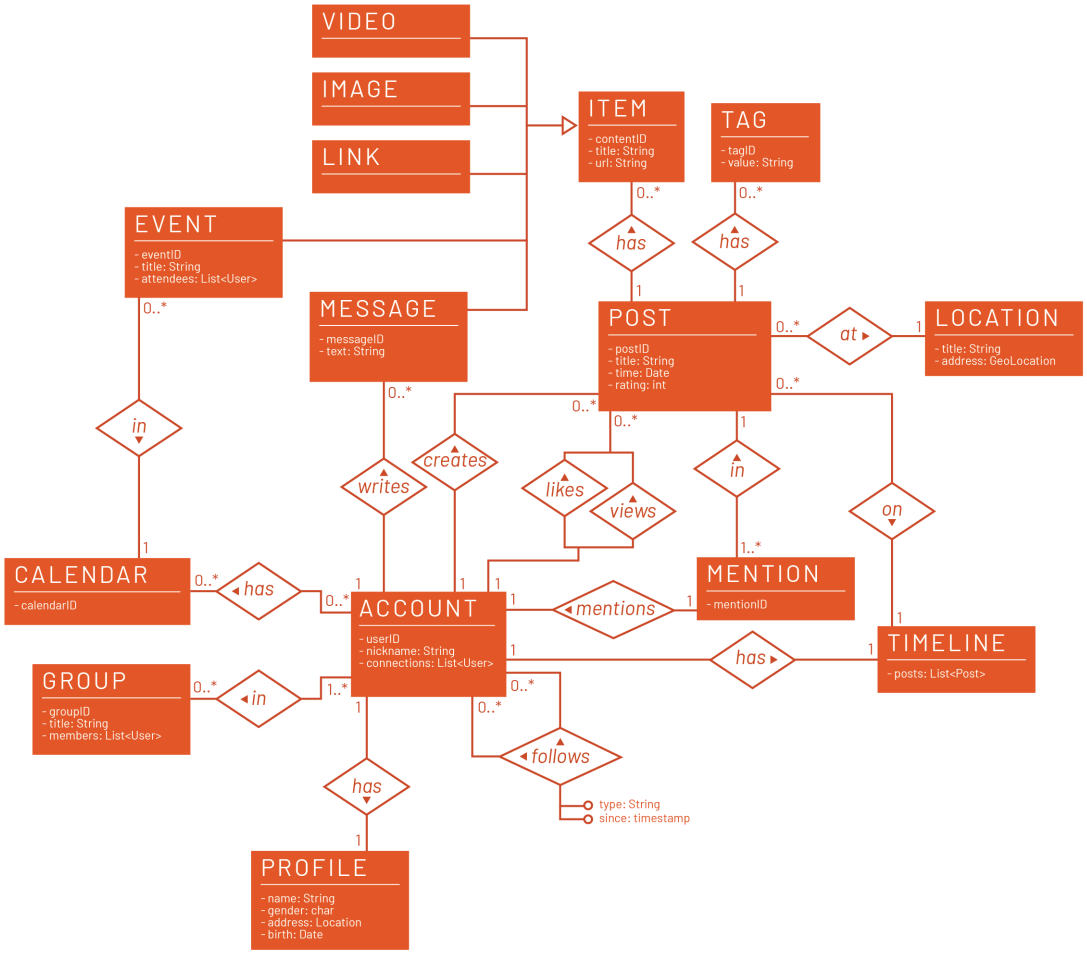



Figure 3.10: Conceptualization of social media activity

### 3.1.6 Instantiating the Conceptual Data Models

In order to determine whether the conceptual data models in Figures 3.9 and 3.10 are a good representation of the domains of energy-consuming activities and social media activity, they were filled with sets of instances. A couple of social media posts were selected and used as a base for the instantiation. Object diagrams are graphs of instances that focus on particular sets of objects and attributes, and the links between these instances. Since these are mainly used to show examples of data structure, these seemed to be a good means for the instantiation of the social media posts, which are displayed in Figures 3.11 to 3.14.

The first example (Figures 3.11 and 3.12) involves an Instagram post an anonymous user account, which belongs to some individual. The post is composed of a message (including tags and a user mention) and an image. The tag `#drinks` denote a food consumption activity, whereas the tags `#music` and `#fun` indicate a leisure activity. In addition, drinks (indicating a food consumption activity), a television, and

a game console (both indicating a leisure activity) are recognized. No place is added to the post. Given this information, we may assume the social media post refers to multiple energy-consuming activities. Firstly, we can derive that the individual is gaming along with watching television. The television and game controller are the appliances that are used for this activity that belongs to both the dwelling and leisure category. Secondly, we deduce the individual is drinking juice (= food) from a glass (= tableware). Hence, numerous of the competency questions from Table 3.1 can be answered, which is displayed in Table 3.2.

The second example (Figures 3.13 and 3.14) involves an Instagram post by another anonymous user account. The post is composed of a message (including a tag), an image, and a place. The message's text token `borrelen` (which could be translated to "having drinks and snacks" in English) denotes a food consumption activity, which is also endorsed by the dining table and multiple food objects recognized in the image. The place (`Maastricht, Netherlands`) does not indicate a specific energy-consuming activity; yet, it does provide more context about the activity that is performed by the individual. Based on the previous assumptions, we imply that the social media post refers to an energy-consuming activity of type food consumption. We deduce the individual is having drinks and snacks; more specifically, roasted (= process) bread (= food) from a platter (= tableware), spread (= food) from a bowl (= tableware), and wine (= food) from a glass (= tableware).

No problems were encountered while instantiating the social media posts using object diagrams. For these posts, the conceptual data models make sense and can be used for the intended uses described before. However, a more thorough evaluation than the instantiation of a couple of social media posts is necessary to validate our ontology. This will be described in more detail in Section *3.1.9 Evaluation*.

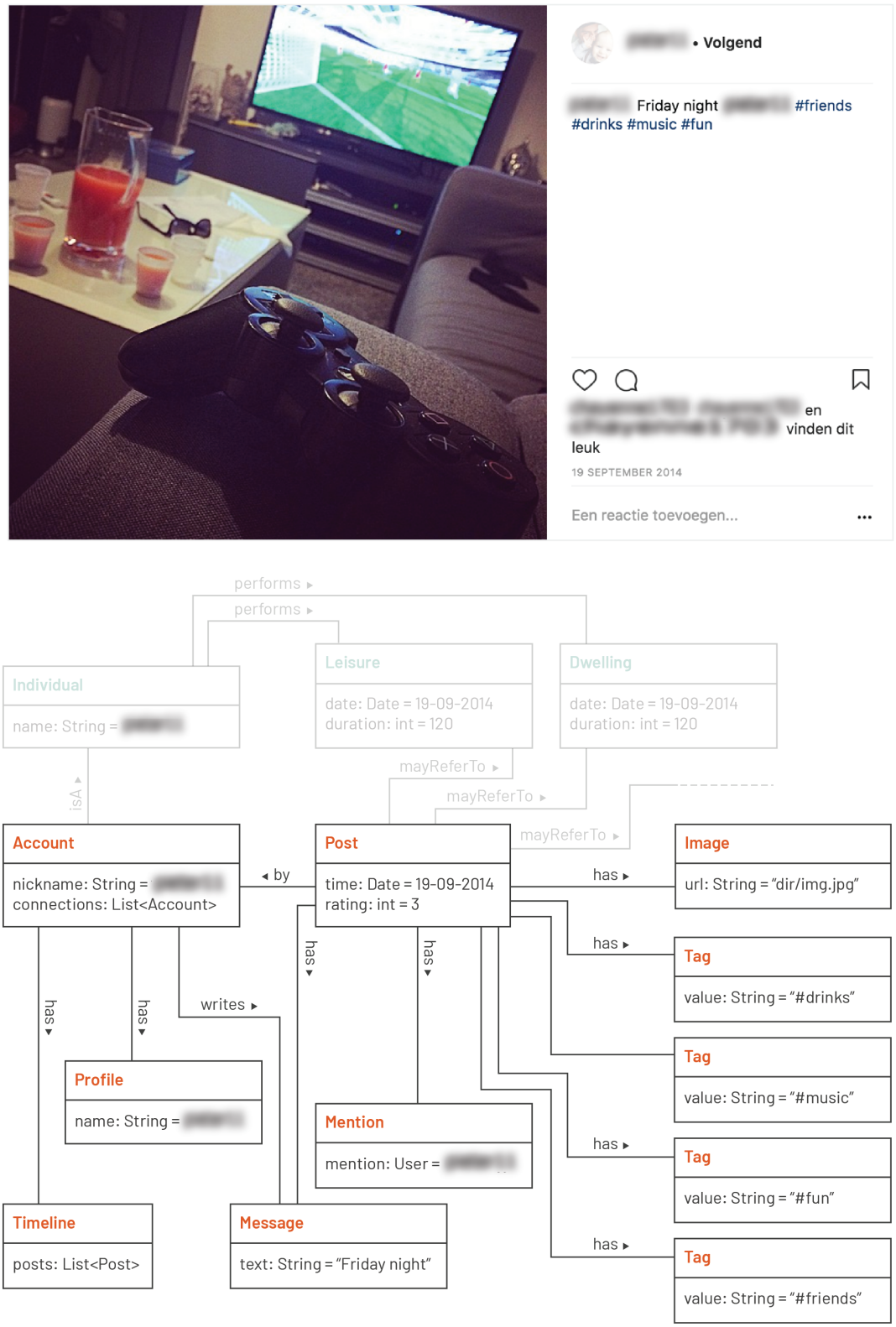



Figure 3.11: Example of instantiating the ontology (1a)

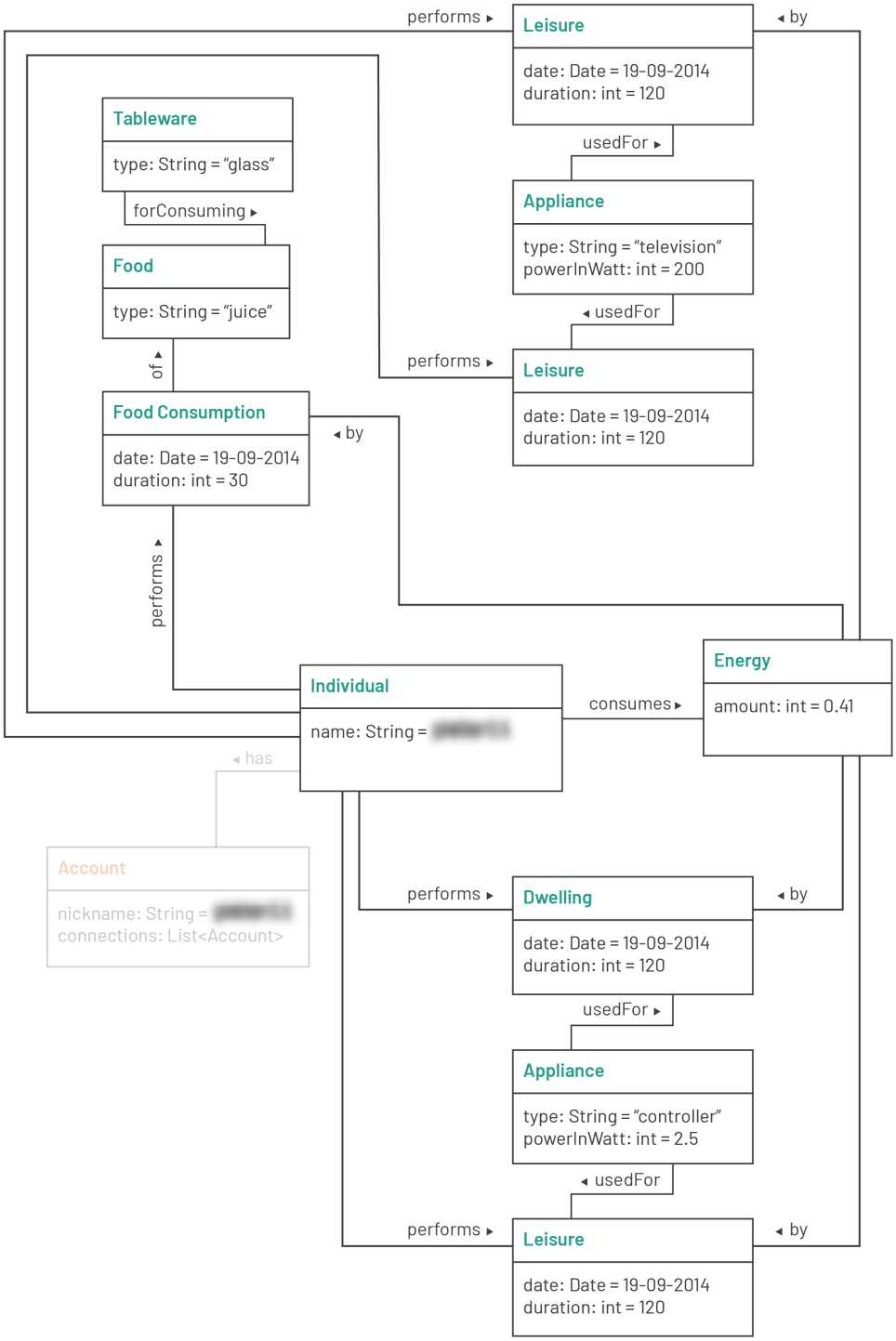



Figure 3.12: Example of instantiating the ontology (1b)

| # | Competency Question (CQ) | Answer |
|---|---|---|
| **1** | Does the individual perform an energy-consuming activity? | Yes |
| **2** | If so, what type (or category) of energy-consuming activity is performed by the individual? | Dwelling, leisure, and food consumption |
| **4** | At what time is the activity performed by the individual? | September 19th, 2014 |
| **5** | What is the duration of the activity? | 120 minutes (*estimated average*) |
| **6** | Does the individual use an object to perform this activity?<br>(i) *If so, what kind of object(s)?* | Yes<br>(i) *Television, game controller, glass* |
| **8** | In case a leisure activity is performed, what kind of artifact(s) is (are) used?<br>(i) *In case the artifact is an appliance, what is its power?* | Television, game controller<br>(i) *200W, 2.5W* |
| **9** | In case a dwelling activity is performed, what kind of appliance(s) is (are) used?<br>(i) *What is the power of this appliance?* | Television, game controller<br>(i) *200W, 2.5W* |
| **10** | In case of a food consumption activity, what kind of food is consumed? | Juice |

Table 3.2: Answers to CQ's for example in Figure 3.11

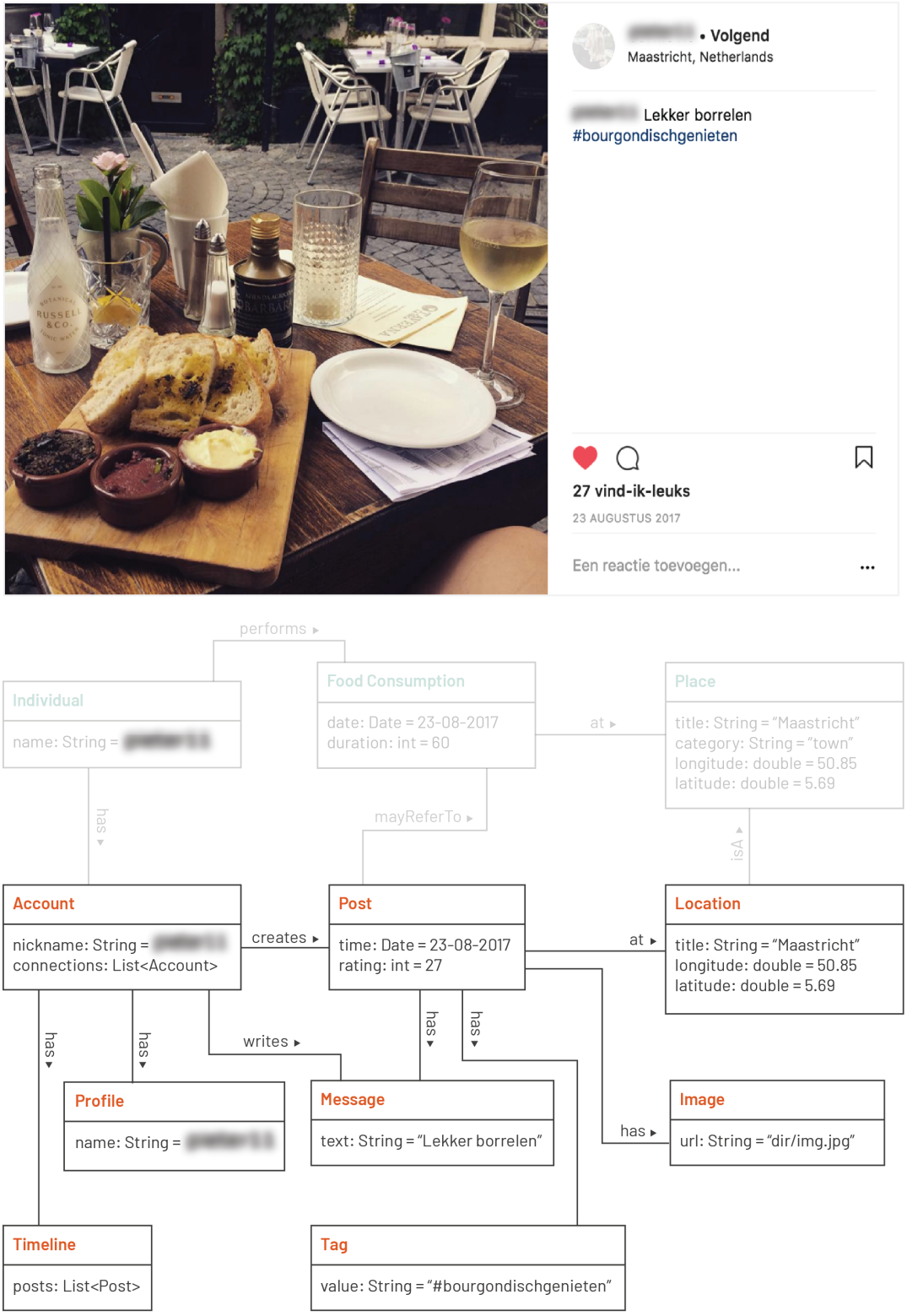



Figure 3.13: Example of instantiating the ontology (2a)

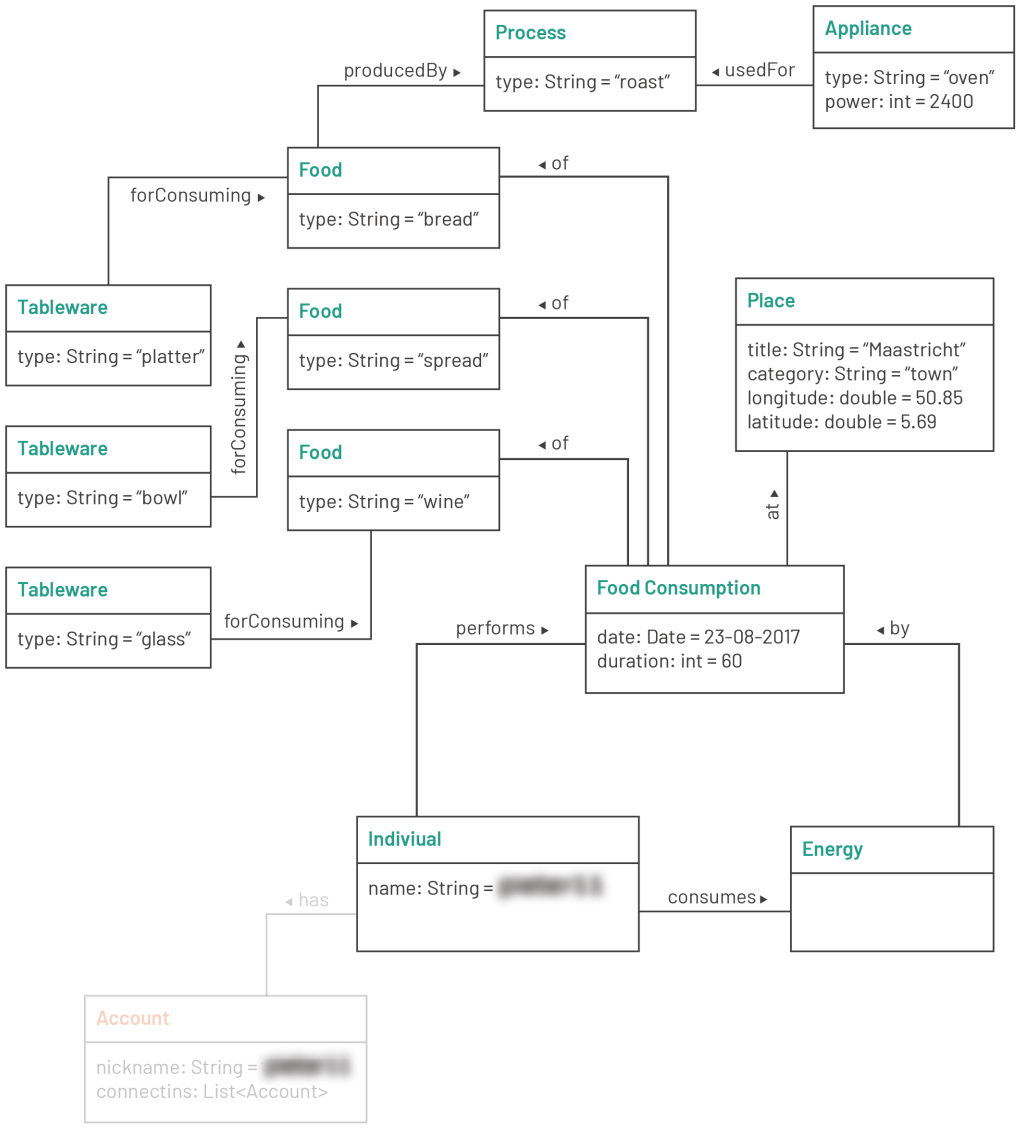



Figure 3.14: Example of instantiating the ontology (2b)

| # | **Competency Question (CQ)** | **Answer** |
|---|---|---|
| **1** | Does the individual perform an energy-consuming activity? | Yes |
| **2** | If so, what type (or category) of energy-consuming activity is performed by the individual? | Food consumption |
| **4** | At what time is the activity performed by the individual? | August 23rd, 2017 |
| **5** | What is the duration of the activity? | 60 minutes (*estimated average*) |
| **6** | Does the individual use an object to perform this activity?<br>(i) *If so, what kind of object(s)?* | Yes<br>(i) *Platter, bowl, glass* |
| **10** | In case of a food consumption activity, what kind of food is consumed<br>(i) *How (= through which process) is this food processed?*<br>(ii) *Does this process require an appliance? If so, what kind of appliance?* | Bread, spread, and wine<br>(i) *Roasting*<br>(ii) *Oven (2400W)* |

Table 3.3: Answers to CQ's for example in Figure 3.13

### 3.1.7 Integration of existing ontologies

To prevent a proliferation of ontologies covering the same entities and relationship, it is important to determine which existing ontologies can be integrated and extended to develop our ontology. Numerous existing ontologies on energy consumption and social media activity were explored in order to discover which entities and relationship can be reused.

**Energy-consuming activities** For the domain of energy-consuming activities, multiple ontologies appeared to be relevant, among which SUMO [65], SEMANCO [58], EnergyUse [18], the BBC Food Ontology[6], MESCO [68], and the Travel Ontology[7]. In Table 3.5 for each ontology is indicated to what extent the entities within the high-level concepts (energy activity, location, dwelling, food consumption, leisure, and mobility) are yet covered. A "+" indicates the entity occurs in the ontology, a "+/−" indicates the entity is covered to some extent, and a "−" indicates the ontology does not include the entity. Each of these existing ontologies is described in more detail below.

The Suggested Upper Merged Ontology (SUMO) and its domain ontologies have been designed as a foundation ontology and is the largest formal public ontology today, used for research and applications in search, linguistics and reasoning (in computer information processing systems). It defines a hierarchy of classes and related rules and relationships. Since this upper ontology covers most of the concepts of our conceptual data model of energy-consuming activities, it is used as the foundation to be extended for our SSMO ontology.

In the SEMANCO Energy Model concepts captured from diverse sources related to the domains of urban planning and energy management are covered. It includes concepts derived from standards, use cases, activity descriptions, and other data sources. It focuses on terms and attributes describing energy consumption and CO2 emission indicators for regions, cities, neighborhoods and buildings, along with climate and socioeconomic factors affecting energy consumption. Compared to the SUMO ontology (which it is built upon), SEMANCO comprises the energy consumption concept to a greater extent. Yet, the concepts of location, food consumption, and mobility are covered to a smaller extent than the SUMO ontology does.

Data from smart plugs is collected in the EnergyUse (EU) platform. Appliance consumption information and community generated energy tips are exported as linked data. It is built upon the PowerOnt [14] ontology, which provides information of energy consumption for numerous household appliances, and extends the DogOnt [13] ontology, which aims to model intelligent domotic environments. The EU ontology mainly covers the concept of an energy consumption activity, as well as the appliance and location (limited to physical locations in a building) entities.

The OntoENERGY [55] ontology focuses on energy units and the corresponding consumption but does not take the energy-consuming activities into account. Since we are more interested in the description of energy-consuming activities than the

[6] https://www.bbc.co.uk/ontologies/fo

[7] http://www.cs.man.ac.uk/ stevensr/ontology/c23.owl

quantification of energy consumption, the OntoENERGY ontology did not seem relevant to integrate in our SSMO ontology.

As the SEMANCO and EU ontologies primarily focus on domestic energy-consuming activities (and thereby cover the concept of dwelling), other existing (non-energy related) ontologies were explored for a better coverage of the concepts of location, food consumption, leisure, and mobility. The BBC Food Ontology (FO) encompasses information about recipes, foods they are composed of, along with suitable diets, menus, seasons, courses and occasions. Entities on the activity of food consumption, food, and the food chain (methods and techniques used to modify the food) are promising for the integration in the SSMO ontology. FO does not cover the tableware entity; yet, this is not problematic since the SUMO ontology yet covers this entity. The MEat Supply Chain Ontology (MESCO) covers the concept of food consumption as well, specifically for the traceability in the meat supply chain. However, the ontology itself was not published online and could thereby not be reused in the development of the SSMO ontology. Furthermore, the Travel Ontology by Stevens, covers most of the relevant entities within the mobility concept, except for the actual mobility activity itself.

| | **SSMO** | **SUMO** | **Semanco** | **EU** | **FO** | **MESCO** | **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 transport | + | + | - | - | - | + | + |

Table 3.4: Overview of the current state-of-the-art related ontologies with a focus on the previously distinguished domains of energy-consuming activities ($+$: included; $+/-$: covered to some extent; $-$: not included)

**Social media activity** To integrate the social media activity in the SSMO ontology, we explored multiple existing ontologies covering this concept. The user account, post, item, mention, and location identities were identified to be the most relevant entities. Whereas the EU ontology yet integrates the commonly used FOAF [35] and SIOC [16] ontologies, our work extends these as well. Moreover, the authors of the SIOC ontology indeed propose the possibility to link to and reuse the FOAF ontology.

In [35], the semantics of the Friend of a Friend (FOAF) ontology are studied to discover how this affects the network structure of multiple online social networks. The SIOC (Semantically-Interlinked Online Communities) ontology reuses terms from existing vocabularies and extends those with new terms needed to describe the relationships between the concepts in the rich world of online community sites. The user account, post, and item entities are covered by both FOAF and SIOC. Yet, the mention entity only recurs in the SIOC ontology, whereas the location entity can merely be found in the FOAF ontology.

| | **SSMO** | **FOAF** | **SIOC** |
|---|---|---|---|
| **Social Media** | | | |
| - User account | + | + | + |
| - Post | + | + | + |
| - Item | + | + | + |
| - Mention | + | - | + |
| - Location | + | + | - |

Table 3.5: Overview of the current state-of-the-art related ontologies with a focus on the previously distinguished domains of energy-consuming activities (+: included; −: not included)

### 3.1.8 Implementation

To a great extent, the SSMO ontology can be built upon existing ontologies, as can be deduced from the overview in Table 3.5; many entities (classes) can be reused. However, some new relationships between those classes have to be defined. In Table 3.6 an overview is created in which the entities are depicted that are reused from existing ontologies. Some existing ontologies have a different purpose than identifying energy consumption; thereby they use a different terminology for entities than we propose in SSMO. For that reason, we created new SSMO entities for these concepts which have an equivalence relationship to the previously mentioned entities from existing ontologies (Table 3.7). Not all entities from the conceptual data models can be covered by existing ontologies. The new entities that had to be created for the SSMO are listed in Table 3.8.

To implement the ontology, we use the Web Ontology Language (OWL) which is introduced below. Protégé[8], Stanford University's free, open-source ontology editor

[8] https://protege.stanford.edu

and framework for building intelligent systems, was used to create/implement the ontology. It allows to export the ontology as an OWL file.

| | **Ontology** | **Prefixed class name** |
|---|---|---|
| **Energy activity** | | |
| - Energy | Semanco | SEMANCO:Energy_Quantity_And_Emission |
| - Individual | SUMO; Semanco | SUMO:Human; SEMANCO:Household_Member |
| **Location** | | |
| - Place | SUMO | geo:SpatialThing |
| - Path | TO; SUMO | upper.owl#Pattern; SUMO:TransitRoute |
| **Dwelling** | | |
| - Activity | SUMO | SUMO:Cooking |
| - Appliance | EU | dogont:Appliances |
| **Food consumption** | | |
| - Activity | SUMO | SUMO:Cooking |
| - Food | FO | fo/Food |
| - Ingredient | FO | fo/Ingredient |
| - Modification | FO | fo/Technique |
| **Leisure** | | |
| - Artifact | SUMO | SUMO:Artifact |
| **Mobility** | | |
| - Activity | SUMO | SUMO:Motion |
| - Mode of transport | TO | travel.owl#ModeOfTransport |
| - Vehicle | SUMO; TO | SUMO:Vehicle; travel.owl#VehicleTransport |
| **Social Media** | | |
| - User account | FOAF | foaf:OnlineAccount |
| - Post | FOAF; SIOC | foaf:Document; ns1:Post |
| - Mention | SIOC | sioc:link |
| - Location | FOAF | foaf:based_near |

Table 3.6: Overview of the entities in the SSMO ontology reused from existing ontologies

**Web Ontology Language (OWL)** One of the aims of the Semantic Web is to improve the current World Wide Web, among others through semantic enrichment of existing Web pages. It relies on the use of semantic markup, typically in the form of metadata described by ontology-like schemadata. The required semantic markup can also be produced by social mechanisms in communities that provide large-scale human-produced markup [6].

Ontology languages allow users to write explicit, formal conceptualizations of domain models. OWL (Web Ontology Language) is the proposed standard for Web

| | Ontology | Prefixed class name |
|---|---|---|
| **Energy activity** | | |
| - Energy | SSMO; Semanco | ssmo:Energy ≡ SEMANCO:Energy_Quantity_And_Emission |
| - Individual | SSMO; SUMO | ssmo:Individual ≡ SUMO:Human |
| **Location** | | |
| - Place | SSMO; SUMO | ssmo:Place ≡ geo:SpatialThing |
| - Path | SSMO; TO | ssmo:Path ≡ upper.owl#Pattern |
| **Food consumption** | | |
| - Modification | SSMO; FO | ssmo:Modification ≡ fo/Technique |
| **Mobility** | | |
| - Mobility | SSMO; SUMO | ssmo:MobilityActivity ≡ SUMO:Motion |

Table 3.7: Overview of the new entities equivalent to reused entities in the SSMO ontology

| | Ontology | Prefixed class name |
|---|---|---|
| **Location** | | |
| - Location | SSMO | ssmo:Location |
| **Dwelling** | | |
| - Activity | SSMO | ssmo:DwellingActivity |
| **Food consumption** | | |
| - Activity | SSMO | ssmo:FoodConsumption |
| - Process | SSMO | ssmo:Process |
| - Tableware | SSMO | ssmo:Tableware |
| **Leisure** | | |
| - Activity | SSMO | ssmo:LeisureActivity |
| - Artifact | SSMO | ssmo:Artifact |

Table 3.8: Overview of the new entities in the SSMO ontology

ontologies. It allows us to describe the semantics of knowledge in a machine-accessible way. It builds upon RDF and RDF Schema. There are two kinds of properties: (i) object properties, which relate objects to other objects (e.g., isPerformedBy), and (ii) data type properties, which relate object to data type values (e.g., title, and age). After the definition of these properties, they can be enriched with a couple of facets: cardinality, required values, and relational characteristics [6].

**Protégé** Stanford's Protégé fully supports the latest OWL 2 Web Ontology Language and RDF specifications from the World Wide Web Consortium, which makes it a good tool for the implementation of the ontology. It allows to create classes (equivalent to entities), object properties (equivalent to relationships), and data type properties (equivalent to properties or attributes). In addition, it enables to add restrictions, such as cardinality restrictions, to the classes and properties.

**Populating the ontology** Manually instantiating the ontology is rather time-intensive and thereby not very efficient. This can be avoided by automatically instantiating (or populating) the ontology. Multiple tools (e.g., Python's rdflib[9]) exist that offer this functionality, and might be considered for future work.

### 3.1.9 Evaluation

The rather complex structure of ontologies makes it hard to evaluate the ontology as a whole; often it is more practical to focus on the evaluation of different levels of the ontology separately.

Moreover, multiple approaches exist for the evaluation of ontologies: the golden standard, the application-based evaluation, the data-driven evaluation and assessment by humans [15]. As for the assessment by humans, the ontology is evaluated by the author of this work by examining to what extent the requirements (from Table 3.1) are met. For the ontology instantiations all questions can be answered; thereby, for these instantiations all posed requirements are met. Yet, this human assessment of just a couple of social media posts is not sufficient for the evaluation.

Since our ontology is created in the context of our Social Smart Meter framework, we also adopt the application-based evaluation approach; this allows us to evaluate how effective the SSMO is in the context of our framework by looking at how the results are affected by the use of the ontology [17]. The framework's outputs and performance are partly dependent on the ontology that is used for it. Hence, the ontology may be evaluated by using it for the framework and evaluating the results of the framework [15]. The performance of our framework will be discussed in more detail in Chapter 5 (*Evaluating the Framework*).

### 3.1.10 Documentation

The documentation of the SSMO ontology was automatically generated by the Live Owl Documentation Environment (LODE)[10], a service that automatically extracts classes, object properties, data properties, named individuals, annotatino properties, general axioms and namespace declarations from our OWL ontology. The corresponding ordered lists, together with their textual definitions are rendered in a human-readable HTML page that enables browsing and navigation by means of embedded links. The documentation page can be found in the Social Smart Meter Ontology's repository[11].

[9] https://github.com/RDFLib/rdflib
[10] http://www.essepuntato.it/lode
[11] https://www.github.com/redekok/social-smart-meter-ontology

### 3.1.11 Maintenance

New ontologies are developed or existing ontologies are updated on a regular basis. To maintain the SSMO ontology, it is important to frequently check if the model is still up to date. Also, one should keep track of new or updated ontologies that can be linked to or re-used in our ontology. Thus, a maintenance process - either an automatic or manual methodology - is required. In previous work, multiple methodologies have been presented. For instance, in [80] an incremental ontology maintenance methodology is proposed which exploits ontology population [80]. Furthermore, the TEXT-TO-ONTO framework [59] applies machine learning techniques to semi-automatically learn ontologies from domain-specific texts. In future work, the different maintenance approaches should be examined in order to identify the most appropriate one for our ontology.

# Chapter 4

# Describing Energy-Consuming Activities using Social Media Data

As we hypothesize that user-generated data is a complementary source of information to describe energy-consuming activities, a framework is developed to automatically process the social media data in order to extract the main characteristics of energy-consuming activities identified in the previous chapter. For this point in time, the user-generated data sources are scoped to social media data sources, which appear to be most promising for this research. In Paragraph 4.3.1 (*Identifying Promising Data Sources*) a further explanation for the selection of our data sources is provided.

The framework, predominantly comprising the social media data processing pipeline, will be the main focus of this scientific contribution. To capture the knowledge and characteristics of the domain of energy-consuming activities–which are present in social media–in our pipeline, we propose to include a rule-based machine learning approach. Over the past years, deep-learning methods have arisen as the state-of-the-art in machine learning. These are methods "with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level" [50]. Recently, new types of systems are introduced that combine representation learning with complex reasoning that should lead to a major progress in artificial intelligence [50]. Thus, this framework aims to combine state-of-the-art machine learning approaches with rule-based reasoning in order to enhance our results.

## 4.1 Framework overview

The data processing framework (Figure 4.1) is composed of three modules (a collection, an enrichment, and a classification module), which are separated in order to ensure these can all operate separately. Based on Figure 4.1 the different modules are introduced below.

During the first stage, the data is collecting through the APIs of our selected data source. Hereafter, the data is pre-processed (in order to get the data in correct format)

and stored in our document-oriented database. Both data (image, and text data) and meta data (user, time, and place data) are collected.

In the second stage, different enrichment steps are included. First, for each social media post, computer vision and natural language processing techniques are applied to respectively the text and image data in order to enrich these data types - e.g., we assign annotations to the image. Then, other data sources are used to enrich our meta data by searching for more details on the user and place data. The last enrichment step is performed through a rule-based reasoning approach. Based on a set of rules, created by reasoning, connections between the data types are examined and conclusions are drawn - e.g., if the individual is at some place other than home, it must have travelled to get there.

Once the data is enriched, the third stage is entered in which the data might be classified to one of the categories of energy-consuming activities. A rule- and dictionary-based approach is used for this classification. Furthermore, a confidence score is assigned to each classification in order to determine how confident we are about it.

**Data types** Text, images, users, places, and time are the data types that are primarily available in social media posts. These data types are also the ones that are most interesting for indicating energy-consuming activities, based on our conceptual data model presented in the previous chapter. Text tokens and image annotations could denote concepts related to energy-consuming activities whereas details on the user, place, and time provide more context about the specific activities. Hence, text, images, users, places, and time are distinguished as the relevant data types for our processing framework. Thereby, these form the main information pillars, recurring in each module (collection, enrichment, and classification).

**Granularity** Social media posts are the input data for our framework, which makes the creator (the individual) of such a post one of our units of analysis. Yet, the analysis of the output of the framework can be performed at different levels of granularity. With the classification output of a single social media post (created by an individual) as a starting point, we could also aggregate the outputs of social media posts by groups of individuals. For instance, all outputs of social media posts created by individuals living in a particular neighborhood could be aggregated to perform an analysis at neighborhood level.

**Output** The framework should be able to classify whether a social media post refers to an energy-consuming activity or not, resulting in a boolean output. Along with this boolean value, the category of energy-consuming activity, the classification confidence, and the terms indicating this activity, should be part of the output. Based on the relevant terms identified in this post, the SSMO ontology can be instantiated to provide a better understanding of the energy-consuming activities. Hereafter, the count of social media posts related to energy-consuming activities (i.e., with a positive boolean), can be used to compare the amount of energy-consuming activities for our different units of analysis (i.e., at individual or group level).


Figure 4.1: High-level overview of framework

## 4.2 Orders of Data

For our framework, we introduce an $N$-order (meta)data enrichment approach. The basic idea is that $N+1$ order data can be derived from $N$ order data - e.g., by processing imagery data (first order data), it can be enriched with a set of annotations (second order data). We distinguish data and meta data, where the latter one describes the first - e.g., data about the creator (i.e., the user) of the social media post is considered to be meta data since it describes the social media post's data (i.e., text and image). Figure 4.2 shows that a social media post is composed of first order data (text, image) and first order meta data (user, time, and place). The second order data (relevant text tokens and image annotations) and second order meta data (more details on user and place) can be derived from our first data order by respectively applying state-of-the-art data processing techniques and accessing the APIs of our enrichment data sources. The third order meta data is derived from our first order data and meta data, as well as from our second order data and meta data, by applying a set of rules, which are defined by reasoning.

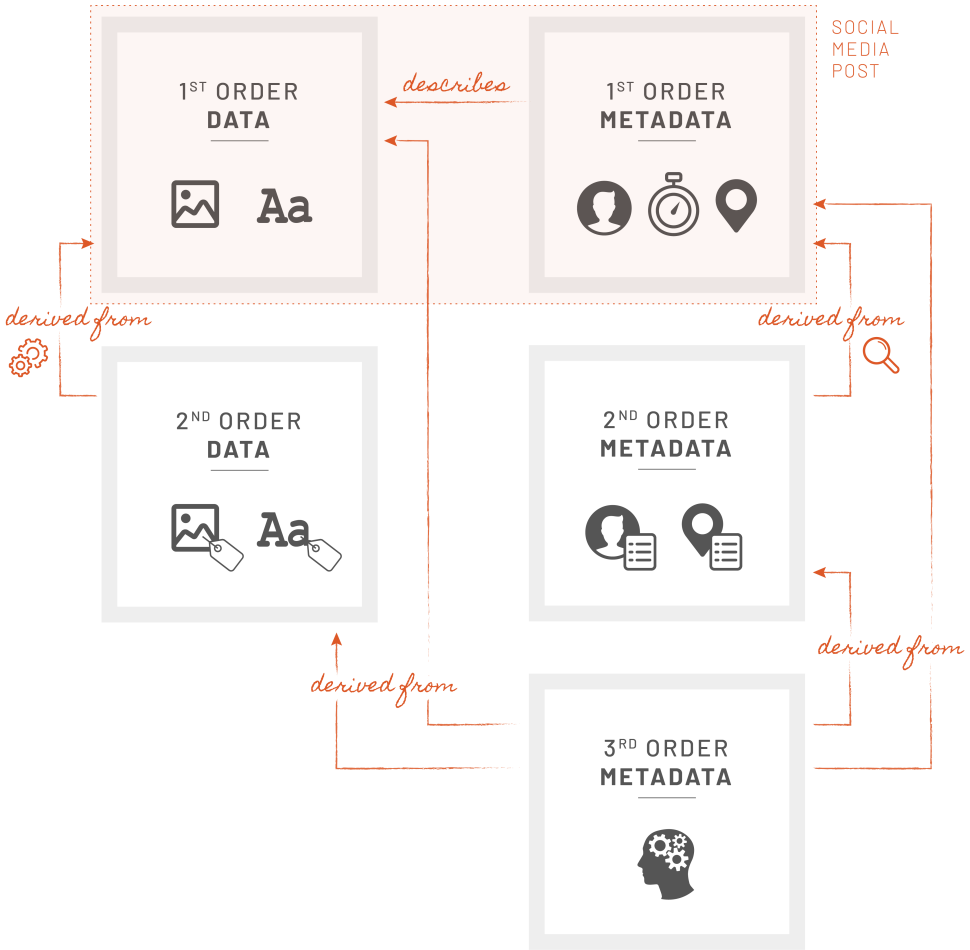



Figure 4.2: $N$ order (meta) data enrichment

**First order data** involves the social media post data (text and imagery data) which is retrieved from the social media data sources.

**Second order data** is derived from the text and imagery data (first order data) by using Natural Language Processing and Computer Vision techniques to retrieve word vectors and image annotations.

**First order metadata** involves the data (including user, time, and place data) describing the social media post data (first order data) which is retrieved from the social media data sources.

**Second order metadata** is derived from the user and place data (first order metadata) by collecting more details on this data using social media and enrichment data sources.

**Third order metadata** is derived from the first and second order (meta)data by reasoning - e.g., if an individual performs an (energy-consuming) activity at some place other than his or her home, we could infer that he or she must have travelled this distance by using a suitable transportation means.

## 4.3 Data Collection

During the data collection, the application programming interfaces (APIs) of our selected data sources are used to gather all the relevant data. Subsequently, the data is pre-processed and stored in a database, in the form of a document.

### 4.3.1 Identifying Promising Data Sources

Today, an abundance of social data sources is available, through which a lot of information on individuals can be retrieved. Specifically for the domain of energy-consuming activities, numerous social data sources were identified to be promising. Through a structured comparison (tables 4.1 and 4.2) this selection of data sources was analyzed and compared on different aspects, such as the usage among Dutch individuals, the availability of an API, the relevance for each of the four categories of energy-consuming activities, or the existence and availability of a variety of (social media) entities such as text, images, user, place, and time (i.e., our types of data).

Social media data sources provide information about the individual user and his or her posts, which allows us to gather our input (meta) data. Enrichment data sources are used to enrich the yet collected meta data (on the user, time, and place). For instance, if we know from our social media data source that a user created a post including a location-based check-in (= meta data), we can use our enrichment data source to enrich this data with place details such as the place categories. Numerous social media data sources are analyzed in Table 4.1, whereas multiple relevant enrichment data sources are analyzed in Table 4.2. Different characters ($+$ and $-$) are used to indicate the availability of data types. A $+$ indicates the data type is available in

the data source, whereas a $-$ indicates its unavailability. The character inside the parentheses indicates the accessibility of the data type through the data source's API - e.g., a user's user details are present on Facebook but these are not accessible through its public API, which would result in $+(-)$. In Table 4.2 the $+$ and $-$ characters are also used to indicate how informative we consider a data source for each of the categories of energy-consuming activities. Based on these comparisons, Instagram and Twitter appeared to be the most promising social media data sources due to their publicly available API and the fact that we can gather information on all our relevant entities through these APIs. In addition, Foursquare and Google Place were considered to be the most promising enrichment data sources.

Furthermore, Newcom Research & Consultancy has been conducting Netherlands' largest research into social media since 2010. Results from the 2017 report (8194 responses) show that:

- WhatsApp remains the largest social media platform (10.9M Dutch users), followed by Facebook (10.4M), YouTube (7.5M), LinkedIn (4.3M), Instagram (3.2M), and Twitter (2.6M).
- The use of WhatsApp and Facebook keeps growing in 2017, the use of Twitter continues to decrease, and relatively, Instagram experienced the largest growth.
- WhatsApp and Facebook are used (on daily basis) by all age groups; YouTube, Instagram and Snapchat are more popular among younger people (age groups 15-19 and 20-39).

The relevant content of WhatsApp and Facebook is not accessible. Facebook's graph API could be interesting for analyzing its social network though, which might be relevant for future work. YouTube's content does not seem sufficient relevant for the data analysis in our field, which leaves us with Instagram and Twitter as most promising (social) data sources. Twitter's API is publicly accessible and Instagram's API requires a license, which could be obtained.

| Data source | Usage | Public API | User details | Text | Images | Geolocation | Place details | Timestamp |
|---|---|---|---|---|---|---|---|---|
| Facebook | 10.4M | + | + (-) | + (-) | + (-) | + (-) | + (+) | + (-) |
| Instagram | 3.2M | + | + (+) | + (+) | + (+) | + (+) | + (+) | + (+) |
| Twitter | 2.6M | + | + (+) | + (+) | + (+) | + (+) | + (+) | + (+) |
| Google+ | < 0.5M | + | + (+) | + (+) | + (+) | + (+) | + (+) | + (+) |
| Steam | < 0.5M | + | + (+) | + (-) | - (-) | - (-) | - (-) | + (+) |
| Nextdoor | < 0.5M | - | + (-) | + (-) | + (-) | - (-) | - (-) | + (-) |
| Peerby | < 0.5M | - | + (-) | + (-) | + (-) | - (-) | - (-) | + (-) |

Table 4.1: Social media data sources
($+$: available; $(+)$: accessible; $-$: unavailable; $(-)$: inaccessible)

| Data source | Public API | Dwelling | Food | Leisure | Mobility | User(s) details | Place details |
|---|---|---|---|---|---|---|---|
| Airbnb | + | - | - | + | - | - (-) | + (+) |
| Foursquare | + | - | + | + | + | - (-) | + (+) |
| Google Places | + | - | + | + | + | - (-) | + (+) |
| Tomtom | + | - | + | + | + | - (-) | + (+) |
| Tripadvisor | + | - | + | + | + | - (-) | + (+) |
| Uber | + | - | - | - | + | + (-) | + (+) |
| Waze | + | - | - | - | + | - (-) | + (+) |
| Yelp | + | - | + | + | + | - (-) | + (+) |
| OV-chipkaart | - | - | + | + | + | + (-) | - (-) |

Table 4.2: Enrichment data sources
(+: available; (+): accessible; −: unavailable; (−): inaccessible)

### 4.3.2 Social media data sources

Following up on the conclusions drawn above, Twitter and Instagram were identified to be the most promising and relevant social media data sources to collect the first order input data for this research, due to the fact that these are widely used in the Netherlands, and provide public APIs to retrieve the data entities (text, images, places, time, user) we are interested in.

**Twitter** Twitter, a social network for microblogging, offers different types of API families: the REST APIs, which allow queries to endpoints, and Streaming APIs, which allow tweet queries in real-time. For this research, we were interested in historical data for which we made use of the Twitter REST (and specifically, the Standard search) API. Through this API tweets can be queried on different parameters, for instance location bounding box (circle specified by the center of coordinates and radius), and time. Tweepy[1], the Python library to access the Twitter API, was used in our data crawler script.

**Instagram** Instagram, a photo and video-sharing social networking service (owned by Facebook), offers an API Platform, a REST API through which a variety of endpoints can be accessed. In particular, the media search allows to query Instagram posts using parameters such as a location bounding box (circle specified by the center of coordinates and radius), and time. Unfortunately, multiple functionalities have been deprecated since April 2018 due to the upheavel around Facebook along with the stricter regulations regarding privacy.

[1] http://www.tweepy.org/

### 4.3.3 Enrichment data sources

To collect the first order metadata that should contribute to the enrichment of the yet retrieved input data, Foursquare and Google Places are used as our enrichment data sources.

**Foursquare** Foursquare, a local search-and-discovery service mobile app which enables its users to update their connections about the venues they have checked in. Its Places API allows to get detailed information on these places (such as the location, categories, etc.). In this framework Foursquare is used to enrich our place data with the corresponding categories.

**Google Places** Google Places, a service by Google, allows to query for detailed place information on a variety of categories by searching for proximity or text string. The Google Places API is used as a second enrichment data source to get first order metadata in case the metadata is not available through Foursquare.

## 4.4 Data Enrichment

After the data collection by the social media and enrichment data sources is accomplished, this data is ready to be enriched. To maintain a clear overview of all different types of (meta)data, an *N* order (meta)data enrichment is used (Figure 4.2). Using this approach, the collected textual, imagery and place data can be enriched through different steps, which are described below. The different enrichment steps that are applied to each data type are also exemplified in Figures 4.4 to 4.7, based on the example social media post in Figure 4.3.


Figure 4.3: Example social media post

### 4.4.1 Text Processing

In order to enrich the text (derive the second order data from the first order one), it is processed using a variety of state-of-the-art techniques. By applying Natural Language Processing techniques, relevant terms can be identified after which we can determine to what extent they match the terms in our dictionary, i.e., to what extent they are informative for the performance of an energy-consuming activity.

If a social media post contains any message, it is stored in the corresponding document in our database. This message might be very noisy, containing slang, hashtags or mentions, which is why several text pre-processing techniques (stopword removal, removal of hashtags and other special characters, stemming, etc.) are applied along with the process of tokenization (word segmentation of the message). This results in a set of tokens that might refer to an energy-consuming activity. To determine whether this is indeed the case, we perform two more enrichment steps. Firstly, the synonym sets of the token are retrieved to provide insights into the context and meaning of the token in order to obviate word sense ambiguity. Secondly, similarity scores between the tokens and dictionaries are calculated.

**Disambiguation** Depending on a sense that is intended for a word, its meaning may vary widely, resulting in some words having an ambiguous sense. Lesk introduced an algorithm to handle this word sense disambiguation using machine readable dictionaries [51]. This Lesk algorithm assumes that words in a particular text section (i.e., a message in our case) are likely to share a common topic. The Adapted Lesk algorithm, implemented by the NLTK library, incorporates WordNet[2]'s lexical database. This algorithm would take a social media message as its input in which a single target word–in this case that would be a term that seems to refer to an energy-consuming activity–occurs as the input, and based on information (derived from WordNet) about this term and a few directly surrounding words, it will output a WordNet sense for this target [9].

To evaluate all tokens of a text message, and its corresponding meaning, as appropriately as possible, Lesk's word sense disambiguation algorithm is used to determine the token's intended sense and if this relates to the performance of an energy-consuming activity. Relevant words (indicating such an energy-consuming activity) are stored along with the intended synonym (using WordNet's synsets).

**Word embeddings for similarity scoring** Word2Vec (a model used for learning vector representations of words, called "word embeddings") is used to calculate a similarity score between the term and our text dictionaries, by computing distances from the given entity (our relevant term) to all entities in the dictionary. These scores are required for calculating the classification confidence.

### 4.4.2 Image Processing

In order to enrich the imagery data (derive second order data from first order data), state-of-the-art image processing techniques are applied to provide annotations on objects and scenes that are recognized in the images.

[2]https://wordnet.princeton.edu/

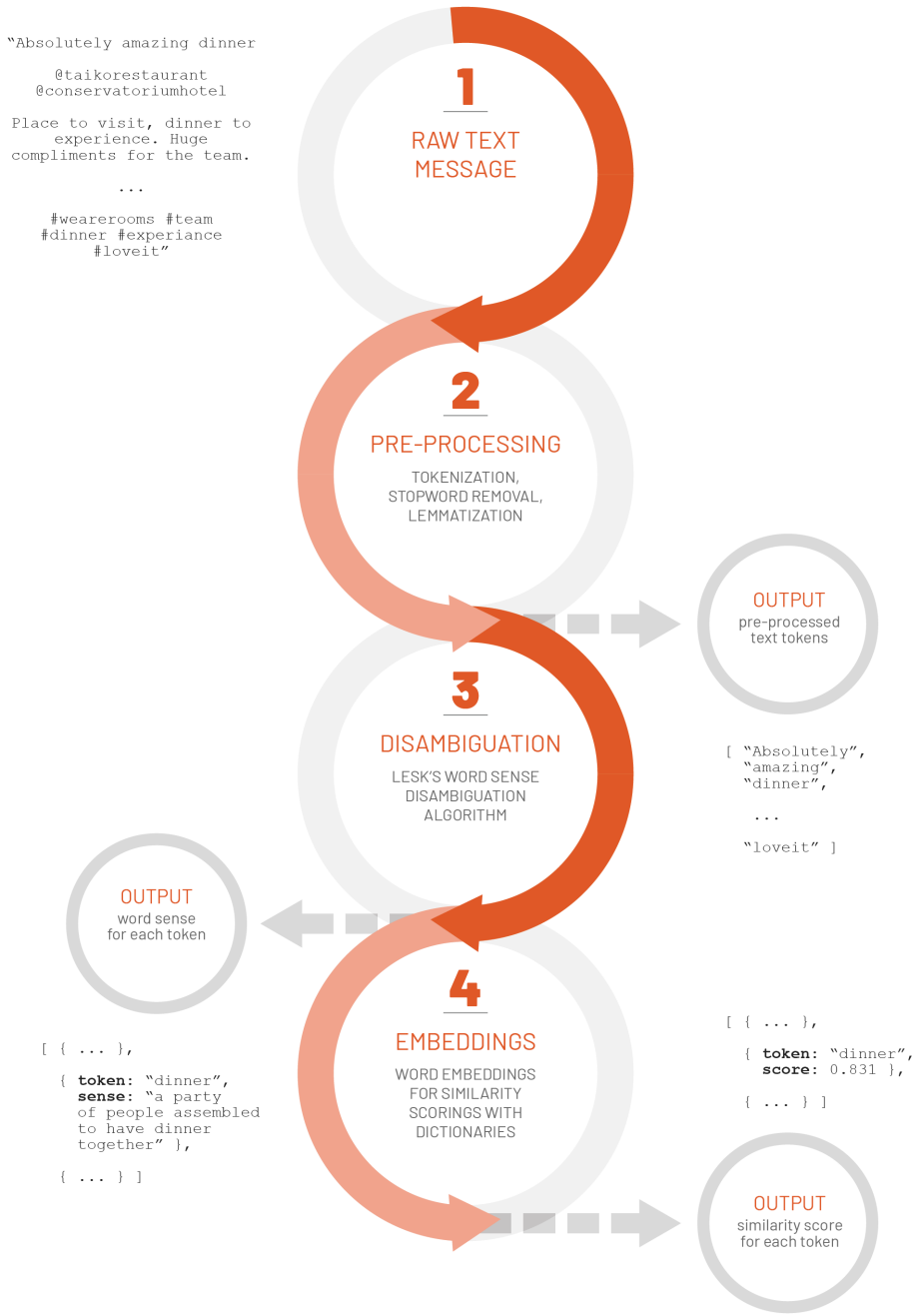



Figure 4.4: Overview of text enrichment steps

We are interested in annotations that indicate an energy-consuming activity performed by the individual that posted the picture along with a social media post. Today, different types of computer vision techniques exist that assign annotations to images. For this framework, both object recognition models and scene recognition models should be included. Given the two examples in Figure 4.6, the difference between the annotations of scene and object recognition becomes very clear. In the

example in Figure 4.6c, the platter of food–which is recognized by the object recognition model–clearly indicates a food consumption activity. However, the objects recognized in the example in Figure 4.6a (multiple persons), do not indicate a leisure activity of visiting a theatre. The scene recognition in Figure 4.6b on the other hand does recognize a theater scene. Hence, the two different computer vision techniques are considered as complementary sources of information, and should thereby both be integrated in the framework.

Another factor influencing the type of annotations is the dataset on which the computer vision model is trained. For that reason, multiple pre-trained computer vision models should be included in our image enrichment module in order to recognize as many different aspects from the image as possible.

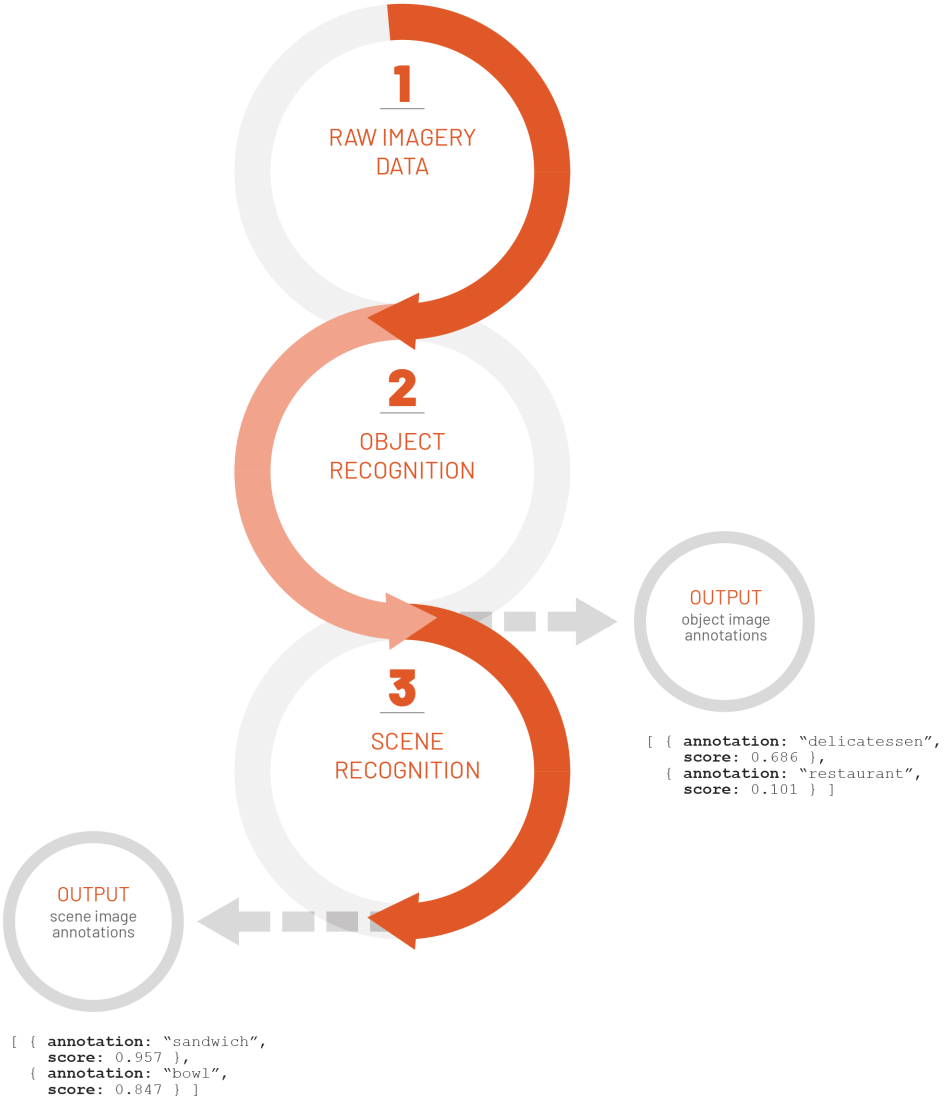
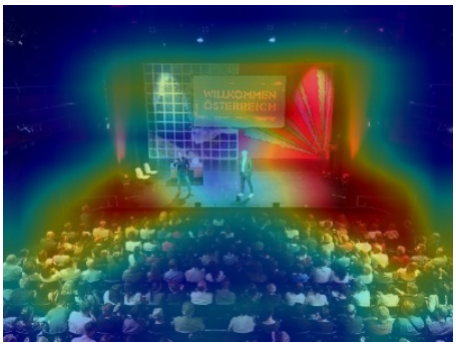



Figure 4.5: Overview of image enrichment steps


(a) Object recognition in example 1 (annotations: person (79x)




(b) Scene recognition in example 1 (annotations: movie theater (indoor), auditorium, orchestra pit)

(c) Object recognition in example 2 (annotations: person, dining table, cup (2x), knife (2x), bowl (5x))

(d) Scene recognition in example 2 (annotations: dining hall, cafetaria, delicatessen)

Figure 4.6: Differences in computer vision techniques

### 4.4.3 Place Processing

Processing the place, retrieved along with the collection of a social media post, could result in a lot of informative information on whether an (and what type of) energy-consuming activity is performed. Enriching a place with its place category (derivation of second order meta data from first order meta data) could be an indicator for the category of energy-consuming activities. In addition, by comparing the user's current location to its previous (or home) one, more information on (other) energy-consuming activities could be deduced.

Since each energy-consuming activity is performed at a particular place, we are interested in the place (in the form of a geolocation) contained by the social media post. The corresponding place details could be another indicator for one or more energy-consuming activities. The information acquired by the enrichment of the place data is two-folded. On the one hand, the place category of the place from the post

may infer a certain energy-consuming activity. For instance, if the individual has created a post including a location-based check-in at a restaurant, this indicates a food consumption activity. On the other hand, if we know the user has checked in to a place that is not equal to either his or her home or the last place he or she checked into, a mobility activity is implied. For instance, say a user checked in at a restaurant at 6PM and checks in at a theater at 9PM that same evening, we may imply that the user traveled the distance between those two places. Concluding, the place enrichments involve: i) include more place details such as the place category, ii) infer an individual's home (and work) location by either getting this information from the user's profile details, or by clustering the locations of his or her posts [87, 90], iii) determine the distance between a user's location-based check-ins.

**Place matching** To enrich our first order place metadata, the available data (name and coordinates) could be used to find a place match with the Google Places and Foursquare places. Numerous studies have looked into place matching yet; [61] found that the mean great circle distance between two matched POIs was equal to 62.8 meters and in [43] a buffer area with a radius of 25 meters (per POI) was used to reduce geocoding errors. Based on these values, we will use a radius of 50 meters in this research. Thus, place matching is performed by applying a few criteria and passing those as parameters for the place query by the enrichment data sources:

- Pass the place name as query or place name.
- Pass the coordinates as center of the (circular) bounding box.
- Set the radius to 0.05 km (to ensure that our bounding box is small enough to find the place we are looking for).

If a match is found, the corresponding place details are requested in order to collect one ore more place categories. In the classification stage these categories will be analyzed to determine whether they indicate one or more energy-consuming activities - e.g., a movie theater indicates a leisure activity.

**Distance between posts** If a user creates multiple posts a day in which he or she checks in at different places, this implies the user has traveled between those places, which also indicates an energy-consuming activity of type mobility. By determining the distances between those places, we could infer how far the user has traveled. Thus, for each user we keep track of the posts created on a day. For each post in this set, we check if the location-based check-in differs from the previous one. If so, the distance between those places is determined. If this distance is larger than a particular threshold ($> 0.2$ km), the concerning post is (also) classified to refer to a mobility activity.

**Home detection** Once we have an overview of all the places a user has checked in, one could infer the user's home location by spatial clustering. When the home location is detected, distances between the home and other location check-ins could be estimated.

In this work, the density-based spatial clustering of applications with noise (DBSCAN, [27]) seems to be an appropriate data clustering algorithm. Opposed to other clustering algorithms, DBSCAN takes the neighborhood size as a parameter and

is applicable to data sets with uneven cluster sizes. It separates high-density clusters from low-density ones and marks outliers points lying alone in low-density areas (whose nearest neighbors are too far away). This matches our demand for a clustering algorithm that is able to determine a user's home, since we assume that the location of a user's home will be a relatively small-sized, high density area, whereas at other places less check-ins take place, resulting in areas of low density.

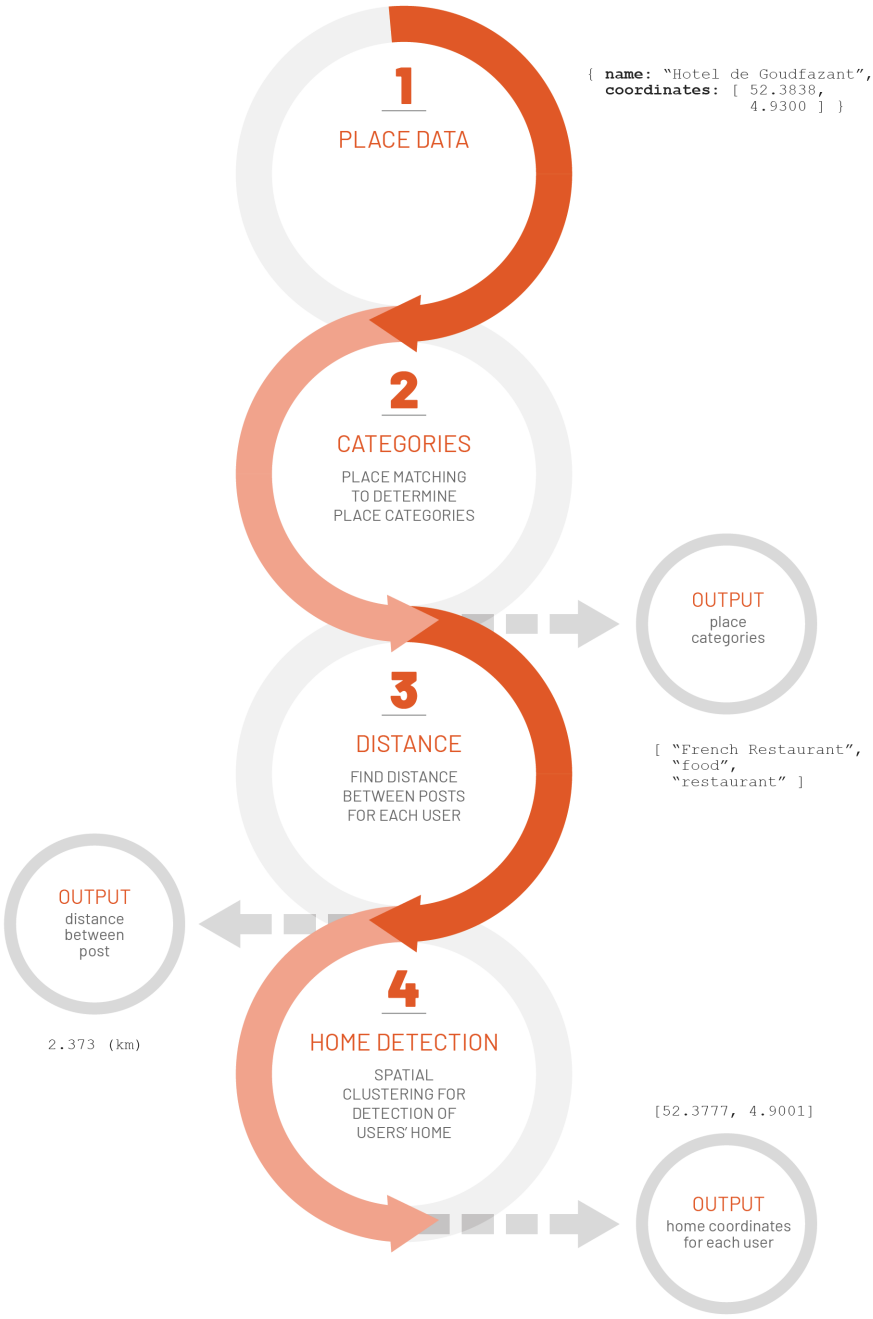



Figure 4.7: Overview of place enrichment steps

## 4.5 Data Classification

Once the social media data post is enriched, a rule-based approach is applied to classify the data for the description of possible energy-consuming activities. These activities can not always be recognized explicitly from the social media post (by matching text tokens, image annotations or place categories to our dictionary). Thereby, some reasoning is necessary to recognize implicit energy-consuming activities as well. For instance, if there is both evidence for a food consumption activity and evidence for the individual performing this activity at home, we could imply that this activity is associated with dwelling as well. If the individual performs a food consumption activity at some place other than home, this activity could instead additionally be linked to leisure. Thus, the rule-based approach should facilitate the inclusion of this human reasoning in order to enhance our model's output results.

### 4.5.1 Rule-Based Approach

Reasoning plays a big role in our *N*-order (meta)data enrichment, particularly in how third order meta data is derived from first and second order (meta) data. To apply some structure in how this reasoning is machine-readable, a set of rules (in the shape of workflows which are composed of step-wise activities, actions and decisions) is introduced, underlying this reasoning.

In order to create a graphical representation of the workflow of the step-wise activities and actions resulting from this rule-based approach, an activity diagram (Figure 4.8) was created. The arrows indicate the direction of the workflow and represent the order in which activities happen, diamonds represent decisions, rounded rectangles represent actions, and bars represent the start (split) or end (join) of concurrent activities. These all illustrate how the workflow should be followed.

On the one hand, we determine whether the user is at home (by comparing the user's location to his or her detected home. Then, relevant terms in the post are analyzed. For each term, we identify if it is evidence for one or more energy-consuming activities (dwelling, leisure, food consumption and/or mobility). In case a leisure or food consumption activity is performed at home, we could reason this activity is classified to dwelling as well. Furthermore, if a food consumption activity is performed at some place other than home, we could reason this also classifies as a leisure activity.

On the other hand, we determine the user's distance to his or her previous post (as described in Paragraph 4.4.3 *Place Processing*). If the distance exceeds the threshold of 0.2 km, we consider it to be a mobility activity. Along with that, we analyze whether a vehicle was required to bridge this distance. If so, the mode of transport can be inferred - e.g., if the distance traveled in a day is more than 5000 kilometers, it is very likely the individual traveled by aircraft to cover that distance.

### 4.5.2 Dictionaries

To determine whether a social media post refers to one or more energy-consuming activities (and if so, to what type(s) of activities), a dictionary-based classification approach is applied. We define a dictionary as a set of terms related to a specific

energy-consuming activity type - e.g. ingredients or cooking utensils are associated with the food consumption category. Thus, each category of energy-consuming activities has a distinct dictionary, or set of related terms. The basic idea is to compare the terms extracted from the message (text tokens), image (annotations), and place (categories) to the terms in the dictionary. For now, a distinct dictionary for each of these types of data is constructed. Undoubtedly, this comes along with some hassle but it also rules out disambiguity to some extent - e.g., the text token “tram" might infer a mobility activity whereas the image annotation “tram" could also point at some tram in the background which might not be related to the user’s activity. For that reason, in this work the dictionaries are separated.

**Text dictionaries** In [60], the authors use a hybrid dictionary-similarity distant supervision with the purpose of classifying Twitter content to energy consumption-related content. Since the work in [60] is very similar to this research, the dictionary used in their work is adopted to this research. Using only a dictionary would not be a sufficient classification means, since social media posts typically contain relatively short messages, and a single word within such a message can have different meanings according to the context. For that reason, not only the words themselves are compared to the dictionary but also the corresponding synonym that is determined by the word sense disambiguity algorithm. All terms in our text dictionary should therefore also be enriched with the synonym that reflects the intended sense (or meaning) of the word.

Further, as described in a previous paragraph, a similarity score between a token and the relevant dictionary are calculated, which is one of the input parameters of our classification confidence model.

**Image dictionaries** The social media image is subject to different pre-trained computer vision models. Each of these models is trained to recognize one or more classes from a predefined list of classes. These lists of classes (obviously) differ for each pre-trained model. For each list, the classes are manually classified to none, one or more of the different categories of energy-consuming activities - e.g. “television" relates to both dwelling and leisure and thus is part of both dictionaries, whereas “person" does not indicate any energy-consuming activity and is thereby not included in any dictionary.

**Place dictionaries** Alike the image annotations, the sets of place categories are also pre-defined. As all place categories that could possibly be assigned to a place are known, these can be categorized in the same manner as the image annotation classes, i.e., by manually classifying which place category relates to which dictionary category - e.g., a “restaurant" place category is part of both food consumption and leisure dictionaries.

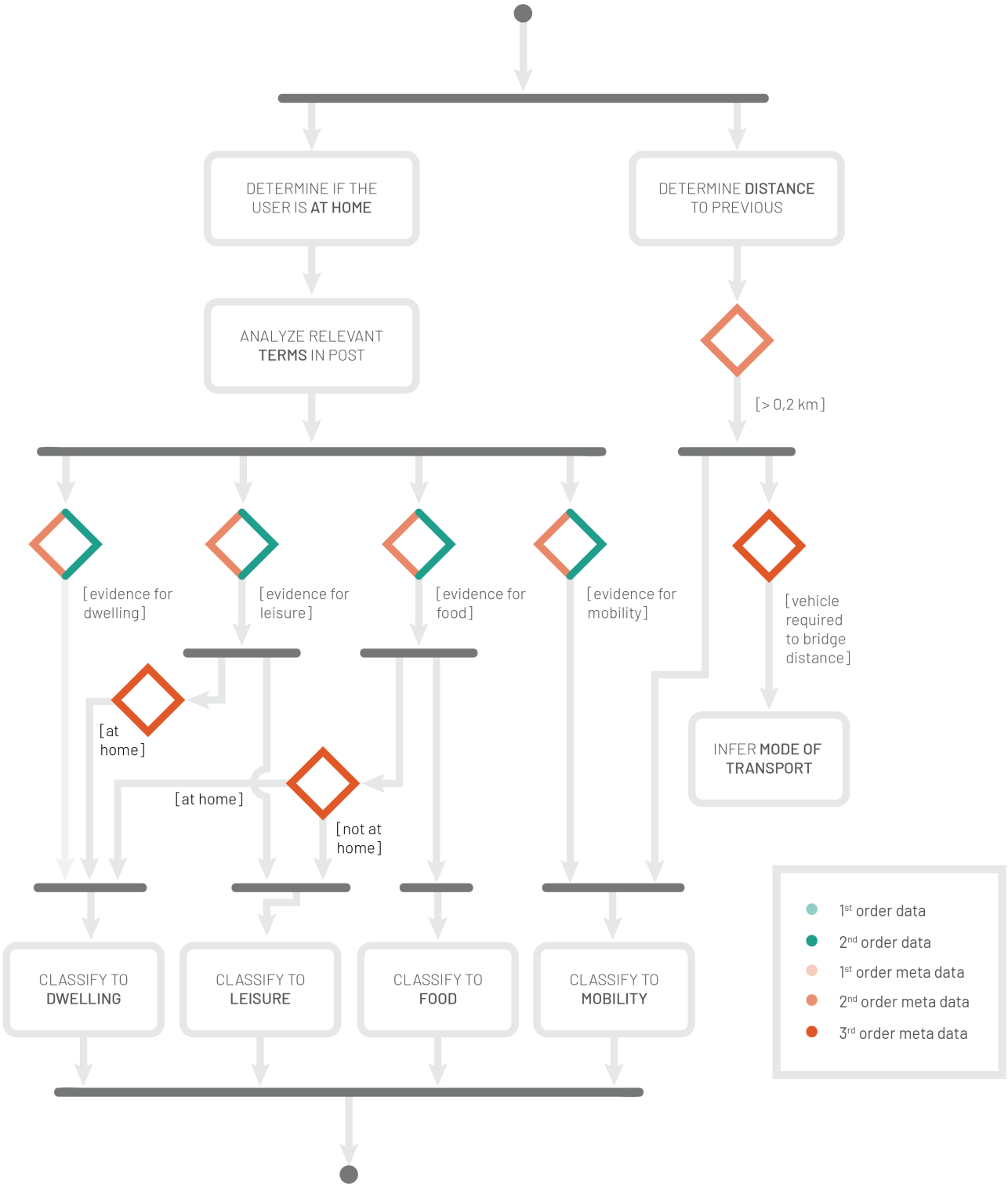



Figure 4.8: Activity diagram of the rule-based approach

### 4.5.3 Confidence Model

As mentioned before, social media data is often very noisy; thereby, it is difficult to separate relevant from irrelevant information. Hence, it might be the case that our framework classifies a social media post falsely to one or more energy-consuming activities. In order to handle this error proneness, we suggest to assign a confidence to each classification.

In order to quantify this confidence, a rule-based classification confidence model is proposed. Relevant parameters affecting the confidence are (i) the ratio of relevant tokens, distinguished on type of data (text, image, place), (ii) for each term some kind of score indicating the similarity (or relatedness) to the category of energy-consuming activities, and (iii) a weighted factor that represents to what extent the type of data is an indicator for this category of energy-consuming activities - e.g., indicative objects related to food consumption are more often recognized in images than objects related to mobility; thereby image annotations related to food consumption grant a higher confidence for this category compared to image annotations related to mobility.

The confidence model is only used for the classification by dictionaries. Hereafter, an initial threshold of 0.5 is applied in order to determine to which categories of energy-consuming activities the social media post is classified. This threshold value should eventually be tuned (based on a user-based evaluation) in order to optimalize the framework's performance. Subsequently, the classification by the rule-based approach is performed.

**Ratio of relevant terms** The ratio of relevant terms (per data type) related to activity category $x$ compared to the total number of relevant terms indicates to what extent the social media post refers to $x$, and thereby also how confident we are about this category. If the data type's relevant terms refer to more than one category of energy-consuming activities, we are less confident about the classification of each category.

**Term scores** The scores for relevant terms are determined separately for each type of data. For a text token this score is equal to the score representing the similarity to the concerned dictionary, whereas for an image annotation this is equal to the annotation score assigned by the model. Since multiple computer vision models are used to enrich the image with annotations, each is accountable for scoring its own annotations ($\in [0,1]$). For a place category this score is binary (either 0 or 1), depending on whether the place category occurs in the dictionary.

**Data type weights** Depending on the social media post, one of the data types (image, text, or place) might be more informative for an energy-consuming activity than the other ones. Therefore, we suggest to include data type weights for each category of energy-consuming activities. For instance, it is hard to recognize a mobility activity from an image, since individuals do not often post images of objects such as a transportation means while traveling. A check-in which is based at a mobility-related place such as an airport or train station would be way more indicative in that situation. On the contrary, if individuals perform a food consumption activity, they are more likely to post images in which food objects can be recognized. Hence, we include a weighted factor ($w_{x,y}$) in the confidence model to

indicate to what extant the term is indicative for the category of energy-consuming activities, depending on from which type of data it is derived. For now, values are assigned to these weights based on intuition (Table 4.3). However, these should eventually be evaluated by involving the crowd.

| **Type of data** ($y$) / **Category** ($x$) | Text | Image | Place |
|---|---|---|---|
| Dwelling (D) | 0.33 | 0.33 | 0.33 |
| Food consumption (F) | 0.33 | 0.33 | 0.33 |
| Leisure (L) | 0.33 | 0.33 | 0.33 |
| Mobility (M) | 0.33 | 0.33 | 0.33 |

Table 4.3: Weights used in the classification confidence model

**Calculating the confidence** Taking all of the above into account, the calculation of our classification confidence can be formulated as in the following equation (4.1):

$$\begin{aligned} \text{confidence}_x &= \sum_y \left(\frac{N_{relevant,x,y}}{N_{relevant,y}} \cdot w_{x,y} \cdot \frac{1}{N_{relevant,x,y}} \sum_x \text{scores}_{x,y}\right) \\ &= \sum_y \left(\frac{1}{N_{relevant,y}} \cdot w_{x,y} \cdot \sum_x \text{scores}_{x,y}\right) \end{aligned} \tag{4.1}$$

where $N_{relevant}$ is the number of relevant terms, $w$ is the weighted factor, $x$ is the type of energy-consuming activity, $y$ is the type of data (text, image, or place), and $\vec{\text{scores}}$ is the vector of the scores ($\in [0,1]$) of all relevant terms. Thus, if the classification confidence for an energy-consuming activities category is calculated, the average of the confidence outputs of all types of data determine the overall confidence score. In future work we could examine whether other strategies (such as taking the maximum of minimum instead of the average) results in better results.

If we take another look at the examples from Figures 4.9 to 4.11, we could calculate the corresponding classification confidences for the category of food consumption:

| | **Example post 4.9** | **Example post 4.10** | **Example post 4.11** |
|---|---|---|---|
| ***Text*** | - | breakfast (F, 0.7404)<br>apple (F, 0.6439)<br>banana (F, 0.6705)<br>yoghurt (F, 0.6551)<br>waffle (F, 0.6999) | Cycling (M, 0.8104)<br>ferry (M, 0.7936)<br>home (D, 0.9450)<br>training (L, 0.8823)<br>workout (L, 0.8765)<br>bike (M, 0.7371) |
| ***Image*** | laptop (D, 0.9994)<br>keyboard (D, 0.9846)<br>mouse (D, 0.9548)<br>keyboard (D, 0.8914)<br>mouse (D, 0.8913) | laptop (D, 0.9995)<br>cup (F, 0.9970)<br>bowl (F, 0.9884)<br>dining table (F, 0.9788)<br>bowl (F, 0.9198) | - |
| ***Place*** | - | - | Boat or Ferry (M, 1) |

Table 4.4: Relevant terms for different categories of energy-consuming activities for multiple social media post examples (D: dwelling, F: food consumption, L: leisure, M: mobility)

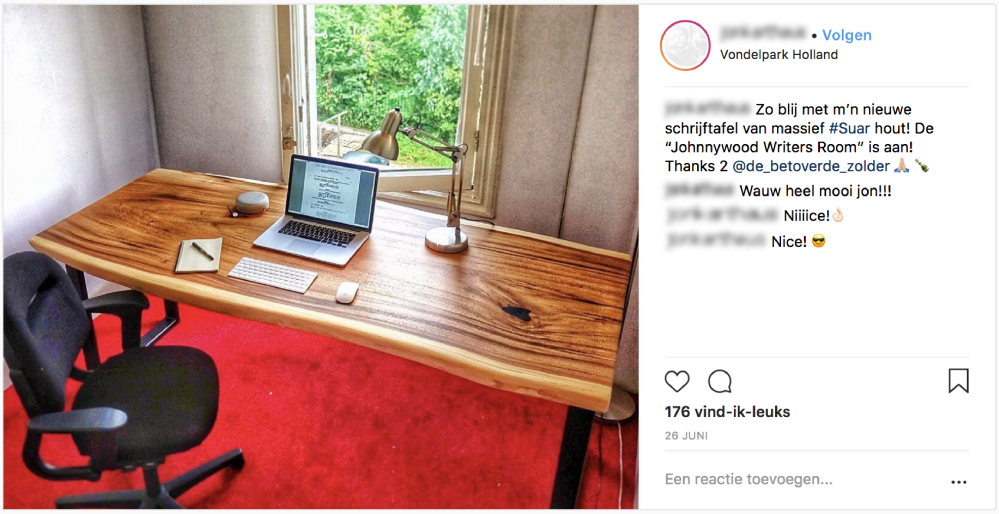



Figure 4.9: Social media post that has evidence for a dwelling activity

*Example 4.9*:

$$\begin{aligned}
\text{confidence}_D &= \sum_y \left(\frac{1}{N_{relevant,y}} \cdot w_{x_D,y} \cdot \sum_{x_D} \text{scores}_{x_D,y}\right) \\
&= \text{confidence}_{D,text} + \text{confidence}_{D,image} + \text{confidence}_{D,place} \\
&= 0 \\
&+ \frac{1}{5} \cdot 0.33 \cdot \sum(0.994, 0.9846, 0.9548, 0.8914, 0.8913) \\
&+ 0 \\
&= 0.3113
\end{aligned} \tag{4.2}$$

*Example 4.10*:

$$\begin{aligned}
\text{confidence}_D &= \sum_y \left(\frac{1}{N_{relevant,y}} \cdot w_{x_D,y} \cdot \sum_{x_D} \text{scores}_{x_D,y}\right) \\
&= \text{confidence}_{D,text} + \text{confidence}_{D,image} + \text{confidence}_{D,place} \\
&= 0 \\
&+ \frac{1}{5} \cdot 0.33 \cdot \sum(0.995) \\
&+ 0 \\
&= 0.0657
\end{aligned} \tag{4.3}$$


Figure 4.10: Social media post that has evidence for dwelling and food consumption activities

$$\begin{aligned}
\text{confidence}_F &= \sum_y \left(\frac{1}{N_{relevant,y}} \cdot w_{x_F,y} \cdot \sum_{x_F} \text{scores}_{x_F,y}\right) \\
&= \text{confidence}_{F,text} + \text{confidence}_{F,image} + \text{confidence}_{F,place} \\
&= \frac{1}{5} \cdot 0.33 \cdot \sum(0.7404, 0.6439, 0.6705, 0.6551, 0.6999) \\
&+ \frac{1}{5} \cdot 0.33 \cdot \sum(0.9970, 0.9884, 0.9788, 0.9198) \\
&+ 0 \\
&= 0.4765
\end{aligned} \tag{4.4}$$

*Example 4.11*:

$$\begin{aligned}
\text{confidence}_D &= \sum_y \left(\frac{1}{N_{relevant,y}} \cdot w_{x_D,y} \cdot \sum_{x_D} \text{scores}_{x_D,y}\right) \\
&= \text{confidence}_{D,text} + \text{confidence}_{D,image} + \text{confidence}_{D,place} \\
&= \frac{1}{6} \cdot 0.33 \cdot \sum(0.9450) \\
&+ 0 \\
&+ 0 \\
&= 0.0520
\end{aligned} \tag{4.5}$$


Figure 4.11: Social media post that has evidence for dwelling, leisure and mobility activities

$$\begin{aligned}
\text{confidence}_L &= \sum_y \left(\frac{1}{N_{relevant,y}} \cdot w_{x_L,y} \cdot \sum_{x_L} \text{scores}_{x_L,y}\right) \\
&= \text{confidence}_{L,text} + \text{confidence}_{L,image} + \text{confidence}_{L,place} \\
&= \frac{1}{6} \cdot 0.33 \cdot \sum(0.8823, 0.8765) \\
&+ 0 \\
&+ 0 \\
&= 0.0967
\end{aligned} \tag{4.6}$$

$$\begin{aligned}
\text{confidence}_M &= \sum_y \left(\frac{1}{N_{relevant,y}} \cdot w_{x_M,y} \cdot \sum_{x_M} \text{scores}_{x_M,y}\right) \\
&= \text{confidence}_{M,text} + \text{confidence}_{M,image} + \text{confidence}_{M,place} \\
&= \frac{1}{6} \cdot 0.33 \cdot \sum(0.8104, 0.7936, 0.7371) \\
&+ 0 \\
&+ \frac{1}{1} \cdot 0.33 \cdot \sum(1) \\
&= 0.6288
\end{aligned} \tag{4.7}$$

The confidence scores for the previous examples show that example 4.9 has a (too) low score for dwelling, example 4.10 has (too) low scores for dwelling and food, and example 4.11 has a (sufficiently) high score for mobility and (too) low scores for dwelling and leisure.

# Chapter 5

# Evaluating the Framework

In the previous chapter the design of the framework and all its components are explained; the actual implementation and resources that are necessary for this will be further explained in detail in the next section. In addition to the data processing model, a front-end Web application is developed to aggregate (by time and place) and visualize the framework's classification output. This application enables users to explore information on the characteristics of energy-consuming activities extracted from the social media posts generated in different cities.

In order to evaluate to what extent social media can be used as a complementary data source for the description of energy-consuming activities, the framework's performance is evaluated through a case study in Amsterdam and Istanbul. The users' opinion will be involved in our evaluation of both the framework's performance and the values we set for the data type weights.

## 5.1 Implementation

To comply with the aim of describing the output of our data processing framework at group level as well, the architecture in Figure 5.1 is proposed, which consists of multiple interconnected components: the data processing component, the data analytics component, the database, and the Web API. Separating these components facilitates different internal representations of information accepted from and presented to the user, which allows for efficient code reuse and parallel development. Furthermore, this form of modularity (separating the components), makes it easier to understand and modify each component, without any knowledge about the other components [48].

The data processing component primarily involves the data processing framework which is discussed in the previous chapter. This component collects, enriches and classifies the social media data after which it stores the processed and classified data in the database. The Web API allows the user to explore and request information on the energy-consuming activities and its characteristics. These requests by the Web API are handled by the data analytics component; when the user requests information through the Web API, the data analytics component is actuated and gathers the relevant data from the database. Then, when the component receives the requested data, it processes the data in such a way that the Web API can visualize it to the user.

**Resources** The data is stored in a document-oriented manner, using the MongoDB database. Instagram and Twitter are selected as the social media data sources (for the input data), whereas Foursquare and Google Places are chosen for the first order enrichment of our input data. The Python programming language is used for the development of the data processing framework in the model component. In addition, the Python microframework Flask[1] was used to build the Web application (view component), which was based on a Bootstrap template (which is built in HTML, CSS and JavaScript). Furthermore, the Highcharts and Leaflet libraries were used for the front-end visualizations. In Figure 5.2 and Table 5.1 an overview of all resources is provided, which will be described in more detail in the concerned sections.

### 5.1.1 Data Processing Component

As mentioned before, the data processing component primarily involves the data processing framework that is designed for extracting the main characteristics of energy-consuming activities. Different types of data, which are at first derived from our social media data sources, are processed in order to classify the social media posts. The textual, imagery and place data is subject to multiple processing techniques for which a variety of state-of-the-art algorithms is used. For each type of data, these techniques are described in more detail below.

**Text processing** If a social media post contains a text message, it is subject to several text processing techniques. The Python-based Natural Language Toolkit (NLTK[2]) is used to (pre-)process (tokenization, stopword removal, lemmatization, etc.) the text. NLTK is one of the most popular libraries in the Natural Language Processing (NLP) community; it has incorporated most of the NLP tasks, it is very elegant and easy to work with [37]. An additional advantage of using the NLTK library is that it provides implementations of both the Lesk algorithm for word sense disambiguation, and a corpus reader for WordNet for looking up a term's synonym set. Furthermore, the calculations to determine similarity scores between a term and a dictionary are based on Word2Vec's word embeddings. This neural network-based language model is a popular embedding model for learning word vectors via a neural network with a single hidden layer [63, 86]. The gensim[3] package, containing methods to calculate the distance between a term and another set of terms, was used for this. Furthermore, the pre-trained Word2Vec model on the Google News corpus[4] was used to obtain the word vector representations.

**Image processing** For object recognition, deep convolutional networks (ConvNets) have replaced the former engineered features, due to the face that (i) they are capable of representing higher-level semantics and (ii) that they are more robust to variance in

[1] http://flask.pocoo.org
[2] https://www.nltk.org/
[3] https://radimrehurek.com/gensim/models/word2vec.html
[4] https://code.google.com/archive/p/word2vec/

scale. Recently, the Feature Pyramid Network (FPN) has enabled new top results in the object detection track of the COCO competition [38, 54]. For our image processing component, multiple computer vision techniques are used to process the image included by a social media post. For the image object recognition we use two state-of-the-art pre-trained models (TensorFlow [1] and Mask R-CNN [38]) which are each trained on a different data set. The TensorFlow model is trained on Google's Open Images data set[5] (approximately nine million images annotated with one or more of the 545 selected object classes) using the `faster_rcnn_inception_resnet_v2_atrous_oid` weights. The Mask R-CNN method is an implementation on Python3, Keras and TensorFlow and extends the Faster R-CNN method. It is based on the FPN and the ResNet101 backbone. The model is trained on the Microsoft COCO dataset by using the pre-trained `mask_rcnn_coco.h5` weights.

Many scene-centric datasets (among others, the Scene15 database [49], the MIT Indoor67 database [69], and the SUN database [83]) are relatively small in size compared to current object datasets such as OpenImages. For that reason, the Places database was presented, containing more than 10 million images comprising 363 unique scene categories, which was large enough to train algorithms that require huge amounts of data, such as CNNs [88]. For the scene recognition component in our model, we incorporated the CNN model based on the ResNet50 backbone[6], which is pre-trained on the Places[7] data set.

**Place processing** In order to enrich a place (including coordinates and a place name) based on the *N* order (meta) data approach–introduced the previous chapter–the distance between posts (i.e., the distance between a user's current post's check-in and the one from the previous post), and home detection methods are to be implemented. To determine the user displacement, GeoPy[8]'s `distance` method is used. Furthermore, the DBSCAN algorithm [27] is used to determine the spatial clusters from which the user's home is derived. Along with those two enrichment methods, Python's shapely library is used to determine to which area the set of coordinates belongs. For Amsterdam, we used the pre-defined 97 neighborhood combinations combinations[9] as areas. As for Istanbul, we stick to the 39 districts identified by the Central Disemmination System[10].

[5] https://storage.googleapis.com/openimages/Web/index.html
[6] https://github.com/CSAILVision/places365
[7] http://places2.csail.mit.edu/index.html
[8] https://github.com/geopy/geopy
[9] http://decentrale.regelgeving.overheid.nl/cvdr/images/Amsterdam/i224394.pdf
[10] https://biruni.tuik.gov.tr/medas/?kn=95&locale=en

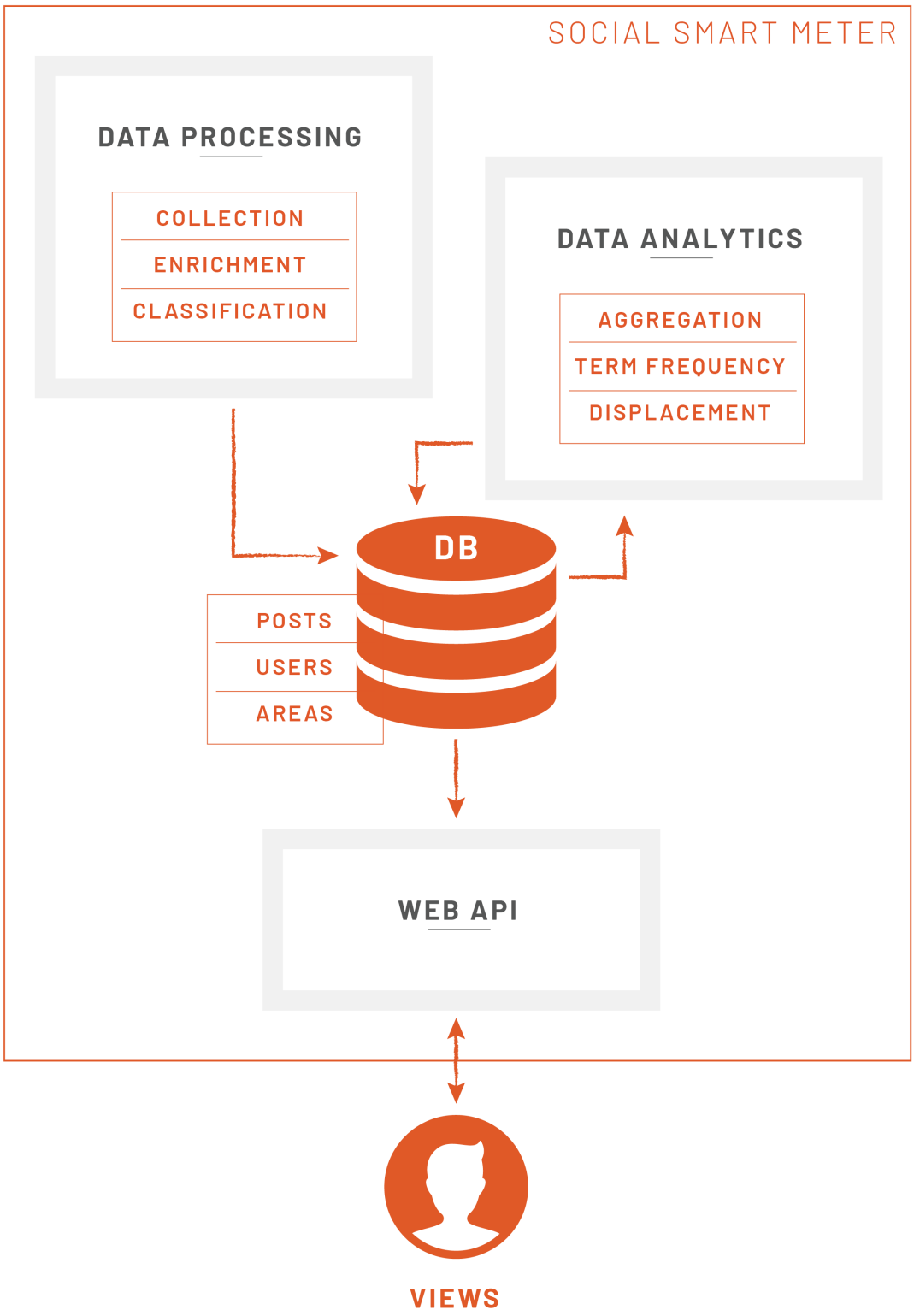



Figure 5.1: Overview of the architecture of the framework

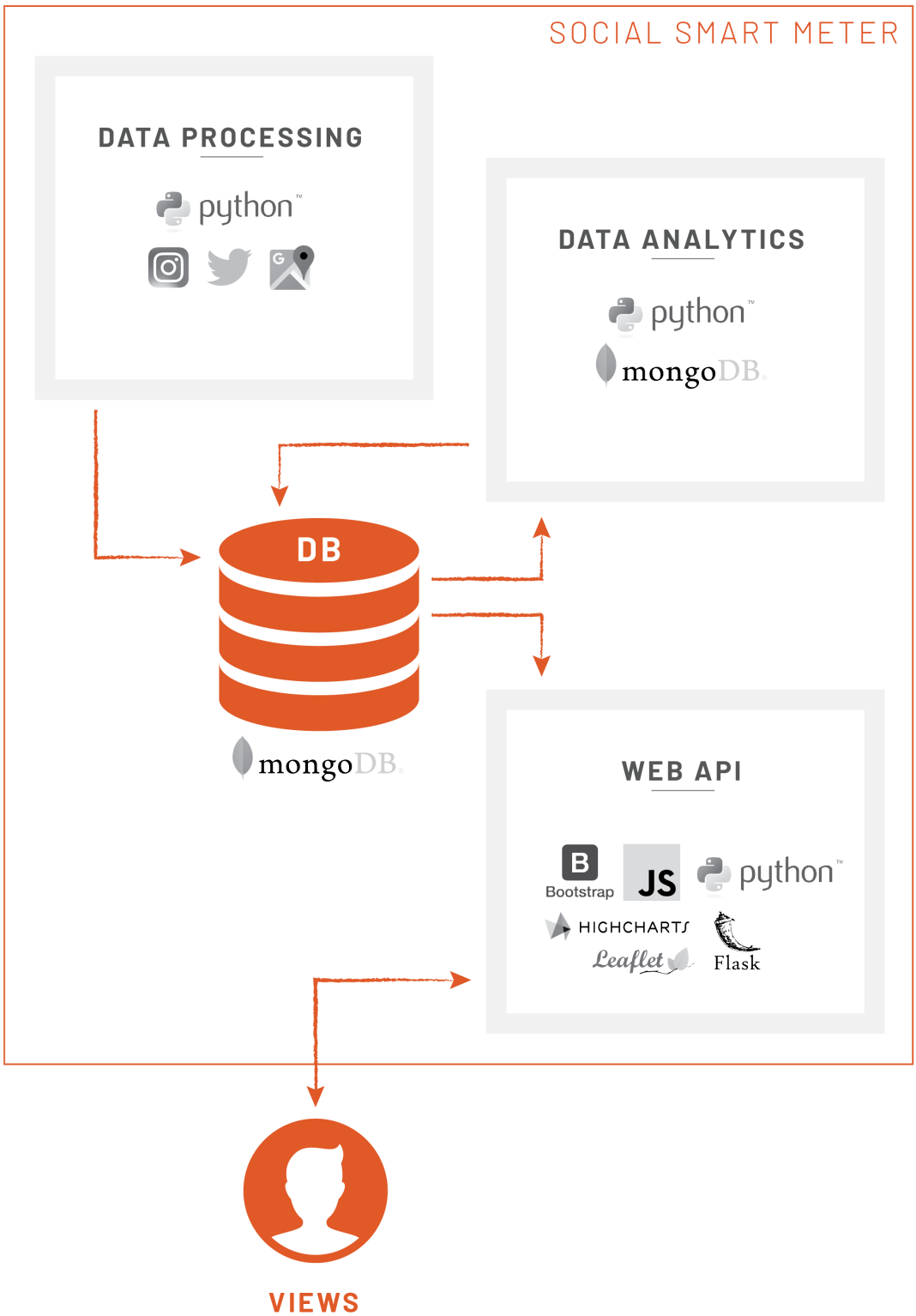



Figure 5.2: Overview of the resources used within the framework

### 5.1.2 Data Analytics Component

The data analytics component is responsible for preparing the data in such way that it can be retrieved by the Web API, including (i) aggregation of the classified data by time and area, (ii) determining the relevant term frequency per category of energy-consuming activities per data type, and (iii) determining the displacement frequency (based on the distance between posts per user). When a user makes a request through the Web API to explore data aggregated by a particular area and time span, these parameters are passed to the data analytics component. Here, a MongoDB view on the data base collection is created which enables the data to be mapped in such a way that it can be prepared for the aggregation, term frequency, and displacement analytics.

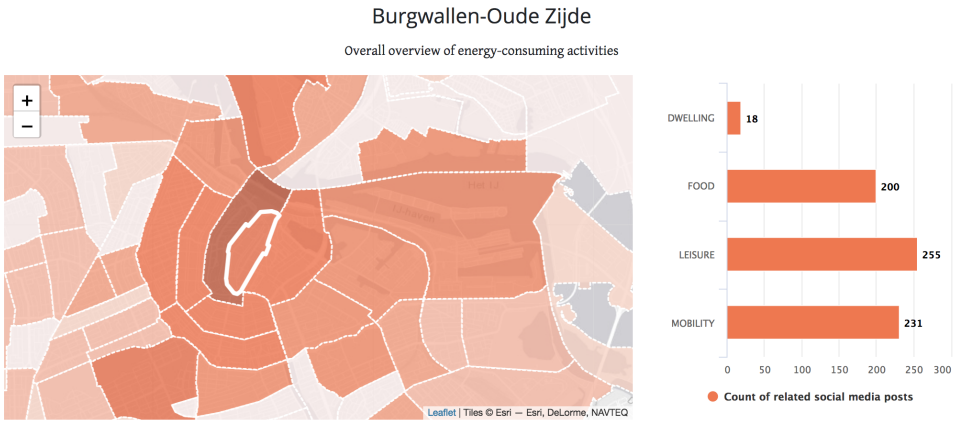
**Aggregation by time and area** Each classified social media post is characterized by a timestamp and a set of coordinates; this allows to categorize the social media posts per area (i.e., one of the target city's neighborhoods) and per time span. Our Web application allows the viewer to make requests for a particular neighborhood per day. Therefore, the classified output is aggregated per neighborhood per day.

**Term frequency** In order to provide more insights into the different energy-consuming activities per category, the relevant term frequency per category per data type is determined - both at neighborhood and at city level. Per data type the top 10 most frequent relevant terms are selected for each category of energy-consuming activities after which they are passed to the Web application.

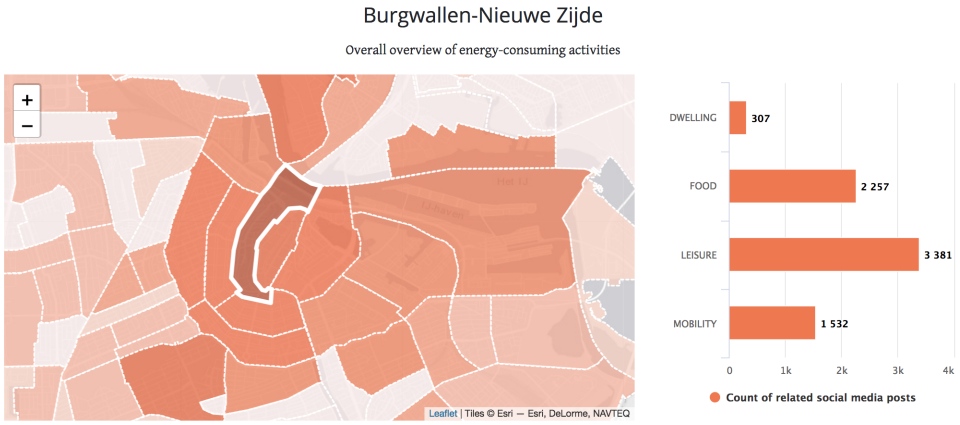
**Displacement frequency** For each social media post our data processing component determines the distance to the previous post by the creator of this post. Based on these distances, we can determine the displacement (per kilometer) frequency per area per day. These numbers provide more insights into the type (and extent) of the mobility activities referred to by the social media posts.

### 5.1.3 Web API

To present the results, interactive visualizations are shown through the Web API, which is the base for our front-end Web application. Interactive maps are created using the Leaflet JavaScript-based library; for each neighborhood in a city, the social media posts related to energy-consuming activities are aggregated. The number of related social media posts indicate the extent of color saturation of this neighborhood; it is scaled based on the number of social media posts related to energy-consuming activities - e.g., from Figure 5.3 we can deduce that more social media posts related to energy-consuming activities are created in Burgwallen-Nieuwe Zijde than in Burgwallen-Oude Zijde.




(a) Neighborhood (Burgwallen-Oude Zijde) with a low(er) count of social media posts classified to energy-consuming activities




(b) Neighborhood (Burgwallen-Nieuwe Zijde) with a high(er) count of social media posts classified to energy-consuming activities

Figure 5.3: Comparing the count of social media posts classified to energy-consuming activities between two neighborhoods in Amsterdam

When the user requests information through the application, the Web API actuates the data analytics component, which accesses the relevant data from the data base through the view. Hereafter, the visualizations are updated based on this aggregated data view. The application enables the user to select a city (Amsterdam or Istanbul), possibly one of its neighborhoods, and a time span (per day).

Furthermore, for both levels of granularity (neighborhood versus entire city) the ratio of the different categories of energy-consuming activities is displayed, using a bar chart. These statistics provide more insights into how many social media posts are created per neighborhood or city that are related to one of the categories of energy-consuming activities. If these numbers and ratios indeed reflect the ones of energy-consuming activities performed in the physical world should be evaluated through our case study (Section 5.3 *Experimental Case Study*).

Besides the interactive maps and bar charts described above, other bar charts are displayed as well. Besides the count of classified social media posts, the top 10 most frequent relevant terms derived from the classified social media posts are presented as well. These are shown for each category of energy-consuming activities (dependent on which category is requested by the viewer) or for the total of all categories. For the mobility category, a bar chart representing the displacement frequency is also shown.

| **DATA PROCESSING** | | |
|---|---|---|
| **Module** | **Component** | **Resource** |
| Collection | Collecting social media data<br>Collecting enrichment data | Instagram API, Twitter API<br>Foursquare API, Google Places API |
| Enrichment (text) | Pre-processing text messages<br>Word sense disambiguation<br>Similarity scoring<br>Translation of terms | NLTK library<br>Lesk algorithm<br>Word2Vec model<br>Googletrans |
| Enrichment (images) | Image object recognition<br>Image scene recognition | Mask R-CNN, Tensorflow<br>Places-365 CNN |
| Enrichment (places) | Place matching<br>Distance between posts<br>User home detection | -<br>GeoPy library<br>DBSCAN algorithm |
| Classification | Dictionaries<br>Rule-based approach | -<br>- |
| **DATA ANALYTICS** | | |
| **Module** | **Component** | **Resource** |
| Statistics | Database views | (Py)Mongo |
| **WEB API** | | |
| **Module** | **Component** | **Resource** |
| Front-end application | Template<br>Web framework | Bootstrap<br>Flask |
| Visualizations | Maps<br>Charts | Leaflet<br>Highcharts |

Table 5.1: Overview of the resources used for the implementation of the framework

## 5.2 User-Based Evaluation

The results of our case study (i.e., the classified social media posts generated in Amsterdam and Istanbul) can be used for the evaluation of the framework's performance. This performance can be evaluated using several metrics, such as accuracy, precision, recall, and F1-score. Before we can calculate these metrics, we need to find the ground truth for the classification of the social media posts. Since this is a subjective matter, we suggest to involve a set of users to determine the ground truth of the classification of the social media posts. By asking a sufficient large set of users to assess whether a social media post relates to an energy-consuming activity (and if so, to what category (or categories) of energy-consuming activities), we could use the average of these votes (wisdom of the crowd) to determine ground truth. Hereafter, this ground truth could be compared to the output of our framework. Based on these comparisons, the evaluation metrics could be calculated.

In addition, we have introduced data type weights as input parameters for our classification confidence model in the previous chapter. The values of these weights are only based on the intuition of a single expert and should thereby be evaluated. For this evaluation we also suggest to use the wisdom of the crowd, by asking users to rank the informativeness of each data type (text, image, place) for each category of activities the social media post is classified to. A sample of 100 social media posts was used for the user evaluation surveys. For each social media post multiple questions were asked, as shown in Figures 5.4 and 5.5. To divide the work load for the users, the sample was divided into five sub-samples. For each sub-sample, a survey was created. In total (for all user evaluation surveys), 43 responses were collected - i.e., on average, approximately nine responses per survey were received.

The majority (> 90%) of the respondents were Master students of Delft University of Technology from different faculties. Each survey was composed of mainly (88% on average) social media posts that were classified - i.e., classification confidence that exceeded a threshold of 0.3, based on the initial data type weights - to at least one of the categories of energy-consuming activities. Respectively, 13, 38, 56, and 32 social media posts were classified to the categories of dwelling, food consumption, leisure, and mobility.

### 5.2.1 Evaluation of Data Type Weights

The initial data type weights have been assigned equal values - i.e., all values were initialized to 0.33. By incorporating the users' opinion, the values can be evaluated and tuned to some extent. The users were asked to rank how informative each data type is (on a scale from 0 to 10) for each classified category in each social media post (Figure 5.5).

The users' average rankings are displayed in Table 5.2 and were adopted as data type weights in the classification module in the data processing pipeline for our case study in Amsterdam and Istanbul. The weight values do not deviate a lot from each other. Yet, we observe that the users find images most and places least informative to describe dwelling activities. The same applies for food consumption activities.

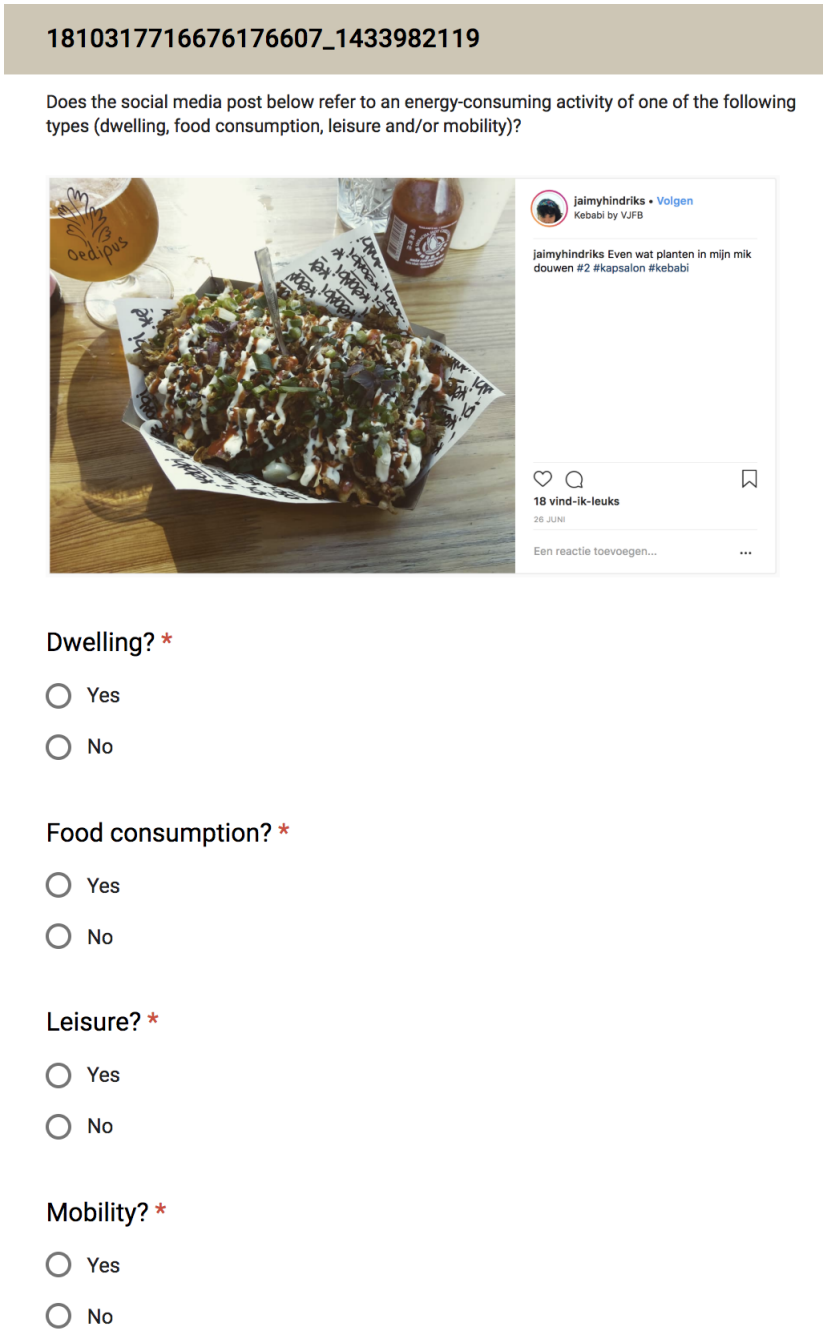



Figure 5.4: Example question for the evaluating users (ground truth)

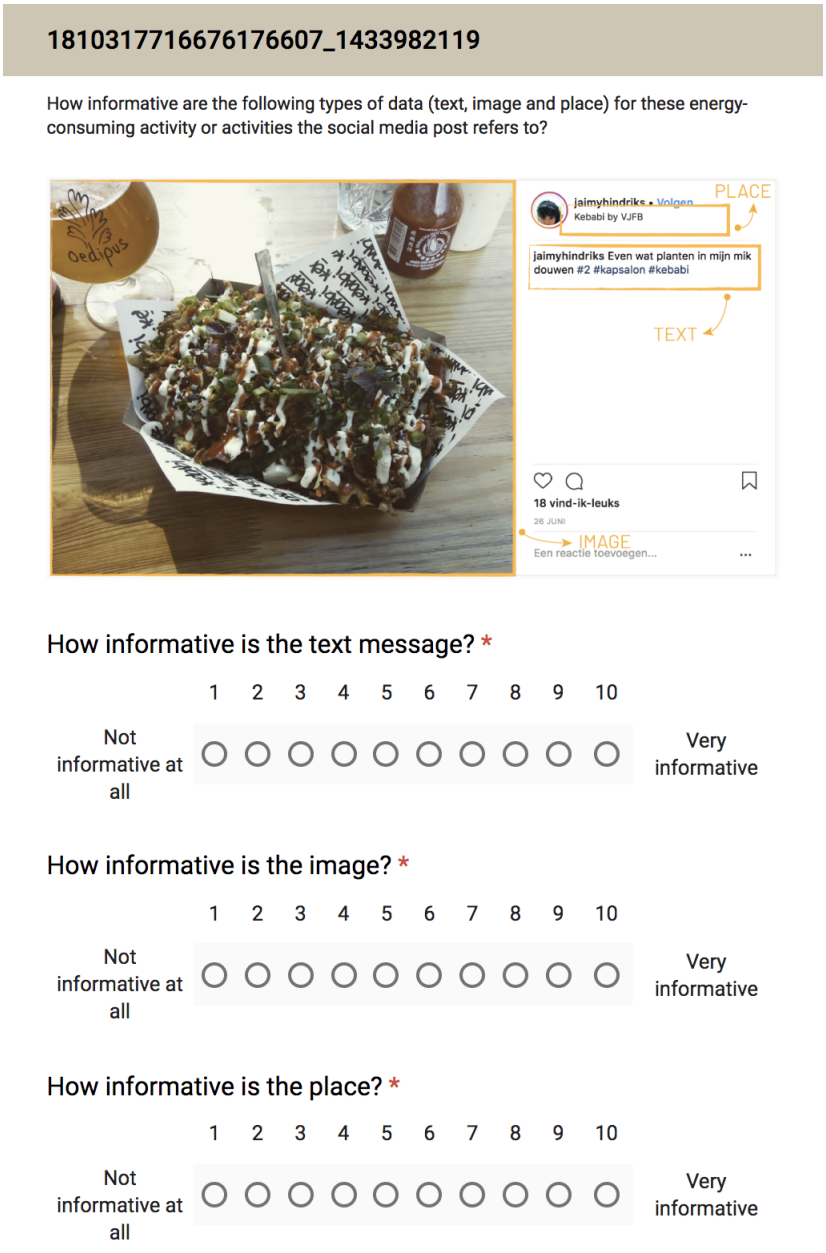



Figure 5.5: Example question for the evaluating users (data type weights)

For leisure activities, text was found most informative whereas images were found least informative. And as for mobility activities, again text was ranked to be most informative whereas places were ranked to be least informative.

| Category \ Data type | Text | Image | Place |
|---|---|---|---|
| Dwelling | 0.35 | 0.40 | 0.25 |
| Food consumption | 0.33 | 0.37 | 0.30 |
| Leisure | 0.35 | 0.32 | 0.33 |
| Mobility | 0.37 | 0.33 | 0.30 |
| **Total (average)** | **0.35** | **0.35** | **0.29** |

Table 5.2: Average (normalized) numbers of the users' vote for informativeness per data type and category of energy-consuming activities

The users' opinion regarding the informativeness of the data types differed from our hypothesis - i.e., the values set in Table 4.3 (Chapter 4). For each category, the informativeness values for the different data types deviate less from each other compared to the values in Table 4.3. Furthermore, the set of users had a different judgment for which data types are most and least informative per category - e.g., the users judged the text data type to be most informative to identify a leisure activity whereas we considered the place data type to be most informative.

Furthermore, the set of users faced a couple of possible difficulties during the survey. In case the users did not understand the language, they were likely to select a rating of 1 for the informativeness of the text. Besides that, it is feasible that the users did not know all places included in the social media posts. In case one did not search for the place to determine the place categories, it is likely he or she rated the informativeness of the place with 1.

### 5.2.2 Evaluation of the Framework's Performance

The framework's performance is measured by evaluating its accuracy, precision, recall, and the F1-score. To determine these metrics, we ask the set of users (through a survey) to assess whether the social media post is related to an energy-consuming activity and if so, to which category or categories (Figure 5.4). Only social media posts that have a classification confidence score higher than the threshold are classified to energy-consuming activities. Based on the evaluation metrics, the threshold with the best scores can be determined.

In Tables 5.3 to 5.6 the evaluation metric scores are provided for each category of energy-consuming activities individually, as well as for the total of energy-consuming activities. The evaluation metric scores are calculated for different classification thresholds (0.3, 0.4, 0.5, 0.6 and 0.7), in order to find the best-performing one. The

| **Metric** / **Category** | Accuracy | | | | | |
|---|---|---|---|---|---|---|
| *Threshold* | *0.30* | *0.35* | *0.40* | *0.50* | *0.60* | *0.70* |
| Dwelling | 0.85 | 0.87 | 0.89 | 0.90 | 0.90 | 0.91 |
| Food consumption | 0.82 | 0.84 | 0.86 | 0.85 | 0.78 | 0.73 |
| Leisure | 0.60 | 0.57 | 0.56 | 0.54 | 0.48 | 0.37 |
| Mobility | 0.81 | 0.82 | 0.82 | 0.82 | 0.80 | 0.74 |
| **Total** | **0.77** | **0.78** | **0.78** | **0.78** | **0.74** | **0.69** |

Table 5.3: Evaluation metrics (accuracy) of the framework's performance

| **Metric** / **Category** | Precision | | | | | |
|---|---|---|---|---|---|---|
| *Threshold* | *0.30* | *0.35* | *0.40* | *0.50* | *0.60* | *0.70* |
| Dwelling | 0.23 | 0.27 | 0.20 | 0.00 | 0.00 | 0.00 |
| Food consumption | 0.68 | 0.79 | 0.95 | 0.95 | 0.92 | 1.00 |
| Leisure | 0.80 | 0.88 | 0.89 | 0.87 | 0.89 | 1.00 |
| Mobility | 0.63 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| **Total** | **0.59** | **0.73** | **0.76** | **0.70** | **0.70** | **0.75** |

Table 5.4: Evaluation metrics (precision) of the framework's performance

framework's overall accuracy varies from 0.69 to 0.78. The accuracy for the classification of leisure activities is relatively low compared to the other categories due to many false negatives - i.e., social media posts that are not classified to leisure while, based on ground truth, they should be. Furthermore, the precision for dwelling activities is rather low whereas the accuracy is relatively high due to many true negatives - i.e., social media posts that (based on ground truth) do not refer to dwelling activities and are indeed not classified to this category by our classification model.

In Figure 5.6 the evaluation metric scores are plotted for the different threshold values. As expected, the recall scores decrease while increasing the threshold - i.e., less and less relevant social media posts have sufficient high confidence scores to exceed the threshold. As for the precision, we observe that the scores are fluctuating for different threshold values. Increasing the threshold results in less true positives, as well as less false positives. However, the numbers of true and false positives do not decrease proportionally. Also, there are very few social media posts with a high confidence score for dwelling. For a threshold greater than 0.4, the precision is zero for dwelling. Both aforementioned reasons contribute to the fluctuating precision scores.

| Metric / Category | Recall | | | | | |
|---|---|---|---|---|---|---|
| *Threshold* | *0.30* | *0.35* | *0.40* | *0.50* | *0.60* | *0.70* |
| Dwelling | 0.38 | 0.38 | 0.13 | 0.00 | 0.00 | 0.00 |
| Food consumption | 0.81 | 0.69 | 0.59 | 0.56 | 0.34 | 0.16 |
| Leisure | 0.61 | 0.49 | 0.46 | 0.45 | 0.34 | 0.15 |
| Mobility | 0.74 | 0.33 | 0.33 | 0.33 | 0.26 | 0.04 |
| **Total** | **0.63** | **0.47** | **0.38** | **0.34** | **0.24** | **0.09** |

Table 5.5: Evaluation metrics (recall) of the framework's performance

| Metric / Category | F1-score | | | | | |
|---|---|---|---|---|---|---|
| *Threshold* | *0.30* | *0.35* | *0.40* | *0.50* | *0.60* | *0.70* |
| Dwelling | 0.29 | 0.32 | 0.15 | - | - | - |
| Food consumption | 0.74 | 0.73 | 0.73 | 1.00 | 0.50 | 0.27 |
| Leisure | 0.69 | 0.63 | 0.61 | 0.33 | 0.49 | 0.26 |
| Mobility | 0.68 | 0.50 | 0.50 | 0.50 | 0.41 | 0.07 |
| **Total** | **0.60** | **0.54** | **0.50** | **0.60** | **0.47** | **0.20** |

Table 5.6: Evaluation metrics (F1-score) of the framework's performance

Based on Figure 5.6 a threshold of either 0.30 or 0.35 appears to result in the best evaluation metric scores. For a threshold of 0.30, a precision of 0.59 is obtained whereas a threshold of 0.35 results in a precision of 0.73. Furthermore, these thresholds (0.30 and 0.35) respectively result in recall scores of 0.63 and 0.47 and in F1-scores of 0.60 and 0.54. Based on the F1-score, a threshold of 0.30 seems to be better performing. Yet, it is dependent on the context whether it is more important to have a higher precision or recall score - i.e., whether it is more important to classify as many social media posts as possible correctly or to discover as many as possible that are referring to energy-consuming activities. In case the quantity of energy (in terms of kWh consumption or CO2 emission) during an activity is analyzed, a higher precision is considered more beneficial. However, when a qualitative overview of all energy-consuming activities performed by an individual is required, it is more advantageous to have a higher recall score. For our case study in Section 5.3, a threshold of 0.35 was selected.

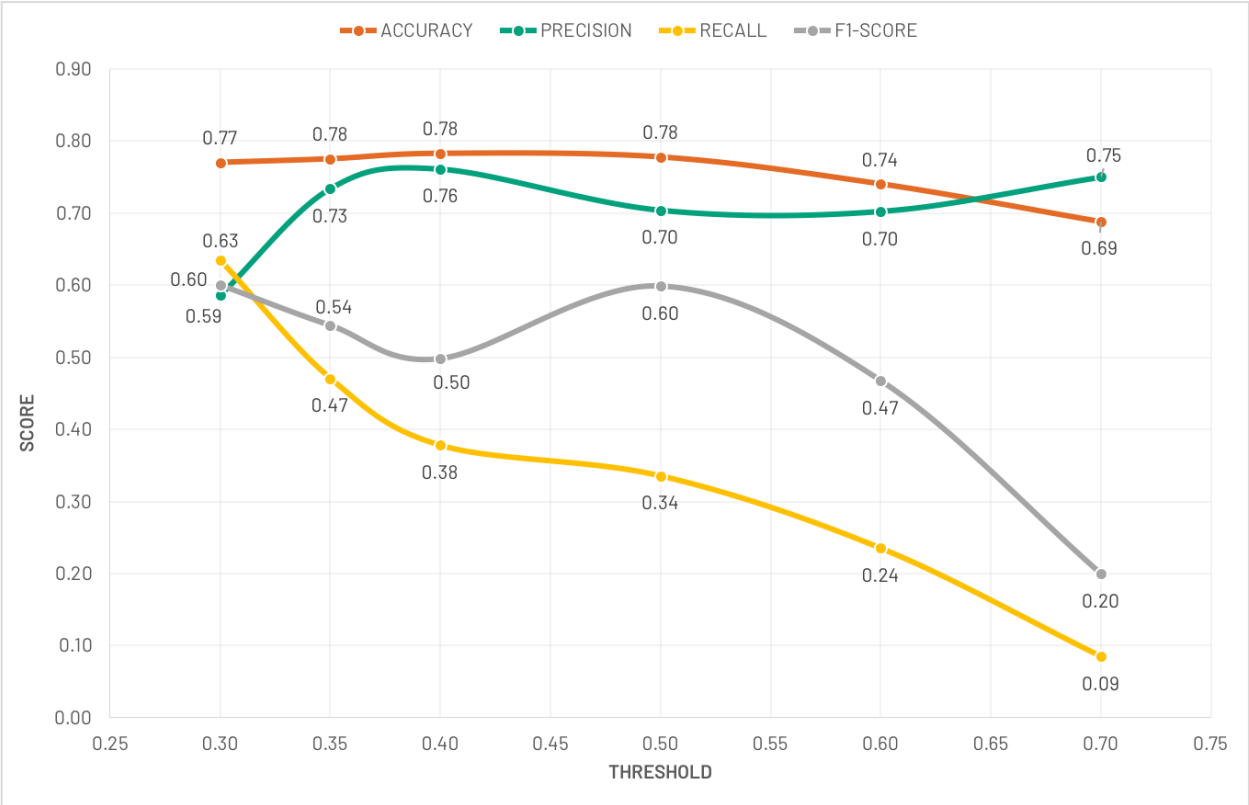



Figure 5.6: Evaluation metrics

### 5.2.3 Evaluation of the Social Smart Meter Ontology

In Chapter 3 the application-based evaluation approach was adopted for the evaluation of our Social Smart Meter Ontology. The framework's outputs are partly dependent on the ontology. Hence, its performance can be used as an indicator for the quality of the ontology. The high precision score (0.73) may indicate that the terms in our ontology are indeed relevant for energy-consuming activities. The lower recall score might also imply that our ontology lacks some concepts. However, it is hard to determine whether the low recall score is caused by the lack of relevant concepts in our ontology or the inability of our data enrichment methods and techniques to identify those concepts in the social media data.

The application-based evaluation approach also has some downsides that should be taken as a side note. It is difficult to generalize the observation about the quality of the ontology in a particular task in the framework - e.g., determining the relationship between an object recognized in the image and the corresponding energy-consuming activity. Moreover, the output of the framework is just partly dependent on the ontology; thereby, the ontology's effect on the framework's performance might be small and indirect.

## 5.3 Experimental Case Study

In this work we use a case study as a means to evaluate our framework. Applying the framework to real social media posts and analyzing the results helps to understand if the framework makes sense. Since the behavior regarding creating social media posts might differ between cities with a different culture, we propose to perform a case study for two different cities: Amsterdam and Istanbul. This allows us to evaluate our framework under different conditions.

### 5.3.1 Data Set Creation

Amsterdam and Istanbul are the target cities for our case study; therefore, a data set should be created in which social media posts (on Instagram and Twitter) generated in either one of the cities (i.e., the post's coordinates should correspond to either one of the cities) are collected for a particular time span. For this study, we collected data from June 22nd until June 27th, and July 27th until July 28th, 2018. At first, only social media posts created in Amsterdam was collected to provide a first round of insights into the Social Smart Meter framework. Hereafter, social media posts created in Istanbul were collected as well in order to compare the results between the two cities. An overview of the numbers of collected social media posts is provided in Table 5.7. In line with our expectations from Chapter 4, we can see that Instagram is used more often to create social media posts than Twitter.

| **Date** | **Amsterdam** | | **Istanbul** | |
|---|---|---|---|---|
| | *Instagram* | *Twitter* | *Instagram* | *Twitter* |
| 22/06/2018 | 16099 | 3602 | - | - |
| 23/06/2018 | 15794 | 3220 | - | - |
| 24/06/2018 | 16365 | 2594 | - | - |
| 25/06/2018 | 15426 | 3024 | - | - |
| 26/06/2018 | 14985 | 3685 | 19887 | 4476 |
| 27/06/2018 | 16966 | 1929 | 28346 | 8931 |
| 27/07/2018 | 17854 | 1684 | 22127 | 4818 |
| 28/07/2018 | 17779 | 3656 | 21082 | 11522 |
| **Total** | **131268** | **23394** | **91442** | **29747** |

Table 5.7: Number of collected social media posts per day

Besides that, we observe that, in general, more social media posts are created in Istanbul than in Amsterdam. Given that Istanbul's population (approximately 15 million inhabitants) is more than 15 times as large as Amsterdam's population (approximately 1 million inhabitants), this is not very strange.

Furthermore, certain events in either of the cities of Amsterdam and Istanbul might influence the number of posts that refer to a particular category of energy-consuming

activities - e.g., the public transport strike in Amsterdam might have led to more people creating social media posts about this strike which results in a higher number of social media posts classified to mobility. However, no significant differences in the data were found during our selected time span.

### 5.3.2 Changes Compared to Framework Design

While running the Social Smart Meter model on the case study's data, we came across a few difficulties. This resulted in some changes in the model compared to the framework design proposed in Chapter 4. These changes (and the underlying reasons) are described in more detail below.

**Image object recognition** In Section 5.1 (*Implementation*) which elaborates on the framework's implementation, we proposed to include two different image object recognition models in the framework (since both models were trained on different data sets). Yet, we found that this significantly affects the time it costs to process a social media post. Since (i) the Mask R-CNN model processes images a lot faster than the TensorFlow model, and (ii) Tensorflow's image object classes did not seem to be a significant added value compared to the Mask R-CNN ones in order to classify the images to energy-consuming activities, we only integrated the first one in our model for this case study.

**Home detection** In Section 4.4.3 (*Place Processing*) we proposed to use a DBSCAN clustering method to detect a user's home. For each user we collected all sets of coordinates (or locations) derived from the social media post this user created. In practice, we found that the timespan of our collected data set was too short. This resulted in small lists of locations per user, which were not sufficient large for our clustering algorithm to function properly. Moreover, a few users did have sufficient locations. However, for these users we found that the centers of the high-dense clusters were often near the city's centroid. This presumably has two reasons: (i) many social media users check into the city (thus, not at a specific venue) where they are at while creating the post resulting in a set of coordinates equal to the centroid of the city, and (ii) few social media users specifically check in at their home place (i.e., with an accurate set of coordinates) when they create a social media post at home.

**Word Sense Disambiguation** Since a word can have multiple meanings, we proposed to match the detected words from a message to our dictionary by means of its word sense. Yet, the Lesk algorithm that was integrated in our framework was only applicable to English text messages. Since the translation of the foreign-language messages to English messages resulted in more difficulties than expected, the word senses were too hard to capture. Hence, for this case study, the word sense disambiguation was omitted.

### 5.3.3 Analysis of the Framework's Results

In our data set there seem to be few differences between the data of different dates. Hence, we take all collected data into account per city to perform the analyses. For

both Amsterdam and Istanbul different visualizations are created. First, an overall overview of the count of social media posts that are classified to any of the categories of energy-consuming activities is provided in Figure 5.7. Hereafter, for each category, the results for both cities are visualized using a map and bar charts. Per category, we describe our observations regarding the similarities and differences between both cities.

| Category | Amsterdam | Istanbul |
|---|---|---|
| Dwelling | 3.25% (1,326) | 4.18% (589) |
| Food consumption | 20.36% (8,312) | 21.99% (3,100) |
| Leisure | 44.75% (18,274) | 41.49% (5,850) |
| Mobility | 31.64% (12,921) | 32.35% (4,561) |
| **Total** | **100% (40,833)** | **100% (14,100)** |

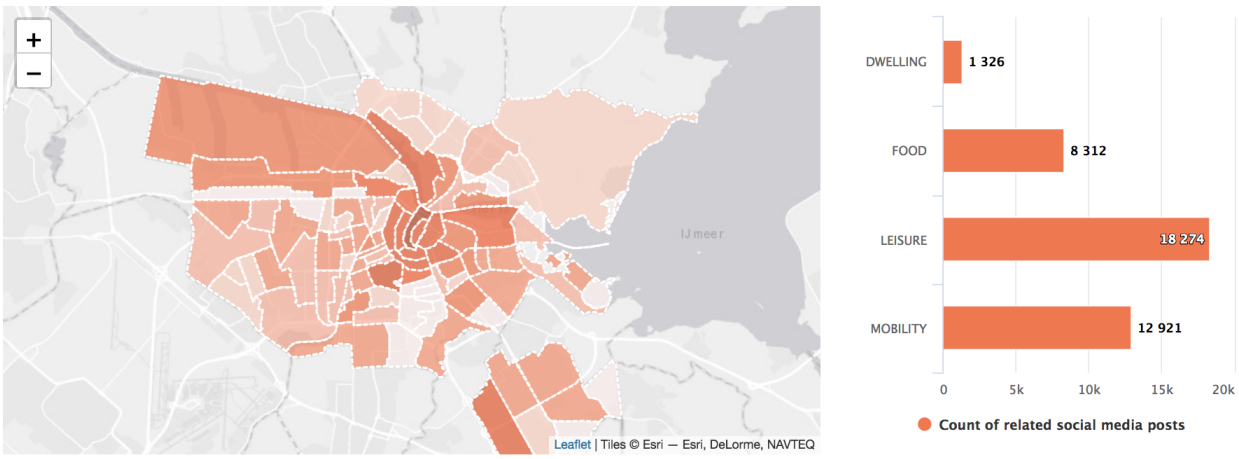
Table 5.8: Percentage of classified social media posts per category of energy-consuming activity

Table 5.8 shows the percentage of each category of energy-consuming activities for both cities. In general, we observe that few social media posts are classified to dwelling. Our rule-based classification approach demands evidence for the user being at home before it classifies a post to dwelling. It is very difficult to derive this evidence from the social media post's (meta)data, which is why so few posts are classified to this category. Furthermore, the mobility category has the largest share in both cities. Most of the posts in this category are classified to mobility by the rule-based approach - i.e., a significant distance to the previous post by this user was determined.

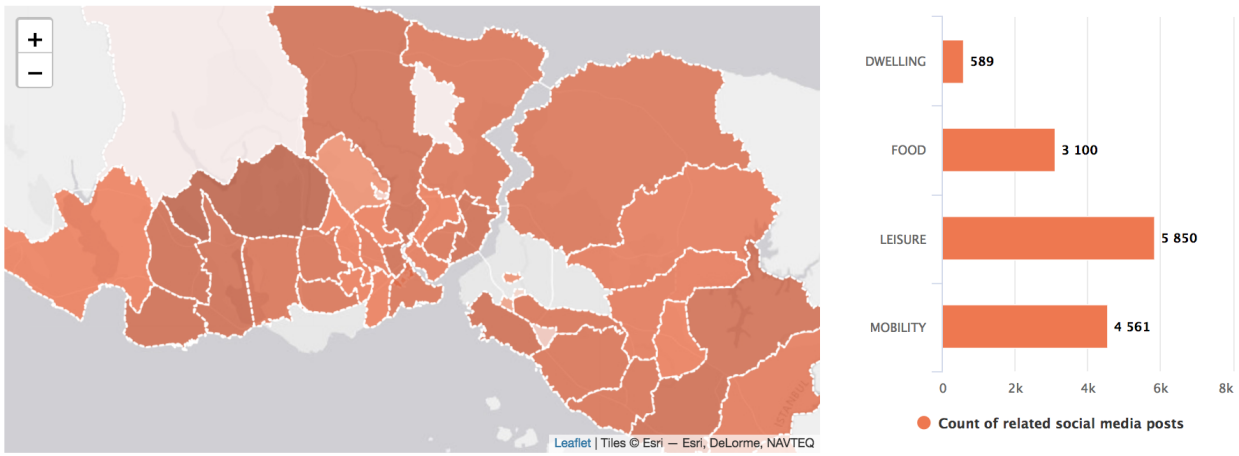
For both Amsterdam and Istanbul, the leisure category has the largest share (approximately 40%) compared to the other categories. The mobility category has the second largest share (approximately 30%). The category of food consumption has a rather small share (approximately 20%). However, nearly all social media posts that are classified to food consumption are also classified to leisure based on the rule-based approach - a food consumption activity that is performed at some other place than home is also considered a leisure activity. This explains why the share of the leisure category is more than twice as large as the share of the food consumption category.

The distribution of social media posts classified to energy-consuming activities cities also differs between both cities. For Amsterdam, most social media posts are created around the city center - the neighborhood with the highest density (Burgwallen-Nieuwe Zijde) also covers the coordinates of the city's centroid. For Istanbul, the social media posts are more evenly distributed over the different neighborhoods than for Amsterdam.

**Dwelling** For both cities, few social media posts are classified to dwelling, presumably due to the explanation that was previously described - i.e., no evidence




(a) Overall overview Amsterdam




(b) Overall overview Istanbul

Figure 5.7: Overall overviews of Amsterdam and Istanbul

indicating that a user is at home. For Amsterdam (Figure 5.8), the few posts in this category were mainly created in the city center. For Istanbul (Figure 5.18), the posts are more evenly distributed. In case no specific place is specified for a post, the post might be created at home. Yet, these posts cannot be mapped to a particular neighborhood, neither can they be visualized on either of the maps in Figures 5.8a and 5.9a.

The text terms that are most informative for a dwelling activity in Amsterdam are “House", “TV", and “gaming". In images, “tv", “laptop", and “keyboard" are the most frequent recognized objects that indicate a dwelling activity for both cities. No informative place terms were found in the social media posts.

**Food Consumption** Again, for both cities, we observe a peak of food consumption-classified social media posts in the neighborhoods that cover the centroid’s coordinates of the city. Based on the top frequent terms in Figures 5.10b

and 5.11b, images seem to be most informative to identify food consumption activities; compared to text and place terms, there are significantly more relevant image annotations indicating a food consumption activity. Furthermore, “food" and “coffee" were the top frequent text terms indicating a food consumption activity in both cities, whereas “cup", “dining table" and “bowl" were the top frequent image annotations. Besides that, individuals appear to create food consumption-related post most often while checking in at a “Bar" (Amsterdam), “Cafe" (both cities) or “Restaurant" (both cities).

**Leisure** The distribution of the leisure-related social media posts over Istanbul’s neighborhoods is rather similar to the food consumption-related distribution: most dense in the center (including the centroid) and west of the center somewhat more dense than east of the center.

In Figure 5.12a the distribution of social media posts in Amsterdam classified to leisure activities seems to be somewhat more distributed over the different neighborhoods. When zooming in on a few particular neighborhoods (Burgwallen-Nieuwe Zijde, Museum- kwartier, and Amstel III/Bullewijk) in Figures 5.13 to 5.15, some interesting observations are made.

In general, the city center (Burgwallen-Nieuwe Zijde) is characterized by many tourists, who are partying, visiting the flower markets, going to museums or enjoying the canals, among other things. This is reflected in the top frequent text terms in Figure 5.13b: “night", “holiday", “party" (text), “Flower Shop", “Art Museum", and “Hotel" (terms) are some terms that comply with these activities.

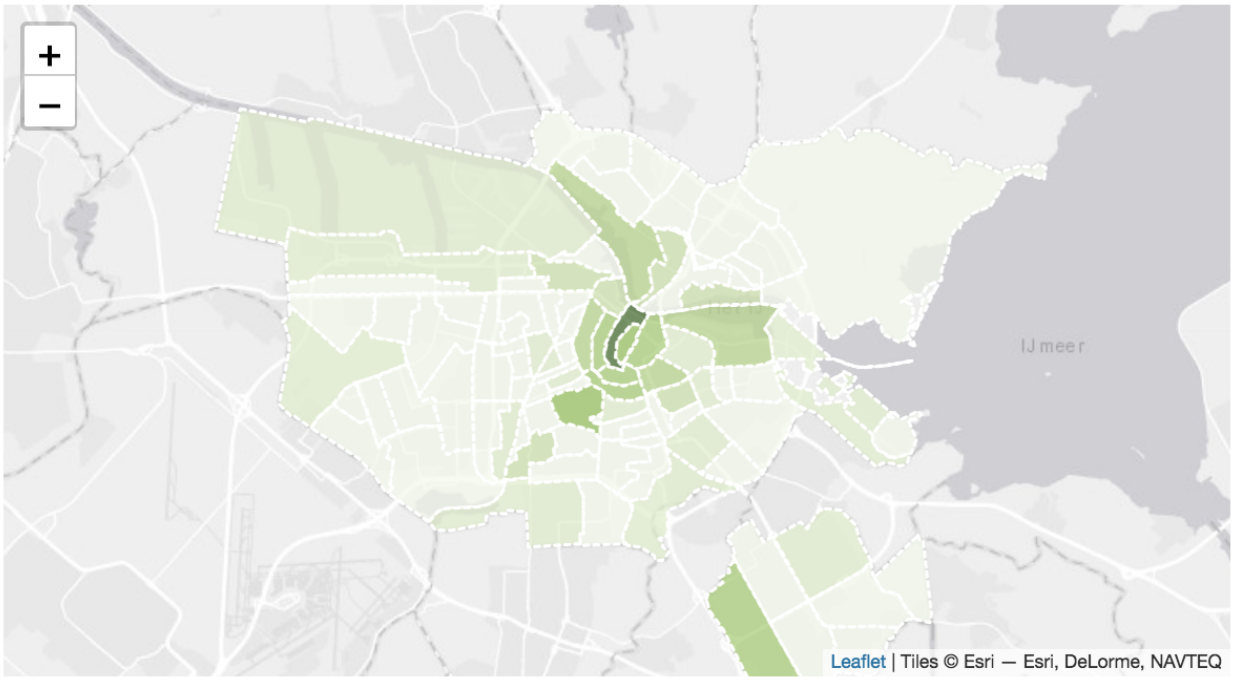
Museumkwartier is the neighborhood where many of Amsterdam’s most famous museums are situated. The bar charts in Figure 5.14b match the hypothesis that in this neighborhood a lot of social media posts related to these museums are created: “museum" (text), “art_gallery" and “museum/indoor" (image), and “Art Museum" (place) are the top frequent terms indicating leisure activities.

Amstel III/Bullewijk is known for Amsterdam’s soccer stadium and the major concert halls. The bar charts in 5.15b reflect this: “concert" and “music" (text), “arena/performance" and “stage/indoor" (image), and “Concert Hall" and “Soccer Stadium" (place) are the most frequent terms.

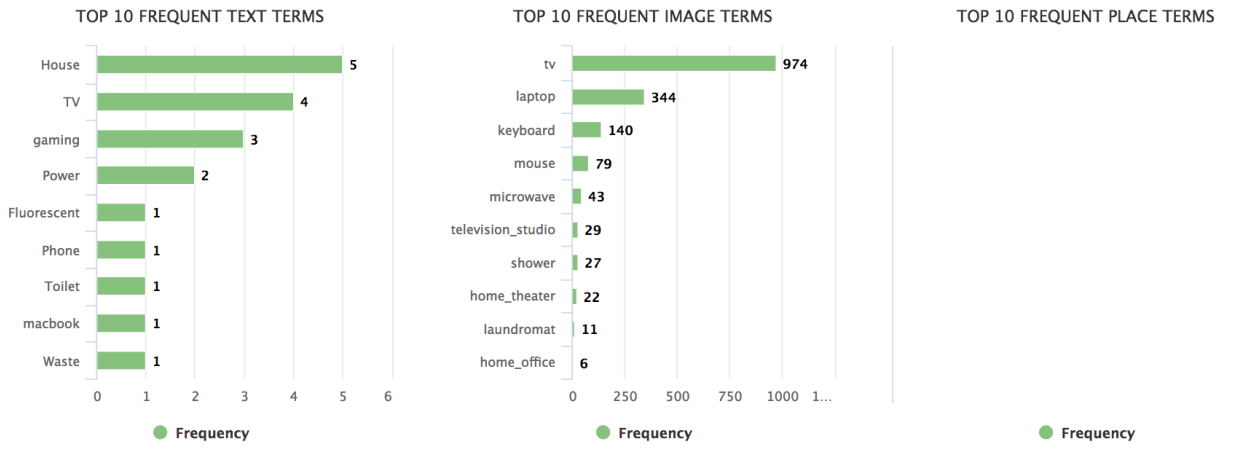
**Mobility** Since Amsterdam’s train station is situated in the city center, it makes sense that this neighborhood is most dense regarding the count of social media posts classified to mobility (Figure 5.17a). This is also due to the canal trips in the city center that individuals (mainly tourists) tend to post about.

In Figure 5.18a one of the western neighborhoods (Basaksehir) is most dense regarding mobility activities. Multiple highways run through this neighborhood, as well as a large highway junction. In line with this, “Gas Station" is the top frequent place term for this neighborhood.

If we compare the frequencies of displacements of both cities (Figure 5.19) we can observe that the displacements (or distance between posts) cover larger distances in Istanbul. Since Istanbul is significantly larger in size than Amsterdam, this is in line with our expectations.

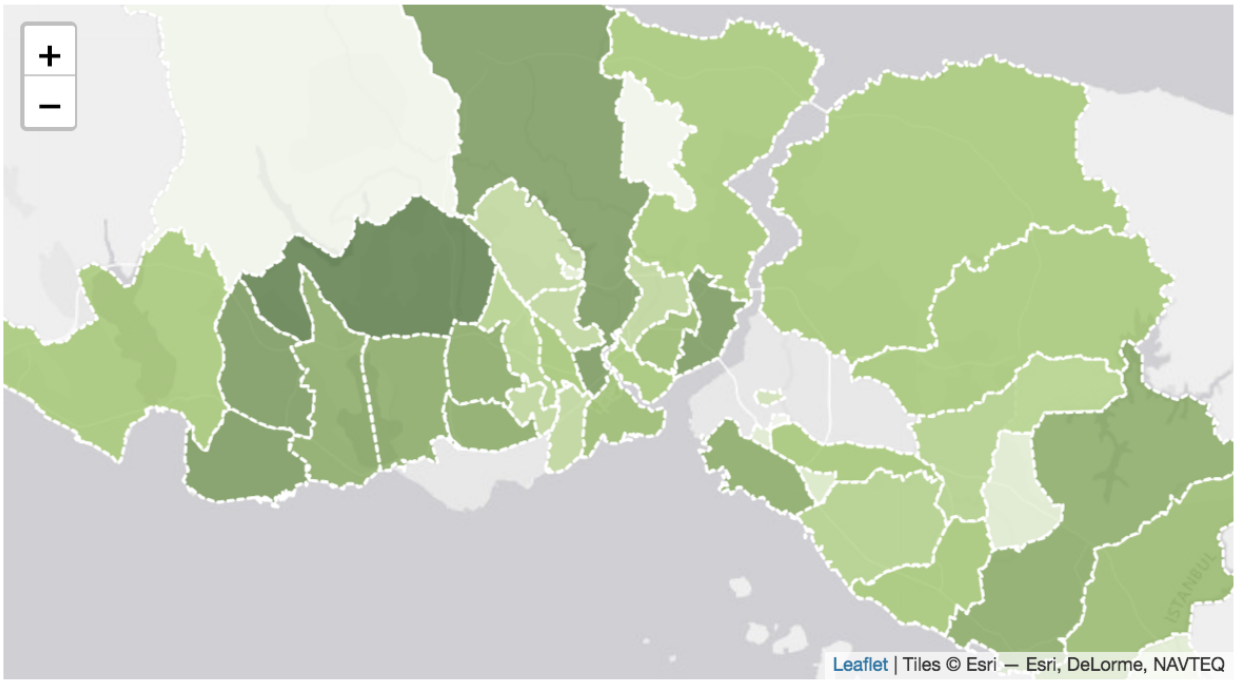



(a) Map visualizing the count of social media posts classified to dwelling activities in Amsterdam

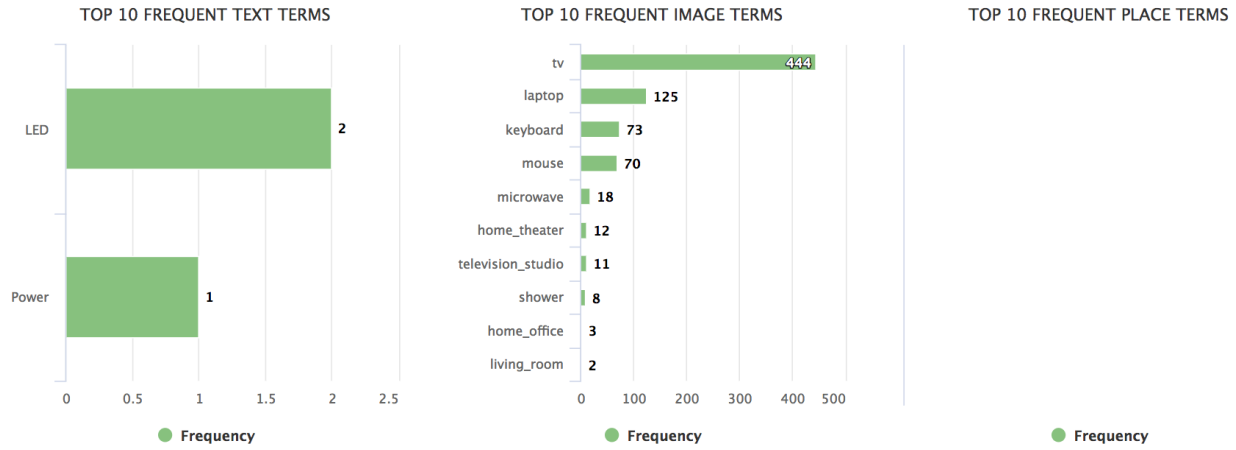



(b) Bar charts visualizing the count of social media posts classified to dwelling activities in Amsterdam

Figure 5.8: Overview of the count of social media posts classified to dwelling activities in Amsterdam

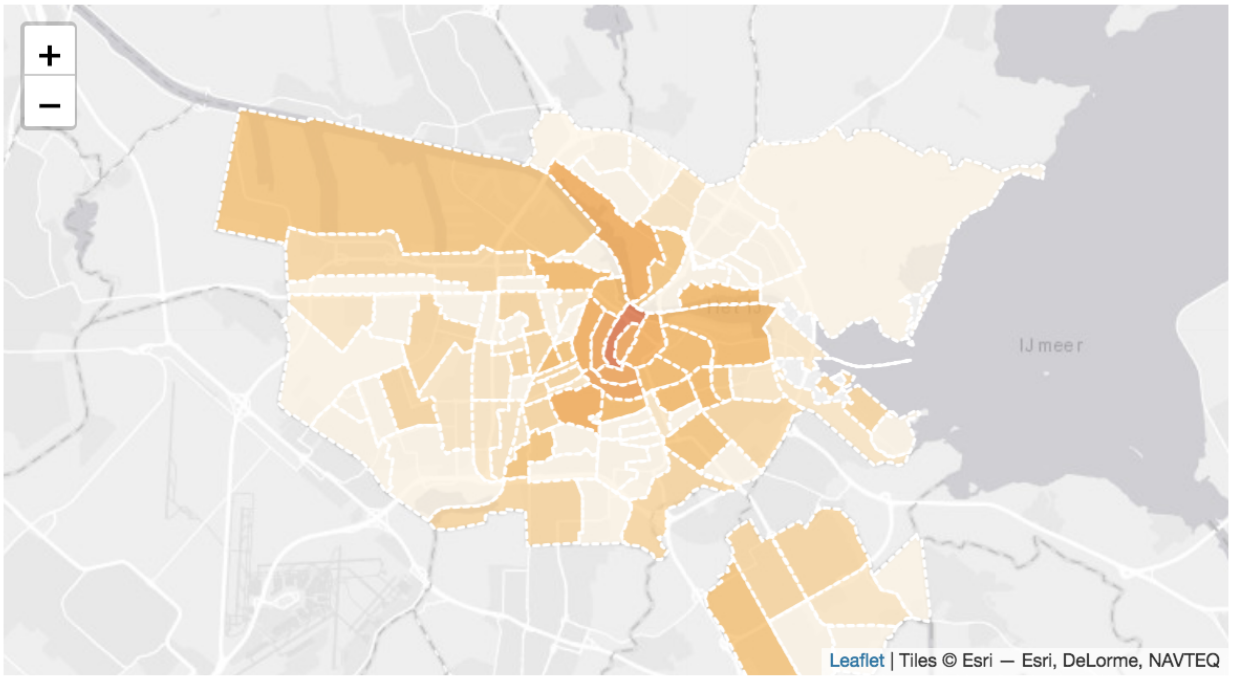



(a) Map visualizing the count of social media posts classified to dwelling activities in Istanbul

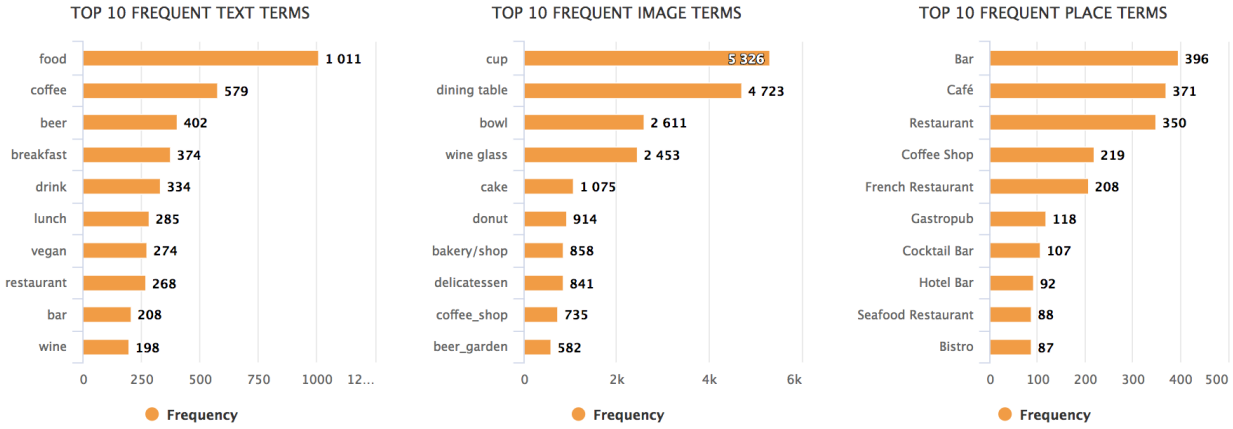



(b) Bar charts visualizing the count of social media posts classified to dwelling activities in Istanbul

Figure 5.9: Overview of the count of social media posts classified to dwelling activities in Istanbul

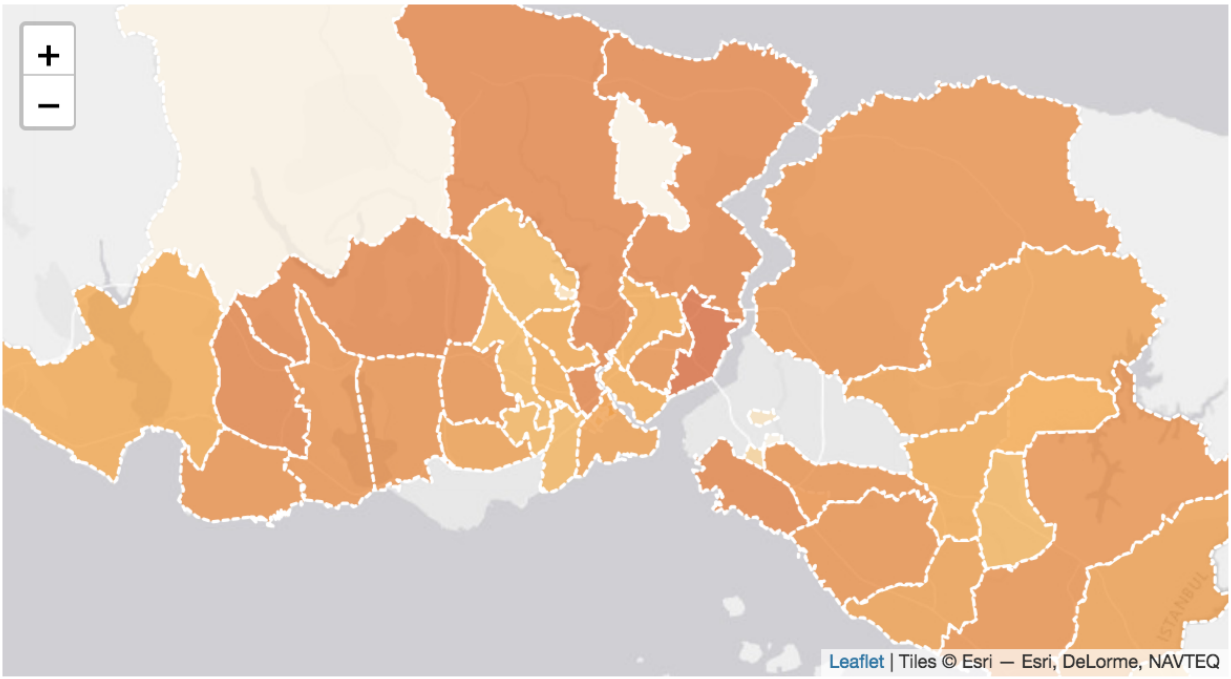



(a) Map visualizing the count of social media posts classified to food consumption activities in Amsterdam

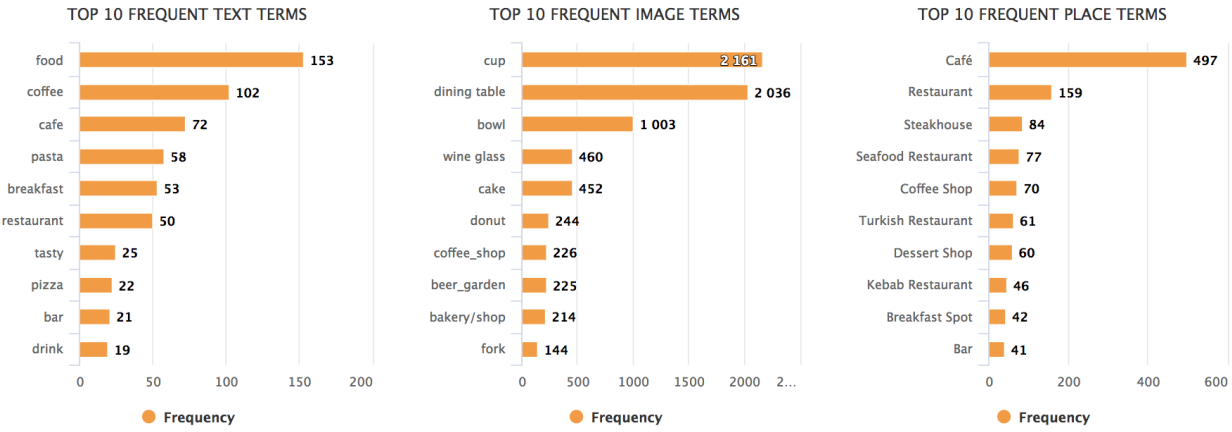



(b) Bar charts visualizing the count of social media posts classified to food consumption activities in Amsterdam

Figure 5.10: Overview of the count of social media posts classified to food consumption activities in Amsterdam

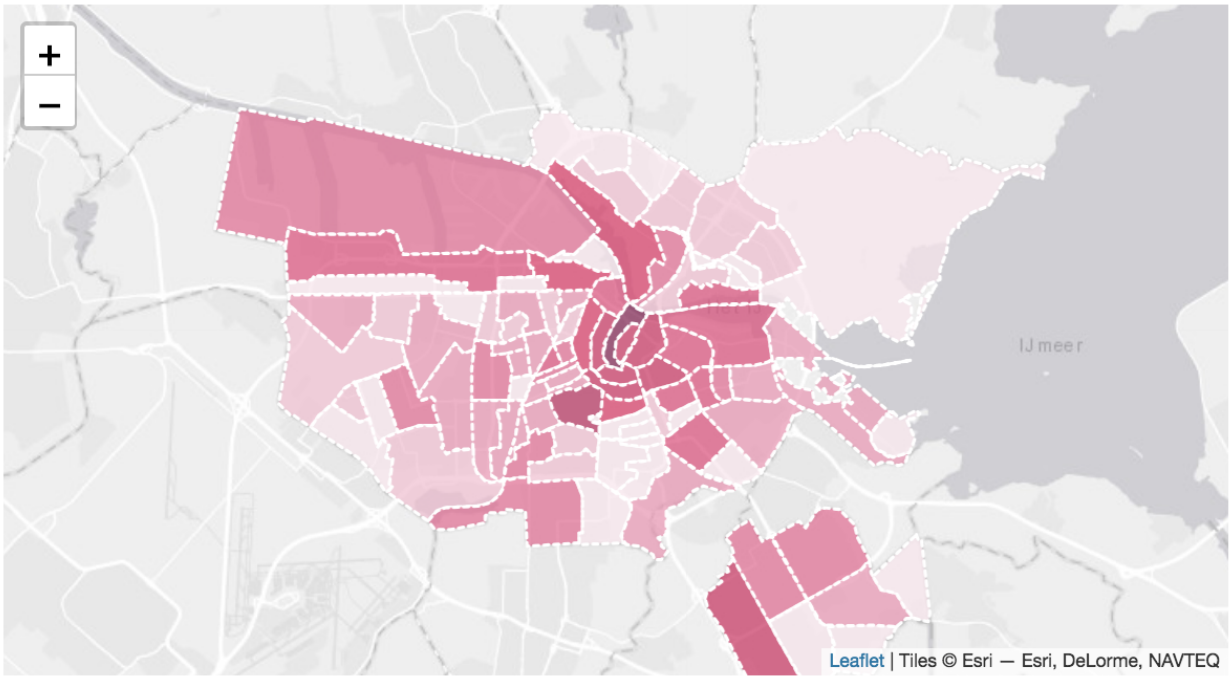



(a) Map visualizing the count of social media posts classified to food consumption activities in Istanbul

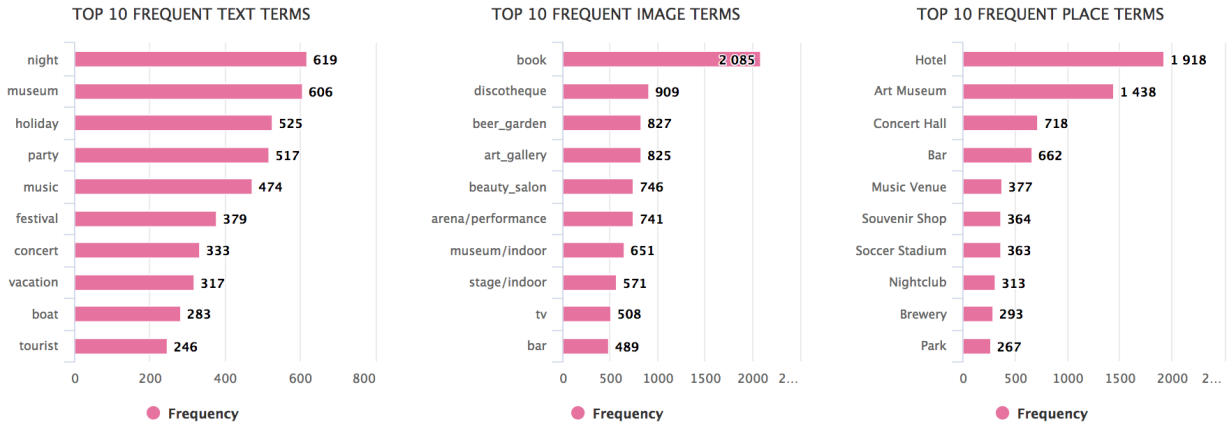



(b) Bar charts visualizing the count of social media posts classified to food consumption activities in Istanbul

Figure 5.11: Overview of the count of social media posts classified to food activities in Istanbul

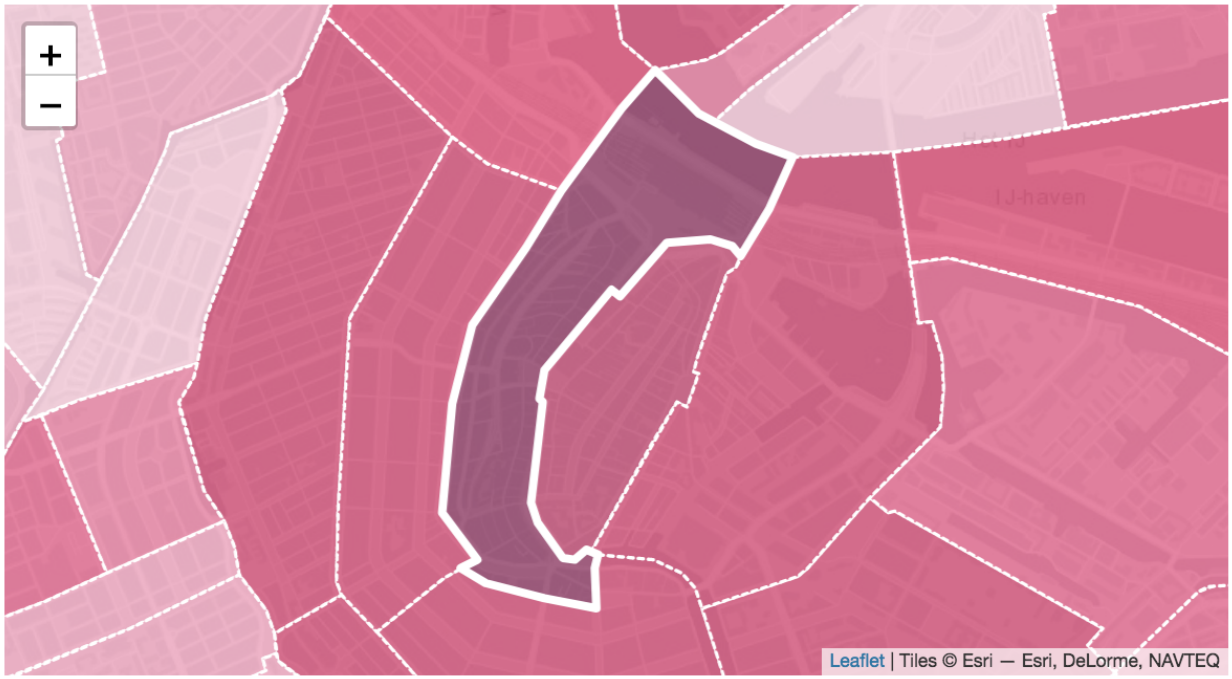



(a) Map visualizing the count of social media posts classified to leisure activities in Amsterdam

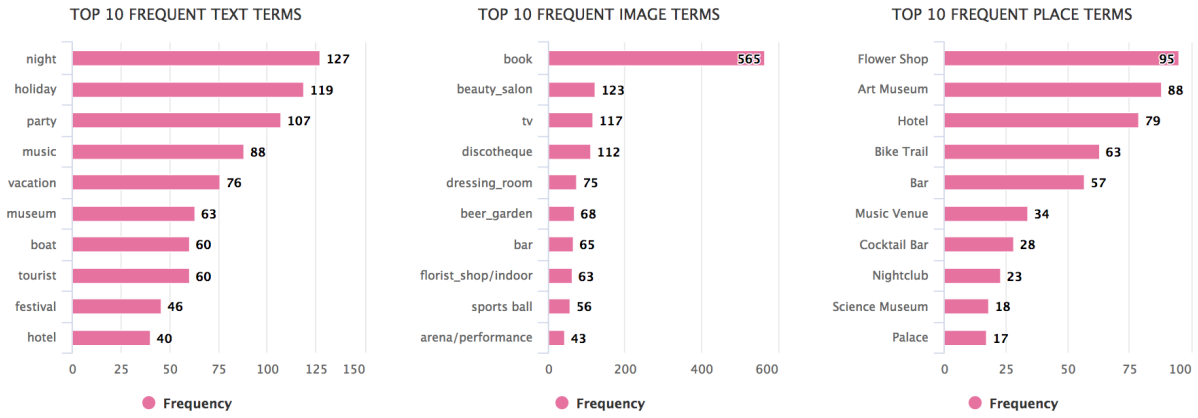



(b) Bar charts visualizing the count of social media posts classified to leisure activities in Amsterdam

Figure 5.12: Overview of the count of social media posts classified to leisure activities in Amsterdam

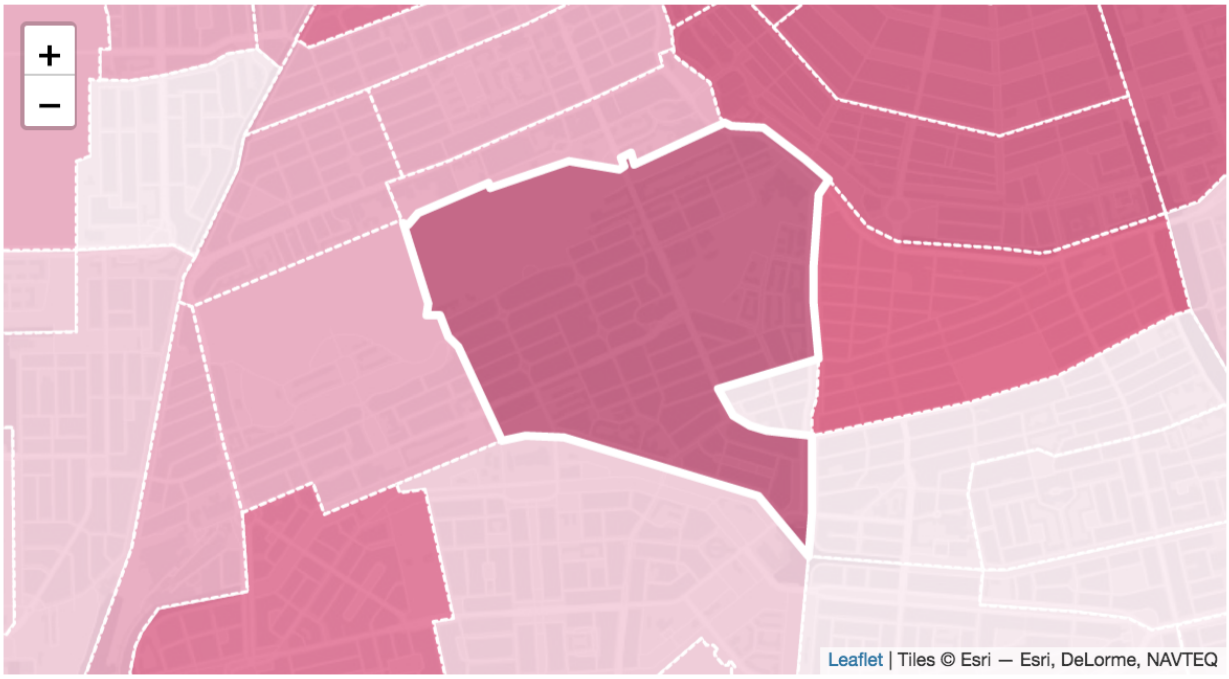



(a) Map visualizing the count of social media posts classified to leisure activities in Burgwallen-Nieuwe Zijde (Amsterdam)

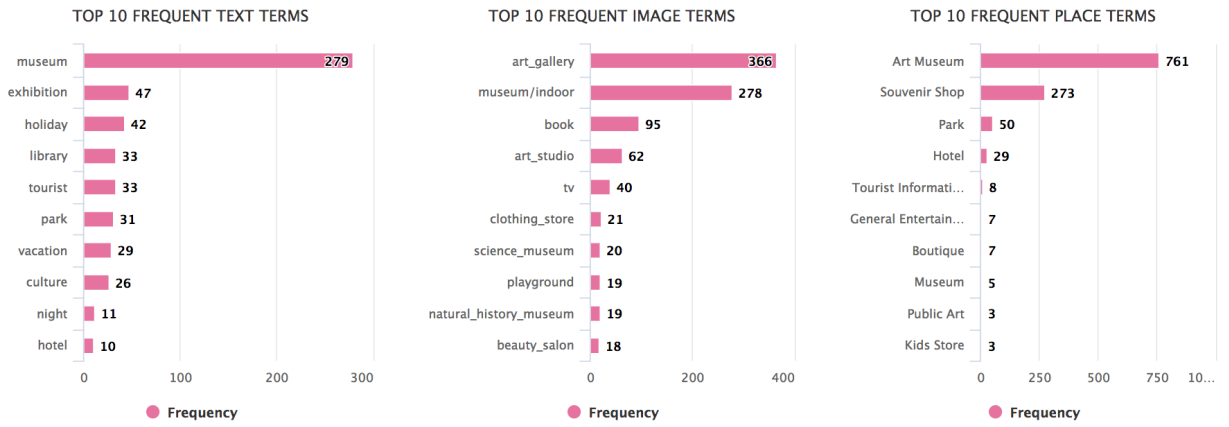



(b) Bar charts visualizing the count of social media posts classified to leisure activities in Burgwallen-Nieuwe Zijde (Amsterdam)

Figure 5.13: Overview of the count of social media posts classified to leisure activities in Burgwallen-Nieuwe Zijde (Amsterdam)

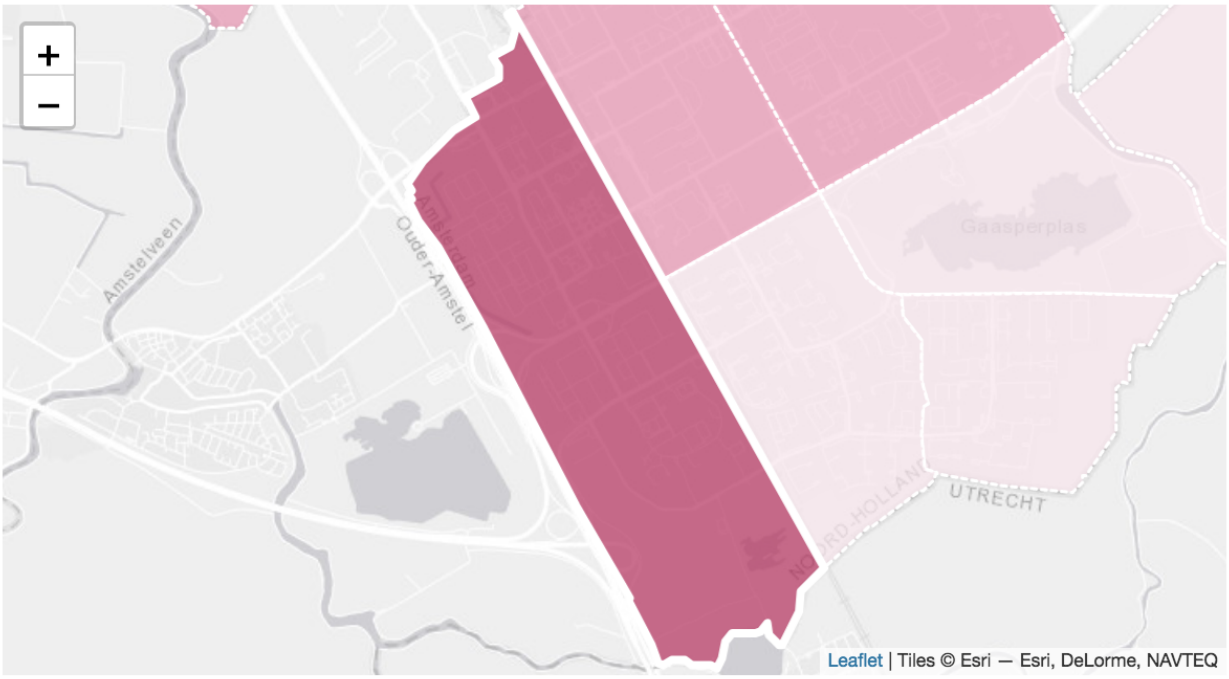



(a) Map visualizing the count of social media posts classified to leisure activities in Museumkwartier (Amsterdam)

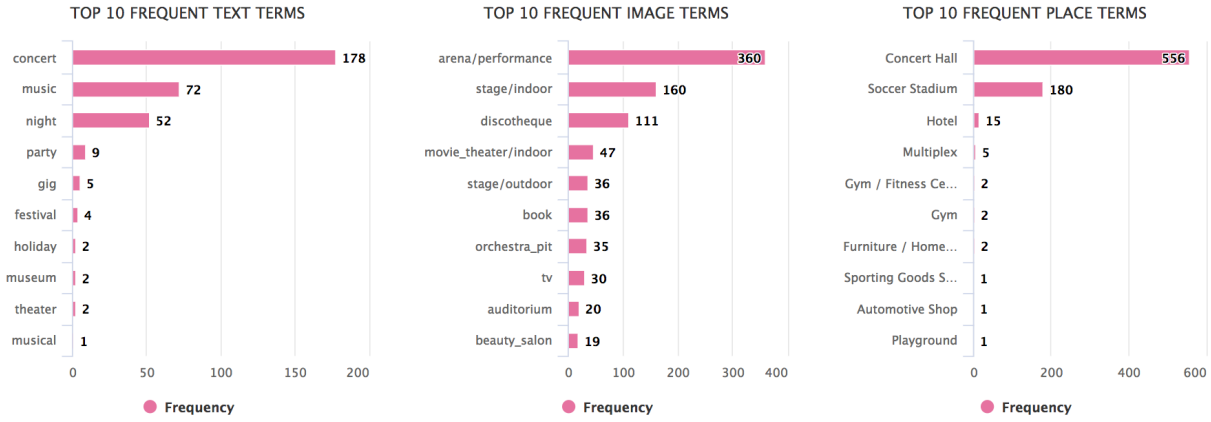



(b) Bar charts visualizing the count of social media posts classified to leisure activities in Museumkwartier (Amsterdam)

Figure 5.14: Overview of the count of social media posts classified to leisure activities in Museumkwartier (Amsterdam)

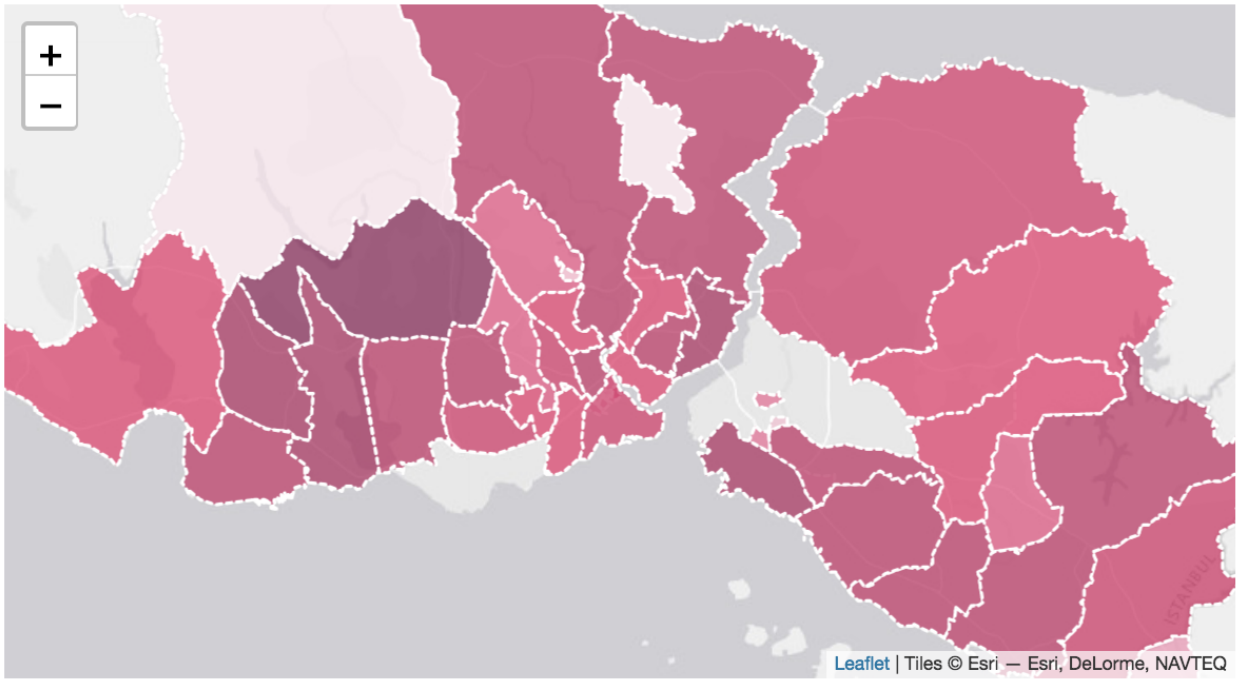



(a) Map visualizing the count of social media posts classified to leisure activities in Amstel III/Bullewijk (Amsterdam)

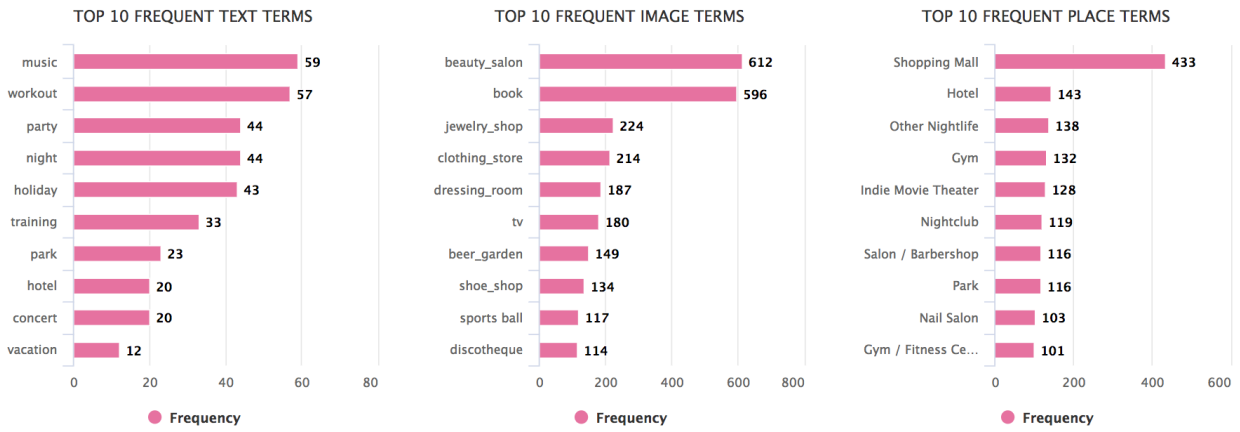



(b) Bar charts visualizing the count of social media posts classified to leisure activities in Amstel III/Bullewijk (Amsterdam)

Figure 5.15: Overview of the count of social media posts classified to leisure activities in Amstel III/Bullewijk (Amsterdam)

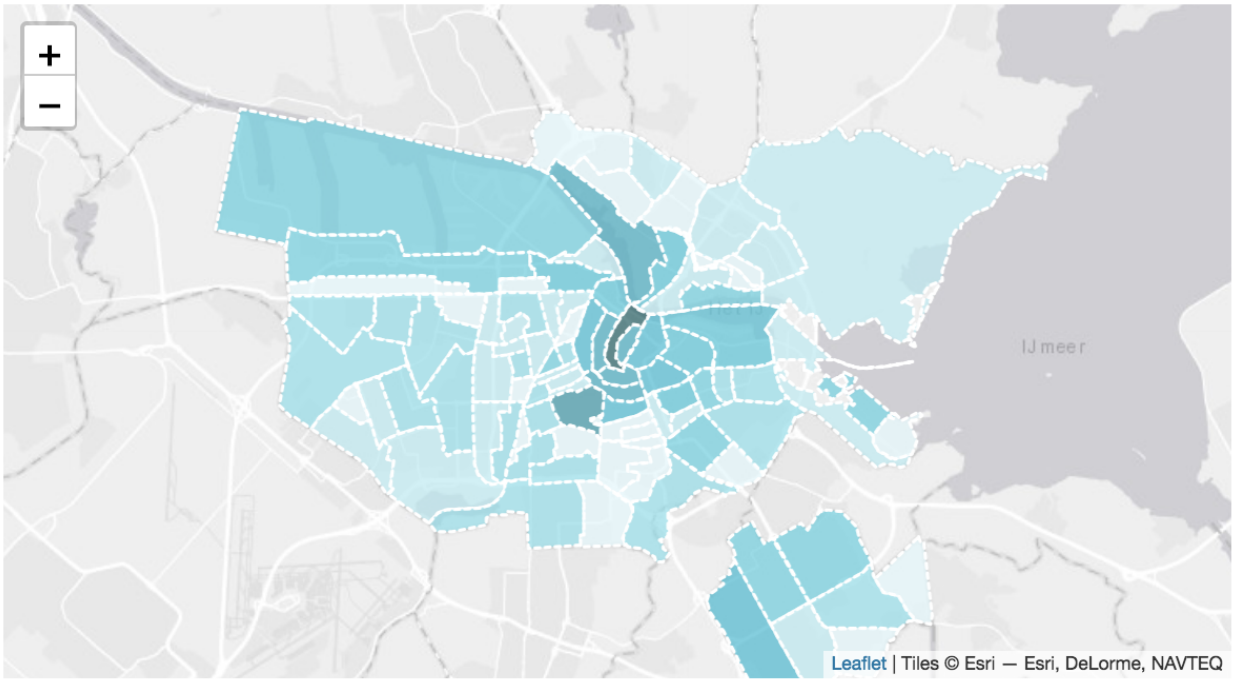



(a) Map visualizing the count of social media posts classified to leisure activities in Istanbul

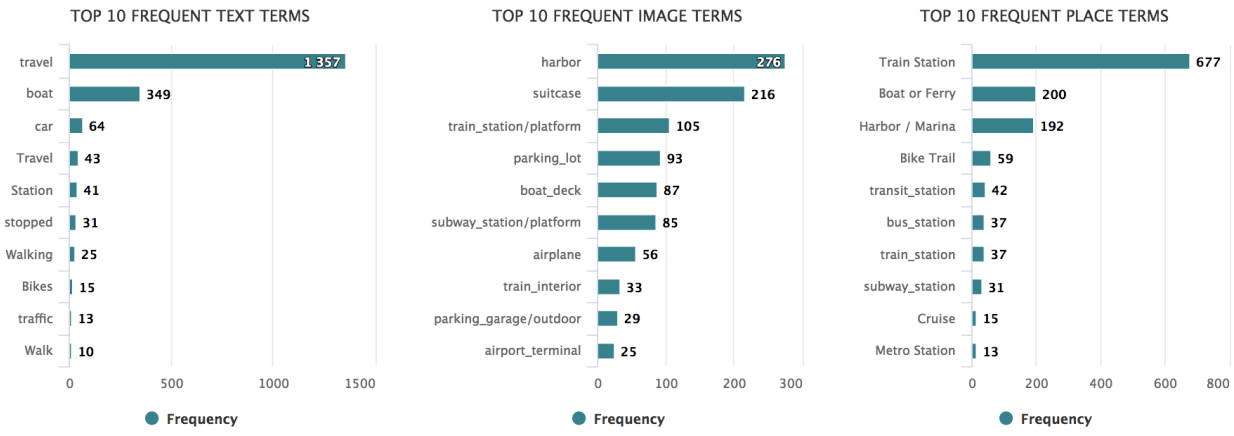



(b) Bar charts visualizing the count of social media posts classified to leisure activities in Istanbul

Figure 5.16: Overview of the count of social media posts classified to leisure activities in Istanbul

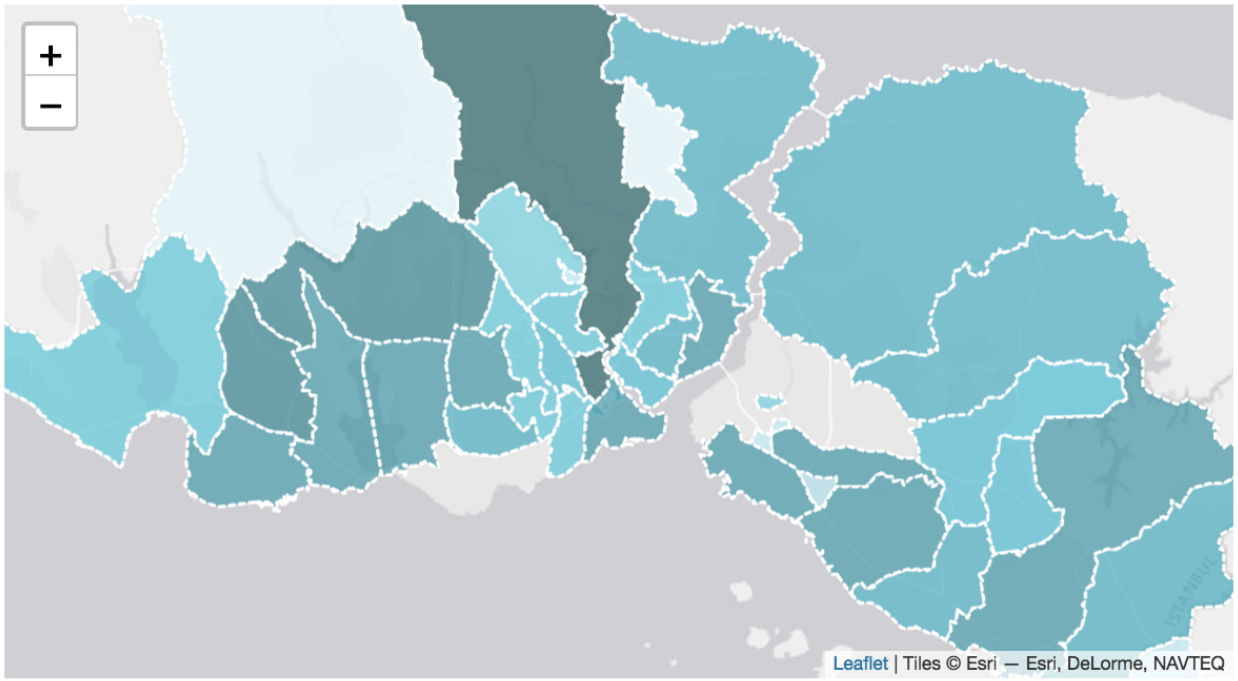



(a) Map visualizing the count of social media posts classified to mobility activities in Amsterdam

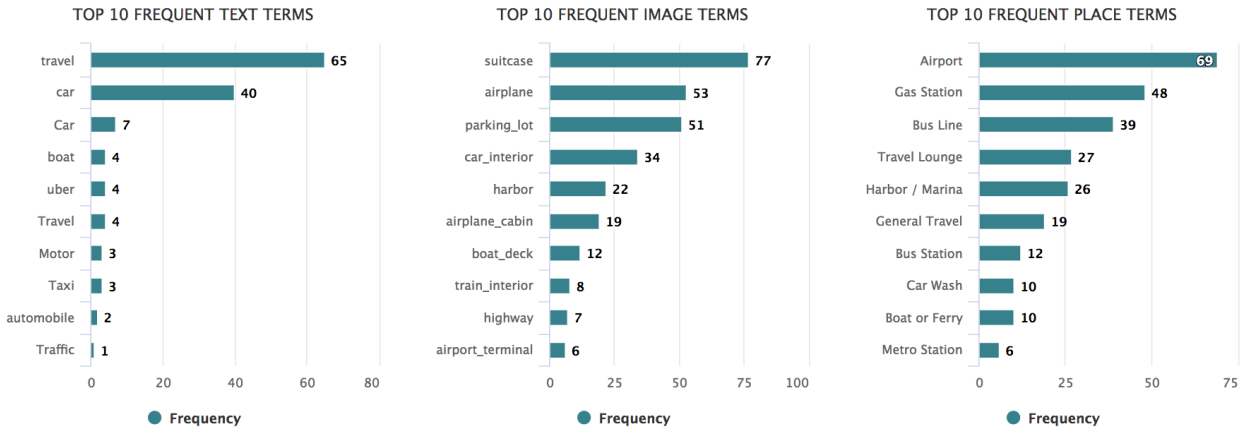



(b) Bar charts visualizing the count of social media posts classified to mobility activities in Amsterdam

Figure 5.17: Overview of the count of social media posts classified to mobility activities in Amsterdam

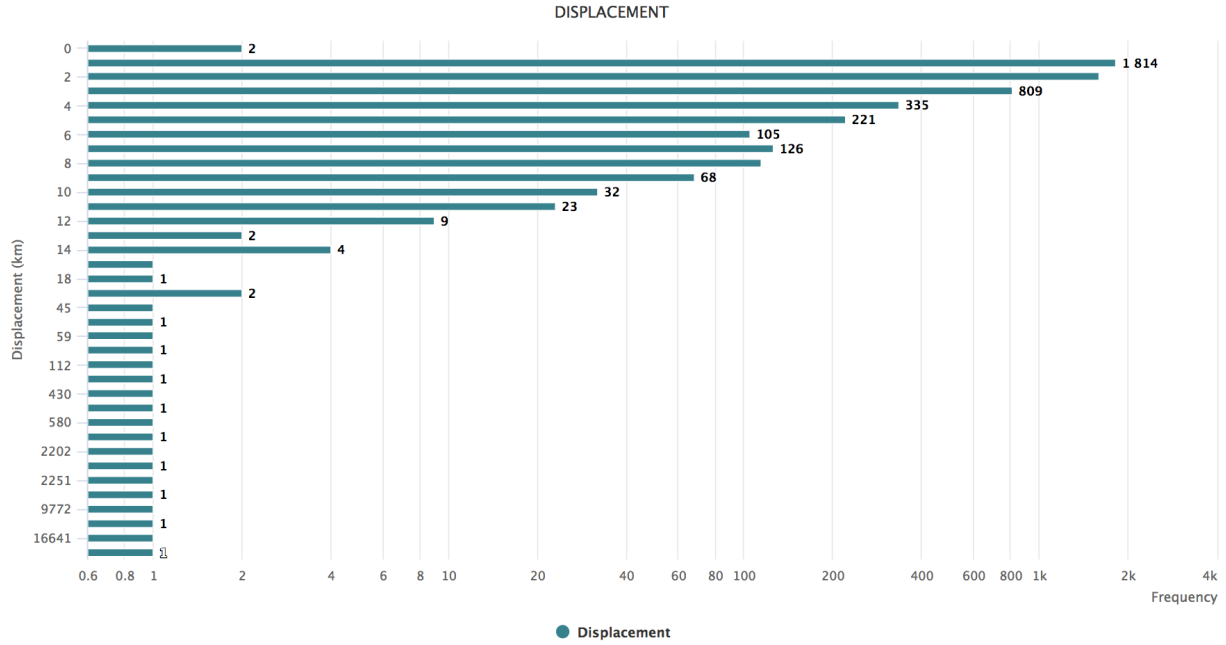



(a) Map visualizing the count of social media posts classified to mobility activities in Istanbul

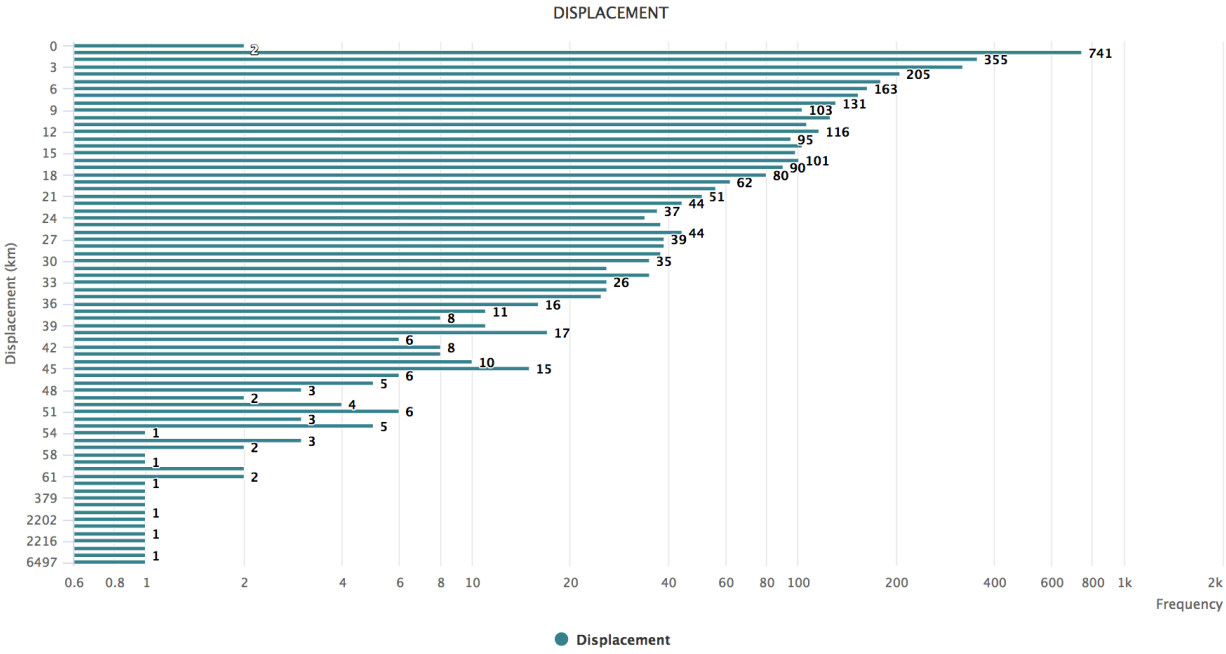



(b) Bar charts visualizing the count of social media posts classified to mobility activities in Istanbul

Figure 5.18: Overview of the count of social media posts classified to mobility activities in Istanbul




(a) Bar chart visualizing the frequency of displacements (distance between posts in kilometers) in Amsterdam




(b) Bar chart visualizing the frequency of displacements (distance between posts in kilometers) in Istanbul

Figure 5.19: Overview of the displacements (distance between posts in kilometers)

### 5.3.4 Threat of Validity

In this section we will reflect upon the results in terms of aspects that affect the validity of the framework. Multiple aspects regarding the design and implementation of the Social Smart Meter framework will be discussed, among which are semantic ambiguity in social media data, the purpose of social media, the data set size and covered time span, and the dictionaries.

**Semantic ambiguity in social media** In Chapter 1 we name th e challenges that are faced when using social media data. We attempted to address the semantic ambiguity of word meanings with the word sense disambiguation algorithm. However, despite the use of multiple natural language processing techniques, we still encountered some difficulties separating noise from relevant information. Furthermore, the discrepancy in time between a social media post and the physical world - e.g., when a user recalls a memory from a while ago - are not addressed by our Social Smart Meter framework yet.

**Using social media data sources** The social media data that was collected for our case study's target cities was not representative for the entire population. For instance, Instagram is more popular among younger people (age groups 15-19 and 20-39). Moreover, the social media APIs are only able to retrieve public data. Yet, a lot of social media users have private profiles resulting in private social media data that is not accessible. Besides that, numerous public social media accounts are held by (commercial) organizations and do not represent an individual. Furthermore, social media posts are created for other purposes - e.g., to share (leisure) experiences or to communicate with others - than providing insights into one's energy-consuming activities. Hence, the aggregated results per city or neighborhood should not be considered to be an entirely representative reflection of all energy-consuming activities performed in that particular area.

**Data set size** For the case study we collected social media data created in the time span of a couple of days. This resulted in approximately 275K social media posts, a large amount of data which was very time costly to process and classify. However, ideally, the data set would have covered a larger time span to prevent single events (such as the public transport strike in Amsterdam) to significantly affect and bias our classification results. Furthermore, the short time span did not allow for the collection of many posts by a particular social media user. This complicated the framework's home detection component; too few location points per user were collected which affected the clustering algorithm's performance. A larger data set (i.e., covering a larger time span) would presumably enhance the framework's home detection component to some extent.

**Limitations of the dictionaries** In this framework, the text dictionaries are considerably limited. These are created based on the classifier's output of [60]. However, synonyms and related concepts are not taken into account yet. Furthermore, the terms in the dictionaries are now limited to only a couple of languages (Dutch, English, and Turkish). Besides that, most methods in our text processing libraries -

such as the Lesk algorithm for word sense disambiguation - are only applicable to English terms. In future work, this language barrier should be handled, and the dictionaries should be (automatically) expanded. Also, the dictionaries might be place-biased - e.g., There are no wild monkeys in Amsterdam. Hence, the detection of a monkey implies a visit to the zoo, which is a leisure activity. In a city like Rio de Janeiro, this could be a regular street view. Hence, it might be an option to adjust the dictionaries per area.

**Evaluation of the framework** Since there is no existing data set yet that captures the energy-consuming activities as we aim for with our Social Smart Meter framework, it is difficult to evaluate the results. By means of crowdsourcing, the ground truth for a single social media post's categories of energy-consuming activities can be determined. However, this does not provide information about to what extent these classified social media posts reflect the energy-consuming activities performed in the physical world.

**Cost in terms of time and space** The pre-trained models used for our framework were rather time and space costly. The social media post's image had to be downloaded from the Web before it could be processed; moreover, the processed images were stored locally as well. Given that a large data set was processed, this was rather space costly. Furthermore, the image processing techniques were rather costly in terms of time as well. It took approximately 10 seconds to process the entire social media post. Given that - for Amsterdam - around 20000 social media post were collected per day, it takes longer than a day to process all the social media data. Thus, currently, the Social Smart Meter framework is too time costly to process the social media data real-time. By using other (less performing) pre-trained models and smaller-sized images, the processing time could be reduced to some extent.

### 5.3.5 Reflection on Results

In addition to the evaluation of the framework, which was briefly mentioned in the previous section, a reflection on the case study's results is provided below. Due to the multi-disciplinary nature of the domain of energy-consuming activities and the lack of domain experts, it is difficult to say to what extent the framework is adherent to reality. Since the results cannot be compared to ground truth, it is also hard to compare the framework's output to the actual numbers of energy-consuming activities performed in the physical world. For each category of energy-consuming activities, some hypotheses on the framework's adherence to reality are posed.

**Dwelling** Few social media content referring to dwelling activities were captured by the framework. This may be due to the fact that social media users do not consider their regular domestic activities interesting enough to share with other social media users. Thus, at this moment, the framework is not very adherent to reality for identifying dwelling activities in social media content. Other platforms - e.g., the

Steam[11] community for games or the Spotify[12] music stream provider - are more likely to be used for sharing data on dwelling activities, such as gaming or playing music, than social media. Furthermore, dwelling activities are most often performed outside working hours. This has not been taken into account yet by the framework but could be included as a rule in the rule-based approach.

**Food consumption** There are multiple types of food consumption; we presume that social media users generate less posts regarding regular dinners at home compared to dining activities that are also labeled as leisure activities. Hence, following up on the reflection on dwelling activities, the framework is not very adherent to reality (yet) for capturing regular dinners at home. Opposed to that, the framework is more adherent to reality when food consumption is also considered as leisure, since social media users are more likely to generate social media content on leisure activities.

**Leisure** Social media is most often used to capture and share leisure activities, since this is considered interesting to share with others. At a city's (touristic) hot spots, the majority of the social media content is generated by tourists. Yet, tourists form the majority of the people present at those locations, especially in the city center of Amsterdam. Since our results show that many social media posts that are generated in an area refer to a leisure activity that characterizes that particular area, our framework's results seem to be a rather good reflection of the activities that are actually performed in the physical world. Thus, the framework is adherent to reality for capturing leisure activities.

**Mobility** People do not often create very explicit social media content about their mobility activities. When they are traveling, they are more likely to read content generated by others or create content about (leisure) activities they performed before. However, when a social media user generates content on regular basis, his or her displacement can be derived from the geolocations included by the posts. Inferring the type of vehicle needs extra attention in an improved version of the framework. Once the framework's output is enriched with more information on the type of vehicle, more insights can be provided on the corresponding energy consumption of that particular trip.

**Overall adherence to reality** As mentioned before, it is difficult to assess the framework's adherence to reality. However, it might not be necessary to provide results that are fully adherent to reality. Some energy-consuming activities (mainly within the dwelling category) are yet captured by traditional data sources. Thus, by providing insights into the remaining energy-consuming activities, the Social Smart Meter could be a complementary source of information for providing insights into the total of energy-consuming activities. More thorough analyses in this complementary part of the domain of energy-consuming activities are left for future work.

[11] https://steamcommunity.com/
[12] https://www.spotify.com/nl/

# Chapter 6

# Conclusions and Future Work

In this work, we proposed the Social Smart Meter framework that identifies and describes energy-consuming activities by processing user-generated content (scoped to social media data). It is composed of different elements: (i) the Social Smart Meter ontology that provides a better understanding of the domain of energy-consuming activities, (ii) the data processing pipeline that enables the classification of social media posts (at the individual level) to the different categories of energy-consuming activities, and (iii) the Social Smart Meter Web application that allows viewers to request this information at group level - i.e., at city or neighborhood level.

## 6.1 Contributions

The contributions (depicted by “C") of this work are as follows:

**Overview of the state of the art (C1)** A literature review is conducted to explore the state of the art in the field of the description and recognition of energy-consuming activities; strengths and weaknesses of recent work are identified in order to determine which methods and tools to include in our Social Smart Meter framework.

**Social Smart Meter Ontology (C2)** The SSMO[1] represents the domain of energy-consuming activities. It is built upon numerous existing ontologies, such as the SUMO, Semanco, and EnergyUse ontologies.

**Analysis of data sources (C3A)** In order to identify the most promising data sources to use for our framework, a structured analysis of data sources - including multiple social media and enrichment data sources - was performed. The data sources have been compared based on multiple aspects, such as usage, API, and available data types.

**Data processing pipeline (C3B)** This model[2] automatically processes social media data for the description of energy-consuming activities at individual level. It processes the social media post’s different data types using multiple state-of-the-art

[1] https://www.github.com/redekok/social-smart-meter-ontology
[2] https://www.github.com/redekok/social-smart-meter

data processing techniques. Hereafter, it classifies the posts to the different categories of energy-consuming activities using a dictionary- and rule-based approach.

**Social Smart Meter Web application (C4)** The data processing pipeline's results were aggregated and analyzed using the implementation of the Social Smart Meter Web application[3]. It allows viewers to request information by time (per day) and place (city or neighborhood level). The implementation was evaluated through a case study for the cities of Amsterdam and Istanbul.

## 6.2 Discussion and Conclusions

Our main research question "How can we automatically process user-generated content to describe energy-consuming activities at individual and group level?" was broken up into four research sub-questions. In order to answer this main research question, we will answer and discuss all sub-questions.

**RQ1: How are energy-consuming activities studied by the state of the art?**

Energy-consuming activities have been studied in plenty of previous studies; these often rely on traditional data sources (energy surveys, smart sensor data, etc.) though. Few studies have incorporated social media data yet. Furthermore, in all of our fields of interest - i.e., the categories of energy-consuming activities - multiple studies have included social media data sources, though with a different purpose than providing energy consumption-related information. Nevertheless, the state-of-the-art techniques from these studies could be re-used in this work.

**RQ2: What are the main characteristics of an individual's energy-consuming activities?**

Conceptual data models (which eventually formed the basis for the SSMO) were developed to provide insights into the domain of energy-consuming activities and to identify its main characteristics. Individuals perform energy-consuming activities at some place at some time. The activities can be of type (i.e., category) dwelling, food consumption, leisure and/or mobility. Each category of energy-consuming activities has its own relevant related concepts - e.g., a dwelling activity is performed using an appliance.

**RQ3: How can we extract the characteristics of energy-consuming activities from social media data?**

A framework was developed to automatically process social media data. It collects a social media post's interesting and available data: text, image, place, user, and time (meta)data. Then, multiple enrichment steps are applied to the data. Once the data is enriched, a dictionary- and rule-based classification model is used to classify whether the social media post refers to either of the categories of energy-consuming activities. Also, classification confidence scores are assigned to the classification of each category.

[3] https://www.github.com/redekok/social-smart-meter-webapp

**RQ4: To what extent can social media serve as a complementary data source to understand energy-consuming activities?**

The framework's classification model provides insight into the energy-consuming activities at individual level - by classifying an individual's social media posts. In addition, the Web application's visualizations and charts provide insights into the energy-consuming activities at group (i.e., city and neighborhood) level. Based on a preliminary user-based evaluation, the informativeness of the different (main) data types (text, images, and places) was evaluated and the corresponding (averaged and normalized) ratings were adopted as data type weights for our classification model. Text was found to be most informative for the classification of leisure and mobility activities, images most informative for dwelling and food consumption activities. Overall, text and images were found more informative than places.

The classifier's performance was evaluated by means of the user-based evaluation as well. Several evaluation metrics were calculated for multiple values of our classification threshold. By tuning the threshold, either a higher precision or recall score can be obtained. Furthermore, the accuracy varies from 0.69 to 0.78, which is considered an acceptable performance. For our case study we selected a threshold of 0.35 (with a corresponding accuracy of 0.78). The accuracy scores for dwelling, food consumption and mobility are rather high for this threshold (respectively, 0.87, 0.84, and 0.72) whereas leisure has a relatively low accuracy score (0.57) due to many false negatives. Overall, the framework's precision was moderately high for this threshold (overall, for the total of all categories, 0.73), though low recall scores (overall, 0.47) were obtained. This resulted in an overall F1-score of 0.54.

Furthermore, most social media posts are classified to leisure and food based on the dictionary classification approach. This is in line with our hypothesis that social media posts are mainly created by users to share fun, leisure experiences. Very few social media posts are classified to dwelling, which may imply that users do not create posts at home very often. Moreover, it is very hard to detect if a social media post is created at home. The place data type rarely has any indication for a user's home, particularly when the user's home address is not known. However, previous works that try to capture dwelling activities performed by the user use traditional data sources - e.g., energy-related surveys, smart meters, smart plugs, etc.

The other categories of energy-consuming activities (food consumption, leisure, and mobility), which are harder to capture using traditional data sources, are captured to some extent by using social media. Food consumption activities are recognized very well (and are barely ambiguous); images and places are very informative for this category of energy-consuming activities. Leisure activities are hard to recognize from image objects; scene recognition and text are moderately informative though. In addition, places are even more informative. Mobility activities do not explicitly occur in social media posts very often. When they occur, places are most informative opposed to text and images which often do not have explicit evidence for a mobility activity

performed by the creator of the post. Implicit mobility activities are more often identified by applying our rule-based approach, which takes the distance between posts into account. Hence, our Social Smart Meter framework works best for identifying food consumption and leisure activities (despite a low recall score) in social media content. It could be improved in the field of identifying dwelling and mobility activities, especially derived by the dictionary-based approach. Mobility activities are yet identified moderately well by the rule-based approach - i.e., by determining the distance between posts.

A case study was performed for Amsterdam and Istanbul to evaluate the framework's results. It showed that the ratio between the different categories of energy-consuming activities was nearly similar for both cities. Besides that, the classified social media posts were more evenly distributed over the different neighborhoods in Istanbul compared to Amsterdam.

In this preliminary quantitative analysis the content of the classified social media posts mainly seemed to reflect the neighborhood in terms of performed leisure activities. For instance, in the Museumkwartier neighborhood in Amsterdam, many social media content referring to musea and art were identified and classified to leisure activities.

Often, semantic ambiguity rises in social media data, which brings along challenges to be faced. Not all challenges are addressed by our framework yet, such as distinguishing relevant information from noise, and handling discrepancies in time between the physical world and the creation of the social media post. Furthermore, social media data is not representative for the entire population of a city. Along with that, many social media accounts are private (and thereby not accessible) or held by (commercial) organizations. Moreover, social media posts are created for other purposes than providing information on a user's energy-consuming activities. Thus, the framework's findings do not entirely represent all energy-consuming activities performed in a particular area and is thereby not completely adherent to reality yet.

Currently, no data set exists that can be used to find the ground truth energy-consuming activities in the physical world. This complicates the evaluation of our framework since there is no existing ground truth to compare the output to. In Section 6.3 (*Future Work*) we describe an attempt to create a data set for such an evaluation.

In this work, a preliminary study is performed to examine how to use user-generated content to describe energy-consuming activities at individual and group level. Due to the multi-disciplinary and complex nature of this domain, it is difficult to assess the framework's adherence to reality. However, a full adherence to reality might not be necessary; traditional data sources are yet able to capture part of all energy-consuming activities, especially the ones that are hardly captured by our framework. Hence, the Social Smart Meter framework provides complementary insights on the total of energy-consuming activities. Since our preliminary results are promising, more thorough analyses are left for future work.

## 6.3 Future work

Given the reflection on our framework in the previous section, a lot of future work possibilities arise. These are described in more detail below.

**Ontology population and evaluation** Another approach for the evaluation of our ontology is the data-driven evaluation approach [17]. A SPARQL endpoint should be provided for this ontology, which allows for automatic population of the ontology. This would also enable our Social Smart Meter framework to generate an instantiated ontology as output for each social media post. This would provide more insights into an energy-consuming activity than just the corresponding category. Also, by evaluating the instantiated ontologies, we could evaluate the SSMO by means of the application-based ontology evaluation approach.

**Cross-platform** Currently, only the Instagram and Twitter social media platforms are integrated in our framework. In a next version of the framework, more social media platforms could be integrated for a more complete overview of the social media activity in a particular area. Moreover, by matching the users across these social media platforms, more (relevant) information about the users could be collected.

**Data enrichment** The enrichment module of our Social Media Meter framework could be extended with multiple other enrichment steps. For instance, for text tokens we could determine whether they refer to a place (instead of only taking the place check-in into account), and we could look into the correlation between hashtags and images. The authors of [72] attempted to discover image content using user-generated hashtags in Instagram posts. The hashtags were used to further discover categories which may not have been present in pre-labeled image categories. Thus, we might be able to get more relevant information from images when taking the hashtags into account while processing the image.

Besides that, the set of rules for our rule-based approach could be expanded to gather more implicit information on energy-consuming activities.

**Infer mode of transport** The origin and destination of a user's mobility activity can be derived from his or her history of social media posts. Currently, the direct distance between two posts - the coordinates of the former post represent the origin whereas the coordinates of the latter one represent the destination - is determined in the framework. However, by using the Google Directions API[4] for example, the directions (including the estimated travel time) for several modes of transport can be retrieved. The time between the creation of the two posts can be compared to the estimated travel times for each mode of transport. When an estimated travel time is larger than the time between the two posts, the corresponding mode of transport can be excluded. In this way, the output of the mobility activity can be enriched with more information on the used mode of transport, which also provides more information on the extent of consumed energy during this activity.

[4] https://developers.google.com/maps/documentation/directions/start

**Finding ground truth and optimize data type weights** Once the ground truth of categories of energy-consuming activities for social media posts is determined through thorough crowdsourcing, the data type weights could be optimized; given the ground truth, the error could be minimized which results in the optimal data weights.

**Dictionary expansion** As briefly mentioned in our discussion, the dictionaries used are still rather limited and should be expanded. For instance, WordNet[5] could be used to find synonyms for dictionary terms that could be included in the concerned dictionary as well. Furthermore, ConceptNet[6] provides concepts that are related to a particular term - e.g., an *oven* is used for *cooking*. In future work, the most relevant relationships for concepts could be identified, after which the dictionaries could be expanded with those related concepts as well.

**Framework evaluation** Yet, no relevant data set exists yet to evaluate the results of our Social Smart Meter framework. In future work, it might be a possibility to ask a (sufficient large and varied) set of participants to keep track (through a log) of all the energy-consuming activities they perform during the day (for a certain period of time that has to be determined). In addition, these participants are asked to provide access to their social media accounts. By classifying the participant's social media posts and comparing the results to the participant's log, we could gather more insights into to what extent their social media activity reflects the physical world in terms of the performance of energy-consuming activities.

The CODALoop project[7] aims to "contribute to a paradigm shift within the field of energy consumption, which focuses not only on reducing energy consumption but also on reducing energy needs" and provides, among other contributions, "a tested prototype of an interactive web-based platform for sharing data about individual and community energy consumption choices". Since this work is performed in context with the CODALoop Project, a collaboration could yield the participants needed for the framework evaluation above.

[5] https://wordnet.princeton.edu/
[6] https://conceptnet.io/
[7] https://jpi-urbaneurope.eu/project/codaloop/

# Bibliography


[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[2] Sofiane Abbar, Yelena Mejova, and Ingmar Weber. You tweet what you eat: Studying food consumption through twitter. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3197–3206. ACM, 2015.

[3] Wokje Abrahamse, Linda Steg, Charles Vlek, and Talib Rothengatter. The effect of tailored information, goal setting, and tailored feedback on household energy use, energy-related behaviors, and behavioral antecedents. *Journal of environmental psychology*, 27(4):265–276, 2007.

[4] Morteza Akbari Fard, Hamed Hadadi, and Alireza Tavakoli Targhi. Fruits and vegetables calorie counter using convolutional neural networks. In *Proceedings of the 6th International Conference on Digital Health Conference*, pages 121–122. ACM, 2016.

[5] Baris Aksanli, Alper Sinan Akyurek, and Tajana Simunic Rosing. User behavior modeling for estimating residential energy consumption. In *Smart City 360Â°*, pages 348–361. Springer, 2016.

[6] G Antoniou, F Van Harmelen, and A Semantic Web Primer. *A semantic web primer*. MIT Press.

[7] Julia Backhaus, Sylvia Breukers, M Paukovic, R Mourik, and O Mont. Sustainable lifestyles. today's facts and tomorrow's trends. d1. 1 sustainable lifestyles baseline report. 2012.

[8] Kenneth Baclawski, Mieczyslaw K Kokar, Paul A Kogut, Lewis Hart, Jeffrey Smith, William S Holmes, Jerzy Letkowski, and Michael L Aronson. Extending uml to support ontology engineering for the semantic web. In *International Conference on the Unified Modeling Language*, pages 342–360. Springer, 2001.

[9] Satanjeev Banerjee and Ted Pedersen. An adapted lesk algorithm for word sense disambiguation using wordnet. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 136–145. Springer, 2002.

[10] Marco Aurelio Beber, Carlos Andres Ferrero, Renato Fileto, and Vania Bogorny. Individual and group activity recognition in moving object trajectories. *Journal of Information and Data Management*, 8(1):50, 2017.

[11] Brent Bleys, Bart Defloor, Luc Van Ootegem, and Elsy Verhofstadt. The environmental impact of individual behavior: Self-assessment versus the ecological footprint. *Environment and Behavior*, page 0013916517693046, 2017.

[12] Todd Bodnar, Matthew L Dering, Conrad Tucker, and Kenneth M Hopkinson. Using large-scale social media networks as a scalable sensing system for modeling real-time energy utilization patterns. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2017.

[13] Dario Bonino and Fulvio Corno. Dogont-ontology modeling for intelligent domotic environments. In *International Semantic Web Conference*, pages 790–803. Springer, 2008.

[14] Dario Bonino, Fulvio Corno, and Luigi De Russis. Poweront: An ontology-based approach for power consumption estimation in smart homes. In *Internet of Things. User-Centric IoT*, pages 3–8. Springer, 2015.

[15] Janez Brank, Marko Grobelnik, and Dunja Mladenić. A survey of ontology evaluation techniques. 2005.

[16] John G Breslin, Andreas Harth, Uldis Bojars, and Stefan Decker. Towards semantically-interlinked online communities. In *European Semantic Web Conference*, pages 500–514. Springer, 2005.

[17] Christopher Brewster, Harith Alani, Srinandan Dasmahapatra, and Yorick Wilks. Data driven ontology evaluation. 2004.

[18] Grégoire Burel, Lara SG Piccolo, and Harith Alani. Energyuse-a collective semantic platform for monitoring and discussing energy consumption. In *International Semantic Web Conference*, pages 257–272. Springer, 2016.

[19] Paul Burger, Valéry Bezençon, Basil Bornemann, Tobias Brosch, Vicente Carabias-Hütter, Mehdi Farsi, Stefanie Lena Hille, Corinne Moser, Céline Ramseier, Robin Samuel, et al. advances in understanding energy consumption behavior and the governance of its change–outline of an integrated framework. *Frontiers in Energy Research*, 3:29, 2015.

[20] Balakrishnan Chandrasekaran, John R Josephson, and V Richard Benjamins. What are ontologies, and why do we need them? *IEEE Intelligent Systems and their applications*, 14(1):20–26, 1999.

[21] Jui-Sheng Chou and Ngoc-Tri Ngo. Time series analytics using sliding window metaheuristic optimization-based machine learning system for identifying building energy consumption patterns. *Applied Energy*, 177:751–770, 2016.

[22] Wei-Ta Chu and Jia-Hsing Lin. Food image description based on deep-based joint food category, ingredient, and cooking method recognition. In *Multimedia & Expo Workshops (ICMEW), 2017 IEEE International Conference on*, pages 109–114. IEEE, 2017.

[23] Gianluigi Ciocca, Paolo Napoletano, and Raimondo Schettini. Food recognition: a new dataset, experiments, and results. *IEEE journal of biomedical and health informatics*, 21(3):588–598, 2017.

[24] Stephen Cranefield, Stefan Haustein, and Martin Purvis. Uml-based ontology modelling for software agents. In *Proceedings of the Workshop on Ontologies in Agent Systems, 5th International Conference on Autonomous Agents*, pages 21–28. Montreal, Canada:[sn], 2001.

[25] Lidija Čuček, Jiří Jaromír Klemeš, and Zdravko Kravanja. A review of footprint analysis tools for monitoring impacts on sustainability. *Journal of Cleaner Production*, 34:9–20, 2012.

[26] Vincius Cezar Monteiro de Lira, Rossano Venturini, Chiara Renso, and Valeria Cesário Times. Mining human mobility data and social media for smart services. 2016.

[27] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[28] Mariano Fernández-López, Asunción Gómez-Pérez, and Natalia Juristo. Methontology: from ontological art towards ontological engineering. 1997.

[29] Corinna Fischer. Feedback on household electricity consumption: a tool for saving energy? *Energy efficiency*, 1(1):79–104, 2008.

[30] Enrico Franconi. Description logics for conceptual design, information access, and ontology integration: Research trends. *Networks*, 2:25–32, 2003.

[31] Piero Fraternali, Sergio Herrera, Jasminko Novak, Mark Melenhorst, Dimitrios Tzovaras, Stelios Krinidis, Andrea Emilio Rizzoli, Cristina Rottondi, and Francesca Cellina. encompassâan integrative approach to behavioural change for energy saving. In *Global Internet of Things Summit (GIoTS), 2017*, pages 1–6. IEEE, 2017.

[32] Daniel Fried, Mihai Surdeanu, Stephen Kobourov, Melanie Hingle, and Dane Bell. Analyzing the language of food on social media. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 778–783. IEEE, 2014.

[33] Jon Froehlich, Eric Larson, Sidhant Gupta, Gabe Cohn, Matthew Reynolds, and Shwetak Patel. Disaggregated end-use energy sensing for the smart grid. *IEEE Pervasive Computing*, 10(1):28–39, 2011.

[34] Yong Geng, Liming Zhang, Xudong Chen, Bing Xue, Tsuyoshi Fujita, and Huijuan Dong. Urban ecological footprint analysis: a comparative study between shenyang in china and kawasaki in japan. *Journal of cleaner production*, 75:130–142, 2014.

[35] Jennifer Golbeck and Matthew Rothstein. Linking social networks on the web with foaf: A semantic web case study. In *AAAI*, volume 8, pages 1138–1143, 2008.

[36] Jeroen Guinée, Reinout Heijungs, Arjan De Koning, Lauran Van, Theo Geerken, Mirja Van Holderbeke, Bart Jansen Vito, Peter Eder, and Luis Delgado. Environmental impact of products (eipro) analysis of the life cycle environmental impacts related to the final consumption of the eu25. 2006.

[37] Nitin Hardeniya, Jacob Perkins, Deepti Chopra, Nisheeth Joshi, and Iti Mathur. *Natural Language Processing: Python and NLTK*. Packt Publishing Ltd, 2016.

[38] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.

[39] Panikos Heracleous, Pongtep Angkititraku, and Kazuya Takeda. Stochastic modeling and disaggregation of energy-consumption behavior. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 8277–8281. IEEE, 2014.

[40] Luis Herranz, Shuqiang Jiang, and Ruihan Xu. Modeling restaurant context for food recognition. *IEEE Transactions on Multimedia*, 19(2):430–440, 2017.

[41] Christina Hughes, Emilio Zagheni, Guy J Abel, Alessandro Sorichetta, Arkadius Wi'sniowski, Ingmar Weber, and Andrew J Tatem. Inferring migrations: Traditional methods and new approaches based on mobile phone, social media, and other big data: Feasibility study on inferring (labour) mobility and migration in the european union from big data and social media data. 2016.

[42] Mustafa Jarrar, Jan Demey, and Robert Meersman. On using conceptual data modeling for ontology engineering. In *Journal on data semantics i*, pages 185–207. Springer, 2003.

[43] Shan Jiang, Ana Alves, Filipe Rodrigues, Joseph Ferreira Jr, and Francisco C Pereira. Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Computers, Environment and Urban Systems*, 53:36–46, 2015.

[44] Andreas Kamilaris, Andreas Pitsillides, and Christos Fidas. 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, 2016.

[45] Yoshiyuki Kawano and Keiji Yanai. Food image recognition with deep convolutional features. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 589–593. ACM, 2014.

[46] Pavlos Kefalas, Panagiotis Symeonidis, and Yannis Manolopoulos. A graph-based taxonomy of recommendation algorithms and systems in lbsns. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):604–622, 2016.

[47] Nasrin Khansari, Ali Mostashari, and Mo Mansouri. Conceptual modeling of the impact of smart cities on household energy consumption. *Procedia Computer Science*, 28:81–86, 2014.

[48] Glenn E Krasner, Stephen T Pope, et al. A description of the model-view-controller user interface paradigm in the smalltalk-80 system. *Journal of object oriented programming*, 1(3):26–49, 1988.

[49] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *null*, pages 2169–2178. IEEE, 2006.

[50] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[51] Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM, 1986.

[52] Liangda Li and Hongyuan Zha. Energy usage behavior modeling in energy disaggregation via marked hawkes process. In *AAAI*, pages 672–678, 2015.

[53] Defu Lian and Xing Xie. Collaborative activity recognition via check-in history. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, pages 45–48. ACM, 2011.

[54] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[55] Tobias Linnenberg, Andreas W Mueller, Lars Christiansen, Christian Seitz, and Alexander Fay. Ontoenergy-a lightweight ontology for supporting energy-efficiency tasks-enabling generic evaluation of energy efficiency in the engineering phase of automated manufacturing plants. In *KEOD*, pages 337–344, 2013.

[56] Chang Liu, Yu Cao, Yan Luo, Guanling Chen, Vinod Vokkarane, Yunsheng Ma, Songqing Chen, and Peng Hou. A new deep learning-based food recognition system for dietary assessment on an edge computing service infrastructure. *IEEE Transactions on Services Computing*, 2017.

[57] Lingna Liu and Yalin Lei. An accurate ecological footprint analysis and prediction for beijing based on svm model. *Ecological Informatics*, 2018.

[58] Leandro Madrazo, A Sicilia, and Gonzalo Gamboa. Semanco: Semantic tools for carbon reduction in urban planning. In *Proceedings of the 9th European Conference on Product and Process Modelling*, 2012.

[59] Alexander Maedche and Raphael Volz. The ontology extraction & maintenance framework text-to-onto. In *Proc. Workshop on Integrating Data Mining and Knowledge Management, USA*, pages 1–12, 2001.

[60] Andrea Mauri, Achilleas Psyllidis, and Alessandro Bozzon. Social smart meter: Identifying energy consumption behavior in user-generated content. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pages 195–198. International World Wide Web Conferences Steering Committee, 2018.

[61] Grant McKenzie, Krzysztof Janowicz, and Benjamin Adams. A weighted multi-attribute method for matching user-generated points of interest. *Cartography and Geographic Information Science*, 41(2):125–137, 2014.

[62] Robert Meersman. Ontologies and databases: More than a fleeting resemblance. *STAR*, page 03, 2001.

[63] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[64] O Mont. Concept paper for the task force on sustainable lifestyles. *Third International Expert Meeting on Sustainable Consumption and Production*, pages 1–14, 2007.

[65] Ian Niles and Adam Pease. Towards a standard upper ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001*, pages 2–9. ACM, 2001.

[66] Bei Pan, Yu Zheng, David Wilkie, and Cyrus Shahabi. Crowd sensing of traffic anomalies based on human mobility and social media. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 344–353. ACM, 2013.

[67] Ali Parsa, Tooraj Abbasian Najafabadi, and Farzad Rajaei Salmasi. Implementation of smart optimal and automatic control of electrical home appliances (iot). In *Smart Grid Conference (SGC), 2017*, pages 1–6. IEEE, 2017.

[68] Teresa Pizzuti, Giovanni Mirabelli, Giovanni Grasso, and Giulia Paldino. Mesco (meat supply chain ontology): An ontology for supporting traceability in the meat supply chain. *Food Control*, 72:123–133, 2017.

[69] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 413–420. IEEE, 2009.

[70] Varun Rai and Adam Douglas Henry. Agent-based modelling of consumer energy choices. *Nature Climate Change*, 6(6):556–562, 2016.

[71] Taha H Rashidi, Alireza Abbasi, Mojtaba Maghrebi, Samiul Hasan, and Travis S Waller. Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. *Transportation Research Part C: Emerging Technologies*, 75:197–211, 2017.

[72] Jaclyn Rich, Hamed Haddadi, and Timothy M Hospedales. Towards bottom-up analysis of social food. In *Proceedings of the 6th International Conference on Digital Health Conference*, pages 111–120. ACM, 2016.

[73] Paula Rocha, Afzal Siddiqui, and Michael Stadler. Improving energy efficiency via smart building energy management systems: A comparison with policy measures. *Energy and Buildings*, 88:203–213, 2015.

[74] Martha G Russell, June Flora, Markus Strohmaier, Jan Pöschko, Rafael Perez, and Neil Rubens. Semantic analysis of energy-related conversations in social media. *Communicating Sustainability for the Green Economy*, page 223, 2013.

[75] Mariya Sodenkamp, Ilya Kozlovskiy, Konstantin Hopf, and Thorsten Staake. Smart meter data analytics for enhanced energy efficiency in the residential sector. 2017.

[76] Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, and Boris Villazón-Terrazas. How to write and use the ontology requirements specification document. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 966–982. Springer, 2009.

[77] Lukas G Swan and V Ismet Ugursal. Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable and sustainable energy reviews*, 13(8):1819–1835, 2009.

[78] Jacopo Torriti. 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, 2017.

[79] Mike Uschold and Michael Gruninger. Ontologies: Principles, methods and applications. *The knowledge engineering review*, 11(2):93–136, 1996.

[80] Alexandros G Valarakos, Georgios Paliouras, Vangelis Karkaletsis, and George Vouros. Enhancing ontological knowledge through ontology population and enrichment. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 144–156. Springer, 2004.

[81] Iana Vassileva, Fredrik Wallin, and Erik Dahlquist. Understanding energy consumption behavior for future demand response strategy development. *Energy*, 46(1):94–100, 2012.

[82] Markus Weiss, Adrian Helfenstein, Friedemann Mattern, and Thorsten Staake. Leveraging smart meter data to recognize home appliances. In *Pervasive Computing and Communications (PerCom), 2012 IEEE International Conference on*, pages 190–197. IEEE, 2012.

[83] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE, 2010.

[84] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. *Social media mining: an introduction*. Cambridge University Press, 2014.

[85] Emilio Zagheni, Venkata Rama Kiran Garimella, Ingmar Weber, et al. Inferring international and internal migration patterns from twitter data. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 439–444. ACM, 2014.

[86] Hamed Zamani and W Bruce Croft. Relevance-based word embedding. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 505–514. ACM, 2017.

[87] Zhenhua Zhang, Qing He, and Shanjiang Zhu. 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, 2017.

[88] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.

[89] Kaile Zhou and Shanlin Yang. Understanding household energy consumption behavior: The contribution of energy big data analytics. *Renewable and Sustainable Energy Reviews*, 56:810–819, 2016.

[90] Wen-Yuan Zhu, Yu-Wen Wang, Chin-Jie Chen, Wen-Chih Peng, and Po-Ruey Lei. A bayesian-based approach for activity and mobility inference in location-based social networks. In *Mobile Data Management (MDM), 2016 17th IEEE International Conference on*, volume 1, pages 152–157. IEEE, 2016.

[91] Zack Zhu, Ulf Blanke, and Gerhard Tröster. Recognizing composite daily activities from crowd-labelled social media data. *Pervasive and Mobile Computing*, 26:103–120, 2016.