User engagement analysis for smart buildings based on social trend tracking

Jiao Jiao
Fraunhofer Institute for Systems and Innovation Research ISI
Breslauer Str. 48
76139 Karlsruhe
Germany
jiao.jiao@isi.fraunhofer.de

Heike Brugger
Fraunhofer Institute for Systems and Innovation Research ISI
Breslauer Str. 48
76139 Karlsruhe
Germany
heike.brugger@isi.fraunhofer.de

Michael Behrisch
Utrecht University
Princetonplein 5
3584 CC Utrecht
The Netherlands
m.behrisch@uu.nl

Wolfgang Eichhammer
Fraunhofer Institute for Systems and Innovation Research ISI
Breslauer Str. 48
76139 Karlsruhe
Germany
Wolfgang.Eichhammer@isi.fraunhofer.de

Keywords

domestic energy efficiency, smart buildings, user engagement

Abstract

The building sector is one of the major contributors to energy demand. Therefore, policy makers and the public are increasingly aware of the need to address global energy sustainability challenges in the building sector. With the growing prevalence of digitalization and Artificial Intelligence (AI), energy supplying infrastructures are undergoing profound changes by the emergence of smart buildings. Smart buildings aim to not only control devices and appliances inside, but can also create active micro-grids or virtual power plants by adapting to grid's conditions and communicating with other buildings. The great potential to provide flexibility for the integration of fluctuating renewable energy make smart buildings one of the central elements in the future energy system. Although smart building technologies are widely introduced and discussed in academia and industry, its services in real life are still rare. If smart buildings should unfold their potential contribution to the future energy system, one important barrier to overcome is the adoption of these technologies by households and investors. To analyse where the barriers in the adoption decision lie, we apply multiple visual analysis approaches. First, keyword extraction together with network visualization is used to identify the most trending sub-topics of smart building concept. Second, a text mining approach is applied to discover how smart building concepts diffuse over time and spatially based on scraped social media posts and online news. Third, this paper further analyses the user engagement degree in the diffusion of smart building concepts by user profile mining. The paper concludes with deriving recommendations on how (nudging) policies could be designed and how smart building technologies could be enhanced to increase user adaptation.

Introduction

Thanks to the boom of digitalization and AI, a new era of smartness is ushered in the building sector. As an integral part of smart grids and smart cities, smart buildings become one of the central elements in the future transformation of the energy system towards climate neutrality. The concept of smart buildings represents the evolution of modern infrastructure and its coverage has expanded from automatic control systems for comfort and safety of occupants to layout design, long-term maintenance and sustainability (Panchalingam 2019). In particular, the public increasingly recognises the importance of smartness in buildings to address the challenges of global environmental and energy sustainability. After all, the energy supplying infrastructure is undergoing profound changes with the emergence of smart buildings which potentially also supply energy and provide flexibility by managing and storing additional energy from the demand side.

Energy efficiency and usage flexibility are not only a question of optimizing components, but also of understanding and addressing user behaviour. Although the concept of smart buildings has been widely introduced in the academia and the industry, its services are still rare in practice, especially for residential buildings. The question of how decision-makers can be nudged to participate and invest in smart building solutions remains, and it is a key factor in the sector's contribution to the energy transition. One way to gather information about the factors which impede actual adaptation is through surveys. While surveys have many advantages, in practice it is difficult to gather information from a large number of respondents (Hojjati 2016, Baudier 2020). Additionally, through surveys it is challenging to capture the real perceptions, since respondents are prone to answer with a social-desirability bias.

With the growing prevalence of online news and social media, large amount of user-generated data is collected with high temporal and spatial resolution, which provides much more real information than previously available and strengthens the progressively expanding role of data science in the sociology domain (Wirtz 2016, Gefen 2017, Chuan 2019). Therefore, tracking the evolution and coverage of concepts on the online news and social media is of theoretical and practical significance for social scientists, policy makers and business managers, because it can sharpen our insights into the convergence and divergence of concepts in the society by taking advantages of user experiences, which plays a vital role in nudging the user engagement of those concepts.

Since the public has been engaged in both knowledge exchange and idea production via social media and online news, society entered a new era of information diffusion, which poses an enormous challenge for social trend analysis, especially for complex concepts like smart buildings. A fuzzy concept often has hundreds of related keywords with millions of user-generated texts. A visual analysis approach can provide an easy-to-understand metaphor to display all of their relationships at once without visual clutter and ambiguity. In the meantime, concept diffusion is a spatio-temporal phenomenon, which is technically demanding to design an intuitive national-wide visualization across a long time span. Visual analysis approaches can efficiently and effectively represent the concept spreading over a long timeframe, which allows experts to focus on a selected time period or region of interest while keeping others in context. Besides, user-generated texts always accompany with high bias. Combining with more objective data sources, such as publications, patents and so on, expert knowledge from the academia and the industry can be made full use of to avoid the one-sidedness of self-claimed information.

In this paper, we combine visual analysis with multiple data driven approaches to analyse how the concept of smart buildings diffuses in Germany and in which topics users are highly engaged. First, keyword co-occurrences together with networks and sunburnt charts are used to identify the most trending topics in the context of smart buildings. Second, to understand how the smart building concept diffuses, a spatio-temporal analysis is applied based on social media posts and online news which were collected in an automated manner (i.e. web scraping). Third, user profile mining further represents the user engagement degree in the diffusion. Based on the casual inference from above visual analysis, this paper concludes with policy recommendations on nudging user engagement in smart buildings, especially smart residential buildings.

Iterative Data Collection based on Web Scraping and Keyword Extraction

To assess the public interest, this paper selects two data sources: social media and online news. The former one can represent the daily attitude of end users, while the latter stands for the macro dimension, rather than individuals. According to the survey of social media ranking in Germany (Hootsuite 2019), Facebook and Twitter are the most popular text based social networks. However, during manual pre-research on Facebook, keywords, e.g., "smart buildings", "energy-efficient buildings" and "smart home", only appear in the posts from large cities. Twitter has a better data availability in small cities than Facebook. Thus, Twitter is chosen as the social media source for further web scraping. As for online news, considering the comprehensive coverage of the Nexis database (LexisNexis 2020), an online platform including over 500 regional and national German newspapers, our online news are collected based on its query results in the workflow presented in Figure 1.

The data collection is a multi-step iterative procedure as described in Figure 1. To avoid the bias of user-generated texts, we first search for the keyword "smart building" in the PatBase patent database (Minesoft Ltd and RWS 2020). PatBase is a global full-text patent database covering over 100 authorities, which

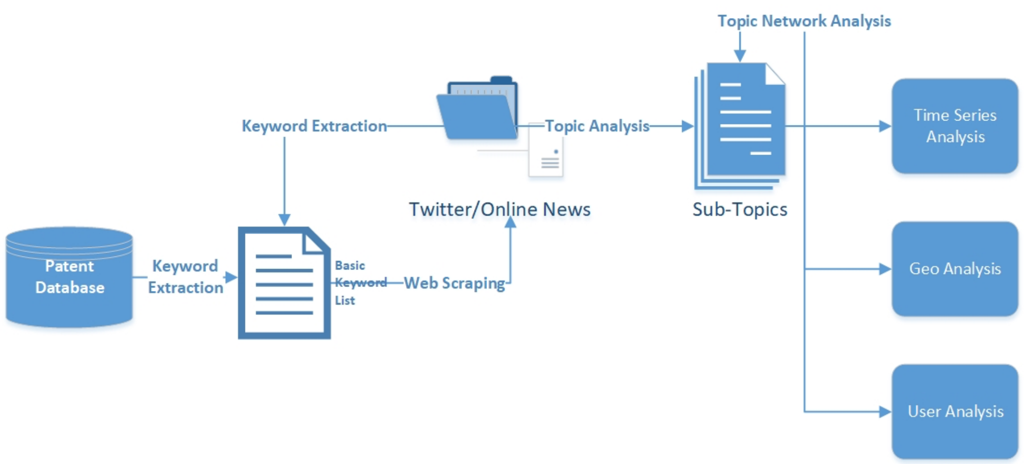

Figure 1. Our analysis workflow allows us to model topics, which we subsequently correlate with temporal, geographical, and user profile aspects, in order to investigate the diffusion of those topics.

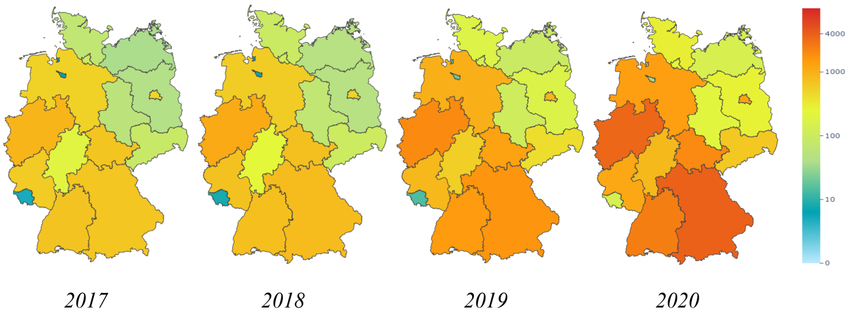

Figure 2. Example for Small Multiples of Choropleth Maps.

hosts more than 140 million patents organized into different categories. Thus, keywords in the both categories and full-texts of the search results are extracted. Since our texts are mixed with German and English, even Spanish and French, we apply the Rapid Automatic Keyword Extraction (RAKE) algorithm to extract keywords. This algorithm is designed as an unsupervised, domain-independent, and language-independent method from individual documents, which can deal with multiple languages at the same time (Rose 2010). RAKE selects keywords based on their interconnectedness, i.e., how strongly the words are linked to parts of the article, and how frequently the words appear in the article. Therefore, with a simple set of stopwords (words without real meaning in the specific context), RAKE only needs go through the whole text once, which makes it suitable and efficient for massive collections (Rose 2010). The top ten extracted keywords together with their categories in the patent database are regarded as the basic keyword list.

With this basic keyword list, we start scraping texts from the above two data sources, Twitter and Nexis. With our self-developed Python tool binding to Twitter developer account, specific historical Tweets within Germany are scraped. Furthermore, related online news published in Germany are scraped by the download function provided by Nexis. Among all the scraped texts, we again extract keywords with the RAKE algorithm and use those keywords to scrap other Tweets and online news respectively, until there is no more new words to scrape in their separated top 300 keywords. As a technical contribution in this work, we improve the RAKE algorithm for Twitter keyword extraction, in terms of its special feature: hashtags. Since hashtags are by design topic terms, we gave them higher priority during the keyword extraction. Specifically we do this by adding more weight in our RAKE algorithm calculation. Overall, there are three things we try to avoid. First, we want to circumvent duplications in the web scraping. If a top 300 keyword is the substring of another, then this keyword is excluded from the scraping procedure but we increment its keyword counter. Second, only keywords focusing on the smart building concept will be considered. For example, although "renewable energy" is core of smart buildings, we do not take it in consideration, since it still appears in other scenarios like automobiles, biomass energy, wave and tidal. If "renewable energy" is counted into our keyword list, many texts only remotely linked with the smart building context would be included in the corpus. Lastly, we need to handle multiple keyword entities, e.g., "energy efficient smart buildings" and "intelligent comfort and energy control". We decide to keep keywords within 3-grams (within a sequence of three keywords). Otherwise, it may merge too many keywords into one phrase.

After ten iterations, in total 659,985 German Tweets and 262,793 online news published in Germany have been scraped based on our keyword extraction strategy. It appears that our iterative data collection method based on Twitter and online news scraping can provide at least millions of user-generated data within a narrowed concept, which can largely broaden data sources for analysis. Besides, compared to surveys, our method is more likely to reflect the real opinion of end users, since they speak out their ideas for their own intention, instead of being asked. We even highlighted keywords defined by users (hashtags) in our algorithm, which takes user experiences as a reference and enhances the accuracy of our method. Furthermore, we combine keywords extracted from patent database and patent categories, rather than starting from scratch in Twitter and online news, which takes the perspectives and the expert knowledge from the academia and the industry into consideration. But with the further iterative keyword extension, keywords from patent database do not have a dominant influence on the final results.

Visual Analysis of Concept Diffusion

One central idea of this paper is to combine data driven techniques with visual analysis concepts. As the reader might not be aware of the vocabulary in these special domains, we want to add short description along with the terminology and add a reference to related work for the interested reader.

Geo-Spatial Visualization: Concept diffusion is a spatio-temporal phenomenon, which requires a visualization concept that combines the spatial and temporal dimension. One established approach is (1) choropleth maps, i.e., thematic maps in which areas are colored in proportion to a statistical variable, and (2) small multiples, i.e., series of similar charts using the same scale and axes, allowing them to be easily compared (Ward 2010). We demonstrate this concept in Figure 2. Each map displays the number of Tweets at a specific timestamp (here per year), which helps us to compare the differences among regions. Ordering multiple maps in a juxtaposition (continuously next to each other) we can see how concepts gradually spread from one region to other regions.

Network Visualization: As complex concepts might contain hundreds of keywords, discovering all of their relationships and pointing out the promising topics can be difficult in a textual form. Networks visualizations (Ribecca 2017), as here in node-

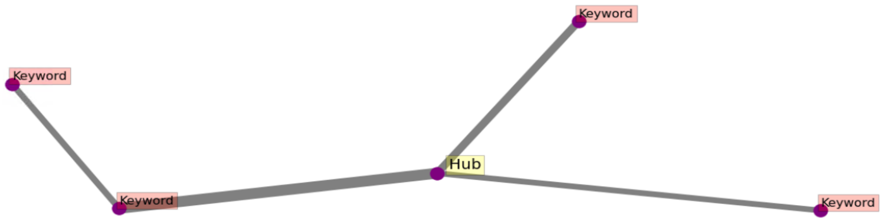

Figure 3. Example for Network Visualization.

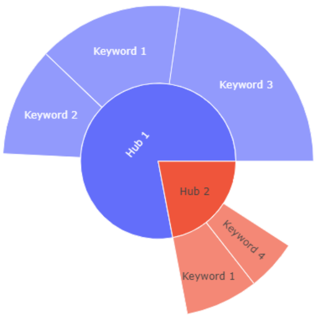

Figure 4. Example for Sunburnt Chart.

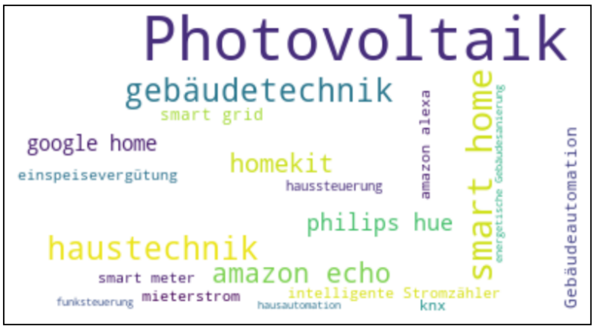

Figure 5. Extracted Keywords from Twitter.

Figure 6. Extracted Keywords from Online News.

link diagrams, help in this case (Wasserman 1994). Keyword co-occurrences, i.e., counts how often keywords exist simultaneously in a document, can represent the relationships among keywords (Isenberg 2016). The higher co-occurrences documents have, the more they should be grouped as the same topic. Thus, we build up the network of keywords based on their co-occurrences. In our case, nodes are keywords and the arcs/edges represent their co-occurrences. The thicker the edges are, the more co-occurrences those two keywords have (Figure 3). The degree of each node can be mapped to how many links they have with other nodes. The higher degree the node has, the higher possibility it can be regarded as the center of a cluster (Hub).

Sunburst Chart: Since our keywords are inherently hierarchical, one example is "Building Integrated Photovoltaics (BIPV)" is contained in "photovoltaics", we also feature a sunburst chart visualization (Munzner 2014). This chart uses nested rings to convey layers of a hierarchy (Figure 4). Each ring is subdivided based on the number of nodes it contains and the percentage of each node takes up. In our case, the first level is divided by hub keywords and the keywords at the second level or higher levels have co-occurrences with the corresponding hub keywords more than the threshold. The size of each node at the first level depends on the word frequencies of hub keywords, while nodes at other levels are distributed in terms of the co-occurrences among hub and keywords on the path to these nodes. Sunburst charts are effective space-filling visualizations, thus, complex relationships can be displayed within limited space effectively and clearly.

Visual Analysis of the Smart Building Concept Diffusion in Germany

KEYWORD ANALYSIS

In general, after stemming (reducing a word variant to the word stem) (Jivani 2011), 84 Twitter keywords and 70 online news keywords within smart building concept are extracted respectively. As depicted in Figure 5 and Figure 6, keywords from Twitter are daily life oriented, while online news keywords are more formal in their style. Although both of them are highly related to smart home and photovoltaics, Tweets are tend to discuss smart products (e.g. Amazon Echo, HomeKit), whereas online news puts forward more abstract concepts in the energy domain (eg. energetische Sanierung, nachhaltiges Bauen). In other words, smart building diffusion among Twitter is driven by enterprises in an active way. And when the involved companies have reached a certain amount, it will have influence on the society and policy making, which can be regarded as a bottom-up mode. On the contrary, online news are published in terms of policies or government strategies, where information goes through top-down direction. Therefore, technology keywords extracted from Twitter are more up-to-date than that of online news.

TOPIC ANALYSIS

As explained in the last chapter, network visualization can be the basis for analysing the relationships between and clusters of keywords. Tweets can only contain 160 characters, while online news are generally longer. At the same time, Tweets are often more optimized concerning keywords (e.g. hashtags) than news articles. Therefore, different thresholds for keyword co-occurrences were necessary to account for the different characteristic of the sources. While the threshold for Twitter was set to 100 co-occurrences between keywords, the threshold for online news was set to 500 to identify meaningful patterns.

As presented in Figure 7, Twitter keyword network can be divided into 3 clusters, whose centroids are smart home ("smart* home*"), smart grid ("smart* grid*") and photovoltaics ("Photovoltaik"). As the largest cluster, smart home consists of two parts: smart appliances and house technology ("Haustechnik*"). The online news network consist of two clusters: own electricity ("eig* Strom*") and smart appliances ("smart* Geräte*"). Obviously, due to the text length, keywords extracted from online news have much more co-occurrences than those of Twitter. However, each of the above cluster is in a star shape, which means edge keywords are almost directly connected to hub (central keywords). Then, if we search for keywords frequently appearing together with a second level keyword, we always get other second level keywords. Therefore, we only consider two keyword co-occurrences for topic distribution and those central keywords are regarded as the topics within the context of smart buildings.

After recognising the hubs from network visualization, a sunburst chart can explicitly reveal the topic distributions. In order to get a clear overview, we set their co-occurrence thresholds for visualization as 200 and 1,000 respectively for Twitter and online news. In general, topic distributions of Twitter and online news are quite similar. Both of them lay emphasis on energy production and smart home, especially energy production takes the largest share. The only difference is that Twitter has a clear small topic, smart grid, while online news merges it into smart home direction (Figure 8). Within the energy production topic, Twitter focuses on photovoltaics, while online news has a broad horizon. Compared with energy production focused by online news, Twitter users pay more attention to prosumer behaviours and novel photovoltaic formats. As for the smart home topic, those keywords of Tweets encompass the majority of smart products, such as Google Home, HomeKit and so on. Although

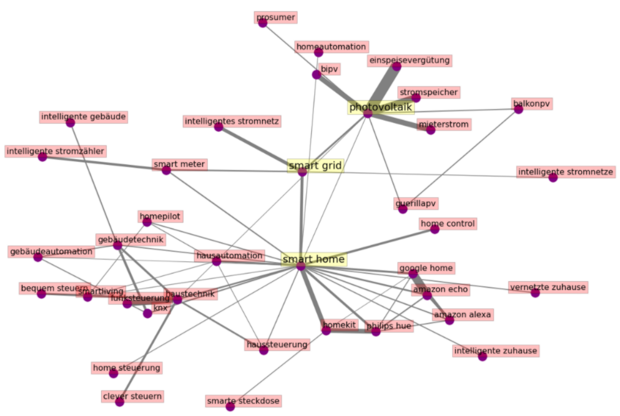

Figure 7. Twitter Keyword Network.

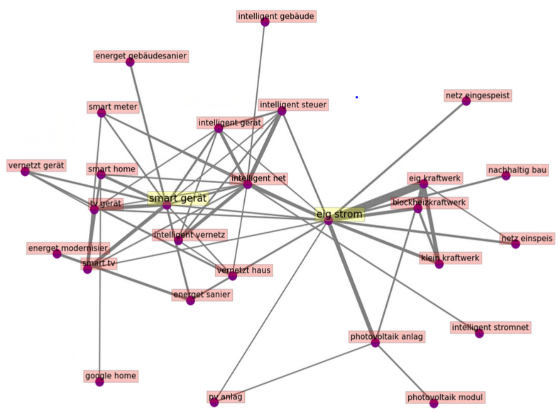

Figure 8. Online News Keyword Network.

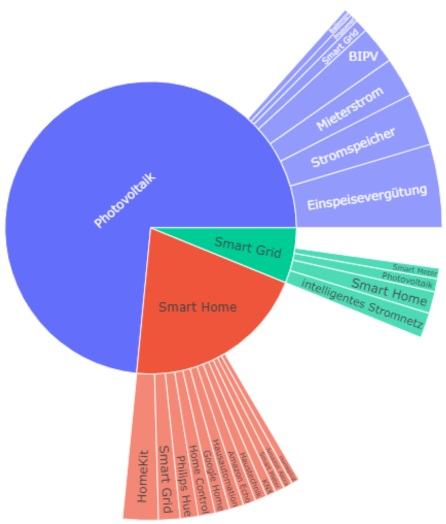

Figure 9. Twitter Keyword Co-occurrence Distribution.

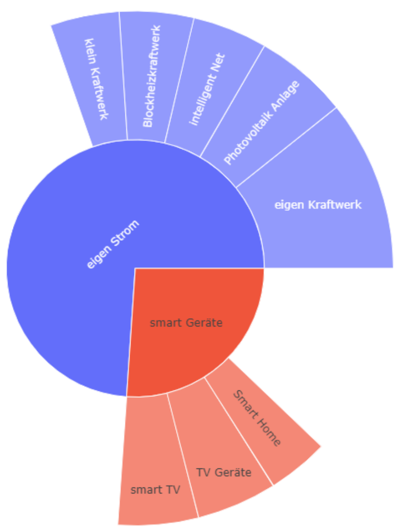

Figure 10. News Keyword Co-occurrence Distribution.

photovoltaics takes the largest share, smart home even has the most keyword pairs in Twitter, since there are more innovations and active anchors in this direction. However, when online news discuss smart appliance ("smart* Geräte*"), they mainly refer to smart TV. Obviously, these differences among keyword pairs are caused by the time sensitivity and publishers of both platforms, as discussed in Keyword Analysis section.

TEMPORAL DIFFUSION ANALYSIS

Although topic distributions of Twitter and online news are similar, their temporal developments in Germany are quite different. Generally online news has an exponential growth, although there are two stable phases: from the beginning of 2013 to the beginning of 2016 and from 2018 to 2019 (Figure 12). However, the number of Tweets has two clear peaks: March 2012 and November 2017 (Figure 11). Statistically, it is difficult to explain this trend. Inspired by time series analysis, we decomposed this trend into different components in terms of topics. Each topic includes the hub and their top keyword pairs in Figure 9.

The topic of photovoltaics leads to the first peak, which takes up almost 80 % of the total Tweets during the topic surging periods (Figure 15). According to 2009 Amendment of the Renewable Energy Sources Act (EEG 2009) (Federal Ministry of the Environment, Nature Conservation and Nuclear Safety 2009) and 2012 Amendment of the Renewable Energy Sources Act (EEG 2012) (Federal Ministry of the Environment, Nature Conservation and Nuclear Safety 2012), photovoltaic systems that went into operation between 2009 and 1 April 2012 could also receive feed-in tariffs for the electricity they consume themselves. This policy and its amendments largely boosted the installation of photovoltaic systems in households, which perfectly explains why the amount of Tweets starts increasing from 2009 and achieves the first peak in the March 2012. Although Twitter is daily life oriented, strategic planning policies, especially for fiscal/financial incentives, can still have a strong and real-time influence on the discussion. As for the second peak in Figure 11, it is caused by the smart home topic (Figure 14). Between September 2016 and December 2017, Amazon, Apple, Google and other technology anchors released their new smart home products, which contributes to the soaring Tweets. However, compared to the earlier policy impacts, the influence of these key events do not lead to one peak but rather a higher level of Tweets over several months. Thus, we can conclude that the commercial events had a relative long propagation period, while the policy events had a more direct and immediate impact on the social diffusion.

In the meantime, the above reasons can also explain the stable periods (2013–2016 and 2018–2019) of diffusion among online news. However, there are much less online news publishers than Twitter users and they have fixed frequencies as well as content for publishing. Thus, the number of online news per month is less volatile. Both policies and key events have less influence on its long-term development. Nevertheless, the instant effect of policies and crucial events at the specific month is still apparent (Figure 12).

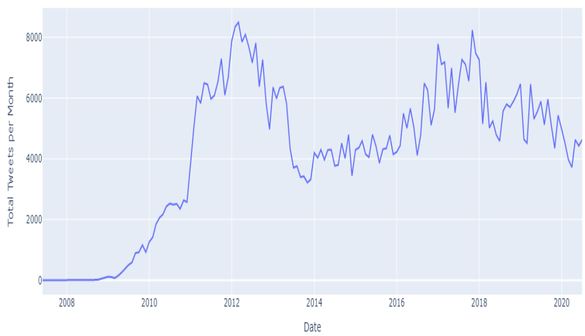

Figure 11. Number of Tweets per Month.

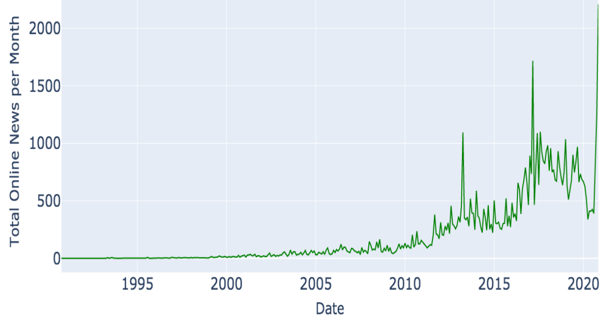

Figure 12. Number of Online News per Month.

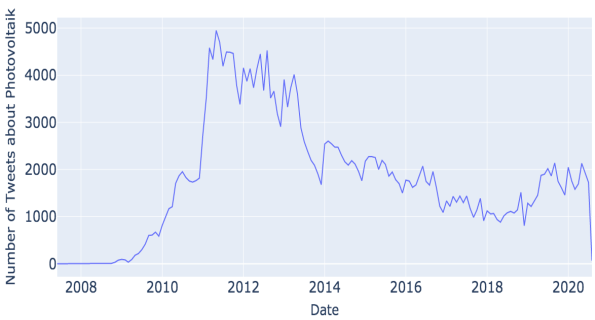

Figure 13. Number of Photovoltaic Tweets per Month.

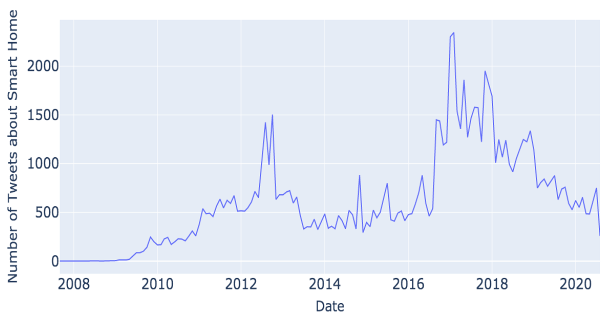

Figure 14. Number of Smart Home Tweets per Month.

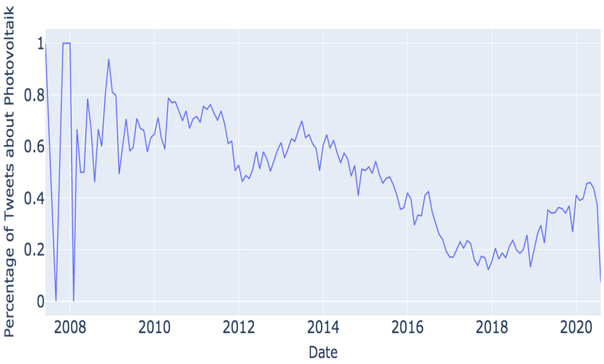

Figure 15. Percent of Photovoltaic in Twitter Corpus.

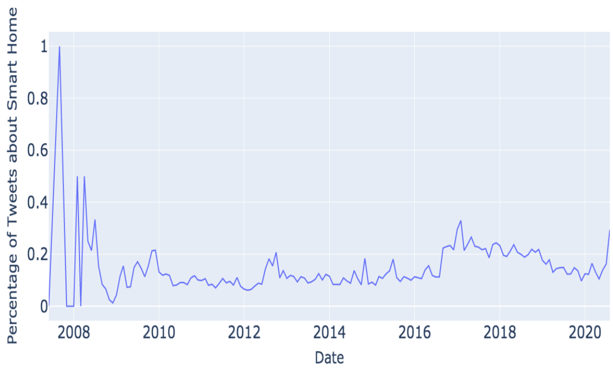

Figure 16. Percent of Smart Home in Twitter Corpus.

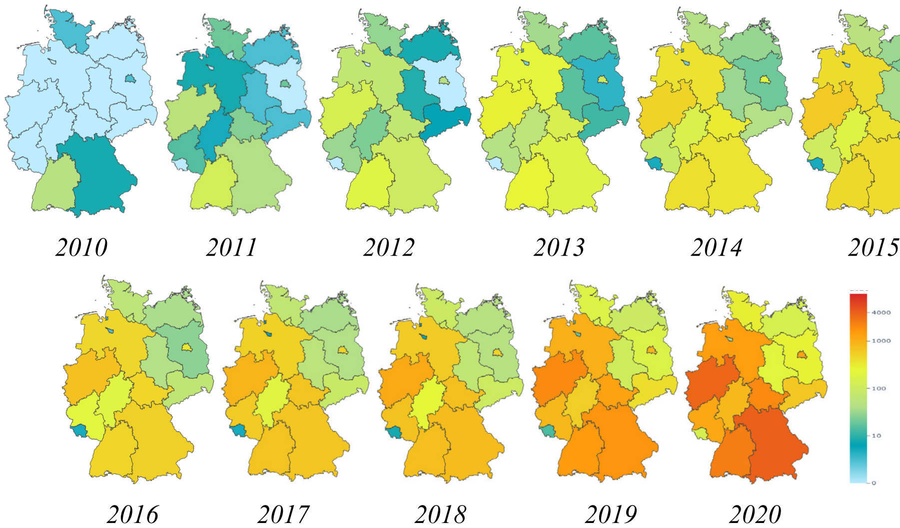

Figure 17. Number of Smart Building Tweets Distributed among 16 States per Year.

GEOGRAPHICAL DIFFUSION ANALYSIS

Since the majority of online news publishers in the Nexis Database are at the national or international level, the analysis of their locations does not represent the real geographical distribution of news content. Therefore, we only consider Twitter with geographical information into our diffusion analysis.

As depicted in Figure 15, before 2010, photovoltaics dominated the smart building concept. Thus, Baden-Württemberg and Bavaria, where have more solar resources and better industrial foundation, lead the concept diffusion. In particular, there are various solar research institutes located in Baden-Württemberg (e.g. Fraunhofer Institute for Solar Energy Systems ISE, DLR the Institute of Solar Research), which drives Baden-Württemberg as the first place. As the booming of photovoltaics boosted by EEG 2009, the differences among the whole of Germany is reducing. Nordrhein-Westfalen, Berlin and Niedersachsen, which have strong technology foundation, catch them up from 2011. With the growing percentage of smart home topic, Nordrhein-Westfalen even stood out from 2014, since it has a centralized digitalization policy, which strengthens their citizen usage of digitalization. On the contrary, with rich renewable energy resources, Schleswig-Hollstein still can maintain its second position in the first years. However, the slow digitalization process directly leads to its development lagging behind others from 2016.

During the whole concept diffusion from 2010 to 2020, the number of Tweets in Bremen and Saarland are always inferior to other states. Considering it may be caused by a population factor, we calculate the number of Tweets per ten million residents per year in Figure 18. Unfortunately, both of them still perform worse than others, which can be inferred that they are caused by their regional policies, instead of the size of the population. Going through the Operational Program 2014–2020 for the European Regional Development Fund (OP ERDF) among all the states in Germany, Bremen and Saarland pay more attention to economy, rather than innovation and sustainability. On the opposite, Thüringen lays more emphasis on innovation and sustainability. It not only refines their sustainability tasks in OP ERDF, but also publishes extra policy document encouraging intelligent technologies for sustainability, RIS3 Thüringen (Regional Research and Innovation Strategy for Smart Specialisation), in 2014, while other regions put forward their general innovation strategies in 2020. Hence, there is no doubt that Thüringen shines out in the smart building diffusion in Germany (Figure 18).

In conclusion, relative natural resources and industrial foundation play a vital role in the early stage of smart building con-

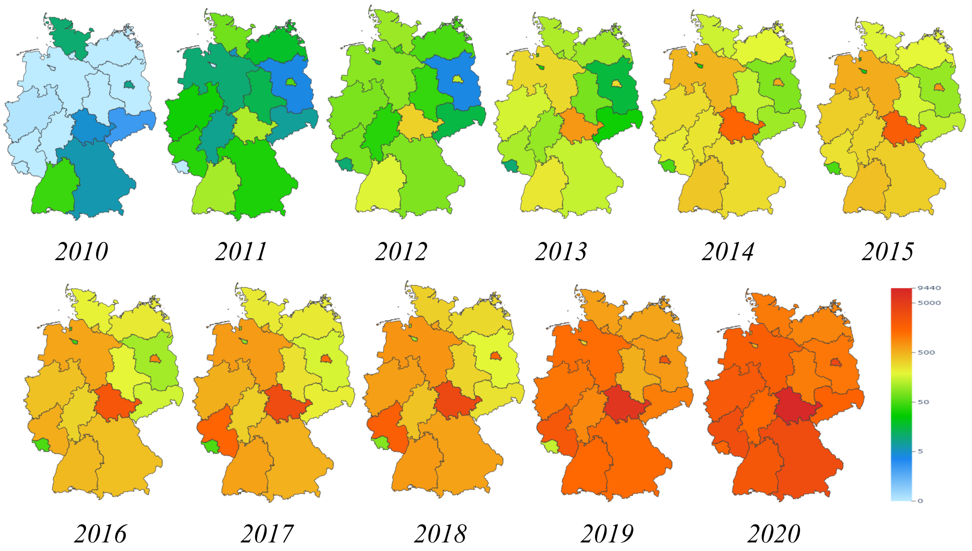

Figure 18. Number of Smart Building Tweets per Ten Million Residents Distributed among 16 States per Year.

Table 1. Top Ten User Accounts based on the Number of Smart Building Tweets.

Username	Tweets	Category	Topic
WorkInSolarJobs	24,894	Job advertisement	PV
NEUE_ENERGIE_VS	21,336	Start-up	PV
Solarserver	18,913	News	PV
photovoltaik	10,823	News	PV
spotHunter	10,565	News	PV
Solar_Reinigung	9,885	Start-up	PV
solenergie	9,731	News	PV
TeamHomeplaza	8,752	Start-up	Smart home & Smart grid
WorkInGreen_Job	8,587	Job advertisement	PV
GermanPersonnel	7,313	Job advertisement	Smart home & PV

cept diffusion. However, both national and regional policies have impacts that are more significant on its diffusion process. As the concept matures, differences among regions become less prevalent (Figure 18).

USER PROFILE ANALYSIS

To have a closer view of the degree of user engagement in the context of smart buildings, we finally analyze user profiles. In general, we have 84,579 users within the smart building concept. Most of them only have sent one (49,577 users) or two (12,569 users) Tweets about smart buildings, although the average number is 8.25. The range of Tweets per user is quite large. The most hardworking account has even sent around 25,000 Tweets for smart buildings.

All the top ten users (seen in Table 1) are public accounts for publishing news or job positions. They either post news for their own start-ups or the whole sector, especially in the photovoltaics domain. Although these accounts focusing on smart building send the most Tweets, it does not mean they have huge impact on its concept diffusion, since they have relatively few followers and thus a rather small reach. For example, user "WorkInSolarJobs" sent out the most Tweets only has 231 followers, no retweets and no favorites in total. As for top five accounts with most followers, all of them are news publishers. As shown in Table 2, their average retweets and favorites are reducing in terms of their follower numbers. However, they also do not play an important role in the user engagement, since their average retweets and favorites are also relative low.

An interesting finding is that four out of the top five accounts with the highest retweets of their posts are private users (Table 3), which means end users themselves play a prominent role in the user engagement. Without a strong tone of commercial promotion, interesting graphics or links can attract the public's attention easily. However, this situation also reflects that there is a lack of anchors in Germany in the smart building context. Neither big companies nor policy makers specifically work on this concept to nudge its diffusion towards end users. The average Retweets (0.3) and the average favorites (0.6) also point out that although there are plenty of Tweets about smart buildings, they stop spreading out maximum within one or two steps, rather than diffusing further.

Recommendations on User Engagement for Smart Buildings in Germany

Based on the results of our analysis, we derive four recommendations to nudge user engagement for smart buildings in Germany.

FOCUS MORE ON PROSUMAGER BUSINESS MODELS

Policy-oriented online news focus on energy production technologies and abstract concepts of energy efficiency. They widely ignore the biggest advantage of smart buildings in the energy transition: flexibility, which is contrarily the most popular topic discussed on Twitter. For end users financial aspects are very relevant, this is reflected by the enormous impact of the EEG in 2009 and the amendment of the EEG in 2012. Business models that leverage flexibility and set financial incentives could thus be beneficial.

Thanks to the maturity of photovoltaic, there is a tendency that buildings are converted from energy consumers to prosumers (i.e. both consumers and producers of energy). The role of prosumers can be further extended to prosumaging, which creates additional flexibility and convenience through managing and storing options. When policy makers or energy enterprises highlight the advantages of combining photovoltaics and electrical vehicle charging or intelligent energy management and provide corresponding business models, more and more end users would be attracted by smart buildings services. Policy makers can tackle these opportunities by more strongly supporting the development of prosumager business models and by emphasizing the advantages of prosumager business models towards the public.

SPEED UP THE PROVISION OF DIGITAL INFRASTRUCTURE

A good digital infrastructure provides the foundation for smart buildings. Regions with a weak digital infrastructure cannot follow the pace of adopting intelligent systems. Apart from providing crucial digital infrastructures, policy measures should also boost training for digital tools, in particular for relative practitioners (eg. installers, sellers and so on). Nudging user engagement towards intelligent systems needs to be based on technology which is easy to use for the end-user. But before these new technologies come to end-users, practitioners should be trained how they work out.

SUPPORTING START-UP AND NEW DIGITAL BUSINESS-MODELS

There are plenty of start-ups in both the photovoltaics and the smart home domain. However, they do not have significant influence on user engagement and need support to broaden their markets. Apart from the direct financial support, building up connections among start-ups, even with established enterprises, can also strengthen their power to attract end users, since it becomes more trustworthy than a single start-up.

PROVIDE ACCESSIBLE, FACT-BASED INFORMATION ON SOCIAL MEDIA

Although online news can interpret policies and crucial events, there is a delay and content gap between the publisher of the information and the end users. After all, news publication needs time and journalists have their own opinions, which can lead readers into their preferred directions. Twitter and other social media can break up this bottleneck. However, until now, only start-ups have made full use of social media platforms in the context of smart buildings. Governmental entities could set up official social media accounts to provide accessible and neutral information on smart buildings so that end-users can get access to reliable information and have more willingness to adopt the smart building concept.

Conclusion

With the objective to advance the state-of-the-art concept diffusion by integrating data-driven approaches, this paper proposes an analysis framework with iterative web scraping and visual analysis based on keyword extraction from Twitter and online news. We combine topic analysis with network visualization and sunburst chart, temporal analysis, geographical

Table 2. Top Five User Accounts based on the Number of Followers.

Username	Type	Followers	Number of Tweets	Retweets	Favorite	Avg. Retweets	Avg. Favorite
mashable	News	9,710,000	1	29	46	29	46
tagesschau	News	2,751,000	20	547	480	27.35	24
derspiegel	News	2,664,000	16	339	399	21.1875	24.9375
zeitonline	News	2,239,000	25	308	430	12.32	17.2
BILD	News	1,745,000	6	7	15	1.166666667	2.5

Table 3. Top Five User Accounts with the Highest Retweets of Their Posts.

Username	Type	Followers	Retweets	Favorite
FSHornschild	Private	3,809	2,335	3,465
labymod	Start-up	31,000	1,101	943
Schisslaweng	Private	7,123	1,060	2,587
ahoi_polloi	Private	9,244	843	1,914
FFriesecke	Private (Shareholder of start-up)	2,984	739	2,570

analysis based on small multiples of choropleth maps, and user profile analysis for a comprehensive user engagement analysis.

From a data perspective, we show in this paper that Twitter is an informative data source to track trends and concept diffusion in the public, especially representing real opinions and relationships of end users. It is actively driven by enterprises, particularly by start-ups. Our analysis finds that online news follows a different logic. They provide news, which are tightly related to policies or crucial commercial strategies and concepts therefore diffuse in a top-down manner.

From a methodology perspective, this paper introduces web scraping and visual analysis into the energy domain, in particular for the analysis of user engagement. First, our iterative web scraping approach based on keyword extraction can provide millions of user-generated data observations and hundreds of related keywords within a narrowed concept. This can largely broaden the data sources compared to more traditional analytical approaches, such as surveys. Compared to web scraping from scratch, our data collection method takes user experiences and expert knowledge from academia and the industry into consideration, which improves its accuracy. Second, we import three new types of visualizations for effectively and clearly displaying a complex concept and its diffusion. Small multiples of choropleth maps help us easily to compare the differences among regions at a specific timestamp and reveal how concepts gradually spread from one region to other regions. Network visualizations and sunbursts chart can greatly contribute to discovering all the relationships among hundreds of keywords within a complex concept and exploring its central topics. Third, inspired by time series analysis, our topic decomposition helps to explain the causes of the temporal diffusion of a complex concept. Fourth, a user profile analysis can accurately unveil the degree of user engagement without the bias introduced by commercial actors.

From a content perspective, this papers draws the following five conclusions: First, topic distributions of Twitter and online news are similar. Both of them put the emphasis on energy production and smart homes, whereby energy production takes the largest share. However, the represented keywords on Twitter are more up-to-date and innovative than those in online news, due to the time sensitivity and publishers of both platforms. Second, the concept of smart buildings has gradually changed from energy-efficient buildings to intelligently-controlled buildings. However, their combination, the intelligent energy management, has gotten less attention. Third, the temporal developments of Twitter and online news in Germany are quite different. The growth of the observations in online news is exponential, although there are two stable phases: from the beginning of 2013 to the beginning of 2016 and from 2018 to 2019. However, the number of Tweets has two clear peaks: March 2012 and November 2017, which are caused by policy measures and key commercial events respectively. Commercial events have a relative long propagation period, while policies always have direct impacts on the social diffusion from their effective dates. Compared to Twitter, both policies and key events have less influence on the long-term discussion about smart buildings in online news. Nevertheless, the instant effect of policies and crucial events in the specific month is still apparent. Fourth, relative natural resources and industrial foundation play a vital role in the early stage of smart building concept diffusion. Thus, Baden-Württemberg and Bavaria lead the concept diffusion. However, both national and regional policies have impacts that are more significant on the diffusion process of smart building concept so that Thüringen stands out in the smart building diffusion in Germany. The third factor is the prevalence of digitalization, which drives the success of Nordrhein-Westfalen. As the concept matures, the differences among regions become less prevalent. Fifth, there is a wide range of users publishing Tweets in the context of smart buildings. However, most of them do not engage in the conversation over a longer period of time. The users who are focusing on this domain over a longer period of time are public accounts, which publish news or job positions. But they do not have a huge impact on the user engagement. Surprisingly, end users themselves play a prominent role in the user engagement, which reflects that there is a lack of anchors in Germany in the smart building context.

Lastly, although our proposed approach has already brought various technical and thematic contributions, there are two areas in particular which could be explored further: First, selecting the keyword pairs with the highest keyword co-occurrences as topics and picking up the hubs by the degree of nodes could lead to ignore the promising keywords with less frequencies. Topic modelling as a clustering approach can solve this issue properly, which will be conducted in our future work. Second, user engagement in the concept diffusion is not enough for energy efficiency and usage flexibility. The user adoption is the final aim for smart buildings. In future we intend to consider sentiment analysis of Twitter and online news for further understanding the degree of user adoption of the corresponding technologies.

References

Baudier P, Ammi C, Deboeuf-Rouchon M. Smart home: Highly-educated students' acceptance [J]. Technological Forecasting and Social Change, 2020, 153: 119355.

Chuan C H, Tsai W H S, Cho S Y. Framing artificial intelligence in American newspapers [C]//Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. 2019: 339–344.

Federal Ministry of the Environment, Nature Conservation and Nuclear Safety. 2009 Amendment of the Renewable Energy Sources Act (EEG 2009). Retrieved from: https://www.iea.org/policies/4653-2009-amendment-of-the-renewable-energy-sources-act-eeg-2009 [accessed on May 22nd 2021].

Federal Ministry of the Environment, Nature Conservation and Nuclear Safety. 2012 Amendment of the Renewable Energy Sources Act (EEG 2012). Retrieved from: https://www.iea.org/policies/5117-2012-amendment-of-the-renewable-energy-sources-act-eeg-2012 [accessed on May 22nd 2021].

Gefen D, Larsen K R. Controlling for lexical closeness in survey research: A demonstration on the technology acceptance model [J]. Journal of the Association for Information Systems, 2017, 18 (10): 1.

Hojjati S N, Khodakarami M. Evaluation of factors affecting the adoption of smart buildings using the technology acceptance model [J]. International Journal of Advanced Networking and Applications, 2016, 7 (6): 2936.

Hootsuite. 2019. Social Media Statistiken 2019 in Deutschland. Retrieved from: https://blog.hootsuite.com/de/social-media-statistiken-2019-in-deutschland/ [accessed on May 22nd 2021].

Isenberg P, Isenberg T, Sedlmair M, et al. Visualization as seen through its research paper keywords [J]. IEEE Transactions on Visualization and Computer Graphics, 2016, 23 (1): 771–780.

Jivani A G. A comparative study of stemming algorithms[J]. Int. J. Comp. Tech. Appl, 2011, 2 (6): 1930–1938.

LexisNexis. Nexis. Retrieved from: https://www.lexisnexis.com/en-us/gateway.page [accessed on May 22nd 2021].

Minesoft Ltd and RWS. PatBase. Retrieved from: https://www.patbase.com/login.asp [accessed on May 22nd 2021].

Munzner T. Visualization analysis and design [M]. CRC press, 2014.

Panchalingam R, Chan K C. A State-Of-The-Art Review On Artificial Intelligence For Smart Buildings [J]. Intelligent Buildings International, 2019: 1–24.

Ribecca S. Network Diagram. Retrieved from: https://datavizcatalogue.com/methods/network_diagram.html [accessed on May 22nd 2021].

Rose S, Engel D, Cramer N, et al. Automatic keyword extraction from individual documents [J]. Text mining: applications and theory, 2010, 1: 1–20.

Ward M O, Grinstein G, Keim D. Interactive data visualization: foundations, techniques, and applications [M]. CRC Press, 2010.

Wasserman S, Faust K. Social network analysis: Methods and applications [J]. 1994.

Wirtz B W, Göttel V. Technology acceptance in social media: Review, synthesis and directions for future empirical research [J]. Journal of Electronic Commerce Research, 2016, 17 (2): 97.