Social Smart Meter: Identifying Energy Consumption Behavior in User-Generated Content

Andrea Mauri
Delft University of Technology
Delft, The Netherlands
a.mauri@tudelft.nl

Achilleas Psyllidis
Delft University of Technology
Delft, The Netherlands
a.psyllidis@tudelft.nl

Alessandro Bozzon
Delft University of Technology
Delft, The Netherlands
a.bozzon@tudelft.nl

ABSTRACT

Having a thorough understanding of energy consumption behavior is an important element of sustainability studies. Traditional sources of information about energy consumption, such as smart meter devices and surveys, can be costly to deploy, may lack contextual information or have infrequent updates. In this paper, we examine the possibility of extracting energy consumption-related information from user-generated content. More specifically, we develop a pipeline that helps identify energy-related content in Twitter posts and classify it into four categories (dwelling, food, leisure, and mobility), according to the type of activity performed. We further demonstrate a web-based application – called *Social Smart Meter* – that implements the proposed pipeline and enables different stakeholders to gain an insight into daily energy consumption behavior, as well as showcase it in case studies involving several world cities.

CCS CONCEPTS

• **Information systems → Web searching and information discovery**; • **Human-centered computing → Social media**;

KEYWORDS

social media, energy consumption, machine learning

ACM Reference Format:
Andrea Mauri, Achilleas Psyllidis, and Alessandro Bozzon. 2018. Social Smart Meter: Identifying Energy Consumption Behavior in User-Generated Content. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3184558.3186977

1 INTRODUCTION

The performance of day-to-day human activities requires considerable amounts of different forms of energy. People consume energy resources while utilizing various home appliances, but also while moving from home to work or to other places relating to leisure activities. Energy consumption levels increase in places where people agglomerate, such as in cities and metropolitan regions. In shaping sustainable energy systems, it is important to gain an insight into individual energy consumption behavior [1].

Such insights can be used to raise awareness about energy consumption [3] or to inform policies about how energy is being consumed. However, the variety of energy forms, in combination with the multitude of human activities, renders the quantitative measurement of individual consumption behavior non-trivial.

Traditionally, measurements of energy consumption are derived primarily from two sources: 1) *Smart meters*: Sensor devices that record and measure consumption of – primarily electric – energy in different time intervals; and 2) *Surveys*: Questionnaires or structured interviews (e.g. household travel surveys), focusing on qualitative characteristics of energy consumption behavior [7]. While being the gold standard when it comes to reliable quantitative measurements of and qualitative information about energy consumption, such data sources also have several drawbacks. Smart meters are costly to deploy, and the data they generate are usually proprietary and lack contextual information. Surveys, though semantically rich, are costly, non-scalable, and infrequently conducted.

In recent years, the proliferation of social data has given rise to a new source of information about peoples' daily spatiotemporal behavior [2, 4, 6]. While other sources, such as mobile phones, satellite imagery, and traffic sensors are increasingly used in deriving information about energy consumption behavior, in this work we focus on user-generated content from social media. Although biased and noisy, data from these sources are becoming widely available, inexpensive to collect, dynamic and frequently updated. Given that they are the byproduct of – or refer to – daily human activities, it is reasonable to assume that information about energy consumption could be embedded in their semantic signatures.

Contribution. In this demo, we introduce *Social Smart Meter*, a web-based application that offers different users the opportunity to gain an insight into four types of energy consumption behavior (dwelling, food, leisure, and mobility), calculated at the neighborhood level. The demo showcases the extent to which data from social media could be used as a complementary source of information about individual energy consumption behavior in several world cities such as Amsterdam, Jakarta, and Boston.

2 SOCIAL DATA PROCESSING PIPELINE

Social Smart Meter aims to identify energy-related content in social media posts and further classify it into four categories, namely:

- *Dwelling*: it refers to the consumption of energy due to the usage of home appliances (e.g., washing machine, gaming console)
- *Food*: it refers to the use of resources associated with the preparation and processing of food.
- *Mobility*: it refers the energy required for moving from one place to another.

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3186977

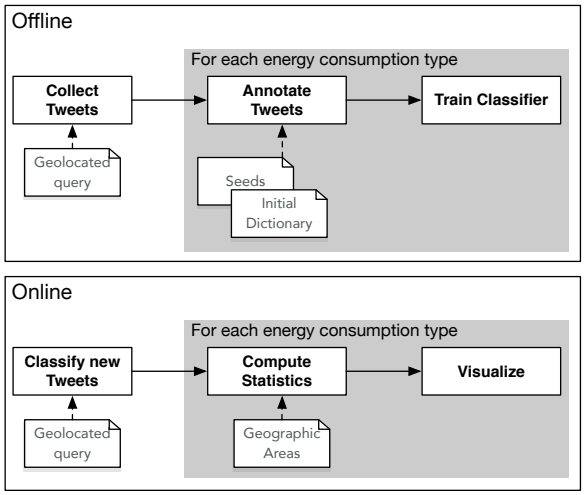

Figure 1: Overview of the Social Smart Meter pipeline

- *Leisure*: it refers to the energy required for performing leisure activities (e.g., watching TV, playing video-games, partying)

These categories cover a considerable spectrum of the activities impacting on the energy footprint of an individual's lifestyle [3].

As shown in Figure 1, the proposed pipeline can be divided into an *offline* and an *online* mode. In the offline mode, social media posts are collected and used for training a classifier. Given that manual annotation of the posts is not scalable, we use a distant supervision learning approach. The labeling function is a hybrid dictionary-similarity heuristic, in which a post is annotated according to a dictionary, as well as according to the similarity of the training dataset to a set of energy-related posts. In the online mode, the Social Smart Meter web-application is deployed. The application identifies in real time social media posts relating to a particular energy consumption activity, while also performing statistical analyses that are fed to an interactive visualization.

2.1 Data collection

In this work we use Twitter as a source of social media data mainly for two reasons: it provides easy access to the data through the Stream API; and, it often collects content that people post on other social networks (e.g. music consumption, gaming activities, visit of places). We are only interested in geo-referenced posts, so the tweets are collected by querying the Twitter Stream API using a bounding box corresponding to the area of interest.

2.2 Annotation and classification of tweets

In the past, distant supervision learning has been successfully used for classifying Twitter content (e.g. detection of political trends [5]). In our work we use a hybrid dictionary-similarity distant supervision. The main intuition is that using only a dictionary might lead to a highly noisy dataset, especially in short texts like tweets, since a single word can have different meanings according to the context. Examples include slang words or proverbs (e.g., "pot calling the kettle black").

The process for the creation of the training data considers two sets: 1) *Dictionary*: a set of terms related to a specific consumption type – e.g. ingredients or cooking utensils are associated with the *food* category. And 2) *Seeds*: a set of relevant tweets for a specific energy consumption type. This set is manually built by selecting the tweets from the ones retrieved in the data collection step.

Both the *Dictionary* and the *Seeds* datasets are language dependent. For machine learning purposes, the considered features are: 1) the vector representation of the tweets obtained by training a `Doc2Vec` model on the tweets corpus; and 2) the vector representation of the words obtained from the pre-trained `Word2Vec` model on the Google News corpus [1].

The process starts by retrieving a set of candidate tweets, considering the ones that contain a *dictionary* term. Then it computes the similarity between the candidates and the centroid of the *seeds*. Given a confidence range $[h, l]$, it labels as positive example the tweets whose similarity is greater than h, and as negative if it's lower then l, leaving the others unlabeled.

Then, the pipeline includes a module devoted to dictionary expansion. This is to account for incompleteness, either due to the labor-intensiveness of the manual creation of an extensive set of words related to a particular energy consumption type, or due to the lack of knowledge about all the relevant terms. A set of candidate words is created by taking all the terms contained in tweets labeled as positive. Then, the process considers a word valid if a dictionary term appears on the list of similar words obtained from the `Word2Vec` model. The main intuition is that, in this way, we are able to select only the words that belong to the same context of the dictionary. The process is iterated until either all the tweets are labeled, or no more candidate tweets can be found.

Finally, the annotated dataset is used to train a classifier for each of the energy consumption types. We employ a logistic regression classifier, using as feature the vector representation obtained from the `Doc2Vec` model.

Annotation Performance. We collected 219,436 geo-referenced tweets in forty days through the Twitter Stream API without any keyword filtering. The collection has been performed between Jan. 23, 2017 and Feb. 26, 2017, within the bounding box area of Amsterdam. The starting dictionaries were created (both in English and Dutch) by manually adding words and crawling e-commerce website and Wikipedia pages. The tweets were annotated with the process described in the previous section.

Next, we trained the logistic regression classifiers. Given that the positive and negative examples were not balanced, we attempted to balance the dataset by randomly sampling 100,000 tweets for the training set and 1000 for the test set.

Table 1 and Table 2 summarize the obtained results, respectively in terms of dictionary size and classifiers performance. Our approach led to and increase of the dictionary size by about ten times. This step is essential in our methodology, as it is unlikely for a single person to build, even with the help of external sources, an appropriate dictionary, and to maintain it. The performance of both the tested classifiers is promising, with satisfactory AUC ROC and accuracy values.

[1] https://code.google.com/archive/p/word2vec/

Table 1: Dictionary Size

Type	Starting	Expanded
Leisure	145	1418
Mobility	167	1420

Table 2: Performances of the classifier in identifying the positive class

Metric	Leisure	Mobility
ROC	0.75	0.8
Accuracy	0.76	0.8
Precision	0.62	0.56
Recall	0.70	0.74
F1	0.66	0.65

Table 3: Example of labeled tweets

Text	Label
So i drove myself from Ijegun to Yaba today. It might not seem like much but that blasted car is manual and i was so scared of driving in Lagos	Mobility
Brands rushing to throw "chemikilz" under the bus for "clean" "healthy" deserve to fail.	Not Relevant
party 50vip grap is gemaakt tour van orgaan-klap! #fun #nightout #party #sugarfactory	Leisure

Table 3 shows some notable examples of tweets classified by our pipeline. It can be noted that it correctly classified both Dutch and English tweets, but it also recognized and discarded the tweets containing proverbs (i.e. “to throw under the bus”).

2.3 Analysis and Visualization

Social Smart Meter uses the annotated tweets to perform a series of analyses. It aggregates the tweets into various spatial units (e.g., neighborhoods, census tracts, postcode areas, etc..) and into hourly time slots. These aggregations helps observe the concentration of different types of energy consumption activities at various scales and times of the day.

To better characterize a specific energy consumption type, Social Smart Meter computes the frequency of occurrence of relevant terms, aggregated by area and time. Moreover, for the mobility energy consumption type, it calculates the user displacement and radius of gyration as implemented in [6].

Social Smart Meter provides views to explore the extracted information. In particular, it projects the overall energy consumption onto a map (Figure 2). The color of the different areas corresponds to the amount of energy consumption registered, measured as the number of tweets posted in each zone. The pie chart shows the percentage of the different energy consumption types. It, further, shows the same information (Figure 3) for each energy consumption type, together with a histogram of the most frequent terms appearing in the tweet. Moreover, for the *Mobility* type, as depicted in Figure 3(d), it shows the histograms of the user displacement and radius of gyration. All the visualizations allow to filter the statistics by area and time range.

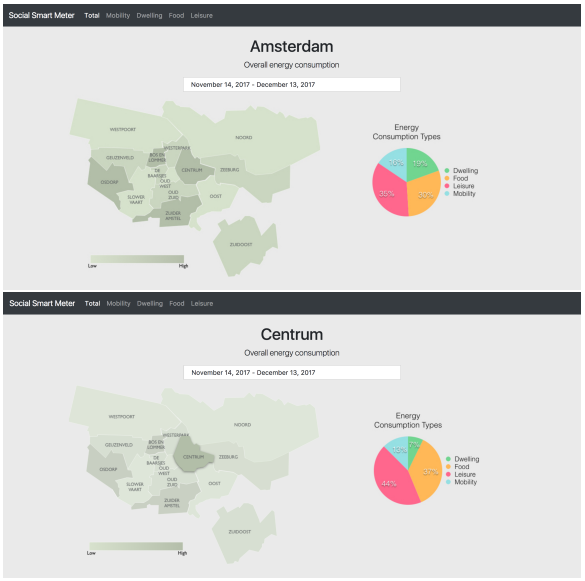

Figure 2: Social Smart Meter visualization of the total energy consumption (top). By clicking on a region the corresponding statistics appear (bottom)

3 SOCIAL SMART METER ARCHITECTURE

The Social Smart Meter is implemented in Python. The tweets annotation phase uses the *gensim*[2] package for handling the Word2Vec and the Doc2Vec models. The web application is built using the Python microframework Flask[3].

Figure 4 shows the Social Smart Meter architecture. First, the Twitter Stream API is used to retrieve the geo-referenced tweets within a specific area. The tweets are, then, classified using the model trained in the *offline* phase (Figure 1); non-relevant tweets are discarded, while the relevant ones are stored in a MongoDB database. Next, the *Energy Behavior Analytics* module periodically analyzes tweets, as described in Section 2.3. The tweets are aggregated into spatial units (also stored in the database), while calculating term frequencies, user displacement, and the radius of gyration. These statistics are accessible through a Web API, which is used by the *Views* component to build the visualizations.

4 DEMONSTRATION

In the demonstration, we will focus on a set of world cities (e.g. Amsterdam, Jakarta, Boston). Attendees will be able to explore, in real time, energy consumption information extracted from the tweets generated in the aforementioned cities, using the interfaces shown in Section 2.3. In addition, the demonstration will include a step-by-step explanation of the social data processing pipeline, as well as a visual exploration of the different dictionaries. Additional information and resources associated with the demonstrator are available at the following address: http://social-glass.tudelft.nl/social-smart-meter/

[2]https://radimrehurek.com/gensim/
[3]http://flask.pocoo.org/

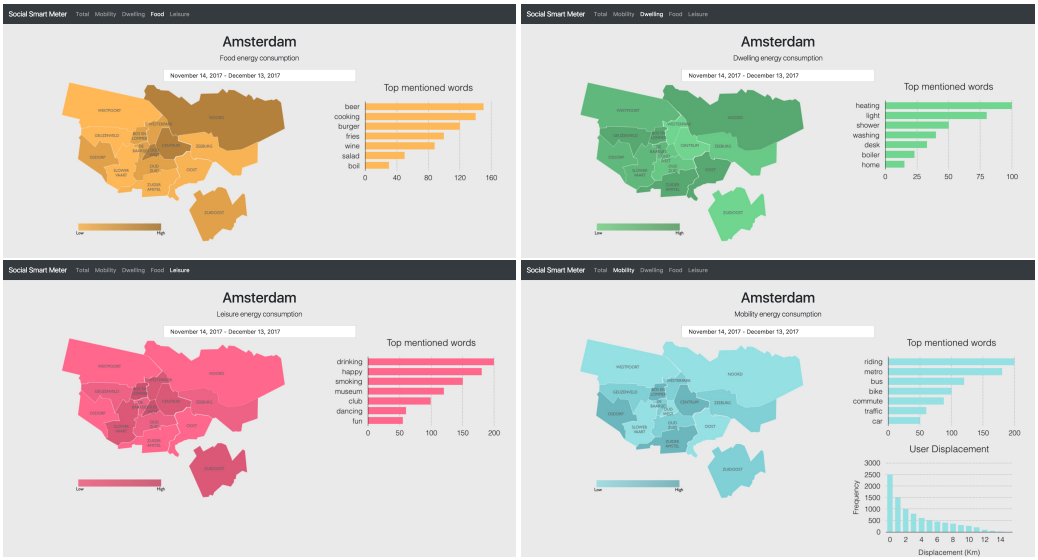

Figure 3: Social Smart Meter disaggregated visualizations of energy consumption. In clockwise order, from top-left: Food, Dwelling, Leisure, and Mobility categories.

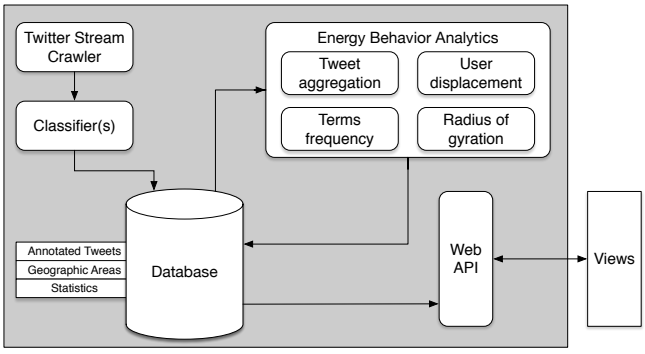

Figure 4: Social Smart Meter architecture

5 CONCLUSIONS AND FUTURE WORK

In this paper, we described a pipeline that identifies and classifies energy consumption-related content in user-contibuted posts, with emphasis on Twitter. We further demonstrated how the proposed pipeline is implemented in the *Social Smart Meter* Web application, and instantiated over several world cities.

Future work will emphasize the effectiveness of the proposed pipeline and web application, with regard to estimating energy consumption behavior from social media. Particular attention will be given to precision, with emphasis on the value of the confidence interval in the annotation phase, the size of the initial seed set, feature engineering (using e.g. Word2Vec, TF-IDF, n-grams etc.), and the use of other classifiers. We also aim to test the adaptability of the pipeline, by deriving data from different social media platforms (e.g. Instagram). The correspondence between energy consumption information identified in social media and that from traditional sources (e.g. smart meters, household travel surveys) will serve as a validation of our approach, subject to availability of fine-grained (i.e. at a similar spatiotemporal resolution) energy data. We will investigate methods to link social media posts with concrete values of energy consumption (in terms of e.g. kWh or CO_2 emissions). Moreover, we aim to test the scalability of the proposed pipeline, by deploying it in different urban and regional contexts.

ACKNOWLEDGMENTS

This work is supported by the JPI Urban Europe Project CODALoop (Project no. 646453) and the Amsterdam Institute for Advanced Metropolitan Solutions (AMS Institute).

REFERENCES

[1] M. A. Alrowaily and M. Kavakli. 2015. The Use of Smart Meters and Social Media in Promoting Conservation Behaviour. In *2015 8th International Conference on u- and e-Service, Science and Technology (UNESST)*. 50–56. https://doi.org/10.1109/UNESST.2015.24

[2] Michela Arnaboldi, Marco Brambilla, Beatrice Cassottana, Paolo Ciuccarelli, and Simone Vantini. 2017. Urbanscope: A Lens to Observe Language Mix in Cities. *American Behavioral Scientist* 61, 7 (2017), 774–793. https://doi.org/10.1177/0002764217717562

[3] Pineda Revilla B., Bertolini L., Pfeffer K., and Savini F. 2016. Changing Energy Needs. Pursuing 'energy conscious lifestyles' through data-driven social learning and individual behavior adaptation processes. (2016).

[4] T. Bodnar, M. L. Dering, C. Tucker, and K. M. Hopkinson. 2017. 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* 47, 10 (Oct 2017), 2627–2640. https://doi.org/10.1109/TSMC.2016.2618860

[5] Micol Marchetti-Bowick and Nathanael Chambers. 2012. Learning for Microblogs with Distant Supervision: Political Forecasting with Twitter. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL '12)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 603–612.

[6] Achilleas Psyllidis, Alessandro Bozzon, Stefano Bocconi, and Christiaan Titos Bolivar. 2015. *A Platform for Urban Analytics and Semantic Data Integration in City Planning*. Springer Berlin Heidelberg, Berlin, Heidelberg, 21–36. https://doi.org/10.1007/978-3-662-47386-3_2

[7] A. G. Ruzzelli, C. Nicolas, A. Schoofs, and G. M. P. O'Hare. 2010. Real-Time Recognition and Profiling of Appliances through a Single Electricity Sensor. In *2010 7th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*. 1–9. https://doi.org/10.1109/SECON.2010.5508244