Journal of Physics:
Conference Series

PAPER • OPEN ACCESS

Review of natural language processing techniques for characterizing positive energy districts

To cite this article: Mengjie Han *et al* 2023 *J. Phys.: Conf. Ser.* **2600** 082024

Review of natural language processing techniques for characterizing positive energy districts

Mengjie Han[1*], Juveria Shah[1], Xingxing Zhang[1]

[1] School of Information and Engineering, Dalarna University, 791 88 Falun, Sweden

* Corresponding author: mea@du.se

Abstract. The concept of Positive Energy Districts (PEDs) has emerged as a crucial aspect of endeavours aimed at accelerating the transition to zero carbon emissions and climate-neutral living spaces. The focus of research has shifted from energy-efficient individual buildings to entire districts, where the objective is to achieve a positive energy balance over a specific timeframe. The consensus on the conceptualization of a PED has been evolving and a standardized checklist for identifying and evaluating its constituent elements needs to be addressed. This study aims to develop a methodology for characterizing PEDs by leveraging natural language processing (NLP) techniques to model, extract, and map these elements. Furthermore, a review of state-of-the-art research papers is conducted to ascertain their contribution to assessing the effectiveness of NLP models. The findings indicate that NLP holds significant potential in modelling the majority of the identified elements across various domains. To establish a systematic framework for AI modelling, it is crucial to adopt approaches that integrate established and innovative techniques for PED characterization. Such an approach would enable a comprehensive and effective implementation of NLP within the context of PEDs, facilitating the creation of sustainable and resilient urban environments.

Keywords: positive energy districts, natural language processing (NLP), modelling, PED elements, NLP task

1. Introduction

There have been several initiatives and pilot projects focused on Positive Energy Districts (PEDs), but the path towards constructing 100 PEDs in Europe by 2025 remains intricate [1]. Designing an energy-efficient district, such as a PED, is a complex process that necessitates collaboration among multiple stakeholders who share a common goal. Numerous studies and pilot projects have already endeavored to develop the concept of PEDs and provide viable solutions to their existing challenges [2]. One crucial area of focus lies in the activities required to scale up or replicate successful PEDs. Since replicating a PED is not a straightforward task, maximizing the replication potential of a PED during the early design stages becomes essential [3]. By leveraging the characteristics of existing PEDs, tailored solutions can be developed for other local contexts. This necessitates gathering comprehensive scientific knowledge about PEDs and identifying the best practices for their operation. Advanced analytical methods can be employed to extract detailed characteristics of PEDs and create a virtual reference model for PEDs at a more granular level.

Text data for PED has seldom been analyzed. A study conducted by Hedman et al. [4] generated a keyword cloud using research papers focusing on PEDs or related topics. They revealed that current PED research predominantly concentrates on innovations at the building level, such as zero energy buildings (ZEBs), intelligent buildings, energy efficiency, and renewable energy sources (RES). In another study, Neumann et al. [5] employed 25 guidelines to structure their collection of opinions concerning spatial scale, target audience, and primary content. They created a series of keyword and topic clusters. These endeavors have provided valuable foundations for utilizing text mining techniques in characterizing PEDs. However, the specific algorithms employed in these studies remain undisclosed, and the potential contributions of natural language processing (NLP) techniques to PED characterization are still unknown. NLP offer the potential to analyze vast volumes of data and extract valuable insights regarding the performance of PEDs. These approaches have demonstrated success across various applications, encompassing load predictions, energy pattern profiling, regional energy consumption mapping, benchmarking building stocks, and analyzing the impact of retrofit strategies. However, to date, no research has been conducted on leveraging NLP methods to extract more detailed information about the characteristics of PEDs. Although NLP models are increasingly applied in numerous domains, the uniqueness of data, models, and model parameters used in most PED studies hinders the broad applicability of their findings.

2. PED elements

While numerous studies have emerged addressing the technical, economic, and social aspects of PEDs, there is limited research focusing on characterizing PEDs. Alpagut et al. [6] proposed a methodology for characterizing PEDs, involving the definition of indicators to measure their performance in geographical, financial, and social aspects. They also developed a "PED Solution Catalogue" comprising both technical and non-technical technologies. Their approach employed a thematic approach utilizing various data collection methods, such as surveys, interviews, and energy simulations. Zhang et al. [2] characterized 60 existing PED projects in Europe, identifying key features and challenges. Their qualitative content analysis approach revealed common themes, patterns, and trends across the projects. Furthermore, studies have explored specific aspects of PEDs, such as barriers, opportunities, and social and environmental justice implications. Despite various terminologies and focused elements, a need for clear, comprehensible, and structured definitions, including key performance indicators and system boundaries, has been emphasized to guide the design, operation, and assessment of PED programs. Therefore, integrating these common features and attributes is crucial to establish a comprehensive digital reference for characterizing PEDs. In this paper, three elements will be used for analysis: energy efficiency, renewable energy and environment. A literature search is made by associating NLP to the elements.

3. Natural Language Processing (NLP)

Natural Language Processing (NLP) is an interdisciplinary field that lies at the intersection of computer science, artificial intelligence, and linguistics. Its primary goal is to enable computers to understand, interpret, and generate human language in a way that is both meaningful and useful. NLP algorithms and techniques empower machines to process, analyze, and derive insights from vast amounts of natural language data, such as text documents, social media posts, emails, and spoken conversations. At its core, NLP involves breaking down human language into its constituent parts and extracting meaningful information from it. This process begins with basic tasks like tokenization, which involves splitting a text into individual words or phrases, and morphological analysis, which examines the internal structure of words to identify their root forms and derivations. NLP algorithms then move on to higher-level tasks such as part-of-speech tagging, syntactic parsing, and semantic analysis, which enable the understanding of sentence structure, grammar, and the relationships between words.

One of the fundamental challenges in NLP is the ambiguity inherent in natural language. Words and phrases can have multiple meanings depending on the context in which they are used. NLP techniques address this challenge through the use of statistical models and machine learning algorithms that learn from vast amounts of annotated data. These models help in disambiguating words and phrases by considering the surrounding words, the overall context, and the statistical patterns observed in the data. In recent years, NLP has been further advanced by the emergence of large-scale pre-trained language models like GPT-3. These models, trained on massive amounts of text data, have demonstrated the ability to generate coherent and contextually relevant responses, making significant strides in natural language generation.

NLP techniques can facilitate effective communication and collaboration among stakeholders involved in PED projects. This improves the accessibility of project-related information, streamlines communication processes, and enhances stakeholder engagement. NLP can assist in analyzing policy documents and regulatory frameworks relevant to PEDs. By processing and interpreting legal texts, NLP algorithms can identify key provisions, regulations, and obligations that pertain to energy efficiency, renewable energy, and sustainable urban development. This can help policymakers, researchers, and project developers gain insights into the regulatory landscape, understand compliance requirements, and inform policy decisions related to PED implementation and governance. NLP can also support the dissemination of knowledge and best practices within the PED domain. By analyzing and summarizing research papers, technical reports, and case studies, NLP algorithms can extract key information, generate concise summaries, and facilitate knowledge sharing among stakeholders. This promotes the transfer of valuable insights and lessons learned, contributing to the continuous improvement and replication of successful PED models. Thus, PED initiatives can benefit from enhanced decision-making, improved energy management, increased stakeholder engagement, and a more comprehensive understanding of energy-efficient urban development with the help of NLP.

4. Results

The keywords used for literature search is the combination of PED elements and NLP tasks. All relevant research papers were read and analysed to obtain the following results. As can be seen in Figure 1, the main NLP techniques that have been used for modelling PED elements are large language models, topic modelling and word2ved. These techniques are based on and trained on a massive amount of text data to understand, represent and generate human-like language. The functions comprise research trend analysis, sentiment classification, stakeholder matching, document comparison, utilization of building metadata, and load modeling. The applications focus on energy efficiency, energy production, and sustainability. However, there are still uncovered PED elements using NLP, which need to be further addressed. For example, occupant comfort is an important element that affects the overall performance of PED since it aims to improve the quality of life for individuals living or working within the district. An NLP method could be designed for utilizing social media or questionnaire data to improve user feedback.

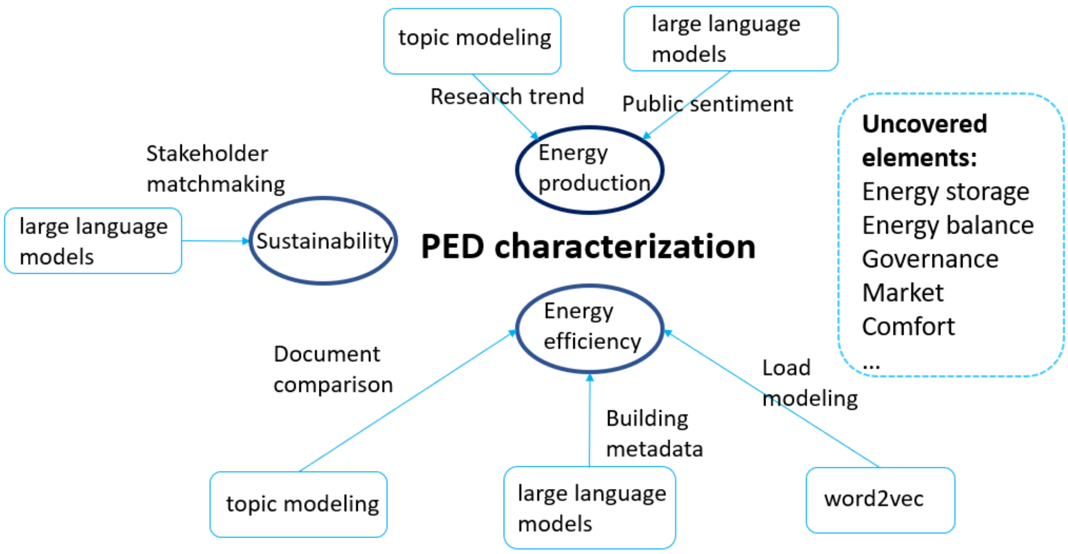

Figure 1. NLP techniques for characterizing PED elements

4.1. Energy efficiency
Research papers have been utilized not only as a corpus for rapid co-occurrence analysis but also as a primary data source for training word2vec models. For instance, Abdelrahman et al. conducted an analysis on a collection of 30,000 papers with the objective of exploring the interconnections between data science and energy efficiency across four categories: data, data science, energy efficiency, and phase [7]. They employed word2vec to represent each word within these categories as a 300-dimensional vector. The subsequent extraction of usability relations revealed that passive design, demand-controlled ventilation, model predictive controls, fault detection and diagnosis, and retrofit analysis frequently utilize data, whereas the same cannot be said for measurement and verification, and operation and maintenance. Moreover, the applications of Generative Adversarial Networks, dimensionality reduction, segmentation, and anomaly detection show promise in modeling energy efficiency. Additionally, another analysis involved clustering concepts based on word2vec, which helped identify key clusters related to data science. This approach was complemented by Latent Dirichlet Allocation (LDA) to quantify semantic relationships between keywords, facilitating the detection of communities of interest. Heating, ventilation, and air conditioning (HVAC) emerged as a central topic, while factors such as thermal environment, indoor illumination, and occupant behaviors were recognized as important considerations for modeling energy efficiency [8].

Improving energy efficiency is greatly facilitated by modeling building energy load, as it provides valuable insights into energy usage patterns and identifies areas where energy savings can be achieved. This information is crucial for pinpointing inefficient energy utilization and optimizing occupant comfort. A novel approach, known as energy2vec, was proposed based on the word2vec model. This approach utilized time series building load data, captured at one-minute intervals, along with appliance status as input [9]. By embedding vectors, the energy2vec model captured contextual information about the energy profile, effectively characterizing residents' habits and their appliance usage patterns. To account for different operating cycles of appliances, various sliding window lengths for word embedding can be applied, enabling the decomposition of load [10]. However, encoding categorical attributes and extracting the most relevant ones pose challenges for load prediction. To overcome this issue, a combination of word2vec and an attention-based Long Short-Term Memory (LSTM) model can be employed. This approach proves advantageous for medium- and long-term load forecasting, as it effectively addresses the encoding and selection of categorical attributes [11].

4.2. Renewable energy
The acceptance of renewable energy by the public plays a crucial role in determining the success of an energy transition. A positive attitude towards renewable energy not only drives increased demand and investment but also garners greater popular support for these sustainable technologies. In a study conducted by Kim et al., tweet data was leveraged to fine-tune the RoBERTa model, enabling a comprehensive understanding of public sentiment regarding the adoption of solar energy in the United States [12]. The findings revealed that states with consumer-friendly net metering policies and a well-established solar market exhibited more positive sentiments towards solar energy. Similarly, Jeong et al. analysed over 18,000 questions posted on Korea's largest portal site to identify public concerns surrounding renewable energy. They employed the TF-IDF technique to map central words and categories onto a word map [13]. Furthermore, a cosine similarity analysis based on word2vec was conducted to measure the similarity between words. The study confirmed that the public is primarily concerned with gaining an understanding of the characteristics, advantages, and disadvantages associated with different renewable energy resources. Other applications of NLP in the realm of renewable energy include comparing features between two countries using term extraction [14] and categorizing hydrothermal biomass conversion techniques using word-code matrices [15]. These NLP-driven approaches contribute to a better understanding of public perceptions and concerns, ultimately aiding in the effective promotion and deployment of renewable energy technologies.

4.3. Environment
PEDs have emerged as a prominent strategy for advancing sustainable urban development, particularly within sustainable energy systems. PEDs play a crucial role in reducing carbon emissions, enhancing energy security, and promoting resilience. In a study conducted by Saheb et al., a combination of topic modelling, BERT (Bidirectional Encoder Representations from Transformers), and clustering techniques was employed to gain insights into the research focus on AI-based sustainable energy [16]. By leveraging concatenated vectors from LDA (Latent Dirichlet Allocation) and BERT, the study revealed a significant emphasis on sustainable building design and the utilization of AI to optimize energy usage. At a sentence level, the fine-tuned BERT model showcased its capability to identify Sustainable Development Goals (SDGs) from various documents, enabling a predictive relational co-occurrence map of the SDGs [17]. Such a map facilitates matchmaking activities, allowing stakeholders to identify potential matches, bridge specific needs, and propose relevant solutions. Green buildings, known for their resource efficiency and reduced environmental impact, play a vital role in supporting sustainability.

5. Conclusions
This paper examines the application of natural language processing (NLP) techniques in characterizing Positive Energy Districts (PEDs). A review of research papers was conducted to explore how NLP methods can be utilized to model PED elements. NLP techniques, including topic modelling, word embedding, and training large language models like BERT, were identified as valuable tools for tasks such as stakeholder matching, sentiment analysis, and metadata analysis. By integrating the outcomes achieved by NLP models with a comprehensive analysis of PEDs' comparative performance using other relevant metrics, this approach brings us closer to realizing the vision of creating sustainable and resilient urban environments. The utilization of NLP methods facilitates decision-making and foster the development of effective strategies for sustainable urban development.

Acknowledgments
The authors thank the funding provided in part by the Joint Programming Initiative (JPI) Urban Europe framework. The work for this paper also receives funding support from the strategic innovation program 'Viable Cities' financed by Vinnova, the Swedish Energy Agency, and Formas (P2022-01000), The Scientific and Technological Research Center of Turkey (Türkiye), and Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMK). This paper is also funded by the Swedish Energy Agency with grant number 8569501, which is part of the IEA EBC Annex 83 positive energy district.

References
[1] Bossi, S.; Gollner, C.; Theierling, S. Towards 100 Positive Energy Districts in Europe: Preliminary Data Analysis of 61 European Cases. Energies 2020, 13, 6083, doi:10.3390/en13226083.
[2] Zhang, X.; Penaka, S.; Giriraj, S.; Sánchez, M.; Civiero, P.; Vandevyvere, H. Characterizing Positive Energy District (PED) through a Preliminary Review of 60 Existing Projects in Europe. Buildings 2021, 11, 318, doi:10.3390/buildings11080318.
[3] Uspenskaia, D.; Specht, K.; Kondziella, H.; Bruckner, T. Challenges and Barriers for Net-Zero/Positive Energy Buildings and Districts—Empirical Evidence from the Smart City Project SPARCS. Buildings 2021, 11, 78, doi:10.3390/buildings11020078.
[4] Hedman, Å.; Rehman, H.U.; Gabaldón, A.; Bisello, A.; Albert-Seifried, V.; Zhang, X.; Guarino, F.; Grynning, S.; Eicker, U.; Neumann, H.-M.; et al. IEA EBC Annex83 Positive Energy Districts. *Buildings* **2021**, *11*, 130, doi:10.3390/buildings11030130.
[5] Neumann, H.-M.; Garayo, S.D.; Gaitani, N.; Vettorato, D.; Aelenei, L.; Borsboom, J.; Etminan, G.; Kozlowska, A.; Reda, F.; Rose, J.; et al. Qualitative Assessment Methodology for Positive Energy District Planning Guidelines. In *Sustainability in Energy and Buildings*

2021; Littlewood, J.R., Howlett, R.J., Jain, L.C., Eds.; Smart Innovation, Systems and Technologies; Springer Singapore: Singapore, 2022; Vol. 263, pp. 507–517.
[6] Alpagut, B.; Akyürek, Ö.; Mitre, E.M. Positive Energy Districts Methodology and Its Replication Potential. In Proceedings of the Sustainable Places 2019; MDPI, July 22 2019; p. 8.
[7] Abdelrahman, M.M.; Zhan, S.; Miller, C.; Chong, A. Data Science for Building Energy Efficiency: A Comprehensive Text-Mining Driven Review of Scientific Literature. *Energy and Buildings* **2021**, *242*, 110885, doi:10.1016/j.enbuild.2021.110885.
[8] Ding, Z.; Liu, R.; Li, Z.; Fan, C. A Thematic Network-Based Methodology for the Research Trend Identification in Building Energy Management. *Energies* **2020**, *13*, 4621, doi:10.3390/en13184621.
[9] Nalmpantis, C.; Krystalakos, O.; Vrakas, D. Energy Profile Representation in Vector Space. In Proceedings of the Proceedings of the 10th Hellenic Conference on Artificial Intelligence; ACM: Patras Greece, July 9 2018; pp. 1–5.
[10] Nie, Z.; Yang, Y.; Xu, Q. An Ensemble-Policy Non-Intrusive Load Monitoring Technique Based Entirely on Deep Feature-Guided Attention Mechanism. Energy and Buildings 2022, 273, 112356, doi:10.1016/j.enbuild.2022.112356.
[11] Peng, J.; Kimmig, A.; Wang, J.; Liu, X.; Niu, Z.; Ovtcharova, J. Dual-Stage Attention-Based Long-Short-Term Memory Neural Networks for Energy Demand Prediction. *Energy and Buildings* **2021**, *249*, 111211, doi:10.1016/j.enbuild.2021.111211.
[12] Kim, S.Y.; Ganesan, K.; Dickens, P.; Panda, S. Public Sentiment toward Solar Energy—Opinion Mining of Twitter Using a Transformer-Based Language Model. *Sustainability* **2021**, *13*, 2673, doi:10.3390/su13052673.
[13] Jeong, S.-Y.; Kim, J.-W.; Joo, H.-Y.; Kim, Y.-S.; Moon, J.-H. Development and Application of a Big Data Analysis-Based Procedure to Identify Concerns about Renewable Energy. *Energies* **2021**, *14*, 4977, doi:10.3390/en14164977.
[14] Milia, M.F. Global Trends, Local Threads. The Thematic Orientation of Renewable Energy Research in Mexico and Argentina between 1992 and 2016. *JSCIRES* **2021**, *10*, s32–s45, doi:10.5530/jscires.10.1s.20.
[15] Acaru, S.F.; Abdullah, R.; Lai, D.T.C.; Lim, R.C. Hydrothermal Biomass Processing for Green Energy Transition: Insights Derived from Principal Component Analysis of International Patents. *Heliyon* **2022**, *8*, e10738, doi:10.1016/j.heliyon.2022.e10738.
[16] Saheb, T.; Dehghani, M.; Saheb, T. Artificial Intelligence for Sustainable Energy: A Contextual Topic Modeling and Content Analysis. *Sustainable Computing: Informatics and Systems* **2022**, *35*, 100699, doi:10.1016/j.suscom.2022.100699.
[17] Matsui, T.; Suzuki, K.; Ando, K.; Kitai, Y.; Haga, C.; Masuhara, N.; Kawakubo, S. A Natural Language Processing Model for Supporting Sustainable Development Goals: Translating Semantics, Visualizing Nexus, and Connecting Stakeholders. *Sustain Sci* **2022**, *17*, 969–985, doi:10.1007/s11625-022-01093-3.